Biostatistics (2000), **1**, 4, *pp*. 389–402
Printed in Great Britain

Estimating the immunity coverage required to prevent epidemics in a community of households

TOM BRITTON*

Department of Mathematics, Uppsala University, S-751 06 Uppsala, Sweden
tom.britton@math.uu.se

NIELS G. BECKER

National Centre for Epidemiology and Population Health,
Australian National University, Canberra ACT 0200, Australia

SUMMARY

An estimation of the immunity coverage needed to prevent future outbreaks of an infectious disease is considered for a community of households. Data on outbreak size in a sample of households from one epidemic are used to derive maximum likelihood estimates and confidence bounds for parameters of a stochastic model for disease transmission in a community of households. These parameter estimates induce estimates and confidence bounds for the basic reproduction number and the critical immunity coverage, which are the parameters of main interest when aiming at preventing major outbreaks in the future. The case when individuals are homogeneous, apart from the size of their household, is considered in detail. The generalization to the case with variable infectivity, susceptibility and/or mixing behaviour is discussed more briefly. The methods are illustrated with an application to data on influenza in Tecumseh, Michigan.

Keywords: Basic reproduction number; Disease transmission; Epidemic data; Herd immunity; Household community; Household outbreaks; Infection parameters; Maximum likelihood estimation; Preventing epidemics; Stochastic epidemic model; Vaccination coverage.

1. INTRODUCTION

The ultimate goal of any vaccination programme is to eliminate the target disease from the community and to prevent major outbreaks from subsequent importations of the disease. It is known that this is possible by reaching, and maintaining, a sufficiently high level of immunity in the community. The level of immunity required for success is called the *critical immunity coverage*. It depends on properties of the disease, characteristics of the community and how immunity is spread across the community. Previous estimates of the critical immunity coverage (see Anderson and May (1991) for a review) assume a population of uniformly mixing individuals and sometimes allow for age-specific transmission rates. Here we allow the community to consist of households. For many diseases it is important to acknowledge that human communities are structured into households, because of the greater contact rate between household members. We also take the extra step of furnishing estimates with confidence bounds. Other papers concerned with immunization policies in a community with households are, for example, Ball *et al.* (1997)

*To whom correspondence should be addressed

Table 1. *Frequency of outbreak sizes of influenza in households of various sizes*

Number	Number initially susceptible in the household				
infected	1	2	3	4	5
0	110	149	72	60	13
1	23	27	23	20	9
2		13	6	16	5
3			7	8	2
4				2	1
5					1
Total	133	189	108	106	31

and Hall and Becker (1996). A recent survey of statistical studies of infectious disease incidence is given in Becker and Britton (1999).

It is not always the smallest achievable critical immunity coverage that is of interest, because ethics and practical considerations often impose constraints. For example, some individuals may be opposed to vaccination, while others will insist on all household members being vaccinated. We therefore also estimate the critical immunity coverages for certain classes of immunization programmes.

Allowing for households in a community clearly makes it necessary to have data on outbreaks in households. Such data are not collected routinely. In particular, surveillance data do not contain details on household size and household outbreak size for registered cases. It is therefore also important to establish what kind of data are needed to estimate the critical immunity coverage for a community of households. We consider data comprised of outbreak sizes for a random sample of households. More specifically, our discussion is motivated by, and illustrated with reference to, the data on outbreaks of influenza in a sample of households in Tecumseh, Michigan, shown in Table 1. These data, taken from Addy *et al.* (1991), are part of a larger study and may not satisfy every assumption we make in our application but they serve well for both illustrating our method and demonstrating that it is practicable to obtain the requisite data.

The method for obtaining estimates of the critical immunity coverage consists of three steps. First transmission models need to be used to quantify the critical immunity coverage in terms of parameters of disease transmission and community structure. Then the various parameters that define the critical immunity coverage need to be estimated. Finally, inferences for the various parameters need to be translated to inferences about the critical immunity coverage.

Section 2 gives the assumptions made about disease transmission in the community and gives an expression for the reproduction number for infectives in this community setting. In Section 3 we address the question: who and how many individuals should be immunized to prevent future outbreaks in the community? Three different vaccination strategies are treated: *individuals* are selected randomly for vaccination, *households* are selected randomly for vaccination (with all members of selected households being vaccinated) and, finally, the optimal allocation of vaccinees, which aims at reducing the maximum number of susceptibles residing in a household. For each of the three strategies we quantify the critical immunity coverage. While we speak of vaccination strategies, it should be remembered that we are actually quantifying the critical immunity coverage. This may differ from the vaccination coverage required to prevent epidemics, because vaccines are not 100% effective in the field. Here we do not address the question of what vaccination coverage is required to achieve the desired immunity coverage. For a discussion of the effect of a variable vaccine response on the vaccination coverage required to prevent epidemics, see

Becker and Starczak (1998). The estimation of model parameters is considered in Section 4. In Section 5 we apply the methods to the data reported in Table 1.

Having discussed and illustrated the analysis for the case when individuals are homogeneous with regard to disease transmission, apart from different household sizes, we then turn our attention to heterogeneous individuals. In Section 6 we allow individuals to vary in infectivity, susceptibility and/or mixing rate by introducing different types of individual. In applications, such types might be different age groups, for example. We discuss how the estimation procedures can be extended, but omit details to avoid technicalities. A complication in the multi-type setting is that some parameters are not estimatable. Assumptions are needed to overcome this identifiability problem. The proposed method is illustrated briefly with reference to the data of Table 1, when individuals are classified into children and adults.

2. Model assumptions and notation

The choice of assumptions is guided by the model of Ball *et al.* (1997).

2.1. Disease transmission between and within households

An infective exerts a force of infection β on each susceptible household member for a period I, the random duration of the infectious period. The durations of the infectious period for different individuals are assumed to be independent. Each susceptible household member avoids infection from an infected household member with probability $\phi(\beta) = E[\exp(-\beta I)]$.

Suppose that n, the total number of households, is very large and let ν denote the mean number of individuals per household. Then the number of individuals in the community is νn. An infective exerts a force of infection $\lambda/\nu n$ on each individual of the community who is not in the same household. Each susceptible avoids infection by an infective, who is not a member of the same household, with probability $E[\exp(-\lambda I/\nu n)]$. As all infections occur from contacts with infectious individuals there have to be some initial infectives for an epidemic to occur.

To quantify the critical immunity coverage and the likelihood function corresponding to data of the form shown in Table 1, we need the probability distribution for the eventual size of household outbreaks. The dependence inherent in the above infection process makes this distribution inaccessible. We therefore adopt an assumption that reduces the dependence.

2.2. A pragmatic assumption

The probability that a susceptible avoids infection from outside the household, throughout the course of the epidemic, depends on the number infected in the community, which is random. We reduce the dependence in the infection process by adopting the assumption, made in Addy *et al.* (1991), that this probability is q, the same constant for each susceptible. The consequence of this assumption is that now all outbreak sizes in households are independent. It is known that this simplifying assumption is effective at enabling inferences to be made about the within-household transmission rate, in a way that makes allowance for the possibility of infection from outside the household. However, results established by Ball *et al.* (1997) show that in certain circumstances this assumption also enables inferences to be made about the rate of transmission between households. Specifically, they show that when the community consists of a large number of households and a large epidemic is initiated by a few infectious individuals, then the distributions of the final sizes of the two epidemics satisfy the same law of large numbers. This

holds true if the parameters q and λ of the two models are calibrated by the equation

$$q = e^{-\lambda\iota\tau}, \tag{1}$$

where τ is the proportion infected and $\iota = E[I]$, the mean duration of the infectious period. The expression for q given in (1) has a very natural interpretation. The law of large numbers implies that the number of individuals eventually infected is approximately $\nu n\tau$, and the probability of avoiding infection from all those individuals is $E[\exp(-\lambda(I_1 + \cdots + I_{\nu n\tau})/\nu n)] \approx \exp(-\lambda\iota\tau)$.

2.3. Final size distribution for household outbreaks

A way of computing the distribution for the final number infected in a household is given by Addy *et al.* (1991) under the pragmatic assumption that q is a constant. It is specified by the recursive formula

$$P_{s,a}(j) = \binom{s}{j}\phi[\beta(s-j)]^{a+j}q^{s-j} - \sum_{r=0}^{j-1}\binom{s-r}{j-r}P_{s,a}(r)\phi[\beta(s-j)]^{j-r}, \quad j = 0, \ldots, s \tag{2}$$

where $P_{s,a}(j)$ is the probability that, starting with s susceptible and a newly infected individuals, there are eventually j further infections. For the application we have in mind, most households will have no infectives initially, so our interest is primarily in $P_{s,0}(j)$.

To compute $P_{s,a}(j)$, all probabilities $\{P_{s,a}(r);\ r \leq j\}$ have to be computed. This can lead to numerical difficulties for very large 'household' sizes, say >100. For realistic household sizes the recursive formula produces numerical values quickly. However, in order to make an estimation, one has to maximize these expressions over some parameter space, and this may cause computational problems even for sizes relevant to households. For example, Addy *et al.* (1991) report computational difficulties for household sizes greater than five.

3. QUANTIFYING THE CRITICAL IMMUNITY COVERAGE

The key to specifying the critical immunity coverage is the existence of reproduction numbers, or epidemic threshold parameters, for large communities. The parameter R_0 is said to be a reproduction number for disease transmission in a community if the probability of a major epidemic is zero if $R_0 \leq 1$, and positive if $R_0 > 1$. Various epidemiologically meaningful reproduction numbers can be defined for a community of households, see Becker and Dietz (1996).

The most commonly used reproduction number relates to branching processes in the following way. In a large community the initial stages of an epidemic can, for many epidemic models, be approximated by a branching process. For the case of a community of households this approximation (Ball *et al.*, 1997) treats households as macro-individuals, and during the initial stages, when most households are uninfected, infected households infect other households independently of each other. The epidemic thus behaves approximately like a multi-type branching process where the type is specified by the size of the household. The event of a major outbreak corresponds to an infinite population in the branching process. It is well known (e.g. Jagers, 1975) that a branching process has positive probability of becoming infinite if and only if the largest positive eigenvalue of the matrix of mean off-spring distribution exceeds one. A natural definition of R_0 is thus to define it as the largest eigenvalue of the matrix with elements being the expected numbers of type-specific households one household of a given type infects (assuming all other households are uninfected).

To this end, let

$$\mu_s = \sum_{j=0}^{s-1} (j+1) P_{s-1,1}(j), \tag{3}$$

using the $P_{s-1,1}(j)$, defined by (2), for the case with $q = 1$ (i.e. no infection from outside). This makes μ_s the expected final number of cases, including the primary case, in a household starting with one primary case and $s-1$ susceptibles, and assuming that there is *no* external force of infection. On average, each of these individuals infects $\lambda\iota$ individuals outside the household. Among those so infected, a proportion $r\pi_r/\nu$ will reside in households having r susceptible individuals, where π_r denotes the proportion of *households* having r initially susceptible individuals and $\nu = \sum_r r\pi_r$ is the average household size. The matrix of expected number of infected households of various types is thus $\{\mu_s \lambda\iota r\pi_r/\nu\}$. Because each element of the matrix is a product of one factor depending on the row and the other depending on the column, the largest eigenvalue is simply the sum of the diagonal elements:

$$R_0 = \frac{\lambda\iota}{\nu} \sum_s s\mu_s\pi_s. \tag{4}$$

A more detailed derivation of R_0 is found in Ball *et al.* (1997).

The expression (4) is used to specify the critical immunity coverage. Although not essential, we assume that μ_s depends on the household size only through s, the number initially susceptible. That is, μ_s does not depend on the number of immune household members present, if any. Then the π_s are the only terms in (4) affected by vaccinating individuals. When individuals are vaccinated the mass of the distribution $\{\pi_s\}$ is shifted towards smaller values of s, thereby decreasing the value of the reproduction number R_0. The critical immunity coverage is specified by the level of immunity that gives $R_0 = 1$. We illustrate this with reference to three specific immunization schemes.

When calculating the critical immunity coverage we assume that all immunity is induced by vaccination. This is appropriate because when we have a vaccination strategy capable of preventing epidemics then there will be no significant immunity due to disease, and vaccination alone must achieve herd immunity.

3.1. Immunizing individuals at random

Assume first that the vaccine renders all vaccinees immune. Suppose each *individual* is, independently, selected for vaccination with probability v. This means that v is the immunity coverage, when the community is large. We refer to this as immunization strategy $\mathcal{I}$. With this strategy

$$\pi_s = \sum_{r \geq s} h_r \binom{r}{s} (1-v)^s v^{r-s},$$

where $\{h_r\}$ is the distribution of the household size in the community, and the reproduction number, now viewed as a function of v, becomes:

$$R(v) = \frac{\lambda\iota}{\nu} \sum_s s\mu_s \sum_{r \geq s} h_r \binom{r}{s} (1-v)^s v^{r-s}. \tag{5}$$

Note that when $v = 0$ this reproduction number coincides with the basic reproduction number, i.e. $R(0) = R_0$. The critical immunity coverage v_C is the solution of $R(v) = 1$ and may be written $v_C = R^{-1}(1)$.

Exactly the same equations apply for the alternative scenario in which every individual is vaccinated and each vaccination, independently, induces immunity with probability v and is a primary vaccine failure with probability $1 - v$.

3.2. Immunizing households at random

Again assume that the vaccine renders all vaccinees immune. Suppose now that a proportion v of all *households* is selected at random, and all individuals of selected households are vaccinated. 'At random' means that each household has the same probability v of being selected, not depending on size. This strategy is denoted $\mathcal{H}$. It is easy to see that the proportion of *individuals* vaccinated under this strategy is also v, giving immunity coverage v. However, vaccinees are distributed differently over households when compared to strategy $\mathcal{I}$.

The effect of strategy $\mathcal{H}$ on the number of susceptibles per household is expressed by

$$\pi_s = (1-v)h_s, \qquad s \geq 1,$$

since for each household size a proportion v of households now has no susceptibles. The reproduction number for this strategy has the expression:

$$R(v) = \frac{\lambda\iota}{\nu}\sum_s s\mu_s(1-v)h_s = (1-v)R(0) \tag{6}$$

which gives the explicit expression

$$v_C = 1 - 1/R(0)$$

for the critical immunity coverage (remember that v_C is the solution to $R(v) = 1$).

By use of Jensen's inequality, it can be shown that the critical immunity coverage under strategy $\mathcal{H}$ is generally higher than the coverage under strategy $\mathcal{I}$. Next we look at the optimal strategy, i.e. the strategy with smallest critical immunity coverage.

3.3. Immunizing optimally selected individuals

Again assume that the vaccine renders all vaccinees immune. The way to choose vaccinees optimally under the present assumptions is derived by Ball *et al.* (1997) for certain distributions of the infectious period, and is conjectured to hold for arbitrary distribution. This strategy, labelled $\mathcal{O}$, aims at reducing the largest number of susceptibles per household. If, hypothetically, individuals are chosen sequentially for vaccination, one starts by vaccinating one susceptible individual from each of the households with the largest number of susceptibles, m, say. The *maximum* number of susceptibles per household is thus reduced by one unit. If this is not enough to achieve $R_0 \leq 1$, one individual from all households with $m-1$ susceptible individuals (including the households selected in the previous step) is selected for vaccination. This selection procedure continues until the resulting reproduction number does not exceed one. One obtains a reproduction number exactly equal to one by vaccinating an individual from only a fraction p of the households at the last step.

Suppose this procedure has been performed so that the maximum number of susceptibles per household is now k. Then the resulting distribution of susceptibles in households can be written $\pi_s = h_s$, for $s < k-1$, $\pi_{k-1} = h_{k-1} + p\sum_{r\geq k} h_r$ and $\pi_k = (1-p)\sum_{r\geq k} h_r$. The corresponding reproduction number is

$$R(v) = \frac{\lambda\iota}{\nu}\left[\sum_{s<k-1} s\mu_s h_s + (k-1)\mu_{k-1}\left(h_{k-1} + p\sum_{r\geq k} h_r\right) + k(1-p)\mu_k\sum_{r\geq k} h_r\right]. \tag{7}$$

The critical immunity coverage is $v_C = [ph_k + (p+1)h_{k+1} + (p+2)h_{k+2} + \cdots]/\nu$ where p and k are obtained by solving the equation $R(v) = 1$ for k and then p.

4. PARAMETER ESTIMATION

For each strategy, a numerical value can be found for v_C once we specify values for the various parameters contained in the expression $R(v) = 1$. We are required to specify β, λ, ι and any other parameters of the distribution for the infectious period, and the distribution of the household size $\{h_r\}$. The equation $R(v) = 1$ does not contain the parameter q, but q plays an essential role in the estimation of other parameters through (1). We now consider how these parameters can be estimated from data.

The distribution of household size can often be obtained precisely from census data. In any event, estimation of the h_r is a straightforward problem and we simply assume that they are known.

For the estimation of other parameters suppose that we have a sample of households selected randomly from a community containing a large number of households. For each household in the sample we observe the eventual number of cases arising during the course of a major epidemic. Feasible estimation from such data requires a parametric form for the distribution of I, the length of the infectious period, whose moment generating function has an explicit expression. Data on the size of household outbreaks are not very informative about parameters of the distribution of I. We therefore recommend that the distribution chosen for I contains at most two unknown parameters, which we may choose to be $\iota = E[I]$ and σ_I, the standard deviation of the infectious period. Furthermore, we may fix ι at some value because the final size distribution depends only on the product $\beta\iota$, so that ι is not estimatable from such data.

Let n_{sj} denote the number households in the sample having initially s susceptibles of which eventually j become infected, $j = 0, \ldots, s$. The log-likelihood for such data is

$$\ell(\beta, \sigma_I, q) = \text{ constant } + \sum_{j,\, s} n_{sj} \log P_{s,0}(j). \tag{8}$$

Here it is assumed that the households in the sample contain no, or very few, initial infectives. This is not an essential assumption, but we make it because it simplifies the discussion, and in applications it is often the case.

In principle, maximizing the log-likelihood (8) is straightforward. In practice, this is not always easy because of the numerically unstable recursive formula (2) for the final size probabilities. For example, Addy *et al.* (1991) report numerical problems for household sizes exceeding five. An alternative approach to maximizing complicated expressions is to use Markov chain Monte Carlo (MCMC) methods. Recently O'Neill *et al.* (2000) used this approach on the Tecumseh data of Table 1 and obtained parameter estimates that coincide with the maximum likelihood estimates obtained by Addy *et al.* (1991) when the same models are treated. The MCMC approach seems likely to cope with larger household sizes.

Maximizing the likelihood (8) provides maximum likelihood estimates for β, σ_I and q. They are approximately normally distributed and their variance–covariance matrix can be estimated by the inverse of the observed Fisher-information matrix. We now wish to translate these inferences into inferences about the critical immunity coverage. For each vaccination strategy, v_C has the form $R^{-1}(1)$, where R depends on β, σ_I and $\lambda\iota$, and also on the strategy. To estimate $\lambda\iota$ we use (1) with the maximum likelihood estimate for q obtained from (8). It remains to specify τ, the proportion of individuals eventually infected. It may be that the eventual proportion of individuals infected in the community is also observed, but we do not assume this here. Instead, we note that (1) is an asymptotic relationship, so that τ can be written:

$$\tau = \nu^{-1} \sum_{j,\, s} j\pi_s P_{s,0}(j). \tag{9}$$

By substituting

$$\lambda\iota = \frac{-\log q}{\tau}, \tag{10}$$

with τ given by (9), into the expression for $R(v)$ we see that $R(v)$ depends only on the parameters β, σ_I and q, for which likelihood inferences are available via (8). By substituting the maximum likelihood

estimates of β, σ_I and q into $R(v)$, thus obtaining $\widehat{R}(v)$, we can solve for v to obtain an estimate $\widehat{v}_C$ of $v_C = R^{-1}(1)$.

A standard error for $\widehat{v}_C$ can be obtained by using the delta-method; see Rao (1973, p 388). First, the standard error for $\widehat{R}(\widehat{v}_C)$ is obtained by the delta method where $R(v)$ is differentiated with respect to β, σ_I and q, and using the observed Fisher-information matrix for these parameters. To get a standard error for $\widehat{v}_C = \widehat{R}^{-1}(1)$, we apply the delta method again and use the fact that the derivative of an inverse function is the reciprocal of the derivative of the function itself. More specifically, the standard error is s.e.$(\widehat{v}_C)$ = s.e.$[\widehat{R}(\widehat{v}_C)]/|\widehat{R}'(\widehat{v}_C)|$, where $\widehat{R}'(\widehat{v}_C)$ is the derivative of $R(v)$ with respect to v, evaluated at $v = \widehat{v}_C$ and replacing β, σ_I and $\lambda\iota$ by their estimates. Using the large sample normality of the estimate we can furnish the critical immunity coverage with a 95% upper confidence bound given by $\widehat{v}_C + 1.645$ s.e.$(\widehat{v}_C)$.

5. APPLICATION TO THE TECUMSEH INFLUENZA DATA

We now use the data presented in Table 1 to illustrate the estimation. The data are part of a larger study on respiratory illness; see Monto *et al.* (1985). They do not correspond in every detail to the format envisioned for our approach, but the main characteristics are present, so that a meaningful illustration is possible. The data consist of a 10% cross-sectional sample of households from the community which was followed prospectively, and each individual in the sample was, among other things, tested for antibodies before and after an influenza outbreak. At the end of Section 6.3 we return to the same data, only then admitting some individual heterogeneities.

A gamma distribution is a suitable description for the variation in the infectious period. As mentioned previously, the mean length ι cannot be separated from β. We fix its value at $\iota = 4.1$ days, this being a plausible value; see Elveback *et al.* (1976). It turns out that the likelihood varies little with the remaining parameter of the gamma distribution, as has been observed by Addy *et al.* (1991). The explanation for this is that only households with several infected individuals contain information on the infectious period, and in this data set only 21 of the 567 households have three or more eventual cases. We proceed as in Addy *et al.* (1991) by choosing the shape parameter equal to $k = 2$, so that the infectious period is modelled as the sum of two exponential variates, each with mean 2.05. This choice is very close to optimal but, as mentioned, the log-likelihood is very flat with respect to the parameter k. This distribution is considered as given, rather than estimated, in what follows. Its moment generating function is therefore completely specified, with $\phi(\alpha) = E[\exp(-\alpha I)] = 1/(1 + 2.05\alpha)^2$.

The log-likelihood (8) then depends only on the parameters β and q. Substituting the data from Table 1 (for example, $n_{42} = 16$) and maximizing with respect to β and q, using statistical software, gives the maximum likelihood estimates $\widehat{\beta} = 0.0446$ and $\widehat{q} = 0.8674$. These estimates of β and q enable us to estimate the μ_s via (2) and (3), giving $\mu_1 = 1$, $\mu_2 = 1.161$, $\mu_3 = 1.361$, $\mu_4 = 1.612$ and $\mu_5 = 1.924$.

We substitute these estimates for β and q into the observed Fisher information matrix, and then invert the matrix to obtain the estimated variance–covariance matrix:

$$I^{-1}(\widehat{\beta}, \widehat{q}) = 10^{-5} \begin{pmatrix} 9.41 & 1.20 \\ 1.20 & 5.01 \end{pmatrix}.$$

All estimates agree well with those obtained by Addy *et al.* (1991).

As mentioned, we would normally assign values to the $\{h_r\}$ from census data. We do not have access to the appropriate census data for Tecumseh, and use the data in Table 1 to assign values to the $\{h_r\}$ in this illustration. Taking the initial number of susceptibles to be the household size, we assign the values $h_1 = 133/567$, $h_2 = 189/567$, $h_3 = 108/567$, $h_4 = 106/567$ and $h_5 = 31/567$, giving the mean household size $\nu = 2.49$.

It remains to use (10) to estimate $\lambda\iota$. This requires an estimate of τ, the eventual proportion of individuals infected in the community. We can do this by use of expression (9), but a more direct way of

Table 2. *Estimates of the critical immunity coverage v_c for the data in Table 1 under various immunization strategies. The last two estimates are obtained when individuals have been categorized into two types, depending their age, and immunization consists of selecting households at random. See Section 6.3 for details*

Immunization strategy	Parameter estimates		
	$\widehat{v}_C$	s.e.$(\widehat{v}_C)$	95% upper bound
One-type model			
Individuals at random	9.0%	0.6%	10.1%
Households at random	11.9%	0.9%	13.5%
Optimally selected	6.5%	1.2%	8.4%
Two-type model			
Assuming $\lambda_{ij} = \lambda_j$	14.7%	0.8%	16.0%
Assuming $\lambda_{ij} = 0,\ i \neq j$	14.0%	–	–

obtaining an estimate is to use $250/1414$, the eventual proportion of individuals infected in our sample of households. Then (10) gives the estimate $\widehat{\lambda\iota} = 0.196$.

We are now ready to estimate the critical immunity coverage for each vaccination strategy. The results are summarized in Table 2.

Immunizing individuals at random

By substituting the various estimates into the equation $R(v) = 1$, with R given by (5), and solving for v we obtain the estimate $\widehat{v}_C = 0.0901$ for the critical immunity coverage under strategy $\mathcal{I}$. The delta method gives the standard error s.e.$(\widehat{v}_C) = 0.0064$. A 'safe' coverage for strategy $\mathcal{I}$ is therefore $\widehat{v}_C + 1.645\,\text{s.e.}(\widehat{v}_C) = 0.1010$. This means that if 10.1% of the community is immune we can be 95% confident that the community is not at risk of a major influenza outbreak.

Immunizing households at random

The same approach, but this time using the expression (6) for R, gives the estimate $\widehat{v}_C = 0.1191$, with standard error s.e.$(\widehat{v}_C) = 0.0095$. This gives a 'safe' coverage for strategy $\mathcal{H}$ of 13.5%, which is a little higher than that obtained for strategy $\mathcal{I}$, as expected.

Immunizing optimally selected individuals

Using the expression (7) for R, we estimate the optimal critical immunity coverage to be $\widehat{v}_C = 0.0635$, with standard error s.e.$(\widehat{v}_C) = 0.0123$, implying that a 'safe' coverage is 8.4%. This is achieved by immunizing no one in households of size three or smaller, and vaccinating individuals in households of size four and five, so that 43% of these household have three susceptible members and 57% have four susceptibles.

6. INDIVIDUALS OF DIFFERENT TYPE

We now briefly discuss how to extend the above methods to the situation where individuals may vary in susceptibility, infectivity and/or mixing patterns, by partitioning the individuals of the community into different *types*. In applications there is often a need to allow for age-specific differences, for example, but types might also be specified by gender, disease history, etc.

6.1. Multi-type model

We consider a natural extension of the model of Ball *et al.* (1997), presented in Section 2, in which parameters and variables of that section are furnished with a subscript to indicate type. Suppose there are k distinguishable types of individual, labelled $1, \ldots, k$. It is convenient to introduce vector notation. For example, let s_i denote the number of susceptibles of type i initially in a household. Then $\mathbf{s} = (s_1, \ldots, s_k)$ indicates the numbers of individuals of the various types that are initially susceptible in the household, and we refer to households of type $\mathbf{s}$. Transmission rates need two subscripts. An infective of type i exerts a force of infection β_{ij} on each susceptible household member of type j and also exerts a force of infection on each external susceptible of type j of $\lambda_{ij}/\nu n$, where (as before) νn is the total size of the community.

We proceed, as in Section 2.2, by making the pragmatic assumption that susceptible individuals of type i avoid infection from outside the household with probability q_i, which is *constant*, but now type-specific. This assumption makes the outcomes in households independent and provides a manageable likelihood function for inferences. Asymptotically, as the number of households of each size increases and a major epidemic is initiated by a small number of infectious individuals, we conjecture that the model with this simplifying assumption produces the same outbreak-size distribution for households as the multi-type version of the model of Ball *et al.* (1997), provided the parameters are calibrated by the equations

$$q_j = \exp\left(-\sum_i \alpha_i \tau_i \iota_i \lambda_{ij}\right), \qquad j = 1, \ldots, k, \tag{11}$$

the analogue of equation (1). Here α_i denotes the proportion of type-i individuals in the community.

With this pragmatic assumption, the model reduces to that considered by Addy *et al.* (1991), who give the analogue of equation (2) as

$$P_{\mathbf{s},\mathbf{a}}(\mathbf{j}) = \binom{\mathbf{s}}{\mathbf{j}} \prod_i \phi_i \left(\sum_k (s_k - j_k)\beta_{ik}\right)^{j_i + a_i} \mathbf{q}^{\mathbf{s}-\mathbf{j}} - \sum_{\mathbf{r}<\mathbf{j}} \binom{\mathbf{s}-\mathbf{r}}{\mathbf{j}-\mathbf{r}} P_{\mathbf{s},\mathbf{a}}(\mathbf{r}) \prod_i \phi_i \left(\sum_k (s_k - j_k)\beta_{ik}\right)^{j_i - r_i}, \mathbf{0} \le \mathbf{j} \le \mathbf{s},$$

where we have used the compact notation

$$\binom{\mathbf{a}}{\mathbf{b}} = \prod_i \binom{a_i}{b_i}, \qquad \mathbf{a}^{\mathbf{b}} = \prod_i a_i{}^{b_i} \qquad \text{and} \qquad \sum_{\mathbf{a}<\mathbf{b}} = \sum_{a_1 \le b1} \cdots \sum_{a_k \le b_k}$$

with strict inequality in at least one index.

6.2. Critical immunity coverage

We now quantify the immunity requirement for preventing major epidemics (more details in a related set-up are in Becker and Hall, 1996). Let $\mathbf{e}_j$ denote the k-vector whose jth element is unity, while all

other elements are zero. We define

$$\mu_{\mathbf{s}j} = \sum_{\mathbf{i}} (\mathbf{i} + \mathbf{e}_j) P_{\mathbf{s}-\mathbf{e}_j, \mathbf{e}_j}(\mathbf{i}),$$

which is the multi-type analogue of μ_s of equation (3). That is, $\mu_{\mathbf{s}j}$ denotes the *vector* giving the mean final size of each type in a household with one initial infective of type j and $\mathbf{s} - \mathbf{e}_j$ initial susceptibles of the various types, assuming there is no infection force acting from outside the household (i.e. computed with $\mathbf{q} = \mathbf{1}$). Let $\mu_{\mathbf{s}j}(m)$ denote the mth component of $\mu_{\mathbf{s}j}$ and $\pi_{\mathbf{s}}$ the proportion of all households having $\mathbf{s}$ initially susceptibles. Then

$$\kappa_{\mathbf{r}i,\mathbf{s}j} = \sum_{m=1}^{k} \mu_{\mathbf{r}i}(m) \frac{\iota_m \lambda_{mj}}{\nu} \pi_{\mathbf{s}} s_j \tag{12}$$

gives the mean number of households of type $\mathbf{s}$ with a primary type-j infective, to get infected by one $\mathbf{r}$ household whose primary infective was of type i, neglecting all other external forces of infection. Equation (12) defines a mean matrix $(\kappa_{\mathbf{r}i,\mathbf{s}j})$ for the multi-type setting and, by approximating the multi-type epidemic model by a multi-type branching process, one can show that the probability of a major outbreak is 0 if and only if R_0, the largest eigenvalue to this matrix, is ≤ 1.

As the quantification of the critical immunity coverage is closely tied to R_0, our task is very much facilitated by having an explicit expression for R_0. This makes the assumption that λ_{ij} is separable, in the form $\lambda_{ij} = \lambda_i \lambda'_j$, very attractive. This separability is also attractive for interpreting parameters, because λ_i becomes a relative infectivity for type-i individuals while λ'_j is a relative susceptibility for type-j individuals. The expression on the right-hand side of (12) then has the form

$$[\text{factor depending on row of } (\kappa_{\mathbf{r}i,\mathbf{s}j})] \times [\text{factor depending on column of } (\kappa_{\mathbf{r}i,\mathbf{s}j})],$$

and the largest eigenvalue of such a matrix is given by the sum of the diagonal terms. Therefore, the critical immunity coverage v_C is specified by the solution of

$$R_0 = \sum_{\mathbf{s}} \sum_{i} \kappa_{\mathbf{s}i,\mathbf{s}i} = \sum_{\mathbf{s}} \sum_{i} \sum_{m} \mu_{\mathbf{s}i}(m) \frac{\iota_m \lambda_m \lambda'_i}{\nu} \pi_{\mathbf{s}} s_i = 1, \tag{13}$$

where only $\pi_{\mathbf{s}}$ and s_i depend on the way immunity is distributed in the community. In particular, when a fraction v of households are selected for vaccination under strategy $\mathcal{H}$ we require

$$R(v) = (1 - v) \sum_{\mathbf{s}} \sum_{i} \sum_{m} \mu_{\mathbf{s}i}(m) \frac{\iota_m \lambda_m \lambda'_i}{\nu} h_{\mathbf{s}} s_i = 1,$$

giving

$$v_C = 1 - 1/R(0).$$

The class of immunization strategies of interest is richer when individuals are of different types, as one may now choose to vaccinate only certain types of individual. The optimal immunization strategy has no simple explicit algorithm in the multi-type case. There are several competing factors determining which choice is optimal. As before one should aim to immunize individuals from larger households, but one should also select individuals who are effective spreaders of the disease.

6.3. Estimation

The log-likelihood function under the pragmatic assumption is a straightforward multi-type version of (8). This enables an estimation of the β_{ij}, $\mathbf{q}$ and parameters of the distribution of the infectious period,

as described by Addy *et al.* (1991). We set the values of the ι_i to one, because they cannot be estimated separately from data of the assumed form. The values of the $h_{\mathbf{r}}$ and α_i are assumed to be available from census data. It remains to estimate the λ_{ij}. The estimate $\widehat{\mathbf{q}}$, together with equation (11), gives estimating equations for the λ_{ij}. However, there are only k equations for the k^2 parameters, implying an identifiability problem that seems to require a reduction in the number of parameters. Here we consider two ways of expressing the λ_{ij} in terms of fewer parameters that provide consistent estimates.

The first reduction assumes that for between-household transmission only susceptibility varies between individuals, i.e. we assume $\lambda_{ij} = \lambda'_j$. This is a particular case of the situation described above, where R_0 has the explicit expression (13) with each $\lambda_m = 1$. A consistent estimator for $\lambda'_j\iota$ is then obtained from (11) and given by $-\log(\widehat{q}_j)/\tilde{\tau}$, where $\tilde{\tau}$ is the overall proportion infected.

The second reduction is $\lambda_{ij} = 0$ if $i \neq j$, which implies that individuals mix only with individuals of the same type outside their own household. While this assumption is never entirely true, it can be a useful approximation for childhood diseases when types consist of age groups. It then follows from equation (11) that $\iota_j\lambda_{jj}$ is estimated consistently by $-\log(\widehat{q}_j)/\alpha_j\tilde{\tau}_j$. Under this assumption the elements in the mean matrix become $\kappa_{\mathbf{r}i,\mathbf{s}j} = \mu_{\mathbf{r}i}(j)\frac{\iota_j\lambda_{jj}}{\nu}\pi_{\mathbf{s}}s_j$. An estimate $\widehat{R}(0)$ is obtained by numerically deriving the largest eigenvalue of the matrix $(\kappa_{\mathbf{r}i,\mathbf{s}j})$, in which parameters are replaced by their estimates.

Parameter estimation requires more detailed data when heterogeneity between individuals is acknowledged. For the data of Table 1, Addy *et al.* (1991, Table 4) classify individuals into two types, namely children (ages 0–17 years) or adults. Accordingly, each household has a type determined by the number of children and the number of adults it contains. For each type of household, the frequency distribution of reported cases is given, although not in full detail in order to save space (the complete data has been obtained from Ira M. Longini). Using these data with the assumption that $\lambda_{ij} = \lambda'_j$ we obtain the estimate $\widehat{R}(0) = 1.173$, giving an estimated critical immunity coverage of $\widehat{v}_{\mathrm{C}} = 0.147$ under strategy $\mathcal{H}$. The standard error for the estimate is s.e.$(\widehat{v}_{\mathrm{C}})$ which has been derived using similar methods as those described for the case with homogeneous individuals. On the other hand, if we assume that children only have contact with children (and adults only with adults) outside their own household, the resulting estimates are $\widehat{R}(0) = 1.163$ and $\widehat{v}_{\mathrm{C}} = 0.140$, under strategy $\mathcal{H}$ (a standard error for $\widehat{v}_{\mathrm{C}}$ was not derived due to the complicated structure of $R(0)$). Note that both of these estimates are larger than 0.119, the estimate of v_{C} obtained under strategy $\mathcal{H}$ when we assumed individuals are homogeneous. It is not known if, under the present set-up, it is a general relation that neglecting heterogeneities produces underestimates of the critical immunity coverage. Such a relation holds when considering several other types of communities (Becker and Utev, 1998; Britton, 1998, e.g.).

A goodness-of-fit test for the Tecumseh data, to see if the model assuming homogeneous individuals explains the data adequately, was performed in Addy *et al.* (1991). The likelihood ratio statistic gives $\chi^2 = 21.81$ when the heterogeneous model is compared with the homogeneous model. There are four degrees of freedom because there are six estimatable parameters when there are two types, compared with two parameters in the case of only one type. The conclusion is hence that individual heterogeneity is significant and should be accounted for, which means that the estimates of v_{C} of the present section are more in agreement with data.

7. DISCUSSION

Our illustration with the Tecumseh data assumes that the community distribution of household sizes is the same as the distribution of susceptibles over households in the sample of data. Because the sample data contained some, but not many, initially immune individuals the community is expected to have slightly larger households, which affects R_0 and v_{C}. In practice, the distribution of those initially susceptible in households and the household sizes should be distinguished, and in case they are estimated, one should adjust standard errors to accommodate the extra uncertainty that this induces.

Methods of statistical analysis provide guidance on the requisite data. The present paper shows that quite informative inferences can be made about the critical immunity coverage from prospective studies of infectious diseases in a sample of households. The number of households in the Tecumseh study is large, but feasible in practice. The requisite number of households in the sample, and their preferred sizes, is likely to depend on how infectious the disease is, both within and between households. Data on who is and who is not initially susceptible to the disease are crucial.

Some assumptions simplified the discussion but are not needed to make the methods valid. For example, it is not necessary to assume that an individual exerts a constant force of infection over the duration of the infectious period. The recursive formula (2) actually holds when infectivity varies randomly, subsequent to infection, as long as the infectivity processes for different individuals are independent and identically distributed. The only change is that ϕ is defined differently. In particular, a latent period with arbitrary distribution has no effect on the distribution of the final size.

ACKNOWLEDGEMENT

We are grateful to Ira M. Longini and Owen D. Lyne for sending us the Tecumseh data in a form which made estimation possible in the multi-type setting, and the referees for suggesting several improvements of the paper. Support from the Swedish Natural Science Research Council and the Australian Research Council is gratefully acknowledged.

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

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

BECKER, N. G. AND DIETZ, K. (1996). Reproduction numbers and critical immunity levels for epidemics in a community of households. In HEYDE, C. C., PROHOROV, Y. V., PYKE, R. AND RACHE, S. T. (eds), *Athens Conference on Applied Probability and Time Series, Volume 1: Applied Probability*, Lecture Notes in Statistics **114**, 267–276.

BECKER, N. G. AND HALL, R. (1996). Immunization levels for preventing epidemics in a community of households made up of individuals of different types. *Mathematical Biosciences* **132**, 205–216.

BECKER, N. G. AND STARCZAK, D. N. (1998). The effect of random vaccine response to prevent epidemics. *Mathematical Biosciences* **154**, 117–135.

BECKER, N. G. AND UTEV, S. (1998). The effect of community structure on the immunity coverage required to prevent epidemics. *Mathematical Biosciences* **147**, 23–39.

BECKER, N. G. AND BRITTON, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society*, Series B **61**, 287–307.

BRITTON, T. (1998). On critical vaccination coverage in multi-type epidemics. *Journal of Applied Probability* **35**, 1003–1006.

ELVEBACK, L. R., FOX, J. P., ACKERMAN, E., LANGWORTHY, A., BOYD, M. AND GATEWOOD, L. (1976). An influenza simulation model for immunization studies. *American Journal of Epidemiology* **103**, 152–165.

HALL, R. AND BECKER, N. G. (1996). Preventing epidemics in a community of households. *Epidemiology and Infection* **117**, 443–455.

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

MONTO, A. S., KOOPMAN, J. S. AND LONGINI, I. M. JR. (1985). Tecumseh study of illness. XIII. Influenza infection and disease, 1976–1981. *American Journal of Epidemiology* **121**, 811–822.

O'NEILL, P., BALDING, D., BECKER, N., EEROLA, M. AND MOLLISON, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Applied Statistics*, in press.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

[*Received September 29, 1999*; *revised March 24, 2000*; *accepted for publication March 30, 2000*]