



Mathematical Statistics
Stockholm University

Estimating vaccine efficacy from small outbreaks

Niels G. Becker
Tom Britton

Research Report 2003:2

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

<http://www.matematik.su.se/matstat>



Estimating vaccine efficacy from small outbreaks

Niels G. Becker¹ and Tom Britton²

January 2003

Abstract

Let C_V and C_0 denote the number of cases among vaccinated and unvaccinated individuals, respectively, and let v be the proportion of individuals vaccinated. The quantity $\widehat{VE} = 1 - (1 - v)C_V / (vC_0) = 1 - (\text{relative attack rate})$ is the most used estimate of the effectiveness of a vaccine to protect against infection. For a wide class of vaccine responses, a family of transmission models and three types of community settings, this paper investigates what \widehat{VE} actually estimates. It does so under the assumption that the community is large and the vaccination coverage is adequate to prevent major outbreaks of the infectious disease, so that only data on minor outbreaks are available. For a community of homogeneous individuals who mix uniformly, it is found that \widehat{VE} estimates a quantity with the interpretation 1- (the mean reduction in susceptibility induced by the vaccine). We provide a standard error for \widehat{VE} in this setting. For a community with some heterogeneity \widehat{VE} can be a very misleading estimate of the effectiveness of the vaccine. When individuals have inherent differences, \widehat{VE} estimates a quantity that depends also on the inherent susceptibilities of different types of individual and on the vaccination coverage for different types. For a community of households, \widehat{VE} estimates a quantity that depends on the rate of transmission within households and on the reduction in infectivity induced by the vaccine. In communities that are structured, into households or age-groups, it is possible that \widehat{VE} estimates a value that is negative even when the vaccine reduces both susceptibility and infectivity.

KEY WORDS: Epidemics; Estimation; Infectivity; Measles outbreak data; Multi-type branching process; Susceptibility; Vaccine efficacy.

¹National Center for Epidemiology and Population Health, Australian National University, Canberra ACT 0200, Australia. E-mail: niels.becker@anu.edu.au

²Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tomb@matematik.su.se.

1 Introduction

An ideal vaccine induces complete protection against infection in every vaccinee. In practice, some vaccinated individuals become cases, so this ideal is not achieved. It is therefore important to have measures of vaccine efficacy and methods for estimating these from data in a variety of settings. Here we consider such estimation from data on outbreaks in a community having a vaccination coverage that is adequate to prevent major epidemics. For example, countries such as Finland and some Pan-American countries have achieved vaccination coverages adequate to eliminate measles. Current measles incidence in such countries is due to minor outbreaks arising from importation of infected individuals and the subsequent chains of transmission. De Serres *et al.* [9] propose the use of such data for monitoring the process of elimination for the disease.

It may seem to be of minor importance to estimate the effectiveness of a vaccine when it is effective enough to achieve the state of elimination. However, it is thought that immunity may wane over time unless it is occasionally boosted by exposure to disease and it is therefore quite important to monitor the effectiveness of a vaccine when the opportunity for exposure to disease has been reduced dramatically.

We introduce our assumptions about the transmission of infection for a community of homogeneous susceptible individuals who mix uniformly in Section 2. Later sections allow individuals to differ and allow the community to be structured into households. For a more thorough presentation of these types of models see, for example, Andersson and Britton [1]. In Section 3 we describe a model for the way individuals respond to vaccination, in terms of how susceptible to infection vaccinated individuals are and how infectious they are, should they become infected. Some measures of vaccine efficacy, based on mean responses, are given. The rest of the paper investigates what

$$\widehat{VE} = 1 - \frac{C_V/v}{C_0/(1-v)}, \quad (1)$$

which is used very widely in epidemiological studies to estimate 'vaccine efficacy', estimates in our model formulations. Here v is the known vaccination coverage, C_V is the observed number of vaccinated cases and C_0 is the observed number of unvaccinated cases. In Section 4 we look at this estimate for a community of homogeneous individuals who mix uniformly. We find that \widehat{VE} estimates a measure of vaccine efficacy that quantifies the reduction in susceptibility induced by the vaccine and we provide a standard error for the estimate corresponding to this setting. In the derivation of the standard error we obtain expressions, interesting in their own right, for the variance-covariance matrix of the total progeny of a multitype Galton-Watson branching process. Section 5 proposes a way to combine data from a number of minor outbreaks and the method is applied to measles outbreak data from Victoria, Australia. In Section 6 we show that, in a heterogeneous community, \widehat{VE} estimates a quantity that depends on the relative susceptibility, both inherent and vaccine-induced, of different types of individual and on the vaccination coverages of different types. In Section 7 we show that in a community of households \widehat{VE} estimates a quantity that depends on the reduction in susceptibility induced by the vaccine, but also on the rate of transmission within households and the reduction in infectivity induced by the vaccine.

The importance of the results lies in showing that in this setting, with homogeneous

individuals, the widely-used \widehat{VE} gives a useful estimate of vaccine efficacy for a wide class of vaccine responses. This does not happen generally; see Smith *et al.* [16]. It is also important to point out that in many real-world settings \widehat{VE} can seriously mislead us and to quantify this bias.

Throughout the paper we consider a closed community of n individuals, where n is assumed to be large.

2 Transmission of infection

We begin by outlining the assumptions made about transmission of infection, with reference to a community of homogeneous individuals who mix uniformly. Assume, for the moment, that there is no vaccine-induced immunity in the community. Individuals may be susceptible (S) to infection, infected (I) or removed (R). Removed individuals are those who acquired immunity after recovering from infection. Define an infectious contact by an infected individual as one that is sufficiently close for the infection to be transmitted if the contacted person is a susceptible individual. Following infection, an infected individual has infectious contacts with other individuals at a rate depending on the progress of the infection within the infected individual. We describe this rate by an infectiousness function $\beta(x)$, where x is the time since infection; see Becker [5, Section 3.1]. The infection has a latency period if $\beta(x) = 0$ for small x . For larger x the infectiousness function typically rises to a peak after which it declines to 0, at which time the individual becomes a removal. In general, $\beta(x)$ is any individual-specific non-negative stochastic process defined for positive x . Infectiousness functions corresponding to different individuals are assumed to be mutually independent. It is known, see for example Ball and Clancy [2], that the final outcome of SIR epidemic models depend on $\beta(x)$ only through the distribution of $B_0 = \int_0^\infty \beta(x)dx$, the total area under the infectiousness function. In particular, a latency period does not affect the distribution of the final size of the epidemic. In the literature, a common choice for $\beta(x)$ is

$$\beta(x) = \beta, \quad 0 \leq x \leq T$$

where β is a constant. Setting $T \equiv 1$ gives the *Reed-Frost* model in continuous time, for which $B_0 \equiv \beta$, while $T \sim \text{Exp}(1)$ gives the *general stochastic epidemic*, for which $B_0 \sim \text{Exp}(1/\beta)$; see for example [1].

As mentioned, a susceptible individual who gets infected has infectious contacts at a rate depending on his/her infectiousness function. When the infected individual mixes uniformly with other community members each individual is equally likely to be contacted at any point in time, so the contact rate exerted on a specific individual x time units after infection is $\beta(x)/n$. Contacts with individuals who are already infected or removed have no effect. The epidemic process continues until every infected individual has become a removal. The final size of the epidemic is given by the number of individuals infected during the course of the epidemic.

3 Response to vaccination

The response to vaccination consists of a change in the way an individual's immune system responds to invasion by, and internal development of, the infection agent. Our concern lies with the way this is reflected by a change in the susceptibility to infection and a change in the individual's infectiousness function. We first describe the effect of vaccination on susceptibility. At time t , let the probability that a susceptible individual makes an infectious contact in the small time increment $[t, t+dt)$ be given by $\lambda_t dt$. Here λ_t is the sum of the values of all infectiousness functions at time t , divided by n . The probability that the individual is infected, in this time increment, is $\lambda_t dt$ if he is unvaccinated and $A\lambda_t dt$ if he is vaccinated. The quantity A is the proportionate reduction in the probability that a vaccinated individual is infected, per contact, relative to an unvaccinated individual. The fact that individuals respond differently to vaccination is accommodated by taking A to be a random variable.

It may be that an individual in whom vaccination induces an immunological response acquires a milder form of the illness, if ever infected, making them less infectious. Recall that the potential an infected individual has to infect others is captured by the total area under the infectiousness function. To quantify a possible reduction in potential to infect others let B_0 denote this area when the infected individual is unvaccinated and $B_V = BB_0$ when the infected individual is vaccinated. The relative infection potential B lies between zero and unity when vaccination reduces infectivity. In our formulation B is taken to be random, to allow for variation in vaccine response, and for simplicity assumed to be independent of B_0 . Note that a value of $B < 1$ can arise in many ways. Perhaps the duration of the infectious period or the level of infectiousness achieved at different times, or both, are reduced by vaccination.

The effect of vaccination on susceptibility *and* infectivity is described by the joint probability distribution

$$\Pr(A = a_i, B = b_i) = p_i, \quad i = 1, 2, \dots, k. \quad (2)$$

This discrete formulation can provide a good approximation to any realistic distribution by suitable choices of k , a_i , b_i and p_i . It is likely that A and B are correlated in practice.

There are several different ways to measure the efficacy of a vaccine; see Halloran *et al.* [11] for a survey. For example, $VE_S = 1 - E(A)$ measures the reduction in susceptibility, $VE_I = 1 - E(B)$ measures the reduction in infectivity and $VE_{SI} = 1 - E(AB)$ is a combination of the reduction in susceptibility and infectivity. The measure of *vaccine efficacy* given by

$$VE = 1 - \frac{\text{prob(a vaccinated individual becomes infected)}}{\text{prob(an unvaccinated individual becomes infected)}} = 1 - \frac{E(C_V)/v}{E(C_0)/(1-v)} \quad (3)$$

is widely used in epidemiological studies; see Orenstein, Bernier *et al.* [13, 14]. A disadvantage of the measure VE is that it is not clear how it relates to the parameters $\{a_i, b_i, p_i\}$ which describe the effect of vaccination on disease transmission. The popularity of VE stems from its simple interpretation and the fact that it has an obvious estimator, namely

$$\widehat{VE} = 1 - \frac{\text{proportion of cases among vaccinated individuals}}{\text{proportion of cases among unvaccinated individuals}} = 1 - \frac{C_V/v}{C_0/(1-v)}.$$

Below we express VE in terms of model parameters for three community settings with the aim of determining what \widehat{VE} estimates in those setting.

4 Community of homogeneous individuals

Consider a large population of homogeneous individuals who mix uniformly and who are all susceptible to a certain infectious disease, with a transmission process as described in Section 2. The expected number of infectious contacts an initial infected individual has during the course of his infection is $R_0 = E(B_0)$, where B_0 is defined in Section 2. The reproduction number of infectives is less than R_0 , the basic reproduction number, when some community members have acquired immunity from previous exposure to the infection or by vaccination.

Now suppose that a fraction v of individuals are vaccinated and that their responses, in terms of susceptibility and infectivity, vary independently in the manner described in Section 3. Individuals are then of type 0, 1, 2, ..., k , where label 0 refers to unvaccinated individuals and 1, 2, ..., k refer to vaccinated individuals with different responses to vaccination. An outbreak in such a population can be described by a multitype epidemic model, where type is specified by vaccination status and vaccine response. Type- i individuals have susceptibility a_i and infectivity b_i , with $a_0 = b_0 = 1$ by definition. Then π_i , the proportion of community members of type i , is given by

$$\pi_0 = 1 - v \quad \text{and} \quad \pi_i = v p_i \quad \text{for } i = 1, 2, \dots, k.$$

For this setting, the reproduction number after vaccination is given by

$$R_V = R_0 \sum_{i=0}^k a_i b_i \pi_i = R_0 [1 - v + v E(AB)], \quad (4)$$

see for example Becker and Marschner [7], Becker and Starczak [8], and below. Assume that the vaccination coverage is high enough to prevent epidemics, that is, assume $R_V \leq 1$. Then the size of an outbreak initiated by a small number of introductory cases can be approximated by the total progeny in a branching process, with infections interpreted as births; see Ball and Clancy [2]. A heuristic argument for this approximation is that, when few individuals are infected it is unlikely that a contact occurs with a previously infected individual. As a consequence, each infectious individual infects new individuals essentially independently, and whenever individuals 'give birth' to new individuals independently we are in the branching process paradigm. Let $C_i^{(n)}$ denote the eventual number of type- i individuals who are infected in a community of size n , $i = 0, \dots, k$. We approximate the distribution of these random variables by the corresponding random variables of a multitype branching process, using C_i to denote the total progeny of type i . Below we drop the index n and interchangeably mean the epidemic and the approximating branching process, the approximation relying on n being large. We are interested in the distribution of

$$\mathbf{C} = \begin{pmatrix} C_0 \\ C_1 \\ \vdots \\ C_k \end{pmatrix} = \mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2 + \cdots ,$$

where \mathbf{Z}_t denotes the vector of the various types of infected individuals in generation t . Generation 0 consists of the primary cases of the outbreak. Generation 1 consists of the individuals infected by the primary cases, and so on.

With the aim of obtaining an expression for the VE of (3), we derive an expression for the mean of \mathbf{C} , namely

$$E(\mathbf{C}) = E(\mathbf{Z}_0) + E(\mathbf{Z}_1) + E(\mathbf{Z}_2) + \cdots , \quad (5)$$

which gives the expected number eventually infected for each of the various types of individual. To this end we introduce the matrix $\mathbf{M} = (m_{ij})$, where

m_{ij} = mean number of type- i individuals infected by a single type- j infective.

From the formulation for the transmission and the vaccine response it follows that

$$m_{ij} = R_0 a_i b_j \pi_i.$$

To see this note that the infection rate between an infectious type- j individual and a given susceptible type- i individual is $a_i b_j \beta(x)/n$, x time units after infection of the former. The probability of infection for this pair is therefore $1 - E\left(e^{-a_i b_j \int \beta(x) dx/n}\right) = 1 - E\left(e^{-a_i b_j B_0/n}\right) \approx E(a_i b_j B_0/n)$. Since there are $n\pi_i$ type- i individuals the expected number of infectious contacts with type- i individuals is $a_i b_j E(B_0/n)n\pi_i = m_{ij}$. In matrix form, \mathbf{M} satisfies

$$\mathbf{M} = R_0 \begin{pmatrix} \pi_0 & b_1 \pi_0 & \cdots & b_k \pi_0 \\ a_1 \pi_1 & a_1 b_1 \pi_1 & \cdots & a_1 b_k \pi_1 \\ \vdots & \vdots & \ddots & \vdots \\ a_k \pi_k & a_k b_1 \pi_k & \cdots & a_k b_k \pi_k \end{pmatrix} = \begin{pmatrix} \pi_0 R_0 \\ a_1 \pi_1 R_0 \\ \vdots \\ a_k \pi_k R_0 \end{pmatrix} \begin{pmatrix} 1 & b_1 & \cdots & b_k \end{pmatrix} = \mathbf{d}_v \mathbf{b}^T, \text{ (say).}$$

The subscript in \mathbf{d}_v indicates that this formulation is for a community of uniformly-mixing individuals.

We now return to the expected final size of the epidemic given in Equation (5). Using $E(\mathbf{Z}_{i+1}|\mathbf{Z}_i) = \mathbf{M}\mathbf{Z}_i$, we find $E(\mathbf{Z}_i) = \mathbf{M}^i E(\mathbf{Z}_0)$ and

$$E(\mathbf{C}) = (\mathbf{I} + \mathbf{M} + \mathbf{M}^2 + \cdots)E(\mathbf{Z}_0) = (\mathbf{I} - \mathbf{M})^{-1}E(\mathbf{Z}_0), \quad (6)$$

where the last equality requires the largest eigenvalue of \mathbf{M} to be less than 1. The largest eigenvalue of \mathbf{M} is $R_V = R_0 \sum_i a_i b_i \pi_i = \mathbf{d}_v^T \mathbf{b}$ as noted in Equation (4). The assumption $R_V < 1$ simply assures that a major outbreak cannot occur. Further, it is easily verified that the form for \mathbf{M} implies that

$$(\mathbf{I} - \mathbf{M})^{-1} = \mathbf{I} + \frac{1}{1 - \mathbf{d}_v^T \mathbf{b}} \mathbf{M} = \mathbf{I} + \frac{1}{1 - R_V} \mathbf{M}.$$

A natural assumption is that the types of the initial infectives depend on their relative susceptibility and their abundance. That is, $E(\mathbf{Z}_0) = c \mathbf{d}_U$ is plausible, where c is some constant. A little algebra then shows that

$$E(\mathbf{C}) = (\mathbf{I} - \mathbf{M})^{-1} E(\mathbf{Z}_0) = c \left(\mathbf{I} + \frac{1}{1 - R_V} \mathbf{d}_U \mathbf{b}^T \right) \mathbf{d}_U = \frac{c}{1 - R_V} \mathbf{d}_U. \quad (7)$$

We now use this to obtain an expression for the vaccine efficacy VE in terms of model parameters. Note that

$$\begin{aligned} VE &= 1 - \frac{\text{prob}(\text{a vaccinated individual becomes infected})}{\text{prob}(\text{an unvaccinated individual becomes infected})} \\ &= 1 - \frac{\text{expected proportion of vaccinated individuals who become infected}}{\text{expected proportion of unvaccinated individuals who become infected}} \\ &= 1 - \frac{E(C_1 + \dots + C_k)/nv}{E(C_0)/n(1-v)} \\ &= 1 - \frac{E(C_V)/v}{E(C_0)/(1-v)} \end{aligned}$$

where $C_V = C_1 + \dots + C_k$ is the number of vaccinated individuals who become cases. Using

$$E(C_0) = (1, 0, \dots, 0) E(\mathbf{C}) \quad \text{and} \quad E(C_V) = (0, 1, \dots, 1) E(\mathbf{C}),$$

and Equation (7) in the expression for VE we get

$$VE = 1 - \sum_{i=1}^k a_i p_i = 1 - E(A) = VE_S.$$

We have shown that if enough individuals are vaccinated in a community of homogeneous individuals, thus ruling out the possibility of a major outbreak, then VE is equivalent to VE_S which is a measure of the reduction in susceptibility due to vaccination.

The result is important in two ways. First, it means that in this setting VE has a clear and useful interpretation as a measure of the protection vaccination provides against infection, and this is true for a wide range of vaccine responses. In this setting VE does not depend on the vaccine's effect on infectivity nor on the distribution of the infection potential B_0 . Second, this useful measure of vaccine efficacy is estimated by

$$\widehat{VE} = 1 - \frac{\text{proportion of cases among vaccinated individuals}}{\text{proportion of cases among unvaccinated individuals}} = 1 - \frac{C_V/v}{C_0/(1-v)},$$

when the vaccination coverage v is known and the eventual number of cases among vaccinated and unvaccinated individuals (C_V and C_0) are observed.

In the Appendix an approximate standard error for \widehat{VE} , valid under the assumption that $\mathbf{Z}_0 \equiv c \mathbf{d}_U$, is shown to be

$$\text{s.e.}(\widehat{VE}) = \sqrt{\frac{(1 - \widehat{VE})(1 - v\widehat{VE})}{C_+} \left(1 - \frac{i_0}{C_+}\right) \left(\frac{1 - \widehat{VE}}{1 - v} + \frac{1}{v}\right)},$$

where $C_+ = C_V + C_0 = \sum_{i=0}^k C_i$ is the observed total number of cases and i_0 is the initial number of infectious individuals. If i_0 is unknown a conservative estimate is obtained by setting $i_0 = 1$. The derivation of the above standard error requires that $\text{Var}(B_0)$ is finite. It is interesting that the standard error, derived using the delta method, does not depend on $\text{Var}(B_0)$ even though the variance-covariance matrix of \mathbf{C} does depend on $\text{Var}(B_0)$.

5 Combining several different outbreaks

The estimator \widehat{VE} considered in the previous section is based on one minor outbreak and its variance is inversely proportional to the outbreak size and therefore *may* not be very small for minor outbreaks. Note, however, that a minor outbreak, by definition, is the total progeny of a subcritical branching process which only guarantees that it is finite. One way to obtain an estimate with higher precision is to combine data from several separate outbreaks, which we now outline briefly.

Suppose we have data from J different minor outbreaks from a community in which a proportion v are vaccinated with the same type of vaccine. Each such minor outbreak gives an estimate of the vaccine efficacy

$$\widehat{VE}_j = 1 - \frac{C_V^{(j)}/v}{C_U^{(j)}/(1-v)},$$

$j = 1, 2, \dots, J$, where $C_V^{(j)}$ and $C_U^{(j)}$ are the number of vaccinated and unvaccinated cases in outbreak j . Since these are minor outbreaks it is natural to combine the C_V 's and C_U 's of different outbreaks by the averages of the different outbreaks. Our combined estimator is hence given by

$$\widehat{VE}_C = 1 - \frac{\overline{C}_V/v}{\overline{C}_U/(1-v)},$$

where $\overline{C}_V = \sum_{j=1}^J C_V^{(j)}/J$ and $\overline{C}_U = \sum_{j=1}^J C_U^{(j)}/J$.

A standard error for \widehat{VE}_C can be derived by the *delta*-method, using the fact that outbreaks are independent. It turns out to have the form of the standard error for \widehat{VE} based on only one single outbreak divided by \sqrt{J} . That is

$$\text{s.e.}(\widehat{VE}_C) = \sqrt{\frac{(1 - \widehat{VE}_C)(1 - v\widehat{VE}_C)}{J\overline{C}_+} \left(1 - \frac{\bar{i}_0}{\overline{C}_+}\right) \left(\frac{1 - \widehat{VE}_C}{1 - v} + \frac{1}{v}\right)},$$

where $\overline{C}_+ = \sum_{j=1}^J C_+^{(j)}/J$ is the average total number of cases and $\bar{i}_0 = \sum_{j=1}^J i_{0j}/J$ is the average number of initial cases. If unknown, i_{j0} , the initial number of infectives in outbreak j , can be set to 1 implying that $\bar{i}_0 = 1$ to provide a conservative standard error.

Example 1. Measles was eliminated in the State of Victoria, Australia, following a state-wide campaign of vaccination in 1998, and the State practices an enhanced surveillance for measles. Importations of measles infection has led to thirty-three outbreaks of measles since then, including fourteen with secondary infections. The total number of cases was 208, of which 29 had been vaccinated.

From the Australian Childhood Immunisation Register, serological surveys and past data on immunisation schedules it is estimated that the vaccination coverage in Victoria is 82%, with most susceptible individuals falling into the 15-35 year age group. The vaccination coverage can be assumed constant since 1998, as an effective routine vaccination is in place to vaccinate infants at age 12 months. The community is assumed to be homogeneous.

This leads to the estimate $\widehat{VE}_C = 0.964$, with standard error $\text{s.e.}(\widehat{VE}_C) = 0.0066$. The small standard error indicates that this kind of estimation is of practical value.

6 Community of individuals with inherent differences

Now consider individuals who differ even in the absence of vaccination. In other words, there is an inherent difference between individuals, which may be a biological difference, a difference in social mixing or both. Suppose there are ℓ types of individual, labelled $1, \dots, \ell$. We refer to them as *base* types to distinguish them from vaccine induced differences.

Disease transmission in such a heterogeneous community is described by a multitype epidemic model. In such models the distribution of B_0 , the total area under the infectiousness function, may differ between types of individual. The rate of mixing between individuals of different type may also differ. See Andersson and Britton [1] for a detailed description of the multitype model.

To introduce our notation, which is analogous to that of Section 4, suppose for the moment that no one has acquired immunity from either vaccination or previous exposure to the infection. Let $m_{rs} = \alpha_r \omega_r \beta_s$ be the mean number of individuals of type r that an initial infective of type s infects, where ω_r is the proportion of individuals of type r . The assumption that each m_{rs} splits into a product of the form

$$(\text{term depending on } r \text{ only}) \times (\text{term depending on } s \text{ only})$$

is known as proportionate, or separable, mixing. In Section 4 we saw that our assumption about vaccine response created different types of individual for which the mean matrix is of the proportionate mixing form. Here we add the proportionate mixing assumption for base types. Strictly speaking, a constraint is needed to make the parameters $\{\alpha_r\}$ and $\{\beta_r\}$ identifiable, but our calculations do not require that this constraint be specified. The proportionate mixing assumption ensures that the basic reproduction number R_0 has the simple form

$$R_0 = \sum_{r=1}^{\ell} \alpha_r \beta_r \omega_r,$$

see [7].

Now suppose that a fraction v_r of base type- r individuals are vaccinated and that individuals may respond differently to vaccination. We introduce k_r possible vaccine responses for individuals of type r . Post-vaccination we label the types of individual ri , where the first letter indicates the base type and the second letter indicates the type of vaccine response. There are then a total of $\sum_{r=1}^{\ell} k_r$ types of vaccinated individual and ℓ types of unvaccinated individual. The vaccination responses are described by

$$\Pr(A_r = a_{ri}, B_r = b_{ri}) = p_{ri}, \quad i = 1, 2, \dots, k_r, \quad r = 1, 2, \dots, \ell$$

and $a_{r0} = b_{r0} = 1$, for each r , because 0 refers to unvaccinated individuals. The proportion of type ri individuals is given by

$$\pi_{ri} = \begin{cases} (1 - v_r) \omega_r & \text{when } i = 0, \\ v_r p_{ri} \omega_r & \text{otherwise.} \end{cases} \quad (8)$$

In this notation, the post-vaccination reproduction number is

$$R_V = \sum_{r=1}^{\ell} \alpha_r \beta_r \left(\sum_{i=0}^{k_r} \pi_{ri} a_{ri} b_{ri} \right), \quad (9)$$

which is the multitype analogue of (4). Assume that the vaccination coverage is adequate to prevent epidemics, that is $R_V \leq 1$. Then the size of an outbreak initiated by a small number of introductory cases can be approximated by the total progeny in a multitype branching process. We are interested in the distribution of the final size vector \mathbf{C} , whose typical element gives the eventual number of cases among those of base type r with vaccine response i . As before, $\mathbf{C} = \mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2 + \dots$, where \mathbf{Z}_t denotes the vector of the various base and vaccine responses types of infected individuals in generation t . Generation 0 consists of the introductory cases of the outbreak. The individuals infected by direct contact with the introductory cases make up Generation 1, and so forth.

The mean $E(\mathbf{C})$ satisfies equation (6), but the expressions for the elements of \mathbf{M} now reflect inherent heterogeneity as well as heterogeneity due to varying vaccine response. In the present setting the elements of the mean matrix \mathbf{M} are given by $m_{ri,sj} = \alpha_r a_{ri} \pi_{ri} \beta_s b_{sj}$.

We may write

$$\mathbf{M} = \mathbf{d}_{\text{MT}} \mathbf{b}_{\text{MT}}^T,$$

where the $(\ell + \sum_{r=1}^{\ell} k_r)$ -vectors \mathbf{d}_{MT} and \mathbf{b}_{MT} are given by

$$\mathbf{d}_{\text{MT}}^T = (\alpha_1 a_{10} \pi_{10}, \dots, \alpha_1 a_{1k_1} \pi_{1k_1}, \alpha_2 a_{20} \pi_{20}, \dots, \alpha_2 a_{2k_2} \pi_{2k_2}, \dots, \alpha_{\ell} a_{\ell 0} \pi_{\ell 0}, \dots, \alpha_{\ell} a_{\ell k_{\ell}} \pi_{\ell k_{\ell}})$$

and

$$\mathbf{b}_{\text{MT}}^T = (\beta_1 b_{10}, \dots, \beta_1 b_{1k_1}, \beta_2 b_{20}, \dots, \beta_2 b_{2k_2}, \dots, \beta_{\ell} b_{\ell 0}, \dots, \beta_{\ell} b_{\ell k_{\ell}}).$$

The subscript in \mathbf{d}_{MT} and \mathbf{b}_{MT} indicates that this formulation is for multitype individuals. The largest eigenvalue of the matrix $\mathbf{M} = \mathbf{d}_{\text{MT}} \mathbf{b}_{\text{MT}}^T$ is simply $\mathbf{d}_{\text{MT}}^T \mathbf{b}_{\text{MT}}$ which is exactly R_V , as given in Equation (9). The assumption $R_V = \mathbf{d}_{\text{MT}}^T \mathbf{b}_{\text{MT}} < 1$ is equivalent to assuming that the multitype branching process is subcritical.

It is reasonable to assume that $E(\mathbf{Z}_0)$ is proportional to \mathbf{d}_{MT} thus reflecting both relative susceptibility and relative abundance of the different types. This leads to

$$E(\mathbf{C}) = (\mathbf{I} - \mathbf{M})^{-1} E(\mathbf{Z}_0) = c \left(\mathbf{I} + \frac{1}{1 - R_V} \mathbf{d}_{\text{MT}} \mathbf{b}_{\text{MT}}^T \right) \mathbf{d}_{\text{MT}} = \frac{c}{1 - R_V} \mathbf{d}_{\text{MT}},$$

provided $R_V = \mathbf{d}_{\text{MT}}^T \mathbf{b}_{\text{MT}} < 1$.

To obtain $E(C_0)$, the expected number of unvaccinated cases, we sum those elements of $E(\mathbf{C})$ corresponding to unvaccinated individuals of the various types. This gives

$$E(C_0) = \frac{c}{1 - R_V} \sum_{r=1}^{\ell} \alpha_r \pi_{r0}.$$

To obtain $E(C_V)$, the expected number of vaccinated cases, we sum those elements of $E(\mathbf{C})$ corresponding to vaccinated individuals of the various types, which gives

$$E(C_V) = \frac{c}{1 - R_V} \sum_{r=1}^{\ell} \sum_{i=1}^{k_r} \alpha_r a_{ri} \pi_{ri}.$$

Substituting π_{ri} , as given by (8), leads to

$$VE = 1 - \frac{\text{expected proportion vaccinated individuals to become infected}}{\text{expected proportion unvaccinated individuals to become infected}}$$

$$\begin{aligned}
&= 1 - \frac{E(C_V)/v}{E(C_0)/(1-v)} \\
&= 1 - \frac{\sum_{r=1}^{\ell} \alpha_r v_r \omega_r E(A_r)/v}{\sum_{r=1}^{\ell} \alpha_r (1-v_r) \omega_r / (1-v)}
\end{aligned} \tag{10}$$

where $v = \sum_{r=1}^{\ell} v_r \omega_r$ is the overall vaccination coverage.

As for a community of homogeneous individuals VE is only affected by the vaccine's reduction of susceptibility and not by the vaccine's effect on infectivity once infected. However, the VE in (10) also depends on the inherent susceptibility of different types of base individual, their relative abundance and the vaccination coverage. Therefore, the quantity estimated by

$$\widehat{VE} = 1 - \frac{C_V/v}{C_0/(1-v)}$$

in a multitype setting has an interpretation that is somewhat obscure. The following two examples give an indication where care is needed.

Example 2. Suppose that the same fraction of each base type of individual is vaccinated, that is $v_r = v$ for all r . Then

$$VE = 1 - \sum_{r=1}^{\ell} \gamma_r E(A_r),$$

where $\gamma_r = \alpha_r \omega_r / \sum_{i=1}^{\ell} \alpha_i \omega_i$. The last term is a weighted average of the $E(A_r)$, with weights γ_r reflecting the relative susceptibility and relative abundance of the base types, and therefore a sensible concept of vaccine efficacy. In the particular case when the vaccine induces the same expected reduction in susceptibility for all base types, that is $E(A_r) = E(A)$ for all r , we obtain

$$VE = 1 - E(A).$$

Therefore \widehat{VE} estimates a characteristic that depends only on the vaccine when it induces the same mean reduction in susceptibility for all base types and an equal fraction of each base type is vaccinated. In other words, the estimate is then robust with respect to differences in the α_r and in the β_r among base types. However, \widehat{VE} estimates a quantity that depends on the heterogeneity inherent within this community when vaccination coverages v_r differ.

Example 3. Suppose now that there are two types of individual, and that the vaccine responses are the same for both types. That is, $\ell = 2$ and $E(A_1) = E(A_2) = E(A)$. Then Equation (10) becomes

$$VE = 1 - \frac{\alpha_1 v_1 \omega_1 + \alpha_2 v_2 \omega_2}{\alpha_1 (1-v_1) \omega_1 + \alpha_2 (1-v_2) \omega_2} \frac{1 - (v_1 \omega_1 + v_2 \omega_2)}{v_1 \omega_1 + v_2 \omega_2} E(A).$$

Without loss of generality assume $\alpha_1 < \alpha_2$, so type-2 individuals are more susceptible. Then VE increases with v_1 and decreases with v_2 . This implies that VE can take any value between two extremes. For $v_1 = 0$ and $v_2 = 1$ we have $VE = 1 - \frac{\alpha_2}{\alpha_1} E(A)$, its lowest value, while $v_1 = 1$ and $v_2 = 0$ gives $VE = 1 - \frac{\alpha_1}{\alpha_2} E(A)$, its highest value.

It is seen that VE can even attain a negative value, when α_2 is large enough, the vaccine effect is modest and most vaccinations are of type-2 individuals. For example, when $\alpha_2/\alpha_1 = 5$, $E(A) = 0.5$, $v_1 = 0$ and $v_2 = 1$ we find $VE = -1.5$. This negative value for VE is misleading since the vaccine is clearly protective; it reduces susceptibility by 50% for both types. The explanation lies in the fact that vaccinees are type-2 individuals who, even after vaccination, are more susceptible than unvaccinated type-1 individuals. The proportion infected among vaccinated individuals will therefore be higher than the proportion infected among unvaccinated individuals, thus giving a negative VE . This example shows that VE is not always a sensible measure of vaccine efficacy when applied to outbreaks in multitype communities.

A more suitable assessment of vaccine efficacy in the case of multitype outbreaks is to measure VE separately for each type, assuming types are distinguishable. This requires that the vaccination coverage for each type is known. The corresponding estimates are then

$$\widehat{VE}_r = 1 - \frac{C_{Vr}/v_r}{C_{0r}/(1-v_r)} \quad r = 1, \dots, \ell,$$

where C_{Vr} and C_{0r} are the number of type- r cases among vaccinated and unvaccinated individuals, respectively. The quantity \widehat{VE}_r estimates $VE_r = 1 - E(A_r)$. If a single measure of the vaccine efficacy is desired the weighted average $\widehat{VE}_C = \sum_r \omega_r \widehat{VE}_r$ should be used. It estimates $VE_C = 1 - \sum_r \omega_r E(A_r)$, the mean reduction in susceptibility in the community.

7 Community of households

Consider homogeneous individuals who reside in a community consisting of a large number of households. A detailed presentation of an epidemic model for a community of households is given by Ball *et al.* [3]. We outline only the assumptions needed for our focus on estimating vaccine efficacy. Let ν_i denote the proportion of households of size i , $i = 1, \dots, h$, so that $\bar{\nu} = \sum_i i\nu_i$ is the mean household size. Contacts are more frequent between two household members than they are between two individuals who do not share the same household. The transmission assumptions of sections 2 and 4 are modified for this scenario in the following way. At x time units after infection, an infected individual has infectious contacts with a specific individual outside the household at a rate $\beta(x)/n$. Additionally, he has contacts with each individual of his household at a rate $\theta_w \beta(x)$, which is independent of the household size. The constant θ_w reflects the enhanced rate of contacts within the household.

Imagine one infected individual in this community when no one has acquired immunity from either vaccination or previous exposure to the infection. Let $R_c = E(B_0)$ be the expected number of individuals from other households to whom infection is transmitted by this infective. Further, assuming the infective resides in a household of size i , let μ_i denote the expected eventual number of cases in the household neglecting further external infections. Expressions for μ_i can be obtained using certain recursive formulae, see for example Andersson and Britton [1, Section 2.4]. Several different basic reproduction numbers can be defined for this initial infective, depending on the way we attribute infections to an

individual. Becker and Dietz [6] define four different basic reproduction numbers for infectives in an SIR epidemic model for a community of households. Here we attribute to an infective all of the primary infectives he generates (by direct contact) in other households, *as well as* all secondary cases arising in those households. In other words, the secondary cases are not attributed to the infectives within the household, whose contacts actually caused their infections, but instead to the infective who generated the primary case of the household. This choice is made because the resulting basic reproduction number has an explicit expression, which is

$$R_0 = R_c \sum_{i=1}^k \frac{iV_i}{\bar{V}} \mu_i,$$

see Hall and Becker [10] and Ball *et al.* [3], and a major outbreak is possible if and only if $R_0 > 1$.

Now suppose that a fraction v of all individuals, chosen among the households in some specified way, are vaccinated and that individuals may respond differently to vaccination. As before, individuals can then be of type 0, 1, 2, ..., k , where 0 indicates unvaccinated individuals and the proportion of different types of vaccine response is specified by $\{p_r\}$, as in the distribution (2) above. We now have types of individual and types of household. Distinguish the type of household by label $\mathbf{n} = (n_0, n_1, \dots, n_k)$, where n_i is the number of type- i individuals in the household. For any given external contact made by an infective, let $\pi_{\mathbf{n}r}$ be the probability that this contact is with a type- r individual from a household of type \mathbf{n} . These probabilities depend on the response probabilities $\{p_r\}$ and on the way vaccinations are allocated among households. The probability that the type- r individual is infected by this contact is a_r . Let $\mu_{\mathbf{n}ri}$ denote the mean number of cases of type i that result within the household of kind \mathbf{n} , given that the primary household case is of type r and neglecting further external infection. Their expressions can be obtained using recursive formulas for a multitype epidemic model, see Andersson and Britton [1, Section 6.1]. In this notation the reproduction number after vaccination is

$$R_V = R_c \sum_{i=0}^k b_i \sum_{r, \mathbf{n}} a_r \pi_{\mathbf{n}r} \mu_{\mathbf{n}ri}, \quad (11)$$

see Becker and Starczak [8]. A major outbreak is possible if and only if $R_V > 1$.

Assume v is large enough to prevent epidemics, that is $R_V < 1$, so that the size of an outbreak initiated by a small number of introductory cases can be approximated by the total progeny in a suitably-defined subcritical branching process. One way to define such a branching process is in terms of the proliferation of infected households, as is done by Bartoszyński [4]. For our purpose it is more appropriate to think in terms of the proliferation of infected individuals, with infections attributed to infectives as described above. Let C_i denote the eventual number of type- i individuals who are infected. We are interested in $\mathbf{C} = \mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2 + \dots$, where \mathbf{Z}_t denotes the vector of the various types of infected individuals in generation t . Generation 0 consists of the introductory cases and all secondary cases in their households. Generation 0 cases infect a number of individuals from different households. These newly infected individuals *and* all secondary cases in their households, make up Generation 1. Generation 2 consists of all individuals directly infected by Generation 1 cases as well as all cases in the households of those new primary infectives, and so forth.

The mean $\mathbf{E}(\mathbf{C})$ satisfies equation (6), but the expressions for the elements of $\mathbf{M} = (m_{ij})$ now reflect the household structure of the community and the mean sizes of household outbreaks. The element m_{ij} is the expected number of type- i individuals infected (infectives of the next generation) by a single type- j infective. It may be expressed as

$$m_{ij} = b_j R_c \sum_{r, \mathbf{n}} a_r \pi_{\mathbf{n}r} \mu_{\mathbf{n}ri}$$

and we may write

$$\mathbf{M} = \mathbf{d}_{\mathbf{H}} \mathbf{b}^T,$$

where element $i + 1$ of the $(k + 1)$ -vector $\mathbf{d}_{\mathbf{H}}$ is $R_c \sum_{r, \mathbf{n}} a_r \pi_{\mathbf{n}r} \mu_{\mathbf{n}ri}$ and element $i + 1$ of the $(k + 1)$ -vector \mathbf{b} is b_i , $i = 0, 1, \dots, k$. The subscript in $\mathbf{d}_{\mathbf{H}}$ indicates that this formulation is for a community of households. As before, the form of the matrix \mathbf{M} implies that its largest eigenvalue is $\mathbf{d}_{\mathbf{H}}^T \mathbf{b}$, which is exactly R_V , as given in (11), so the assumption $R_V < 1$ is identical to saying that this multitype branching process is subcritical.

For this form of \mathbf{M} we can easily verify that

$$(\mathbf{I} - \mathbf{M})^{-1} = \mathbf{I} + \frac{1}{1 - R_V} \mathbf{M},$$

provided $R_V = \mathbf{d}_{\mathbf{H}}^T \mathbf{b} < 1$.

Note that \mathbf{Z}_0 gives the numbers of type $0, 1, 2, \dots, k$ individuals in the households infected initially. Assume that the community outbreak is initiated by a few individuals, possibly one, making a contact with an external infective. If external contacts are made randomly it is reasonable to assume that $\mathbf{E}(\mathbf{Z}_0)$ is proportional to $\mathbf{d}_{\mathbf{H}}$. This leads to

$$\mathbf{E}(\mathbf{C}) = (\mathbf{I} - \mathbf{M})^{-1} \mathbf{E}(\mathbf{Z}_0) = c \left(\mathbf{I} + \frac{1}{1 - R_V} \mathbf{d}_{\mathbf{H}} \mathbf{b}^T \right) \mathbf{d}_{\mathbf{H}} = \frac{c}{1 - R_V} \mathbf{d}_{\mathbf{H}},$$

provided $R_V < 1$.

Therefore $\mathbf{E}(C_0) = (1, 0, \dots, 0) \mathbf{E}(\mathbf{C})$ and $\mathbf{E}(C_V) = (0, 1, \dots, 1) \mathbf{E}(\mathbf{C})$ lead to

$$VE = 1 - \frac{\mathbf{E}(C_V)/v}{\mathbf{E}(C_0)/(1 - v)} = 1 - \frac{(1 - v) \sum_{r, \mathbf{n}} a_r \pi_{\mathbf{n}r} \sum_{i=1}^k \mu_{\mathbf{n}ri}}{v \sum_{r, \mathbf{n}} a_r \pi_{\mathbf{n}r} \mu_{\mathbf{n}r0}}. \quad (12)$$

Contrary to the community setting of sections 4 and 6, in the present setting VE depends also on the b_i 's, the vaccine's effect on infectiousness once infected. This is because $\mu_{\mathbf{n}ri}$, the expected number of infected type- i individuals in a household of type \mathbf{n} , in an outbreak initiated by one primary household infective of type r and neglecting external infection, depends on the b_i 's. For a simple illustration of this fact, suppose $b_r = 0$. Then the primary case will not spread the disease further, implying that

$$\mu_{\mathbf{n}ri} = \begin{cases} 1 & \text{for } i = r, \\ 0 & \text{for } i \neq r. \end{cases}$$

Now consider Equation (12) for some specific examples. The first example merely shows that Equation (12) reduces to the VE of Section 4 under appropriate specific assumptions. The remaining two examples are intended to reveal the extent to which various factors influence the expression (12), to help us understand what \widehat{VE} is estimating in this setting.

Example 4. Suppose all households are of size 1. Then 'household type' coincides with 'individual type', giving

$$\pi_{nr} = \begin{cases} 1 - v & \text{when } r = 0, \\ vp_r & \text{otherwise,} \end{cases} \quad \text{and} \quad \mu_{nri} = \begin{cases} 1 & \text{when } r = i, \\ 0 & \text{otherwise,} \end{cases}$$

so that VE reduces to $1 - E(A)$, as it should since the setting is then equivalent to that considered in Section 4, namely a community of homogeneous individuals who mix uniformly.

Example 5. Suppose all households are of size 2, that exactly one individual per household is vaccinated and that the Reed-Frost model applies; see Section 2. Denote the household types $(0, j)$, $j = 1, 2, \dots, k$, where 0 refers to unvaccinated individuals and j refers to the type of vaccinated individual in the household. Then

$$\pi_{(0,j)r} = \begin{cases} p_j/2 & \text{when } r = 0 \text{ or } j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mu_{(0,j)ri} = \begin{cases} 1 & \text{when } r = i = 0 \text{ or } r = i = j, \\ 1 - q_w^{a_j} & \text{when } r = 0 \text{ and } i = j, \\ 1 - q_w^{b_j} & \text{when } r = j \text{ and } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

where q_w is the probability that one susceptible avoids infection from his infected household partner when both are unvaccinated individuals. Substituting these quantities and $v = 0.5$ into Equation (12) gives

$$VE = 1 - \frac{E[A + (1 - q_w^A)]}{E[1 + A(1 - q_w^B)]} = \frac{E(q_w^A - Aq_w^B)}{E(1 + A - Aq_w^B)}. \quad (13)$$

Under the natural assumption that $A \leq 1$ and $B \leq 1$, implying that vaccination never increases susceptibility or infectivity, it can be shown that $VE \leq 1 - E(A)$.

This result may tempt us to use \widehat{VE} as a conservative estimate of $1 - E(A)$, but this is unwise. Firstly, because the inequality is strict except for some trivial cases. More importantly, VE can be substantially less than $1 - E(A)$. For example, with $\Pr(A = .9, B = .1) = 1$ and $q_w = .1$ we get $VE = -0.479$. Not only is this far removed from $1 - E(A) = 0.1$, but VE is negative although the vaccine reduces both susceptibility and infectivity.

The negative value of VE is explained as follows. The stipulated effect of the vaccine's reduction in susceptibility is modest, while the reduction in infectivity is substantial. Unvaccinated individuals have a slightly higher chance of being infected by an external contact, than their vaccinated partners, but they have a much lower chance of being infected by a within-household contact because of the low infectivity of their vaccinated partner. As a consequence, more vaccinated than unvaccinated individuals tend to be infected in this setting. Therefore \widehat{VE} underestimates $VE_S = 1 - E(A)$ and the greater the reduction in infectivity, in other words the better the vaccine performs, the more is VE_S underestimated to the extent that a negative effect may be indicated. In short,

the measure VE and its estimator \widehat{VE} can be misleading when based on data from a structured community.

We now consider more specific versions of (13) that help to quantify the extent of any underestimation of VE_S .

a) When there is no transmission within households, that is $q_w = 1$, the process behaves like a uniformly mixing community. We indeed get $VE = 1 - E(A) = VE_S$ when we substitute $q_w = 1$.

b) Suppose the vaccine has an all-or-nothing effect, with the probability of vaccine failure being p_F . That is, $\Pr(A = B = 1) = p_F = 1 - \Pr(A = 0)$. Then (13) becomes

$$VE = (1 - p_F) \left(\frac{1}{1 + p_F(1 - q_w)} \right) \leq 1 - p_F = VE_S,$$

which is never negative. However the deflation factor can be as low as $1/2$, tending towards this as $p_F \rightarrow 1$ and $q_w \rightarrow 0$.

c) Suppose the vaccine affects susceptibility, but not infectivity. That is, A may have any distribution, but $B \equiv 1$. Then (13) becomes

$$VE = \frac{E(q_w^A) - q_w E(A)}{1 + (1 - q_w)E(A)}.$$

If we also assume that $A \leq 1$, so one cannot become more susceptible from vaccination, then simple calculus gives $0 \leq VE \leq 1 - E(A)$. The two extremes are obtained when $A \equiv 1$ and $A \equiv 0$ respectively.

d) Suppose instead that the vaccine has no effect on susceptibility. That is, B may have any distribution, but $A \equiv 1$. Equation (13) then becomes

$$VE = \frac{q_w - E(q_w^B)}{2 - E(q_w^B)}.$$

If we also assume that $B \leq 1$, so the vaccine does not increase infectivity, then it is straightforward to show that $-(1 - q_w) \leq VE \leq 1 - E(A) = 0$. The upper bound is attained for $B \equiv 1$, corresponding to no vaccine effect. The lower bound is attained for $B \equiv 0$, corresponding to a vaccine that blocks transmission completely. In other words, \widehat{VE} estimates a negative quantity, even though the vaccine reduces disease transmission.

Example 6. Again suppose that all households are of size two, but now a proportion v of households have both members vaccinated and the remaining households have no vaccinated members. The possible household types are $(0, 0)$ and (i, j) for $1 \leq i \leq j \leq k$.

$$\pi_{(i,j)r} = \begin{cases} 1 - v & \text{when } i = j = r = 0; \\ vp_i^2 & \text{when } i = j = r = 1, 2, \dots, k; \\ vp_i p_j & \text{when } 1 \leq i < j \leq k \text{ and } r = i \text{ or } j; \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mu_{(i,j)rs} = \begin{cases} 1 + (1 - q_w^{a_r b_r}) & \text{when } i = j = r = s = 0, 1, 2, \dots, k; \\ 1 & \text{when } 1 \leq i < j \leq k \text{ and either } r = s = i \text{ or } r = s = j; \\ 1 - q_w^{a_s b_r} & \text{when } 1 \leq i < j \leq k \text{ and } r = i, s = j \text{ or } r = j, s = i; \\ 0 & \text{otherwise.} \end{cases}$$

Substituting these into Equation (12) gives

$$VE = 1 - \frac{E\left(A_1(2 - q_w^{B_1 A_2})\right)}{1 + p_w},$$

where (A_1, B_1) and (A_2, B_2) are independent and identically distributed random variables with distribution (2). Under the natural assumption that $A \leq 1$ and $B \leq 1$ it can be shown that $1 - E(A) \leq VE \leq 1 - E(A)/(1 + p_w)$. Note that the inequality with respect to $1 - E(A)$ here is in the opposite direction to that in Example 5. This means that VE not only depends on the effect of the vaccine and on the community structure, but also on how vaccinations are distributed over the community.

In the particular case of an all-or-nothing response we obtain

$$VE = 1 - \left(\frac{1 + p_F p_w}{1 + p_w}\right) p_F,$$

which increases with p_w and decreases with the failure rate p_F . The lower and upper bounds on VE are obtained in the limit as p_F tends to 1 and 0 respectively.

A Appendix

Here we derive a standard error for \widehat{VE} for a community of homogeneous individuals who mix uniformly. We begin by deriving expressions for the variance-covariance matrix of the total progeny in a multitype branching process, a quantity interesting in its own right. For a single-type branching process the variance for the total progeny is known, see Jagers [12, p. 39] for example, but we have not found the variance matrix for the total progeny of a multitype Galton-Watson branching process in the literature.

A.1 The total progeny of a multitype branching process

Consider a multitype Galton-Watson process with $k + 1$ different types of individual labelled $0, 1, \dots, k$. Bold font letters denote $(k + 1)$ -vectors unless referred to as a matrix, in which case the dimension is $(k + 1) \times (k + 1)$. Let $p^{(j)}(\mathbf{r})$ denote the offspring distribution of a type- j individual, $j = 0, \dots, k$. Let $\mathbf{Z}_t^{(j)} = (Z_{t0}^{(j)}, \dots, Z_{tk}^{(j)})'$ be the vector specifying the number of individuals of the different types in generation t assuming the process started with one type- j individual in generation zero. That is, $\mathbf{Z}_0^{(j)} = \mathbf{e}_j$, the $(k + 1)$ -vector whose j 'th element is unity and all its other elements are zero. The vector $\mathbf{Z}_1^{(j)}$ has distribution $p^{(j)}(\mathbf{r})$, so its probability generating function is

$$f^{(j)}(\mathbf{s}) = E\left(\mathbf{s}^{\mathbf{Z}_1^{(j)}}\right) = E\left(s_0^{Z_{10}^{(j)}} \cdot \dots \cdot s_k^{Z_{1k}^{(j)}}\right) = \sum_{\mathbf{r}} p^{(j)}(\mathbf{r}) \mathbf{s}^{\mathbf{r}}.$$

Define the mean matrix $\mathbf{M} = (m_{ij})$ by $m_{ij} = E(Z_{1i}^{(j)}) = f_i^{(j)}(\mathbf{1})$, where $f_i^{(j)}$ denotes the partial derivative of $f^{(j)}$ with respect to s_i . It is well-known, see [12], that

$$E(\mathbf{Z}_t^{(j)}) = \mathbf{M}^t \mathbf{e}_j.$$

Define also the second order moments of the offspring distribution

$$c_{uv}^{(j)} = \mathbb{E} \left(Z_{1u}^{(j)} Z_{1v}^{(j)} \right) - m_{uj} m_{vj} = \begin{cases} f_{uu}^{(j)}(\mathbf{1}) + f_u^{(j)}(\mathbf{1}) \left[1 - f_u^{(j)}(\mathbf{1}) \right], & u = v, \\ f_{uv}^{(j)}(\mathbf{1}) - f_u^{(j)}(\mathbf{1}) f_v^{(j)}(\mathbf{1}), & u \neq v, \end{cases} \quad (14)$$

where $f_{uv}^{(j)}(\mathbf{s})$ is the second derivative of $f^{(j)}(\mathbf{s})$ with respect to s_u and s_v .

The total number of individuals up to generation t is defined by $\mathbf{Y}_t^{(j)} = \sum_{r=0}^t \mathbf{Z}_r^{(j)}$. Each component of $\mathbf{Y}_t^{(j)}$ is increasing in t , so that $\mathbf{Y}_t^{(j)}$ tends to a limit $\mathbf{Y}_\infty^{(j)}$ as $t \rightarrow \infty$. From standard theory for branching processes, see [12], it is known that the largest eigenvalue R of the matrix \mathbf{M} determines whether $\mathbf{Y}_\infty^{(j)}$ can be infinite or not. For $R < 1$, known as the subcritical case, $\mathbf{Y}_\infty^{(j)}$ is finite almost surely and so is its expected value $\mathbb{E} \left(\mathbf{Y}_\infty^{(j)} \right)$. From now on assume that we are in the subcritical case $R < 1$, implying that the branching process will surely die out.

Let $h_t^{(j)}(\mathbf{s}) = \mathbb{E} \left(\mathbf{s}^{\mathbf{Y}_t^{(j)}} \right)$, the probability generating function for $\mathbf{Y}_t^{(j)}$. By conditioning on what happens in the first generation and using the fact that each of the individuals of the first generation initiates an independent multitype branching processes, we obtain the recursion equation

$$h_t^{(j)}(\mathbf{s}) = s_j f^{(j)} \left[\mathbf{h}_{t-1}(\mathbf{s}) \right].$$

This recursion can be used to derive the mean vector and variance-covariance matrix of the total progeny. Taking the partial derivative of $h_t^{(j)}(\mathbf{s})$ with respect to s_i and setting $\mathbf{s} = \mathbf{1}$ gives recursive formulae whose solution, in vector notation, is

$$\mathbb{E} \left(\mathbf{Y}_t^{(j)} \right) = (\mathbf{I} - \mathbf{M}^{t+1})(\mathbf{I} - \mathbf{M})^{-1} \mathbf{e}_j.$$

Since the largest eigenvalue of \mathbf{M} is less than unity, we can take limits as $t \rightarrow \infty$ to obtain

$$\mathbb{E} \left(\mathbf{Y}_\infty^{(j)} \right) = \mathbf{G} \mathbf{e}_j,$$

where $\mathbf{G} = (\mathbf{I} - \mathbf{M})^{-1}$ has typical element $= g_{ij} = \mathbb{E} \left(Y_{\infty i}^{(j)} \right)$.

The elements of the variance matrix for $\mathbf{Y}_\infty^{(j)}$ are derived in a similar fashion, using second order partial derivatives of $h_t^{(j)}(\mathbf{s})$. The result is

$$\sigma_{ir}^{(j)} = \text{Cov} \left(Y_{\infty i}^{(j)}, Y_{\infty r}^{(j)} \right) = \sum_{u=0}^k \sum_{v=0}^k c_{uv}^{(j)} g_{ui} g_{vr} + \sum_{u=0}^k m_{uj} \sigma_{ir}^{(u)} \quad \text{for all } i, j, r,$$

which may be written

$$\sigma_{ir}^{(j)} = \sum_{s=0}^k \sum_{u=0}^k \sum_{v=0}^k g_{js} g_{ui} g_{vr} c_{uv}^{(s)}. \quad (15)$$

The variance-covariance matrix for a process started by more than one individual is simply the sum over such elements, one for each ancestor, since all ancestors start *independent* branching processes.

Now consider some forms for \mathbf{M} and $\{c_{uv}^{(j)}\}$ of particular interest in our epidemic context.

A.2 Proportionate mixing

All community settings considered in this paper have a mean matrix of the form $\mathbf{M} = \mathbf{d}\mathbf{b}^T$. Its largest eigenvalue R , denoted R_V in previous sections, is then given by

$$R = \mathbf{d}^T \mathbf{b} = \sum_{i=0}^k d_i b_i.$$

It is assumed to be less than 1 and then

$$\mathbf{G} = (\mathbf{I} - \mathbf{M})^{-1} = \frac{1}{1 - R} \mathbf{d}\mathbf{b}^T + \mathbf{I}.$$

This means that

$$g_{ij} = \mathbb{E} \left(Y_{\infty i}^{(j)} \right) = \begin{cases} \frac{d_j b_j}{1 - R} + 1, & i = j, \\ \frac{d_i b_j}{1 - R}, & i \neq j. \end{cases}$$

These are the first moments of the total progeny assuming there was one ancestor of type j . It is often not known who started the epidemic. In previous sections we assume that the expected number of initially infected individuals of the different types is given by $\mathbb{E}(\mathbf{Z}_0) = c\mathbf{d}$ for some constant c . Let \mathbf{C} denote the vector of total progeny under this assumption, then the mean vector of \mathbf{C} has elements

$$\mathbb{E}(C_i) = c \sum_{j=0}^k d_j \mathbb{E} \left(Y_{\infty i}^{(j)} \right) = \frac{c d_i}{1 - R}, \quad (16)$$

a result given in previous sections.

To obtain expressions for the variance-covariance terms $\sigma_{ir}^{(j)}$ defined by (15) we need to know the variance-covariance terms defined by (14). For epidemics these are determined by the model and the distribution of the area under the infectiousness function. For an individual, given his value of B_0 and that, upon vaccination, his response makes him type j , the numbers of infectious contacts this individual has with each of the different types follow independent Poisson variables. The Poisson number of contacts with type- u individuals has mean $b_j d_u B_0 / \mathbb{E}(B_0)$, $u = 0, \dots, k$ (the denominator comes from the definition that the unconditional expected number of contacts is $b_j d_u$). When we drop the conditioning on B_0 the number of contacts with the different types become dependent, unless B_0 is deterministic. It can be shown that the variances and covariances for these number of contacts are

$$c_{uu}^{(j)} = b_j^2 d_u^2 \tau^2 + b_j d_u \quad \text{and} \quad c_{uv}^{(j)} = b_j^2 d_u d_v \tau^2, \quad u \neq v,$$

where $\tau^2 = \text{Var}(B_0) / [\mathbb{E}(B_0)]^2 = \text{Var}(B_0) / R_0^2$ is the squared coefficient of variation. Making use of the assumed structure of \mathbf{M} in (15) we find

$$\sigma_{ir}^{(j)} = \text{Cov} \left(Y_{\infty i}^{(j)}, Y_{\infty r}^{(j)} \right) = \begin{cases} \frac{d_i^2 b_j}{(1 - R)^3} \left[(1 + \tau^2)z + (1 - R)(\tau^2 b_j + 2b_i) + \frac{(1 - R)^2}{d_i} \right], & i = r, \\ \frac{d_i d_r b_j}{(1 - R)^3} \left[(1 + \tau^2)z + (1 - R)(\tau^2 b_j + b_i + b_r) \right], & i \neq r, \end{cases}$$

where $R = \sum_{u=0}^k b_u d_u$ and $z = \sum_{u=0}^k b_u^2 d_u$.

These are the variance-covariance terms for the vector $\mathbf{Y}_\infty^{(j)}$, i.e. given that there was one ancestor of type j . In the previous sections we assumed that the expected number of initial infectives of the different types is proportional to susceptibility and the abundance of the type. Here we make a slightly stronger assumption by assuming this relation but for the actual numbers rather than the expected numbers, i.e. that $\mathbf{Z}_0 \equiv c\mathbf{d}$ for some c . Since all ancestors constitute independent branching processes it follows that the covariance terms are then given by

$$\sigma_{ir} = \text{Cov}(C_i, C_r) = c \sum_{j=0}^k d_j \sigma_{ir}^{(j)} = \begin{cases} \frac{cRd_i^2}{(1-R)^3} \left[\left(1 + \frac{\tau^2}{R}\right)z + (1-R)2b_i + \frac{(1-R)^2}{d_i} \right], & i = r, \\ \frac{cRd_i d_r}{(1-R)^3} \left[\left(1 + \frac{\tau^2}{R}\right)z + (1-R)(b_i + b_r) \right], & i \neq r \end{cases} \quad (17)$$

The classical Reed-Frost and General Epidemic models correspond to $\tau^2 = 0$ and $\tau^2 = 1$ respectively.

A.3 Standard error of \widehat{VE} for a homogeneous community

We now derive an approximate standard error for the estimate \widehat{VE} in the setting of Section 4. Recall that $d_i = a_i \pi_i R_0$ where $\pi_0 = 1 - v$ and $\pi_j = v p_j$, $j = 1, \dots, k$, and that the estimate of the vaccine efficacy was defined as

$$\widehat{VE} = 1 - \frac{1-v}{v} \left(\frac{C_1 + \dots + C_k}{C_0} \right) = h(C_0, \dots, C_k), \quad \text{say.}$$

We apply the δ -method, see Rao [15, Section 6a.2], to $h(C_0, \dots, C_k)$ to obtain the approximate expression

$$\text{Var}(\widehat{VE}) \approx \sum_{i=0}^k \sum_{j=0}^k h_i(\mu_0, \dots, \mu_k) h_j(\mu_0, \dots, \mu_k) \text{Cov}(C_i, C_j),$$

where h_i denotes the partial derivative of $h(c_0, \dots, c_k)$ with respect to c_i , and $\mu_i = \text{E}(C_i)$. Inserting expressions for the partial derivatives we get

$$\begin{aligned} \text{Var}(\widehat{VE}) \approx & \left(\frac{1-v}{v} \right)^2 \left[\frac{(\mu_1 + \dots + \mu_k)^2}{\mu_0^4} \text{Var}(C_0) + \frac{1}{\mu_0^2} \sum_{i=1}^k \text{Var}(C_i) \right. \\ & \left. - 2 \frac{\mu_1 + \dots + \mu_k}{\mu_0^3} \sum_{i=1}^k \text{Cov}(C_0, C_i) + \frac{1}{\mu_0^2} \sum_{i \neq j \geq 1}^k \text{Cov}(C_i, C_j) \right]. \end{aligned}$$

The first and second order moments of the C_0, \dots, C_k are given by (16) and (17), respectively. This gives a variance expression containing c , which comes from assuming $\mathbf{Z}_0 \equiv c\mathbf{d}$. Let the total number of initial infectives be $i_0 = c \sum_{j=0}^k d_j$, then the variance formula becomes

$$\text{Var}(\widehat{VE}) \approx \frac{R(1-R)\text{E}(A)[1-v+v\text{E}(A)]}{i_0} \left[\frac{\text{E}(A)}{1-v} + \frac{1}{v} \right].$$

From this expression it is seen that the approximate variance of \widehat{VE} does not depend on τ , the coefficient of variation of the area under the infectiousness function. In other words, the variance of $\widehat{VE} = 1 - (1 - v)C_V/vC_0$ does not depend on τ even though the variances of C_0 and C_V do. Admittedly, this is only an approximate conclusion due to omitting higher order terms in the Taylor expansion when applying the δ -method. This result holds more generally, when the variance of B_0 depends on the type of the individual. In the expression for $\text{Var}(\widehat{VE})$, the quantities $E(A)$ and R are unknown and must be estimated. Naturally, $E(A)$ is estimated by $1 - \widehat{VE}$. We estimate R by replacing $\sum_{j=0}^k \mu_j$ in the expression $\sum_{j=0}^k \mu_j = i_0/(1 - R)$, see (16), by the corresponding observed value C_+ . This gives the estimate $\hat{R} = 1 - i_0/C_+$ and with these estimates we obtain the standard error

$$\text{s.e.}(\widehat{VE}) = \sqrt{\frac{(1 - \widehat{VE})(1 - v\widehat{VE})}{C_+} \left(1 - \frac{i_0}{C_+}\right) \left(\frac{1 - \widehat{VE}}{1 - v} + \frac{1}{v}\right)}.$$

If i_0 , the initial number of infectives, is unknown a conservative estimate is obtained by replacing i_0 by 1.

The authors thank Kerry-Ann O'Grady, Communicable Diseases Section, Department of Human Services, Victoria, Australia, for providing the data on measles outbreaks and immunity used for Example 1. Support from National Health and Medical Research Council (Project Grant 148919) and from The Swedish Research Council is gratefully acknowledged.

References

- [1] ANDERSSON, H., BRITTON, T. (2000). *Stochastic models and their statistical analysis*. Springer Lecture Notes in Statistics **151**. Springer, New York.
- [2] BALL, F & CLANCY, D. (1993). The final size and severity of a generalised stochastic multitype epidemic model. *Adv. Appl. Prob.* **25**, 721-736.
- [3] BALL, F., MOLLISON, D. & SCALIA-TOMBA, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Prob.* **7**, 46-89.
- [4] BARTOSZYŃSKI, R. (1972). On a certain model of an epidemic. *Applicationes Mathematicae* **13**, 139-151.
- [5] BECKER, N.G. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall, London.
- [6] BECKER, N.G. & DIETZ, K. (1996). Reproduction numbers and critical immunity levels for epidemics in a community of households. In *Athens Conference on Applied Probability and Time Series, Volume 1: Applied Probability*, Eds. Heyde C.C., Prohorov Yu V., Pyke R. and Rachev S.T. Lecture Notes in Statistics **114**, 267-276.

- [7] BECKER, N.G. & MARSCHNER, I.C. (1990). The effect of heterogeneity on the spread of disease, in *Stochastic Processes in Epidemic Theory*, Gabriel, J.-P., Lefèvre, C. and Picard, P. (Eds). *Springer Lecture Notes in Biomathematics* **86**:90-103.
- [8] BECKER, N.G. & STARCZAK, D.N. (1998). The effect of random vaccine response on the vaccination coverage required to prevent epidemics. *Mathematical Biosciences* **154**, 117-135.
- [9] DE SERRES, G., GAY, N.J., & FARRINGTON, C.P. (2000). Epidemiology of transmissible diseases after elimination. *Am. J. Epidem.* **11**, 1039-1048.
- [10] HALL, R. & BECKER N.G. (1996). Preventing epidemics in a community of households. *Epidemiol. Infect.* **117**, 443-455.
- [11] HALLORAN, M.E., LONGINI, I.M. & STRUCHINER, C.J. (1999). Design and interpretation of vaccine field studies. *Epidemiologic Reviews* **21**, 73 - 88.
- [12] JAGERS, P. (1975). *Branching Processes with Biological Applications*. Wiley, London.
- [13] ORENSTEIN, W.A., BERNIER, R.H., DONDERO, T.J. *et al.* (1985). Field evaluation of vaccine efficacy. *Bulletin of the World Health Organization* **63**, 1055-1068.
- [14] ORENSTEIN, W.A., BERNIER, R.H. & HINMAN, A.R. (1988). Assessing vaccine efficacy in the field: further observations. *Epidemiologic Reviews* **10**, 212-241.
- [15] RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd edition). New York: Wiley.
- [16] SMITH, P. G., RODRIGUES, L. C. and FINE, P. E. M. (1984). Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int. J. Epidemiology* **13**, 87-96.