Estimation of basic reproduction numbers: individual heterogeneity and robustness to mis-specification of the contact function

C. Paddy Farrington [1] Steffen Unkel [2]
Karim Anaya-Izquierdo [3]

[1]Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, United Kingdom. E-mail: c.p.farrington@open.ac.uk.

[2]Medical Statistics Group, Institute of Medical Informatics, Justus-Liebig University Giessen, Heinrich-Buff-Ring 44, 35392 Giessen, Germany. Email: steffen.unkel@informatik.med.uni-giessen.de.

[3]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT. Email: karim.anaya-izquierdo@lshtm.ac.uk.

Abstract

The basic reproduction number of an infection in a given population, R_0, is inflated by individual heterogeneity in contact rates. Recently, new methods for estimating R_0 using social contact data have been proposed. These methods, like most of their predecessors, ignore individual heterogeneity, and are sensitive to mis-specification of the contact function. Using a frailty framework, we derive expressions for R_0 in the presence of age-varying heterogeneity. In this case, R_0 is the spectral radius of a population version of the next generation operator, which involves the variance function of the age-dependent frailty. This variance can be estimated within a shared frailty framework from paired data on two infections transmitted by the same route. We propose two estimators of R_0 for infections in endemic equilibrium. We investigate their performance by simulation, and find that is generally more robust than the other to mis-specification of the effective contact function. These methods are applied to data on varicella zoster virus infection. Using paired data on parvovirus B19 infection, we show that there is evidence of individual heterogeneity in contact rates. The R_0 estimate for varicella zoster infection increases appreciably when this heterogeneity is taken into account.

KEYWORDS: eigenfunction; eigenvalue; infectious disease; operator; social contact; time-varying frailty.

1 Introduction

The basic reproduction number of an infection, R_0, is the expected number of individuals infected by a single 'typical' individual during their infectious period, within an otherwise susceptible population (Diekmann, Heesterbeek and Metz 1990). If $R_0 > 1$ the infection may become endemic within the population, and the larger the value of R_0, the harder it is to eliminate the infection. For these reasons, estimating R_0 is a problem of practical relevance to public health.

Several contrasting methods have been proposed for estimating R_0; see Dietz (1993) and Farrington, Kanaan and Gay (2001). More recently, attempts have been made to combine serological survey data with contact functions estimated from social contact surveys to estimate reproduction numbers for endemic infections (Wallinga, Teunis and Kretzschmar 2006, Goeyvaerts et al. 2010, Melegaro et al. 2011). Most statistical approaches used to date, with the exception of that of Farrington, Kanaan and Gay (2001), take no account of heterogeneities in individual circumstances and behaviors, the presence of which inflates the value of R_0. All methods, even when based on social contact data, are sensitive to mis-specification of the contact function (Greenhalgh and Dietz 1994, Goeyvaerts et al. 2010).

We propose to incorporate individual heterogeneity from unmeasured age-dependent covariates, which may be achieved naturally within a frailty framework (Aalen, Borgan and Gjessing 2008, Chapter 6). These developments build on ideas first introduced in Farrington, Kanaan and Gay (2001), but allow individual heterogeneity to vary with age, in line with empirical observations (Unkel and Farrington 2012, Unkel et al. 2012). The approach yields two contrasting estimators of R_0, which make different use of the contact function. We study the robustness of these estimators to mis-specification of the contact function.

2 The basic reproduction number

2.1 Context

Throughout we consider an infection transmitted directly and non-sexually from person to person, in endemic equilibrium within a large population. The infection is spread by contacts between infected and susceptible individuals, the rate at which contacts occur being governed by the effective contact rate function. This usually depends on measured covariates such as age, gender or location, and unmeasured, possibly occult, covariates related to individual behaviors, characteristics or environments which may enhance or inhibit the transmission of infection.

2.2 Next generation operator with individual heterogeneity

For simplicity, we assume that age is the only measured covariate, denoted x or y. We suppose that the age-dependent effect of the unmeasured covariates can be compounded into a single positive random variable of unit mean, whose distribution and support may depend on x. We denote such random variables $u(x)$, and assume that

$$u(x) = w(x, u_1, \ldots, u_k)$$

where $w(.)$ is a deterministic function and $u_1, \ldots, u_k$ are age-invariant random variables representing unmeasured individual characteristics. Similarly, when we need to consider distinct individuals, we shall write $v(y) = w(y, v_1, \ldots, v_k)$. Note that, while $u_1, \ldots, u_k$ are age-invariant, it is still possible to accommodate age-dependence through $u(x)$. Of key importance in describing the degree of heterogeneity is the variance of $u(x)$; let $\gamma(x) = \mathrm{var}\{u(x)|x\}$. The range of $u(x)$ is determined only by x and the function w.

The effective contact rate function $\beta(x, u(x); y, v(y))$ denotes the contribution by an infected person with characteristics $(y, v(y))$ to the infection hazard of susceptible persons with characteristics $(x, u(x))$.

Let $f(x, u(x))$ denote the joint density of age x and $u(x)$. This may be written $f(u(x)|x)f(x)$ where $f(x)$ is the marginal age distribution of the population.

To keep matters simple, we shall assume that infection-related mortality is ignorable, and that the infectious period D of the infection and its latency period are short. These assumptions can be weakened; see for example Farrington, Kanaan and Gay (2001). For each $y > 0$ let $(a(y), b(y))$ denote the range of $v(y)$, where $b(y)$ can be $+\infty$, and let $\mathbb{D} \subseteq \mathbb{R}^+ \times \mathbb{R}^+$ denote the set of values $(x, u(x))$. Define the following operator B on functions $h : \mathbb{D} \to \mathbb{R}^+$:

$$B[h](x, u(x)) = Df(x, u(x)) \int_0^\infty \int_{a(y)}^{b(y)} \beta(x, u(x); y, v(y))h(y, v(y))\mathrm{d}v(y)\mathrm{d}y. \quad (1)$$

The operator B is clearly linear and positive. It is the next generation operator of the infection: if at a given instant, the frequency distribution of infectious individuals is h, where h integrates to the number of infectives present, then $B[h]$ denotes the expected frequency distribution of the individuals they infect during their infectious period D. These individuals constitute the next generation of infectives. The literature refers more usually to the next generation matrix, of which R_0 is the leading eigenvalue (Diekmann, Heesterbeek and Metz 1990). In the present setting the variable $u(x)$, and possibly also x, is continuous, whence the use of operators (Greenhalgh 1990).

2.3 Frailty model for age-dependent individual heterogeneity

To make further progress, we introduce a multiplicative age-dependent frailty perspective and assume that

$$\beta(x, u(x); y, v(y)) = u(x)\beta_0(x, y)v(y)$$

where $u(x)$, $v(y)$, for fixed x and y, are independent age-dependent random variables of unit mean with identical conditional densities $f(u(x)|x)$ and $f(v(y)|y)$, and $\beta_0(x, y)$ is a deterministic non-negative bivariate function. Thus, $u(x)$ and $v(y)$ are multiplicative

frailties. The next generation operator from equation (1) is then

$$B[h](x,u(x)) = Df(x,u(x))u(x)\int_0^\infty \int_{a(y)}^{b(y)} \beta_0(x;y)v(y)h(y,v(y))\mathrm{d}v(y)\mathrm{d}y. \qquad (2)$$

Since in general we can have $b(y) = \infty$, some care is required in what follows to ensure the operator is bounded. Note that the operator B maps into the following subset of functions on $\mathbb{D}$:

$$\mathbb{S} = \{f(u(x)|x)u(x)g(x) : g \text{ bounded and integrable on } \mathbb{R}^+\}.$$

This is a complete subspace of the set of bounded integrable functions on $\mathbb{D}$. Equipped with the L_1 norm, it is therefore a Banach space. Henceforth we shall restrict B and regard it as an operator from $\mathbb{S}$ to $\mathbb{S}$. For $h \in \mathbb{S}$, let h^* be the univariate function such that $h(x,u(x)) = f(u(x)|x)u(x)h^*(x)$. Then, in the respective L_1 norms,

$$|h| = \int_{\mathbb{D}} |f(u(x)|x)u(x)h^*(x)|\mathrm{d}u(x)\mathrm{d}x = \int_{\mathbb{R}^+} |h^*(x)|\mathrm{d}x = |h^*|.$$

Suppose now that the kernel $Df(x)\beta_0(x,y)\{1+\gamma(y)\}$ is positive, integrable and bounded above by K. It then follows that $|B[h]| \leq K|h^*| = K|h|$, and hence the operator B is bounded. Let $\rho(B) = \lim_{m\to\infty} |B^m|^{1/m}$ denote its spectral radius (Jörgens 1982, page 69). This spectral radius is the basic reproduction number, R_0 (Diekmann, Heesterbeek and Metz 1990, Greenhalgh 1990).

In practice, the individual frailties $u(x)$ are not observed: as in survival analysis, we only observe population quantities averaged over the frailties. Consider the operator B^* on the Banach space of bounded and integrable functions on $\mathbb{R}^+$ defined by

$$B^*[h^*](x) = Df(x)\int_0^\infty \beta_0(x,y)\{1+\gamma(y)\}h^*(y)\mathrm{d}y.$$

We show that B^* has the same spectral radius as B, so that henceforth inference about R_0 can be focused on B^*.

PROPOSITION. The operators B and B^* have the same spectral radii.

PROOF. Substituting $h(y, v(y)) = f(u(x)|x)u(x)h^*(x)$ in equation (2) and integrating out $u(x)$ we have $B[h](x, u(x)) = f(u(x)|x)u(x)B^*[h^*](x)$. Hence $B[h]^* = B^*[h^*]$. Iterating, we have, for any positive integer m,

$$B^m[h]^* = B^{*m}[h^*].$$

Since $|h| = |h^*|$ it follows that

$$\begin{aligned} |B^m| &= \sup\{|B^m[h]| : |h| = 1\} \\ &= \sup\{|B^m[h]^*| : |h| = 1\} \\ &= \sup\{|B^{*m}[h^*] : |h^*| = 1\} \\ &= |B^{*m}|, \end{aligned}$$

hence $\rho(B) = \rho(B^*)$.

The relationship of B^* to the next generation operator B is akin to the relationship between the population survivor function and the individual survivor function in frailty survival models (Aalen, Borgan and Gjessing 2008, page 235). In this sense, it can be regarded as the population next generation operator induced by B. It shares some of the identifiability issues of survival models with frailty. Thus, given data on a single infection, it is not possible to say whether the data arose from heterogeneous contacts of variance $\gamma(x)$ with next generation operator B, or whether they arose from non-heterogeneous contacts with next generation operator B^*. Later, however, we will see how it is possible to identify the correct scenario with the help of data on other infections.

Suppose now that B^* and its transpose are compact. Note that since B^* is bounded, B^* is compact if it is finite-dimensional, which can be assumed in applications since age is usually discretized. It then follows that, under suitable conditions (see Theorem 8.6 of Jörgens (1982) pages 183 and 238), $\rho(B)$ equals the leading eigenvalue of B^*. Note also that this is also equal to the leading eigenvalue of the operator B_1^* with kernel

$\beta_0(x,y)\{1+\gamma(y)\}f(y)$ and the operator B_2^* with kernel $f(x)\{1+\gamma(x)\}\beta_0(x,y)$. Thus $\rho(B) = \rho(B^*) = \rho(B_1^*) = \rho(B_2^*)$. Henceforth we shall refer to this common value as R_0.

2.4 Endemic infections

The hazard of infection acting on a susceptible individual of characteristics $(x, u(x))$ at time t is

$$\lambda(x,u(x),t) = N^{-1}u(x)\int_{\mathbb{D}}\beta_0(x,y)v(y)I(y,v(y),t)\mathrm{d}v(y)\mathrm{d}y$$

where N is the population size and $I(.,.,t)$ denotes the frequency distribution of infectious individuals at time t. When the infection is in endemic equilibrium, t may be dropped, and

$$I(y,v(y)) = ND\lambda(y,v(y))S(y,v(y))f(y,v(y))$$

where $S(y,v(y))$ is the probability that an individual of characteristics $(y,v(y))$ is susceptible. For an infection conferring long-lasting immunity, this is

$$S(y,v(y)) = \exp\Big(-\int_0^y \lambda(z,v(z))\mathrm{d}z\Big).$$

Thus, at the endemic equilibrium, the hazard of infection is of the form

$$\lambda(x,u(x)) = u(x)\lambda_0(x) \tag{3}$$

with

$$\lambda_0(x) = D\int_0^\infty \beta_0(x,y)\{1+\gamma(y)\}f(y)S_0(y)\lambda_0(y)\mathrm{d}y \tag{4}$$

where

$$S_0(y) = \{1+\gamma(y)\}^{-1}\int_{a(y)}^{b(y)} v(y)^2\exp\Big(-\int_0^y v(z)\lambda_0(z)\mathrm{d}z\Big)f(v(y)|y)\mathrm{d}v(y). \tag{5}$$

Equation (3) defines an age-dependent multiplicative frailty model for the hazard of infection, with frailty $u(x)$ and baseline hazard $\lambda_0(x)$. Note also that if $\lambda_0(x)$ is

identically zero, then $S_0(y) \equiv 1$, and that otherwise $S_0(y) < 1$ for all $y > 0$. Thus, equation (4) is an integral equation of the type used to represent infections in endemic equilibrium without involving frailties or individual heterogeneity (Greenhalgh 1990), and so existing theory for such equations may be applied. Notably, if $\lambda_0(x) > 0$ then $R_0 > 1$; and if $R_0 > 1$ then there exists a nonzero solution $\lambda_0(x)$.

2.5 An explicit expression for the basic reproduction number

Let $l_1(x)$ denote the leading left eigenfunction of the operator B_1^*, which has kernel $D\beta_0(x,y)\{1+\gamma(y)\}f(y)$. A simple manipulation of equation (4) yields

$$R_0 = \frac{\int_0^\infty l_1(x)\lambda_0(x)\mathrm{d}x}{\int_0^\infty l_1(x)\lambda_0(x)S_0(x)\mathrm{d}x}. \tag{6}$$

Note that if $l_2(x)$ is the leading left eigenfunction of the operator B_2^*, which has kernel $Df(x)\{1+\gamma(x)\}\beta_0(x,y)$, then we also have

$$R_0 = \frac{\int_0^\infty l_2(x)\{1+\gamma(x)\}\lambda_0(x)f(x)\mathrm{d}x}{\int_0^\infty l_2(x)\{1+\gamma(x)\}\lambda_0(x)S_0(x)f(x)\mathrm{d}x}. \tag{7}$$

This expression generalizes one derived in Farrington, Kanaan and Gay (2001); it can be easier to handle than (6) as the term $\{1+\gamma(x)\}S_0(x)$ in the denominator simplifies. Equations (6) and (7) provide the basis for a new estimation method, described in the next section. Henceforth it is assumed that the population age structure $f(x)$ is known; it can readily be derived from a life table for the population.

3 Estimation of R_0

3.1 Social contact surveys and serological survey data

We begin by briefly describing the types of data from which we shall endeavour to draw inferences on R_0.

A social contact function $C(x,y)$, relevant to the transmission of the infection of interest, is a non-negative bivariate function proportional to the mean effective contact

rate function $\beta_0(x, y)$. Thus there is a positive constant q such that

$$\beta_0(x, x) = qC(x, y). \tag{8}$$

An estimate $\hat{C}(x, y)$ of $C(x, y)$ may be obtained from a contact survey, in which 'contacts' are defined using a suitably chosen proxy variable (Wallinga, Teunis and Kretzschmar 2006, Mossong et al. 2008). Thus, $\hat{C}(x, y)$ denotes the estimated number of proxy contacts per unit time that a typical individual of age y makes with a single typical individual of age x.

Serological survey data are a readily available source of information on individual infection histories, and will be used, notably, to estimate the baseline infection hazard $\lambda_0(x)$. Serum samples taken from individuals of age x are tested for the presence of antibodies to one or more infections. A positive test result indicates that the individual has been infected in the past; a negative result indicates that the individual has not been infected. Serological data are thus current status data.

3.2 Estimation of R_0 with individual heterogeneity

In order to estimate R_0 in the presence of individual heterogeneity, an estimate is required for $\gamma(x)$. This may be obtained from a shared frailty model for two infections: the infection of primary interest, and a second infection transmitted by the same route (Farrington, Unkel and Anaya-Izquierdo 2012, Unkel et al. 2012). Since the two infections share the same route of transmission, it may be reasonable to assume that the individual frailties relevant to the transmission of infection are the same for the two infections and, furthermore, that the effective contact functions are proportional. It then follows from equation (3) that the hazards of infection for the infection of interest, $\lambda_1(x, u(x))$ and for the other infection, $\lambda_2(x, u(x))$, are linked via the shared frailty model:

$$\lambda_1(x, u(x)) = u(x)\lambda_{01}(x), \quad \lambda_2(x, u(x)) = u(x)\lambda_{02}(x),$$

where $\lambda_{01}(x)$ and $\lambda_{02}(x)$ are the baseline hazards for infections 1 and 2, respectively, and $u(x)$ is the shared frailty. In addition, if $u(x)\beta_{01}(x,y)v(y)$ and $u(x)\beta_{02}(x,y)v(y)$ denote the effective contact functions for each infection, we have

$$\begin{aligned}\beta_{01}(x,y) &= q_1 C(x,y),\\ \beta_{02}(x,y) &= q_2 C(x,y)\end{aligned}$$

for distinct values $q_1, q_2 > 0$.

Estimation of the variance of $u(x)$, and hence of $\gamma(x)$, then proceeds by specifying a suitable parametric model for the distribution of $u(x)$. To estimate R_0, two methods are available; one is based on equation (7); the other extends existing methods applicable when there is no individual heterogeneity.

3.3 The eigenfunction method

We begin with the new method, which does not require q_1 or q_2 to be estimated. Suppose that serum samples on n_x individuals of age x are available, for $x = 1, \ldots, M$. Let n_{x00} denote the number of individuals of age x not previously infected by either infection, n_{x01} the number previously infected by infection 2 but not by infection 1, n_{x10} the number previously infected by infection 1 but not infection 2, and n_{x11} the number previously infected by both infections. Thus the serological data comprise 4-tuples $(n_{x00}, n_{x01}, n_{x10}, n_{x11})$. The expected proportions in the four cells are given by the population survivor functions from the shared frailty model:

$$\begin{aligned}p_{x00} &= E\Big\{\exp\Big(-\int_0^x u(y)(\lambda_{01}(y)+\lambda_{02}(y))\mathrm{d}y\Big)\Big\},\\ p_{x01} &= E\Big\{\exp\Big(-\int_0^x u(y)\lambda_{01}(y)\mathrm{d}y\Big)\Big\} - p_{x00},\\ p_{x10} &= E\Big\{\exp\Big(-\int_0^x u(y)\lambda_{02}(y)\mathrm{d}y\Big)\Big\} - p_{x00},\\ p_{x11} &= 1 - p_{x00} - p_{x01} - p_{x10}.\end{aligned}$$

The expectations are taken over the random variable $u_1, \ldots, u_k$. The baseline hazards $\lambda_{01}(x)$ and $\lambda_{02}(x)$ are specified parametrically. To allow for extra-multinomial

variation owing to assay variability, we model the data as Dirichlet-multinomial, with dispersion parameter $\phi \in (0, 1]$, the limit $\phi = 0$ corresponding to the multinomial. Setting $\psi = (1 - \phi)/\phi$, the log likelihood kernel is then

$$l = \sum_{x=1}^{M} \Big\{ \log\Big(\frac{\Gamma(\psi)}{\Gamma(n_x + \psi)}\Big) + \log\Big(\frac{\Gamma(n_{x00} + \psi p_{x00}))}{\Gamma(\psi p_{x00})}\Big) + \log\Big(\frac{\Gamma(n_{x01} + \psi p_{x01}))}{\Gamma(\psi p_{x01})}\Big) + \log\Big(\frac{\Gamma(n_{x10} + \psi p_{x10}))}{\Gamma(\psi p_{x10})}\Big) + \log\Big(\frac{\Gamma(n_{x11} + \psi p_{x11}))}{\Gamma(\psi p_{x11})}\Big)\Big\}. \quad (9)$$

Maximization of the log likelihood yields estimates $\hat{\lambda}_{01}(x), \hat{\lambda}_{02}(x)$ and $\hat{\gamma}(x)$. We then obtain the left leading eigenvector $\hat{l}_2(x)$ of $f(x)\{1 + \hat{\gamma}(x)\}\hat{C}(x, y)$, and the estimated functions $\hat{S}_{01}(x), \hat{S}_{02}(x)$ from equation (5). Finally, these components are assembled using equation (7) to obtain the reproduction number of the infection of primary interest:

$$\hat{R}_0 = \frac{\int_0^\infty \hat{l}_2(x)\{1 + \hat{\gamma}(x)\}\hat{\lambda}_{01}(x) f(x) \mathrm{d}x}{\int_0^\infty \hat{l}_2(x)\{1 + \hat{\gamma}(x)\}\hat{\lambda}_{01}(x)\hat{S}_{01}(x) f(x) \mathrm{d}x}. \quad (10)$$

In practice, all quantities present in these equations are discretized, preferably in narrow age intervals, typically one year, up to some maximum age. Thus the integrals in equations (10) are replaced by sums.

3.4 The q-factor method

We now describe the other estimation method. This extends the method used for low-dimensional contact matrices (Farrington, Kanaan and Gay 2001).

The basic reproduction numbers R_{0i}, $i = 1, 2$, are the leading eigenvalues of the operators $Df(x)\beta_{0i}(x, y)\{1+\gamma(y)\}$. The idea is to obtain the R_{0i} directly by substituting estimated values into this operator, suitably discretized as a matrix. To this end, however, estimates of the constants q_i are required, since $\beta_{0i}(x, y)$ must be estimated by $\hat{q}_i\hat{C}(x, y)$.

The q_i are estimated by calibrating the estimated social contact function against the serological data. Substituting $q_i C(x,y)$ in equation (4), we obtain the integral equations

$$\lambda_{0i}(x) = q_i D \int_0^\infty C(x,y)\{1+\gamma(y)\}S_{0i}(y)\lambda_{0i}(y)f(y)\mathrm{d}y, \tag{11}$$

for $i = 1, 2$, where

$$S_{0i}(y) = \{1+\gamma(y)\}^{-1} \int_{a(y)}^{b(y)} v(y)^2 \exp\Big(- \int_0^y v(z)\lambda_{0i}(z)\mathrm{d}z\Big) f(v(y)|y)\mathrm{d}v(y).$$

Since the infections are endemic, equations (11) have nonzero solutions $\lambda_{0i}(x)$. For given values q_i and $\gamma(x)$, these solutions may be found by iterating equations (11). The corresponding values of $p_{00}(x), p_{01}(x), p_{10}(x)$ and $p_{11}(x)$ may then be obtained, whence the log likelihood from equation (9) may be evaluated. Note that this log likelihood is now a function only of q_1, q_2 and the parameters in $\gamma(x)$. The maximum likelihood estimators so derived may then be used to obtain an estimate of R_0 as the leading eigenvalues of the operator $\hat{q}_1 D f(x)\hat{C}(x,y)(x,y)\{1+\hat{\gamma}(y)\}$.

4 Robustness

Both methods of estimation are maximum likelihood methods. The q-factor method requires estimation of fewer parameters, so is likely to be more efficient. However, a key factor in deciding which of the two methods to use is their robustness to misspecification of the social contact structure $C(x,y)$.

The eigenfunction method requires knowledge only of the leading left eigenfunction of $f(x)\{1+\gamma(y)\}C(x,y)$, rather than approximating $\beta_0(x,y)$ by $qC(x,y)$, where q must be estimated. The estimate of q is a complicated function of $C(x,y)$ involving its whole spectral structure. To guarantee unbiasedness, the q-factor method requires the strong assumption (8), restated for convenience as

$$\begin{gathered} C(x,y) \text{ and } \beta_0(x,y) \\ \text{are proportional.} \end{gathered} \tag{12}$$

In contrast, the eigenfunction method requires only that

$$f(x)\{1+\gamma(x)\}C(x,y) \text{ and } f(x)\{1+\gamma(x)\}\beta_0(x,y)$$
$$\text{have proportional leading left eigenfunctions.} \quad (13)$$

Assumption (12) implies assumption (13), but not vice-versa. This suggests that the eigenfunction method may be more robust to mis-specification of the social contact function $C(x,y)$. In practice neither assumption is likely to be met exactly: social contact functions are estimated by measuring proxy variables for the elusive 'effective contacts' to be described, and so are at best approximate. General results from perturbation theory exist about the impact of matrix errors on eigenvalues and eigenvectors, but are not well-suited for our present purpose (Golub and Van Loan 1996, pages 320-330). Therefore, to investigate further the robustness of the two approaches, we proceed by simulation.

Since our focus is on mis-specification of $C(x,y)$, and to keep the simulations simple, we undertook simulations in the absence of individual heterogeneity, so that $\gamma(x)=0$. In this case it is not necessary to use data on a second infection. We took $\beta_0(x,y)$ to be proportional to the 16×16 social contact matrix for England and Wales estimated in Farrington et al. (2009), each age group corresponding to a 5-year age band. Let C_{true} denote this matrix. The density $f(x)$ was uniform $U(0,80)$. The matrix $\beta_0(x,y)$ was scaled to obtain a true value of R_0 of 3.02. The expected infection hazard and survivor function were then obtained by solving the integral equation (4).

We studied 12 distinct, realistic scenarios. In the first, we took $C(x,y)=C_{\text{true}}$, so that $\beta_0(x,y)$ and $C(x,y)$ were proportional - as specified by assumption (12). In the other 11 scenarios we perturbed $C(x,y)$: five random perturbations, and six systematic perturbations. For all the random and five of the systematic perturbations, assumption (12) is violated; in the remaining scenario the scaling factor between $\beta_0(x,y)$ and $C(x,y)$ was doubled. For each scenario, we generated 100 simulated serological data sets with annual sample sizes $n_x=100, 500$ or 1000, and estimated R_0 using the q-factor

and eigenfunction methods. The bias and empirical standard error were computed for each combination of scenario, sample size and estimation method. The results are in Tables 1 and 2, where the perturbations are defined, c_{ij} referring to the $(i,j)^{\text{th}}$ entry of C_{true}.

Both estimation methods produce unbiased results when assumption (12) is satisfied, though as expected the standard error is greater using the eigenfunction method. In all the perturbation scenarios except two, the bias in R_0 from using the q-factor method is greater in absolute value than the bias in R_0 from using the eigenfunction method. One exception is scenario 4 in Table 1, for which the bias is very small for both methods. The other exception, for which the bias is greater, is scenario 10 in Table 2. In this scenario, the c_{ij} are doubled for $i, j \leq 2$, that is, the assumed contact rates in children aged 0 to 10 years are overestimated by a factor of 2 compared to other ages. In practice it is far more likely that the contact rates in this age group are underestimated: in this scenario (scenario 9) the eigenfunction method slightly outperforms the q-factor method. The largest bias for the q-factor method is 30% in relative terms, which occurs when the diagonal entries c_{ii} are overstated (scenario 8); the largest bias for the eigenfunction approach is 20%, when the entries c_{ij} with $i, j \leq 2$ are overstated (scenario 10).

5 Application to varicella zoster virus infection

We illustrate these methods with data from England and Wales on varicella zoster virus (VZV) infection, which causes varicella and shingles. A vaccine against VZV infection exists, but is not currently recommended for routine use in the UK. The critical immunization threshold at birth which is required to interrupt transmission of the infection is $1 - R_0^{-1}$, whence the interest in estimating R_0.

We use serological data for persons aged 1 - 20 years collected by the Health Protection Agency in 1996. To estimate heterogeneity in transmission we use paired data

on parvovirus B19, an infection which, like VZV, is transmitted by droplets and close contact with respiratory secretions. The VZV data are described in Vyse et al. (2004), and the parvovirus B19 data in Vyse et al. (2007). The full data comprise 1747 serum samples for which paired assay results on both VZV and B19 are available, and 344 for which results on VZV only are available.

The social contact data were obtained from a survey of conversational contacts undertaken in England and Wales, as part of a wider European project (Mossong et al. 2008). The contact matrix we used is stratified in 5-year age groups from ages 0-4 to 75+ (Farrington et al. 2009), and is regarded here as a fixed quantity. To match this, we parameterized the baseline hazards of infection, $\lambda_{01}(x)$ and $\lambda_{02}(x)$, as piecewise constant on 5-year age groups.

The age distribution $f(x)$ was obtained from the life table for England for 1995-1997 from the Office for National Statistics (www.ons.gov.uk).

The heterogeneity was described by the following time-varying frailty model, introduced in Farrington, Unkel and Anaya-Izquierdo (2012) and investigated in more detail in Unkel et al. (2012):

$$u(x) = w(x, u_1, u_2) = \{1 + (u_1 - 1)\exp(-\rho x^2)\}u_2,$$

where u_1 and u_2 are independent random variables distributed gamma with mean 1 and variances γ_1 and γ_2, and $\rho \geq 0$. Thus,

$$\gamma(x) = \gamma_1(1 + \gamma_2)\exp(-\rho x^2) + \gamma_2.$$

This model is motivated by empirical considerations, and the need to allow flexibly for the presence of different types of individual heterogeneity at different ages (Farrington, Unkel and Anaya-Izquierdo 2012, Unkel et al. 2012). Thus, the term involving u_1 describes the heterogeneity of childhood behaviors and environments, such as nursery attendance, which tends to decline rapidly with age, as represented by the $\exp(-\rho x^2)$ function, whereas the term u_2 describes potential long-lasting heterogeneity in adult

behavior. With this choice, $S_0(y)$ in equation (5), pre-multiplied by $\{1+\gamma(y)\}$, is

$$\begin{aligned}\{1+\gamma(y)\}S_0(y) &= \int_0^\infty \Big[\{1-h(y)^2\}\{1+uP(y)\gamma_1\}^{-1/\gamma_1} \\ &\quad +2h(y)\{1-h(y)\}\{1+uP(y)\gamma_1\}^{-1-1/\gamma_1} \\ &\quad +h(y)^2(1+\gamma_1)\{1+uP(y)\gamma_1\}^{-2-1/\gamma_1}\Big] \\ &\quad \times\Gamma(1/\gamma_2)^{-1}u^{1+1/\gamma_2}(1/\gamma_2)^{1/\gamma_2}\exp\{-uQ(y)+u/\gamma_2\}\mathrm{d}u,\end{aligned}$$

where $h(y)=\exp(-\rho y^2)$, $P(y)=\int_0^y h(z)\lambda_0(z)\mathrm{d}z$, and $Q(y)=\int_0^y h(z)\{1-h(z)\}\lambda_0(z)\mathrm{d}z$. These integrations are done numerically.

The q-factor method requires the integral equations (11) to be solved for λ_{0i}, $i=1,2$, for given values of q_1, q_2 and the parameters in $\gamma(x)$. This is readily achieved by iterating equations (11).

Models were fitted using both the q-factor and the eigenfunction method, with and without heterogeneity. Note that, for each estimation method, the model without heterogeneity is nested within the model with heterogeneity. The models for the two estimation methods are not nested. Confidence intervals for R_0 were obtained using the profile likelihood, using a Lagrange multiplier applied to $\log(R_0)$ (Critchley, Ford and Rijal 1988).

The results are shown in Table 3. They indicate that the q-factor and eigenfunction methods yield broadly comparable results, though the confidence intervals are slightly wider for the eigenfunction method, as expected. The q-factor method without heterogeneity produces a rather bad fit to the serological data: the hazard of infection from equation (11) is too low at young ages and too high at older ages. This suggests that the social contact matrix $C(x,y)$ is mis-specified for these infections, as also found by Goeyvaerts et al. (2010). Allowing for heterogeneity produces a marked improvement in fit, notably for the q-factor method. Doubts about the appropriateness of $C(x,y)$ suggest that the generally more robust estimator of R_0 based on the eigenfunction method should be preferred, and evidence of improved fit suggest that the effect of heterogeneity should be taken into account.

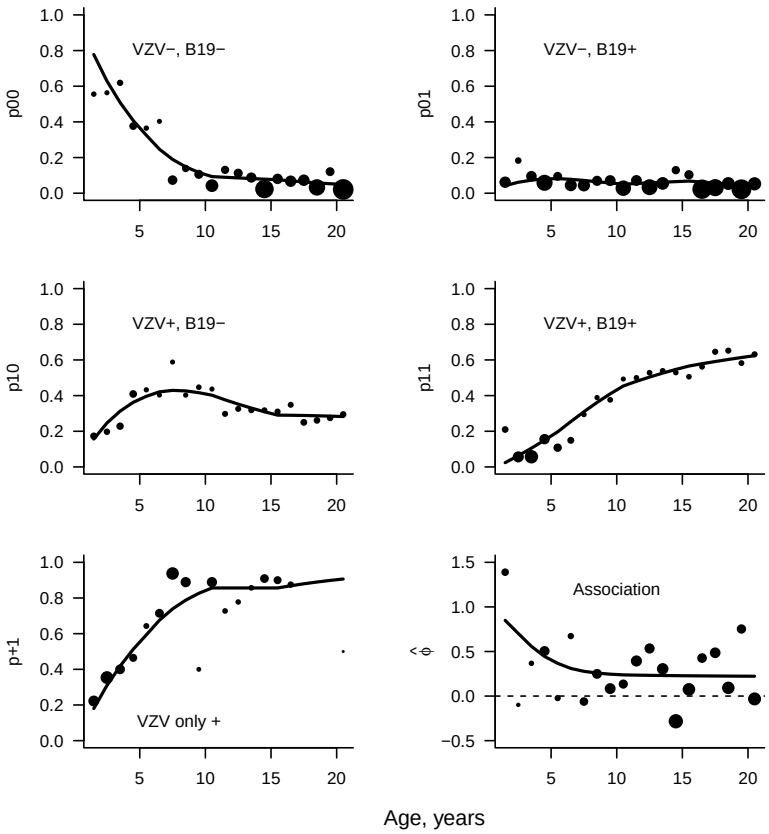

Figure 1: Data (dots) and fit (full lines) of the eigenfunction model with heterogeneity (see text).

Figure 1 shows the data and the fitted model using the eigenfunction method with heterogeneity. The areas of the plotting symbols are proportional to the empirical precisions; for the top four panels they are on the same scale.

The top four panels show the distribution of the 4-tuples $(n_{x00}, n_{x01}, n_{x10}, n_{x11})$ at $x = 1, \ldots, 20$, along with the fitted values, while the bottom left panel shows the marginal fit for the 344 samples with only VZV information. The model fits are generally good.

The bottom right panel of Figure 1 shows the observed and fitted values of the measure of association ϕ described in Unkel and Farrington (2012); the dashed line at

0 represents zero association, corresponding to $\phi = \gamma(x) = 0$. Most points and the fitted line lie above the line $\phi = 0$, indicating a positive association. There is some weak evidence of stronger association (larger ϕ values) and hence greater heterogeneity at lower ages, as observed with many other infections (Unkel et al. 2012).

Our best estimate for R_0 for varicella zoster virus infection in England and Wales is therefore 5.3, with 95% confidence interval (3.5, 10.5). It follows that the critical immunization level required to interrupt endemic transmission of VZV infection in England and Wales is in excess of 81%, with 95% confidence interval $(71\%, 90\%)$.

6 Final comments

The individual heterogeneity attending effective contacts relevant to the transmission of infection is generally ignored in the estimation of reproduction numbers, and thus R_0 estimates have consequently tended to be underestimated. While the problem is well recognized, there has so far been little work to remedy it. The same applies to sensitivity of estimation of R_0 to mis-specification of the contact function.

Our approach suffers from several limitations; we mention three. The first is the restriction to frailties of the form $u(x) = w(x, u_1, \ldots, u_k)$ where $u_1, \ldots, u_k$ are time-invariant random variables. Extensions to more general stochastic processes are desirable in principle, such as those discussed by Aalen, Borgan and Gjessing (2008), Chapters 10 and 11, though more informative data than the current status data we have used would likely be required to identify such models. A second modelling limitation is the multiplicative frailty framework we have employed, which in effect imposes a proportional mixing structure on the random effects. Third, while the eigenfunction method is often less biased than the q-factor method, it is not clear to what types of mis-specification of the social contact function $C(x, y)$ it is, or is not, robust. Further work to elucidate this issue would be welcome.

The key novel ingredient required for estimating R_0 in the presence of heterogeneity

is the term $\gamma(x)$, the variance of $u(x)$. The estimation method we propose makes use of shared frailty models, applied to paired data on infections sharing a transmission route. Further development of this approach, and of different methods for estimating $\gamma(x)$, are desirable.

Acknowledgements

We thank Dr Richard Pebody of the Health Protection Agency for permission to use the serological data. This research was supported by a project grant from the Medical Research Council and a Royal Society Wolfson Research Merit Award.

References

Aalen, O.O., Borgan, Ø. and Gjessing, H. (2008), *Survival and Event History Analysis: A Process Point of View*. New York: Springer.

Critchley, F., Ford, I. and Rijal, O. (1988), "Interval estimation based on the profile likelihood: strong Lagrangian theory with applications to discrimination", *Biometrika*, 75, 21-28.

Diekmann, O., Heesterbeek, J.A.P. and Metz, J.A.J. (1990), "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations", *J. Math. Biol.*, 28, 365-382.

Dietz, K. (1993), "The estimation of the basic reproduction number for infectious diseases", *Statist. Meth. Med. Res.*, 2, 23-41.

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

Farrington, C.P., Unkel, S. and Anaya-Izquierdo, K. (2012), "The relative frailty variance and shared frailty models", *J. R. Statist. Soc.*, Series B, 74, In Press. Pub-

lished online 15 Feb 2012, DOI: 10.1111/j.1467-9868.2011.01021.x.

Farrington, C.P., Whitaker, H.J., Wallinga, J. and Manfredi, P. (2009), "Measures of disassortativeness and their application to directly transmitted infections", *Biometrical J.*, 51, 387-407.

Goeyvaerts, N., Hens, N., Ogunjimi, B. et al. (2010), "Estimating infectious disease parameters from data on social contacts and serological markers", *J. R. Statist. Soc*, Series C, 59, 255-277.

Golub, G. H. and Van Loan, C. F. (1996), *Matrix Computations* (3rd ed.), Baltimore: Johns Hopkins University Press.

Greenhalgh D. (1990), "Vaccination campaigns for common childhood diseases", *Math. Biosci.*, 100, 201-240.

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

Jörgens, K. (1982), *Linear Integral Operators*. London: Pitman Press.

Melegaro, A., Jit, M., Gay, N., Zagheni, E. and Edmunds, W.J. (2011), "What contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns", *Epidemics*, 3, 143-151.

Mossong, J., Hens, N., Jit, M. et al. (2008), "Social contacts and mixing patterns relevant to the spread of infectious diseases", *PLoS Med.*, 5, E74.

Unkel, S. and Farrington, C.P. (2012), "A new measure of the time-varying association for shared frailty models with bivariate current status data", *Biostatistics*, In Press. Published online 23 April 2012, DOI: 10.1093/biostatistics/kxs010.

Unkel, S., Farrington, C.P., Whitaker, H. and Pebody, R. (2012), "Time-varying frailty models and the estimation of heterogeneities in transmission of infectious diseases", Open University Statistics Group Technical Report 12/03. Available from http://statistics.open.ac.uk/technical_report.

Vyse A.J., Gay N.J., Hesketh L.M., Morgan-Capner P., and Miller E. (2004), "Seroprevalence of antibody to varicella zoster virus in England and Wales in children and

young adults", *Epidemiol. Infect.*, 132, 1129-1134.

Vyse A.J., Andrews N.J., Hesketh L.M. and Pebody R. (2007), "The burden of parvovirus B19 infection in women of childbearing age in England and Wales", *Epidemiol. Infect.*, 135, 1354-1362.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006), "Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents", *Am. J. Epidemiol.*, 164, 936-944.

Table 1: Bias and empirical standard error (s.e.) of $\hat{R}_0$ for five random perturbation structures and error-free case and three different sample sizes n_x obtained from $N = 100$ simulated samples or the q-factor method (Q) and eigenfunction method (E).

Scenario		Method		$n_x = 100$	$n_x = 500$	$n_x = 1000$
1	No error	Q	bias	-0.0008	0.0073	0.0054
			empirical s.e.	0.0380	0.0193	0.0134
		E	bias	0.0120	0.0137	0.0093
			empirical s.e.	0.0608	0.0295	0.0205
2	$c_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim U(0, 0.5)$	Q	bias	0.2877	0.2833	0.2855
			empirical s.e.	0.0495	0.0176	0.0149
		E	bias	0.1841	0.1293	0.1255
			empirical s.e.	0.1078	0.0352	0.0277
3	$c_{ij}(1 + \epsilon_{ij})$ with $\epsilon_{ij} \sim U(0, 2.0)$	Q	bias	0.1510	0.1500	0.1455
			empirical s.e.	0.0427	0.0204	0.0160
		E	bias	0.0984	0.0900	0.0835
			empirical s.e.	0.0743	0.0363	0.0279
4	$c_{ij}(1 + \epsilon_{ij})^{-1}$ with $\epsilon_{ij} \sim U(0, 0.5)$	Q	bias	-0.0137	0.0156	0.0064
			empirical s.e.	0.0444	0.1783	0.1674
		E	bias	0.0395	0.0205	0.0122
			empirical s.e.	0.0743	0.0328	0.0202
5	$c_{ij}(1 + \epsilon_{ij})^{-1}$ with $\epsilon_{ij} \sim U(0, 2.0)$	Q	bias	0.3207	0.3169	0.3189
			empirical s.e.	0.0470	0.0209	0.0177
		E	bias	0.2499	0.2208	0.2265
			empirical s.e.	0.0884	0.0379	0.0323
6	$c_{ij}(1 - \epsilon_{ij})^{-1}$ with $\epsilon_{ij} \sim U(0, 0.5)$	Q	bias	0.4900	0.4779	0.4887
			empirical s.e.	0.1313	0.0257	0.1142
		E	bias	-0.2084	-0.2243	-0.2289
			empirical s.e.	0.0557	0.0291	0.0205

Table 2: Bias and empirical standard error (s.e.) of $\hat{R}_0$ for six systematic perturbation structures and three different sample sizes n_x obtained from $N = 100$ simulated samples for the q-factor method (Q) and eigenfunction method (E).

Scenario		Method		$n_x = 100$	$n_x = 500$	$n_x = 1000$
7	$c_{ij} \times 2$ $\forall\ i, j$	Q	bias	0.0054	0.0085	0.0066
			empirical s.e.	0.0418	0.0232	0.0178
		E	bias	0.0273	0.0141	0.0097
			empirical s.e.	0.0645	0.0304	0.0209
8	$c_{ij} \times 2$ for $i = j$	Q	bias	0.9053	0.9030	0.9150
			empirical s.e.	0.1599	0.1237	0.1663
		E	bias	-0.3234	-0.3307	-0.3351
			empirical s.e.	0.0689	0.0290	0.0179
9	$c_{ij} \times 0.5$ for $i, j \leq 2$	Q	bias	0.6753	0.6717	0.6672
			empirical s.e.	0.0651	0.0299	0.0196
		E	bias	0.5415	0.5153	0.5095
			empirical s.e.	0.1001	0.0444	0.0278
10	$c_{ij} \times 2$ for $i, j \leq 2$	Q	bias	-0.2887	-0.2852	-0.2870
			empirical s.e.	0.0355	0.0155	0.0112
		E	bias	-0.6111	-0.6081	-0.6103
			empirical s.e.	0.0506	0.0248	0.0159
11	$c_{ij} + 2$ for $j \geq i - 1$ and $j \leq i + 1$	Q	bias	0.1679	0.1605	0.1633
			empirical s.e.	0.0457	0.0254	0.0160
		E	bias	0.1380	0.1206	0.1208
			empirical s.e.	0.0697	0.0296	0.0236
12	$c_{ij} \times 0.5$ for $j \geq i - 1$ and $j \leq i + 1$	Q	bias	0.5914	0.5927	0.5896
			empirical s.e.	0.0569	0.0265	0.0152
		E	bias	0.2682	0.2491	0.2395
			empirical s.e.	0.0921	0.0374	0.0238

Table 3: Estimated values of R_0 for varicella zoster virus in the absence and presence of heterogeneity, obtained using the q-factor method (Q) and eigenfunction method (E).

Setting	Method	R_0	95% CI	-2 loglik	no. parameters
Ignoring heterogeneity	Q	3.93	(3.44, 4.48)	4281.43	3
	E	3.18	(2.17, 4.52)	4226.95	9
Allowing heterogeneity	Q	5.12	(2.59, 8.89)	4220.43	6
	E	5.33	(3.49, 10.47)	4213.20	12