Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2010)
59, *Part* 2, *pp.* 255–277

Estimating infectious disease parameters from data on social contacts and serological status

Nele Goeyvaerts,

Hasselt University, Diepenbeek, Belgium

Niel Hens,

Hasselt University, Diepenbeek, and University of Antwerp, Belgium

Benson Ogunjimi,

University of Antwerp, Belgium

Marc Aerts and Ziv Shkedy

Hasselt University, Diepenbeek, Belgium

and Pierre Van Damme and Philippe Beutels

University of Antwerp, Belgium

[Received August 2008. Final revision July 2009]

Summary. In dynamic models of infectious disease transmission, typically various mixing patterns are imposed on the so-called 'who acquires infection from whom' matrix. These imposed mixing patterns are based on prior knowledge of age-related social mixing behaviour rather than observations. Alternatively, we can assume that transmission rates for infections transmitted predominantly through non-sexual social contacts are proportional to rates of conversational contact which can be estimated from a contact survey. In general, however, contacts reported in social contact surveys are proxies of those events by which transmission may occur and there may be age-specific characteristics that are related to susceptibility and infectiousness which are not captured by the contact rates. Therefore, we model transmission as the product of two age-specific variables: the age-specific contact rate and an age-specific proportionality factor, which entails an improvement of fit for the seroprevalence of the varicella zoster virus in Belgium. Furthermore, we address the effect on the estimation of the basic reproduction number, using non-parametric bootstrapping to account for different sources of variability and using multimodel inference to deal with model selection uncertainty. The method proposed makes it possible to obtain important information on transmission dynamics that cannot be inferred from approaches that have been traditionally applied hitherto.

Keywords: Basic reproduction number; Bootstrap procedure; Model averaging; Model selection; Social contact data; Transmission parameters; Who acquires infection from whom matrix

1. Introduction

The first approach in modelling transmission dynamics of infectious diseases, and more particularly in estimating age-dependent transmission rates, was described by Anderson and May

Address for correspondence: Nele Goeyvaerts, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Campus Diepenbeek, Agoralaan 1 Gebouw D, B-3590 Diepenbeek, Belgium.
E-mail: nele.goeyvaerts@uhasselt.be

 0035–9254/10/59255

(1991). The idea is to impose different mixing patterns on the so-called 'who acquires infection from whom' (WAIFW) matrix β_{ij}, thereby constraining the number of distinct elements for identifiability reasons, and to estimate the parameters from serological data. Many researchers have elaborated on this approach of Anderson and May (1991), among whom are Greenhalgh and Dietz (1994), Farrington *et al.* (2001) and Van Effelterre *et al.* (2009). However, estimates of important epidemiological parameters such as the basic reproduction number R_0 turn out to be sensitive with respect to the choice of the mixing pattern imposed (Greenhalgh and Dietz, 1994).

An alternative method was proposed by Farrington and Whitaker (2005), where contact rates are modelled as a continuous contact surface and estimated from serological data. Clearly, both methods involve a somewhat *ad hoc* choice, namely the structure for the WAIFW matrix and the parametric model for the contact surface. Alternatively, to estimate age-dependent transmission parameters, Wallinga *et al.* (2006) augmented seroprevalence data with auxiliary data on self-reported numbers of conversational contacts per person, while assuming that transmission rates are proportional to rates of conversational contact. The social contact surveys that were conducted as part of the 'Improving public health policy in Europe through modelling and economic evaluation of interventions for the control of infectious diseases' (known as 'POLYMOD') project (Mossong *et al.*, 2008b; Hens *et al.*, 2009a) allow us to elaborate on the methodology that was presented by Wallinga *et al.* (2006).

The paper is organized as follows. In the next section, we outline the build-up of the Belgian social contact survey and the information that is available for each contact. Further, we briefly explain the epidemiological characteristics of varicella zoster virus (VZV) and the serological data from Belgium that we use. In Section 3, we illustrate the traditional approach of imposing mixing patterns to estimate the WAIFW matrix from this serological data set. In Section 4, a transition is made to the novel approach of using social contact data to estimate R_0. We show that a bivariate smoothing approach allows for a more flexible and better estimate of the contact surface compared with the maximum likelihood (ML) estimation method of Wallinga *et al.* (2006). Further, some refinements are proposed, among which is an elicitation of contacts with high transmission potential and a non-parametric bootstrap approach, assessing sampling variability and accounting for age uncertainty, as suggested by Halloran (2006).

Our main result is the novel method of disentangling the WAIFW matrix into two components: the contact surface and an age-dependent proportionality factor. The method proposed, as described in Section 5, tackles two dimensions of uncertainty. First, by estimating the contact surface from data on social contacts, we overcome the problem of choosing a completely parametric model for the WAIFW matrix. Second, to overcome the problem of model selection for the age-dependent proportionality factor, concepts of multimodel inference are applied and a model-averaged estimate for R_0 is calculated. Some concluding remarks are provided in the last section. The data sets and R code that are used in this paper are available from the authors on request.

2. Data

2.1. Belgian contact survey

Several small-scale surveys were made to gain more insight into social mixing behaviour that is relevant to the spread of close contact infections (Edmunds *et al.*, 1997, 2006; Beutels *et al.*, 2006; Wallinga *et al.*, 2006; Mikolajczyk and Kretzschmar, 2008). To refine on contact information, a large multicountry population-based survey was conducted in Europe as part of the POLYMOD project (Mossong *et al.*, 2008b).

In Belgium, this survey was conducted in a period from March until May 2006. A total of 750 participants, selected through random-digit dialling, completed a diary-based questionnaire about their social contacts during one randomly assigned weekday and one randomly assigned day in the weekend (not always in that order). In this paper, we follow the sampling scheme of the POLYMOD project and consider only one day for each participant (Mossong *et al.*, 2008b). The data set consists of participant-related information such as age and gender, and details about each contact: age and gender of the contacted person, and location, duration and frequency of the contact. In case the exact age of the contacted person was unknown, participants had to provide an estimated age range and the mean value is used as a surrogate. Further, a distinction between two types of contact was made: non-close contacts, defined as two-way conversations of at least three words in each other's proximity, and close contacts that involve any sort of physical skin-to-skin touching.

Teenagers (9–17 years old) filled in a simplified version of the diary and were closely followed up to anticipate problems of interpretation. For children (under 9 years old), a parent or exceptionally another adult caregiver filled in the diary. One adult respondent made over 1000 contacts and was considered an outlier to the data set. This person is likely to be very influential and therefore was excluded from the analyses that are presented here. Analyses are based on the remaining 749 participants. Using census data on population sizes of different age by household size combinations, weights are given to the participants to make the data representative of the Belgian population. In total, the 749 participants recorded 12775 contacts of which three are omitted from analysis owing to missing age values for the contacted person. For a more detailed perspective on the Belgian contact survey and the importance of contact rates on modelling infectious diseases, we refer to Hens *et al.* (2009a).

2.2. Serological data

Primary infection with VZV, which is also known as human herpes virus type 3, results in varicella, which are commonly known as chickenpox, and mainly occur in childhood. Afterwards, the virus becomes dormant in the body and may reactivate at a later stage, resulting in herpes zoster, which is commonly known as shingles. Infection with VZV occurs through direct or aerosol contact with infected people. A person who is infected with chickenpox can transmit the virus for about 7 days. Following Garnett and Grenfell (1992) and Whitaker and Farrington (2004), we ignore chickenpox cases resulting from contact with people who are suffering from shingles. Zoster indeed has a limited effect on transmission dynamics when considering large populations with no immunization programme (Ferguson *et al.*, 1996).

In a period from November 2001 until March 2003, 2655 serum samples in Belgium were collected and tested for VZV. Together with the test results, gender and age of the individuals were recorded. In the data set, age ranges from 0 to 40 years and six individuals were younger than 6 months. Belgium has no mass vaccination programme for VZV. Further details on the data set can be found in Hens *et al.* (2008, 2009b).

3. Estimation of R_0 by imposing mixing patterns

3.1. Estimating transmission rates

To describe transmission dynamics, a compartmental maternally derived immunity–susceptible–infectious–recovered model for a closed population of size N is considered. By doing so, we explicitly take into account the fact that, in the first phase, newborns are protected by maternal antibodies and do not take part in the transmission process. We assume that mortality due to

infection can be ignored, which is plausible for VZV in developed countries, and that infected individuals maintain lifelong immunity after recovery. Further, demographic and endemic equilibria are assumed, which means that the age-specific population sizes remain constant over time and that the disease is in an endemic steady state at the population level. For simplicity, we assume type I mortality defined as

$$\exp\left\{-\int_0^a \mu(s)\,\mathrm{d}s\right\} = \begin{cases} 1, & \text{if } a < L, \\ 0, & \text{if } a \geqslant L, \end{cases}$$

where $\mu(a)$ denotes the age-specific mortality rate. This implies that everyone survives up to age L and then promptly dies, which is a reasonable assumption when describing transmission dynamics for VZV in Belgium (see also Whitaker and Farrington (2004)). We make a similar assumption for the age-specific rate $\gamma(a)$ of losing maternal antibodies, which we shall denote as 'type I maternal antibodies':

$$\exp\left\{-\int_0^a \gamma(s)\,\mathrm{d}s\right\} = \begin{cases} 1, & \text{if } a \leqslant A, \\ 0, & \text{if } a > A, \end{cases} \tag{1}$$

meaning that all newborns are protected by maternal antibodies until a certain age A and then move to the susceptible class instantaneously. Under these assumptions, the proportion of susceptible individuals is given by

$$x(a) = \exp\left\{-\int_A^a \lambda(s)\,\mathrm{d}s\right\}, \qquad \text{if } a > A, \tag{2}$$

where $\lambda(a)$ denotes the age-specific force of infection, and $x(a) = 0$ if $a \leqslant A$.

If the mean duration of infectiousness D is short compared with the timescale on which the transmission and mortality rate vary, the force of infection can be approximated by (Anderson and May, 1991)

$$\lambda(a) = \frac{ND}{L}\int_A^\infty \beta(a, a')\,\lambda(a')\,x(a')\,\mathrm{d}a', \tag{3}$$

where $\beta(a, a')$ denotes the transmission rate, i.e. the *per capita* rate at which an individual of age a' makes an effective contact with a person of age a, per year. Formula (3) reflects the so-called 'mass action principle', which implicitly assumes that infectious and susceptible individuals mix completely with each other and move randomly within the population.

Estimating transmission rates by using seroprevalence data cannot be done analytically since the integral equation (3) in general has no closed form solution. However, it is possible to solve this numerically by turning to a discrete age framework, assuming a constant force of infection in each age class. Denote the first age interval $(a_{[1]}, a_{[2]})$ and the jth age interval $[a_{[j]}, a_{[j+1]})$, $j = 2, \ldots, J$, where $a_{[1]} = A$ and $a_{[J+1]} = L$. Making use of formula (2), the prevalence of immune individuals of age a is now well approximated by (Anderson and May, 1991)

$$\pi(a) = 1 - \exp\left\{-\sum_{k=1}^{j-1}\lambda_k(a_{[k+1]} - a_{[k]}) - \lambda_j(a - a_{[j]})\right\}, \tag{4}$$

if a belongs to the jth age interval. Note that we allow the prevalence of immune individuals to vary continuously with age and that we do not summarize the binary seroprevalence outcomes into a proportion per age class. Further, the force of infection for age class i equals ($i = 1, \ldots, J$)

$$\lambda_i = \frac{ND}{L}\sum_{j=1}^{J}\beta_{ij}\left[\exp\left\{-\sum_{k=1}^{j-1}\lambda_k(a_{[k+1]} - a_{[k]})\right\} - \exp\left\{-\sum_{k=1}^{j}\lambda_k(a_{[k+1]} - a_{[k]})\right\}\right], \tag{5}$$

where β_{ij} denotes the *per capita* rate at which an individual of age class j makes an effective contact with a person of age class i, per year. The transmission rates β_{ij} make up a $J \times J$ matrix: the so-called WAIFW matrix.

Once the WAIFW matrix has been estimated, following Diekmann *et al.* (1990) and Farrington *et al.* (2001), the basic reproduction number R_0 can be calculated as the dominant eigenvalue of the $J \times J$ next generation matrix with elements $(i, j = 1, \ldots, J)$:

$$\frac{ND}{L}(a_{[i+1]} - a_{[i]})\beta_{ij}. \tag{6}$$

R_0 represents the number of secondary cases that are produced by a typical infected person during his or her entire period of infectiousness, when introduced into an entirely susceptible population with the exception of newborns who are passively immune through maternal antibodies. In the next section, we illustrate the traditional approach of imposing mixing patterns to estimate the WAIFW matrix from seroprevalence data.

3.2. Imposing mixing patterns

The traditional approach of Anderson and May (1991) imposes different, somewhat *ad hoc*, mixing patterns on the WAIFW matrix. Note that, in the previous section, we ended up with a system of J equations with $J \times J$ unknown parameters (5) and thus restrictions on these patterns are necessary. Among the proposals in the literature, one distinguishes between several mixing assumptions such as homogeneous mixing ($\beta(a, a') = \beta$), proportional mixing ($\exists u : \beta(a, a') = u(a)\, u(a')$), separable mixing ($\exists u, v : \beta(a, a') = u(a)\, v(a')$) and symmetry ($\beta(a, a') = \beta(a', a)$). Note that the last two mixing assumptions require additional restrictions to be made. As illustrated by Greenhalgh and Dietz (1994) and Van Effelterre *et al.* (2009), the structure that is imposed on the WAIFW matrix has a large effect on the estimate of R_0. In this section, we assume that the transmission rate is constant within six discrete age classes ($J = 6$). We follow Anderson and May (1991), Van Effelterre *et al.* (2009) and Ogunjimi *et al.* (2009) and consider the following mixing patterns, based on prior knowledge of social mixing behaviour, to model the WAIFW matrix for VZV:

$$\left.\begin{aligned}
W_1 &= \begin{pmatrix} \beta_1 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} \beta_1 & \beta_1 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_3 & \beta_3 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 & \beta_6 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix} \\
W_3 &= \begin{pmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_3 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 & \beta_6 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix}, \qquad W_4 = \begin{pmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_1 & \beta_1 & \beta_1 \\ \beta_2 & \beta_2 & \beta_2 & \beta_2 & \beta_2 & \beta_2 \\ \beta_3 & \beta_3 & \beta_3 & \beta_3 & \beta_3 & \beta_3 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_4 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix} \\
W_5 &= \begin{pmatrix} \beta_1 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 \end{pmatrix}, \qquad W_6 = \begin{pmatrix} \beta_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_6 \end{pmatrix}.
\end{aligned}\right\} \tag{7}$$

To estimate the transmission parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^{\mathrm{T}}$ from seroprevalence data, we follow an iterative procedure from Farrington *et al.* (2001) and Kanaan and Farrington (2005). First, we assume plausible starting values for $\boldsymbol{\beta}$ and solve equation (5) iteratively for the piecewise constant force of infection $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_6)^{\mathrm{T}}$, which in its turn can be contrasted with the serology. Second, this procedure is repeated under the constraint $\boldsymbol{\beta} \geqslant \mathbf{0}$, until the Bernoulli log-likelihood

$$\sum_{i=1}^{n} [y_i \log\{\pi(a_i)\} + (1 - y_i) \log\{1 - \pi(a_i)\}]$$

has been maximized. Here, n denotes the size of the serological data set, y_i denotes a binary variable indicating whether subject i had experienced infection before age a_i and the prevalence $\pi(a_i)$ is obtained from equation (4).

3.3. Application to the data

For the remainder of the paper, the following parameters, which are specific for Belgium in 2003 (Eurostat, 2007; Federale Overheidsdienst Economie Afdeling Statistiek, 2006), are kept constant when estimating the WAIFW matrix and R_0: size of the population aged 0–80 years, $N = 9\,943\,749$; life expectancy at birth, $L = 80$. The mean duration of infectiousness for VZV is taken as $D = 7/365$. Type I mortality and type I maternal antibodies with age $A = 0.5$ are assumed. Removing individuals who were younger than 6 months, the size of the serological data set becomes $n = 2649$.

In this application, the population is divided into six age classes taking into account the schooling system in Belgium, following Van Effelterre *et al.* (2009): (0.5,2), [2,6), [6,12), [12,19),

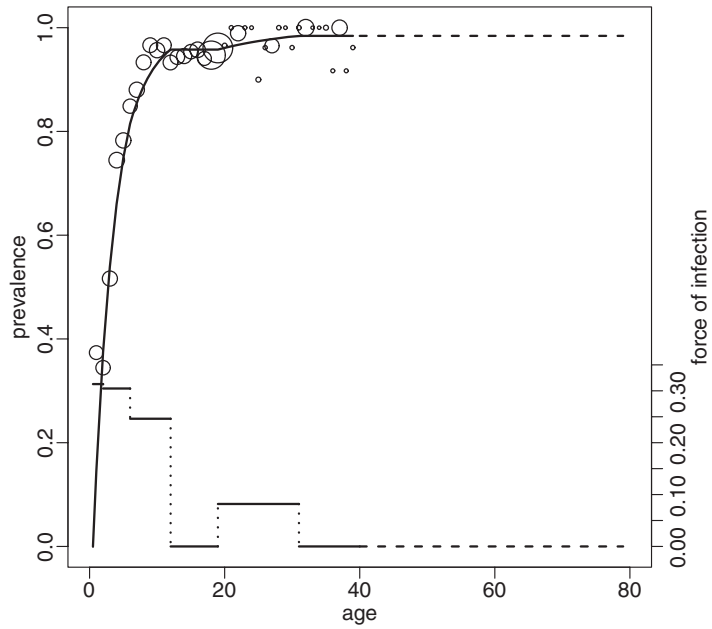

Fig. 1. Estimated prevalence (upper curve) and force of infection (lower curve) for VZV, assuming a piecewise constant force of infection; **o**, observed serological data with size proportional to the corresponding sample size; -------, estimated prevalence and force of infection for the age interval [40,80) years, which lacks serological information

Table 1. Estimates for the transmission parameters (multiplied by 10^4) and for R_0, obtained by imposing mixing patterns W_2, W_3 and W_4 on the WAIFW matrix

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{R}_0$	*95% confidence interval for* R_0	*AIC*
W_2	1.413	1.335	1.064	0.000	0.343	0.000	3.51	[3.07, 13.42]	1372.819
W_3	1.362	1.441	0.873	0.000	0.343	0.000	3.37	[2.81, 13.38]	1372.819
W_4	1.334	1.298	1.049	0.000	0.349	0.000	4.21	[3.69, 13.13]	1372.756

[19,31) and [31, 80) years. The last age class has a wide range because the serological data set only contains information for individuals up till 40 years. The following ML estimate for $\boldsymbol{\lambda}$ is obtained by assuming a piecewise constant force of infection and using constrained optimization to ensure monotonicity ($\pi'(a) \geqslant 0$): $\hat{\boldsymbol{\lambda}}^{\mathrm{ML}} = (0.313, 0.304, 0.246, 0, 0.082, 0)^{\mathrm{T}}$. A graphical display of the fit is presented in Fig. 1 and a broken line is used to indicate the estimated prevalence and force of infection for the age interval [40, 80) years, which lacks serological information.

During the estimation process, non-identifiability problems occur for mixing patterns W_1, W_5 and W_6, which are related to the fact that $\hat{\lambda}_4^{\mathrm{ML}} = \hat{\lambda}_6^{\mathrm{ML}} = 0$. Therefore, these mixing patterns are left from further consideration. For the remaining three, ML estimates for $\boldsymbol{\beta}$ and R_0 are presented in Table 1. Note that mixing pattern W_4 has a regular configuration for the data, whereas W_2 and W_3 are non-regular since unconstrained ML estimation induces negative estimates for β_4 (Farrington *et al.*, 2001). The estimate of R_0 ranges from 3.37 to 4.21. A 95% bootstrap-based percentile confidence interval for R_0 is presented as well, applying a non-parametric bootstrap by taking $B = 1000$ samples with replacement from the serological data. The fit of the three mixing patterns can be compared by using model selection criteria, such as Akaike's information criterion AIC and Bayes information criterion BIC (Schwarz, 1978). As can be seen from Table 1, the AIC-values (equivalent to BIC here) are virtually equal and do not provide any basis to guide the choice of a mixing pattern.

These results differ somewhat from those obtained by Van Effelterre *et al.* (2009), where a different data set for VZV serology was used, which was collected from a large laboratory in the city of Antwerp between October 1999 and April 2000.

4. Estimation of R_0 by using data on social contacts

4.1. Constant proportionality of the transmission rates

In the previous section, we have illustrated some *caveats* that are involved in the traditional approach of imposing mixing patterns on the WAIFW matrix. In general, the choice of the structures as well as the choice of the age classes are somewhat *ad hoc*. Since evidence for mixing patterns is thought to be found in social contact data, i.e. governing contacts with high transmission potential, an alternative approach to estimate transmission parameters has emerged: augmenting seroprevalence data with data on social contacts. In Wallinga *et al.* (2006), it was argued that $\beta(a, a')$ is proportional to $c(a, a')$, the *per capita* rate at which an individual of age a' makes contact with a person of age a, per year:

$$\beta(a, a') = q\, c(a, a'). \tag{8}$$

We shall refer to this assumption as the 'constant proportionality' assumption, since q represents a constant disease-specific factor. Translating this assumption into the discrete framework with

age classes $(a_{[1]}, a_{[2]}), [a_{[2]}, a_{[3]}), \ldots, [a_{[J]}, a_{[J+1]})$ is straightforward $(i, j = 1, \ldots, J) : \beta_{ij} = q\, c_{ij}$, where c_{ij} denotes the *per capita* rate at which an individual of age class j makes contact with a person of age class i, per year.

The proportionality factor and the contact rates are not identifiable from serological data only. Therefore, to estimate the WAIFW matrix, we first need to estimate the contact rates c_{ij} by using social contact data. Following the Belgian contact survey, 'making contact with' is then defined as a two-way conversation of at least three words in each other's proximity and/or any sort of physical skin-to-skin touching (Section 2.1). In Section 4.3.1, we shall refine this definition and consider specific types of contact with high transmission potential. In the second step, keeping the estimated contact rates fixed, we estimate the proportionality factor from serological data by using the estimation method that was described in Section 3.2.

4.2. Estimating contact and transmission rates

Consider the random variable Y_{ij}, i.e. the number of contacts in age class j during 1 day as reported by a respondent in age class i $(i, j = 1, \ldots, J)$, which has observed values $y_{ij,t}$, $t = 1, \ldots, T_i$, where T_i denotes the number of participants in the contact survey belonging to age class i. Now define $m_{ij} = E(Y_{ij})$, i.e. the mean number of contacts in age class j during 1 day as reported by a respondent in age class i. The elements m_{ij} make up a $J \times J$ matrix, which is called the 'social contact matrix'. Now, the contact rates c_{ij} are related to the social contact matrix as follows:

$$c_{ij} = 365\, \frac{m_{ji}}{w_i},$$

where w_i denotes the population size in age class i, obtained from demographical data. When estimating the social contact matrix, the reciprocal nature of contacts needs to be taken into account (Wallinga *et al.*, 2006):

$$m_{ij} w_i = m_{ji} w_j, \tag{9}$$

which means that the total number of contacts from age class i to age class j must equal the total number of contacts from age class j to age class i.

4.2.1. Bivariate smoothing

The elements m_{ij} of the social contact matrix are estimated from the contact data by using a bivariate smoothing approach as described by Wood (2006). In contrast with the ML approach as presented by Wallinga *et al.* (2006), the average number of contacts is modelled as a two-dimensional continuous function over age of respondent and contact, giving rise to a 'contact surface'. The basis is a tensor product spline derived from two smooth functions of the respondent's and contact's age, ensuring flexibility:

$$Y_{ij} \sim \text{NegBin}(m_{ij}, k), \qquad g(m_{ij}) = \sum_{l=1}^{K} \sum_{p=1}^{K} \delta_{lp}\, b_l(a_{[i]})\, d_p(a_{[j]}), \tag{10}$$

where g is some link function, δ_{lp} are unknown parameters and b_l and d_p are known basis functions for the marginal smoothers. To allow for overdispersion, we assume that the contact counts Y_{ij} are independently negative binomial distributed with mean m_{ij}, dispersion parameter k and variance $m_{ij} + m_{ij}^2/k$.

The basis dimension K should be chosen sufficiently large to fit the data well, but sufficiently small to maintain reasonable computational efficiency (Wood, 2006). For tensor product smoothers, the upper limit of the degrees of freedom is given by the product of the K values

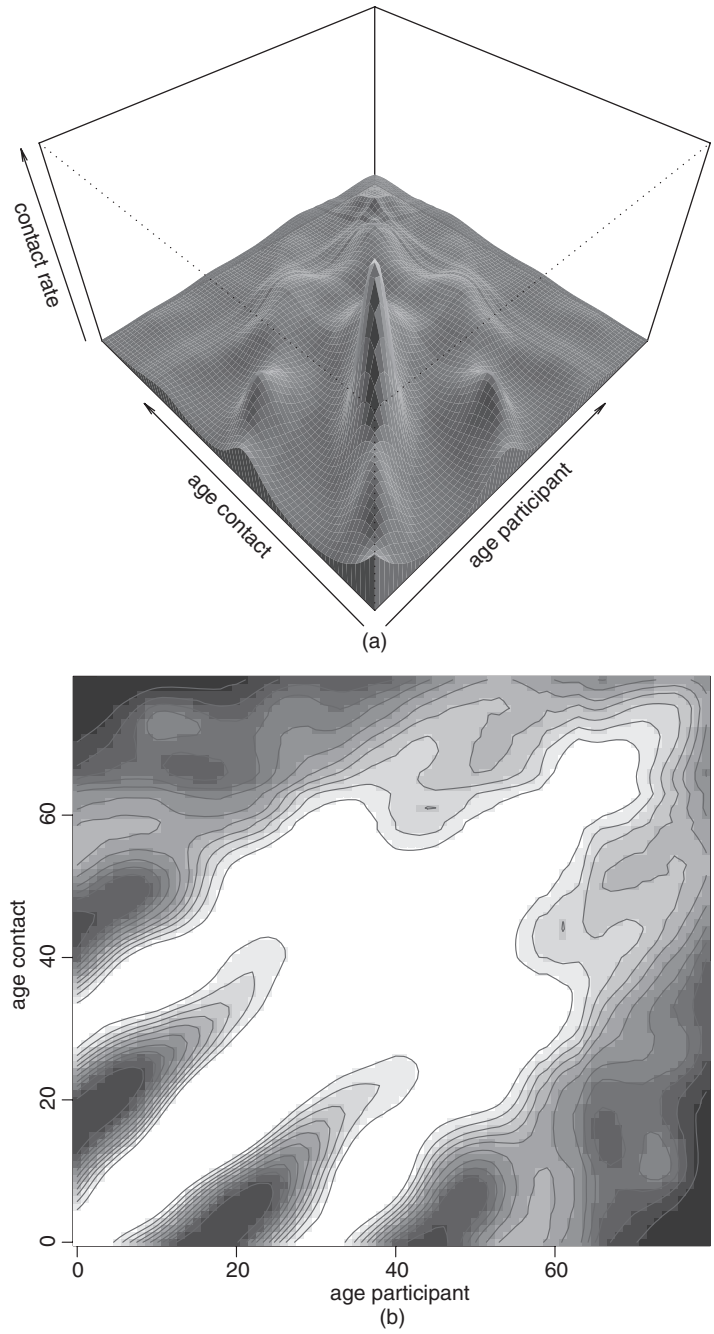

Fig. 2. (a) Perspective and (b) image plot of the estimated contact rates c_{ij} obtained with bivariate smoothing: the x- and y-axes represent the age of the respondent and the age of the contacted person respectively

provided for each marginal smooth, minus 1, for the identifiability constraint. However, the actual effective degrees of freedom are also controlled by the degree of penalization that is selected during fitting.

Thin plate regression splines are used to avoid the selection of knots and a log-link is used in model (10). Diary weights, as discussed in Section 2.1, are taken into account in the smoothing process. By applying a smooth-then-constrain approach as proposed by Mammen *et al.* (2001), the reciprocal nature of contacts (9) is taken into account.

4.2.2. *Estimating the contact rates*

The smoothing is performed in R with the `gam` function from the `mgcv` package (Wood, 2006), considering 1-year age intervals, $[0, 1), [1, 2), \ldots, [100, 101)$ years. An informal check (by comparing the estimated degrees of freedom and the basis dimension) shows that $K = 11$ is a satisfactory choice of basis dimension for the Belgian contact data. In Fig. 2, the estimated contact surface that is obtained with the bivariate smoothing approach is displayed. The smoothing approach seems well able to capture important features of human contact behaviour. Three components clearly arise in the smoothed contact surface. First, we can see a pronounced assortative structure on the diagonal, representing high contact rates between individuals of the same age. Second, an off-diagonal parent–child component comes forward, reflecting a very natural form of contact between parents and children, which might be important in modelling certain childhood infections such as parvovirus B19 (Mossong *et al.*, 2008a). Finally, there even seems to be evidence for a grandparent–grandchild component.

Except for the assortativeness, these features are not reflected by the contact rates, estimated by maximizing the likelihood of the 'saturated model' that was proposed by Wallinga *et al.* (2006), considering the same six age classes as used in Section 3.3 (the results have been omitted here). Furthermore, the AIC- and BIC-criteria indicate that the smoothing method outperforms the saturated model of Wallinga *et al.* (2006), showing improved estimation of the contact surface by using non-parametric techniques.

4.2.3. *Estimating R_0*

Under the constant proportionality assumption (8), we can now estimate the WAIFW matrix for VZV by using serological data. Keeping the estimated contact rates $\hat{c}_{ij}$ fixed, we estimate the proportionality factor q by using the estimation method that was described in Section 3.2. In Table 2, estimates for q and R_0 together with their corresponding 95% profile likelihood confidence intervals, and AIC-values, are presented for the bivariate smoothing approach and the saturated model that was proposed by Wallinga *et al.* (2006). The results are fairly similar,

Table 2. ML estimates for the proportionality factor and R_0, obtained from contact rates estimated by bivariate smoothing and the saturated model of Wallinga *et al.* (2006), assuming constant proportionality

Model for c_{ij}	$\hat{q}$	*95% confidence interval for* q	$\hat{R}_0$	*95% confidence interval for* R_0	*AIC*
Smoothing	0.132	[0.124, 0.140]	15.69	[14.74, 16.69]	1386.618
Saturated	0.124	[0.117, 0.132]	14.08	[13.26, 14.94]	1377.146

though the saturated model induces a smaller AIC-value compared with the smoothing approach. As can be seen from both model fits in Fig. 3, contact rate estimates between children will mainly determine the fit to the serological data, limiting the advantage of a better contact surface estimate. Note that the 95% confidence intervals in Table 2 are implausibly narrow, resulting from the fact that the estimated contact rates are held constant.

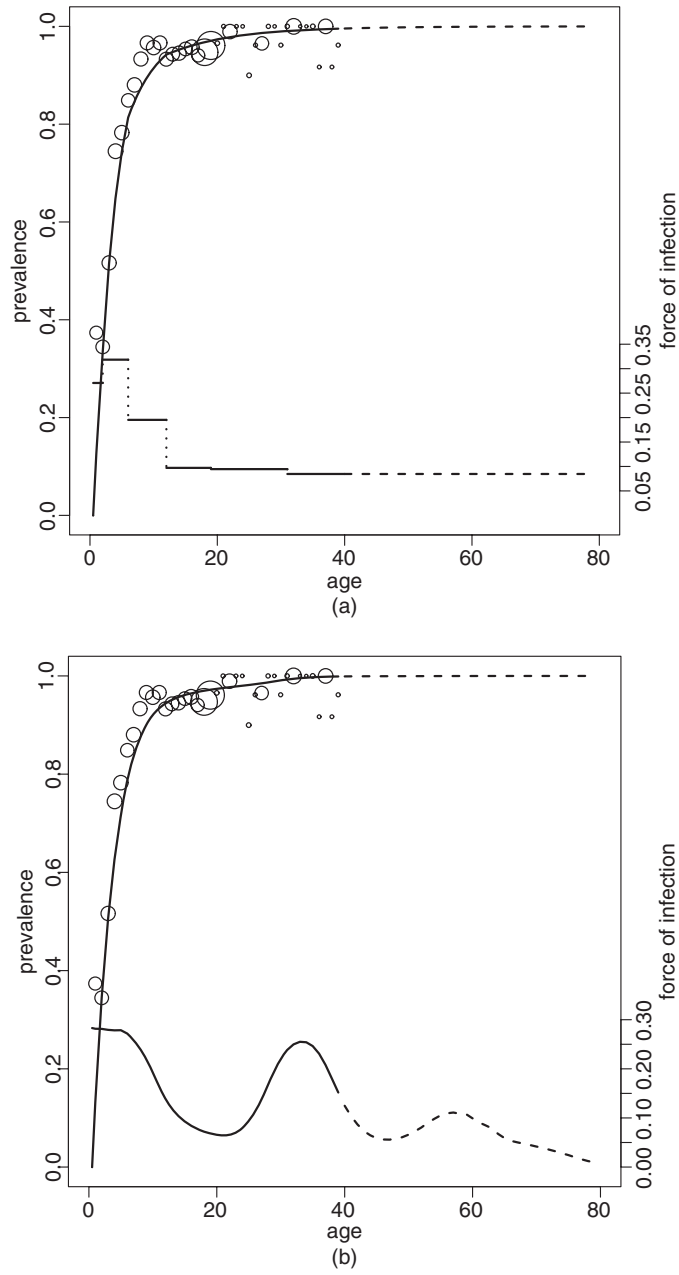

Fig. 3. Estimated prevalence (upper curve) and force of infection (lower curve) obtained from contact rates estimated (a) by using ML for the saturated model of Wallinga *et al.* (2006) and (b) by using bivariate smoothing

4.3. Refinements to the social contact data approach

The aim is to disentangle the WAIFW matrix clearly into the contact process and the transmission potential. Therefore, in what follows, contact rates are estimated by using a bivariate smoothing approach, since this method outperforms the saturated model that was estimated by using ML as proposed by Wallinga *et al.* (2006) (Section 4.2.2). Following Ogunjimi *et al.* (2009) and Melegaro *et al.* (2009), contacts with high transmission potential are filtered from the social contact data. Further, to improve statistical inference, we present a non-parametric bootstrap approach, explicitly accounting for all sources of variability.

4.3.1. Contacts with high transmission potential

The aim is to trace the type of contact which is most likely to be responsible for VZV transmission, thereby exploiting the following details provided on each contact: duration and type of contact, which is either close or non-close (Section 2.1). Five types of contact are considered and we shall explore which one induces the best fit to the serological data. First, the contact rates $c(a, a')$ are estimated by using the complete contact data set as we did in Section 4.2.3 and, further, four specific types of contact with high transmission potential for VZV are selected according to Ogunjimi *et al.* (2009) and Melegaro *et al.* (2009) (Table 3).

Assuming constant proportionality, ML estimates for the transmission parameters q_k ($k = 1, \ldots, 5$) and for the basic reproduction number R_0 together with their corresponding 95% profile likelihood confidence intervals (first entry) are presented in Table 4. For each model C_k, the AIC-value, AIC-difference $\Delta_k = \mathrm{AIC}_k - \mathrm{AIC}_{\min}$, Akaike weight

$$w_k = \exp(-\tfrac{1}{2}\Delta_k) \Big/ \sum_l \exp(-\tfrac{1}{2}\Delta_l)$$

and evidence ratio $w_{\min}/w_k$ are calculated following Burnham and Anderson (2002), where $\mathrm{AIC}_{\min}$ and $w_{\min}$ correspond to the model with the smallest AIC-value. Recall that AIC is an estimate of the expected, relative Kullback–Leibler distance, whereas the Kullback–Leibler distance embodies the information that is lost when an approximating model is used instead of the unknown, true model. A given Akaike weight w_k is considered as the weight of evidence in favour of a model k being the actual Kullback–Leibler best model for the situation at hand, given the data and the set of candidate models considered.

According to the AIC-criterion, although differences in AIC are minor, the contact matrix consisting of close contacts longer than 15 min (model C_3) implies the best fit to the seroprevalence data. A graphical representation of the estimated prevalence and force of infection has been omitted here, since the result is very close to that obtained for model C_1 in Fig. 3. Further, there is evidence for model C_5 as well, having an Akaike weight of 0.329 and an evidence ratio of 1.7. The latter model adds non-close contacts longer than 1 h to model C_3 and therefore these models are closely related.

4.3.2. Non-parametric bootstrap

We explicitly acknowledge that until now, by keeping the estimated contact rates fixed, we have ignored the variability originating from the contact data. To assess sampling variability for the social contact data and the serological data altogether, we shall use a non-parametric bootstrap approach. Furthermore, building in a randomization process, uncertainty concerning age is accounted for. After all, in the social contact data, ages of respondents are rounded up, which is also so for some individuals in the serological data set. Concerning the age of contacts, lower and upper age limits are given by the respondents. Instead of using the mean value of these age

Table 3. Candidate models assuming various types of contact underlying VZV transmission

Model	*Parameter*	*Type of contact*
C_1	q_1	All contacts
C_2	q_2	Close contacts
C_3	q_3	Close contacts >15 min
C_4	q_4	Close contacts and non-close contacts >1 h
C_5	q_5	Close contacts >15 min and non-close contacts >1 h

Table 4. ML estimates for the proportionality factor and R_0, 95% profile likelihood confidence intervals (first entry), 95% bootstrap-based percentile confidence intervals (second entry) and several measures related to model selection, obtained from contact rates estimated by using bivariate smoothing, considering different types of contact C_1–C_5, assuming constant proportionality

Model	$\hat{q}_k$	*95% confidence interval for* q_k	$\hat{R}_0$	*95% confidence interval for* R_0	*AIC*	Δ_k	w_k	*Evidence ratio*
C_1	0.132	[0.124, 0.140]	15.69	[14.74, 16.69]	1386.618	11.660	0.002	340.4
		[0.103, 0.175]		[12.34, 21.41]				
C_2	0.160	[0.150, 0.169]	10.24	[9.65, 10.85]	1379.581	4.623	0.057	10.1
		[0.126, 0.208]		[8.21, 13.68]				
C_3	0.173	[0.163, 0.184]	8.68	[8.18, 9.20]	1374.958	0.000	0.574	1.0
		[0.133, 0.221]		[6.89, 11.34]				
C_4	0.145	[0.136, 0.154]	11.73	[11.05, 12.47]	1380.354	5.396	0.039	14.9
		[0.113, 0.188]		[9.41, 15.95]				
C_5	0.156	[0.147, 0.166]	10.40	[9.79, 11.04]	1376.068	1.110	0.329	1.7
		[0.119, 0.204]		[8.05, 14.10]				

limits, a random draw is now taken from the uniform distribution on the corresponding age interval. In summary, each bootstrap cycle consists of the following six steps:

(a) randomize ages in the social contact data and the serological data set;
(b) take a sample with replacement from the respondents in the social contact data;
(c) recalculate diary weights based on age and size of household of the selected respondents;
(d) estimate the social contact matrix (the smooth-then-constrain approach);
(e) take a sample with replacement from the serological data;
(f) estimate the transmission parameters and R_0.

This bootstrap approach allows us to calculate bootstrap confidence intervals for the transmission parameters and for the basic reproduction number which take into account all sources of variability.

The effect on statistical inference is now illustrated for the models that were considered in the previous section. 900 bootstrap samples are taken from the contact data and from the serological data simultaneously, while ages are being randomized. Merely $B = 587$ bootstrap samples lead to convergence in all five smoothing procedures, which might be induced by the sparse structure of the contact data. However, by individual monitoring of non-converging `gam` functions, convergence was reached after all and a comparison of the bootstrap results showed little difference whether or not these samples were included. 95% percentile confidence intervals

for q and R_0 are calculated on the basis of the $B = 587$ bootstrap samples (see Table 4, second row for each entry). Taking into account sampling variability for the social contact data has a noticeable effect, as can be seen from the wider 95% confidence intervals.

5. Age-dependent proportionality of the transmission rates

The proportionality factor q might depend on several characteristics that are related to susceptibility and infectiousness, which could be ethnic, climate, disease or age specific. Examples of age-specific characteristics that are related to susceptibility and infectiousness include the mean infectious period, secretion of mucus and hygiene. In the situation of seasonal and pandemic influenza this has been established and used in realistic simulation models (see for example Cauchemez *et al.* (2004) and Longini *et al.* (2005)). Furthermore, the conversational and physical contacts that were reported in the diaries serve as proxies of those events by which an infection can be transmitted. For example, sitting close to someone in a bus without actually touching each other may also lead to transmission of infection. In light of these discrepancies, q can be considered as an age-specific adjustment factor which relates the true contact rates underlying infectious disease transmission to the social contact proxies.

In view of this, we shall explore whether q varies with age, an assumption that we shall refer to as 'age-dependent proportionality':

$$\beta(a, a') = q(a, a')\, c(a, a'), \tag{11}$$

which in the discrete framework turns into $\beta_{ij} = q_{ij} c_{ij}$ $(i, j = 1, \ldots, J)$. In the previous section, it was observed that, under the constant proportionality assumption, close contacts longer than 15 min imply the best fit to the serological data for VZV. Therefore, in what follows, the contact rate is modelled by using close contacts longer than 15 min and we shall elaborate on this particular model by assuming age dependence. First, discrete structures are applied to model q as an age-dependent proportionality factor and, second, 'continuous' log-linear regression models are considered for the same purpose. Finally, we assess the level of model selection uncertainty and calculate a model-averaged estimate for the basic reproduction number.

5.1. Discrete structures

The proportionality factor q_{ij} is now allowed to differ between age classes. Discrete matrix structures, involving two transmission parameters γ_1 and γ_2, are explored in modelling q_{ij}. Five models are considered, which fit the following structures for q_{ij} to the seroprevalence data:

$$M_1 = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & \gamma_2 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} \gamma_1 & \gamma_1 \\ \gamma_2 & \gamma_2 \end{pmatrix}, \qquad M_3 = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & \gamma_1 \end{pmatrix},$$
$$M_4 = \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}, \qquad M_5 = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_2 \end{pmatrix}.$$

The population is divided into two age classes, namely $[0.5, 12)$ and $[12, 80)$ years, which is a choice based on the dichotomy of the population according to the schooling system in Belgium (Section 3.3), yielding the smallest AIC-value. Higher order extensions, considering more parameters and/or number of age classes, were fitted to the serological data as well. The improvement in log-likelihood, however, does not outweigh the increase in the number of transmission parameters.

Note that the structures of M_1–M_5 resemble the mixing patterns that were imposed on the WAIFW matrix in the traditional Anderson and May (1991) approach. We emphasize that

Table 5. Candidate models for the proportionality factor together with ML estimates for the transmission parameters and R_0, 95% bootstrap-based percentile confidence intervals and several measures related to model selection

Model	*Parameter*		*95% confidence interval*	$\hat{R}_0$	*95% confidence interval for* R_0	K	*AIC*	Δ_k	w_k	*Evidence ratio*
C_3	$\hat{q}$	0.173	[0.133, 0.221]	8.68	[6.89, 11.34]	1	1374.958	8.884	0.003	84.9
M_1	$\hat{\gamma}_1$	0.185	[0.136, 0.244]	4.79	[4.15, 9.98]	2	1366.306	0.232	0.261	1.1
	$\hat{\gamma}_2$	0.079	[0.006, 0.196]							
M_2	$\hat{\gamma}_1$	0.183	[0.138, 0.240]	5.37	[4.47, 9.68]	2	1366.285	0.211	0.264	1.1
	$\hat{\gamma}_2$	0.078	[0.006, 0.187]							
M_3	$\hat{\gamma}_1$	0.185	[0.136, 0.244]	8.26	[6.82, 11.25]	2	1366.074	0.000	0.293	1.0
	$\hat{\gamma}_2$	0.069	[0.006, 0.199]							
M_6	$\hat{\gamma}_0$	−1.622	[−2.028, −1.212]	5.79	[4.63, 12.60]	2	1368.709	2.635	0.079	3.7
	$\hat{\gamma}_1$	−0.023	[−0.067, 0.016]							
M_7	$\hat{\gamma}_0$	−1.720	[−2.441, −1.182]	5.03	[4.20, 1318.68]	3	1368.325	2.251	0.095	3.1
	$\hat{\gamma}_1$	0.014	[−0.086, 0.305]							
	$\hat{\gamma}_2$	−0.002	[−0.024, 0.001]							
M_8	$\hat{\gamma}_0$	−1.517	[−2.224, −0.446]	3.55	[1.76, 159.96]	2	1374.324	8.250	0.005	61.9
	$\hat{\gamma}_1$	−0.065	[−0.403, 0.064]							

the method that is proposed here differs greatly from the latter, since the WAIFW matrix is now estimated by using the estimated contact rates: $\beta_{ij} = q_{ij}\hat{c}_{ij}$. Hence, in contrast with the approach of Anderson and May (1991) who estimated β_{ij} by fixing the structure of the mixing pattern, in our approach we estimate the contact pattern from the survey data and use several proportionality structures to select the best model from which the β_{ij} are estimated.

Table 5 displays ML estimates for γ_1, γ_2 and the basic reproduction number R_0, together with their corresponding 95% percentile confidence intervals ($B = 603$ bootstrap samples converged out of 700). For model M_4, γ_2 is non-identifiable, and unconstrained optimization of model M_5 would not lead to convergence. According to the AIC-criterion, the remaining models fit equally well and are informative with respect to VZV transmission dynamics. Most likely, this is because the main transmission routes for VZV are between children and from infectious children to susceptible adults, embodied by the first column $(\gamma_1, \gamma_2)^{\mathrm{T}}$. The three models result in approximately the same estimates for γ_1 and γ_2 and consequently the differences in AIC are only minor.

It is clear from Table 5 that we estimate a difference in transmissibility between those younger and older than 12 years (about 0.18 and 0.07 respectively). This difference cannot be solely explained by the estimated contact rates. A possible explanation is that, when infectious children make close contact with susceptible children during a sufficient amount of time, the probability of effective VZV transmission is higher compared with the same situation with susceptible adults. Another potential cause is underreporting of contacts between children. After all, up to the age of 8 years the contact diaries were filled in by the parents, which may have induced some reporting bias (Hens *et al.*, 2009a).

5.2. Continuous modelling

As opposed to previously, the proportionality factor $q(a, a')$ is now allowed to vary continuously over age. Log-linear regression models are considered for $q(a, a')$, since we expect an exponential decline of q over a due to hygiene habits as well as an exponential decline of q over a' due to decreasing secretion of mucus. The following log-linear models are fitted to the data:

$$
\begin{aligned}
M_6, &\quad \log\{q(a)\} = \gamma_0 + \gamma_1 a;\\
M_7, &\quad \log\{q(a)\} = \gamma_0 + \gamma_1 a + \gamma_2 a^2;\\
M_8, &\quad \log\{q(a')\} = \gamma_0 + \gamma_1 a';\\
M_9, &\quad \log\{q(a')\} = \gamma_0 + \gamma_1 a' + \gamma_2 (a')^2;\\
M_{10}, &\quad \log\{q(a,a')\} = \gamma_0 + \gamma_1 a + \gamma_2 a'.
\end{aligned}
$$

Model M_6 models q as a first-degree function of age of the susceptible individual and model M_7 allows for an additional quadratic effect of age, a^2. Models M_8 and M_9 are the analogue of M_6 and M_7 for age of the infectious person, a'. Finally, M_{10} models q as an exponential function of a and a' simultaneously. For model M_9, no convergence was obtained and model M_{10} gives rise to an estimated proportionality factor which is exponentially increasing over a', inducing unrealistically large estimates for q at older ages.

ML estimates for the model parameters and the basic reproduction number R_0 are presented in Table 5, together with the corresponding 95% percentile confidence intervals ($B = 603$ bootstrap samples converged out of 700). According to the AIC-criterion, M_6 and M_7 fit equally well. Allowing the proportionality factor to vary by age of infectious individuals does not seem to improve the model fit substantially, as can be seen by comparing the AIC-values of C_3 and M_8.

Clearly, for models M_7 and M_8, the upper limits of the confidence intervals for R_0 are very large, as a consequence of estimated proportionality factors which are exponentially increasing over a and a' respectively. This result originates from two things: first, there is a lack of serological information for individuals aged 40 years and older and, second, VZV is highly prevalent in the population and most individuals become infected with VZV before the age of 10 years. Mathematically the latter means that, from a certain age on, $\pi(a) \approx 1$ and $\pi'(a) \approx 0$, leading to an indeterminate force of infection $\lambda(a) = \pi'(a)/\{1 - \pi(a)\}$. In Section 5.4, we assess the sensitivity of the results to the former issue, repeating all analyses by using simulated serological data for the age range $[40, 80)$ years.

Fig. 4 displays the estimated prevalence function and force of infection for the discrete model M_3 (Fig. 4(a)) and the continuous model M_7 (Fig. 4(b)). The results are remarkably similar. The effect of making q age dependent is visualized by comparing Fig. 4 with the fit of model C_1, which was very close to model C_3, in Fig. 3(b). The models assuming age-dependent proportionality estimate an initially higher force of infection and a steeper decrease from the age of 10 years, after which the force of infection is reduced by a factor of 2, compared with the constant proportionality model. Whereas the latter model predicts total immunity for VZV at older ages, the age-dependent proportionality models estimate a fraction of seropositive individuals which is below 1 at all times.

5.3. Model selection and multimodel inference

Table 5 presents all candidate models for the proportionality factor q that we have collected until now, among which are the constant proportionality model C_3, the discrete age-dependent proportionality models M_1, M_2 and M_3, and the continuous age-dependent proportionality models M_6, M_7 and M_8. Further for each model, the number of parameters K, the AIC-value, the AIC-difference Δ_k, the Akaike weight w_k and the evidence ratio are displayed.

Model M_3 with an assortative component γ_1 and a background component γ_2 is the 'best' model for q according to the AIC-criterion. However, model selection uncertainty is likely to be high since the selected best model has an Akaike weight of only 0.293 (Burnham and Anderson, 2002). The evidence ratios for M_3 *versus* M_1 and M_2 are both 1.1, which means that there is weak support for the best model. If many independent samples could be drawn, the three

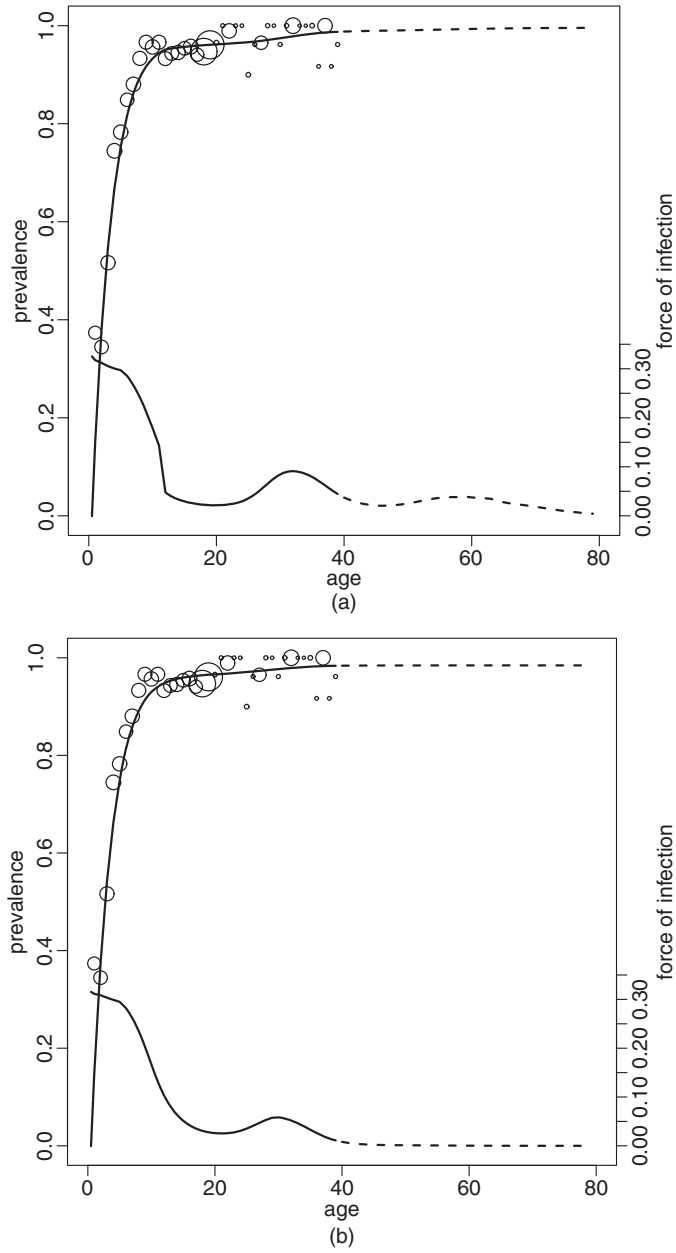

Fig. 4. Estimated prevalence (upper curve) and force of infection (lower curve) for (a) the discrete model M_3 and (b) the continuous model M_7

discrete age-dependent models would probably compete with each other for the best model position. The continuous models M_6 and M_7 have evidence ratios around 3.5, indicating that these models also contribute some information. Models C_3 and M_8 have the largest AIC-difference Δ_k, a very small Akaike weight (0.005 or less) and very large evidence ratios (84.9 and 61.9 respectively), which means that there is little support for these two models.

Since there is no single model in the candidate set that is clearly superior to the others and since the estimate for the basic reproduction number R_0 varies noticeably over the candidate models, we are not inclined to base prediction only on M_3. Applying the concepts of model averaging, as described in Burnham and Anderson (2002), a weighted estimate of R_0 is calculated, based on the model estimates and the corresponding Akaike weights:

$$\hat{\bar{R}}_0 = \sum_{k=1}^{7} w_k (\hat{R}_0)_k = 6.07.$$

With the bootstrap procedure, we obtain a 95% percentile confidence interval for this model-averaged estimate $\hat{\bar{R}}_0$, namely [4.4, 351.6]. Again, there is a large upper limit induced by the same issues as reported in Section 5.2.

5.4. Sensitivity analysis

To assess the lack-of-data problem, we simulate serological data for the age range [40, 80) years by using a constant prevalence $\pi = 0.983$, which is estimated from a thin plate regression spline model for the original serological data. Sample sizes for 1-year age groups are chosen according to the Belgian population distribution in 2003 (Federale Overheidsdienst Economie Afdeling Statistiek, 2006) and the total size of serological data now amounts to $n = 3856$. The seven candidate models for the proportionality factor q are now applied to the 'complete' serological data set.

The results are presented in Table 6 and are, overall, quite similar to the results that were obtained before (Table 5). The 95% percentile confidence intervals for R_0 ($B = 599$ bootstrap samples converged out of 700), however, are narrower since the simulated data for the age range [40, 80) years are 'forcing' the proportionality factor q to follow a natural pace. This is illustrated for model M_7 in Fig. 5, where the estimated function $q(a)$ is depicted for 100 randomly chosen bootstrap samples. In particular, right confidence interval limits for R_0 are smaller, whereas for most models the R_0-estimate seems to have decreased just a little.

Model selection uncertainty is illustrated quite nicely here, since four models (M_7, M_3, M_2 and M_1) have Akaike weights that are close to 0.24 and these models also had the most support

Table 6. Candidate models for the proportionality factor applied to the serological data set augmented with simulated data, together with ML estimates for the transmission parameters and R_0, 95% bootstrap-based percentile confidence intervals and several measures related to model selection

Model	*Parameter*		*95% confidence interval*	$\hat{R}_0$	*95% confidence interval for* R_0	K	*AIC*	Δ_k	w_k	*Evidence ratio*
C_3	$\hat{q}$	0.159	[0.126, 0.195]	7.98	[6.60, 10.19]	1	1618.747	70.774	$\ll 0.0001$	$\gg 10^3$
M_1	$\hat{\gamma}_1$	0.189	[0.137, 0.250]	4.20	[3.88, 5.74]	2	1548.714	0.741	0.201	1.4
	$\hat{\gamma}_2$	0.052	[0.021, 0.095]							
M_2	$\hat{\gamma}_1$	0.186	[0.136, 0.247]	4.74	[4.36, 6.07]	2	1548.627	0.654	0.210	1.4
	$\hat{\gamma}_2$	0.052	[0.020, 0.091]							
M_3	$\hat{\gamma}_1$	0.189	[0.137, 0.250]	8.28	[6.43, 11.52]	2	1548.344	0.371	0.242	1.2
	$\hat{\gamma}_2$	0.044	[0.016, 0.082]							
M_6	$\hat{\gamma}_0$	−1.561	[−1.934, −1.120]	4.96	[4.47, 6.54]	2	1551.321	3.348	0.055	5.3
	$\hat{\gamma}_1$	−0.035	[−0.067, −0.014]							
M_7	$\hat{\gamma}_0$	−1.793	[−2.247, −1.079]	5.22	[4.60, 7.51]	3	1547.973	0	0.292	1.0
	$\hat{\gamma}_1$	0.030	[−0.074, 0.126]							
	$\hat{\gamma}_2$	−0.002	[−0.006, 0.001]							
M_8	$\hat{\gamma}_0$	−1.458	[−2.061, −0.844]	2.69	[2.08, 12.97]	2	1610.113	62.140	$\ll 0.0001$	$\gg 10^3$
	$\hat{\gamma}_1$	−0.103	[−0.254, 0.016]							

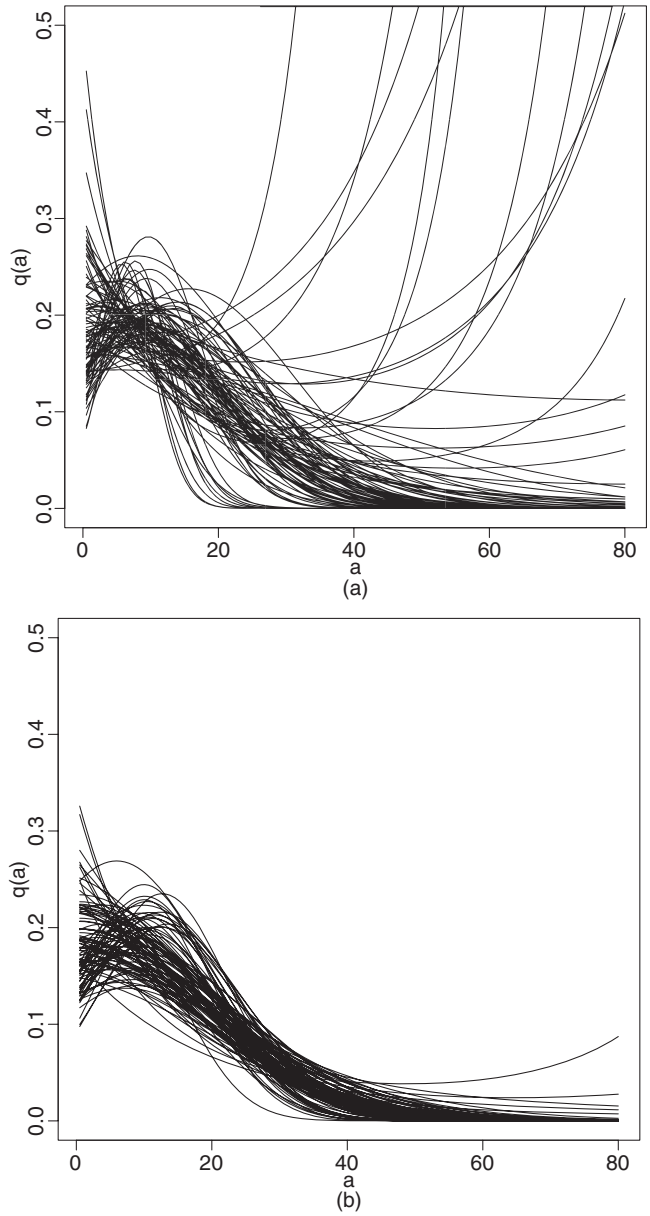

Fig. 5. $q(a)$ estimates for model M_7, shown for 100 randomly chosen bootstrap samples from (a) the original serological data and (b) these data augmented with simulated data for ages [40, 80) years

for the original data set (Table 5). The model-averaged estimate $\hat{\bar{R}}_0$ now equals 5.64 and the 95% bootstrap-based percentile confidence interval is [4.7, 7.5].

6. Concluding remarks

In this paper, an overview of various estimation methods for infectious disease parameters from data on social contacts and serological status was given. The theoretical framework included a

compartmental maternally derived immunity–susceptible–infectious–recovered model, taking into account the presence of maternal antibodies, and the mass action principle, as presented by Anderson and May (1991). An important assumption made was that of endemic equilibrium, which means that infection dynamics are in a steady state. The serological data set that we used was collected over 17 months, averaging over potential epidemic cycles of VZV in Belgium during that period. In Section 3, we have illustrated the traditional basic approach of imposing mixing patterns on the WAIFW matrix to estimate transmission parameters from serological data. In contrast, the novel approach of using social contact data to estimate infectious disease parameters avoids the choice of a parametric model for the entire WAIFW matrix.

The idea is fairly simple: transmission rates for infections that are transmitted from person to person in a non-sexual way, such as VZV, are assumed to be proportional to rates of making conversational and/or physical contact, which can be estimated from contact surveys. Although more time consuming, the bivariate smoothing approach as proposed in Section 4 was better able to capture important features of human contact behaviour, compared with the ML estimation method of Wallinga *et al.* (2006). However, when a non-parametric bootstrap approach was applied to take into account sampling variability, problems of convergence arose, probably due to the large number of 0s in combination with the log-link. Therefore, a mixture of Poisson distributions or a zero-inflated negative binomial distribution could be more appropriate. Further, in Section 4, we dealt with a couple of challenges that were posed by Halloran (2006). The social contact survey contained useful additional information on the contact itself, which allowed us to target very specific types of contact with high transmission potential for VZV. Furthermore, a non-parametric bootstrap approach was proposed to improve statistical inference.

The constant proportionality assumption was relaxed in Section 5 and we have shown that an improvement of fit could be obtained by disentangling the transmission rates into a product of two age-specific variables: the age-specific contact rate and an age-specific proportionality factor. The latter may reflect, for instance, differences in characteristics related to susceptibility and infectiousness or discrepancies between the social contact proxies that were measured in the contact survey and the true contact rates underlying infectious disease transmission. We emphasize that there are probably other models for $q(a, a')$ than those which were considered in Section 5, which fit the data even better. Our choice of a set of plausible candidate models was directed by parsimony on the one hand, limiting the total number of parameters to 3, and prior knowledge on the other hand, considering log-linear models. Furthermore, we restricted analyses to close contacts lasting longer than 15 min, which means that close contacts of short duration and non-close contacts are assumed not to contribute to transmission of VZV.

It is important to note that different assumptions concerning the underlying type of contact as well as different parametric models for $q(a, a')$ are likely to entail different estimates of R_0; however, they may still induce a similar fit to the serological data. To deal with this problem of model selection uncertainty we turned to multimodel inference in Section 5.3. In Fig. 6, estimates of R_0 are presented for the main estimation methods that were considered in this paper: the traditional method of imposing mixing patterns on the WAIFW matrix (W_4) and the method of using data on social contacts, assuming constant proportionality (the saturated model SA, C_1 and C_3) and age-dependent proportionality (M_1, M_2 and M_3). There is a pronounced variability in the estimates of R_0, which is partially captured by the model-averaged estimate MA, calculated from Table 5.

When estimating $q(a, a')$, we were faced with three problems of indeterminacy. First, there is lack of serological information for individuals aged 40 years and older, second, prevalence of VZV rapidly stagnates, leading to an indeterminate force of infection and, third, serological surveys do not provide information related to infectiousness. Models which only expressed age

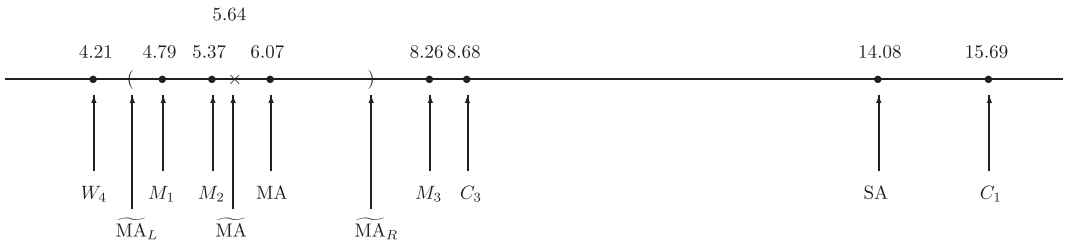

Fig. 6. R_0-estimates for mixing pattern W_4, applied to the serological data in Section 3.3, and for the following models using social contact data: the saturated model SA as proposed by Wallinga *et al.* (2006), applied in Section 4.2.3 assuming constant proportionality, and further bivariate smoothing models, constant proportionality models C_1 and C_3 considering all and close contacts longer than 15 min respectively (Section 4.3.1) and discrete age-dependent proportionality models M_1, M_2 and M_3 (Section 5.1) (the model-averaged estimates for R_0 calculated from Table 5 (MA), based on the original serological data, and from Table 6 ($\widetilde{\mathrm{MA}}$), based on the serological data set augmented with simulated data, are displayed, as well as 95% bootstrap-based percentile confidence interval limits for the latter: $[\mathrm{MA_L}, \mathrm{MA_R}]$)

differences in q for infectious individuals, such as the discrete model M_5 (Section 5.1) and the continuous models M_8 and M_9 (Section 5.2), either did not lead to convergence or induced unrealistically large bootstrap estimates for q at older ages.

A sensitivity analysis in Section 5.4 showed that a lack of serological data had a big effect on confidence intervals for R_0. We simulated data for the age range $[40, 80)$ years, giving rise to a model-averaged estimate $\widetilde{\mathrm{MA}}$ as displayed in Fig. 6 with corresponding confidence interval limits $[\widetilde{\mathrm{MA}}_\mathrm{L}, \widetilde{\mathrm{MA}}_\mathrm{R}]$. The latter problems of indeterminacy might be controlled by combining information on the same infection over different countries or on different airborne infections, assuming that there is a relationship between the country- or disease-specific $q(a, a')$ respectively. This strategy already appeared beneficial when estimating R_0 directly from seroprevalence data, without using social contact data (Farrington *et al.*, 2001).

Further, the effect of intervention strategies, such as school closures, might be investigated by incorporating transmission parameters, estimated from data on social contacts and serological status, in an age–time dynamical setting. Finally, it is important to note that the models rely on the assumptions of type I mortality and type I maternal antibodies to facilitate calculations. Consequently, model improvements could be made through a more realistic approach of demographical dynamics.

Acknowledgements

We thank the Joint Editor and the reviewers for their comprehensive comments that led to an improved version of the manuscript. This study has been made and funded as part of 'Simulation models of infectious disease transmission and control processes', a strategic basic research project that is funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders, project 060081. This work has been partly funded and benefited from discussions held in POLYMOD, a European Commission project funded within the sixth framework programme, contract SSP22-CT-2004-502084. We also gratefully acknowledge support from the Interuniversity Attraction Pole research network P6/03 of the Belgian Government (Belgian Science Policy).

References

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Beutels, P., Shkedy, Z., Aerts, M. and Van Damme, P. (2006) Social mixing patterns for transmission models

of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidem. Infectn*, **134**, 1158–1166.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multi-model Inference: a Practical Information-theoretic Approach*. New York: Springer.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. and Boëlle, P. Y. (2004) A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statist. Med.*, **23**, 3469–3487.

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.

Edmunds, W. J., Kafatos, G., Wallinga, J. and Mossong, J. R. (2006) Mixing patterns and the spread of close-contact infectious diseases. *Emergng Themes Epidem.*, **3**, no. 10.

Edmunds, W. J., O'Callaghan, C. J. and Nokes, D. J. (1997) Who mixes with whom?: a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proc. R. Soc.* B, **264**, 949–957.

Eurostat (2007) Population table for Belgium, 2003. Eurostat, Luxembourg. (Available from `http://epp.eurostat.ec.europa.eu/`.)

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). *Appl. Statist.*, **50**, 251–292.

Farrington, C. P. and Whitaker, H. J. (2005) Contact surface models for infectious diseases: estimation from serologic survey data. *J. Am. Statist. Ass.*, **100**, 370–379.

Federale Overheidsdienst Economie Afdeling Statistiek (2006) Levensverwachting bij de geboorte, per gewest en internationale vergelijking. Federale Overheidsdienst Economie Afdeling Statistiek, Brussels. (Available from `http://mineco.fgov.be/`.)

Ferguson, N. M., Anderson, R. M. and Garnett, G. P. (1996) Mass vaccination to control chickenpox: the influence of zoster. *Proc. Natn. Acad. Sci. USA*, **93**, 7231–7235.

Garnett, G. P. and Grenfell, B. T. (1992) The epidemiology of varicella-zoster virus infections: a mathematical model. *Epidem. Infectn*, **108**, 495–511.

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

Halloran, M. E. (2006) Challenges of using contact data to understand acute respiratory disease transmission. *Am. J. Epidem.*, **164**, 945–946.

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P. and Beutels, P. (2008) Modelling multi-sera data: the estimation of new joint and conditional epidemiological parameters. *Statist. Med.*, **27**, 2651–2664.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2009a) Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infect. Dis.*, **9**, article 5.

Hens, N., Kvitkovicova, A., Aerts, M., Hlubinka, D. and Beutels, P. (2009b) Modelling distortions in seroprevalence data using change-point fractional polynomials. *Statist. Modllng*, to be published.

Kanaan, M. N. and Farrington, C. P. (2005) Matrix models for childhood infections: a bayesian approach with applications to rubella and mumps. *Epidem. Infectn*, **133**, 1009–1021.

Longini, I. M., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D. A. T. and Halloran, M. E. (2005) Containing pandemic influenza at the source. *Science*, **309**, 1083–1087.

Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001) A general projection framework for constrained smoothing. *Statist. Sci.*, **16**, 232–248.

Melegaro, A., Jit, M., Gay, N., Zagheni, E. and Edmunds, W. J. (2009) What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns. *Technical Report*. Health Protection Agency, London.

Mikolajczyk, R. T. and Kretzschmar, M. (2008) Collecting social contact data in the context of disease transmission: prospective and retrospective study designs. *Socl Netwrks*, **30**, 127–135.

Mossong, J., Hens, N., Friederichs, V., Davidkin, I., Broman, M., Litwinska, B., Siennicka, J., Trzcinska, V. P. A., Beutels, P., Vyse, A., Shkedy, Z., Aerts, M., Massari, M. and Gabutti, G. (2008a) Parvovirus B19 infection in five European countries: seroepidemiology, force of infection and maternal risk of infection. *Epidem. Infectn*, **136**, 1059–1068.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Scalia Tomba, G., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, W. J. (2008b) Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.*, **5**, no. 3.

Nardone, A., de Ory, F., Carton, M., Cohen, D., van Damme, P., Davidkin, I., Rota, M., de Melker, H., Mossong, J., Slacikova, M., Tischer, A., Andrews, N., Berbers, G., Gabutti, G., Gay, N., Jones, L., Jokinen, S., Kafatos, G., Martínez de Aragón, M. V., Schneider, F., Smetana, Z., Yargova, B., Vranckx, R. and Miller, E. (2007) The comparative sero-epidemiology of varicella zoster virus in 11 countries in the european region. *Vaccine*, **25**, 7866–7872.

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P. and Beutels, P. (2009) Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. *Math. Biosci.*, **218**, 80–87.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P. and Beutels, P. (2009) Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidem. Infectn*, **137**, 48–57.
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. Epidem.*, **164**, 936–944.
Whitaker, H. J. and Farrington, C. P. (2004) Infections with varying contact rates: application to varicella. *Biometrics*, **60**, 615–623.
Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.