Biostatistics Advance Access published July 2, 2014

Biostatistics (2014), *pp.* 1–14
doi:10.1093/biostatistics/kxu031

Modeling individual heterogeneity in the acquisition of recurrent infections: an application to parvovirus B19

STEVEN ABRAMS*

Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Agoralaan 1 Gebouw D, B3590 Diepenbeek, Belgium

steven.abrams@uhasselt.be

NIEL HENS

Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Agoralaan 1 Gebouw D, B3590 Diepenbeek, Belgium and Centre for Health Economics Research and Modeling Infectious Diseases and Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Universiteitsplein 1, B2610 Wilrijk, Belgium

Summary

In recent years, it has been shown that individual heterogeneity in the acquisition of infectious diseases has a large impact on the estimation of important epidemiological parameters such as the (basic) reproduction number. Therefore, frailty modeling has become increasingly popular in infectious disease epidemiology. However, so far, using frailty models, it was assumed infections confer lifelong immunity after recovery, an assumption which is untenable for non-immunizing infections. Our work concentrates on refining the existing frailty models to encompass complexities of waning immunity and consequently recurrent infections while accounting for individual heterogeneity. Univariate and shared gamma frailty models, frequently used in practice, and correlated gamma frailty models that have proven to be a valuable alternative are considered. We show that incorrectly assuming lifelong immunity when applying frailty models introduces substantial bias in the estimation of both the baseline hazard and the frailty parameters, and consequently of the basic and effective reproduction number. We illustrate our work using cross-sectional serological data on parvovirus B19 (PVB19) from Belgium for which the link with varicella zoster virus is exploited.

Keywords: Reproduction number; Serological data; SIR and SIRS transmission models; Social contact hypothesis; Univariate, Shared and correlated gamma frailty models.

1. Introduction

Frailty models are frequently used in survival analysis to model univariate and multivariate event times. The univariate frailty model, introduced by Vaupel *and others* (1979) and Lancaster (1979) to biostatistics and econometric literature, respectively, incorporates a latent frailty variable representing unobserved

*To whom correspondence should be addressed.

individual characteristics of a specific subject. In multivariate frailty modeling, the concept of frailty terms is used to model associations among event times, which goes back to the early work by Clayton (1978). Such models are based on a common and general approach of assuming independence among multivariate event times conditional on a set of latent variables (random effects). Upon the specification of a distribution for these latent variables, one obtains a multivariate model for the observed data for which the dependence structure arises when common or dependent latent variables enter in the conditional survival functions. Traditionally, frailty models are derived under a conditional independence assumption by specification of latent frailty terms that act multiplicatively on the baseline hazard. As individuals differ in their risk of acquiring an infection, these frailty models have found their way into the field of infectious disease epidemiology. Whereas interest in this paper goes to the specific use of frailty models in infectious disease epidemiology, the findings as reported here are of interest to the use of frailty models in the general survival context.

Coutinho *and others* (1999) were the first to explicitly account for individual heterogeneity in the acquisition of infections when estimating the infection hazard or force of infection. Farrington *and others* (2001) used the shared gamma frailty model in the context of bivariate current status data on measles and mumps in the UK. Hens *and others* (2009) illustrated that the restrictive assumption of a common frailty for both infections can be relaxed by using a correlated gamma frailty model, albeit at the cost of specifying a parametric baseline hazard function. From an epidemiological point of view, traditional frailty models in infectious disease epidemiology assume lifelong immunity after recovery from the infection. For some diseases, however, reinfections with the same pathogen are possible and the assumption of lifelong immunity therefore becomes untenable. Since individual heterogeneity inflates the estimates for both the basic reproduction number and critical vaccination coverage (Farrington *and others*, 2001), a correct assessment of heterogeneity, and therefore a correct specification of the disease processes, is of utmost importance to obtain reliable estimates for these quantities. Goeyvaerts *and others* (2011) considered some of these mathematical transmission models allowing for waning immunity of immunoglobulin G (IgG) antibodies against parvovirus B19 (PVB19) infection, thereby illustrating that models with waning immunity processes greatly improved model fit. Moreover, in general, susceptible-infectious-recovered-susceptible transmission dynamics (SIRS), referring to the different states in the compartmental model, were found to be reasonably successful in describing the observed serological profile of PVB19. Consequently, SIRS transmission models are used to derive frailty models encompassing potential reinfections. Bivariate gamma frailty models are extended and evaluated using bivariate current status data on PVB19 and varicella zoster virus (VZV).

This paper is organized as follows. We start by introducing cross-sectional multisera data on PVB19 and VZV and data on social contacts in Section 2. In Section 3, we introduce the concepts required to model these current status data. Thereafter, refined frailty models are explicitly derived for shared and correlated frailty terms. In addition, the mass action principle is incorporated through which the disease-specific forces of infection are related to the data on social mixing behavior. The application to serological data on PVB19 and VZV is shown in Section 4. Finally, we discuss the conclusions drawn from our data application, and present some avenues for further research in Section 5.

2. DATA

The methodology in this paper is illustrated using serological survey data on PVB19 and VZV infections in Belgium. Residual blood samples were collected in Belgium between 2001 and 2003 and tested for the presence of infection-specific IgG antibodies, reflecting infection experience. These blood samples are tested using a so-called enzyme-linked immunosorbent assay test, thereby classifying samples as either being seropositive or seronegative based on the cut-off level specified by the manufacturer, resulting in

a binary outcome. Hence, the serological status of the individual is a direct measure of his/her immunity against the disease, at least if complete serological protection is agreed upon. In addition, age at the time of data collection was registered. Since the true infection time is unobserved, and infection takes place either between birth and the monitoring time for seropositives or after the sampling time for seronegatives, we are faced with type I interval-censored or current status data. The statistical analysis included in this paper is limited to serological data on 2974 subjects with known immunological status for both PVB19 and VZV.

PVB19 causes a wide variety of diseases of which a childhood rash called fifth disease is the most well-known clinical presentation. On the other hand, primary infection with VZV results in varicella (chickenpox) and mainly occurs in childhood. Upon recovery from varicella, the virus stays dormant within the human body and may be reactivated after years to decades, giving rise to herpes zoster. Despite the fact that herpes zoster is an important aspect of varicella zoster infections, herpes zoster occurrences are unobserved in the serological profile and are therefore left aside. Transmission of both viruses occurs through direct close person-to-person contact. As transmission rates relate to contact rates, data on social mixing from the Belgian POLYMOD survey, a large-scale prospective survey conducted between May 2005 and September 2006, are used in addition to the serology on PVB19 and VZV.

3. Materials and methods

3.1 *Notation*

Consider bivariate cross-sectional serological data (y_1, y_2, a) with y_i the observed immunological status of the individual for infection $i = 1, 2$ and a his/her age at monitoring time. Indices referring to the individual level are suppressed from notation. The immunological status of an individual for infection $i = 1, 2$ is defined as the random variable Y_i which takes value one (zero) if he/she is seropositive (seronegative). The time variable represents the individual's age at data collection, expressed in years. Based on these kind of data, the true event time T_i^* is unknown and censored. In fact, cross-sectional serological data constitute type I interval-censored or current status data for which the true event time T_i^* lies in the interval $[0, T_i)$ or $[T_i, \infty)$, where T_i is the monitoring time with regard to infection i. Henceforth, univariate monitoring times $T_1 = T_2 = T$ equal to the age at sampling, a, are considered in all derivations. The binary response variable Y_i, conditional on the age a, follows a Bernoulli distribution with probability of being seropositive $\pi_i(a)$ ($Y_i|a \sim \text{Binomial}(1, \pi_i(a))$). In this paper, a parametric model for $\pi_i(a)$ is assumed for which the estimation of the model parameters $\boldsymbol{\theta}$ is performed using maximum likelihood estimation. The seroprevalence $\pi_i(a)$ is related to the proportion of susceptible individuals $S_i(a)$ in the population through $\pi_i(a) = 1 - S_i(a)$, where $S_i(a)$ is termed the survival function. The individual contribution of bivariate current status data (y_1, y_2, a) to the log-likelihood function is given by

$$\begin{aligned} ll(y_1, y_2, a|\boldsymbol{\theta}) = {} & y_1 y_2 \log(p_{11}(a|\boldsymbol{\theta})) + y_1(1 - y_2)\log(p_{10}(a|\boldsymbol{\theta})) \\ & + (1 - y_1)y_2 \log(p_{01}(a|\boldsymbol{\theta})) + (1 - y_1)(1 - y_2)\log(p_{00}(a|\boldsymbol{\theta})), \end{aligned} \tag{3.1}$$

where $(p_{11}(a|\boldsymbol{\theta}), p_{10}(a|\boldsymbol{\theta}), p_{01}(a|\boldsymbol{\theta}), p_{00}(a|\boldsymbol{\theta}))$ defines the multinomial probability distribution for $\mathbf{Y}|a$, $\mathbf{Y} = (Y_1, Y_2)$: $p_{11}(a|\boldsymbol{\theta}) = 1 - S_1(a) - S_2(a) + S_{12}(a)$, $p_{10}(a|\boldsymbol{\theta}) = S_2(a) - S_{12}(a)$, $p_{01}(a|\boldsymbol{\theta}) = S_1(a) - S_{12}(a)$, $p_{00}(a|\boldsymbol{\theta}) = S_{12}(a)$. In general, $p_{\alpha_1\alpha_2}(a|\boldsymbol{\theta}) = \Pr(Y_1 = \alpha_1, Y_2 = \alpha_2|a, \boldsymbol{\theta})$, where the indices $\alpha_1, \alpha_2 = 0, 1$ refer to the immunological status with respect to infections 1 and 2, respectively. Hence, the log-likelihood function for n subjects equals $\sum_{j=1}^{n} ll(y_{1j}, y_{2j}, a_j|\boldsymbol{\theta})$. For the sake of simplicity, the dependence of the survival functions on the model parameters $\boldsymbol{\theta}$ is suppressed from notation throughout the rest of the paper. In conclusion, parametric models for the seroprevalence $\pi_i(a)$ result in expressions for the marginal and joint survival functions that are substituted into the log-likelihood function. The log-likelihood function is then maximized to obtain maximum likelihood estimates for the model parameters. Throughout

the paper, the subscript i refers exclusively to a specific infection and individual-specific subscripts are generally omitted.

3.2 *Infection dynamics*

Although a lot of different mathematical transmission models can be found in the infectious disease literature, we will focus on the most simple ones describing disease dynamics for infections either conferring lifelong immunity or not. The most simple compartmental model is the so-called SIR (susceptible-infected-recovered) model. To facilitate a mathematical derivation of the disease dynamics, we make several assumptions (see, e.g. Hens *and others*, 2012). First, we assume that the infection is in endemic equilibrium, meaning that the disease incidence may undergo cyclical epidemics, however, fluctuating around a stationary average over time, which is believed to be a tenable assumption for PVB19 and VZV. Further, we assume that the population has reached a demographic equilibrium implying that the age distribution is stationary. Finally, the number of births and deaths are assumed to be constant over time and exactly balanced, entailing a constant population of size N.

In the time-homogeneous SIR model, individuals of age a flow from the susceptible compartment S to the infectious compartment I at an age-dependent rate $\lambda(a)$. After being infected with the agent, these individuals recover at a rate γ and move towards the recovered class R in which they remain until death. Individuals in each state experience natural mortality at a rate $\mu(a)$. Disease-related mortality is neglected, which is a plausible assumption, at least for the childhood infections under consideration in a developed country such as Belgium. Therefore, the SIR model mimics simple features corresponding to immunizing infections. However, as argued before, not all infectious diseases can be successfully described using an SIR model. Extensions of the basic SIR model are numerous but for the purpose of this paper we consider the slightly more complicated SIRS model. In the SIRS model, recovered individuals of age a are allowed to move back to the susceptible class at a replenishment rate $\sigma(a)$. A description of the SIRS model is included in Appendix A of supplementary material available at *Biostatistics* online.

3.3 *Waning antibodies and individual heterogeneity*

Consider a univariate frailty term Z describing individual heterogeneity. The proportion of susceptible individuals of age a with frailty Z, say $S(a|Z)$, evolves following the ODE:

$$\frac{\mathrm{d}S(a|Z)}{\mathrm{d}a} = -\lambda(a, Z)S(a|Z) + \sigma(a)R(a|Z), \tag{3.2}$$

with $\lambda(a, Z)$ the conditional force of infection, $\sigma(a)$ the replenishment rate and $R(a|Z)$ the proportion of seropositives of age a given the frailty term Z. In this paper, the replenishment rate $\sigma(a)$ is assumed to be independent of Z. The differential equation (3.2) can be solved using $R(a|Z) \approx 1 - S(a|Z)$ to obtain:

$$S(a|Z) = \exp\left(-\int_0^a \{\lambda(u, Z) + \sigma(u)\}\,\mathrm{d}u\right) + \int_0^a \sigma(u)\exp\left(-\int_u^a \{\lambda(v, Z) + \sigma(v)\}\,\mathrm{d}v\right)\mathrm{d}u. \tag{3.3}$$

Note that the expression for $S(a|Z)$ reduces to the general exponential formula for an immunizing infection with hazard rate $\lambda(a, Z)$ if the replenishment rate $\sigma(a)$ equals 0, $\forall a$. We rely on the proportional hazards assumption such that the frailty term Z acts multiplicatively on the baseline force of infection $\lambda_0(a)$, i.e. $\lambda(a, Z) = Z\lambda_0(a)$ (see, e.g. Wienke, 2010). The unconditional survival function $S(a)$ equals the expectation of $S(a|Z)$ with respect to Z, i.e. $S(a) = \mathrm{E}[S(a|Z)]$. The expression for the unconditional

survival function $S(a)$ becomes:

$$S(a) = \exp\left(-\int_0^a \sigma(u)\,\mathrm{d}u\right)\mathbf{L}(M_0(a)) + \int_0^a \sigma(u)\exp\left(-\int_u^a \sigma(v)\,\mathrm{d}v\right)\mathbf{L}(M_0(a) - M_0(u))\,\mathrm{d}u$$
$$= W_\sigma(Q_\sigma(a))\mathbf{L}(M_0(a)) + \int_0^a \sigma(u)W_\sigma(Q_\sigma(a) - Q_\sigma(u))\mathbf{L}(M_0(a) - M_0(u))\,\mathrm{d}u, \quad (3.4)$$

where $W_\sigma(s) = \exp(-s)$, $Q_\sigma(a)$ is the cumulative replenishment rate and $M_0(a)$ is the cumulative baseline hazard function; $\mathbf{L}(s)$ represents the Laplace transform of the random variable Z, for which the explicit form depends on the selected frailty distribution. In many applications, the mathematically convenient gamma frailty distribution is considered, and therefore without loss of generality we focus on gamma distributed frailty terms within this paper. The Laplace transform of a gamma distributed random variable Z with unit mean and variance σ_f^2 (i.e. $Z \sim \Gamma(1/\sigma_f^2, 1/\sigma_f^2)$) has a closed-form expression that takes the form $\mathbf{L}(s) = (1 + \sigma_f^2 s)^{-1/\sigma_f^2}$.

In the univariate frailty setting, the frailty variance represents heterogeneity due to unobserved factors in the population that have an impact on the infection time. However, modeling the occurrence of multiple infections simultaneously allows us to quantify the association in acquisition of these infections and to draw conclusions regarding similarities in routes of transmission.

3.4 *Bivariate shared and correlated gamma frailty models*

First, let Z represent a shared gamma frailty term for both infections under investigation. The frailty term indicates how frail individuals are with respect to the acquisition of both infections. Therefore, the interpretation of the frailty variance is fundamentally different from the one in the univariate setting. The general form of the marginal survival function $S_i(a)(i = 1, 2)$, where i denotes the specific infection under investigation, depends upon the constitution of the underlying infection process. For example, for an immunizing infection, say i, conferring lifelong immunity after recovery, one easily ends up with the following formula for the marginal unconditional survival function (see (3.4)): $S_i(a) = \mathbf{L}(M_{i0}(a))$, where $\mathbf{L}(s)$ is the Laplace transform of $Z \equiv Z_i$ and $M_{i0}(a)$ is the cumulative baseline hazard function corresponding to infection i. For non-immunizing infections, the expression becomes more complicated. Essentially, the marginal survival function $S_i(a)$ simply reduces to the formula in (3.4), except for the specification of infection-specific parameters such as $\sigma_i(a)$, $M_{i0}(a)$ and $Q_{\sigma_i}(a)$ with the same meaning as before. An expression for the joint unconditional survival function $S_{12}(a)$ can be derived relying on conditional independence of the event times given the frailty term Z (Appendix B of supplementary material available at *Biostatistics* online).

The shared gamma frailty model suffers from the restrictive assumption of perfect correlation among infection-specific frailty terms. In some situations, a more flexible alternative is required that relaxes the assumption of perfect correlation and allows for a more general correlation structure. Yashin *and others* (1995) introduced the correlated gamma frailty model accounting for a positive correlation between frailty terms. Recently, Hens *and others* (2009) have illustrated its value in the context of bivariate current status data.

Following the notation by Hens *and others* (2009), frailty terms $Z_i(i = 1, 2)$ are additively decomposed into independent gamma distributed random variables $Y_l^*(l = 0, 1, 2)$ with mean and variance $k_l > 0$: $Z_i = \sigma_{if}^2(Y_0^* + Y_i^*)$, where σ_{if}^2 represents the frailty variance. The superscript * is used to avoid confusion with the infection-dependent immunological status Y_i of an individual. Consequently, the infection-specific frailty terms $Z_i(i = 1, 2)$ are gamma distributed. Assuming the frailty terms to have unit mean, one easily derives expressions for the frailty variances σ_{if}^2 in terms of the component parameters $k_l(l = 0, 1, 2)$: $\sigma_{if}^2 = 1/(k_0 + k_i)$. Hence, the variance-covariance structure imposed by the additive decomposition of the frailties implies a correlation between Z_1 and Z_2. The Pearson product–moment correlation coefficient for

the frailty variables equals $\rho_{12} = k_0/\sqrt{(k_0+k_1)(k_0+k_2)}$ inducing an upper bound on the strictly positive correlation coefficient: $0 \leqslant \rho_{12} \leqslant \min(\sigma_{1f}/\sigma_{2f}, \sigma_{2f}/\sigma_{1f})$ which is rather restrictive when the frailty variances differ greatly. The marginal unconditional survival function $S_i(a)$ can be written as:

$$S_i(a) = \mathbf{L}_i(M_{i0}(a))W_{\sigma_i}(Q_{\sigma_i}(a)) + \int_0^a \sigma_i(u)W_{\sigma_i}(Q_{\sigma_i}(a) - Q_{\sigma_i}(u))\mathbf{L}_i(M_{i0}(a) - M_{i0}(u))\,\mathrm{d}u. \quad (3.5)$$

The only difference compared with the shared frailty setting is the dependency of the Laplace transform $\mathbf{L}_i(s)$ on the specific infection i through its frailty variance σ_{if}^2. Making use of the conditional independence assumption and independence of the gamma components Y_l^*, the unconditional bivariate survival function can be expressed in terms of the component-specific Laplace transforms $\mathbf{L}_{Y_l^*}$ $(l = 0, 1, 2)$ (Appendix B of supplementary material available at *Biostatistics* online).

Correlated gamma frailty models are identifiable at least when specifying parametric baseline forces of infection and replenishment rates. In the next section, we present the mass action principle in the presence of individual heterogeneity, which yields a parametric model for the baseline forces of infection given a known contact structure.

3.5 *Baseline force of infection*

The mass action principle relates the baseline force of infection $\lambda_0(a)$ to an augmented effective contact function $\beta(a, Z; a', Z')$ representing the per capita rate at which an individual of age a' and frailty Z' makes effective contact with an individual of age a and frailty Z. Under a multiplicative assumption, the augmented contact function can be written as $\beta(a, Z; a', Z') = ZZ'\beta_0(a, a')$. The multiplicative assumption implies a proportional hazards assumption for $\lambda(a, Z)$ such that the time homogeneous mass action principle can be rendered as follows (Farrington *and others*, 2001):

$$\lambda(a, Z) = \frac{ND}{L}\int_0^\infty\int_0^\infty ZZ'\beta_0(a, a')\lambda(a', Z')S(a'|Z')\phi(a')f(Z')\,\mathrm{d}a'\,\mathrm{d}Z', \quad (3.6)$$

with $f(Z')$ the probability density function for the frailty variable Z', $S(a'|Z')$ the proportion of susceptible individuals of age a' given frailty Z', N the total population size, D the mean duration of infectiousness, L the life expectancy and $\phi(a') = \exp\left(-\int_0^{a'}\mu(u)\,\mathrm{d}u\right)$ the probability of being alive at age a'. Using the proportional hazards assumption, (3.6) can be simplified to

$$\lambda_0(a) = \frac{ND}{L}\int_0^\infty\int_0^\infty Z'\beta_0(a, a')\lambda(a', Z')S(a'|Z')\phi(a')f(Z')\,\mathrm{d}a'\,\mathrm{d}Z'. \quad (3.7)$$

In order to integrate data on social contact behavior in the estimation of the age-dependent baseline infection hazard, the baseline effective contact function $\beta_0(a, a')$ is written as $q(a, a'|c) \times c(a, a')$ (Wallinga *and others*, 2006). The proportionality factor $q(a, a'|c)$ represents the transmission potential upon a contact between an individual of age a' and one of age a. This might depend on several characteristics related to susceptibility and infectiousness which could be ethnic-, climate-, disease- and/or age-specific. In addition, conversational and physical contacts reported in social contact diaries serve as proxies of those contacts by which the infection is successfully transmitted. Therefore, $q(a, a'|c)$ can be considered as an age-specific adjustment factor relating true contact rates underlying disease transmission to reported proxies. In addition, $c(a, a')$ is the annual rate at which individuals of age a' contact individuals of age a within the population. The contact rates are estimated from Belgian social contact data (Section 2 of Goeyvaerts *and others*, 2011). Solving the mass action principle involves finding a solution without explicit closed-form expression. Consequently, an iterative procedure described in

Kanaan and Farrington (2005) is used to solve the mass action principle numerically after turning to a discrete age framework, assuming a constant force of infection within each age interval. The reader is referred to Appendix C of supplementary material available at *Biostatistics* online for more details on the mass action principle.

3.6 *Reproduction numbers*

For an overview of the key ingredients and a graphical representation of the estimation procedure, we refer to Appendix D of supplementary material available at *Biostatistics* online. Once the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ of the model parameters are obtained, the basic reproduction number R_{i0} with respect to infection i, that is, the average number of secondary infections produced by a single infectious individual during his/her entire infectious period when introduced in a fully susceptible population, is computed as $(1+\hat{\sigma}_{if}^2)$ times the dominant eigenvalue of the next generation matrix $ND_iL^{-1}\phi(a)\beta_{i0}(a,a')$ (Diekmann *and others*, 1990). Furthermore, the effective reproduction number R_i is the equivalent of R_{i0} in a population which is not entirely susceptible, and is therefore defined as the leading eigenvalue of the matrix $ND_iL^{-1}\phi(a)S_i(a)\beta_{i0}(a,a')$ (see Appendix E of supplementary material available at *Biostatistics* online).

4. Results

In this section, we describe the results of fitting the shared and correlated gamma frailty models described in Section 3.4 to bivariate serological data on PVB19 and VZV. Despite the general formulation of the models in terms of infection-specific age-dependent proportionality factors $q_i(a,a'|c)$, these terms are considered to be age-invariant in all fitted models as age-dependent proportionality in transmission did not reveal any improvement in model fit. Drawing especially on the conclusions of Goeyvaerts *and others* (2011), the transmission characteristics for PVB19 infection are modeled through an SIRS transmission model in which reinfections are possible. In contrast to potential reinfections with PVB19, varicella zoster infections are assumed to confer lifelong immunity as more complex disease dynamics did not improve model fit. Especially, allowing for reinfections with VZV resulted in an estimated replenishment rate which was not significantly different from zero (not shown).

Table 1 links the model definitions in Section 3.4 with the candidate models presented here. Frailty models relying on the assumption of lifelong immunity for both infections are denoted by *UGF-1*, *SGF-1* and *CGF-1* for univariate, shared and correlated gamma frailty models, respectively. In addition, *UGF-2a*, *SGF-2a* and *CGF-2a* models allow for replenishment of the susceptible compartment at a constant rate σ_1, at least for PVB19. Finally, *UGF-2b*, *SGF-2b* and *CGF-2b* models simply extend the previous models by introducing an age-dependent dichotomous replenishment for PVB19 based on a cut-off value H (σ_{11} for $a<H$ and σ_{12} for $a\geqslant H$). The optimal cut-off value is selected to be $H=35$ years based on a grid search. As advocated previously, infections with VZV are considered to confer lifelong immunity leading to $\sigma_2(a)=0, \forall a$. Furthermore, the joint survival function $S_{12}(a)$ in univariate gamma frailty models simply reduces to the product of $S_1(a)$ and $S_2(a)$, indicated by "$S_1(a)\times S_2(a)$" in Table 1. The general expressions for the unconditional survival functions simplify according to the structure of the replenishment rates. The statistical analysis was performed using R, version 2.14.2 (R Core Team, 2012). Documented R-code is made available in Appendix J of supplementary material available at *Biostatistics* online.

In Table 2, the parameter estimates for the model parameters are presented together with bootstrap 95% percentile confidence limits based on $B=500$ bootstrap samples. Univariate (UGF), shared (SGF) and correlated (CGF) gamma frailty models are fitted to the serological data at hand under different scenarios for the transmission dynamics of PVB19. The results in Table 2 clearly indicate that the univariate, shared

Table 1. *Definition of the candidate models fitted to the bivariate serological survey data on PVB19* $(i = 1)$ *and VZV* $(i = 2) : I_H = I(a < H)$. *Simple SIR dynamics are assumed for VZV, giving rise to a replenishment* $\sigma_2(a) = 0, \forall a$.

Model specification	Frailty	Dynamics PVB19	Replenishment $\sigma_1(a)$	Survival functions		
				$S_1(a)$	$S_2(a)$	$S_{12}(a)$
UGF-1	Univariate	SIR	0	Equation (3.4)	Equation (3.4)	$S_1(a) \times S_2(a)$
UGF-2a	Univariate	SIRS	σ_1	Equation (3.4)	Equation (3.4)	$S_1(a) \times S_2(a)$
UGF-2b	Univariate	SIRS	$\sigma_{11} I_H + \sigma_{12}(1 - I_H)$	Equation (3.4)	Equation (3.4)	$S_1(a) \times S_2(a)$
SGF-1	Shared	SIR	0	Equation (3.4)	Equation (3.4)	Equation (B.1)†
SGF-2a	Shared	SIRS	σ_1	Equation (3.4)	Equation (3.4)	Equation (B.1)
SGF-2b	Shared	SIRS	$\sigma_{11} I_H + \sigma_{12}(1 - I_H)$	Equation (3.4)	Equation (3.4)	Equation (B.1)
CGF-1	Correlated	SIR	0	Equation (3.5)	Equation (3.5)	Equation (B.2)†
CGF-2a	Correlated	SIRS	σ_1	Equation (3.5)	Equation (3.5)	Equation (B.2)
CGF-2b	Correlated	SIRS	$\sigma_{11} I_H + \sigma_{12}(1 - I_H)$	Equation (3.5)	Equation (3.5)	Equation (B.2)

† Appendix B of supplementary material available at *Biostatistics* online.

and correlated gamma frailty models with SIRS dynamics for PVB19 outperform their SIR counterparts assuming lifelong immunity after recovery for both infections. Moreover, based on the Akaike Information Criterion (AIC) it turns out that the SGF-2b model is the best fitting model, whereas the Bayesian Information Criterion (BIC) selects the SGF-2a model. Despite the fact that both information criteria select different models, the key message lies within the improved fit when accounting for potential PVB19 reinfections compared with the traditional approach in which lifelong immunity is presumed. It is not completely clear what causes the decreased replenishment rate for people aged 35 years and above. It could reflect the general observation that infection or boosting through exposure to individuals who are infectious with PVB19, elicits higher antibody responses in mature immune systems, which could prolong the process of antibody waning (Goeyvaerts *and others*, 2011). Although more flexible than shared frailty models, using correlated frailty models here shows little advantage over using shared frailty models, at least when accounting for reinfections with PVB19. Given that modeling the underlying infection process decreases the unobserved infection-specific heterogeneity, it is no surprise that SGF-2a and SGF-2b frailty models are preferred over their correlated counterparts (CGF-2a and CGF-2b). Since the estimated correlation, implied by the additive decomposition of the correlated frailties, is bounded by the ratio of the frailty standard deviations, the bootstrap-based 95% confidence intervals for ρ_{12} in CGF-2a and CGF-2b are asymmetric.

The frailty variances in the univariate frailty models (UGF) are not comparable with those in the bivariate models since their interpretation differs. In the univariate setting, the frailty variance reflects heterogeneity with respect to unobserved factors in the population. In contrast, in any bivariate frailty model, frailty terms impose a correlation structure among infections. In Figure 1, the estimated marginal seroprevalences for PVB19 and VZV are graphically displayed based upon the three bivariate shared gamma frailty models (SGF-1, SGF-2a and SGF-2b). The differences in estimated seroprevalence are more pronounced for PVB19 compared with VZV due to the fact that the models only differ in assumptions regarding PVB19 transmission. Fitting models with SIRS dynamics for both infections leads to an equivalent performance in terms of log-likelihood value at the cost of one extra parameter which is estimated not significantly different from 0. This coincides with our previous arguments to settle for SIR features with regard to VZV.

In Figure 2, the estimated proportions of individuals having one of the four possible serological profiles, determined by their status with respect to both infections, are displayed. These multinomial probabilities

Table 2. *Maximum likelihood estimates for the model parameters as well as for the basic and effective reproduction numbers R_{i0} and R_i, respectively, and together with 95% bootstrap percentile confidence intervals in square brackets. The corresponding AIC- and BIC-values (minima underlined), obtained under the assumption of Type I mortality. Application of univariate, bivariate shared and correlated gamma frailty models to serology on PVB19 ($i = 1$) and VZV ($i = 2$).*

Model				$\hat{R}_0$		$\hat{R}$		AIC	BIC
UGF-1	q_{10}	0.086	[0.079; 0.094]	5.27	[4.47; 6.22]	1.831	[1.568; 2.142]	4506.27	4530.26
	σ^2_{1f}	0.435	[0.316; 0.560]						
	q_{20}	0.168	[0.159; 0.187]	8.36	[7.92; 9.77]	1.143	[1.133; 1.259]		
	σ^2_{2f}	3.0e−6	[3.4e−7; 0.053]						
	ρ_{12}	0.000	–						
UGF-2a	q_{10}	0.071	[0.068; 0.074]	3.02	[2.89; 3.52]	1.054	[1.049; 1.229]	4481.84	4511.82
	σ_1	0.011	[0.007; 0.015]						
	σ^2_{1f}	9.7e−7	[3.4e−7; 0.117]						
	q_{20}	0.168	[0.159; 0.187]	8.36	[7.92; 9.77]	1.143	[1.133; 1.259]		
	σ^2_{2f}	3.0e−6	[3.4e−7; 0.053]						
	ρ_{12}	0.000	–						
UGF-2b	q_{10}	0.071	[0.068; 0.074]	3.04	[2.92; 3.16]	1.063	[1.055; 1.072]	4477.00	4512.99
	σ_{11}	0.017	[0.012; 0.023]						
	σ_{12}	0.008	[0.005; 0.012]						
	σ^2_{1f}	1.5e−6	[3.4e−7; 4.0e−6]						
	q_{20}	0.168	[0.159; 0.187]	8.36	[7.92; 9.77]	1.143	[1.133; 1.259]		
	σ^2_{2f}	3.0e−6	[3.4e−7; 0.053]						
	ρ_{12}	0.000	–						
SGF-1	q_{10}	0.073	[0.069; 0.077]	3.59	[3.27; 3.90]	1.278	[1.189; 1.368]	4537.28	4555.27
	q_{20}	0.209	[0.189; 0.232]	12.07	[10.46; 13.74]	1.516	[1.368; 1.664]		
	σ^2_f	0.158	[0.102; 0.210]						
	ρ_{12}	1.000	–						
SGF-2a	q_{10}	0.072	[0.068; 0.075]	3.17	[2.94; 3.43]	1.106	[1.052; 1.178]	4477.98	<u>4501.97</u>
	σ_1	0.011	[0.007; 0.014]						
	q_{20}	0.177	[0.162; 0.196]	9.15	[8.07; 10.53]	1.221	[1.139; 1.333]		
	σ^2_f	0.036	[5.4e−7; 0.086]						
	ρ_{12}	1.000	–						
SGF-2b	q_{10}	0.072	[0.069; 0.075]	3.13	[2.95; 3.38]	1.093	[1.057; 1.165]	<u>4474.39</u>	4504.38
	σ_{11}	0.016	[0.010; 0.022]						
	σ_{12}	0.008	[0.005; 0.012]						
	q_{20}	0.173	[0.161; 0.191]	8.82	[8.01; 10.13]	1.189	[1.136; 1.300]		
	σ^2_f	0.021	[3.6e−7; 0.071]						
	ρ_{12}	1.000	–						
CGF-1	q_{10}	0.086	[0.079; 0.094]	5.26	[4.44; 6.20]	1.827	[1.563; 2.135]	4505.62	4535.61
	q_{20}	0.180	[0.163; 0.200]	9.40	[8.19; 10.92]	1.246	[1.142; 1.375]		
	σ^2_{1f}	0.433	[0.310; 0.558]						
	σ^2_{2f}	0.048	[3.5e−7; 0.098]						
	ρ_{12}	0.332	[0.001; 0.501]						

(Continued)

Table 2. *Continued.*

Model				$\hat{R}_0$		$\hat{R}$		AIC	BIC
CGF-2a	q_{10}	0.072	[0.068; 0.075]	3.17	[2.94; 3.50]	1.106	[1.052; 1.210]	4481.98	4517.96
	σ_1	0.011	[0.007; 0.014]						
	q_{20}	0.177	[0.162; 0.196]	9.15	[8.07; 10.53]	1.221	[1.139; 1.333]		
	σ^2_{1f}	0.036	[4.6e−7; 0.105]						
	σ^2_{2f}	0.036	[4.6e−7; 0.086]						
	ρ_{12}	1.000	[0.565; 1.000]						
CGF-2b	q_{10}	0.071	[0.067; 0.074]	3.08	[2.89; 3.34]	1.077	[1.040; 1.148]	4478.53	4520.51
	σ_{11}	0.017	[0.011; 0.022]						
	σ_{12}	0.009	[0.006; 0.012]						
	q_{20}	0.173	[0.161; 0.192]	8.82	[8.01; 10.20]	1.188	[1.135; 1.304]		
	σ^2_{1f}	0.021	[3.5e−7; 0.071]						
	σ^2_{2f}	0.021	[3.5e−7; 0.072]						
	ρ_{12}	1.000	[0.653; 1.000]						

Single underline indicates a minimum in the univariate frailty setting (models UGC-1, UGC-2a and UGC-2b).

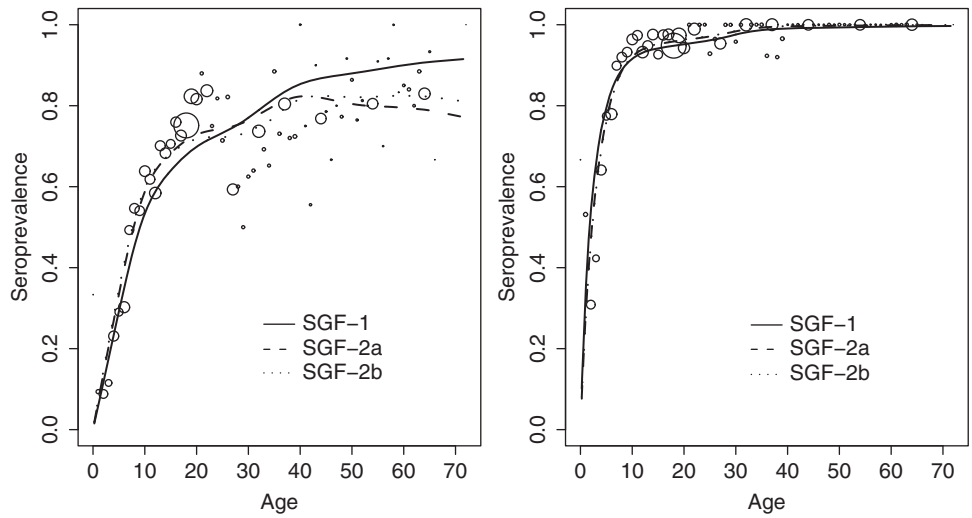

Fig. 1. The observed seroprevalence (dots with size proportional to the number of observations) and estimated seroprevalence of PVB19 (left panel) and VZV (right panel) based on three bivariate shared gamma frailty models: the SGF-1 model (solid line), the SGF-2a model (dashed line) and the SGF-2b model (dotted line).

are equal to $p_{\alpha_1\alpha_2}(a|\hat{\boldsymbol{\theta}})$, $\alpha_1, \alpha_2 = 0, 1$, as defined in Section 3.1, and $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimates regarding the model parameters. For example, $p_{11}(a|\hat{\boldsymbol{\theta}})$ denotes the probability to be infected with both infections before the age of a years. In line with the fitted marginal seroprevalence curves, one easily observes that the shared gamma frailty models with SIRS dynamics for PVB19 (SGF-2a and SGF-2b) improve the fit compared with the fit corresponding to the SGF-1 model. Especially, the largest improvements are seen with respect to $p_{11}(a|\hat{\boldsymbol{\theta}})$ and $p_{01}(a|\hat{\boldsymbol{\theta}})$ in the left upper and left lower panel, respectively. Interest lies in the quantification of the impact of misspecifying the underlying infection process on estimates for the reproduction numbers R_0 and R. In Table 2, the shared SIR gamma frailty model (SGF-1)

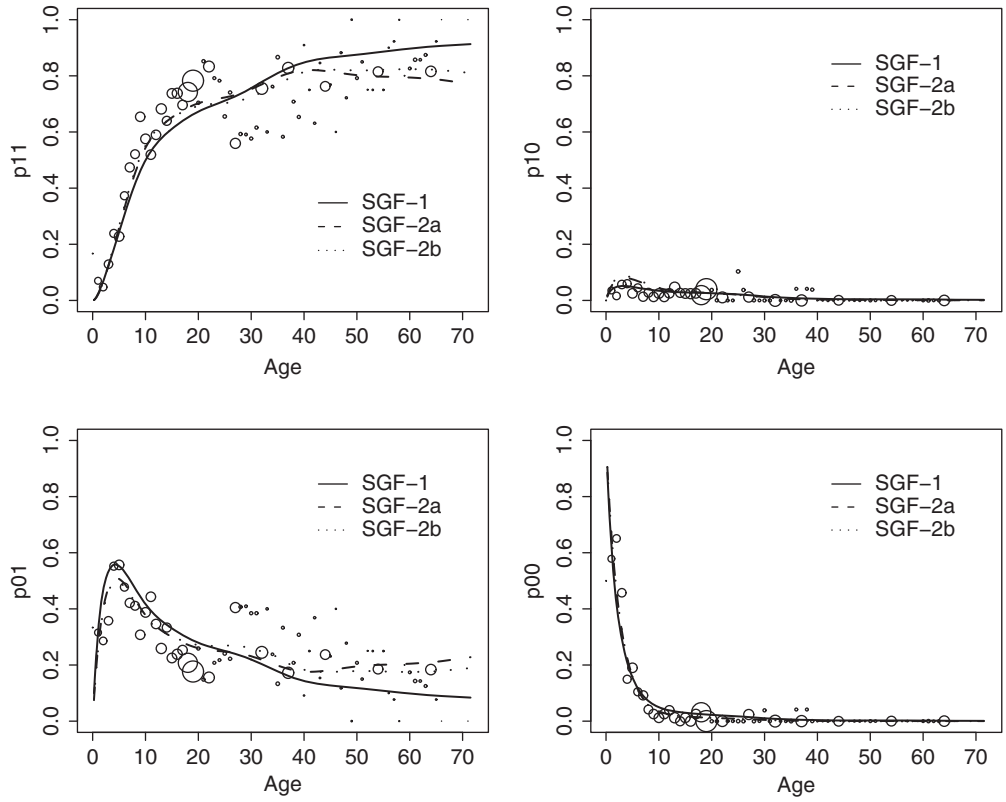

Fig. 2. The observed prevalence with respect to one of the four serological profiles (dots with size proportional to the number of observations) and estimated prevalences based on three bivariate shared gamma frailty models: the SGF-1 model (solid line), the SGF-2a model (dashed line) and the SGF-2b model (dotted line); $p_{11}(a|\hat{\boldsymbol{\theta}})$ (upper left panel), $p_{10}(a|\hat{\boldsymbol{\theta}})$ (upper right panel), $p_{01}(a|\hat{\boldsymbol{\theta}})$ (lower left panel) and $p_{00}(a|\hat{\boldsymbol{\theta}})$ (lower right panel).

yields larger R_0 estimates for both PVB19 and VZV (3.59 [3.27; 3.90] and 12.07 [10.46; 13.74], respectively, with 95% confidence limits between squared brackets) when compared with the estimates of the SGF-2a and SGF-2b models, encompassing reinfection dynamics for PVB19. Estimates for R_0 and R in the shared and correlated SIRS models (SGF-2a, SGF-2b, CGF-2a and CGF-2b) are almost identical (Table 2). Simulation results are shown in Appendix F of supplementary material available at *Biostatistics* online.

5. Discussion

Recently, frailty modeling gained much attention in infectious disease epidemiology due to the seminal work by Farrington *and others* (2001) and the work by Hens *and others* (2009). Although mathematical models describing various infection processes are extensively used to study the spread of infectious diseases, characteristics of infection processes that deviate from traditional SIR features are not yet incorporated in the frailty setting. In this paper, we illustrate the necessity to account for such features when infections are unlikely to confer lifelong immunity. The traditional frailty models are extended to encompass other disease dynamics when compared with lifelong immunity after infection. The impact of these refinements on the estimation of the basic and effective reproduction number is quantified in the context

of serological data on PVB19 and VZV. Our findings support the hypothesis of PVB19 recurrences in humans (see, e.g. Goeyvaerts *and others*, 2011), illuminating important epidemiological consequences if these dynamics are not appropriately accounted for. Furthermore, applying traditional frailty models to PVB19 serology induces the frailty variance to be largely overestimated, and as demonstrated, estimates for the basic reproduction number R_0 derived thereof are biased. Modeling the infection process of PVB19 more accurately decreases the amount of unexplained heterogeneity, and equivalently reduces the frailty variance associated therewith. Consequently, the rather restrictive upper bound on the correlation coefficient in the correlated frailty model is elevated. This feature adds to the general usefulness of the correlated gamma frailty model.

Although the refined shared gamma frailty models describe the age-dependent serological profile of PVB19 reasonably well, the decrease at the ages 20–30 is not fully captured. A more flexible shape for the age-dependent replenishment rate allows us to improve model fit further (not shown), albeit that the biological interpretation becomes more complicated. Nevertheless, additional work on this matter is required to ascertain whether the decline in seroprevalence results solely from age-dependent differences in replenishment. In the absence of any individual-specific covariates, individual heterogeneity is accounted for through the specification of latent gamma frailty variables. In spite of the mathematical convenience of the gamma frailty distribution, many other frailty distributions can be used in the presented models. However, such an analysis would not change our conclusions with respect to the infection dynamics of PVB19 (see Appendix G of supplementary material available at *Biostatistics* online). Furthermore, standard parametric models for the baseline hazard functions such as Gompertz or Weibull hazards are easily implemented as well, albeit that for the data at hand these models seem unable to capture the biological complexity.

Although we have limited our analyses to include only subjects with complete immunological information for both infections, one is able to extend the likelihood to accommodate for individuals with incomplete data. This direct-likelihood approach is described in Appendix H of supplementary material available at *Biostatistics* online and yields almost identical estimates. As paired serological data give rise to multinomial observations, a higher empirical variance might be observed than can be accommodated for in the presented models. One way to overcome this is to use a Dirichlet-Multinomial likelihood with an additional overdispersion parameter (Farrington *and others*, 2013). Further research is required to investigate the effects of simultaneous modeling individual heterogeneity in the acquisition of infections and overdispersion. In this paper, we aim at extending existing models encompassing frailties in the acquisition of infectious diseases to recurrent infections and at underlining the importance of a correct specification of the underlying infection process which is identified as the most important source of misspecification and for which the effects on the estimated reproduction numbers are far more pronounced. Note that Farrington *and others* (2013) studied robustness against misspecification of the contact function of which the impact is believed to be rather limited compared with the underlying infection process.

In addition, the use of cross-sectional serological survey data to model the seroprevalence relies on the assumption of perfect classification of individuals into a seropositive and seronegative group. Nevertheless, perfect testing is seldom achieved which could affect estimates with regard to the epidemiological parameters of interest. Although, to date, frailty modeling based on serology is performed without explicitly accounting for imperfect test results and without quantifying its impact on derived estimates, test sensitivity and specificity are useful quantities to adjust for potential misclassification (Appendix I of supplementary material available at *Biostatistics* online). Further research is needed to allow for age-dependent frailty terms in the context of recurrent infections. Recently, Farrington *and others* (2013) considered such an age-dependency in addition to the assumptions of lifelong immunity and shared frailty terms. Furthermore, an extension of the correlated frailty concept to more flexible decompositions of the frailty variables seems valuable. As pointed out earlier, correlated gamma frailty models rely on an additive decomposition of the latent frailties thereby restricting the range of the correlation between them. Although the correlation in these models does not describe the correlation among original infection times, it can be interpreted

as a valid correlation measure. Therefore, the ability to impose more flexible correlation structures could reveal additional information on the dependence among infections.

The approach undertaken in this paper allows us to model recurrent infections, even if the underlying infection process is known to be more complex than the one proposed here. However, cross-sectional serological data deviate from general recurrent event data in the sense that only a single observation per subject is available. In an epidemiological setting, acquiring recurrent event data rely on continuously monitoring subjects and these data are therefore hard to collect. Upon knowledge of the underlying infection mechanism, these models convey an alternative way of assessing potential violations of the lifelong immunity assumption and thereby retrieve the recurrent nature from cross-sectional serology. Whereas the use of the refined frailty models is illustrated on type I interval-censored data with a single observation per subject, these models offer the opportunity to model recurrent event data in the presence of any type of censoring, any parametric baseline hazard, any frailty distribution, and are therefore important in the general survival context.

Supplementary Material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Funding

This work was supported by the Research Fund of Hasselt University (BOF11NI31 to S.A.). N.H. acknowledges support from the University of Antwerp Scientific Chair in Evidence Based Vaccinology, financed by a gift from Pfizer.

Acknowledgments

For the data processing, we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government – department EWI. *Conflict of Interest*: None declared.

References

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**(1), 141–151.

Coutinho, F. A. B., Massad, E., Lopez, L. F., Burattini, M. N., Struchiner, C. and Azevedo-Neto, R. (1999). Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling* **30**(1–2), 97–115.

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. *Journal of Mathematical Biology* **28**(4), 365–382.

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. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(3), 251–292.

Farrington, C. P., Unkel, S. and Anaya-Izquierdo, K. (2013a). Estimation of basic reproduction numbers: individual heterogeneity and robustness to perturbation of the contact function. *Biostatistics* **14**(3), 528–540.

FARRINGTON, C. P., WHITAKER, H. J., UNKEL, S. AND PEBODY, R. (2013b). Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases. *American Journal of Epidemiology* **177**(5), 474–486.

GOEYVAERTS, N., HENS, N., AERTS, M. AND BEUTELS, P. (2011). Model structure analysis to estimate basic immunological processes and maternal risk for parvovirus B19. *Biostatistics* **12**(2), 283–302.

HENS, N., SHKEDY, Z., AERTS, M., FAES, C., VAN DAMME, P. AND BEUTELS, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*, Volume 63, Statistics for Biology and Health. New York: Springer.

HENS, N., WIENKE, A., AERTS, M. AND MOLENBERGHS, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: an illustration for cross-sectional serological data. *Statistics in Medicine* **27**(14), 2785–2800.

KANAAN, M. N. AND FARRINGTON, C. P. (2005). Matrix models for childhood infections: a Bayesian approach with applications to rubella and mumps. *Epidemiology and Infection* **133**(6), 1009–1021.

LANCASTER, T. (1979). Econometric models for the duration of unemployment. *Econometrica* **47**(4), 939–956.

R CORE TEAM. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

VAUPEL, J., MANTON, K. AND STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**(3), 439–454.

WALLINGA, J., TEUNIS, P. AND KRETZSCHMAR, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology* **164**(10), 936–944.

WIENKE, A. (2010). *Frailty Models in Survival Analysis*. Boca Raton, USA: Chapman & Hall/CRC.

YASHIN, A. I., VAUPEL, J. W. AND IACHINE, I. A. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* **5**(2), 145–159.

[*Received November 06, 2013; revised April 28, 2014; accepted for publication May 27, 2014*]