Time-varying frailty models and the estimation of heterogeneities in transmission of infectious diseases

Steffen Unkel

Justus Liebig University Giessen, Giessen, Germany

C. Paddy Farrington and Heather J. Whitaker

The Open University, Milton Keynes, UK

Richard Pebody

Health Protection Agency, London, UK

Summary. In this paper, a frailty modelling framework is presented for representing and making inference on individual heterogeneities relevant to the transmission of infectious diseases, including heterogeneities that evolve over time. Central to this framework is the use of multivariate data on several infections. We explore new simple but flexible families of time-dependent frailty models, in which the frailty is modulated over time in a deterministic fashion. Methods of estimation, issues of identifiability and model choice are discussed. Results from such models are interpreted in the light of concomitant information on routes of transmission. Applications to paired serological survey data on a range of infections with same and different routes of transmission are presented.

Keywords: Current status data; Frailty; Heterogeneity; Infectious Diseases; Serological survey; Time-varying frailty models; Transmission routes

1. Setting the scene

It is well established that individuals in a population show variation with respect to properties that are relevant to the transmission of infectious diseases (Diekmann and Hesterbeek, 2001). There is ample empirical evidence of variation between individuals, also known as individual heterogeneity, for most measurable attributes such as age, gender, sexual activity, family size, geographical location and genetic characteristics (Anderson and May, 1991). For example, heterogeneity in sexual activity levels are likely to be important in the spread of sexually transmitted infections; personal hygiene is likely to be relevant in the spread of infections transmitted by the faecal-oral route; the propensity of individuals to associate together may be a relevant factor in the spread of infections transmitted by airborne droplets or close contact.

The presence of heterogeneity, and its degree, can have a substantial bearing on the transmission of infection, inflating contact rates within certain subgroups and thus increasing reproduction numbers, with consequences on the likely impact of mass vaccination programmes and other control measures. However, while the importance of heterogeneity in

Address for correspondence: Steffen Unkel, Medical Statistics Group, Institute of Medical Informatics, Justus Liebig University Giessen, Heinrich-Buff-Ring 44, 35392 Giessen, Germany.
E-mail: Steffen.Unkel@informatik.med.uni-giessen.de

the transmission of infection is well understood, and its presence in most cases is well accepted, this recognition has not been translated adequately into infectious disease models. Commonly, such models ignore completely any relevant heterogeneities, or explicitly take into account a small number of key and measurable sources of heterogeneity - such as age effects for childhood infections, or variation in sexual practices for sexually transmitted infections (Anderson and May, 1991; Vynnycky and White, 2010). Models that seek to account for more than a single source of heterogeneity are uncommon (Farrington and Whitaker, 2005).

This gap between the mathematical theory of infectious disease transmission and the statistical models actually used is due to three factors: (i) data on known sources of heterogeneity are usually limited; (ii) other potential sources of heterogeneity can only be guessed at and discussed qualitatively rather than integrated into quantitative modelling; and (iii) there may be further, unsuspected sources of heterogeneity, which by their very nature remain imponderable.

The data on sources of heterogeneities are difficult to obtain for two main reasons. First, because it is not enough to collect data on individuals who are infected, it is also necessary to collect data on the individuals who infected them, and indeed on the individuals with whom they make contact. Such data are not easy to obtain. This difficulty is compounded by the second problem, which is that it is often not possible to define what constitutes a contact. A contact is commonly understood as an event during which transmission of infection between two individuals *could* occur (Diekmann and Hesterbeek, 2001), if one was infected and the other susceptible. However, only for infections for which contact events are well-defined, such as 'having sexual intercourse' for sexually transmitted infections, can we expect to make successful attempts at quantifying the extent of relevant heterogeneities. Unfortunately, for most types of infection including waterborne infections, foodborne infections, community-acquired respiratory infections or any infection only requiring spatio-temporal proximity for transmission, there is no event that can be clearly or uniquely defined as a contact. For these infections it is usually necessary to define a contact by some proxy variable. For example, for infections transmitted via the close-contact route, social contact surveys have been used to quantify rates of proxy contacts to distinguish between close and non-close contacts (Edmunds et al., 2006; Mossong et al., 2008).

In the present paper, in view of the paucity of data on sources of heterogeneities available, we attempt to address the problem of modelling heterogeneities in the transmission of infectious diseases without knowing in advance the source of these heterogeneities, and without a precise definition of what constitutes a contact.

Farrington et al. (2001) suggested to make inferences on heterogeneities indirectly using the fact that they leave an epidemiological footprint or signature through the associations they induce between different infections in the same individuals. This fact can be exploited using multivariate serological survey data. Serological data are a key resource in infectious disease epidemiology and are obtained by testing blood serum residues for the presence of antibodies to one or more infections. A positive (negative) result indicates prior infection (susceptibility to infection), giving rise to current status data (Sun, 2006). Current status data in this context means that if the two survival variables of interest T_1 and T_2 represent the ages at the onset of infection by two distinct infectious agents, then their onset can only be determined to lie below or above the age x when the survey was undertaken. In this context, the time scale is age and the defining time point from which times are measured is birth. The association between two infections, which may or may not share the same route of transmission, can be investigated through the use of bivariate serological survey data on

the same individuals. Such data enable us to observe the effects of heterogeneity without explicitly specifying the mechanisms that give rise to them.

Consider two infections and suppose each individual in the population has some measurable characteristics and unobserved (latent) characteristics. The measured characteristics are covariates relevant to the transmission of infectious diseases, such as age, sexual behaviour, needle sharing practice or duration of drug injection. The unmeasured characteristics may represent behavioural or environmental factors (such as sociability, personal hygiene, family size or nursery attendance) which affect the transmission of both infections, or immunological factors which affect individual susceptibilities to both infections. To facilitate the notation, it is assumed in the sequel that age x is the only measured attribute of an individual and that these unmeasured attributes are described by a random variable $U > 0$ with density $f(U)$ and $\mathrm{E}(U) = 1$. Inspired by the work of Coutinho et al. (1999), Farrington et al. (2001) showed how bivariate serological survey data on two infections could be used to estimate the degree of heterogeneity using shared frailty models (Duchateau and Janssen, 2008; Wienke, 2011) for the force of infection j ($j = 1, 2$):

$$\lambda_j(x, U) = U\, \lambda_{0j}(x) \ ,$$

for an individual of age x and positive random effect (frailty) U, where the baseline forces of infection $\lambda_{0j}(x)$ are independent of U and describe the age effect. The random variation in U induces the association between the failure times T_1 and T_2; T_1 and T_2 are conditionally independent given $U = u$. However, the work of Farrington et al. (2001) does not use the available information on how the strength of association, and hence the degree of heterogeneity, varies with age. Such information is important as it can suggest pointers to the source of the heterogeneity, for example if the association is sustained in adulthood it may reflect a common source of transmission for the two infections.

The frailty modelling framework applied to paired serological survey data previously by Farrington et al. (2001), Farrington and Whitaker (2005) and Hens et al. (2009) has recently been extended to incorporate age-dependence in the heterogeneity (Farrington et al., 2012). This leads to a shared frailty model of the form

$$\lambda_j(x, U(x)) = U(x)\, \lambda_{0j}(x) \ , \tag{1}$$

for $j = 1, 2$ with shared frailty $U(x)$, which may vary with age. We argue that such an approach provides valuable insights into the presence, strength, source and variation with age of heterogeneities that are relevant to the transmission of infectious diseases.

Frailty models are notoriously difficult to identify, a problem compounded by the introduction of age dependence. An important question is whether paired current status data contain enough information to distinguish information between competing plausible models for the time-varying frailty. Farrington et al. (2012) proposed simple families of time-varying frailty models, but did not investigate these comprehensively. The primary purpose of the present paper is to determine whether these families – which can include models with non-gamma distributions for the frailties at baseline $x = 0$, different correlation structures, and both additive and multiplicative structures – are identifiable in practice from paired serological survey data. We investigate this using data on a range of infections.

The remainder of the paper is organized as follows. In Section 2, recently developed families of time-varying frailty models are presented along with a discussion of identifiability issues for shared frailty models. Methods of estimation are given in Section 3. In Section 4, the methods developed in this paper are applied to paired serological survey data on several

pairs of infection with same and different modes of transmission. Concluding comments are given in Section 5.

2. Modelling time-varying frailties and identifiability issues

Several ways of modelling time-dependent frailties have been proposed, including stochastic processes (Aalen et al., 2008, Chapter 11). For an overview of the literature, the interested reader is referred to Wienke (2011) (pp. 232–234) and the references therein.
Recently, Farrington et al. (2012) proposed simple but flexible families of time-varying frailty models, in which the frailties are modulated over time in a deterministic fashion (Farrington et al., 2012). Suppose that $U(x) = w(x, Z_1, \ldots, Z_q)$ for some known function w which may involve parameters to be estimated, where $Z_1, \ldots, Z_q$ are independent time-invariant frailties of unit mean. With our applications of interest, the interpretation of the Z_j $(j = 1, \ldots, q)$ should be motivated by epidemiological considerations.
Our focus of attention is on the frailty variance, which describes the degree of unmeasured heterogeneity in the population. We require the frailty variance, $\mathrm{var}(U(x))$, to be time-dependent and simultaneously, for identification purposes, the frailty mean, $\mathrm{E}(U(x))$, to be equal to one. For pairs of infections with a similar mode of transmission, we expect $\mathrm{var}(U(x))$ to remain above a nonzero threshold at all ages, representing individual heterogeneity in contact rates for this common transmission route. For pairs of infections with dissimilar modes of transmission, we expect $\mathrm{var}(U(x))$ to tend to zero in adulthood. In childhood, we expect $\mathrm{var}(U(x))$ to be nonzero, as most non-sexual transmission routes are largely confounded owing to the closeness and intensity of contacts between children. The heterogeneity, and hence $\mathrm{var}(U(x))$, is likely to decline with age owing to the homogenising effects of social factors such as school attendance and increasing social distance.
The model setting of Farrington et al. (2012) does include frailty models with different correlation structures such as piecewise frailty models. Paik et al. (1994) and Wintrebert et al. (2004) proposed piecewise frailty models with nested structures. Ignoring the nesting, one could build piecewise-constant frailty models on disjoint age intervals $I_j = (x_{j-1}, x_j]$ for $j = 1, \ldots, q$ with $x_0 = 0$ and $x_q < \infty$ as follows. Let

$$U(x) = \sum_{j=1}^{q} Z_j I_j(x) \ , \tag{2}$$

where $Z_j > 0$ are identically distributed with unit mean and variance σ_j^2 $(j = 1, \ldots, q)$, and $I_j(x) = 1$ if $x \in I_j$ (with $I_j(x) = 0$ otherwise). If $\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, q$, then the variance of $U(x)$ is constant. A declining frailty variance is obtained by assuming e.g. $\sigma_j^2 = \sigma^2 \exp\left\{-([m_j - m_1]/\rho)^k\right\}$, where m_j is the midpoint of I_j and k is some known positive integer, for example $k = 2$.
Piecewise frailties are a natural extension of age-invariant frailties to capture age-varying heterogeneity. However, piecewise frailties assume that the frailty in age group j is independent from the frailty in age group $j + 1$. At the other extreme, Farrington et al. (2012) proposed the following time-varying frailty model in which the frailties across age groups are perfectly correlated:

$$U(x) = 1 + [Z - 1]h(x) \ , \quad 0 \leq h(x) \leq 1 \ , \tag{3}$$

where $Z > 0$ is an age-invariant frailty of unit mean and $h(x)$ is a deterministic function. To model the early childhood decline in heterogeneity one could choose $h(x)$ as

$$h(x) = \exp\left\{-(x/\rho)^k\right\} \quad . \tag{4}$$

Note that $\mathrm{E}(U(x)) = 1$ and $\mathrm{var}(U(x)) = h(x)^2 \times \mathrm{var}(Z)$. Under general conditions, the model (3) with $h(x)$ chosen as in (4) (and $\rho > 0$) predicts that the heterogeneity and hence the strength of association tends to zero as x increases, unless $h(x) = 1$. For infections with distinct transmission routes, associations that tail off to zero can be expected. But if we observe an association greater than zero at large x, more flexible models are needed.
Farrington et al. (2012) generalized the 1-component frailty model (3) to families that contain more than one age-invariant frailty. Consider the additive frailty model

$$U(x) = \sum_{j=1}^{q} Z_j h_j(x) \ , \quad \sum_{j=1}^{q} h_j(x) = 1, \quad 0 \le h_j(x) \le 1 \ , \tag{5}$$

for $j = 1, \ldots, q$. The restrictions on the functions $h_j(x)$ ensure that $U(x)$ is non-negative with unit mean. Consider a pair of infections that are transmitted by the same route. For such a pair of infections one might expect heterogeneity to decline to some positive constant value. One could choose a 2-component frailty model with

$$h_1(x) = \frac{\exp\left\{-(x/\rho)^2\right\}}{1 + \exp\left\{-(x/\rho)^2\right\}} \quad \text{and} \quad h_2(x) = \frac{1}{1 + \exp\left\{-(x/\rho)^2\right\}} \ ,$$

where the first component Z_1 represents transient (not route-specific) declining heterogeneity in childhood and the second component Z_2 represents persistent (route-specific) heterogeneity in adulthood. The variance of $U(x)$ is $\mathrm{var}(U(x)) = h_1(x)^2 \times \mathrm{var}(Z_1) + h_2(x)^2 \times \mathrm{var}(Z_2)$. Note that the additive family (5) does include the piecewise frailty model (2) when $h_j(x) = I_j(x)$.
Unfortunately, the additive model is a little unsatisfactory; the restrictions imposed on the age-dependent trajectories $h_j(x)$ mean that the presence of one type of frailty does impact on the other. An alternative to the additive model (5) is the multiplicative family

$$U(x) = \prod_{j=1}^{q} \left[1 + (Z_j - 1)\, h_j(x)\right] \ , \quad 0 \le h_j(x) \le 1 \ , \tag{6}$$

for $j = 1, \ldots, q$. For example, one could set $q = 2$ and choose $h_1(x)$ as in (4) and $h_2(x) = 1$. This 2-component multiplicative model is perhaps more easily interpretable than its additive counterpart. The variance of $U(x)$ is $\mathrm{var}(U(x)) = h_1(x)^2 \times \mathrm{var}(Z_1) + \mathrm{var}(Z_2) + h_1(x)^2 \times \mathrm{var}(Z_1) \times \mathrm{var}(Z_2)$. It is easy to see that the 1-component frailty model (3) is a special case both of the additive family (5) with $q = 2$ and $\mathrm{var}(Z_2) = 0$ and of the multiplicative family (6) with $q = 1$.
It is well known that time-invariant frailty models are not identifiable from single samples of survival data, unless strong parametric assumptions are imposed on the baseline hazard (Aalen et al., 2008). Essentially, this is because there is no independent information available to separate the effect of heterogeneity in the population from the shape of the baseline hazard. Shared frailty models applied to bivariate survival data do enable time-invariant frailty models to be identified: the independent information required is contained

in the cross-ratio function (Clayton, 1978; Oakes, 1989). Although paired current status data contain much less information than bivariate right-censored data, the pairing still provides the information required to identify the frailty, namely information on the cross-ratio function expressed as (Hougaard, 1984; Farrington et al., 2012):

$$CRF(x,x) = \frac{\mathrm{var}(U|\, T_1 > x, T_2 > x)}{\mathrm{E}(U|\, T_1 > x, T_2 > x)^2} + 1 \ , \tag{7}$$

which measures how the strength of association varies over time in survivors.
However, further elaboration of the models to include time-varying frailties introduces new identifiability problems. For example, a decreasing cross-ratio function could either arise from a time-invariant inverse Gaussian (or other) frailty, or from a time-varying frailty such as that proposed in equation (3).
There are some limits to this non-identifiability: it can be shown that some association patterns cannot arise from a time-invariant frailty (Farrington et al., 2012). Nevertheless, when interpreting patterns of association in terms of time-varying frailties, it is important to remain open to the possibility that a contributing factor might be a selection effect, rather than a variation in the heterogeneity *per se*. For example, it is wise when observing decreases in heterogeneity to fit models derived both from gamma and inverse Gaussian frailties, and possibly others.
These considerations also limit the complexity of the models that can usefully be fitted. From a conceptual point of view each individual's hazard is most satisfactorily conceived of as a stochastic process evolving through time, the mean of which reflects the baseline hazard and the variance of which reflects the heterogeneity. However, with only one observation per person, and a censored one at that, it is somewhat ambitious to expect to identify such a process without strong modelling assumptions. Thus, we take the more empirical approach embodied in equation (5) or (6), which is to model the baseline hazard and the variance directly, but assume perfect correlation at different times within individuals - an admittedly simplistic assumption. This is akin to a model with common slopes and random intercepts. Nevertheless, a test of the perfect correlation assumption is provided by fitting the piecewise-independence model (2).
Similarly, the assumption of a shared frailty between infections is probably also somewhat simplistic. It is likely that some heterogeneity is shared, but supplemented by heterogeneity specific to each infection. Thus, in certain circumstances, a more compelling conceptual model may be that of a correlated, rather than shared, frailty (Hens et al., 2009). However, lack of identifiability of the unshared components of the frailties or how their variance may vary with age has restrained us from taking that approach. Nevertheless, the heterogeneity we model is most likely only part of the heterogeneity.
Further identifiability issues, of a statistical rather than structural nature, stem from non-linearity of the model. Notably, correlations between the estimates for the parameters in (5) or (6) are likely to render their estimation problematic when, as is sometimes the case, there is rather little information in the data on the strength of association between infections.

3. Estimation

Suppose that paired serological data are available on two infections. Let $\pi_{00}(x)$ be the probability that an individual of age x has been infected by neither infection and π_{01} the probability that an individual of age x has been infected by infection 2 but not infection 1,

and similarly define $\pi_{10}(x)$ and $\pi_{11}(x)$. Recall the shared age-dependent frailty model (1) for the force of infection j ($j = 1, 2$). For this model the four probabilities $\pi_{00}(x)$, $\pi_{01}(x)$, $\pi_{10}(x)$ and $\pi_{11}(x)$ at age x are computed as

$$\pi_{00}(x) = \mathrm{E}\left(\exp\left\{-\int_0^x U(x)\left[\lambda_{01}(y) + \lambda_{02}(y)\right]\mathrm{dy}\right\}\right) , \tag{8}$$

$$\pi_{01}(x) = \mathrm{E}\left(\exp\left\{-\int_0^x U(x)\lambda_{01}(y)\,\mathrm{dy}\right\}\right) - \pi_{00}(x) , \tag{9}$$

$$\pi_{10}(x) = \mathrm{E}\left(\exp\left\{-\int_0^x U(x)\lambda_{02}(y)\,\mathrm{dy}\right\}\right) - \pi_{00}(x) \tag{10}$$

and

$$\pi_{11}(x) = 1 - \pi_{01}(x) - \pi_{10}(x) - \pi_{00}(x) . \tag{11}$$

The expectations in the expressions (8)–(10) are computed with respect to the random variables $Z_1, Z_2, \ldots, Z_q$ used to define $U(x)$. For mathematical convenience, it is advantageous to assign frailty distributions to $Z_1, Z_2, \ldots, Z_q$ with a simple Laplace transform. Frailty distributions such as the Gamma or inverse Gaussian can be easily expressed by their Laplace transform. In the supporting information that is associated with this paper and available on-line, the three probabilities (8)–(10) are given for a variety of age-dependent frailty models, which are used in the next Section.

Paired serological survey data on n_x individuals of age x give rise to a multinomial observation $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$, where $n_x = \sum_{i,j=0,1} n_{ijx}$ and n_{00x} is the number of individuals of age x in the sample that are uninfected by either infection, n_{10x} is the number of individuals that are uninfected by infection 2 but have been infected by infection 1, and so on. Given parameterizations of $U(x)$ and λ_{0j} ($j = 1, 2$), the shared frailty model (1) is fitted to these data by maximising a product multinomial likelihood. The multinomial log-likelihood kernel is: $l = \sum_x \sum_{i,j=0,1} n_{ijx} \ln\left(\pi_{ij}(x)\right)$, where ln denotes the natural logarithm. To allow for overdispersion, for example resulting from test variability, we also model the data by means of a compound Dirichlet-multinomial distribution with dispersion parameter ν and $0 < \nu < 1$. This inflates the multinomial component variances by the factor $1 + \nu(n_x - 1)$. If $\nu = 0$, then the distribution reduces to the multinomial. Assuming that the observations at different ages are independent, the log-likelihood kernel is

$$\begin{aligned} l_{DM} = \sum_x \Bigg\{ & \ln\left(\frac{\Gamma(\psi)}{\Gamma(n_x + \psi)}\right) + \ln\left(\frac{\Gamma(n_{00x} + \psi\pi_{00}(x))}{\Gamma(\psi\pi_{00}(x))}\right) \\ & + \ln\left(\frac{\Gamma(n_{01x} + \psi\pi_{01}(x))}{\Gamma(\psi\pi_{01}(x))}\right) + \ln\left(\frac{\Gamma(n_{10x} + \psi\pi_{10}(x))}{\Gamma(\psi\pi_{10}(x))}\right) \\ & + \ln\left(\frac{\Gamma(n_{11x} + \psi\pi_{11}(x))}{\Gamma(\psi\pi_{11}(x))}\right)\Bigg\} , \end{aligned} \tag{12}$$

where $\psi = (1 - \nu)/\nu$, hence $\nu = 1/(1 + \psi)$. Individuals with data on only one infection contribute a reduced (binomial or beta-binomial) log-likelihood kernel based on the appropriate two-way margin.

The fitting procedure for a pre-specified model is as follows. For the current set of parameters, obtain the baseline hazards $\lambda_{0j}(x)$ ($j = 1, 2$), compute the probabilities (8)–(11), then

evaluate the log-likelihood, and iterate until convergence. For the Dirichlet-mutinomial model we evaluate the goodness-of-fit by calculating the (scaled) deviance: maximize the log-likelihood (12), fix ψ at its estimated value and then compute $-2\,(l_{DM} - l_{DMSat})$, where l_{DMSat} is the saturated log-likelihood given by

$$\begin{aligned} l_{DMSat} \;=\; \sum_x \Bigg\{ & \ln\left(\frac{\Gamma(\hat{\psi})}{\Gamma(n_x+\hat{\psi})}\right) + \ln\left(\frac{\Gamma(n_{00x}+s00(x))}{\Gamma(\hat{\psi}s00(x))}\right) \\ & + \ln\left(\frac{\Gamma(n_{01x}+\hat{\psi}s01(x))}{\Gamma(\hat{\psi}s01(x))}\right) + \ln\left(\frac{\Gamma(n_{10x}+\hat{\psi}s10(x))}{\Gamma(\hat{\psi}s10(x))}\right) \\ & + \ln\left(\frac{\Gamma(n_{11x}+\hat{\psi}s11(x))}{\Gamma(\hat{\psi}s11(x))}\right)\Bigg\} \;, \end{aligned} \tag{13}$$

with $sij(x) = n_{ijx}/n_x$ $(i,j=0,1)$, that is, the probabilities $\pi_{ij}(x)$ $(i,j=0,1)$ in (12) are replaced in (13) by their observed counterparts. The baseline hazards $\lambda_{0j}(x)$ $(j=1,2)$ were parameterized as piecewise constant on age classes chosen on epidemiological grounds. We investigated sensitivity to the parameterisation of the baseline hazards. Other choices including continuous parametric baselines resulted in worse fits but did not alter the choice of frailty distributions or frailty models. For the multiplicative family (6), the expressions (8)–(10) cannot be computed in closed form. We used the `integrate` function in the software package R version 2.15.1 (R Development Core Team, 2011) to carry out the necessary numerical integration. The function `nlm` is used to maximize the log-likelihood. Since we take a population focus, we use the standard marginal Akaike information criterion (AIC) as a statistical tool for comparison among models (Vaida and Blanchard, 2005).

4. Applications to bivariate serological survey data

In this Section, the statistical models presented in this paper are applied to paired serological survey data on several pairs of infections with similar or different mode of transmission in Subsection 4.1 and Subsection 4.2, respectively. The data have arisen from three large surveys undertaken in the United Kingdom (Data source: Health Protection Agency). All are nationwide surveys of serum samples taken for diagnostic testing for conditions unconnected with the infections studied here. For each infection, a positive (negative) test result indicates prior infection (susceptibility to infection). Equivocal test results are recoded as being positive indicating prior exposure. Owing to ethical restrictions, the only information on each individual is locality of the testing laboratory, gender, age, and test results. Computations were carried out using the software package R version 2.15.1 (R Development Core Team, 2011). The paired mumps and rubella infection data as well as the computer code used to analyze this data set are available upon request.

4.1. Pairs of infections with similar mode of transmission

Toxoplasma and *Helicobacter pylori*

The serological data are from the year 1996. Whereas the study of *Helicobacter pylori* (hereafter abbreviated as *H. pylori*) was reported by Vyse et al. (2002), a publication on the Toxoplasma data is currently in preparation. For 3,632 individuals of 1-84 years of age bivariate data are available on both infections, for 1,243 (3,515) individuals univariate

data are available on Toxoplasma (*H. pylori*) only. Toxoplasmosis is a protozoan zoonosis and *H. pylori* is a bacterial infection of humans. Both infections are transmitted by oral ingestion of contaminated matter (Heymann, D. L. (ed.), 2008, pp. 250-253, 613-617). Heterogeneity in hygiene is likely to result in association between the two infections. The two infections have been studied together by Unkel and Farrington (2012), who introduced a new association measure relevant for shared frailty models with bivariate current status data, denoted $\phi(x)$, whose properties approximate those of the cross-ratio function (7) (Clayton, 1978; Farrington et al., 2012; Oakes, 1989). The value $\phi(x) = 0$ corresponds to independence; $\phi(x) > 0$ corresponds to positive association, notably that resulting from heterogeneity, and $\phi(x) < 0$ to negative association, as may arise owing to cross-immunity. Figure 1 (i) gives the association between times to infection for Toxoplasma and *H. pylori* infections.

* * * Figure 1 about here * * *

In the plot, the areas of the points are proportional to the precision of the estimates. A LOESS (locally weighted scatterplot smoothing) curve is superimposed on the plot to capture trends with age. Figure 1 (i) shows that there is a strong heterogeneity in childhood and that the heterogeneity is declining with age towards some positive constant in adulthood. The declining association may be due to a selection effect caused by a time-invariant (non-Gamma) frailty model, or to temporal variation of the frailty itself. We fitted various shared frailty models to the Toxoplasma and *H. pylori* infection data. Fitting results for a selection of ten of those fitted models are presented in Table 1.

* * * Table 1 about here * * *

Model 1a and model 2a are time-invariant shared frailty models in which the frailty is Gamma and inverse Gaussian distributed, respectively. Fitting results seem not to give evidence that the declining association is due to a selection effect caused e.g. by a inverse Gaussian time-invariant frailty. It is also worth mentioning that piecewise-constant frailty models with constant or declining variance (model 3a and model 4a) fit the data considerably worse than all other models tried. As such, independent piecewise frailties are not supported by the data. But they are also not supported by epidemiological considerations; the assumption that the frailty in age group j is independent from the frailty in age group $j+1$ is too restrictive. Note that for the piecewise models as well as for some of the additive and multiplicative models we also tried inverse Gaussian frailties, which did not result in improved fits.

The best model (lowest AIC and deviance) is the 2-component multiplicative compound Dirichlet-multinomial model, with the two frailties being independently Gamma distributed. This model choice is also supported by epidemiological considerations. The first frailty Z_1 represents transient heterogeneities in childhood. The second frailty Z_2 most likely represents heterogeneity due to differences in hygiene levels in adulthood.

Figure 1 (i) compares the observed and fitted values of the association measure ϕ for four different models. The time-invariant frailty model 1a predicts a constant association. The piecewise independent Gamma model 4a (with declining variance) does not give a good fit to the observed association pattern at early ages. The 1-component frailty model 5a predicts that the heterogeneity tends to zero. The fitted curve that most closely resembles the observed pattern corresponds to the 2-component multiplicative model 10a.

For model 10a, approximate 95% confidence intervals (CIs) for the parameters of interest, obtained by simulating from a multivariate normal distribution with covariance matrix

set equal to a numerical estimate of the observed Fisher information matrix, keeping the dispersion parameter ν fixed at its estimated value, are as follows: for θ_1: $(.0597, .2416)$; for ρ: $(3.9967, 9.7078)$ and for θ_2: $(.8488, 2.5348)$. The maximum variance inflating factor $1 + \hat{\nu}(n_x - 1)$ over all n_x is 1.04 with an average of 1.02.
The degree of unmeasured individual heterogeneity in the population at age x is represented by the variance of the frailty $U(x)$. Using the estimated parameters for the best fitting model, an estimate of $\sqrt{\text{var}(U(x))}$, along with approximate 95% CIs, is given in Figure 2 (i).

* * * Figure 2 about here * * *

By age 18, the fitted frailty standard deviation has reached the positive value (.8201) at which it remains in adulthood.
The contribution of each observation to the overall goodness-of-fit measured by the deviance can be evaluated by computing pointwise deviances. Supplementary Figure 1 (i) in the supporting information displays absolute pointwise deviances for the 2-component multiplicative multinomial model (model 9a) and its Dirichlet counterpart (model 10a). A diagnostic plot such as Supplementary Figure 1 is useful for identifying ill-fitting data points and for visualizing the effect of allowing for overdispersion. Supplementary Figure 1 (i) reveals one outlier at age 1.

Mumps and Rubella
This survey was undertaken in 1986 (Morgan-Capner et al., 1988). Owing to the selective rubella vaccination programme in adolescent girls, which was in place at the time of the survey, the data only comprise males. As the data are rather sparse at higher ages, only 4116 individuals aged 1-40 are included in the analysis. Marginal data are not included as the corresponding number of cases is negligible. Both mumps and rubella are transmitted by close contact and respiratory droplets. Figure 1 (ii) shows the observed and fitted associations between times to infection for mumps virus and rubella virus. Again there is positive association between times to infection, which is declining with age, but the association is lower than for the previous example. The fitting results are presented in Table 2.
Clearly, the best fit and lowest AIC are obtained by the 2-component Dirichlet-multinomial model 10b. Nearly for all observed 4-tuples the Dirichlet-multinomial leads to a reduction in the pointwise deviance (see Supplementary Figure 1 (ii) in the supporting information). The maximum variance inflating factor is 1.39 with an average of 1.18. That is, the multinomial component variances are increased on average by more than 18%. For model 10b, approximate 95% CIs for the parameters of interest are as follows: for θ_1: $(.0510, .3297)$; for ρ: $(2.4243, 4.1544)$ and for θ_2: $(3.1708, 14.8400)$. An estimate of $\sqrt{\text{var}(U(x))}$, along with approximate 95% CIs, is given in Figure 2 (ii). By age 8, the fitted frailty standard deviation has reached the positive value (.3770) at which it remains in adulthood.

Table 2: Fitting results for mumps and rubella infection data.

	Frailty model	Parameter estimates	deviance	df	p-value	AIC
1b	$U \sim \Gamma(\theta, 1/\theta)$	$\hat{\theta} = 8.0969$	162.91	109	.0006	6444.99
2b	$U \sim InvG(1, \theta)$	$\hat{\theta} = 5.8919$	162.19	109	.0007	6444.27
3b	Piecewise independent Gamma with constant variance	$\hat{\theta} = 3.9742$	162.18	109	.0007	6444.26
4b	Piecewise independent Gamma with declining variance	$\hat{\theta} = 3.9743$ $\hat{\rho} = 9904.25$	162.18	108	.0005	6446.26
5b	1-component Gamma	$\hat{\theta} = 0.0806$ $\hat{\rho} = 2.5332$	134.30	108	.0440	6418.38
9b	2-component multiplicative double Gamma	$\hat{\theta}_1 = 0.0682$ $\hat{\rho} = 2.3599$ $\hat{\theta}_2 = 14.6230$	130.68	107	.0597	6416.75
10b	2-component multiplicative double Gamma (Dirichlet multinomial)	$\hat{\theta}_1 = 0.1297$ $\hat{\rho} = 3.3027$ $\hat{\theta}_2 = 7.0297$ $\hat{\nu} = 0.0018$	108.80	106	.4067	6396.87

4.2. Pairs of infections with different mode of transmission

Parvovirus B19 and Cytomegalovirus

This survey was undertaken in 1991. The study of Parvovirus B19 (hereafter referred to as B19) was reported by Gay et al. (1994) and the study of Cytomegalovirus (hereafter abbreviated as CMV) by Vyse et al. (2009). As the data are rather sparse at higher ages, only individuals with age 1-44 are retained for the analysis. For 1268 cases antibody data on both infections are available. For 3839 (757) individuals data included information on the seroprevalence on B19 (CMV) only.

B19 causes erythema infectiosum, commonly known as slapped cheek syndrome or fifth disease. It is clinically similar to rubella. CMV is a member of the herpes virus family and is able to establish latent infection in the host following primary infection, from which it can periodically reactivate. Infection is common and usually asymptomatic. B19 is transmitted by the respiratory route, whereas CMV is transmitted by mucosal contact with any bodily fluid. In childhood, different transmission routes are confounded to some degree, so associations are to be expected at younger ages owing to heterogeneity of individual social contact intensities.

Figure 3 (i) shows the observed association between times to infection for B19 and CMV, along with some fitted association curves.

* * * Figure 3 about here * * *

The fitted association patterns for the 1-component model (5c) and the 2-component model (9c) are virtually identical. The corresponding fitting results are given in Table 3.

The best model (lowest AIC) is the 1-component Gamma model (5c). This is not surprising as the heterogeneity tails off to zero in adulthood, hence a second component that specifies a route-specific persistent heterogeneity in adulthood is not needed. As it can seen from

Table 3, the frailty variance for the second component in model 9c is almost zero. Including an overdispersion parameter in the 1-component model does not improve the model fit for this pair of infections, see also Supplementary Figure 1 (iii) in the supporting information. The 95% CIs for the frailty parameters of model 5c are as follows: for θ: $(.0747, .2542)$ and for ρ: $(3.4651, 4.9134)$. An estimate of $\sqrt{\text{var}(U(x))}$, along with approximate 95% CIs, is given in Figure 2 (iii). By age 10, the fitted frailty standard deviation has virtually reached zero level.

Table 3: Fitting results for B19 and CMV infection data.

	Frailty model	Parameter estimates	deviance	df	p-value	AIC
1c	$U \sim \Gamma(\theta, 1/\theta)$	$\hat{\theta} = 6.8852$	245.11	209	.0440	8739.49
2c	$U \sim InvG(1, \theta)$	$\hat{\theta} = 5.4696$	244.70	209	.0457	8739.09
5c	1-component Gamma	$\hat{\theta} = 0.1434$ $\hat{\rho} = 4.1778$	230.54	208	.1357	8726.93
9c	2-component multiplicative double Gamma	$\hat{\theta}_1 = 0.1436$ $\hat{\rho} = 4.1730$ $\hat{\theta}_2 = \infty$	230.56	207	.1253	8728.95
10c	1-component Gamma (Dirichlet multinomial)	$\hat{\theta} = 0.1434$ $\hat{\rho} = 4.1778$ $\hat{\nu} < 0.0001$	230.54	207	.1254	8728.93

B19 and *H. pylori*

The B19 and *H. pylori* infection data were obtained in the survey that was undertaken in 1996, providing information on seroprevalence on both infections for 1829 individuals of 1-79 years of age. For 1006 (5125) individuals marginal data are available on B19 (*H. pylori*) only. As for this infection pair the main routes of transmission are distinct (respiratory/fecal oral), we could again expect an existing association in childhood to tail off to zero in adulthood. Figure 3 (ii) shows the observed and fitted association patterns. The fitted curve that most closely resembles the observed pattern corresponds to the 1-component Dirichlet-multinomial (model 10d). This is also the best fitting model in terms of the AIC, see Table 4.

Table 4: Fitting results for B19 and *H. Pylori* infection data.

	Frailty model	Parameter estimates	deviance	df	p-value	AIC
1d	$U \sim \Gamma(\theta, 1/\theta)$	$\hat{\theta} = 3.6632$	395.05	346	.0353	8445.81
5d	1-component Gamma	$\hat{\theta} = 0.1070$ $\hat{\rho} = 9.6043$	385.60	345	.0650	8438.36
9d	2-component multiplicative double Gamma	$\hat{\theta}_1 = 0.1176$ $\hat{\rho} = 9.0066$ $\hat{\theta}_2 = 9.5051$	385.09	344	.0630	8439.85
10d	1-component Gamma (Dirichlet multinomial)	$\hat{\theta} = 0.1118$ $\hat{\rho} = 9.5399$ $\hat{\nu} = 0.001$	375.45	344	.1171	8430.21

Fitting the Dirichlet-multinomial leads to inflating the multinomial component variances on

average by about 2% (see also Supplementary Figure 1 (iv) in the supporting information). The 95% CIs for the frailty parameters of model 10d are as follows: for θ: $(.0587, .2032)$ and for ρ: $(6.2095, 12.8092)$. An estimate of $\sqrt{\text{var}(U(x))}$, along with approximate 95% CIs, is given in Figure 2 (iv). Compared to the previous example (Figure 2 (iii)), the fitted frailty standard deviation has tailed off to zero later in adulthood.

5. Discussion

We presented a frailty modelling framework for representing and making inference on individual heterogeneities that are relevant to the transmission of infectious diseases. This framework is able to take possible time-related variation in heterogeneity into account. To incorporate time-dependent heterogeneities via frailty models, we explored new families of models in which the frailty is modulated over time in a deterministic fashion. The new models overcome the disadvantages of independent piecewise frailties, the latter being neither supported by the data nor by epidemiological considerations.

Central to our approach of quantifying relevant heterogeneities and how they evolve over time is the use of paired serological survey data on different infections for the same individuals. For the infection pairs we studied, we found strong heterogeneity (association) in childhood (for mumps and rubella, the association is strong only in early childhood), which is declining with age to a positive constant (for pairs of infections with same mode of transmission) or to zero (for pairs of infections with distinct mode of transmission).

Our model choice is guided by both statistical and epidemiological considerations. For infection pairs with common mode of transmission we advocate 2-component models in which the first (second) component represents transient (persistent) heterogeneities in childhood (adulthood). For infection pairs with different transmission mode a 1-component model to represent the early childhood decline in heterogeneity that tails off to zero seems satisfactory.

Frailty modelling, *a fortiori* time-dependent frailty modelling, is fraught with lack of identifiability, though it is interesting to note that the data contain sufficient information to rule out the piecewise independent Gamma models. It is useful to consider what additional sources of data might help in that respect. One possibility might be to consider multivariate data on more than two infections in the same individuals; such data would need to be plentiful to avoid extreme sparsity in the cross-classification. Higher-variate data would also enable one to make inferences about routes of transmission, when these are unknown or uncertain, exploiting the information on the associations with other infections (Farrington et al., 2013). Another possibility might be to supplement information on immunoglobulin G (IgG) antibody prevalence with information on other antibodies, notably IgM which provide information on recent infection. A third possibility would be to use serological survey data collected at several time points to identify age and temporal effects, although only data from single surveys are usually available. A fourth desirable source of enhanced information would be to include other measurable characteristics (apart from age) in the analysis, e.g. socio-economic variables.

Further work, some of it under way, is required in several areas, some of which have already been mentioned. Firstly, the parametric frailty models used in this paper need to be revisited with a view to fitting non-parametric functions for the baseline forces of infection $\lambda_{0j}(x)$ $(j = 1, 2)$ and the age-dependent trajectories $h_j(x)$ $(j = 1, \ldots, q)$. In the present paper, we restricted ourselves to the use of the Gamma and inverse Gaussian distribution to

model the frailties. In different settings other frailty distributions might be more applicable. For example, for sexually transmitted diseases, in which heterogeneity could be represented by number of sexual partners, discrete frailty distributions (such as the Poisson) may be appropriate.

Secondly, frailties are specific to an individual and for some infections (e.g. sexually transmitted diseases) individual behaviour is of utmost importance. But, for example, for respiratory infections in the home, the infection probability increases with household size and we could expect associations due to variation in household size. Likewise, other small scale social structure may be important in the transmission of some infections. Unfortunately, data sufficiently detailed to allow studying such effects directly are very seldom collected, and in any case are unlikely to exhaust the myriad sources of heterogeneity which are likely to play a role in transmission. The heterogeneity that can be quantified using the approach of the present paper may in some sense be regarded as averaging over such effects.

Thirdly, individual heterogeneity may have the effect of increasing the estimates of the following two key epidemiological parameters: the basic reproduction number R_0 and the critical immunization threshold $\pi_c = 1 - R_0^{-1}$. For example, if there is heterogeneity due to differences in the propensity to make contacts, then R_0 is inflated by a factor involving $1 + \mathrm{var}(U(x))$. The impact of individual heterogeneity on R_0 and π_c for various infections has been investigated by Farrington et al. (2013). Finally, we only considered infections with lifelong immunity, that is, susceptible-infectious-recovered (SIR) infections (Vynnycky and White, 2010). It would be of interest to apply the methodology presented in this paper to susceptible-infectious-susceptible (SIS) infections which confer no immunity or to susceptible-infectious-recovered-susceptible infections conferring temporary immunity.

Supporting Information

Additional supporting information may be found in: 'Supporting information for "Time-varying frailty models and the estimation of heterogeneities in transmission of infectious diseases"'.

Acknowledgments

The Associate Editor and two reviewers made valuable comments and suggestions on the first draft of this paper. We would like to acknowledge Dr Louise Hesketh (Public Health Laboratory Service, Preston, UK) who performed much of the serology. This research was supported by a grant from the UK Medical Research Council, and by a Royal Society Wolfson Research Merit Award to C. Paddy Farrington.

References

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

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

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

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

Diekmann, O. and J. A. P. Hesterbeek (2001). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley: Chichester.

Duchateau, L. and P. Janssen (2008). *The Frailty Model*. Springer: New York.

Edmunds, W. J., G. Kafatos, J. Wallinga, and J. R. Mossong (2006). Mixing patterns and the spread of close-contact infectious diseases. *Emerging Themes in Epidemiology 3:10*.

Farrington, C. P., M. Kanaan, and N. J. Gay (2001). Estimation of the basic reproduction number for infectious diseases from age stratified serological survey data (with discussion). *Journal of the Royal Statistical Society Series C (Applied Statistics) 50*, 251–292.

Farrington, C. P., S. Unkel, and K. Anaya-Izquierdo (2012). The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society Series B 74*, 673–696.

Farrington, C. P. and H. J. Whitaker (2005). Contact surface models for infectious diseases: estimation from serologic survey data. *Journal of the American Statistical Association 100*, 370–379.

Farrington, C. P., H. J. Whitaker, S. Unkel, and R. Pebody (2013). Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases. *American Journal of Epidemiology 177*, in press.

Gay, N. J., L. M. Hesketh, B. J. Cohen, M. Rush, C. Bates, P. Morgan-Capner, and E. Miller (1994). Age-specific antibody prevalence to parvovirus b19: how many women are infected in pregnancy. *Communicable Disease Report Review 4* (Review number 9), R104–R107.

Hens, N., A. Wienke, M. Aerts, and G. Molenberghs (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological survey data. *Statistics in Medicine 28*, 2785–2800.

Heymann, D. L. (ed.) (2008). *Control of Communicable Diseases Manual* (19th ed.). American Public Health Association: Washington D.C.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika 71*, 75–83.

Morgan-Capner, P., J. Wright, C. L. Miller, and E. Miller (1988). Surveillance of antibody to measles, mumps, and rubella by age. *British Medical Journal 297*, 770–772.

Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. Scalia Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinka, and W. J. Edmunds (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine 5(3)*, e74.

Oakes, D. (1989). Bivariate survival models induced by frailities. *Journal of the American Statistical Association 84*, 487–493.

Paik, M. C., W.-Y. Tsai, and R. Ottman (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics 50*, 975–988.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Springer: New York.

Unkel, S. and C. P. Farrington (2012). A new measure of time-varying association for shared frailty models with bivariate current status data. *Biostatistics 13*, in press.

Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-effects models. *Biometrika 92*, 351–370.

Vynnycky, E. and R. G. White (2010). *An Introduction to Infectious Disease Modelling*. Oxford University Press: Oxford.

Vyse, A. J., N. J. Gay, L. M. Hesketh, N. J. Andrews, B. Marshall, H. I. J. Thomas, P. Morgan-Capner, and E. Miller (2002). The burden of *Helicobacter pylori* infection in england and wales. *Epidemiology and Infection 128*, 411–417.

Vyse, A. J., L. M. Hesketh, and R. G. Pebody (2009). The burden of infection with cytomegalovirus in england and wales: how many women are infected in pregnancy? *Epidemiology and Infection 137*, 526–533.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC Press: Boca Raton, FL.

Wintrebert, C. M. A., H. Putter, A. H. Zwinderman, and J. C. van Houwelingen (2004). Centre-effect on survival after bone marrow transplantation: application of time-dependent frailty models. *Biometrical Journal 46*, 512–25.

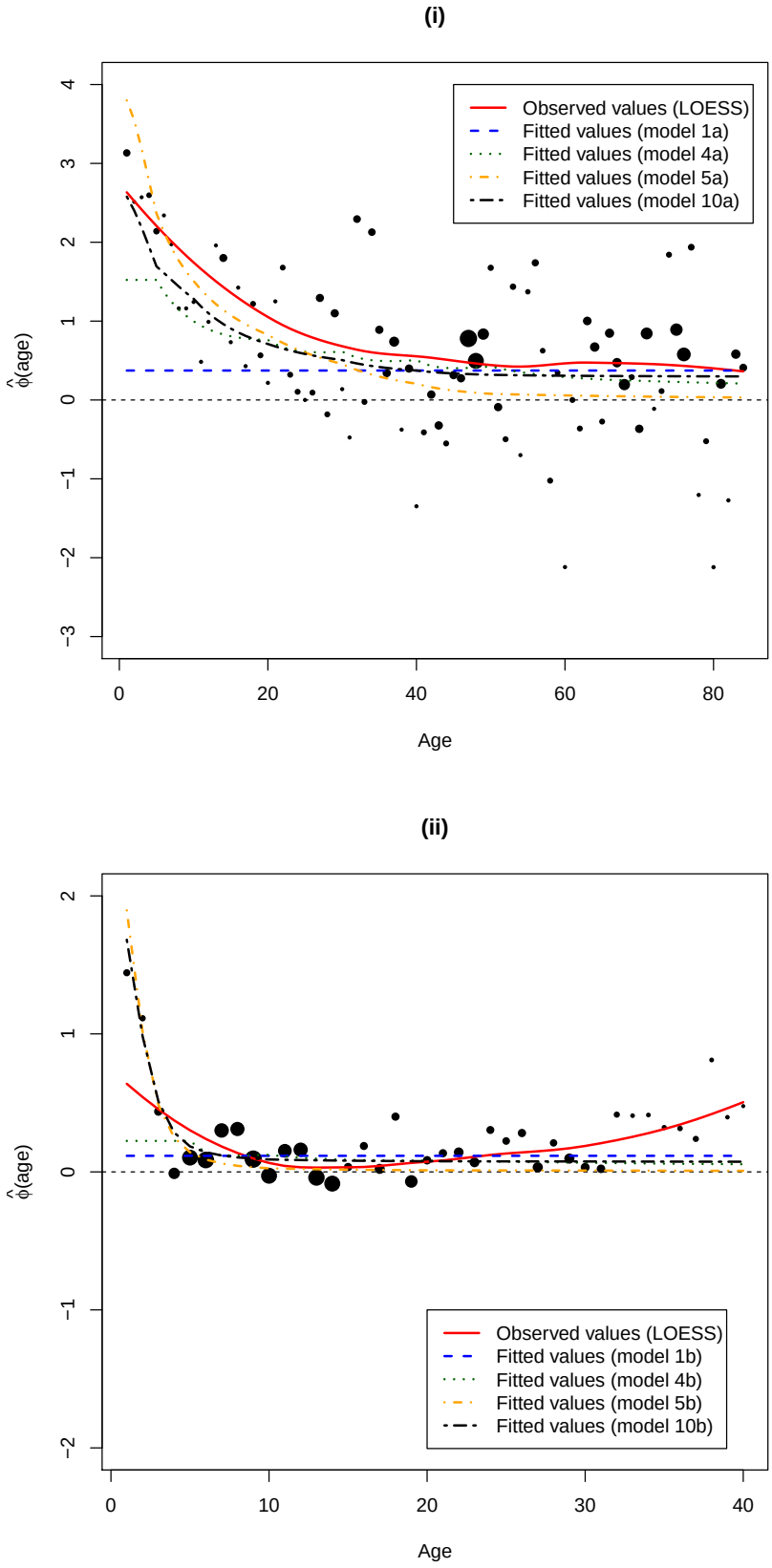


Fig. 1: Association between pairs of infections with similar transmission route: observed and fitted values of ϕ by age for (i) Toxoplasma and *H. pylori* and (ii) mumps and rubella infection data (horizontal dashed line at zero: no association).

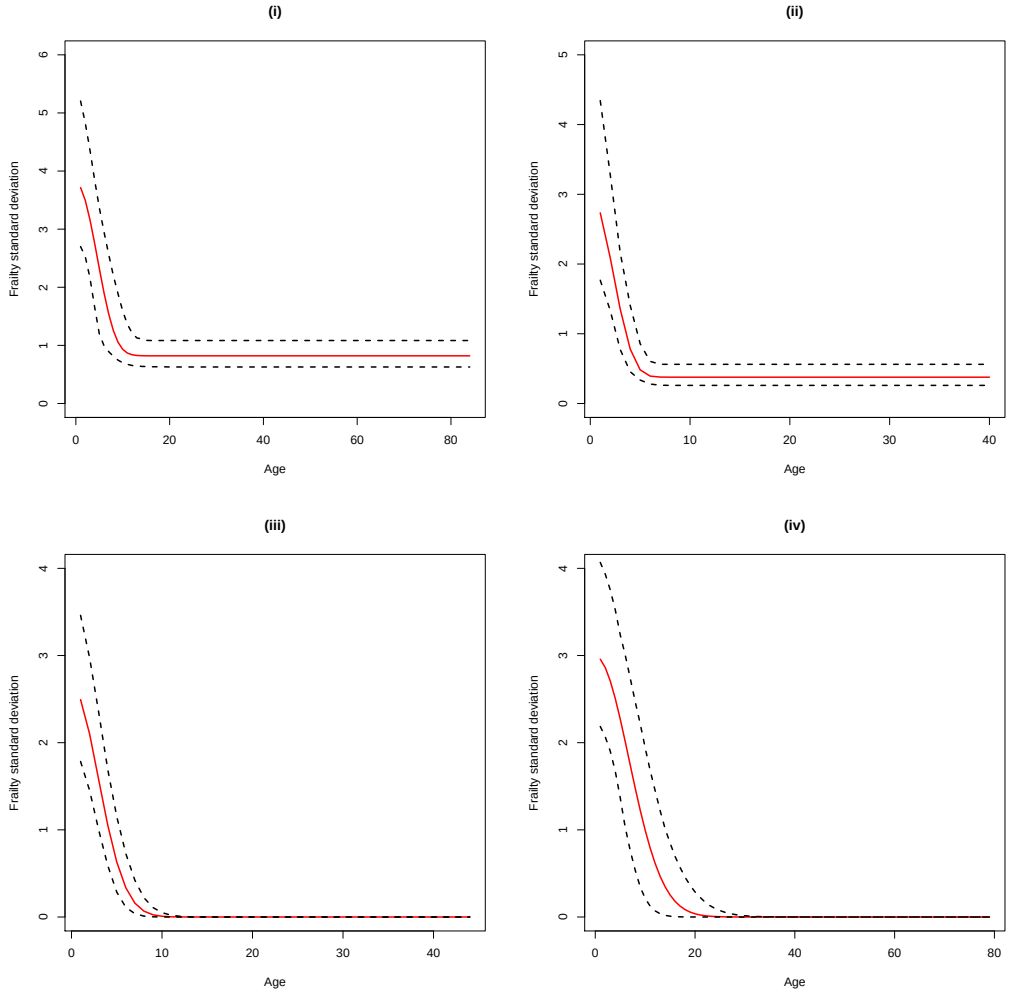


Fig. 2: Standard deviation of the frailty (solid line) with 95% CIs (dashed lines) obtained from fitting the 2-component multiplicative compound Dirichlet-multinomial model to (i) Toxoplasma and *H. pylori* and (ii) mumps and rubella infection data, and the 1-component Dirichlet multinomial model to (iii) B19 and CMV and (iv) B19 and *H. pylori* infection data.

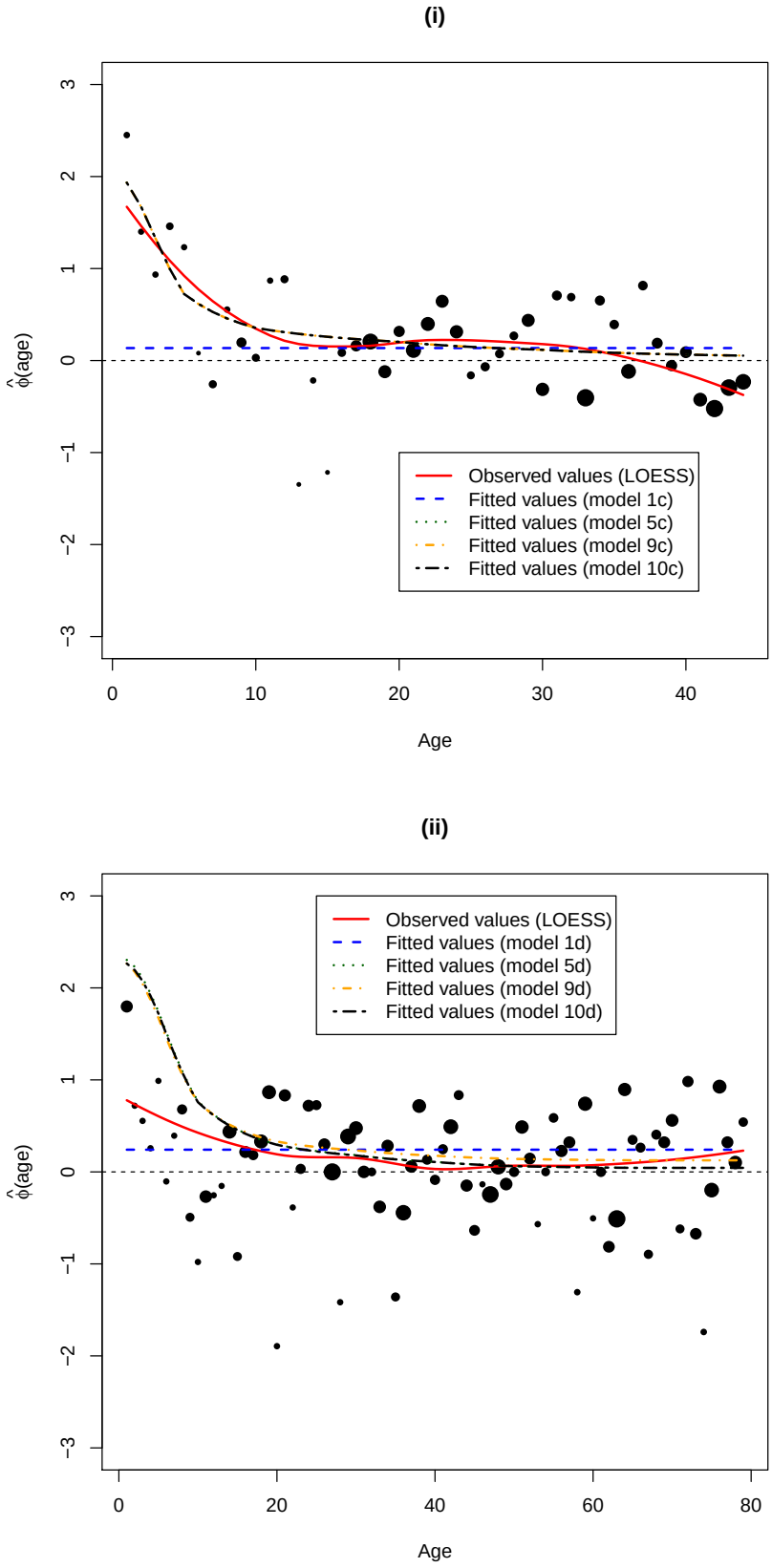


Fig. 3: Association between pairs of infections with different transmission route: observed and fitted values of ϕ by age for (i) B19 and CMV and (ii) B19 and *H. Pylori* infection data (horizontal dashed line at zero: no association).

Table 1: Fitting results for Toxoplasma and *H. pylori* infection data. For the piecewise-constant baseline hazards eight age classes are chosen. In models 3a and 4a, m_j is the midpoint of the jth interval $I_j = (x_{j-1}, x_j]$ $(j = 1, \ldots, 8)$.

	Frailty model	Parameterisation of $h(x)$, $h_1(x)$ and $h_2(x)$	Parameter estimates	deviance	df	p-value	AIC
1a	$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 2.2091$	412.06	366	.0484	8504.62
2a	$U \sim InvG(1, \theta)$		$\hat{\theta} = 1.8794$	411.64	366	.0499	8504.20
3a	Piecewise independent Gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, \ldots, 8)$	$\hat{\theta} = 0.3549$	438.06	366	.0057	8530.62
4a	Piecewise independent Gamma	$\sigma_j^2 = \sigma^2 \exp(-([m_j - m_1]/\rho)^2)$ $(j = 1, \ldots, 8)$	$\hat{\theta} = 0.2780$ $\hat{\rho} = 55.8701$	437.77	365	.0053	8532.33
5a	1-component Gamma	$h(x) = \exp(-(x/\rho)^2)$	$\hat{\theta} = 0.0202$ $\hat{\rho} = 4.8897$	404.58	365	.0756	8499.14
6a	1-component inverse Gaussian	$h(x) = \exp(-(x/\rho)^2)$	$\hat{\theta} = 0.0129$ $\hat{\rho} = 5.6673$	404.41	365	.0758	8498.98
7a	2-component additive double Gamma	$h_1(x) = \frac{\exp(-(x/\rho)^2)}{1+\exp(-(x/\rho)^2)}$ $h_2(x) = \frac{1}{1+\exp(-(x/\rho)^2)}$	$\hat{\theta}_1 = 0.0014$ $\hat{\rho} = 12.9457$ $\hat{\theta}_2 = 3.2327$	397.73	364	.1079	8494.29
8a	2-component additive double Gamma (Dirichlet multinomial)	$h_1(x) = \frac{\exp(-(x/\rho)^2)}{1+\exp(-(x/\rho)^2)}$ $h_2(x) = \frac{1}{1+\exp(-(x/\rho)^2)}$	$\hat{\theta}_1 = 0.0014$ $\hat{\rho} = 12.9498$ $\hat{\theta}_2 = 3.2326$ $\hat{\nu} = 0.00002$	397.52	363	.1025	8496.09
9a	2-component multiplicative double Gamma	$h_1(x) = \exp(-(x/\rho)^2)$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.0272$ $\hat{\rho} = 3.9897$ $\hat{\theta}_2 = 3.2476$	399.53	364	.0967	8496.09
10a	2-component multiplicative double Gamma (Dirichlet multinomial)	$h_1(x) = \exp(-(x/\rho)^2)$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.1220$ $\hat{\rho} = 6.8521$ $\hat{\theta}_2 = 1.4868$ $\hat{\nu} = 0.0005$	383.81	363	.2169	8482.37

Supporting information for "Time-varying frailty models and the estimation of heterogeneities in transmission of infectious diseases"

Steffen Unkel[1], C. Paddy Farrington[2],

Heather J. Whitaker[2] and Richard Pebody[3]

[1]Medical Statistics Group, Institute of Medical Informatics
Justus Liebig University Giessen, Germany

[2]Department of Mathematics and Statistics
The Open University, Milton Keynes, United Kingdom

[3]Health Protection Agency – Colindale, London, United Kingdom

October 4, 2012

1-component age-dependent frailty model (3)

$Z \sim \Gamma(\theta, 1/\theta)$

The three probabilities (8)–(10) are

$$\begin{aligned}
\pi_{00}(x) &= \exp\left\{H^1(x) + H^2(x) - \Lambda_{01}(x) - \Lambda_{02}(x)\right\}\left[1 + \frac{H^1(x) + H^2(x)}{\theta}\right]^{-\theta} , \\
\pi_{01}(x) &= \exp\left\{H^1(x) - \Lambda_{01}(x)\right\}\left[1 + \frac{H^1(x)}{\theta}\right]^{-\theta} - \pi_{00}(x) , \\
\pi_{10}(x) &= \exp\left\{H^2(x) - \Lambda_{02}(x)\right\}\left[1 + \frac{H^2(x)}{\theta}\right]^{-\theta} - \pi_{00}(x) ,
\end{aligned}$$

where $H^j(x) = \int_0^x h(t)\lambda_{0j}(t)\,\mathrm{dt}$ $(j = 1, 2)$.

$Z \sim InvG(1, \theta)$

The three probabilities (8)–(10) are

$$\begin{aligned}
\pi_{00}(x) &= \exp\left\{H^1(x) + H^2(x) - \Lambda_{01}(x) - \Lambda_{02}(x)\right\} \\
&\quad \times \exp\left\{\theta - \left[\theta^2 + 2\theta(H^1(x) + H^2(x))\right]^{1/2}\right\} \ , \\
\pi_{01}(x) &= \exp\left\{H^1(x) - \Lambda_{01}(x)\right\} \exp\left\{\theta - \left[\theta^2 + 2\theta H^1(x)\right]^{1/2}\right\} - \pi_{00}(x) \ , \\
\pi_{10}(x) &= \exp\left\{H^2(x) - \Lambda_{02}(x)\right\} \exp\left\{\theta - \left[\theta^2 + 2\theta H^2(x)\right]^{1/2}\right\} - \pi_{00}(x) \ .
\end{aligned}$$

Additive family (5) with $q = 2$

$Z_1 \sim \Gamma(\theta_1, 1/\theta_1)$ and $Z_2 \sim \Gamma(\theta_2, 1/\theta_2)$

The three probabilities (8)–(10) are

$$\begin{aligned}
\pi_{00}(x) &= \exp\left\{H_1^1(x) + H_2^1(x) + H_1^2(x) + H_2^2(x) - \Lambda_{01}(x) - \Lambda_{02}(x)\right\} \\
&\quad \times \left[1 + \frac{H_1^1(x) + H_1^2(x)}{\theta_1}\right]^{-\theta_1} \left[1 + \frac{H_2^1(x) + H_2^2(x)}{\theta_2}\right]^{-\theta_2} \ , \\
\pi_{01}(x) &= \exp\left\{H_1^1(x) + H_2^1(x) - \Lambda_{01}(x)\right\} \left[1 + \frac{H_1^1(x)}{\theta_1}\right]^{-\theta_1} \left[1 + \frac{H_2^1(x)}{\theta_2}\right]^{-\theta_2} - \pi_{00}(x) \ , \\
\pi_{10}(x) &= \exp\left\{H_1^2(x) + H_2^2(x) - \Lambda_{02}(x)\right\} \left[1 + \frac{H_1^2(x)}{\theta_1}\right]^{-\theta_1} \left[1 + \frac{H_2^2(x)}{\theta_2}\right]^{-\theta_2} - \pi_{00}(x) \ ,
\end{aligned}$$

where $H_i^j(x) = \int_0^x h_i(t)\lambda_{0j}(t)\,\mathrm{dt}$ $(i, j = 1, 2)$.

$Z_1 \sim InvG(1,\theta_1)$ **and** $Z_2 \sim InvG(1,\theta_2)$

The three probabilities (8)–(10) are

$$\begin{aligned}
\pi_{00}(x) &= \exp\left\{H_1^1(x)+H_2^1(x)+H_1^2(x)+H_2^2(x)-\Lambda_{01}(x)-\Lambda_{02}(x)\right\} \\
&\quad \times \exp\left\{\theta_1-\left[\theta_1^2+2\theta_1(H_1^1(x)+H_1^2(x))\right]^{1/2}\right\} \\
&\quad \times \exp\left\{\theta_2-\left[\theta_2^2+2\theta_2(H_2^1(x)+H_2^2(x))\right]^{1/2}\right\} \ , \\
\pi_{01}(x) &= \exp\left\{H_1^1(x)+H_2^1(x)-\Lambda_{01}(x)\right\} \\
&\quad \times \exp\left\{\theta_1-\left[\theta_1^2+2\theta_1 H_1^1(x)\right]^{1/2}\right\} \exp\left\{\theta_2-\left[\theta_2^2+2\theta_2 H_2^1(x)\right]^{1/2}\right\}-\pi_{00}(x) \ , \\
\pi_{10}(x) &= \exp\left\{H_1^2(x)+H_2^2(x)-\Lambda_{02}(x)\right\} \\
&\quad \times \exp\left\{\theta_1-\left[\theta_1^2+2\theta_1 H_1^2(x)\right]^{1/2}\right\} \exp\left\{\theta_2-\left[\theta_2^2+2\theta_2 H_2^2(x)\right]^{1/2}\right\}-\pi_{00}(x) \ .
\end{aligned}$$

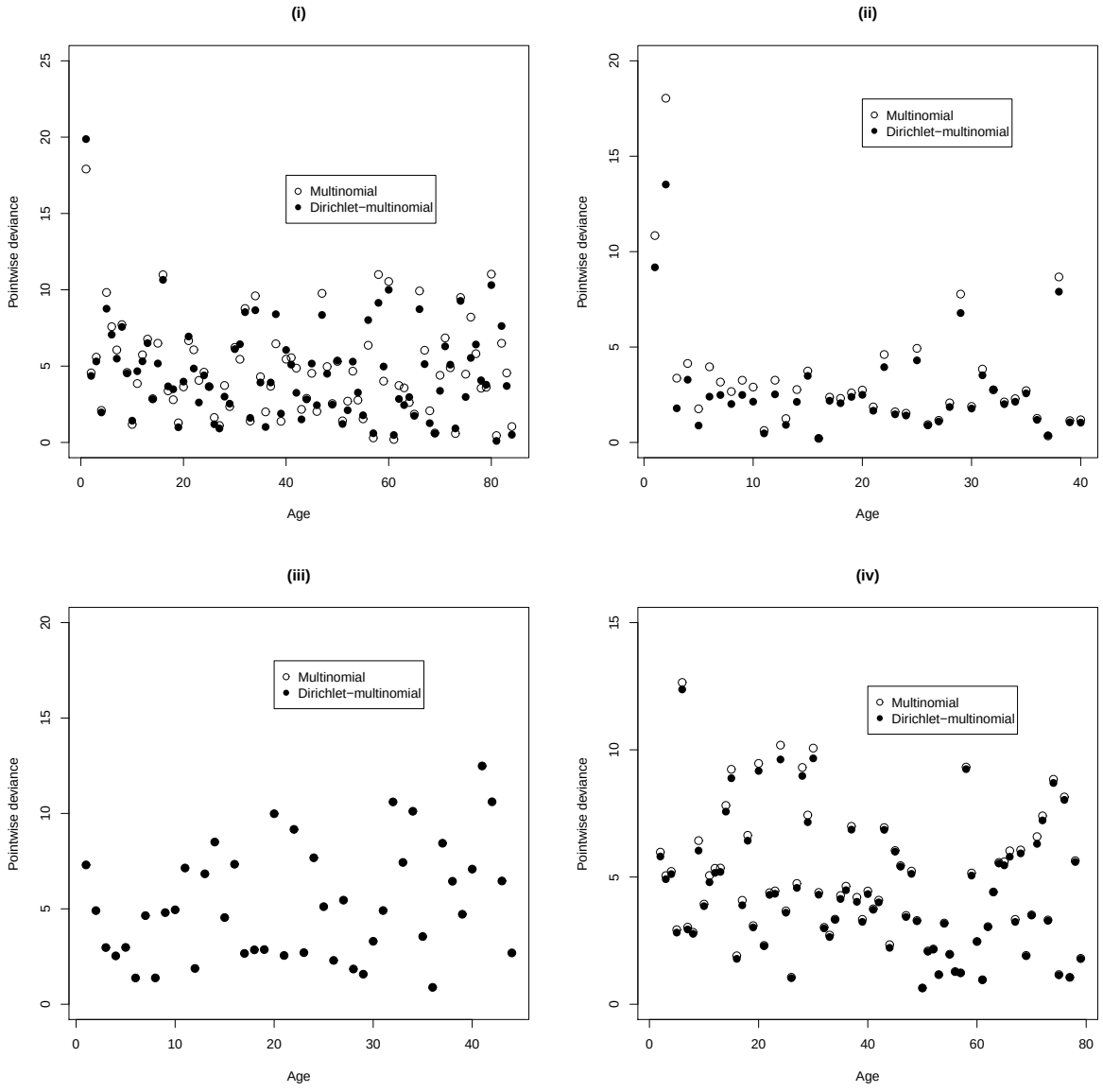
Multiplicative family (6) with $q = 2$

The three probabilities (8)–(10) are

$$\begin{aligned}
\pi_{00}(x) &= \exp\left\{H_1^1(x)+H_2^1(x)+H_1^2(x)+H_2^2(x)-H_{12}^1(x)-H_{12}^2(x)-\Lambda_{01}(x)-\Lambda_{02}(x)\right\} \\
&\quad \times K_{12}(x) \ , \\
\pi_{01}(x) &= \exp\left\{H_1^1(x)+H_2^1(x)-H_{12}^1(x)-\Lambda_{01}(x)\right\} \times K_1(x)-\pi_{00}(x) \ , \\
\pi_{10}(x) &= \exp\left\{H_1^2(x)+H_2^2(x)-H_{12}^2(x)-\Lambda_{02}(x)\right\} \times K_2(x)-\pi_{00}(x) \ ,
\end{aligned}$$

where $H_{ij}^k(x) = \int_0^x h_i(t)h_j(t)\lambda_{0k}(t)\,\mathrm{dt}$ $(i,j,k=1,2)$ and

$$\begin{aligned}
K_{12}(x) &= \mathrm{E}\left(\exp\left\{-U_1\left[H_1^1(x)+H_1^2(x)-H_{12}^1(x)-H_{12}^2(x)\right]\right.\right. \\
&\qquad -U_2\left[H_2^1(x)+H_2^2(x)-H_{12}^1(x)-H_{12}^2(x)\right] \\
&\qquad \left.\left.-U_1U_2\left[H_{12}^1(x)+H_{12}^2(x)\right]\right\}\right) \ , \\
K_1(x) &= \mathrm{E}\left(\exp\left\{-U_1\left[H_1^1(x)-H_{12}^1(x)\right]-U_2\left[H_2^1(x)-H_{12}^1(x)\right]-U_1U_2H_{12}^1(x)\right\}\right) \ , \\
K_2(x) &= \mathrm{E}\left(\exp\left\{-U_1\left[H_1^2(x)-H_{12}^2(x)\right]-U_2\left[H_2^2(x)-H_{12}^2(x)\right]-U_1U_2H_{12}^2(x)\right\}\right) \ .
\end{aligned}$$



Supplementary Figure 1: Pointwise absolute deviances for the multinomial model and its compound Dirichlet-multinomial counterpart applied to (i) Toxoplasma and *H. pylori*, (ii) mumps and rubella, (iii) B19 and CMV and (iv) B19 and *H. pylori* infection data (2-component multiplicative model for (i) and (ii); 1-component model for (iii) and (iv)).