Parametric overdispersed frailty models for current status data.

Peer-reviewed author version

ABRAMS, Steven; AERTS, Marc; MOLENBERGHS, Geert & HENS, Niel (2017) Parametric overdispersed frailty models for current status data.. In: BIOMETRICS, 73(4), p. 1388-1400..

DOI: 10.1111/biom.12692
Handle: http://hdl.handle.net/1942/23638

Biometrics XX, 1–24 DOI: XX.XXXX/XXXXXX.x
XXXXXXX XXXX

Parametric overdispersed frailty models for current status data

Steven Abrams[1,*], Marc Aerts[1], Geert Molenberghs[1,2] and Niel Hens[1,3]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics,
Hasselt University, Diepenbeek, Belgium

[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics,
Katholieke Universiteit Leuven, Leuven, Belgium

[3]Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute (WHO Collaborating Centre),
University of Antwerp, Antwerp, Belgium

email: steven.abrams@uhasselt.be

Summary: Frailty models have a prominent place in survival analysis to model univariate and multivariate time-to-event data, often complicated by the presence of different types of censoring. In recent years, frailty modelling gained popularity in infectious disease epidemiology to quantify unobserved heterogeneity using Type I interval-censored serological data or current status data. In a multivariate setting, frailty models prove useful to assess the association between infection times related to multiple distinct infections acquired by the same individual. In addition to dependence among individual infection times, overdispersion can arise when the observed variability in the data exceeds the one implied by the model. In this paper, we discuss parametric overdispersed frailty models for time-to-event data under Type I interval-censoring, building upon the work by Molenberghs et al. (2010) and Hens et al. (2009). The proposed methodology is illustrated using bivariate serological data on hepatitis A and B from Belgium anno 1993–1994. Furthermore, the relationship between individual heterogeneity and overdispersion at a stratum-specific level is studied through simulations. Although it is important to account for overdispersion, one should be cautious when modelling both individual heterogeneity and overdispersion based on current status data as model selection is hampered by the loss of information due to censoring.

Key words: Correlated frailty models; Gompertz hazards; Infectious disease epidemiology; Overdispersed frailty models; Serological survey data; Current status data.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Frailty models are very popular in survival analysis due to their convenient way of modelling unobserved heterogeneity. In its simplest form, a frailty is a latent random proportionality factor modifying the individual's hazard function, or the one of related individuals. Although the term 'frailty' was introduced by Vaupel et al. (1979) in a univariate setting, the concept goes back to the early work of Greenwood and Yule (1920) on "accident proneness." Due to the seminal work by Clayton (1978), frailty models were highly promoted by their applicability to model multivariate survival data. In general, frailty models extend the well-known Cox proportional hazards model (Cox, 1972) by including random frailty terms allowing for a heterogeneous study population. All sampled individuals differ in their propensity to experience the event under consideration, and consequently have different event hazards. In many cases, unobserved heterogeneity arises from the inability to measure all relevant covariate information for which the event hazard needs to be adjusted. Under the proportional hazards assumption, the frailty acts multiplicatively on a baseline hazard function, defining a random-effects model for time-to-event data. In a multivariate context, the joint frailty distribution imposes a correlation structure among event- and individual-specific frailties, and consequently implies a dependence between event times. To that end, shared and correlated frailty models have been proposed (see, e.g., Wienke, 2010).

In many contemporary statistical analyses, the outcome of interest is the time to a specific event such as death, occurrence of disease, or discharge from hospital. Such time-to-event data are prominent in survival data, both in univariate as well as multivariate settings in which hierarchical structures are often present. In addition to accounting for data hierarchies, there exists a need to account for overdispersion in many data applications (Hinde and Demétrio, 1998a; Molenberghs et al., 2010). Overdispersion arises when the observed variability in the data exceeds the variation predicted by the model. In such situation, the proposed model

and its prescribed mean-variance link are too restrictive to describe the data adequately. In practice, many different causes for overdispersion exist such as cluster sampling, correlation among individual responses, and unobserved covariate information. In general, there exist two groups of models to account for overdispersion: (1) moment-based approaches relying on more flexible forms for the mean-variance relationship, and (2) two-stage models for the response entailing a distribution for one or more parameters of the response model. The latter method leads to compound probability distributions for the response variable enabling, at least in theory, full likelihood estimation of the model parameters. Broad overviews of moment-based and full-distribution approaches for dealing with overdispersion are provided by Hinde and Demétrio (1998a,b) in the context of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). These authors mainly focus on random-effects based solutions to the problem of overdispersion including the beta-binomial model for binomial data relying on beta-distributed random effects, and the negative-binomial model for Poisson counts with the natural parameter following a gamma distribution.

In the last two decades, cross-sectional serological studies, providing insight into population immunity and individual-level past infection experience, have become quintessential to inform infectious disease models. Serological data consists of infection-specific antibody titre concentrations based on which individuals are typically classified as seropositive or -negative, entailing so-called current status data. Individual immunological statuses regarding multiple infections aggregated by age at cross-sectional sampling time constitute multinomial response values. As individuals differ in social contact behaviour, susceptibility to infection and infectiousness upon infection, frailty models are of importance to quantify unobserved variability in the time to the acquisition of infections. Furthermore, the use of bivariate frailty models in infectious disease epidemiology was popularized by the seminal work of Farrington et al. (2001) and the work by Hens et al. (2009) on shared and correlated frailty

models, respectively, applied to serological data on immunizing infections. Recently, several extensions have been proposed focusing on, but not limited to, recurrent infection processes (Abrams and Hens, 2015) and time-varying individual heterogeneity (Farrington et al., 2012; Unkel et al., 2014). In addition, the issue of extra-multinomial variation in immunological response data, due to reasons different from individual variability in event hazards, is addressed by applying Dirichlet-multinomial models (Farrington et al., 2013), extending the well-known beta-binomial model for overdispersed binomial data. The approach undertaken in this paper differs from the aforementioned one in the sense that random effects are introduced at the level of the baseline hazard rather than assuming randomness directly at the probability scale (i.e., by means of the Dirichlet distribution). In essence, our approach extends the work by Molenberghs et al. (2010) to the case of current status data. These authors introduced a general and flexible framework of generalized linear models accounting for both overdispersion and clustering in case of repeated measurements through the use of two separate sets of random effects, and with particular attention given to binary, count and time-to-event data. Particular emphasis is placed on so-called conjugate random effects at the level of the mean for overdispersion, and normal random effects embedded in the linear predictor for clustering (Molenberghs et al., 2007, 2014). We focus on parametric frailty models, implying the specification of a parametric shape for the baseline hazards.

The paper is organized as follows. In Section 2, a motivating example is presented. The methodology is introduced in Section 3, and maximum likelihood estimation is discussed in Section 4. The proposed frailty models are fitted to hepatitis A and B serology for which results are shown in Section 5. Results of an additional data application are briefly discussed therein as well. In Section 6, a simulation study is performed to assess model performance and effects of estimating both individual heterogeneity and overdispersion. The manuscript

ends with a discussion on the implications of modelling multivariate current status data using overdispersed frailty models highlighting avenues for further research.

2. Case studies

Bivariate cross-sectional serology consists of blood serum samples tested for the presence of infection-specific IgG antibodies, reflecting former infection experience. Blood samples are tested using an enzyme-linked immunosorbent assay test, classifying samples (and equivalently individuals) as either being seropositive or -negative based on a pre-specified cut-off value. Hence, the individual's serological status is a direct measure of his/her immunity against the disease, at least if complete serological protection is agreed upon. Since the true infection (event) times are unobserved, and infection takes place either between birth and the observation time for seropositives or thereafter for seronegatives, one is faced with current status data. Hepatitis A and B serological survey data, obtained from a sero-epidemiological study conducted in 1993–1994 in Flanders, Belgium, is used to illustrate the methodology. Hepatitis A is a viral infection of the liver for which symptoms are diarrhoea, nausea, fever, abdominal pain and a yellow skin, and is mainly transmitted via contaminated food or water. Hepatitis B causes liver inflammation, jaundice and in rare cases death, and transmission is mainly driven by sexual and blood contact. In total, 4026 blood samples were drawn from a representative study population, and tested for the presence of hepatitis A and B antibodies. Complete immunological information on hepatitis A and B antibody prevalence was obtained for 3787 subjects, and age at the time of data collection was registered for each of these study subjects. For more details, the reader is referred to Beutels et al. (1997). Furthermore, mumps and rubella serological survey data, obtained from a large survey of prevalence of infection-specific antibodies conducted between November 1986 and December 1987 in the UK, are considered as a second data application (Morgan-Capner et al., 1988).

3. Materials and methods

3.1 *Terminology*

Time-to-event data represent the times to a specific event such as death, failure or infection. The analysis of such data is often hampered by the occurrence of censoring, implying that response values are only partially known. Right-censored observations occur when true event times exceed the follow-up period of individuals, for example, as a result of subjects dropping out before the end of the study or the study ending prior to the occurrence of the event. In order to exemplify right censoring in a bivariate setting with clustering, let T^*_{ijk} represent the true event time, C_{ijk} the censoring time and $\Delta_{ijk} = \mathbf{1}_{T^*_{ijk} \leqslant C_{ijk}}$ the censoring indicator for event $i = 1, 2$ and individual $j = 1, \ldots, N_k$ in stratum $k = 1, \ldots, K$. In case of right-censoring, the observation times T_{ijk} are equal to the true event times T^*_{ijk} only when events occur prior to censoring, i.e., $T^*_{ijk} \leqslant C_{ijk}$, and T_{ijk} equals C_{ijk} otherwise. In general, right-censoring can be considered as a special case of interval-censoring, for which T^*_{ijk} is known to take place in some time window, with time intervals $[C_{ijk}, \infty)$. Finally, in case of current status data, the true event times are unknown, hence $T_{ijk} = C_{ijk}$, for all sampled individuals and both events. The censoring indicator Δ_{ijk} represents event experienced before T_{ijk}, or event status, hence the name current status data. Throughout the paper, all derivations are made under the general assumption of clustering into strata, i.e., each subject is classified into one of the K strata, and the problem of overdispersion is discussed at the subject- as well as stratum-level. Note that we adopt the term 'stratum' throughout this paper, inspired by our data application (i.e., age cohorts), although 'cluster' can be used instead in case of hierarchical data.

3.2 *Generalized linear models*

Let T^*_{ijk} represent the time to event i for individual j in stratum k, ignoring censoring for the time being. The random variable T^*_{ijk} follows an exponential family distribution, i.e., a

member of the class of distribution functions used in a generalized linear model (McCullagh and Nelder, 1989), if the probability density function can be written as

$$f_i(t^*_{ijk}|\eta_i,\phi_i) \;=\; \exp\left[\phi_i^{-1}\left\{t^*_{ijk}\eta_i-\psi_i(\eta_i)\right\}+c_i(t^*_{ijk},\phi_i)\right], \tag{1}$$

where η_i and ϕ_i represent a specific set of unknown parameters, and $\psi_i(\cdot)$ and $c_i(\cdot,\cdot)$ are known functions. The parameter η_i is termed *natural* or *canonical parameter* whereas ϕ_i denotes the *dispersion parameter*. The mean μ_i and variance σ_i^2 of the random variable T^*_{ijk} follow from the function $\psi_i(\cdot)$ through $\mathrm{E}(T^*_{ijk})=\mu_i=\psi_i'(\eta_i)$ and $\mathrm{Var}(T^*_{ijk})=\sigma_i^2=\phi_i\psi_i''(\eta_i)$ (Molenberghs and Verbeke, 2005). In general, the mean and variance are related through $\sigma_i^2=\phi_i\psi_i''\{\psi_i'^{-1}(\mu_i)\}=\phi_i v_i(\mu_i)$, with $v_i(\cdot)$ the so-called variance function corresponding to event i. The variance function describes the mean-variance relationship.

For time-to-event data, the exponential and Weibull distributions are often considered in literature for non-negative response variables (see, e.g., Wienke, 2010; Molenberghs et al., 2010). A flexible alternative to these distributions is the Gompertz distribution encompassing both monotonic increasing and decreasing hazards. The Gompertz distribution has been used by Hens et al. (2009) to analyse the serological data introduced in Section 2. The Gompertz model $T^*_{ijk}\sim\mathcal{G}(\xi_i,\nu_i)$ can be formulated as follows:

$$f_i(t^*_{ijk}) \;=\; \xi_i\exp\left(\nu_i t^*_{ijk}\right)\exp\left[-\frac{\xi_i}{\nu_i}\left\{\exp\left(\nu_i t^*_{ijk}\right)-1\right\}\right], \tag{2}$$

where $\xi_i>0$ and $-\infty<\nu_i<\infty$ are unknown model parameters. Although the Weibull and Gompertz distributions are not part of the exponential family in the conventional fashion, they do belong to the family in a contrived way by considering transformations of the variable T^*_{ijk}. In case of the Gompertz model, one can easily show that the random variable $\nu_i^{-1}\{\exp(\nu_i T^*_{ijk})-1\}$ follows an exponential distribution with parameter ξ_i. Although we focus in the main text on the Gompertz model, expressions corresponding to the exponential and Weibull models can be found in Web Appendix A. Covariate information $\boldsymbol{x}_{ijk}$ for individual j in stratum k can be accounted for by means of the proportional hazards assumption (Cox,

1972): $T^*_{ijk}|\boldsymbol{x}_{ijk} \sim \mathcal{G}(\xi_i\kappa_{ijk}, \nu_i)$, where $\kappa_{ijk} = \exp(\boldsymbol{x}'_{ijk}\boldsymbol{\zeta}_i)$, $\boldsymbol{x}_{ijk}$ is a p-dimensional vector of known covariate values, and $\boldsymbol{\zeta}_i$ is a p-dimensional vector of unknown fixed effects parameters.

3.3 *Overdispersion models*

One elegant route to accommodate overdispersion is by means of a two-stage approach specifying a latent distribution for one of the model parameters. In general, this approach consists of choosing a conditional distribution for the outcome T^*_{ijk}, given an event-specific random effect θ_{ijk} for subject j in stratum k and covariate information $\boldsymbol{x}_{ijk}$, denoted by $f_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk})$, and combined with a distributional model for the random effect, i.e., $f_i(\theta_{ijk})$. Doing so, the marginal model of the outcome $T^*_{ijk}|\boldsymbol{x}_{ijk}$, assuming independence of the random effects θ_{ijk} and $\boldsymbol{x}_{ijk}$ and suppressing dependence on model parameters, becomes:

$$f_i(t^*_{ijk}|\boldsymbol{x}_{ijk}) \;=\; \int_{\mathcal{R}} f_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk}) f_i(\theta_{ijk}) d\theta_{ijk}, \tag{3}$$

where $\mathcal{R}$ represents the range of the overdispersion random variable θ_{ijk}. For the Gompertz setting described previously, the random effect θ_{ijk} can be introduced as $T^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk} \sim \mathcal{G}(\xi_i\theta_{ijk}\kappa_{ijk}, \nu_i)$, for θ_{ijk} a non-negative random variable:

$$f_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk}) \;=\; \xi_i\theta_{ijk}\kappa_{ijk} \exp\left(\nu_i t^*_{ijk}\right) \exp\left[-\frac{\xi_i\theta_{ijk}\kappa_{ijk}}{\nu_i}\left\{\exp\left(\nu_i t^*_{ijk}\right) - 1\right\}\right]. \tag{4}$$

More specifically, the model formulation implies a proportional hazards assumption (see, e.g., Cox, 1972):

$$\lambda_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk}) = \frac{f_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk})}{S_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk})} = \xi_i\theta_{ijk}\kappa_{ijk} \exp\left(\nu_i t^*_{ijk}\right), \tag{5}$$

where θ_{ijk} and κ_{ijk} act multiplicatively on the baseline hazard function $\lambda_{i0}(t^*_{ijk}) = \xi_i \exp(\nu_i t^*_{ijk})$, and $S_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk}) = 1 - F_i(t^*_{ijk}|\theta_{ijk}, \boldsymbol{x}_{ijk})$ is the conditional event-specific survival function. The derivation of the conditional survival function is presented in Web Appendix A.

Various random effects distributions with density $f_i(\theta_{ijk})$ can be considered. The gamma distribution is a popular choice in survival analysis since it is in line with the data range, giving rise to a Gompertz-gamma model. Furthermore, gamma random effects can be mo-

tivated by the concept of conjugacy (Cox and Hinkley, 1974; Lee et al., 2006) exploited in Molenberghs et al. (2010) in the context of repeated measurements. However, this approach is not limited to the use of gamma random effects and other non-negative distributions such as the inverse Gaussian distribution produce tractable expressions for the marginal densities $f_i(t^*_{ijk}|\boldsymbol{x}_{ijk})$ and marginal survival functions $S_i(t^*_{ijk}|\boldsymbol{x}_{ijk})$. In Table 1, the model components for the Gompertz-gamma and Gompertz-inverse Gaussian models are summarized in terms of the Gompertz hazard $\lambda_{i0}(t^*_{ijk})$ and integrated or cumulative Gompertz hazard $\Lambda_{i0}(t^*_{ijk}) = (\xi_i/\nu_i)\{\exp(\nu_i t^*_{ijk}) - 1\}$.

[Table 1 about here.]

Note that the expressions for the unconditional survival functions coincide with the evaluation of the Laplace transform of θ_{ijk} in $\kappa_{ijk}\Lambda_{i0}(t^*_{ijk})$, i.e., $S_i(t^*_{ijk}|\boldsymbol{x}_{ijk}) = \mathcal{L}_{\theta_{ijk}}\{\kappa_{ijk}\Lambda_{i0}(t^*_{ijk})\}$. The Gompertz-gamma and Gompertz-inverse Gaussian models presented in Table 1 define gamma and inverse Gaussian frailty models, respectively, with Gompertz baseline hazard functions (Wienke, 2010). The event-specific random effects θ_{ijk}, $i = 1, 2$, are termed individual frailties and can be assumed (1) independent (univariate frailty model); (2) equal $\theta_{1jk} = \theta_{2jk} = \theta_{jk}$ (shared frailty model); or (3) correlated (correlated frailty model), requiring the specification of a bivariate frailty distribution for $\boldsymbol{\theta}_{jk} = (\theta_{1jk}, \theta_{2jk})'$. One way to define a correlation structure among frailties θ_{1jk} and θ_{2jk} is by means of the 'variable-in-common' method; $\theta_{ijk} = \sigma^2_{\theta_i}(W_{0jk} + W_{ijk})$, where the components W_{cjk} are independent random variables with mean and variance ω_c, $c = 0, 1, 2$. In survival analysis, it is customary to set the mean of the frailties θ_{ijk} equal to one for reasons of identifiability. Therefore, $\alpha_i\beta_i = 1$ or $\beta_i = 1$ for gamma or inverse Gaussian random effects, respectively, leading to frailty variances $\mathrm{Var}(\theta_{ijk}) = \sigma^2_{\theta_i} = \alpha_i^{-1}$. In the correlated frailty setting (3) with additive decomposition, this identifiability constraint implies frailty variances $\mathrm{Var}(\theta_{ijk}) = \sigma^2_{\theta_i} = 1/(\omega_0 + \omega_i)$, and non-negative correlation $\rho_\theta = \omega_0/\sqrt{(\omega_0 + \omega_1)(\omega_0 + \omega_2)}$. The correlation is bounded above by

the minimum of the ratios of the frailty standard deviations, i.e., $0 \leqslant \rho_\theta \leqslant \min\left(\frac{\sigma_{\theta_1}}{\sigma_{\theta_2}}, \frac{\sigma_{\theta_2}}{\sigma_{\theta_1}}\right)$. As the event-specific random effects describe both the between-subject variability as well as the association between event-specific event times, ρ_θ and the random-effects variances are dependent. In the remainder of this paper, the variance-covariance matrix associated with the random vector $\boldsymbol{\theta}_{jk}$ is denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Although dependence between event times within the same individual is imposed by means of the specified covariance structure, the model formulation in (5) implies independence among event times of different individuals, irrespective of the stratum to which they belong. Therefore, overdispersion at the stratum-level is not yet accounted for. Hereunder, two different methods are considered to do so.

3.3.1 *Dirichlet-multinomial model.* First of all, we consider the Dirichlet-multinomial (DM) model to accommodate for extra-multinomial variability at the stratum level. This model has been considered before by Farrington et al. (2013) for the analysis of bivariate serological data. In order to introduce this marginal model, we consider bivariate current status data $(\delta_{1jk}, \delta_{2jk}, t_{1jk}, t_{2jk})$, where δ_{ijk} and t_{ijk} are the observed status and observation times regarding event $i = 1, 2$ for subject $j = 1, \ldots, N_k$ in stratum $k = 1, \ldots, K$, respectively. Let n_{lmk} denote the number of subjects in stratum k having status l and m (l, $m = 0$, 1) for event 1 and 2, respectively. Therefore, the data comprise 4-tuples $\boldsymbol{n}_k = (n_{00k}, n_{10k}, n_{01k}, n_{11k})$ with expected proportions $\boldsymbol{p}_k = (p_{00k}, p_{10k}, p_{01k}, p_{11k})$ in the four cells. The DM model can be used to account for overdispersion at the stratum-level, thereby hypothesizing a mixture distribution directly on the probability scale:

$$\begin{aligned} \boldsymbol{n}_k|\boldsymbol{\pi}_k &\sim \text{Multinomial}\left(N_k, \boldsymbol{\pi}_k\right) \\ \boldsymbol{\pi}_k &\sim \text{Dirichlet}\left(\varphi\boldsymbol{p}_k\right), \end{aligned}$$

resulting in a compound distribution with Dirichlet parameters $\varphi\boldsymbol{p}_k = (\varphi p_{lmk})_{l,m} > 0$, and the marginal density function for $\boldsymbol{n}_k$ given by:

$$f(\boldsymbol{n}_k|\boldsymbol{p}_k,\varphi) \;=\; C_{n_{lmk}}^{N_k}\frac{\Gamma(\varphi)}{\Gamma(N_k+\varphi)}\frac{\prod_{l,m=0}^{1}\Gamma(n_{lmk}+\varphi p_{lmk})}{\prod_{l,m=0}^{1}\Gamma(\varphi p_{lmk})}, \tag{6}$$

where $N_k = \sum_{l,m=0}^{1} n_{lmk}$ and $C_{n_{lmk}}^{N_k} = N_k!/\prod_{l,m=0}^{1} n_{lmk}!$ the normalizing constant. When individual contributions to n_{lmk} are independent, the DM model reduces to a multinomial one for the response vector $\boldsymbol{n}_k$. However, extra-multinomial variation is introduced using the overdispersion parameter $\varphi > 0$ implying correlation $\rho = \sqrt{1/(1+\varphi)}$ among individual multinomial responses within the same stratum ($0 < \rho < 1$).

Suppose that, in an infectious disease context with K age strata, serum samples of N_k individuals of age t_k are available, for $k = 1, \ldots, K$. Furthermore, let n_{lmk} represent the number of individuals of age t_k with status l and m with respect to infection 1 and 2, respectively. Consequently, the expected proportions in the four cells are given by the population survival functions derived from the overdispersed frailty model formulated in equation (5). We will come back to this in Section 4.

3.3.2 *Multiplicative overdispersed frailty models.* Since individuals of the same stratum are likely to be correlated, the model presented in (5) can be extended to incorporate additional stratum-specific random effects υ_{ik} for subjects in stratum $k = 1, \ldots, K$ and event $i = 1, 2$. These random effects can be introduced at the level of the hazard, implying:

$$\begin{aligned}
\lambda_i(t_{ijk}^*|\theta_{ijk},\upsilon_{ik},\boldsymbol{x}_{ijk}) &= \xi_i\theta_{ijk}\upsilon_{ik}\kappa_{ijk}\exp\left(\nu_i t_{ijk}^*\right),\\
f_i(t_{ijk}^*|\theta_{ijk},\upsilon_{ik},\boldsymbol{x}_{ijk}) &= \theta_{ijk}\upsilon_{ik}\kappa_{ijk}\lambda_{i0}(t_{ijk}^*)\exp\left[-\theta_{ijk}\upsilon_{ik}\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)\right],
\end{aligned} \tag{7}$$

again suppressing dependence on model parameters associated with the baseline hazard function $\lambda_{i0}(t_{ijk}^*)$, the frailty distribution $f_i(\theta_{ijk})$ and the covariate model κ_{ijk}. The random vector $\boldsymbol{\upsilon}_k = (\upsilon_{1k}, \upsilon_{2k})'$ has mean vector $\mathbf{1}$, to ensure identifiability, and variance-covariance

matrix $\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}} = \begin{pmatrix} \sigma^2_{\upsilon_1} & \rho_{\upsilon}\sigma_{\upsilon_1}\sigma_{\upsilon_2} \\ \rho_{\upsilon}\sigma_{\upsilon_1}\sigma_{\upsilon_2} & \sigma^2_{\upsilon_2} \end{pmatrix}.$$

Outcomes for event i from subjects in the same stratum are therefore correlated through υ_{ik} while observations from different strata are assumed independent. Although it is not strictly necessary, the two sets of random effects $\boldsymbol{\theta}_{jk}$ and $\boldsymbol{\upsilon}_k$ are assumed to be independent. Distributional assumptions for $\boldsymbol{\theta}_{jk}$ and $\boldsymbol{\upsilon}_k$ produce a marginal model $f_i(t^*_{ijk}|\boldsymbol{x}_{ijk})$. In the next section, the methodology is cast into the maximum likelihood (ML) framework.

4. Maximum likelihood estimation

Fitting the overdispersed frailty model in equation (7) to bivariate uncensored time-to-event data $\boldsymbol{t}^*_{jk} = (t^*_{1jk}, t^*_{2jk}, \boldsymbol{x}_{ijk})$ proceeds by integrating over the latent random effects, or frailties, resulting in the following likelihood contribution for stratum k:

$$L_k(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}|\boldsymbol{t}^*_{jk}, \boldsymbol{x}_{jk}) = \int_{\mathcal{R}} \prod_{j=1}^{N_k} f_{12}(\boldsymbol{t}^*_{jk}|\boldsymbol{\vartheta}, \boldsymbol{\theta}_{jk}, \boldsymbol{\upsilon}_k, \boldsymbol{x}_{ijk}) f(\boldsymbol{\theta}_{jk}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}) f(\boldsymbol{\upsilon}_k|\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}) d\boldsymbol{\theta}_{jk} d\boldsymbol{\upsilon}_k,$$

where $f_{12}(\boldsymbol{t}^*_{jk}|\boldsymbol{\vartheta}, \boldsymbol{\theta}_{jk}, \boldsymbol{\upsilon}_k, \boldsymbol{x}_{jk}) = f_1(t^*_{1jk}|\boldsymbol{\vartheta}_1, \theta_{1jk}, \upsilon_{1k}, \boldsymbol{x}_{1jk}) f_2(t^*_{2jk}|\boldsymbol{\vartheta}_2, \theta_{2jk}, \upsilon_{2k}, \boldsymbol{x}_{2jk})$ under the assumption of conditional independence of the event times T^*_{1jk} and T^*_{2jk} given the random frailties, and $\boldsymbol{\vartheta}_i$ the vector of infection-specific baseline hazard parameters ξ_i and ν_i, and regression parameters $\boldsymbol{\zeta}_i$. Under model (4) the likelihood contribution for stratum k simplifies to the product of individual contributions

$$L_{jk}(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}|\boldsymbol{t}^*_{jk}, \boldsymbol{x}_{jk}) = \int_{\mathcal{R}} f_1(t^*_{1jk}|\boldsymbol{\vartheta}_1, \theta_{1jk}, \boldsymbol{x}_{1jk}) f_2(t^*_{2jk}|\boldsymbol{\vartheta}_2, \theta_{2jk}, \boldsymbol{x}_{2jk}) f(\boldsymbol{\theta}_{jk}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\, d\boldsymbol{\theta}_{jk},$$

which corresponds to assuming a degenerate distribution for the random vector $\boldsymbol{\upsilon}_k$ at $\mathbf{1}$. In general, the likelihood function becomes

$$L\left(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}|\boldsymbol{t}^*_{jk}, \boldsymbol{x}_{jk}\right) = \prod_{k=1}^{K} L_k(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}|\boldsymbol{t}^*_{jk}, \boldsymbol{x}_{jk}) \tag{8}$$

Maximizing the likelihood in (8) is complicated due to the presence of K integrals. Partial marginalization can be considered to overcome the direct maximization problem, in agreement with Molenberghs et al. (2010), thereby integrating out one set of random effects.

Partial marginalization is performed by integrating the conditional density $f_i(t^*_{ijk}|\boldsymbol{\vartheta}_i, \theta_{ijk}, \upsilon_{ik}, \boldsymbol{x}_{ijk})$ over the frailty distribution $f_i(\theta_{ijk})$, leaving the frailty term υ_{ik} untouched. Integrating over θ_{ijk} yields

$$f_i(t^*_{ijk}|\boldsymbol{\vartheta}_i, \upsilon_{ik}, \boldsymbol{x}_{ijk}) = -\frac{d}{dt} S_i(t|\boldsymbol{\vartheta}_i, \upsilon_{ik}, \boldsymbol{x}_{ijk})\Big|_{t=t^*_{ijk}} = -\frac{d}{dt} \mathcal{L}_{\theta_{ijk}} \left[\upsilon_{ik}\kappa_{ijk}\Lambda_{i0}(t)\right]\Big|_{t=t^*_{ijk}} \tag{9}$$

in terms of the survival function $S_i(\cdot)$, or the Laplace transform $\mathcal{L}_{\theta_{ijk}}(\cdot)$ with respect to θ_{ijk}. Closed-form expressions for the Laplace transform are available for gamma and inverse Gaussian random variables θ_{ijk} (see Section 3.3 and Web Appendix B):

$$\mathcal{L}_{\theta_{ij}}(s) = (1+\beta_i s)^{-\alpha_i}, \qquad \mathcal{L}_{\theta_{ij}}(s) = \exp\left\{\frac{\alpha_i}{\beta_i}\left(1-\sqrt{1+\frac{2\beta_i^2 s}{\alpha_i}}\right)\right\},$$

respectively. The product of univariate density functions $f_i(t^*_{ijk}|\boldsymbol{\vartheta}_i, \upsilon_{ik}, \boldsymbol{x}_{ijk})$ produces the joint density $f_{12}(\boldsymbol{t}^*_{jk}|\boldsymbol{\vartheta}, \boldsymbol{\upsilon}_k, \boldsymbol{x}_{jk})$ in the univariate frailty context, thereby assuming independence between θ_{1jk} and θ_{2jk}. In general, deriving the joint density function, conditional on random effects υ_{ik}, involves integration over the joint density $f(\boldsymbol{\theta}_{jk}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, and can be expressed in terms of the joint Laplace transform. Expressions for the partially marginalized univariate and joint density functions under different distributional assumptions for θ_{ijk} are presented in Web Appendix B. The principle of partial marginalization is useful to lower the dimensionality of integration when integrating out random effects numerically, and applies when strong conjugacy holds (see, e.g., Molenberghs et al., 2014).

Since we are faced with current status data, the likelihood function requires modification. Therefore, consider bivariate cross-sectional serological data $(\delta_{1jk}, \delta_{2jk}, t_{1jk}, t_{2jk})$ with δ_{ijk} and t_{ijk} as previously defined, and univariate observation times $t_{1jk} = t_{2jk} \equiv t_{jk}$ for both events. In our application, strata are defined based on age cohorts such that all subjects within the same stratum are observed at the same time, i.e., $t_{1k} = \ldots = t_{N_k k} \equiv t_k$. The random

vector $\boldsymbol{n}_k = (n_{00k}, n_{10k}, n_{01k}, n_{11k})$, where n_{lmk} represents the number of subjects in stratum k with status l and m with regard to event 1 and 2, respectively, follows a multinomial distribution, conditional on observation time $T_{ijk} = t_k$, with probability vector $\boldsymbol{p}_k = \boldsymbol{p}(t_k) = (p_{lm}(t_k|\boldsymbol{\vartheta}, \boldsymbol{\Sigma_\theta}, \boldsymbol{\Sigma_\upsilon}))_{l,m=0,1}$; $p_{lmk} = p_{lm}(t_k|\boldsymbol{\vartheta}, \boldsymbol{\Sigma_\theta}, \boldsymbol{\Sigma_\upsilon}) = \Pr(\Delta_{1jk} = l, \Delta_{2jk} = m \mid \boldsymbol{\vartheta}, \boldsymbol{\Sigma_\theta}, \boldsymbol{\Sigma_\upsilon})$. The stratum-specific multinomial contribution of bivariate aggregated current status data $(\boldsymbol{n}_k, t_k)$ to the likelihood is given by:

$$L_k(\boldsymbol{n}_k, t_k) \;=\; C_{n_{lmk}}^{N_k} \prod_{l,m=0}^{1} p_{lmk}^{n_{lmk}},$$

suppressing dependence on the model parameters $\boldsymbol{\vartheta}$, $\boldsymbol{\Sigma_\theta}$ and $\boldsymbol{\Sigma_\upsilon}$ for the sake of simplicity. The multinomial probabilities p_{lmk} can be expressed in terms of the marginal univariate and joint survival functions as follows: $p_{11k} = 1 - S_1(t_k) - S_2(t_k) + S_{12}(t_k, t_k)$, $p_{10k} = S_2(t_k) - S_{12}(t_k, t_k)$, $p_{01k} = S_1(t_k) - S_{12}(t_k, t_k)$, $p_{00k} = S_{12}(t_k, t_k)$, where marginal survival functions are obtained after integrating out the frailties, and derivation of the joint survival function relies on the conditional independence assumption. The likelihood contribution for aggregated age-group data $\boldsymbol{n}_k$ in group k in the DM model equals $f(\boldsymbol{n}_k|\boldsymbol{p}_k, \varphi)$ in equation (6). In both cases, the likelihood functions are constructed by taking the product of contributions L_k, $k = 1, \ldots, K$.

5. Data Applications

5.1 *Hepatitis A and B, Flanders, Belgium*

In this section, the methodology is illustrated on hepatitis A and B serological data from Flanders, Belgium, anno 1993–1994 for which the baseline hazard can plausibly be assumed to be of the Gompertz type (Web Appendix C). The event times of interest T_{ijk}^* are infection times with regard to hepatitis A ($i = 1$) and hepatitis B ($i = 2$). Univariate observation times T_{jk} and censoring indicators Δ_{ijk} refer to the age at the time of data collection and the immunological status, respectively, for individual $j = 1, \ldots, N_k$ in age-group (stratum) k. Age strata of length one are considered in the analyses giving rise to a total of K = 95

strata. Stratification in age cohorts is a natural one since behavioural and/or environmental conditions with respect to the acquisition of infections could vary across age cohorts, e.g., overdispersion resulting from test variability (Unkel et al., 2014). Individual covariate information is absent in our data application, thereby excluding κ_{ijk} from the general model formulations in (6) and (7). The models are fitted using the SAS procedure NLMIXED. All SAS code is made available on the I-BioStat website (https://ibiostat.be/online-resources/).

5.1.1 *Overdispersed frailty models.* In Table 2, an overview of the overdispersed frailty models (with Gompertz baseline hazards) is presented with distributional assumptions regarding infection- and indvidual-specific frailties θ_{ijk} and stratum-specific random effects υ_{ik}. Traditional frailty models result from assuming a degenerate distribution at $\mathbf{1}$ for $\boldsymbol{\upsilon}_k$. Alternatively, lognormal random effects $\boldsymbol{\upsilon}_k$ with mean vector $\mathbf{1}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}$ are considered to accommodate extra-multinomial variability at the age-group level. Note that the specification of lognormal random effects υ_{ik} is equivalent to defining a random vector $\boldsymbol{b}_k \sim N_2(\boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\Sigma}_{\boldsymbol{b}})$ with $\upsilon_{ik} = \exp(b_{ik})$ and $\boldsymbol{\mu}_{\boldsymbol{b}} = \left(0.5\sigma^2_{b_1}, 0.5\sigma^2_{b_2}\right)'$ ensuring that $\mathrm{E}(\upsilon_{ik}) = 1$. Furthermore, the variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ takes the form:

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \begin{pmatrix} \log(\sigma^2_{\upsilon_1} + 1) & \log(\rho_\upsilon \sigma_{\upsilon_1} \sigma_{\upsilon_2} + 1) \\ \log(\rho_\upsilon \sigma_{\upsilon_1} \sigma_{\upsilon_2} + 1) & \log(\sigma^2_{\upsilon_2} + 1) \end{pmatrix}.$$

Recall that the frailty variances are related to the parameters α_i through $\sigma^2_{\theta_i} = 1/\alpha_i$.

[Table 2 about here.]

In Table 3, ML estimates for the model parameters in the shared and correlated Gompertz-gamma and -inverse Gaussian frailty models are shown together with AIC- and BIC-values (upper part). The correlated gamma frailty model outperforms all other fitted models based on both information criteria, implying a correlation between individual infection times which differs from unity (shared frailty model). Despite the fact that the data favours a model

accounting for correlation among event times ($\rho_\theta \neq 0$), results from univariate Gompertz-gamma and -inverse Gaussian frailty models are provided in Web Appendix C.

[Table 3 about here.]

Models combining individual- and stratum-specific random effects are presented in the middle part of Table 3. Gompertz-gamma-lognormal frailty models clearly perform better than the traditional frailty models, indicating that overdispersion at the group-level exists. Again, the combined model with correlated individual-level random effects θ_{ijk}, either unrestricted or constrained with $\sigma_{\upsilon_1} = \sigma_{\upsilon_2}$, yields a better fit to the serological data as compared to univariate and shared alternatives. Based on AIC- and BIC-values, the analysis corresponding to the constrained correlated gamma-lognormal model can be viewed as the final one. Although inverse Gaussian-lognormal models have been considered, these models did not outperform the gamma-lognormal counterparts, and results therefrom are therefore displayed in Web Appendix C. The model fit of the best model based on AIC- and BIC-values is graphically depicted in Figure 1, displaying the model-based multinomial probabilities $\boldsymbol{p}_k = (p_{00k}, p_{10k}, p_{01k}, p_{11k})$ together with the observed proportions. The added value of this analysis compared to the previous one reported by Hens et al. (2009) is a more reliable assessment of the amount of unobserved heterogeneity, which is quintessential for the estimation of epidemiological parameters. Furthermore, Hens et al. (2009) did not discuss the implications of heterogeneity on the estimation of these parameters. In Web Appendix C, we illustrate the estimation of the (basic) reproduction number and critical vaccination coverage, two commonly used epidemiological measures to describe a pathogens transmission potential and the effort required to avoid outbreaks by means of vaccination, respectively, in the presence of individual heterogeneity. Ignoring such heterogeneity leads to a substantial underestimation of both quantities. On top of that, an appreciable difference was found

between estimates derived from models with and without accounting for overdispersion thereby underlining the importance of modelling overdispersion.

[Figure 1 about here.]

5.1.2 *Dirichlet-multinomial models.* We fit the DM Gompertz-gamma models introduced in Section 3.3 to the hepatitis A and B serology. These models differ from the Gompertz-gamma-lognormal models in the sense that they introduce randomness directly on the multinomial probability scale. Larger values for the overdispersion parameter ρ, or equivalently smaller values for φ, imply more evidence in favour of a model allowing for overdispersion compared to the multinomial model. The correlated frailty model outperforms all other DM models (see lower part of Table 3) and model fit is almost equivalent to the one for the less parsimonious correlated Gompertz-gamma-lognormal model (see middle part of Table 3). In conclusion, it seems sensible to account for overdispersion when modelling the serology under study, albeit that equivalent overdispersed frailty models in terms of model fit provide different heterogeneity estimates. In Section 6, a simulation approach is used to assess performance of the described methodology in light of these findings.

5.2 *Rubella and mumps, UK*

In Web Appendix D, we illustrate the application of the proposed models to serological data on mumps and rubella in the UK, anno 1986–1987 (Farrington et al., 2001). In general, the novel overdispersed frailty models outperform the traditional shared and correlated ones. Our results are in line with those reported by Farrington et al. (2001) in terms of the marginal forces of infection and common transmission route for mumps and rubella. However, the current analysis extends the aforementioned one by combining both overdispersion and the inclusion of individual-specific frailty terms thereby capturing the dependence between the two infections as well as accounting for individual heterogeneity. Furthermore, the novel

models are more flexible compared to the DM shared gamma frailty model, both in terms of the overdispersion process as well as the correlation structure for the individual-specific frailty terms.

6. Simulation study

In this simulation study, we aim at evaluating the performance of the Gompertz-gamma-lognormal model in case of current status data. We mainly focus on the estimation of the heterogeneity parameters and quantify the impact of different model assumptions thereon. The simulation set-up was carried out under different scenarios, enabling the investigation of the effects of censoring, implications when jointly estimating heterogeneity and overdispersion, and effects of sample size and parametric baseline hazards. For an overview of the simulation steps and details concerning the general simulation protocol, we refer to Web Appendix E. In the simulation approach presented in this paper, Gompertz baseline hazards ($\lambda_{i0}(t) = \xi_i \exp(\nu_i t)$) are considered. Simulation results identifying the impact of sample size and information loss due to censoring are presented in Web Appendix E. In addition, performance of the models in case of exponential event times, implying baseline hazards $\lambda_{i0}(t) = \lambda_{i0}$, is discussed there as well.

6.1 *Misspecification of overdispersion process*

In this section, misspecification of the overdispersion process at the stratum-level is studied. We relied on the correlated gamma-lognormal model to generate current status data, and investigated the performance of the correlated gamma and Dirichlet-multinomial correlated gamma frailty models. Alternatively, one can assess the performance of the correlated gamma-lognormal model when simulating data under the DM model. The latter approach is undertaken in Web Appendix E. Table 4 shows the true values, mean parameter estimates (Mean), relative bias (Rel. Bias) and empirical standard error (e.s.e.) estimates, convergence

rate (CR), and AIC- and BIC-selection percentages (AIC % and BIC %) for current status data under the Gompertz-gamma-lognormal model (based on 500 runs). The choice of the Gompertz parameters, heterogeneity parameters and sample size $N = 3787$ are inspired by the hepatitis A and B case study, entailing multisera data which typically has a rather large sample size. Estimates for the heterogeneity parameters are biased when ignoring the extra-multinomial variation at the age-group level (i.e., in the correlated gamma model). Both within- as well as between-stratum dependence among bivariate event times are estimated using simulation-based Kendall's τ estimates, denoted by $\hat{\tau}_{WS}$ and $\hat{\tau}_{BS}$, respectively (see Web Appendix E). Based on AIC- and BIC-criteria, dealing with the trade-off between fit and parsimony of the models, the correct model is identified in 98% and 89% of the simulation runs, respectively. Consequently, one needs to be cautious when modelling both individual heterogeneity and unobserved age-group variability based on current status data as selecting the incorrect model affects parameter estimates. More specifically, reliable estimates for the heterogeneity parameters σ_{θ_i} are of importance to derive relevant epidemiological parameters such as the basic reproduction number (Coutinho et al., 1999).

[Table 4 about here.]

6.2 *Misspecification of individual heterogeneity*

Finally, we consider misspecification of the bivariate individual frailty distribution for $\boldsymbol{\theta}_{jk}$. To date, the popular but restrictive shared frailty model is often considered to describe bivariate time-to-event data. Therefore, correlated gamma-lognormal current status data is generated, and shared gamma, shared gamma-lognormal and DM shared gamma frailty models are fitted to the simulated data (see Table 5). Although the estimates of the Gompertz baseline parameters can be considered stable across the various models, the estimated variance parameters σ_{θ_i} differ substantially. Clearly, the heterogeneity parameters σ_{θ_i} are underestimated when incorrectly assuming shared frailties instead of correlated frailties with

equal variances. However, this is in line with the negative correlation between ρ_θ and σ_{θ_i} which has been shown before in the context of bivariate correlated frailty models (Wienke et al., 2005). The shared frailty model, assuming perfect correlation and common frailty variance, yields the lowest variance estimate compared to the models accommodating overdispersion. In conclusion, misspecifying the individual heterogeneity process has a substantial impact on the estimation of both the frailty variances and overdispersion parameters. The simulation-based Kendall's τ estimates for the shared gamma frailty model are in line with what we expect theoretically, i.e., $\hat{\tau}_{WS} = \hat{\tau}_{BS} \approx \hat{\sigma}^2_{\theta_1}/(\hat{\sigma}^2_{\theta_1} + 2) = 0.226$ (Wienke, 2010).

[Table 5 about here.]

7. Discussion

Building upon the work by Hens et al. (2009) and Molenberghs et al. (2010), we have studied parametric overdispersed frailty models, combining gamma or inverse Gaussian distributed individual frailty terms with lognormally distributed stratum-specific random effects, in the context of current status data. Although the choice of frailty distributions is merely inspired by the concepts of partial marginalization and conjugacy through their closed-form expressions for the Laplace transform, other frailty distributions could be considered thereby increasing the computational burden. Indeed, this leads to intractable expressions for the likelihood function and, in combination with lognormal random effects, prevents partial marginalization. Furthermore, (correlated) individual infection-specific frailties impose association between individual event times whereas the lognormal random effects capture overdispersion at the stratum-level. Semi-parametric correlated frailty models in which the hazard functions are left unspecified cannot be considered due to well-known identifiability issues (Iachine, 2004). However, since we are interested in the estimation of the amount of unobserved heterogeneity and the strength of the association between event times, a parametric choice with regard to the baseline hazard function is a natural one. Particular

attention was given to the Gompertz distribution for (bivariate) time-to-event outcomes, albeit that the presented methodology is generally applicable for all kinds of non-negative distributions entailing parametric hazard functions (see Web Appendices A and C).

Although attention is confined to maximum likelihood estimation including partial marginalization regarding the gamma frailties, pseudo-likelihood could be considered as an alternative for which a large advantage in terms of computational stability has been noticed before (Molenberghs et al., 2014). Performance of the general model combining individual frailties and overdispersion random effects is evaluated and contrasted with the Dirichlet-multinomial frailty model. The Dirichlet-multinomial model provides an easy way to accommodate overdispersion in multinomial response data. We combined frailty methodology with the Dirichlet-multinomial distribution to analyse bivariate current status data. Although these models are quite simple in nature, they come with the price of a reduced flexibility as compared to the combined (e.g., gamma-lognormal) frailty models. In addition, the Dirichlet-multinomial model has no straightforward counterpart for the analysis of uncensored (or right-censored) time-to-event data.

Since reliable estimates for (individual) heterogeneity parameters are quintessential in infectious disease epidemiology when deriving important epidemiological parameters such as the basic reproduction number R_0 (Farrington et al., 2001), the correct assessment of both individual heterogeneity and overdispersion is crucial. Our simulation study reveals that one needs to be cautious when modelling both individual heterogeneity and overdispersion, notwithstanding they are acting at different hierarchical levels and playing distinct roles, since misspecifying one process has large consequences with regard to the estimation of the other. Although Dirichlet-multinomial frailty models could outperform traditional bivariate frailty models, the performance of models accommodating overdispersion by means of introducing (log-)normal random effects at the hazard level should be investigated as well.

Furthermore, large sample sizes are required to infer parameters in the overdispersed frailty models due to information loss in case of current status data. Although model selection criteria such as AIC and BIC are used to perform selection, a formal goodness-of-fit test would be very useful to assess whether the models under investigation describe the data adequately. The development of such a test is an interesting topic for further research.

The analyses in this paper rely on the assumption of time-invariant individual heterogeneity. However, extensions towards time-varying frailty models for single events have been proposed by Farrington et al. (2012), and discussed further by Unkel et al. (2014). Overdispersed frailty models encompassing both the concepts of time-varying individual frailties and overdispersion random effects at the age-group level provide an avenue for further research. Furthermore, correlated frailties are constructed by means of the 'variable-in-common' method which is typically used (Wienke, 2010). Nevertheless, in general, bivariate distributions for $\boldsymbol{\theta}_{jk}$ could be imposed implying different association structures among individual event times, albeit potentially at the cost of tractable expressions for the likelihood.

Finally, restricting the variance components in the hierarchical frailty models to be non-negative implies a positive intraclass correlation, meaning that two members of the same cluster are more alike than those from different groups. Although such models prohibit negative dependence among subjects in the same cluster, their induced marginal models do not, thereby being able to account for underdispersion (Molenberghs and Verbeke, 2005, 2011). However, underdispersion seems unrealistic for the type of data presented here, and is therefore considered beyond the scope of this manuscript.

8. Supplementary Materials

Web Appendices A–E, referenced in Sections 3–7, are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was supported by the Research Fund of Hasselt University (grant BOF11NI31 to S.A.). N.H. received support from the University of Antwerp by way of the Scientific Chair in Evidence-based Vaccinology, financed in 2009–2014 by a gift from Pfizer, Inc., New York. The authors gratefully acknowledge financial support from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). This research is part of a project that has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement 682540 – TransMID).

References

Abrams, S. and Hens, N. (2015). Modeling individual heterogeneity in the acquisition of recurrent infections: an application to parvovirus B19. *Biostatistics* **16,** 129–142.

Beutels, M., Van Damme, P., Aelvoet, W., Desmyter, J., Dondeyne, F., Goilav, C., Mak, R., Muylle, L., Pierard, D., Stroobant, A., Van Loock, F., Waumans, P., and Vranckx, R. (1997). Prevalence of hepatitis A, B and C in the Flemish population. *European Journal of Epidemiology* **13,** 275–280.

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,** 141–151.

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **34,** 187–220.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall/CRC, London.

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,** 251–292.

Farrington, C. P., Unkel, S., and Anaya-Izquierdo, K. (2012). The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **74,** 1–24.

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

Greenwood, M. and Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* **83,** 255–279.

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,** 2785–2800.

Hinde, J. and Demétrio, C. G. B. (1998a). Overdispersion: Models and estimation. *Computational Statistics & Data Analysis* **27,** 151–170.

Hinde, J. and Demétrio, C. G. B. (1998b). Overdispersion: Models and estimation. *XIII Sinape, São Paulo* .

Iachine, I. A. (2004). Identifiability of bivariate frailty models. Preprint 5, Department of Statistics, University of Southern Denmark, Odense.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Chapman & Hall/CRC, Boca Raton.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, London.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Molenberghs, G. and Verbeke, G. (2011). A note on a hierarchical interpretation of negative variance components. *Statistical Modelling* **11,** 389–408.

Molenberghs, G., Verbeke, G., and Demétrio, C. G. B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis* **13,** 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G. B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* **25,** 325–347.

Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R., and Demétrio, C. G. B. (2014). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Statistical Methods in Medical Research* **24,** 434–452.

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

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **135,** 370–384.

Unkel, S., Farrington, C. P., Withaker, H. J., and Pebody, R. (2014). Time varying frailty models and the estimation of heterogeneities in transmission of infectious diseases. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **63,** 141–158.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16,** 439–454.

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

Wienke, A., Arbeev, K., Locatelli, I., and Yashin, A. I. (2005). A comparison of different correlated frailty models and estimation strategies. *Mathematical Biosciences* **198,** 1–13.

Received XXXX XXXX. *Revised XXXX* XXXX. *Accepted XXXX* XXXX.

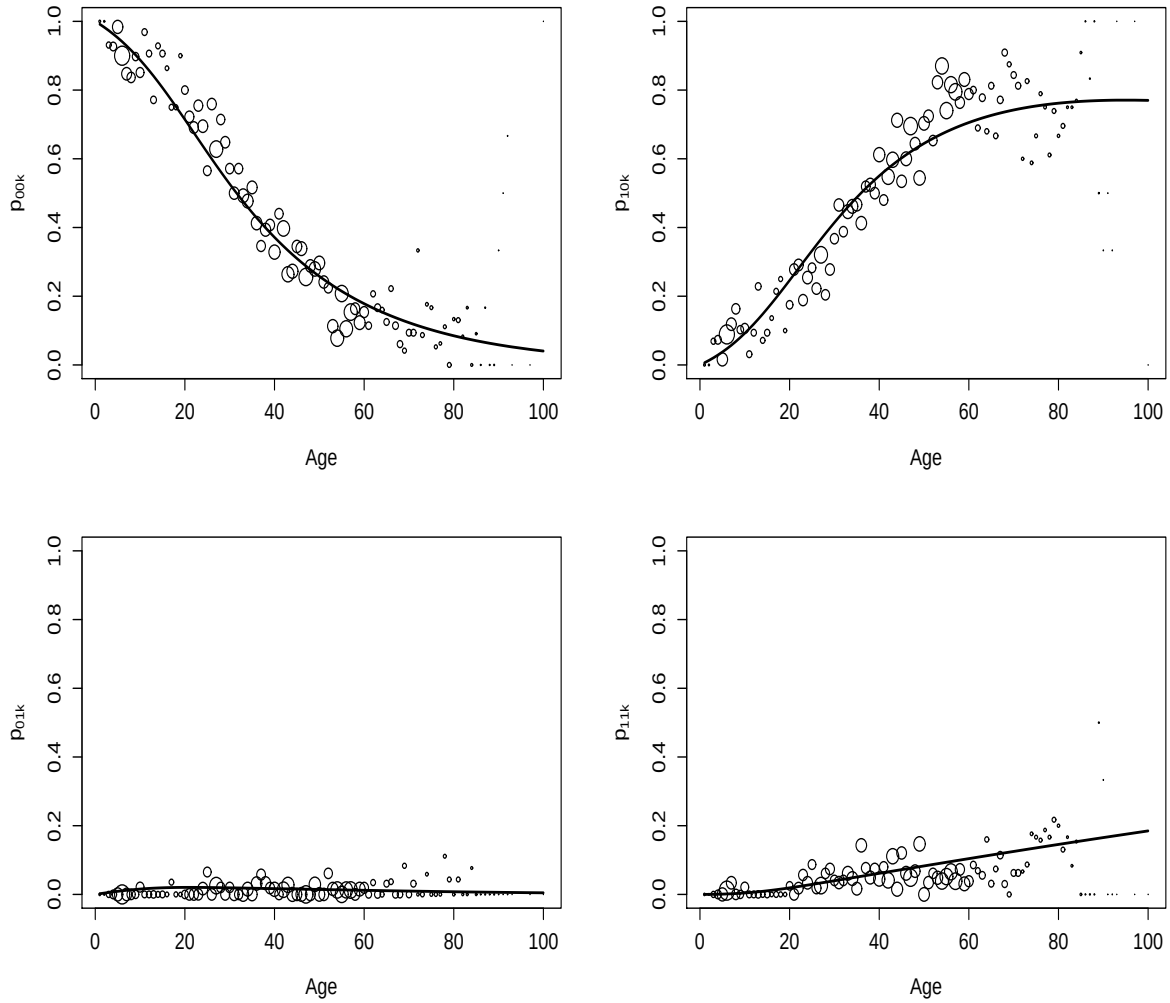

Figure 1. Predicted multinomial probabilities $\boldsymbol{p}_k = (p_{00k}, p_{10k}, p_{01k}, p_{11k})$ (solid lines) corresponding to the correlated Gompertz-gamma-lognormal frailty model and observed proportions (black circles) based on the hepatitis A and B serology with size proportional to the number of observations in each age group.

Table 1
Model elements for the Gompertz-gamma and Gompertz-inverse Gaussian models.

Element	Gompertz-Gamma $\theta_{ijk} \sim \Gamma(\alpha_i, \beta_i)$	Gompertz-Inverse Gaussian $\theta_{ijk} \sim IG(\alpha_i, \beta_i)$
$f_i(\theta_{ijk})$	$\frac{\theta_{ijk}^{\alpha_i - 1} \exp\left(\theta_{ijk}/\beta_i\right)}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}$	$\sqrt{\frac{\alpha_i}{2\pi\theta_{ijk}^3}} \exp\left[-\frac{\alpha_i}{2\beta_i^2\theta_{ijk}}\left(\theta_{ijk} - \beta_i\right)^2\right]$
$f_i(t_{ijk}^*\|\boldsymbol{x}_{ijk})$	$\frac{\lambda_{i0}(t_{ijk}^*)\kappa_{ijk}\alpha_i\beta_i}{\left[1+\beta_i\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)\right]^{\alpha_i+1}}$	$\frac{\nu_i\lambda_{i0}(t_{ijk}^*)\exp\left\{\frac{\alpha_i}{\beta_i}\left[1-\sqrt{1+2\beta_i^2\alpha_i^{-1}\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)}\right]\right\}\kappa_{ijk}\alpha_i\beta_i}{\sqrt{1+2\beta_i^2\alpha_i^{-1}\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)}}$
$S_i(t_{ijk}^*\|\boldsymbol{x}_{ijk})$	$\frac{1}{\left[1+\beta_i\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)\right]^{\alpha_i}}$	$\exp\left\{\frac{\alpha_i}{\beta_i}\left[1-\sqrt{1+2\beta_i^2\alpha_i^{-1}\kappa_{ijk}\Lambda_{i0}(t_{ijk}^*)}\right]\right\}$

Table 2
Distributional assumptions with regard to $\boldsymbol{\theta}_{jk}$ and $\boldsymbol{\upsilon}_k$ in the overdispersed frailty models described in equation (7).

	υ_{ik}	
θ_{ijk}	$\boldsymbol{\upsilon}_k = \mathbf{1}$	$\boldsymbol{\upsilon}_k \sim \log N_2(\mathbf{1}, \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}})$
(1) independence		
$\theta_{ijk} \sim \Gamma\left(\alpha_i, \alpha_i^{-1}\right)$	univariate gamma	univariate gamma-lognormal
$\theta_{ijk} \sim IG(\alpha_i, 1)$	univariate inverse Gaussian	univariate inverse Gaussian-lognormal
(2) shared		
$\theta_{jk} \sim \Gamma(\alpha, \alpha^{-1})$	shared gamma	shared gamma-lognormal
$\theta_{jk} \sim IG(\alpha, 1)$	shared inverse Gaussian	shared inverse Gaussian-lognormal
(3) correlated		
$\boldsymbol{\theta}_{jk} \sim \Gamma_2(\mathbf{1}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$	correlated gamma	correlated gamma-lognormal
$\boldsymbol{\theta}_{jk} \sim IG_2(\mathbf{1}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$	correlated inverse Gaussian	correlated inverse Gaussian-lognormal

Table 3

Estimates and standard errors between brackets for shared and correlated Gompertz-gamma and Gompertz-inverse Gaussian frailty models (upper part), univariate, shared and correlated Gompertz-gamma-lognormal frailty models (middle part) and Dirichlet-multinomial Gompertz-gamma frailty models (lower part) applied to bivariate serology on hepatitis A and B in Flanders, Belgium.

	Gompertz frailty models			
	Shared		Correlated	
Parameter	Gamma	Inverse Gaussian	Gamma	Inverse Gaussian
$\xi_1 \times 10^2$	1.219 (0.103)	1.346 (0.096)	0.668 (0.115)	1.363 (0.113)
$\nu_1 \times 10^2$	3.693 (0.461)	3.278 (0.367)	10.481 (1.686)	3.934 (0.588)
$\xi_2 \times 10^2$	0.172 (0.034)	0.173 (0.034)	0.178 (0.036)	0.175 (0.035)
$\nu_2 \times 10^2$	−0.023 (0.741)	−0.032 (0.745)	0.111 (0.930)	0.002 (1.241)
σ_{θ_1}	0.723 (0.084)	0.800 (0.133)	1.635 (0.175)	1.108 (0.281)
σ_{θ_2}	0.723 (0.084)	0.800 (0.133)	1.417 (1.167)	1.016 (2.849)
ρ_θ	1.000 (−)	1.000 (−)	0.557 (0.457)	0.763 (2.060)
$-2\log(L)$	5687.020	5690.547	5653.495	5688.325
AIC	5697.020	5700.547	5667.495	5702.325
BIC	5709.789	5713.317	5685.372	5720.202

	Gompertz-gamma-lognormal frailty models			
			Correlated	
Parameter	Univariate	Shared	Unrestricted	$\sigma_{\upsilon_1} = \sigma_{\upsilon_2}$
$\xi_1 \times 10^2$	0.785 (0.137)	1.272 (0.115)	0.633 (0.147)	0.612 (0.142)
$\nu_1 \times 10^2$	8.709 (1.819)	3.484 (0.478)	12.730 (3.298)	13.588 (2.618)
$\xi_2 \times 10^2$	0.165 (0.041)	0.168 (0.036)	0.163 (0.036)	0.163 (0.036)
$\nu_2 \times 10^2$	1.464 (3.540)	0.314 (0.766)	0.600 (0.874)	0.632 (0.876)
σ_{θ_1}	1.447 (0.210)	0.715 (0.087)	1.867 (0.300)	1.944 (0.226)
σ_{θ_2}	2.544 (3.461)	0.715 (0.087)	1.356 (0.629)	1.431 (0.613)
ρ_θ	0.000 (−)	1.000 (−)	0.677 (0.297)	0.664 (0.268)
σ_{υ_1}	0.122 (0.197)	0.147 (0.062)	0.346 (0.223)	0.415 (0.114)
σ_{υ_2}	0.529 (0.427)	0.400 (0.107)	0.426 (0.123)	0.415 (0.114)
ρ_υ	−0.590 (1.063)	−0.671 (0.432)	−0.626 (0.464)	−0.595 (0.413)
$-2\log(L)$	5674.394	5672.675	5641.146	5641.252
AIC	5692.394	5688.675	5661.146	5659.252
BIC	5715.379	5709.106	5686.685	5682.237

	Dirichlet-multinomial Gompertz-gamma frailty models		
Parameter	Univariate	Shared	Correlated
$\xi_1 \times 10^2$	0.766 (0.156)	1.286 (0.124)	0.655 (0.134)
$\nu_1 \times 10^2$	9.013 (2.007)	3.282 (0.532)	10.918 (2.059)
$\xi_2 \times 10^2$	0.158 (0.035)	0.163 (0.036)	0.169 (0.037)
$\nu_2 \times 10^2$	0.398 (4.077)	0.254 (0.783)	0.294 (0.908)
σ_{θ_1}	1.503 (0.222)	0.678 (0.102)	1.698 (0.208)
σ_{θ_2}	0.959 (12.515)	0.678 (0.102)	1.277 (1.066)
ρ_θ	0.000 (−)	1.000 (−)	0.611 (0.507)
ρ	0.093 (0.016)	0.092 (0.017)	0.078 (0.017)
$-2\log(L)$	5664.659	5672.014	5644.114
AIC	5678.659	5684.014	5660.114
BIC	5696.536	5699.337	5680.546

Table 4

Averaged parameter estimates (Mean), relative bias (Rel. Bias) and empirical standard errors (Emp. s.e.) for the correlated gamma, correlated Gompertz-gamma-lognormal and correlated Dirichlet-multinomial frailty model applied to 500 simulation sets of size N = 3787 under the Gompertz-gamma-lognormal frailty model.

Parameter	True Value		Gamma	Gamma-Lognormal	Dirichlet-Multinomial
$\xi_1 \times 10^2$	0.600	Mean	0.598	0.613	0.611
		Rel. Bias	−0.004	0.021	0.019
		Emp. s.e.	0.098	0.104	0.104
$\nu_1 \times 10^2$	2.000	Mean	1.816	2.209	1.914
		Rel. Bias	−0.092	0.105	−0.043
		Emp. s.e.	1.675	1.685	1.859
$\xi_2 \times 10^2$	0.200	Mean	0.190	0.196	0.194
		Rel. Bias	−0.050	−0.022	−0.031
		Emp. s.e.	0.040	0.036	0.038
$\nu_2 \times 10^2$	3.000	Mean	3.494	3.444	3.431
		Rel. Bias	0.165	0.148	0.144
		Emp. s.e.	1.398	1.019	1.321
σ_{θ_1}	1.900	Mean	1.729	1.895	1.793
		Rel. Bias	−0.090	−0.003	−0.057
		Emp. s.e.	0.692	0.641	0.723
σ_{θ_2}	1.400	Mean	1.646	1.585	1.640
		Rel. Bias	0.176	0.132	0.171
		Emp. s.e.	0.583	0.422	0.563
ρ_θ	0.700	Mean	0.697	0.709	0.697
		Rel. Bias	−0.004	0.013	−0.004
		Emp. s.e.	0.193	0.148	0.184
σ_{υ_1}	0.350	Mean	–	0.351	–
		Rel. Bias	–	0.003	–
		Emp. s.e.	–	0.162	–
σ_{υ_2}	0.450	Mean	–	0.491	–
		Rel. Bias	–	0.091	–
		Emp. s.e.	–	0.161	–
ρ_υ	−0.600	Mean	–	−0.630	–
		Rel. Bias	–	0.051	–
		Emp. s.e.	–	0.240	–
ρ	–	Mean	–	–	0.104
		Rel. Bias	–	–	–
		Emp. s.e.	–	–	0.017
τ_{WS}	0.321	Mean	0.300	0.322	–
		Emp. s.e.	0.054	0.047	–
τ_{BS}	0.301	Mean	0.300	0.304	–
		Emp. s.e.	0.054	0.043	–
CR	–		0.930	0.980	0.920
AIC %	–		0.000	0.982	0.018
BIC %	–		0.018	0.892	0.089

Table 5

Averaged parameter estimates (Mean), relative bias (Rel. Bias) and empirical standard errors (Emp. s.e.) for the shared gamma, shared gamma-lognormal and shared Dirichlet-multinomial frailty model applied to 500 simulation sets of size $N = 3787$ under the correlated gamma-lognormal frailty model.

Parameter	True Value		Gamma	Gamma-Lognormal	Dirichlet-Multinomial
$\xi_1 \times 10^2$	0.600	Mean	0.615	0.616	0.623
		Rel. Bias	0.026	0.027	0.038
		Emp. s.e.	0.069	0.069	0.069
$\nu_1 \times 10^2$	2.000	Mean	0.854	0.980	0.845
		Rel. Bias	−0.573	−0.510	−0.578
		Emp. s.e.	0.300	0.301	0.300
$\xi_2 \times 10^2$	0.200	Mean	0.219	0.215	0.223
		Rel. Bias	0.094	0.074	0.117
		Emp. s.e.	0.031	0.031	0.031
$\nu_2 \times 10^2$	3.000	Mean	2.098	2.247	2.069
		Rel. Bias	−0.301	−0.251	-0.310
		Emp. s.e.	0.330	0.333	0.326
σ_{θ_1}	1.400	Mean	0.765	0.826	0.784
		Rel. Bias	−0.454	−0.410	−0.440
		Emp. s.e.	0.060	0.059	0.059
σ_{θ_2}	1.400	Mean	0.765	0.826	0.784
		Rel. Bias	−0.454	−0.410	−0.440
		Emp. s.e.	0.060	0.059	0.059
ρ_θ	0.500	Mean	1.000	1.000	1.000
		Rel. Bias	–	–	–
		Emp. s.e.	–	–	–
σ_{υ_1}	0.350	Mean	–	0.258	–
		Rel. Bias	–	−0.262	–
		Emp. s.e.	–	0.057	–
σ_{υ_2}	0.450	Mean	–	0.351	–
		Rel. Bias	–	−0.221	–
		Emp. s.e.	–	0.071	–
ρ_υ	−0.600	Mean	–	−0.679	–
		Rel. Bias	–	0.132	–
		Emp. s.e.	–	0.215	–
ρ	–	Mean	–	–	0.111
		Rel. Bias	–	–	–
		Emp. s.e.	–	–	0.017
τ_{WS}	0.203	Mean	0.226	0.254	–
		Emp. s.e.	0.032	0.031	–
τ_{BS}	0.183	Mean	0.226	0.225	–
		Emp. s.e.	0.032	0.032	–
CR	–		1.000	1.000	1.000
AIC %	–		0.000	0.982	0.018
BIC %	–		0.004	0.916	0.080