On the Addams family of discrete frailty distributions for modeling multivariate case I interval-censored data

Peer-reviewed author version

Bardo, Maximilian; HENS, Niel & Unkel, Steffen (2024) On the Addams family of discrete frailty distributions for modeling multivariate case I interval-censored data. In: Biostatistics,.

DOI: 10.1093/biostatistics/kxae035
Handle: http://hdl.handle.net/1942/44307

Journal Title Here Vol. , No. , pp. 1–25
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

On the Addams family of discrete frailty distributions for modelling multivariate case I interval-censored data

Maximilian Bardo,[1,*] Niel Hens[2,3] and Steffen Unkel[1,4]

[1]Department of Medical Statistics, University Medical Center Göttingen, Street, Postcode, State, Germany, [2]I-BioStat, Data Science Institute, Hasselt University, Belgium, [3]Centre for Health Economics Research and Modelling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium and [4]Faculty V: School of Life Sciences, University of Siegen, Germany

[*]Corresponding author. maximilian.bardo@proton.me

FOR PUBLISHER ONLY Submitted Date Month Year; Revised Date Month Year; revised version accepted Date Month Year

ABSTRACT

Random effect models for time-to-event data, also known as frailty models, provide a conceptually appealing way of quantifying association between survival times and of representing heterogeneities resulting from factors which may be difficult or impossible to measure. In the literature, the random effect is usually assumed to have a continuous distribution. However, in some areas of application, discrete frailty distributions may be more appropriate. The present paper is about the implementation and interpretation of the Addams family of discrete frailty distributions. We propose methods of estimation for this family of densities in the context of shared frailty models for the hazard rates for case I interval-censored data. Our optimization framework allows for stratification of random effect distributions by covariates. We highlight interpretational advantages of the Addams family of discrete frailty distributions and the K-point distribution as compared to other frailty distributions. A unique feature of the Addams family and the K-point distribution is that the support of the frailty distribution depends on its parameters. This feature is best exploited by imposing a model on the distributional parameters, resulting in a model with non-homogeneous covariate effects that can be analysed using standard measures such as the hazard ratio. Our methods are illustrated with applications to multivariate case I interval-censored infection data.

KEYWORDS: Discrete distributions; Frailty; Heterogeneity; Infectious diseases; Multivariate survival data

1. INTRODUCTION

Multivariate time-to-event data are commonly encountered in the life sciences. An example is the time to occurrence of multiple non-lethal events within the same individual. In this setting, the individual can be thought of as forming a cluster in which event times are likely to be correlated. An alternative point of view is that there is heterogeneity across individuals (or clusters) due to characteristics that may be difficult or impossible to measure. Random effect (RE) models for time-to-event data, also known as frailty models, offer a conceptually appealing approach to quantify these associations within clusters and to model unobserved heterogeneity across individuals (or clusters) (Duchateau and Janssen, 2008; Hougaard, 2000; Wienke, 2010).

While the majority of the existing literature assumes continuous frailty distributions, such as the gamma ($\mathcal{G}$), inverse Gaussian, or log-normal, for some applications discrete frailty distributions may be more appropriate. One such application is the study of infectious diseases transmitted via close contact, where unobserved heterogeneity with respect to becoming infected with pathogens of interest may be represented and analysed by latent risk groups.

This paper discusses the interpretation of discrete frailty models with parameter dependent support. We propose interpreting the discrete frailties as ordered latent risk categories. We extend this analysis by allowing the parameters of the discrete frailty distribution to depend on covariates. The ordered nature of discrete frailty facilitates comparing latent risk categories within or across different strata (as defined by covariate values) using hazard ratios, along with probabilities of belonging to specific risk categories. This approach facilitates the analysis of non-homogeneous populations or non-homogeneous covariate effects, which might lead to a deeper understanding of the problem under study.

In this context we discuss the "Addams" family ($\mathcal{AF}$) of discrete frailty distributions introduced by Farrington et al. (2012). This family is homogeneous in the sense that it consists of discrete distributions, except for the continuous $\mathcal{G}$ distribution, which serves as a reference case of constant association within clusters over time. The discrete distributions within the $\mathcal{AF}$ are scaled variants of negative binomial, shifted negative binomial, binomial, and Poisson distributions. For this family, we investigate the hazard ratios conditional on risk group membership within or across different strata. Furthermore, we introduce estimation routines for the $\mathcal{AF}$ in the context of case I interval-censored data (Sun, 2007). Our optimization framework allows for stratification of frailty distributions along covariates. We apply the $\mathcal{AF}$ to data from a serological survey of human papillomaviruses (Mollema et al., 2009; Scherpenisse et al., 2012).

The structure of the paper is as follows: Section 2 presents the time-invariant shared frailty model and reviews the literature on discrete frailty models, comparing the $\mathcal{AF}$ to other discrete frailty distributions. We further discuss the interpretation of discrete frailty models for which the support is parameter dependent. Section 3 examines the conditional time-to-event model resulting from the $\mathcal{AF}$ of distributions.

Section 4 outlines the estimation framework and optimization algorithm. In Section 5, the $\mathcal{AF}$ is applied to multivariate serological survey data. Finally, Section 6 offers concluding remarks.

2. DISCRETE FRAILTY MODELS

Let i refer to a cluster, $i = 1, \ldots, n$, and let n be the number of clusters observed. Each cluster contains J units, and $T_i^{(j)} > 0$ denotes the time-to-event random variable (RV) for the j^{th} unit, $j = 1, \ldots, J$, within the i^{th} cluster. All indices are unique, such that $i \neq i'$ and $j \neq j'$. The random vectors $\boldsymbol{T}_i = [T_i^{(1)}, \ldots, T_i^{(J)}]^T$ and $\boldsymbol{T}_{i'}$ are independent given the covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$. Within a cluster, $T_i^{(j)}$ and $T_i^{(j')}$ are independent given the random and unobservable time-invariant cluster-specific frailty Z_i and covariates. The non-negative RV Z_i has density or probability mass function $g(z|\tilde{\boldsymbol{x}}_i) \equiv g_i(z)$, where the index i is relevant only if the distribution or its parameters depend on (a subset of) covariates denoted by $\tilde{\boldsymbol{x}}_i$. Technically, $\tilde{\boldsymbol{x}}_i$ could include unit-specific covariates. However, as Z_i is shared within the cluster i, $\tilde{\boldsymbol{x}}_i$ will typically also be shared within the cluster i, without containing covariates that differ across the units. For unit-specific covariates, correlated frailties (Hens et al., 2009) may be more appropriate, where the correlation parameter between the frailties of a cluster may depend on cluster-invariant covariates, while unit-specific frailty parameters may depend on unit-specific covariates. Given covariates, Z_i and $Z_{i'}$ are independent. We do not distinguish in language between the random Z_i and a realisation z_i and refer to both as frailty or RE.

The conditional hazard rate, is assumed to be of the form

$$\lambda_i^{(j)}(t|Z_i) = Z_i \exp\{\boldsymbol{x}_i^{(j)^T} \boldsymbol{\beta}^{(j)}\} \lambda_0^{(j)}(t), \tag{1}$$

with cluster- and unit-specific covariate vector $\boldsymbol{x}_i^{(j)}$, and parameters $\boldsymbol{\beta}^{(j)}$, as well as unit-specific baseline hazard rate $\lambda_0^{(j)}(t)$. Note that the covariates $\tilde{\boldsymbol{x}}_i$ are only implicitly included in the equation (1) through the distribution of Z_i, which may also include a potential expectation parameter. Also note that for brevity, the dependence of quantities, such as hazard rates or densities, on $\boldsymbol{x}_i^{(j)}$ or $\tilde{\boldsymbol{x}}_i$ is indicated by a superscript (j) or subscript i, respectively.

In much of the existing literature, Z_i is typically considered a continuous random variable, with common choices being the log-normal or $\mathcal{G}$ distributions. Nonetheless, discrete frailty distributions have also been explored for both univariate and multivariate data. A prominent choice is the K-point distribution (Palloni and Beltrán-Sánchez, 2017; Bijwaard, 2014; Begun et al., 2000; Gasperoni et al., 2020; Pickles and Crouchley, 1994; Choi and Huang, 2012; Choi et al., 2014; Troncoso-Ponce, 2018), i.e. a frailty distribution with (ordered) support parameters $\{z_{(1)}, \ldots, z_{(K)}\} \in \mathbb{R}^K_{\geq 0}$ and corresponding probability parameters $\mathrm{pr}_m = g_i(z_{(m)}) > 0$ for $m = 1, \ldots, K$, and $\sum_{m=1}^K \mathrm{pr}_m = 1$ (Wienke, 2010). Other choices are the binomial ($\mathcal{B}$) (Ata and Özel, 2013), negative binomial ($\mathcal{NB}$) (Ata and Özel, 2013; Caroni et al., 2010), geometric (Caroni et al., 2010; Cancho et al., 2021; Choi and Huang, 2012; Choi et al., 2014), Poisson ($\mathcal{P}$) (Ata and Özel, 2013; Caroni et al., 2010; Cancho et al., 2020b; Choi and Huang, 2012; Choi et al., 2014), and the hyper-Poisson distributions (Mohseni et al., 2020; de Souza et al., 2017). The framework of the zero-inflated and zero-modified power series (ZMPS) distributions, where the probability of $Z = 0$ is modified by an additional parameter as compared to the discrete reference distribution, has been investigated by Cancho et al. (2018, 2020a) and Molina et al. (2021), respectively. In particular, the zero-inflated geometric, $\mathcal{P}$ and logarithmic distributions and the zero-modified geometric and $\mathcal{P}$ distributions are investigated. The Addams family ($\mathcal{AF}$), as conceptualised by Farrington et al. (2012), includes (shifted and) scaled $\mathcal{NB}$, as well as scaled $\mathcal{B}$ and $\mathcal{P}$ distributions, and the $\mathcal{G}$ distribution as a continuous special case.

Modelling a whole family of distributions, such as the $\mathcal{AF}$, is preferable to modelling a given (discrete) distribution, such as the $\mathcal{P}$, since the latter strategy typically severely limits the patterns of heterogeneity as measured by the relative frailty variance ($\mathrm{RFV}(\boldsymbol{t}) = \frac{\mathrm{Var}(Z|\boldsymbol{T}>\boldsymbol{t})}{\mathrm{E}(Z|\boldsymbol{T}>\boldsymbol{t})^2}$, $\boldsymbol{t} = [t^{(1)}, \ldots, t^{(J)}]^T$) to either monotonically increasing or monotonically decreasing trajectories over time; for examples see Farrington et al. (2012) and Bardo and Unkel (2023). It can be shown that the long-term trajectory of the RFV (or association, as measured by the cross-ratio function [$\mathrm{CRF}(\boldsymbol{t}) = 1 + \mathrm{RFV}(\boldsymbol{t})$]) is determined with positive probability by the smallest value of the frailty distribution ($z_{(1)}$). Specifically, if $z_{(1)} = 0$, RFV and CRF approach infinity as time approaches infinity. Conversely, if $z_{(1)} > 0$, the RFV approaches zero and the CRF approaches one. Therefore, the choice of a discrete frailty distribution has an enormous impact on the model, even if $g_i(0) > 0$ or $z_{(1)} > 0$ is very small, and even more so if the model's trajectory of RFV is monotone. The $\mathcal{AF}$ achieves greater flexibility in the trajectory of heterogeneity and association by incorporating discrete frailty distributions for which $z_{(1)} = 0$ and frailty distributions for which $z_{(1)} > 0$, and is thus able to induce increasing and decreasing trajectories of the RFV (CRF). Thus, in the case of $\mathcal{AF}$, the trajectory of the RFV can be informed by the data in the fitting process. This is a rare property among frailty distributions, and for the discrete distributions mentioned above it is only possible for the K-point distribution and the ZMPS. However, in these cases the decision between a decreasing or increasing long-term trajectory lies on the edge of the parameter space (Bardo and Unkel, 2023), which is not the case for the $\mathcal{AF}$. As optimising on the edge of the parameter space is usually difficult, this is an advantage when using the $\mathcal{AF}$. Note, however, that the K-point distribution and the ZMPS are able to induce non-monotone trajectories of the RFV (CRF), which is not the case for the $\mathcal{AF}$. We will discuss the $\mathcal{AF}$ and its special cases in Section 3.

From an interpretive point of view, the $\mathcal{AF}$ and K-point distributions offer a unique perspective because the support is also subject to estimation. The support of other discrete frailty distributions is usually the natural number including zero ($\mathbb{N}_{\geq 0}$). Therefore, the interpretation of discrete frailty models often focuses on the cure rate, i.e. those who are not susceptible to the event of interest (see e.g. de Souza et al. (2017); Cancho et al. (2018, 2020a,b); Mohseni et al. (2020); Cancho et al. (2021); Molina et al. (2021)). Subject-related interpretations are also common. Caroni et al. (2010) suggest interpreting discrete frailties as the unobservable number of flaws in a unit or exposure to damage on an unknown number of occasions. Similar interpretations can be found in Ata and Özel (2013) for time-to-event models on earthquake data. Both studies consider discrete frailty distributions with support on $\mathbb{N}_{\geq 0}$. However, due to the latent nature of the frailty, such concrete interpretations are difficult because the hazard ratio (HR) of, say, two events versus one event would be fixed by the model structure at HR $= 2$ if the support is fixed at $\mathbb{N}_{\geq 0}$. However, if the true HR is less than two, the probability mass of frailty could be shifted to the right relative to the distribution of the number of hits. Consequently, a more abstract interpretation such as the "effective" number of harms would be more appropriate. The K-point frailty distributions are often interpreted as representing sub-populations, such as unobservable carriers of certain disease genotypes (e.g., Pickles and Crouchley (1994); Begun et al. (2000); Wienke (2010); Bijwaard (2014)). Palloni and Beltrán-Sánchez (2017) interpret a delayed binary frailty as the effect of adverse early life conditions on adult mortality. Using the K-point distribution, where K is also subject to estimation, Gasperoni et al. (2020) interpret the discrete frailties as (an unknown number of) latent sub-populations. They suggest calculating HRs between these sub-populations, which is interpretable as the support being subject to estimation.

In the present paper, we endorse the interpretation of discrete frailties as latent subpopulations and expand upon it. Note that we focus on discrete frailty distributions for which the support Ω is parameter dependent, which is mainly the case for the K-point distribution and, although more restricted, the $\mathcal{AF}$, as will be seen in Section 3. Let $\Omega_i = \{z_{i,(1)}, z_{i,(2)}, \dots\}$, with $z_{i,(1)} \geq 0$ and $z_{i,(k)} < z_{i,(k+1)}$, represent the support of discrete RV Z_i, where the distribution parameters might depend on $\tilde{\boldsymbol{x}}_i$. We define that $\tilde{\boldsymbol{x}}_i$ constitutes a stratum of the population. We also consider $z_{i,(k)}$ as the conditional hazard-determining value for an individual in the k^{th} risk category (RC) within the stratum which is defined by $\tilde{\boldsymbol{x}}_i$.

For discrete frailty distributions for which the support depends on parameters, the within-stratum hazard ratio, $\text{HR}_\text{W}(k) = \frac{\lambda_i^{(j)}(t|Z_i = z_{i,(k+1)})}{\lambda_i^{(j)}(t|Z_i = z_{i,(k)})} = \frac{z_{i,(k+1)}}{z_{i,(k)}}$, might be analysed. The $\text{HR}_\text{W}(k)$ compares the hazard of the k^{th} and $(k+1)^{\text{th}}$ RC of stratum i. This is in line with Gasperoni et al. (2020), except for the presence of different strata for the distribution of the frailty, i.e. the HR_W might differ for $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_{i'}$. Moreover, due to the presence of different strata for the frailty distribution, an across-stratum analysis can be conducted by computing the across-stratum hazard ratio, $\text{HR}_\text{A}(k) = \frac{\lambda_i^{(j)}(t|Z_i = z_{i,(k)})}{\lambda_{i'}^{(j)}(t|Z_{i'} = z_{i',(k)})} = \frac{z_{i,(k)}}{z_{i',(k)}}$, for $\tilde{\boldsymbol{x}}_i \neq \tilde{\boldsymbol{x}}_{i'}$, and equality in the remaining covariates $\boldsymbol{x}_i^{(j)}$ and $\boldsymbol{x}_{i'}^{(j)}$. If the support between stratum i and i' is very different, it might be more desirable to analyse $\text{HR}_\text{A}(k; k') = \frac{z_{i,(k)}}{z_{i',(k')}}$ for all k' for which $P(Z_{i'} \leq z_{i',(k')}) \in [P(Z_i \leq z_{i,(k-1)}), P(Z_i \leq z_{i,(k)}))$ or the closest quantiles of $Z_i, Z_{i'}$ if no such k' exists. In order to put the analysis of the HR_W and HR_A into context, they should always be accompanied by reporting the distribution of the RCs. This allows for a separate but accompanying analysis of the distribution of individual heterogeneity via the distribution of RCs and the magnitude related impact of the RCs on the hazard rates.

The approach of imposing a model on the distribution parameters of the frailty has some similarity to generalized additive models for location, scale, and shape parameters for the population time-to-event distribution, i.e. for the time-to-event distribution with the frailty marginalized out. However, modelling the determinants of the randomness of hazard rates via covariates might be more intuitive, as the hazard rates are usually the standard approach for modelling time-to-event data. This approach is not new per se. It can be found, for example, in Aalen et al. (2008) and is also quite common in the field of discrete frailty modelling (Choi and Huang, 2012; Choi et al., 2014; Molina et al., 2021; Cancho et al., 2018, 2020a), and random slopes could also be interpreted in this way. What is new, however, is the type of analysis with within-stratum and across-stratum hazard ratios. Note that this type of analysis has no counterpart for continuous frailty distributions, as there is no k^{th} RC. Nor does such an analysis make sense for discrete frailty models where the support of the frailty distribution is set by assumption, e.g. to $\mathbb{N}_{\geq 0}$, as this would fix the HR_W and HR_A by assumption. A tempting alternative in such a case might be to calculate the hazard ratio between, say, the third and the first quartile of the frailty distribution. However, the quartiles of the frailty distribution vary over time due to selection effects, which is particularly important for hazard ratios which inherit the survival condition. Therefore, the hazard ratio is either meaningless if the quartiles at $t = 0$ are chosen and maintained throughout the analysis, or the analysis is more complicated if the time-varying quartiles of the frailty distribution are chosen.

Therefore, an analysis via the HR_W and HR_A is unique to discrete frailty models, where the support of the frailty distribution varies with (a subset of) distribution parameters that may depend on stratum membership. In this case, the data inform the support of the frailty distribution by likelihood criteria making it suitable to represent latent RCs. This allows the analysis of non-homogeneous covariate effects using common measures such as HRs as described above and probabilities of belonging to a particular RC. This might in particular be helpful in communicating the results of heterogeneous (covariate) effects to an audience outside of statistics such as medical doctors.

3. THE ADDAMS FAMILY OF DISCRETE FRAILTY DISTRIBUTIONS

Let $\Lambda_0^{(j)}(t)$ denote the cumulative baseline hazard rate $\int_0^t \lambda_0^{(j)}(u)du$. Moreover, $\Lambda_i(\boldsymbol{t}) = \sum_{j=1}^J \exp\{\boldsymbol{x}_i^{(j)^T}\boldsymbol{\beta}^{(j)}\}\times\, \Lambda_0^{(j)}(t^{(j)})$. We further ignore that the parameters of the frailty distribution may depend on $\tilde{\boldsymbol{x}}_i$ in the first part of this section and come back to this issue in the latter part of this section.

The RFV that induces the $\mathcal{AF}$ equals $\text{RFV}(\boldsymbol{t}) = \gamma \exp\{\alpha\mu\Lambda_i(\boldsymbol{t})\}$, with $\gamma, \mu = \text{E}(Z) \in \mathbb{R}_{>0}$, and $\alpha \in \mathbb{R}$. As shown in Farrington et al. (2012), the Laplace transform $\mathcal{L}(s) = \int_0^\infty \exp\{-zs\}g(z)dz$ of the $\mathcal{AF}$ equals

$$\mathcal{L}(s) = \begin{cases} \left(\left(1 - \frac{\gamma}{\alpha}\right)\exp\{-\alpha\mu s\} + \frac{\gamma}{\alpha}\right)^{\frac{1}{\alpha-\gamma}} & \text{if } \alpha \neq \gamma, \alpha \neq 0, \\ \exp\left\{\frac{1}{\gamma}\big(\exp\{-\gamma\mu s\} - 1\big)\right\} & \text{if } \alpha = \gamma, \\ (1 + \gamma\mu s)^{-\frac{1}{\gamma}} & \text{if } \alpha = 0. \end{cases} \tag{2}$$

Note that μ may be set to one for identification purposes. Hence, the $\mathcal{AF}$ is a two parameter distribution. We will retain the parameter μ in our notation, as we will explicitly model it via $\mu(\tilde{\boldsymbol{x}}) = \exp\{\tilde{\boldsymbol{x}}^T \boldsymbol{\beta}^{(0)}\}$, where $\boldsymbol{\beta}^{(0)} \in \mathbb{R}$ is an additional vector of parameters. In a setting where the distribution of the frailty is not the focus of the analysis, $\boldsymbol{\beta}^{(0)}$ is typically interpreted as the log-proportional hazard ratio.

Table 1. RFV and distribution parameters of the Addams family and support.

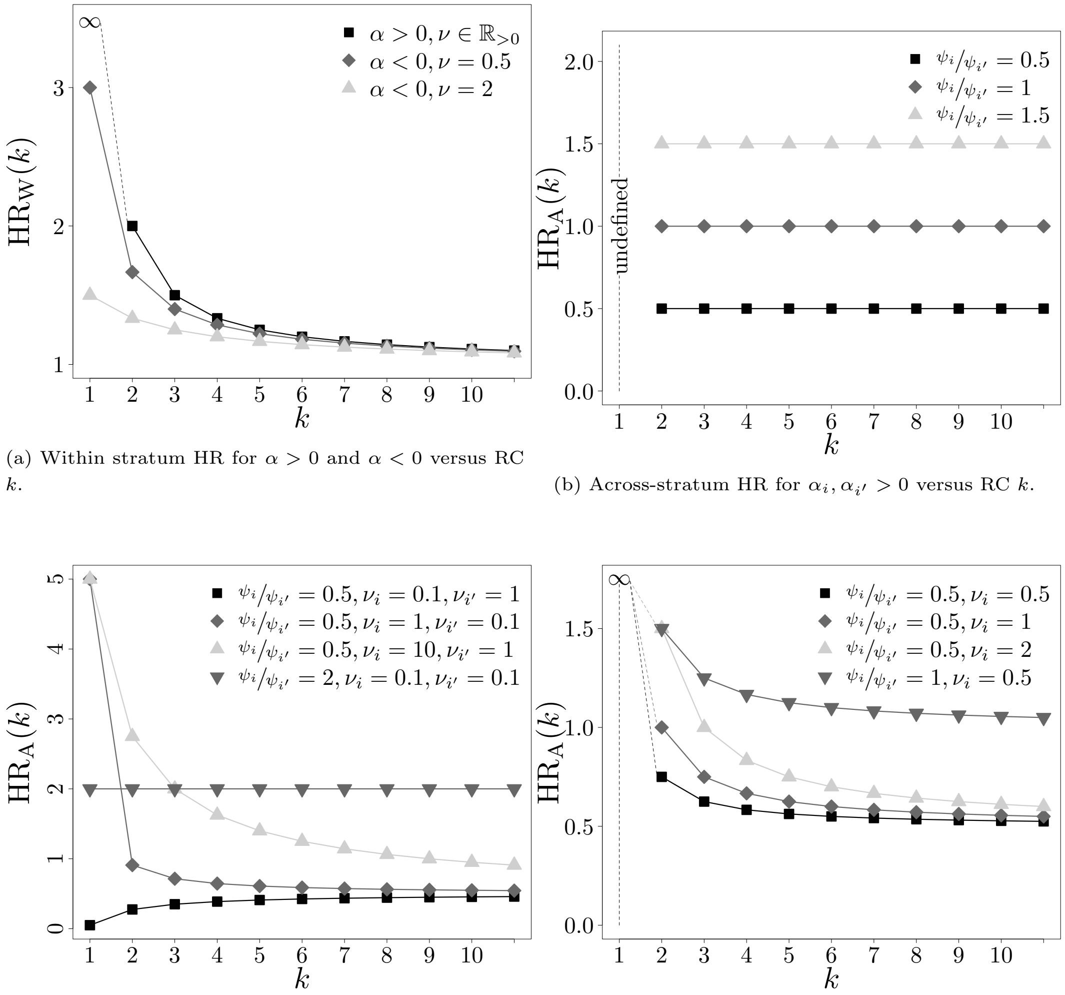
Parameters	$Z \sim$	Distribution Parameters	Support
$\gamma > 0 > \alpha$	$\psi\mathcal{NB}_{>0}(\nu, \pi)$	$\nu = \frac{1}{\gamma-\alpha}$ (number of successes), $\pi = \frac{-\alpha}{\gamma-\alpha}$ (success probability)	$\psi\times$ $\{\nu, 1+\nu, 2+\nu, \ldots\}$
$\gamma > 0 = \alpha$	$\mathcal{G}(\gamma^{-1}, \gamma^*)$	γ^{-1} (shape), $\gamma^* = (\mu\gamma)^{-1}$ (rate)	$\mathbb{R}_{>0}$
$\alpha = \gamma > 0$	$\psi\mathcal{P}(\lambda^*)$	$\lambda^* = \gamma^{-1}$ (rate)	$\psi \times \{0, 1, 2 \ldots\}$
$\gamma > \alpha > 0$	$\psi\mathcal{NB}(\nu, \pi)$	$\nu = \frac{1}{\gamma-\alpha}$ (number of successes), $\pi = \frac{\alpha}{\gamma}$ (success probability)	$\psi\times$ $\{0, 1, 2, \ldots\}$
$\alpha > \gamma > 0$	$\psi\mathcal{B}(b, \pi)$	$b = (\alpha - \gamma)^{-1}$ (number of trials), $\pi = \frac{\alpha-\gamma}{\alpha}$ (success probability)	$\psi\times$ $\{0, 1, \ldots, b\}$

The Laplace transform uniquely determines the distribution of Z, as shown in Table 1 (Farrington et al., 2012). The $\mathcal{AF}$ consists of different scaled and possibly shifted discrete distributions. The scaling parameter ψ of the corresponding discrete distributions is equal to $\mu|\alpha|$, and the parameter α selects one of the distributions of the $\mathcal{AF}$. If $\alpha < 0$, the scaled and shifted negative binomial ($\psi\mathcal{NB}_{>0}$) is chosen, where the support is shifted to the right by the parameter ν, resulting in a model without a latent non-susceptible sub-population. If $\alpha > 0$, a frailty distribution is chosen that includes a non-susceptible latent sub-population. For $\gamma > \alpha > 0$, the non-shifted scaled negative binomial ($\psi\mathcal{NB}$) and for $\alpha = \gamma$ the scaled Poisson ($\psi\mathcal{P}$) distribution is selected. In the case of $\alpha > \gamma$, the frailty distribution is scaled binomial ($\psi\mathcal{B}$), resulting in a model with an upper bound of the frailty that limits the maximum deviation of RC-conditioned survival curves between the susceptible latent sub-populations and a finite number of RCs including a non-susceptible group. Note that for the case $\alpha > \gamma$, $(\alpha - \gamma)^{-1}$ has to be an integer in order for $\mathcal{L}(s)$ to be a valid Laplace transform (Farrington et al., 2012). The continuous exception in the $\mathcal{AF}$ is the $\mathcal{G}$ distribution, which results from $\alpha = 0$.

The parameter α plays a unique role in the context of discrete shared frailty modelling. For discrete shared frailty models the RFV (CRF) either approaches zero (one) or infinity (Bardo and Unkel, 2023). Therefore, it is desirable to have a continuous exception within a family of discrete shared frailty distributions for which the RFV (CRF) does not approach zero (one) or infinity with time approaching infinity. Within the context of the $\mathcal{AF}$ this is the $\mathcal{G}$ which arises for $\alpha = 0$. In that case the RFV (CRF) is constant, a shape that is impossible for a discrete shared frailty model to generate. The nested structure might be utilized to test for a constant RFV (CRF) within the $\mathcal{AF}$.

Moreover, the $\mathcal{AF}$ can choose between a monotonically decreasing or increasing trajectory of the RFV (CRF) through the sign of α. This is unprecedented in the context of discrete shared frailty models. Though the ZMPS distribution is able to create decreasing trajectories of the RFV (CRF), this involves the edge of the parameter space for its deflation/inflation parameter. The opposite is true for the K-point distribution: its RFV (CRF) approaches zero (one) in the long run unless $z_{(1)}$ is equal to zero, which again involves the edge of the parameter space. For the $\mathcal{AF}$, the parameter α chooses between a decreasing or increasing RFV (CRF) by determining the support of the discrete distribution without involving the edge of parameter space. If $\alpha > 0$, $z_{(1)} = 0$ and a cure fraction exists. This induces an increasing trajectory of the RFV (CRF). If $\alpha < 0$ instead, $z_{(1)} > 0$ and no cure fraction exists. This induces a decreasing trajectory of the RFV (CRF); see Bardo and Unkel (2023) for a discussion of the shape of the RFV (CRF) for discrete frailty models.

The feature of the support being dependent on the distribution parameters offers the possibility of a meaningful interpretation of the $\mathrm{HR_W}$. Figure 1a shows examples of the $\mathrm{HR_W}$ for $\alpha > 0$ and $\alpha < 0$. If $\alpha > 0$, $\mathrm{HR_W}(1) \equiv \infty$ and $\mathrm{HR_W}(k) = \frac{k}{k-1}$ for $k \geq 2$. Note that there may be an upper bound on the RCs if $\psi\mathcal{B}$ is chosen. In this case, the upper bound of the frailty as well as the number of RCs chosen through the fitting procedure may be the main component of the analysis, e.g. by comparing the RC-related survival curve of the upper bound versus another RC. If $\alpha < 0$, $\mathrm{HR_W}(k) = \frac{\nu+k}{\nu+k-1}$ which approaches $\frac{k}{k-1}$ for large k. So there is reasonable flexibility for the first few within-stratum HRs and the focus may be on $\mathrm{HR_W}(1)$, where the model is more flexible than for later RCs. This is less flexible than the K-point distribution is, provided that K is large enough, which may even show a non-monotone trajectory of the $\mathrm{HR_W}$. However, modelling the K-point distribution with a large K is difficult given that this involves $2(K-1)$ parameters for a latent distribution, especially if the cluster size is small. On the contrary, within the $\mathcal{AF}$ one is remunerated with $|\Omega| = \infty$ (except for the $\psi\mathcal{B}$ case) which can be important for extreme observations where events occur very early in the lifespan. However, the parametric constraints of the $\mathcal{AF}$ on the trajectory of HR_W should always be taken in consideration and challenged by the K-point distribution whenever possible.



(a) Within stratum HR for $\alpha > 0$ and $\alpha < 0$ versus RC k.

(b) Across-stratum HR for $\alpha_i, \alpha_{i'} > 0$ versus RC k.

(c) Across-stratum HR for $\alpha_i, \alpha_{i'} < 0$ versus RC k.

(d) Across-stratum HR for $\alpha_i < 0, \alpha_{i'} > 0$ versus RC k.

Fig. 1. Within- and across-stratum hazard ratios versus RC k for various members of the Addams family.

This analysis can be extended to the across-stratum HR, HR_A, if there is a model for the distribution of individual heterogeneity within the $\mathcal{AF}$. For that purpose, we allow the parameters to depend on a stratum-specifying set of covariates $\tilde{\boldsymbol{x}}$, which determine the parameters of individual heterogeneity $\alpha(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}^T\boldsymbol{\zeta}, \gamma(\tilde{\boldsymbol{x}}) = \exp\{\tilde{\boldsymbol{x}}^T\kappa\}$, with each parameter in the vectors $\boldsymbol{\zeta}, \boldsymbol{\kappa}$ being an element of $\mathbb{R}$. For the sake of brevity, we denote $\alpha(\tilde{\boldsymbol{x}}_i), \gamma(\tilde{\boldsymbol{x}}_i)$, and $\mu(\tilde{\boldsymbol{x}}_i)$ by α_i, γ_i, and μ_i, respectively. Figure 1 shows the HR_A for varying scenarios of α_i and $\alpha_{i'}$. For $\alpha_i, \alpha_{i'} > 0$ (Figure 1b), the $\text{HR}_\text{A} = \frac{\psi_i}{\psi_{i'}}$ for all $k \geq 2$ and is undefined for $k = 1$. However, for $\alpha_i, \alpha_{i'} < 0$ (Figure 1c), the $\text{HR}_\text{A}(k) = \frac{\psi_i(\nu_i+k-1)}{\psi_{i'}(\nu_{i'}+k-1)}$, which approaches a constant ratio $\text{HR}_\text{A}(k) = \frac{\psi_i}{\psi_{i'}}$ for large k. Note that the HR_A might be greater or less than one for all k, but can also cross the threshold of one with increasing k. If $\tilde{\boldsymbol{x}}$ is for example an experimental treatment indicator (in a univariate context), the HR_A represents a heterogeneous treatment effect which might identify sub-groups within the population for which the treatment is harmful. For $\alpha_i < 0, \alpha_{i'} > 0$ (Figure 1d), $\text{HR}_\text{A}(1) \equiv \infty$, as for stratum i' there is a latent sub-population that is not susceptible to the event of interest, whereas for stratum i all latent sub-populations are susceptible. For $k \geq 2$ and $\text{HR}_\text{A}(k) = \frac{\psi_i(\nu_i+k-1)}{\psi_{i'}(k-1)}$. Another scenario, not explicitly shown in Figure 1b and 1d, is that one or both strata (i and i') might have (different) upper bounds of frailty in the $\psi\mathcal{B}$ case. In such a case, the stratum with the larger upper bound (which could still be ∞), say i', could be considered more vulnerable, since that stratum has a higher proportion of individuals who are expected to experience the event very early, namely those with a frailty value greater than $\psi_i b_i$.

Note that the parameters in the formula of HR_A and HR_W (and hence the parameters as specified in the legends of Figure 1) do not uniquely identify the parameters of the frailty distribution, i.e. for a given trajectory of HR_A, HR_W there is an infinite set of (α, γ) or (α_i, γ_i) and $(\alpha_{i'}, \gamma_{i'})$, respectively, that induce the same trajectory but with a different distributions of the RCs which did not need to be specified for Figure 1. This shows that the analysis of the frailty model has always two branches. The first branch is the analysis of HRs, which indicate the meaning of being in a particular latent RC (in a particular stratum) relative to another latent RC or to another observable stratum in the same latent RC. On the one hand, HR_W can help to assess the importance of individual heterogeneity, e.g. if $\text{HR}_\text{W}(k)$ is large, then latent RC membership has a large effect on expected survival. On the other hand, comparing HR_W across strata or analysing the HR_A gives an account of random covariate effects where, e.g., covariates with a beneficial effect on survival or covariates with partly beneficial, partly detrimental effects might be detected. The second branch of the analysis is the distribution of RCs across strata, which may indicate differences in the distribution of risk-taking behaviour and predisposition across strata, e.g. by indicating a stratum with a heavier tail of vulnerable RCs. Taken together, the analysis of HRs and the distribution of RCs can provide thorough analytical explanations in terms of selection and random covariate effects that can help to explain the trajectories of population survival curves (where the RCs are marginalised out), i.e. explanations for why the survival curves of two strata come closer or even cross over time.

4. ESTIMATION

In this section, all time-dependent quantities are evaluated at the monitored (censored or uncensored) event times of the individuals. We delete the argument from the expression and indicate the corresponding quantity with a subscript, e.g. $\Lambda_i^{(j)} = \exp\{\boldsymbol{x}_i^{(j)^T}\boldsymbol{\beta}^{(j)}\}\Lambda_0^{(j)}(t_i^{(j)})$. Furthermore, let $A \in \mathbb{P}(\{1,\ldots,J\})$, where $\mathbb{P}$ denotes the power set. Then, $\Lambda_i^{(A)} = \sum_{j\in A}\Lambda_i^{(j)}$ and $\Lambda_i^{(-A)} = \sum_{j\notin A}\Lambda_i^{(j)}$. Note that we define $\Lambda_i^{(\emptyset)} = 0$ and $\Lambda_i = \Lambda_i^{(1,\ldots,J)}$.

We develop estimation routines for case I interval-censored data. In the case of case I interval-censored data it is only known whether the event occurred during follow-up or not but the exact event time is unknown. For multivariate cases ($J > 1$) it is easier to understand the likelihood if one starts by exploiting the conditional independence assumption of $T_i^{(j)}$ and $T_i^{(j')}, j \neq j'$, given $Z_i = z$:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\lambda}_0, \boldsymbol{\beta}; \text{data}) &= \prod_{i=1}^{n}\int_0^\infty \prod_{j=1}^{J}(1-\exp\{-z\Lambda_i^{(j)}\})^{d_i^{(j)}}\exp\{-z\Lambda_i^{(j)}\}^{1-d_i^{(j)}}g_i(z)dz \\ &= \prod_{i=1}^{n}\int_0^\infty \sum_{A\in\mathbb{P}(d_i)}(-1)^{|A|}\exp\{-z_i(\Lambda_i^{(A)}+\Lambda_i^{(-d_i)})\}g_i(z)dz \\ &= \prod_{i=1}^{n}\sum_{A\in\mathbb{P}(d_i)}(-1)^{|A|}\mathcal{L}(\Lambda_i^{(A)}+\Lambda_i^{(-d_i)}). \end{aligned} \tag{3}$$

where $d_i^{(j)}$ is the observational unit and target-specific event indicator (equal to one if the event occurred during the follow-up, zero otherwise), and d_i is the set of targets on which the i^{th} observational unit had an event. The vector $\boldsymbol{\lambda}_0$ contains all parameters of the baseline hazard rates, $\boldsymbol{\beta} = [\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(J)}]$, and $\boldsymbol{\theta} = [\boldsymbol{\zeta}, \boldsymbol{\kappa}]$.

Quasi-Newton optimization routines were applied for optimizing the corresponding log-likelihood based on (3). We choose BFGS as the standard method, as implemented in R version 4.2.2 (R Core Team, 2022). Standard errors (SE) are obtained via the Hessian of the log-likelihood, which is approximated by Richardson extrapolation as implemented by Gilbert and Varadhan (2019). The delta method was applied where necessary to obtain SE: confidence intervals (CIs) are based on $\ln$ or $\ln\{-\ln\}$ transformations if the parameter of interest is greater than zero or between zero and one respectively, and are then transformed back to the scale of interest.

We provide algorithms that are able to fit univariate and multivariate frailty models for case I interval-censored data. The frailty distributions can be stratified by a (multi-level) factor. The frailty distribution might either be $\mathcal{AF}$ or from the power variance family (both parameters can be estimated). The baseline hazard can be chosen to be piecewise-constant or the parametric generalized gamma distribution (Cox et al., 2007) or one of its special cases, respectively. Covariates can be added in proportional hazards manner. Overdispersion parameters might be added by means of the Dirichlet compound multinomial distribution. Implementations are available on GitHub (https://github.com/time-to-MaBo/Addamsfamily/).

5. APPLICATIONS

We illustrate the $\mathcal{AF}$ in the context of multivariate case I interval-censored data on the human papillomavirus (HPV), obtained from a serological survey in the Netherlands (PIENTER-2); see Mollema et al. (2009) for details on PIENTER-2 and Scherpenisse et al. (2012) for an investigation of the respective HPV dataset. The data were collected in the years 2006 and 2007 and cover people aged 0 to 79. Participants were asked to complete a questionnaire and to provide a blood sample (Mollema et al., 2009). By means of the blood samples, the level of antibodies regarding the high-risk HPV types 16, 18, 31, 33, 45, 52, and 58 were determined in order to detect past infections. Therefore, at the time of observation, it is only known whether the study participants have had an infection in the past or not, but it is never known exactly when the potential infection occurred, resulting in case I interval-censored data, also known as current status data (Sun, 2007). Note that at the time of data collection the Dutch national immunization programme did not include a vaccine against HPV.

We analysed the nationwide sample including oversampled migrants and applied weighting factors to make the sample representative for the Dutch population. We excluded individuals in their first year of life from the analysis, as maternal antibodies could be transmitted to the infant transplacentally or through breastfeeding (Rintala et al., 2005). This left us with a sample size of $n = 6384$ individuals aged 2 to 80. The weighted proportion of females in the dataset is 49.9% (unweighted: 54.4%).

The observed time is the individuals' age at the date of serological monitoring. The event indicator $d_i^{(j)} = 1$ means that individual i is seropositive with respect to pathogen j, $j \in \{\mathrm{HPV16}, \mathrm{HPV18}, \mathrm{HPV31}, \mathrm{HPV33}, \mathrm{HPV45}, \mathrm{HPV52}, \mathrm{HPV58}\}$, and seronegative and still susceptible if $d_i^{(j)} = 0$. Seroprevalence is interpreted as a proxy for past infections. Note, however, that there is a time-lag between the infection and the time of seroconversion, as well as a difference in the number of individuals who were infected with HPV and those who seroconverted: in previous studies, antibodies could not be detected for about 20-50% of females who were carriers of HPV DNA. However, antibody responses are relatively stable over time and hence, the study of the population's seroprevalence might yield important insights; see Scherpenisse et al. (2012) for a discussion.

We consider the following models for the individual hazard rates, $\lambda_i^{(j)}(t) = Z_i \lambda_0^{(\mathrm{sex}:j)}(t)$, sex $\in \{m, f\}$, where the target-specific baseline hazard $\lambda_0^{(\mathrm{sex}:j)}(t)$ is either sex-stratified (sex-stratified baseline hazard model), or non-stratified $\lambda_0^{(\mathrm{sex}:j)}(t) = \lambda_0^{(j)}(t)$ (non-stratified baseline hazard model). The purpose of stratifying baseline hazards by sex is twofold. The first is to investigate whether it is justified to estimate baseline hazards jointly for both sexes, and the second is to investigate whether a potential difference in the distribution is better explained by stratified baseline hazards than by different distributions of individual heterogeneity. In any case the target (and potentially sex) specific hazard rate is piecewise constant with a unique parameter within the intervals $[0; 5)$, $[5; 10)$, $[10; 20)$, $[20; 30)$, $[30; 40)$, $[40; 50)$, $[50; 65)$, $[65; 80)$. The frailty Z is either sex-stratified, i.e. $Z_i \sim \mathcal{AF}(\alpha_{\mathrm{sex}_i}, \gamma_{\mathrm{sex}_i})$ (sex-stratified RE model), or non-stratified, i.e. $Z_i \sim \mathcal{AF}(\alpha, \gamma)$ (non-stratified RE model). Note that in both cases $\mu_m \equiv 1$, $\mu_f \in \mathbb{R}_{>0}$ except for the stratified-hazard model where μ_f is also set to one for the sake of identifiability. The stratified RE model might be better able to reflect differing patterns in individual heterogeneity due to biological and environmental predisposition as well as a different distribution of risk-related behaviour across males and females. We combine the stratification status of the baseline hazard with the stratification status of the RE.

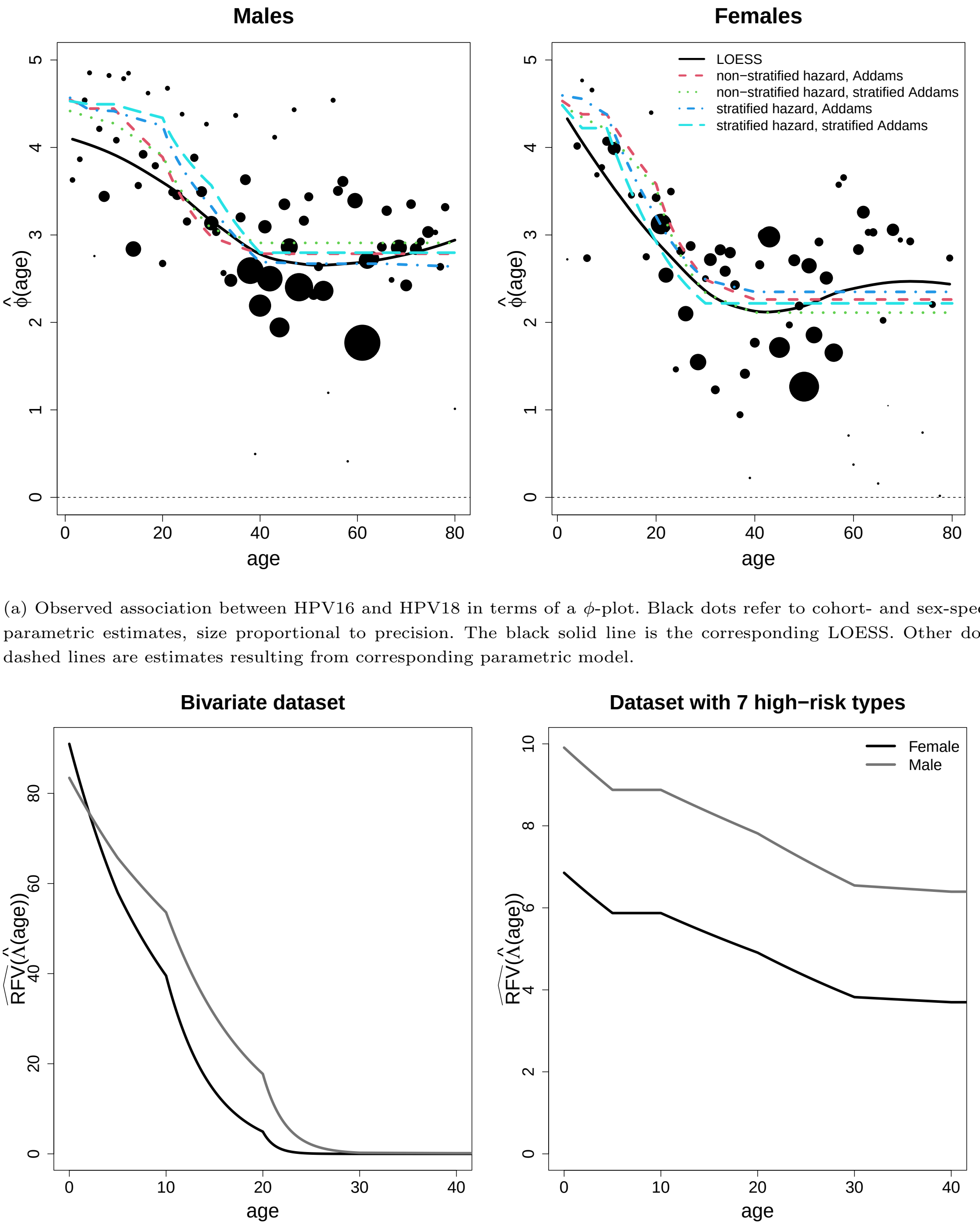
An HPV infection can be transmitted via skin-to-skin contact, often - though not exclusively (see, e.g., Syrjänen (2010), Rintala et al. (2005) or Meyers et al. (2014)) - via sexual intercourse (Gavillon et al., 2010). Therefore, individual – and typically unobserved – behaviour is an important determinant of an individual's risk of contracting HPV. Frailty models have been used previously to incorporate unobservable individual heterogeneity in the transmission of infectious diseases (see, e.g., Unkel et al. (2014) or Hens et al. (2009)). Moreover, we suspect that there may be distinct jumps in the individual hazard rates due to differences in individual behaviour that are relevant for transmission, e.g. comparing individuals who have no sex at all to individuals who have (see, e.g., Richardson et al. (2000), Burchell et al. (2006)), or whether the individuals use condoms or not (see, for example, Lam et al. (2014) or Nielson et al. (2010)). More formally, non-Gaussian and discrete omitted covariates may be the most important drivers of individual heterogeneity, and thus a discrete frailty model may be particularly appropriate here.

We start with a bivariate analysis including HPV16 and HPV18. An extension to higher variate data with data on seven types of HPV follows.

5.1. Bivariate Data Analysis

We begin by exploiting the nested structure of the models for model selection. The stratification of the RE is statistically significant on conventional levels by means of a likelihood ratio test (LRT) no matter the stratification status of the baseline hazard. The null-hypothesis is H_0: $Z_m \sim Z_f \sim \mathcal{AF}(\alpha, \gamma)$ vs. $H_1 : Z_{\text{sex}} \sim \mathcal{AF}(\alpha_{\text{sex}}, \gamma_{\text{sex}})$. Note that the expectation parameter μ_f is not included in the hypothesis. In the case of non-stratified baseline hazards the LRT test statistic equals 29.700 on 2 degrees of freedom (p-value ≈ 0). In the case of sex-stratified baseline hazards the LRT test statistic equals 30.822 on 2 degrees of freedom (p-value ≈ 0). Better performance of stratified RE models is also suggested by the ϕ-plot which can be seen in Figure 2a. The measure ϕ is an association measure for bivariate current status data introduced by Unkel and Farrington (2012), $\phi > 0$ indicates positive and $\phi < 0$ negative association. Additionally, ϕ tracks $\ln\{1+\text{RFV}(\boldsymbol{t})\}$ with a time-lag. It can be observed that the association between HPV16 and HPV18 is higher for females early in life, but declines more strongly than for males. This is likely to be the reason for the success of stratified RE models here as those models are able to choose a distinct intercept and slope of the RFV across the sexes. We choose the sex-stratified RE model for further analysis.

In terms of AIC, the non-stratified baseline hazard model $\lambda_0^{(j)}(t)$, $j \in \{\text{HPV16}, \text{HPV18}\}$ performs better than sex-stratified-baseline hazard model $\lambda_0^{(\text{sex}:j)}(t)$ (9647 vs. 9656). Thus, we choose the non-stratified baseline hazard, stratified RE model for further analysis.



(a) Observed association between HPV16 and HPV18 in terms of a ϕ-plot. Black dots refer to cohort- and sex-specific non-parametric estimates, size proportional to precision. The black solid line is the corresponding LOESS. Other dotted and dashed lines are estimates resulting from corresponding parametric model.

(b) RFV of bivariate (left) and higher variate (right) dataset of non-stratified hazard, stratified RE model. Note the different scales on the y-axis. The curves reach a plateau at around 30 years of age, hence, the x-axis was cut-off after 40 years. Note that also the seroprevalence curves $(1 - P(T^{(j)} > t))$ reach a plateau around this time (not shown).

Fig. 2. Observed association between HPV types in the PIENTER-2 data.

A LRT for a constant RFV (CRF), i.e. $H_0 : Z_{\text{sex}} \sim \mathcal{G}(\gamma_{\text{sex}})$ vs. $H_1 : Z_{\text{sex}} \sim \mathcal{AF}(\alpha_{\text{sex}}, \gamma_{\text{sex}})$ for males and females, yields a test-statistic of 93.514 on 2 degrees of freedom (p-value ≈ 0). Hence, the hypothesis on a constant RFV (CRF) can also be rejected. A constant pattern of the RFV is also not suggested by Figure 2a, where it can be observed that association is constantly falling up to the age of around 50. Considering all tests and pairwise AIC comparisons we choose the non-stratified baseline hazard, stratified $\mathcal{AF}$ model for final analysis.

Table 2. Estimated RFV parameters (above dashed line) and resulting estimated frailty distribution parameters (below dashed line) for non-stratified hazard, stratified RE model. Parentheses below point estimates show 95%-CIs.

	male	female
$\hat{\alpha}_{\text{sex}}$	-0.502 $(-0.809;-0.196)$	-2.882 $(-5.008;-0.757)$
$\hat{\gamma}_{\text{sex}}$	83.447 (66.629;104.509)	90.996 (60.167;137.621)
$\hat{\psi}_{\text{sex}}$	0.502 (0.243;1.038)	0.946 (0.336;2.66)
$\hat{\nu}_{\text{sex}}$	0.012 (0.009;0.016)	0.011 (0.006;0.018)
$\hat{\pi}_{\text{sex}}$	0.006 (0.002;0.015)	0.031 (0.009;0.11)

Table 3. Estimated distribution of RCs and across stratum analysis for stratified RE, non-stratified hazard model. Parentheses below point estimate show 95%-CI.

	$\hat{P}\big(Z_{\text{sex}} \leq \hat{z}_{\text{sex},(k)}\big)$			$\hat{z}_{\text{sex,(k)}}$		
k^{th} RC	males	females	$\frac{\hat{P}\left(Z_f \leq z_{f,(k)}\right)}{\hat{P}\left(Z_m \leq z_{m,(k)}\right)}$	males	females	$\widehat{\text{HR}}_{\text{A}}(k)$
1^{st}	0.941 (0.932;0.949)	0.964 (0.956;0.97)	1.024 (1.013;1.036)	0.006 (0.002;0.015)	0.01 (0.004;0.026)	1.684 (1.411;2.01)
2^{nd}	0.952 (0.949;0.955)	0.974 (0.971;0.976)	1.023 (1.013;1.032)	0.508 (0.245;1.053)	0.956 (0.341;2.684)	1.881 (1.064;3.325)
3^{rd}	0.958 (0.956;0.96)	0.978 (0.977;0.98)	1.022 (1.017;1.027)	1.011 (0.489;2.091)	1.902 (0.677;5.344)	1.882 (1.061;3.336)
4^{th}	0.961 (0.96;0.963)	0.982 (0.98;0.983)	1.021 (1.017;1.025)	1.513 (0.732;3.13)	2.848 (1.014;8.004)	1.882 (1.061;3.34)
5^{th}	0.964 (0.963;0.966)	0.984 (0.983;0.985)	1.02 (1.017;1.024)	2.016 (0.975;4.168)	3.794 (1.35;10.664)	1.882 (1.06;3.341)

The RFV parameter estimates for the stratified RE, non-stratified baseline hazard models can be seen in Table 2. The estimated RFV (CRF) is decreasing for both sexes. The estimated RFV parameters indicate higher heterogeneity across clusters or association within a cluster for females early on, as indicated by the intercept of the RFV ($\hat{\gamma}_f > \hat{\gamma}_m$). However, the descent is more strongly for females ($|\hat{\alpha}_f| > |\hat{\alpha}_m|$) and consequently it is estimated that heterogeneity/association is stronger for males from the 4th year of life onwards; see left-hand panel of Figure 2b. These results are also supported by the non-parametric estimates of ϕ in Figure 2a.

The estimated distribution corresponds to a $\hat{\psi}_{\text{sex}}\mathcal{NB}_{>0}(\hat{\nu}_{\text{sex}}, \hat{\pi}_{\text{sex}})$ for males and females. The mean parameter $\hat{\mu}_f = 0.328$ (95%-CI $[0.091; 1.182]$) and is insignificant as indicated by the 95%-CI. The resulting distribution parameters can also be found in Table 2. The estimated mean $\hat{\mu}_f$ indicates lower expected frailty (and therefore lower population hazard) for females initially. However, the mean parameter has to be interpreted in the context of its distribution. Let $\mu_{\text{sex}}(\boldsymbol{t}) = \mathrm{E}(Z|\boldsymbol{T} > \boldsymbol{t}, \text{sex})$. The limit of the conditional expectation of the frailty is $\mu_{\text{sex}}([\infty, \infty]) = \psi_{\text{sex}}\nu_{\text{sex}}$. With the estimates from Table 2, $\hat{\mu}_f([\infty, \infty]) = 0.01 > \hat{\mu}_m([\infty, \infty]) = 0.006$ follows and the initial order of the expectations is reversed. In this example, this leads to the estimated population seroprevalence $\hat{P}(T^{(\text{HPV16})} \leq \text{age})$ being higher for males early in life but from 12 years of life onwards, females start to catch up and finally cross the curve of males at 25 years of life (not shown). We will discuss the reason for the switching order of $\hat{\mu}_f(\boldsymbol{t})$, and $\hat{\mu}_m(\boldsymbol{t})$ that finally leads to crossing seroprevalence curves by analysing the distribution of the frailties in the paragraphs below.

Table 3 shows an excerpt of the distribution of the RCs. We interpret the distribution of the RCs as the distribution of stratum-relative risk-related behaviour and predisposition. The bulk of the population is estimated to be in the lowest RC, though there is more probability to the right of the lowest RC for males. The ratio of the cumulative probabilities between females and males is always above one, also indicating a more heavy tail for males. The heavier tale of the distribution of latent RCs for males is the reason for $\mu_m > \hat{\mu}_f$.

The numerical value of the frailties then assigns a magnitude related interpretation to the distinct RCs. The estimated support shows that females have a higher category-related hazard in each RC (see the last column Table 3). Across strata, given the k^{th} RC, the conditional or RC-related HR, $\widehat{\text{HR}}_{\text{A}}(k) = \frac{\hat{z}_{f,(k)}\hat{\lambda}_0^{(j)}(t)}{\hat{z}_{m,(k)}\hat{\lambda}_0^{(j)}(t)}$, is 1.684 in the important first category. In this case, the $\widehat{\text{HR}}_{\text{A}}(k)$ approaches its limit (with respect to k), $\frac{\hat{\psi}_f}{\hat{\psi}_m} = 1.883$, fast due to small values of $\hat{\nu}_{\text{sex}}$. Higher RC-related hazard for females is the reason for $\hat{\mu}_f(\boldsymbol{t})$ surpassing $\hat{\mu}_m(\boldsymbol{t})$ with time progressing: the individuals belonging to the tale of the distribution of the RCs start to seroconvert early. This selection effect is more pronounced for males due to the heavier tale of the distribution of RCs. Due to extreme individuals within the male population seroconverting more quickly, the higher RC-related hazard for females causes $\hat{\mu}_f([t,0]) > \hat{\mu}_m([t,0])$ from 12 years of life onwards.

Within stratum, $\widehat{\mathrm{HR}}_{\mathrm{W}}(1) = 94.878$ (95%-CI $[60.448; 148.919]$) for females and 84.949 (95%-CI $[65.475; 110.215]$) for males, i.e. being in the second instead of the lowest RC is estimated to be more hazardous for females than for males even from a relative perspective. The $\widehat{\mathrm{HR}}_{\mathrm{W}}(k)$ then approaches its limit $\frac{k}{k-1}$ immediately because $\hat{\nu}_{\text{sex}}$ is small for males and females.

Differences in unobserved heterogeneity across the sexes are reflected by the support and the distribution of the RCs. Given that HPV is a sexually transmitted disease, the membership to a certain RC is partly governed by stratum-relative (sexual) behaviour in that sense, that having, for example, a higher number of sexual partners than some stratum reference should put one in a higher RC than the reference individual. It is tempting to interpret the difference in magnitude of a given RC on the conditional hazard rate across the sexes. For the human immunodeficiency viruses, for example, it is known that male to female transmission is more likely than female to male transmission (see Nicolosi et al. (1994) or European Study Group on Heterosexual Transmission of HIV (1992)). Assuming that each RC comprises the same set of sexual behaviour across the sexes, $z_{f,(k)} > z_{m,(k)}$ for all k, could also hint on a higher susceptibility of females with respect to an infection with HPV16 and HPV18 per relevant contact. However, the RCs are anchored in the stratum and do not necessarily imply the same behaviour across the sexes. Hence, this interpretation is highly speculative and assumption based.

5.2. Higher Variate Analysis

When including all seven high-risk types of HPV for which we have data, the direction of interpretation is largely similar to that of the bivariate case, and mainly the magnitude changes. The estimated RFV parameters are $\hat{\alpha}_m = -1.359$ (95%-CI $[-1.8; -0.918]$), $\hat{\gamma}_m = 9.908$ (95%-CI $[8.928; 10.995]$), $\hat{\alpha}_f = -2.005$ (95%-CI $[-2.5; -1.509]$), $\hat{\gamma}_f = 6.855$ (95%-CI $[6.143; 7.649]$). The heterogeneity/association is less extreme than in the bivariate case early in life. However, the association remains at larger levels compared to the bivariate case, as shown in Figure 2b, indicating that association remains high throughout life. It can also be seen that the RFV (CRF) of females is always below that of males, indicating greater heterogeneity due to individual factors for males throughout the entire time period. As the level of association differs strongly between the bivariate case above and the higher variate case with seven high-risk HPV types, a shared frailty model might be seen as inadequate to capture the patterns of association between the various types of HPV. Therefore, a correlated frailty model may be more appropriate. The shared frailty model might be chosen for its simplicity, however, if the specific types of HPV are less relevant to the research question but, for example, the prognostic factor of one "anonymous" high-risk type on another "anonymous" high-risk type is investigated.

The (initial) expectation of the frailty is virtually identical for males and females; $\hat{\mu}_f = 0.955$ (95%-CI $[0.883; 1.033]$). In the higher variate case the distribution of the RCs is not as much focused on the lower categories. The tail is again more heavy for males (not shown). The $\mathrm{HR}_\mathrm{A}(k) > 1$ for all k again indicates higher RC-related hazard for females. The $\widehat{\mathrm{HR}}_\mathrm{W}(1)$ is less extreme in the higher variate case than in the bivariate case: 9.86 (95%-CI $[8.627; 11.268]$) for females and 12.267 (95%-CI $[10.786; 13.95]$) for males. Note that the order of $\widehat{\mathrm{HR}}_\mathrm{W}(1)$ for males and females changes when comparing this to the bivariate scenario.

6. CONCLUSION

In this paper, we discuss the Addams family of discrete frailty distributions, which has been conceptualised by Farrington et al. (2012) for modelling individual heterogeneity in time-to-event models. We further examine the properties of the conditional time-to-event model induced by the Addams family and develop estimation routines for multivariate case I interval-censored data.

For discrete frailty distributions, the RFV (CRF) approaches either infinity or zero (one) over the course of time, where the distinction is made by the minimum of the support of the frailty being zero or greater than zero respectively. Few discrete frailty distributions are able to manipulate the support via its parameters to choose the long-term behaviour of the two functions accordingly, but this typically involves the edge of the parameter space; see Bardo and Unkel (2023) for a discussion. For the Addams family of discrete frailty distributions, the minimum of the support can either be zero, resulting in a cure rate model, or greater than zero without involving the edge of the parameter space. Consequently, the RFV (CRF) is either monotonically increasing or monotonically decreasing, again without involving the edge of the parameter space. Through the introduction of a scaling parameter, the Addams family is also able to increase or flatten the slope of the RFV (CRF) and might even approach a constant by approaching its continuous exception, a shape that is impossible for discrete shared frailty model to generate. This makes the Addams family a useful general-purpose modelling approach.

A unique feature of the Addams family is that the support of the discrete frailty distribution varies with its parameters and is hence subject to estimation. We suggest interpreting the support as ordered latent risk categories. This feature allows for a unique analysis of the latent model as the effect of latent risk category membership on the hazard rates can be separated from the distribution of the latent risk category membership. By focusing on the support of the frailty, the latent model can essentially be interpreted analogously to the effect of a covariate, e.g. via time-invariant hazard ratios of different risk categories, which we call the within-stratum hazard ratio. If a model is imposed on the distribution parameters of the Addams family, this analysis can be enriched by the across-stratum hazard ratio, i.e. the hazard ratio of a given latent risk category for different strata that are defined by covariates. In a second step, the distribution of the ordered risk category membership can be examined in order to fully understand the impact of unobserved heterogeneity on observable patterns such as population hazard rates and ratios which are averaged over the risk category membership of survivors. This type of analysis could also be performed with the discrete K-point distribution. However, there is no counterpart to this covariate-style analysis for continuous frailty distributions, or for discrete frailty distributions where the support is fixed. This is because the distribution of frailty cannot be meaningfully separated from the effect of frailty on hazard rates, as one would need to compare, e.g., quartiles of frailty distributions that vary with survival. Consequently, a time-invariant proportional hazards interpretation is not possible because the hazard ratio inherits the survival condition. Thus, the Addams family and the K-point distribution offer the possibility to analyse the latent model, which may include heterogeneous covariate effects, thoroughly with common measures.

The analysis of the latent model via the within-stratum hazard ratio might help to understand the importance of individual heterogeneity. In that sense, individual heterogeneity might be regarded as important if the within-stratum hazard ratios are large and vice versa. The analysis of the across-stratum hazard ratio may reveal structural differences in individual heterogeneity across covariates, prompting a discussion of the reasons for this. In this sense, the covariate-style interpretation may be beneficial for scientific discussion, as hazard ratios and probabilities are a common way of communicating with a non-statistical audience.

We applied the Addams family to multivariate case I interval-censored infection data and allowed the distribution of individual heterogeneity to differ for males and females. Males are found to have a higher probability for more hazardous categories, possibly reflecting a more cautious behaviour in the female population compared to males. However, the estimated hazard in each risk category is higher for females than for males, which might reflect a higher biological burden with respect to the susceptibility of HPV. There was no evidence for the existence of a non-susceptible sub-group, neither in the bivariate data set, including HPV 16 and HPV 18, nor in the data set containing seven high-risk types of HPV.

Competing interests

No competing interest is declared.

Acknowledgments

We are grateful to Fiona van der Klis and Liesbeth Mollema from the Dutch National Institute for Public Health and the Environment (RIVM), for giving us permission to use the data from the second PIENTER study. The authors also thank the anonymous reviewers for their valuable suggestions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) [grant number UN 400/2-1].

REFERENCES

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Event History Analysis: A Process Point of View.* Statistics for Biology and Health. Springer, Dordrecht.

Ata, N. and Özel, G. (2013). Survival functions for the frailty models based on the discrete compound poisson process. *Journal of Statistical Computation and Simulation*, 83(11):2105–2116.

Bardo, M. and Unkel, S. (2023). The shape of the relative frailty variance induced by discrete random effect distributions in univariate and multivariate survival models. *https://arxiv.org/pdf/2303.04915*, pages 1–25.

Begun, A. Z., Iachine, I. A., and Yashin, A. I. (2000). Genetic nature of individual frailty: comparison of two approaches. *Twin Research and Human Genetics*, 3(1):51–57.

Bijwaard, G. (2014). Multistate event history analysis with frailty. *Demographic Research*, 30:1591–1620.

Burchell, A. N., Winer, R. L., de Sanjosé, S., and Franco, E. L. (2006). Chapter 6: Epidemiology and transmission dynamics of genital hpv infection. *Vaccine*, 24

Suppl 3:S3/52–61.

Cancho, V. G., Barriga, G., Leão, J., and Saulo, H. (2021). Survival model induced by discrete frailty for modeling of lifetime data with long-term survivors and change-point. *Communications in Statistics - Theory and Methods*, 50(5):1161–1172.

Cancho, V. G., Macera, M. A. C., Suzuki, A. K., Louzada, F., and Zavaleta, K. E. C. (2020a). A new long-term survival model with dispersion induced by discrete frailty. *Lifetime Data Analysis*, 26(2):221–244.

Cancho, V. G., Suzuki, A. K., Barriga, G. D. C., and do Espirito Santo, Ana P. J. (2020b). A multivariate survival model induced by discrete frailty. *Communications in Statistics - Simulation and Computation*, pages 1–19.

Cancho, V. G., Zavaleta, K. E. C., Macera, M. A. C., Suzuki, A. K., and Louzada, F. (2018). A bayesian cure rate model with dispersion induced by discrete frailty. *Communications for Statistical Applications and Methods*, 25(5):471–488.

Caroni, C., Crowder, M., and Kimber, A. (2010). Proportional hazards models with discrete frailty. *Lifetime Data Analysis*, 16(3):374–384.

Choi, S. and Huang, X. (2012). A general class of semiparametric transformation frailty models for nonproportional hazards survival data. *Biometrics*, 68(4):1126–1135.

Choi, S., Huang, X., and Chen, Y.-H. (2014). A class of semiparametric transformation models for survival data with a cured proportion. *Lifetime Data Analysis*, 20(3):369–386.

Cox, C., Chu, H., Schneider, M. F., and Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23):4352–4374.

de Souza, D., Cancho, V. G., Rodrigues, J., and Balakrishnan, N. (2017). Bayesian cure rate models induced by frailty in survival analysis. *Statistical Methods in Medical Research*, 26(5):2011–2028.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer, Dordrecht.

European Study Group on Heterosexual Transmission of HIV (1992). Comparison of female to male and male to female transmission of hiv in 563 stable couples. european study group on heterosexual transmission of hiv. *BMJ (Clinical research ed.)*, 304(6830):809–813.

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(4):673–696.

Gasperoni, F., Ieva, F., Paganoni, A. M., Jackson, C. H., and Sharples, L. (2020). Non-parametric frailty cox models for hierarchical time-to-event data. *Biostatistics (Oxford, England)*, 21(3):531–544.

Gavillon, N., Vervaet, H., Derniaux, E., Terrosi, P., Graesslin, O., and Quereux, C. (2010). Papillomavirus humain (hpv) : comment ai-je attrapé ça ? *Gynecologie, obstetrique & fertilite*, 38(3):199–204.

Gilbert, P. and Varadhan, R. (2019). numderiv: Accurate numerical derivatives. `https://cran.r-project.org/web/packages/numDeriv/`.

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*, 28(22):2785–2800.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

Lam, J. U. H., Rebolj, M., Dugué, P.-A., Bonde, J., von Euler-Chelpin, M., and Lynge, E. (2014). Condom use in prevention of human papillomavirus infections and cervical neoplasia: systematic review of longitudinal studies. *Journal of Medical Screening*, 21(1):38–50.

Meyers, J., Ryndock, E., Conway, M. J., Meyers, C., and Robison, R. (2014). Susceptibility of high-risk human papillomavirus type 16 to clinical disinfectants. *The Journal of Antimicrobial Chemotherapy*, 69(6):1546–1550.

Mohseni, N., Maboudi, A. A. K., Baghestani, A., and Saeedi, A. (2020). A cure rate model with discrete frailty on hodgkin lymphoma patients after diagnosis. *Archives of Advances in Biosciences*, 11(4):15–22.

Molina, K. C., Calsavara, V. F., Tomazella, V. D., and Milani, E. A. (2021). Survival models induced by zero-modified power series discrete frailty: application with a melanoma data set. *Statistical Methods in Medical Research*, 30(8):1874–1889.

Mollema, L., de Melker, H. E., Hahne, S. J., van Weert, J. W., Berbers, G. A., and van der Klis, F. R. (2009). *PIENTER 2-project: second research project on the protection against infectious diseases offered by the national immunization programme in the Netherlands (Report 230421001/2009)*. Rijksinstituut voor Volksgezondheid en Milieu RIVM.

Nicolosi, A., Maria Léa Corrêa Leite, Musicco, M., Arici, C., Gavazzeni, G., and Lazzarin, A. (1994). The efficiency of male-to-female and female-to-male sexual transmission of the human immunodeficiency virus: a study of 730 stable couples. *Epidemiology*, 5(6):570–575.

Nielson, C. M., Harris, R. B., Nyitray, A. G., Dunne, E. F., Stone, K. M., and Giuliano, A. R. (2010). Consistent condom use is associated with lower prevalence of human papillomavirus infection in men. *The Journal of Infectious Diseases*, 202(3):445–451.

Palloni, A. and Beltrán-Sánchez, H. (2017). Discrete barker frailty and warped mortality dynamics at older ages. *Demography*, 54(2):655–671.

Pickles, A. and Crouchley, R. (1994). Generalizations and applications of frailty models for survival and event data. *Statistical Methods in Medical Research*, 3(3):263–278.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, H., Franco, E., Pintos, J., Bergeron, J., Arella, M., and Tellier, P. (2000). Determinants of low-risk and high-risk cervical human papillomavirus infections in montreal university students. *Sexually Transmitted Diseases*, 27(2):79–86.

Rintala, M. A. M., Grénman, S. E., Puranen, M. H., Isolauri, E., Ekblad, U., Kero, P. O., and Syrjänen, S. M. (2005). Transmission of high-risk human papillomavirus (hpv) between parents and infant: a prospective study of hpv in families in finland. *Journal of Clinical Microbiology*, 43(1):376–381.

Scherpenisse, M., Mollers, M., Schepp, R. M., Boot, H. J., de Melker, H. E., Meijer, C. J. L. M., Berbers, G. A. M., and van der Klis, F. R. M. (2012). Seroprevalence of seven high-risk hpv types in the Netherlands. *Vaccine*, 30(47):6686–6693.

Sun, J. (2007). *The Statistical Analysis of Interval Censored Failure Time Data*. Springer, New York.

Syrjänen, S. (2010). Current concepts on human papillomavirus infections in children. *APMIS*, 118(6-7):494–509.

Troncoso-Ponce, D. (2018). Estimation of competing risks duration models with unobserved heterogeneity using hsmlogit. `https://doi.org/10.2139/ssrn.3114159`.

Unkel, S. and Farrington, C. P. (2012). A new measure of time-varying association for shared frailty models with bivariate current status data. *Biostatistics (Oxford, England)*, 13(4):665–679.

Unkel, S., Farrington, C. P., Whitaker, 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(1):141–158.

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