Advances in frailty models

Balan, T.A.

Advances in Frailty Models

Theodor Adrian Bălan

Cover design: Alexandru Andrei
Printing: Ipskamp Printing, The Netherlands

ISBN: 978-90-9031062-6

Advances in Frailty Models

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 26 september 2018
klokke 13.45 uur

door

Theodor Adrian Bălan

geboren te Boekarest, Roemenië in 1989

Promotor:	Prof. dr. H. Putter
Leden promotiecomssie:	Prof. dr. S. le Cessie Prof. dr. M.J.C. Eijkemans · *Universitair Medisch Centrum Utrecht, Utrecht* Prof. dr. D. Rizopoulos · *Erasmus Medisch Centrum, Rotterdam*

Table of Contents

Chapter 1

Introduction: A tutorial in frailty modeling

1.1 Introduction

Cox's proportional hazards model (Cox, 1972) is one of the most popular regression models for time to event outcomes. The hazard function, which may be used to describe the distribution of event times, is defined as the instantaneous probability of an event, given that the individual has not experienced the event previously. The proportional hazards assumption specifies that the ratio of the hazards between any two individuals is constant in time, and the shape of the hazard is given by a non-parametric "baseline hazard". If a model is perfectly specified, so that all possible relevant covariates are accounted for, then the baseline hazard reflects the randomness of the event time, given the value of covariates.

In practice however, it is rarely possible to account for all relevant covariates. Then the explanatory variables account for *observed heterogeneity*, and the unaccounted part is termed *unobserved heterogeneity*. If this is the case, then the estimated hazard for a specific set of covariates does not have an individual interpretation (Woodbury and Manton, 1977). Rather, it represents an average hazard function, where the average is taken at each time point over the individuals still alive at that time point. The effects of unobserved heterogeneity on life times were collectively referred to as *frailty* in demographic research (Vaupel, Manton, and Stallard, 1979). The frailty is an unobserved

This chapter is part of the manuscript under preparation: T.A. Balan, H. Putter. *A tutorial in frailty models: theory and practice*

individual random effect that acts multiplicatively on the hazard. The estimated spread of this random effect (e.g. variance) is an indication of the amount of unobserved heterogeneity. The frailty model quickly gained popularity in econometrics (Heckman and Singer, 1984), demographics (Vaupel and Yashin, 1985) and biostatistics (Aalen, 1988).

The Cox model, developed originally for *univariate* survival data, has been extended to a more general framework based on counting processes (Andersen and Gill, 1982). The resulting "extended Cox model" easily accommodates more complex data, such as correlated event times (*clustered failures*) or multiple events per individual (*recurrent events*). Frailty models based on the extended Cox model are referred to as *shared* frailty models (Nielsen et al., 1992; Andersen, Borgan, et al., 1993), as opposed to *univariate* frailty models in the simpler univariate survival data scenario.

For clustered failures, the estimated frailty variance describes unobserved heterogeneity between clusters. Within a cluster, the event times are assumed to be independent, given the frailty. Therefore, shared frailty models are often used to model the effect of unobserved risk factors that are specific to the clusters. For recurrent events, the estimated frailty variance describes unobserved heterogeneity between individuals, as in the univariate frailty case. Conditional on the frailty, the event history of an individual is typically modeled as a Poisson or renewal process. In all cases, frailty models involve the conditional specification of the hazard or intensity of the event process, as if the frailty would be observed. Consequently, the estimated covariate effects retain the interpretation of an individual effect.

Most theoretical results in frailty models have focused on the gamma frailty model. In particular, maximum likelihood estimators have been shown to be well behaved (Murphy, 1994; Murphy, 1995b). However, numerous other frailty distributions have been proposed in the literature (Hougaard, 1986a; Hougaard, 2000; Paddy Farrington, Unkel, and Anaya-Izquierdo, 2012). The real frailty distribution is almost impossible to be known in advance. It is therefore of interest to compare the characteristics of different frailty models in terms of the resulting population hazards (for univariate survival data) or within cluster correlation patterns (for clustered survival data).

The aim of this chapter is to provide an overview of theory and practice in the field of frailty models, while offering insight into the problems that are addressed by such models. In Section 1.2, we study the effects of unobserved heterogeneity in survival data, univariate frailty models and different frailty distributions. In Section 1.3, we analyze the effect of unobserved heterogeneity in clustered survival data and introduce the shared frailty model. We study different correlation structures and we discuss frailty models for recurrent events data. In Section 1.4, we discuss estimation and inference procedures for frailty models, we compare available software packages and we examine the representation of event history data in the R statistical software. In Section 1.5 we overview different extensions to the frailty models. Finally, in Section 1.6, we conclude with an outline of the rest of this thesis.

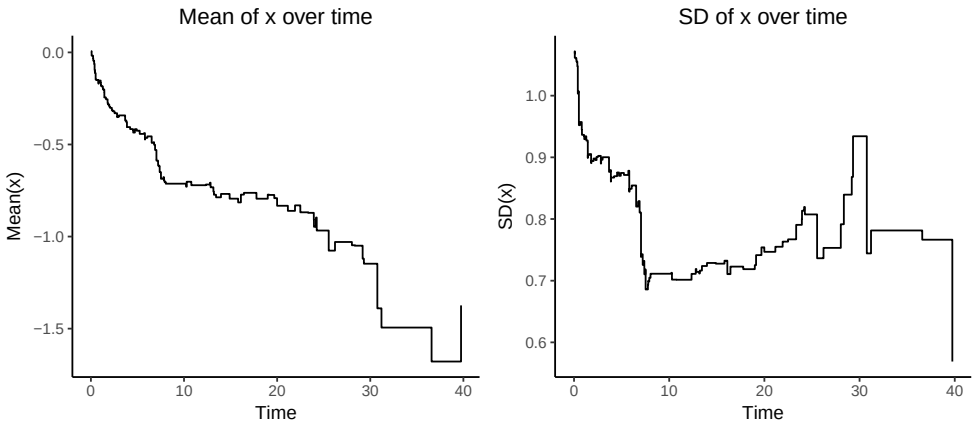

Figure 11: Changes in the mean and variance of a covariate x over time among survivors in a proportional hazards model.

1.2 Univariate frailty models

1.2.1 Heterogeneity in the Cox model

Heterogeneity over time

The Cox model specifies that the hazard as

$$\lambda(t \mid x) = \lambda_0(t) \exp(\beta^\top \mathbf{x}), \tag{1.1}$$

where β is a $p \times 1$ vector of regression coefficients, $\mathbf{x}$, is a $p \times 1$ vector of covariates and $h_0(t)$ is an unspecified baseline hazard function. The hazard functions of two individuals with covariate vectors $\mathbf{x}^*$ and $\tilde{\mathbf{x}}$ are equal only when $\beta^\top \mathbf{x}^* = \beta^\top \tilde{\mathbf{x}}$. The exponent $\exp(\beta_j)$ is the hazard ratio between an individual with $x_j = 1$ and an individual with $x_j = 0$. Time dependent covariates are easily accommodated in (1.1) and are discussed discussed in Section 1.4. Henceforth, we assume that $\mathbf{x}$ is time-constant.

The risk set at time t is composed of individuals that have not yet experienced the event of interest or have not yet been removed for other reasons, such as censoring. The distribution of the covariates among the individuals in the risk set changes in time. We illustrate this by considering the model (1.1) and only one covariate following a standard normal distribution $x \sim N(0, 1)$ and $\beta > 0$, so that individuals with larger values of x have a higher hazard. At time $t = 0$, the mean and variance of x are 0 and 1, respectively. As time passes, the risk set progressively comprises individuals with lower values of x. For this reason, the average value and the sample variance of x among the individuals at risk decreases over time.

This is illustrated by simulating a single sample of size $n = 100$, and a covariate $x \sim N(0, 1)$, with $\beta = 1$, $\lambda_0(t) \equiv 0.1$ and uniform censoring on $(20, 50)$. In this simulated

sample, at time 0, x had mean 0.007 and standard deviation 1.068. The estimated β was 0.943, with a standard error of 0.127. The mean and standard deviation of x among the individuals in the risk set are shown in Figure 11, as a function of time. The message is that, in time, the population of "survivors" (those still at risk) is more homogeneous and of a lower risk than the original risk set at time 0.

Heterogeneity due to missing covariates

The proportional hazards assumption in the Cox model (1.1) specifies that the ratio $\lambda(t \mid \mathbf{x}^*)$ divided by $\lambda(t \mid \tilde{\mathbf{x}})$ equals $\exp(\beta^\top(\mathbf{x}^* - \tilde{\mathbf{x}}))$, which does not depend on time. When this assumption is violated, the covariate effect β is time dependent. The true model is therefore

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp(\beta(t)\mathbf{x})$$

with $\beta(t)$ not constant.

Assume that the model (1.1) is correct and $p \geq 2$. Then, if important covariates are omitted from the model, the proportional hazards assumption does not usually hold for the remaining covariates. This is illustrated by simulating a single large data set with sample size $n = 10,000$. Two independent covariates x_1 and x_2 are considered, both ~ $N(0,1)$, with $\beta_1 = \beta_2 = 1$, $\lambda_0 \equiv 1$ and uniform censoring on $(20, 50)$. The following output is shown from Cox models fitted with the standard **survival** package in R (Therneau and Grambsch, 2000). When both covariates are included into the model, the results are close to the simulation scenario, with both estimated regression coefficients close to 1:

```
## Call:
## c12 <- coxph(formula = Surv(time, status) ~ x1 + x2, data = d)
##
##       coef exp(coef) se(coef)    z      p
## x1 1.0016    2.7225   0.0138 72.7 <2e-16
## x2 1.0240    2.7843   0.0140 73.2 <2e-16
##
## Likelihood ratio test=9014  on 2 df, p=0
## n= 10000, number of events= 8240
```

No evidence of violation of the proportional hazards assumption is found, when a test based on Schoenfeld residuals is used (Grambsch and Therneau, 1994):

```
## Call: cox.zph(c12, transform = "identity")
##              rho  chisq     p
## x1       0.00101 0.0081 0.928
## x2      -0.00357 0.1050 0.746
## GLOBAL        NA 0.1510 0.927
```

However, if x_2 is omitted, the resulting estimate of the effect of x_1 is visibly smaller than 1:

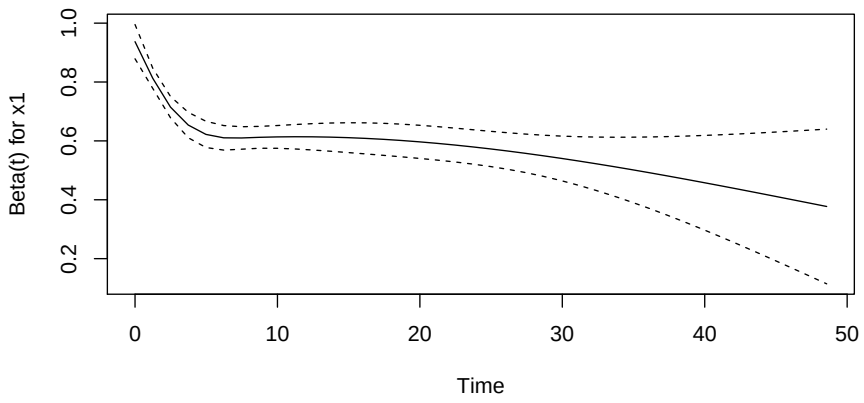

Figure 12: Plot of scaled Schoenfeld residuals based $\beta(t)$ induced by omitting a covariate from a proportional hazards model.

```
## Call:
## c1 <- coxph(formula = Surv(time, status) ~ x1, data = d)
##
##      coef exp(coef) se(coef)    z      p
## x1 0.7028    2.0195   0.0124 56.6 <2e-16
##
## Likelihood ratio test=3271  on 1 df, p=0
## n= 10000, number of events= 8240
```

Moreover, there is clear evidence against the proportional hazards assumption.

```
## Call: cox.zph(c1, transform = "identitiy")
##        rho chisq        p
## x1 -0.0852  55.3 1.06e-13
```

This is also illustrated by the plot of scaled Schoenfeld residuals of $\beta(t)$ in figure 12. It appears that the effect of x starts as close to the true value 1, and then decreases in time. The result given by the Cox model only with x_1 is an "average" effect in this case.

The explanation for the phenomenon illustrated in the simulated example is that the individual hazard is determined by the combined effect of x_1 and x_2. On average, the "high risk" individuals (high x_1, high x_2) are the first to have the event, followed by the "moderate risk" ones (high x_1 and low x_2, or low x_1 and high x_2), and eventually the "low risk" ones (low x_1 and low x_2). In particular, the individuals that survive up to a certain

time are more likely to have lower values of x_2. If x_2 is omitted from the model, this decrease in risk among the survivors must be accounted for only by x_1, thus reducing the perceived effect of the included covariate.

Conditional and marginal hazards

More generally, suppose that the proportional hazards model (1.1) holds for a covariate vector $\mathbf{x} = (\mathbf{x}_{\text{incl}}, \mathbf{x}_{\text{omit}})$ with covariate effects $\beta = (\beta_{\text{incl}}, \beta_{\text{omit}})$, so that the true model is

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp(\beta_{\text{incl}}^\top \mathbf{x}_{\text{incl}} + \beta_{\text{omit}}^\top \mathbf{x}_{\text{omit}}). \tag{1.2}$$

Imagine that a Cox model is fitted only including x_{incl}. This will result in an estimated effect that is biased towards 0 and, usually, a violation of the proportional hazards assumption. In reality, it is possible that some relevant covariates are not measured (here represented by $\mathbf{x}_{\text{omit}}$). In this case, these omitted covariates are said to induce *unobserved heterogeneity*. The differences between individuals that are explained by $\mathbf{x}_{\text{incl}}$ are referred to as *observed heterogeneity*.

The $\lambda(t \mid \mathbf{x})$, as defined in model (1.2), is referred to as the *conditional* hazard, with β_{incl} summarizing the *conditional* effect of $\mathbf{x}_{\text{incl}}$. When unobserved heterogeneity is present, the resulting $\lambda(t \mid \mathbf{x}_{\text{incl}})$ is referred to as the *marginal* hazard (although it is marginal with respect to $\mathbf{x}_{\text{omit}}$ but still conditional on $\mathbf{x}_{\text{incl}}$). The estimated effect from the marginal model does not have an individual interpretation. Namely, $\lambda(t|\mathbf{x}_{\text{incl}})$ represents a weighted average of the individual hazards corresponding to those individuals in the risk set at time t, where the weighing is determined by the distribution of $\mathbf{x}_{\text{omit}}$ in this risk set.

Since the effect of $\mathbf{x}_{\text{omit}}$ cannot be directly observed, one can define the random variable $Z = \exp(\beta_{\text{omit}}^\top \mathbf{x}_{\text{omit}})$. Z is referred to as a "frailty" term that acts multiplicatively on the hazard.

1.2.2 The frailty model

In the univariate frailty model, the hazard of an individual with frailty Z is specified as

$$\lambda(t \mid Z) = Z\lambda(t). \tag{1.3}$$

For identifiability, Z is assumed to be scaled so that $\mathrm{E}Z = 1$. The second term in (1.3), $\lambda(t) \equiv \lambda(t \mid Z = 1)$, is the conditional hazard for an individual with $Z = 1$. We refer to $\lambda(t)$ simply as the conditional hazard. The conditional cumulative hazard is defined as $\Lambda(t) = \int_0^t \lambda(s)ds$. The conditional survival function for an individual with frailty Z is then given by

$$S(t \mid Z) = \exp(-Z\Lambda(t)).$$

The marginal survival curve associated with $\Lambda(t)$ is obtained by taking the expectation of $S(t|Z)$ with respect to Z,

$$\overline{S}(t) = \mathrm{E}[\exp(-Z\Lambda(t))]. \tag{1.4}$$

Unlike $S(t)$, $\overline{S}$ has a population averaged interpretation. If there are no covariates, $\overline{S}$ may be seen as a weighted average of individual survival curves, where the weighing depends on the distribution of Z. The hazard may be derived from the survival function as $\lambda(t) = d/dt[-\log S(t)]$. Therefore, the marginal hazard may be calculated as

$$\begin{aligned}\overline{\lambda}(t) &= \frac{\mathrm{E}\left[Z\exp(-Z\Lambda(t))\right]}{\mathrm{E}\left[\exp(-Z\Lambda(t)\right]}\lambda(t)\\ &= \mathrm{E}[Z|T \geq t]\lambda(t).\end{aligned}$$

A population averaged interpretation may also be given here: $\overline{\lambda}(t)$ may be seen as a weighted average of individual hazards *of individuals alive at time t*, where the weighing depends on the distribution of Z *among the individuals alive at time t*.

The conditional and marginal hazards are equal only if $\mathrm{E}[Z|T \geq t] = 1$ for all t. In other words, if all frailties Z are equal to 1. Otherwise, the frailty distribution among the survivors at time t behaves in a similar fashion with the distribution of an observed covariate among survivors, as shown in Section 1.2.1.

If observed covariates are also present, then it is usually assumed that the proportional hazards assumption holds conditional on the frailty, with $\lambda(t|Z) = Z\exp(\beta^\top \mathbf{x})\lambda_0(t)$. Then, the population averaged interpretations of $\overline{S}$ and $\overline{h}$ hold conditional on $\mathbf{x}$. In other words, for a hypothetical population of individuals with given covariate values $\mathbf{x}$. This is the same as the interpretation that is given to the marginal hazard in Section 1.2.1.

Regardless of whether the differences between individuals come from observed covariates $\mathbf{x}$ or from the frailty, individuals with higher hazards die earlier. Therefore, the population of survivors is more homogeneous and at a lower risk for events than the general population at time 0. The advantage of frailty models is that this is explicitly modeled. Before we further study the relation between marginal and conditional hazards in Section 1.2.4, we first discuss different frailty distributions in Section 1.2.3.

1.2.3 Frailty distributions

The Laplace transform

The distribution of a random variable $Z > 0$ can also be uniquely specified by its Laplace transform,

$$\mathcal{L}(c) = \mathrm{E}\left[\exp(-Zc)\right].$$

It is immediate that $\mathcal{L}(0) = 1$. The expectation of Z may be obtained as minus the derivative of $\mathcal{L}$ calculated in 0, $\mathrm{E}Z = -\mathcal{L}'(0)$. Furthermore, $\mathcal{L}''(0) = \mathrm{E}Z^2$ and higher order moments of Z can be obtained by taking further derivatives of $\mathcal{L}$. Denote the kth derivative of $\mathcal{L}$ as $\mathcal{L}^{(k)}$. The squared coefficient of variation, defined as $\mathrm{CV}^2 = \mathrm{var}[Z]/(\mathrm{E}[Z])^2$, may be expressed as

$$\mathrm{CV}^2[Z] = \frac{\mathcal{L}''(0)}{(\mathcal{L}'(0))^2} - 1.$$

In terms of the Laplace transform, the marginal survival function from (1.4) may be written as

$$\overline{S}(t) = \mathcal{L}(\Lambda(t)),$$

and the marginal hazard as

$$\overline{\lambda}(t) = \frac{d}{dt}[-\log S(t)] = -\frac{\mathcal{L}'(\Lambda(t))}{\mathcal{L}(\Lambda(t))}\lambda(t).$$

The Laplace transform of the frailty distribution of survivors can be obtained from Bayes' theorem:

$$\begin{aligned}\mathcal{L}_{Z|T\geq t}(c) &= \mathrm{E}\left[\exp(-Z\,c)|T \geq t\right] \\ &= \frac{\mathrm{E}[\exp(-Z(c+\Lambda(t)))]}{\mathrm{E}[\exp(-Z\Lambda(t))]} \\ &= \frac{\mathcal{L}(c+\Lambda(t))}{\mathcal{L}(\Lambda(t))}.\end{aligned} \tag{1.5}$$

The expectation, variance and squared coefficinet of variation of $Z|T \geq t$ follow as

$$\begin{aligned}\mathrm{E}[Z|T \geq t] &= -\frac{\mathcal{L}'(\Lambda(t))}{\mathcal{L}(\Lambda(t))}, \\ \mathrm{var}[Z|T \geq t] &= \frac{\mathcal{L}''(\Lambda(t))}{\mathcal{L}(\Lambda(t))} - \left(\frac{\mathcal{L}'(\Lambda(t))}{\mathcal{L}(\Lambda(t))}\right)^2 \\ \mathrm{CV}^2[Z|T \geq t] &= \frac{\mathcal{L}''(\Lambda(t))\mathcal{L}(\Lambda(t))}{(\mathcal{L}'(\Lambda(t)))^2} - 1.\end{aligned}$$

Infinitely divisible distributions

The *infinitely divisible* distributions are a family of distributions with tractable Laplace transform, specified as $\mathcal{L}(c) = \exp(-\alpha\psi(c;\gamma))$ with $\alpha > 0$ and $\gamma > 0$. The expectation and variance can be expressed as

$$\begin{aligned}&\mathrm{E}[Z|T \geq t] = \alpha\psi'(\Lambda(t);\gamma), \\ &\mathrm{var}[Z|T \geq t] = -\alpha\psi''(\Lambda(t);\gamma), \\ &\mathrm{CV}^2[Z|T \geq t] = -\frac{\psi''(\Lambda(t);\gamma)}{\alpha(\psi'(\Lambda(t)))^2}.\end{aligned} \tag{1.6}$$

The **gamma** distribution is a prominent member of the infinitely divisible family. The density of the gamma distribution with parameters $\theta > 0$ and $\eta > 0$ is given by $f(t;\theta,\eta) = \frac{\theta^\eta}{\Gamma(\eta)}t^{\eta-1}e^{-\theta t}$, where $\Gamma(\eta) = \int_0^\infty s^{\eta-1}e^{-s}ds$ is the gamma function. Its Laplace transform is given by

$$\mathcal{L}(c) = \left(\frac{\theta}{\theta + c}\right)^\eta,$$

which, in terms of (1.6), can be expressed as $\alpha = \eta$, $\theta = \gamma$, and $\psi(c;\gamma) = \log(\gamma + c) - \log(\gamma)$. By convention, the expectation of the frailty is fixed to 1, so the restriction $\theta = \eta$ is applied. In this parmaetrization, Z follows a gamma(θ, θ) distribution, with $\mathrm{E}[Z] = 1$ and $\mathrm{var}[Z] = \theta^{-1} = \xi$. The expectation and variance of the frailty distribution of the survivors is given through (1.6), resulting in

$$\mathrm{E}\left[Z|T \geq t\right] = \frac{\theta}{\theta + \Lambda(t)},$$

$$\mathrm{var}\left[Z|T \geq t\right] = \frac{\theta}{(\theta + \Lambda(t))^2}.$$

Both functions reach their maximum at $t = 0$, with expectation 1 and variance θ^{-1}, and decrease over time. For the gamma frailty, it is immediate that $\overline{\lambda}(t) \leq \lambda(t)$. In other words, the marginal hazard is always smaller than the hazard of an individual with frailty 1.

A more general family of infinitely divisible distributions is the **power variance function (PVF)** family, with the Laplace transform $\mathcal{L}$ described by

$$\mathcal{L}(c;\alpha,\gamma,m) = \exp\left(-\alpha\,\mathrm{sign}(m)\left\{1 - \left(\frac{\gamma}{\gamma + c}\right)^m\right\}\right)$$

where $\mathrm{sign}(m)$ is the sign of m, and $m > -1$ and $m \neq 0$. It was proposed in a series of papers (Hougaard, 1984; Hougaard, 1986a; Hougaard, 1986b) and is studied in detail in Hougaard (2000). To obtain $\mathrm{E}[Z] = 1$ and $\mathrm{var}[Z] = \theta^{-1}$, one can set $\alpha = \theta\mathrm{sign}(m)(m + 1)/m$ and $\gamma = \theta(m + 1)$. Particular cases of include:

- The gamma frailty, obtained as a limiting case when $m \longrightarrow 0$ with $m < 0$;
- The inverse Gaussian distribution, when $m = -1/2$;
- The so-called Hougaard distributions, when $m < 0$;
- The compound Poisson distribution, when $m > 0$, which has probability mass at 0. This is consistent with a scenario where a part of the population is not susceptible for the event of interest;
- The positive stable distribution, obtained as a limiting case when $\gamma \longrightarrow 0$. This distribution cannot be scaled to have $\mathrm{E}[Z] = 1$, so usually the $\alpha = 1$ restruction is imposed. Its expectation is infinite and the variance is not defined. However, the resulting Laplace transform takes the simple form $\mathcal{L}(c) = \exp(-\alpha c^\gamma)$, with $\alpha > 0$ and $\gamma \in (0, 1)$.

The **log-normal** distribution has often been used for frailty models, although it it not part of the PVF family. It is infinitely divisible, but the corresponding expression of ϕ cannot be expressed in closed form. Consequently, its Laplace transform and expressions for the distribution of survivors are not easily obtained. Its popularity stems from the

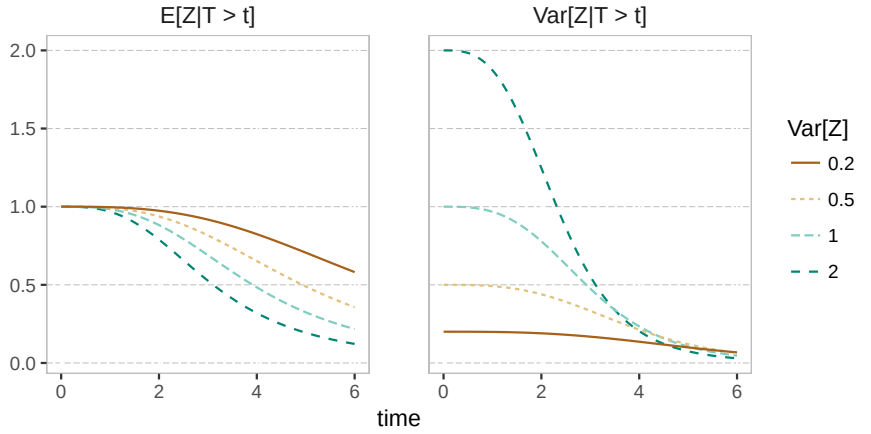

Figure 13: Frailty distribution of survivors, gamma frailty, $\lambda(t) = t^2/20$.

normal random effects in linear models. The log-normal frailty is usually parametrized with the $\mathrm{E}[\log Z] = 0$ and $\mathrm{var}[\log Z] = \sigma^2$, corresponding to a normally distributed random effect on the scale of the covariates. If matched by mean and variance, it is virtually indistinguishable from the inverse Gaussian distribution. Other families of distributions, such as the Addams and Kummer families of distributions were also introduced in the context of frailty models (Aalen, Borgan, and Gjessing, 2008; Paddy Farrington, Unkel, and Anaya-Izquierdo, 2012).

1.2.4 Frailty effects

The different frailty distributions discussed in Section 1.2.3 represent different ways of expressing unobserved heterogeneity. Different frailty distributions lead to different selection effects . Moreover, they impact the marginal effect of the observed covariates in different ways, generalizing the phenomenon illustrated in Section 1.2.1. An advantage of the PVF family of distributions and their closed form Laplace transforms is that it facilitates the study of these phenomenons (Aalen, 1988; Aalen, 1994; Vaupel and Yashin, 1985). An overview may be found in Aalen, Borgan, and Gjessing (2008, ch. 6).

The selection effect In Section 1.2.3, it was shown that, for the gamma frailty model, the expectation and variance of the frailty distribution of the survivors decreases in time. In Figure 13, we show the expectation and the variance of $\mathrm{E}[Z|T \geq t]$, when the conditional hazard is given by $\lambda(t) = t^2/20$, for variances 0.2, 0.5, 1 and 2.

It follows that the marginal hazard appears as a "dragged down" version of the conditional hazard, similar to Figure 11. This selection effect is stronger if the frailty variance is larger. In particular, the marginal hazard may appear to grow, reach a peak and then

decrease beyond a time point, even if the conditional hazard is increasing. As in Section 1.2.1, the explanation is that individuals with a higher hazard die earlier, on average, than individuals with a lower hazard. In particular, this is true for all frailty distributions discussed in Section 1.2.3. For example, for the compound Poisson distribution, when individuals with frailty 0 never experience the event of interest, the marginal hazard will eventually decrease towards 0 after some time point. The point made in Section 1.2.1 is essential here as well: in the presence of unobserved heterogeneity, the marginal hazard has a population averaged rather than an individual interpretation.

The marginal hazard ratio In Section 1.2.1, we illustrated that, when important covariates are omitted in a Cox model, the marginal effect of the remaining covariates is time dependent. The same phenomenon happens with the marginal covariate effect in the case of frailty models. Suppose that only one observed covariate is present, $x \in \{0, 1\}$, and that the frailty model (1.3) is true. Then, e^{β} is the hazard ratio between two individuals with the same frailty, one with $x = 1$, the other with $x = 0$. The marginal hazards for the two groups defined by x are given by

$$\begin{aligned}\overline{\lambda}_0(t) &= \mathrm{E}[Z|T \geq t, x = 0]\, \lambda_0(t),\\ \overline{\lambda}_1(t) &= \mathrm{E}[Z|T \geq t, x = 1]\, e^{\beta}\lambda_0(t).\end{aligned}$$

The marginal effect of x can be quantified by the ratio of these two marginal hazards. This results in

$$e^{\overline{\beta}(t)} = \frac{\overline{\lambda}_1(t)}{\overline{\lambda}_0(t)} = \frac{\mathrm{E}[Z|T \geq t, x = 1]}{\mathrm{E}[Z|T \geq t, x = 0]}\, e^{\beta}.$$

In general, $\overline{\beta}(t)$ is not constant in time. If Z is a gamma frailty with variance θ^{-1}, for example, this is

$$e^{\overline{\beta}(t)} = \frac{\theta + \Lambda_0(t)}{\theta + e^{\beta}\Lambda_0(t)}\, e^{\beta}.$$

If $\beta < 0$, $e^{\overline{\beta}(t)}$ is an increasing function with a minimum of e^{β} and asymptotic maximum of 1. Conversely, if $\beta > 0$, then $e^{\overline{\beta}(t)}$ is a decreasing function with a maximum of e^{β} and asymptotic minimum of 1. The conclusion is that, at the population level, the covariate effect appears to vanish over time. Therefore, the gamma frailty shows a similar behavior with the unobserved covariates scenario that was studied by simulation in Section 1.2.1.

Similar considerations apply for other frailty distributions. For example, for the inverse Gaussian distribution, the marginal hazard ratio is

$$e^{\overline{\beta}(t)} = \left(\frac{\theta + 2\Lambda(t)}{\theta + 2\Lambda_0(t)e^{\beta}}\right)^{1/2}.$$

A peculiar case is that of the positive stable distribution, for which

$$e^{\overline{\beta}(t)} = e^{\gamma\,\beta},$$

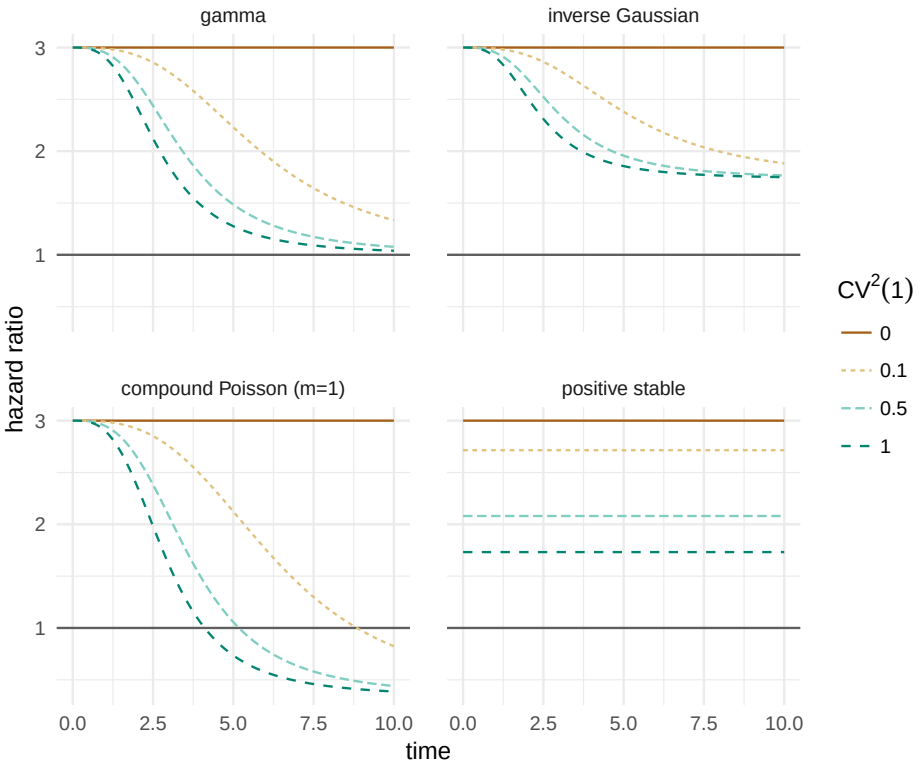

Figure 14: Marginal hazard ratio between two groups of individuals: a high risk one with $\lambda_1(t) = 3\lambda_0(t)$ and a low risk one with $\lambda_0(t) = t^2/20$. For comparability, the distribution are matched by the squared coefficient of variation of the distribution of survivors at time $t = 1$, with $CV^2(1) = \text{var}[Z|T \geq 1]/\text{E}[Z|T \geq 1]^2$.

which does not depend on time, so we have $\overline{\beta}(t) \equiv \overline{\beta} = \gamma\,\beta$. Since $0 < \gamma < 1$, $\overline{\beta}$ is an "attenuated" version of β.

The effect of different frailty distributions on the hazard ratio is illustrated in Figure 14. For the gamma and inverse Gaussian, the marginal hazard ratio approaches 1 with time. For the positive stable distribution, the attenuated marginal effect is observed. For the compound Poisson distribution, a "crossover" is present, where the marginal hazard ratio actually crosses 1. This is the effect of having non-susceptible individuals, represented by the mass at 0 of the distribution. This happens because, in the risk set at large time points, the proportion of non susceptible individuals is higher in the high risk group than in the low risk group.

Implications The shrinking of the hazard ratio in the presence of unobserved heterogeneity has important implications. One is that this might explain moderate familial risks found in clinical studies (Aalen, Valberg, et al., 2014). Moreover, it has an impact for the interpretation of estimated regression coefficients. In the context of a randomized clinical trial with two treatment arms, unobserved heterogeneity induces a loss of balance between the groups. While this may cause an effect as illustrated in Figure 14, it also implies that the estimated marginal hazard ratio does not have a causal interpretation anymore (Aalen, Cook, and Røysland, 2015).

Another phenomenon of interest is the "interruption of treatment", where x may change value at some point, describing the situation where individuals are moved from the treatment to the control group once the treatment does not appear to have any more effect (Aalen, 1994). If the treatment is beneficial, then individuals surviving in the control group will on average have a lower frailty than those in the treatment group. As an artifact, it might seem advantageous to remove individuals from the treatment group after some time, because the control group seems at a lower risk (comprising mostly low-frailty individuals).

1.2.5 Identifiability

In the frailty model, the marginal hazard equals $\overline{\lambda}(t) = \lambda(t)\mathrm{E}[Z \mid T > t]$. If there are no covariates, then only $\overline{\lambda}(t)$ is observed. Without strong parametric assumptions on $\lambda(t)$, is impossible to decide whether frailty is present or not. In other words, the frailty model is not identifiable in this case. The presence of covariates, together with the assumption of proportional hazards conditional on the frailty, make the frailty model identifiable, as long as the frailty distribution has finite expectation. This is due to the marginal non-proportional effect of the observed covariates, as exemplified in Figure 14. This identifiability result has been studied in Elbers and Ridder (1982)

Without further assumptions, observing a time dependent covariate effect of the type shown in Figure 14 is equally compatible with two explanations. One is that the proportional hazards assumption holds in the conditional model, and this effect appears distorted at the marginal level as a result of unobserved heterogeneity. The second is that there is no unobserved heterogeneity, and the observed covariate simply has a time dependent effect. In the first case, the frailty model is the natural choice. In the second case, the modeling strategy would rather include a stratified analysis or an extended Cox model with interactions of covariates with time (Therneau and Grambsch, 2000, ch. 6.5).

In this context, the result of Elbers and Ridder (1982), while theoretically interesting, is of little practical use. Only a firm - and probably naïve - belief in the conditional proportional hazards assumption can substantiate a claim towards the presence of frailty. In principle, this situation changes in the case of clustered survival data, because positive correlation between the event times is induced by the frailty. This is discussed in Section 1.3. The more information on the correlation structure, the easier it is to distinguish the frailty from non proportional hazards. However, when the cluster size is small, the identifiability result, identifying the appropriate model remains a difficult problem.

The positive stable distribution does not have finite expectation, and therefore it does not fall under the Elbers and Ridder (1982) result. As shown in Figure 14, it preserves the proportional hazards assumption at the marginal level. It is not identifiable with univariate survival data, even with covariates. In some sense, this may be seen as an advantage, since it illustrates that the identifiability of univariate frailty models is based on a strong assumption about the mechanism that generated the data. The positive stable distribution does prove useful in the context of clustered failures or recurrent events in Section 1.3.

1.3 Shared frailty models

1.3.1 Missing covariates in paired data

Consider the situation of paired life times, where covariate values are the same for individuals from the same pair. Assume that individuals from a given pair have the same distribution of the event time, denoted as T, with the hazard function $\lambda(t \mid x) = \lambda_0(t)\exp(\beta x)$. Further, assume that x is a realization of a random variable X with density $f_X(x)$. We denote $f(t \mid x)$ and $S(t \mid x)$ as the density and survival function of T, given $X = x$. The marginal survival function of T (where the covariate x is integrated over) is given by $\overline{S}(t) = \int S(t \mid x) f_X(x) dx$.

Consider one pair, with life times T_1 and T_2. The marginal survival function of either T_1 or T_2 is given by $\overline{S}$. However, if $T_1 = t_1$ is observed, the marginal survival function of T_2 will change. Heuristically, if a large life time t_1 is observed, then it is likely that the pair has a low hazard, which in turn makes it more likely that the value of x in that pair is low if $\beta > 0$ (or high if $\beta < 0$). Since x is shared by both individuals, a low hazard for T_1 means that the hazard for T_2 is also low, and that in turn makes it more likely that the corresponding life time t_2 is large as well.

All this leads to positive marginal correlation of the two life times. More specifically, it is straightforward to show that the marginal survival function of T_2, given $T_1 = t_1$, is given by

$$S(t_2 \mid T_1 = t_1) = \int \overline{f}(t_1 \mid x) S(t_2 \mid x) dx,$$

with $\overline{f}(t_1 \mid x) = f(t_1 \mid x) f_X(x) / (\int f(t_1 \mid x) f_X(x) dx)$. Figure 15 shows $S(t_2 \mid T_1 = t_1)$ for $t_1 = 0.1$ and $t_1 = 2$, for the case where the conditional distribution of T_1 and T_2, given $x = 0$, is exponential with mean 1, and $\beta = 1$, and X has a normal distribution with mean 0 and standard deviation σ.

It can be seen that for $t_1 = 2$, the conditional survival curves are higher than the marginal survival curve, while for $t_1 = 0.1$ this is the other way around. For higher standard deviation of the distribution of X, the conditional survival curves are more distinct from the marginal survival function. That means that for higher standard deviation of X, the influence of knowing the value of T_1 is higher, and the correlation between T_1 and T_2 is higher. In fact, one can derive an explicit expression of the correlation between

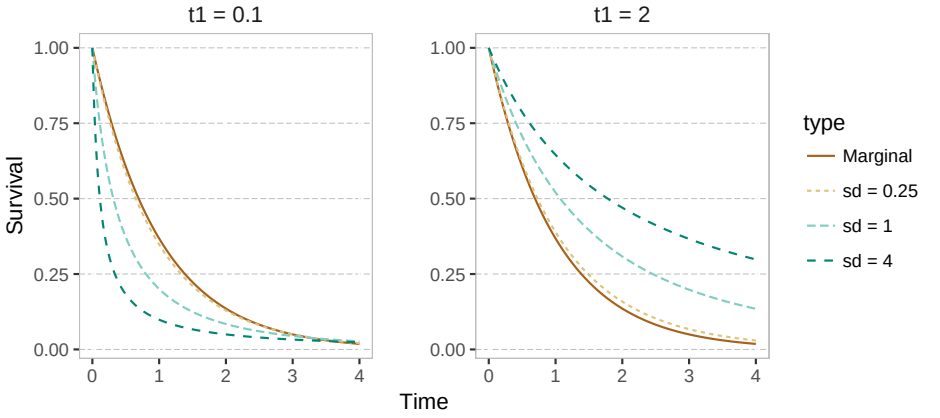

Figure 15: Conditional survival function of T_2, given $t_1 = 0.1$ and given $t_1 = 2$; the conditional distribution of T_1 and T_2 given $X = x$ is exponential with rate $\lambda e^{\beta x}$ and $\beta = 1$, and X has a normal distribution with mean 0 and standard deviation σ^2, with different values of σ.

T_1 and T_2, when the baseline distribution of T_1 is exponential with rate h. It is given by

$$\operatorname{cor}(T_1, T_2) = \frac{e^{2\beta^2\sigma^2} - e^{\beta^2\sigma^2}}{2e^{2\beta^2\sigma^2} - e^{\beta^2\sigma^2}}.$$

A plot of the correlation as a function of σ^2, for $\beta = 1$, is shown in Figure 16.

If the correlation of life times cannot be explained by observed covariates (for example, because x is omitted), then there are two practical approaches. One is marginal modeling, which is in the spirit of general estimating equation (GEE) models. For the Cox model, this involves adjusting the standard errors of the observed covariates (Lin and Wei, 1989). The second is to model the conditional hazard by introducing a "shared" frailty Z, that would take the place of $\exp(\beta x)$ in the previous example. The resulting "shared" frailty model is detailed in Section 1.3.2. The advantage of this approach is that differences between clusters can be quantified, and that the covariate effects have an individual interpretation, as in the case of univariate frailty models.

1.3.2 Clustered failures

The shared frailty model

Assume that there are N clusters and n_i individuals are part of cluster i. The hazard of the jth individual from cluster i is specified as

$$\lambda_{ij}(t|Z_i) = Z_i \exp(\beta^\top \mathbf{x}_{ij})\lambda_0(t). \tag{1.7}$$

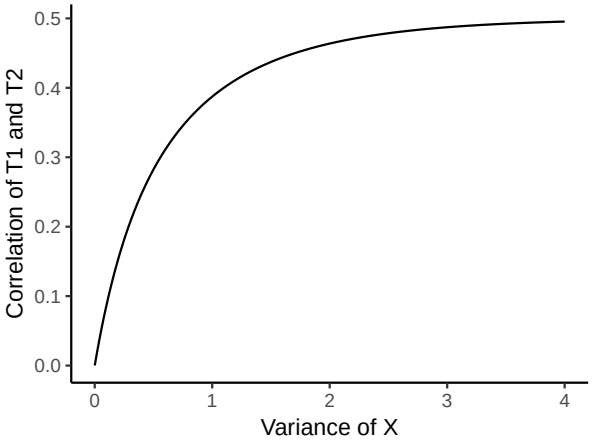

Figure 16: Correlation between T_1 and T_2 as a function of σ^2; the conditional distribution of T_1 and T_2 given $X = x$ is exponential with rate $\lambda e^{\beta x}$ and $\beta = 1$, and X has a normal distribution with mean 0 and variance σ^2.

The individuals in cluster i share the frailty Z_i, and conditional on Z_i their lifetimes are assumed to be independent. While in the univariate case individuals are thought to be a random sample from a larger population of individuals, in the clustered failures case the clusters are thought to be a random sample from a population of clusters.

In the univariate case, the marginal hazard was derived given the individual survival until time t. In the clustered failure case, the marginal hazard is derived given all information about the cluster until time t, including observed events and censorings. This is studied in the following section.

Frailty distributions and clustered failures

Suppose that there are two individuals in a cluster. The conditional cumulative hazard for individuals $j = 1, 2$ is given by

$$\Lambda_j(t) = \int_0^t Y_j(s) \exp(\beta^\top \mathbf{x}_j) \lambda_0(s) ds,$$

where $Y_j(s) = 1$ when individual j is at risk at time s and 0 otherwise. Conditional on Z, the conditional joint survival function of T_1, T_2 is defined as

$$\begin{aligned} S(t_1, t_2|Z) &= P(T_1 > t_1, T_2 > t_2|Z) \\ &= \exp(-Z(\Lambda_1(t_1) + \Lambda_2(t_2))). \end{aligned}$$

The marginal joint survival is obtained by taking the expectation with respect to Z, which results in

$$S(t_1, t_2) = \mathcal{L}(\Lambda_1(t_1) + \Lambda_2(t_2)). \tag{1.8}$$

The Laplace transform of Z, given that individual 1 and 2 are alive at t_1 and t_2, is obtained, with the same arguments as in (1.5), as

$$\mathcal{L}_{Z|T_1>t_1,T_2>t_2}(c) = \frac{\mathcal{L}(c + \Lambda_1(t_1) + \Lambda_2(t_2))}{\mathcal{L}(\Lambda_1(t_1) + \Lambda_2(t_2))}.$$

The only difference from the univariate case is that $\Lambda(t)$ is now replaced by $\Lambda_1(t_1)+\Lambda_2(t_2)$.

Assume now that the event time of the first individual T_1 is observed at t_1. Recall that the density of T is given by $f(t) = \lambda(t)S(t)$. It is then obtained that

$$\begin{aligned}\lim_{dt\downarrow 0} \frac{\mathsf{P}(t_1 \le T_1 < t_1 + dt, T_2 > t_2|Z)}{dt} &= Z\lambda_1(t_1)\exp(-Z(\Lambda_1(t_1) + \Lambda_2(t_2))) \\ &= \frac{\partial}{\partial t_1} S(t_1, t_2|Z).\end{aligned}$$

The Laplace transform of $Z|T_1 = t_1, T_2 > t_2$, defined as

$$\mathcal{L}_{Z\,|\,T_1=t_1,T_2>t_2}(c) = \mathrm{E}[\exp(-cZ)\,|\,T_1 = t_1, T_2 > t_2]$$

can be calculated from Bayes' theorem:

$$\begin{aligned}\mathcal{L}_{Z\,|\,T_1=t_1,T_2>t_2}(c) &= \frac{\mathrm{E}[Z\lambda_1(t_1)\exp(-Z(c + \Lambda_1(t_1) + \Lambda_2(t_2)))]}{\mathrm{E}[Z\lambda_1(t_1)\exp(-Z(\Lambda_1(t_1) + \Lambda_2(t_2)))]} \\ &= \frac{\mathcal{L}'(c + \Lambda_1(t_1) + \Lambda_2(t_2))}{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t_2))}.\end{aligned}$$

More generally, for a cluster of arbitrary size, denote the event history of all its individuals up to some horizon τ as $\mathcal{F}_\tau$. Assume that this comprises $N(\tau)$ observed events, and let

$$\Lambda_\bullet(\tau) = \sum_j \Lambda_j(\tau). \tag{1.9}$$

Denote $\mathcal{L}^{(k)}$ as the k-th derivative of the Laplace transform. The Laplace transform of $Z|\mathcal{F}_\tau$ is given by

$$\mathcal{L}_{Z|\mathcal{F}_\tau}(c) = \frac{\mathcal{L}^{(N(\tau))}(c + \Lambda_\bullet(\tau))}{\mathcal{L}^{(N(\tau))}(\Lambda_\bullet(\tau))}. \tag{1.10}$$

The expectation of this distribution follows as

$$\mathrm{E}[Z\,|\,\mathcal{F}_\tau] = -\frac{\mathcal{L}^{(N(\tau)+1)}(\Lambda_\bullet(\tau))}{\mathcal{L}^{(N(\tau))}(\Lambda_\bullet(\tau))}. \tag{1.11}$$

Therefore, for an individual with covariate vector $\mathbf{x}$ from a cluster where the event history of the cluster is given by $\mathcal{F}_t$, the marginal hazard is

$$\overline{\lambda}(t) = \mathrm{E}[Z|\mathcal{F}_{t-}] \exp(\beta^\top \mathbf{x})\lambda_0(t). \tag{1.12}$$

For the gamma frailty, it is obtained that

$$\mathrm{E}[Z|\mathcal{F}_{t-}] = \frac{\theta + N(t-)}{\theta + \Lambda_\bullet(t-)},$$
$$\mathrm{var}[Z|\mathcal{F}_{t-}] = \frac{\theta + N(t-)}{(\theta + \Lambda_\bullet(t-))^2}.$$

Similar expressions may be derived in a similar fashion for other PVF frailty distributions.

Dependence and the cross-ratio

The estimated frailty variance offers an indication of unobserved heterogeneity between clusters, but it offers little information on the resulting marginal correlation of the event times. Even for paired data, the formulas for the bivariate survival function in (1.8) are difficult to interpret.

One measure of bivariate dependence is Kendall's coefficient of concordance (Kendall's tau). Denote two pairs of individuals as $\{(11), (12)\}$ and $\{(21), (22)\}$, where (ij) refers to individual j of cluster (pair) i. Kendall's tau is defined as

$$\tau_K = \mathrm{E}[\mathrm{sign}(T_{11} - T_{21})(T_{12} - T_{22})],$$

where sign is the sign function. This is proportional to the probability that the order of events are concordant between the two clusters clusters. The median concordance is a similar measure that only involves one pair,

$$\kappa = \mathrm{E}[\mathrm{sign}(T_1 - \mathrm{median}(T_1))(T_2 - \mathrm{median}(T_2))].$$

This is proportional to the probability that the events within the same cluster are concordant, i.e. they occur both before the median survival time or after. In frailty models, both τ_K and κ are positive quantities, since the specification (1.12) only allows for positive dependence. Under independence, both measures would be 0. However, the reverse statement is not usually true.

A more natural way of exploring the within-cluster dependence structure is via the cross-ratio (Oakes, 1989), which compares how the marginal hazard would behave if an event would happen as opposed to an event not happening. Unlike τ_K and κ, this is a local measure of dependence. To illustrate this, we consider one cluster with individuals 1 and 2. Conditional on the frailty, their event times T_1 and T_2 are independent. Denote the marginal hazard of individual 2 if individual 1 is alive at t_1 as

$$\lambda_2(t|T_1 > t_1) = \frac{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t))}{\mathcal{L}(\Lambda_1(t_1) + \Lambda_2(t))}\lambda_2(t),$$

and the marginal hazard of individual 2 if individual 1 had an event at time t_1 as

$$\lambda_2(t|T_1 = t_1) = \frac{\mathcal{L}''(\Lambda_1(t_1) + \Lambda_2(t))}{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t))}\lambda_2(t).$$

These two are marginal hazards under different hypothetical event histories of the other individual in the cluster. They are equal only if there is no dependence between the two individuals. The cross-ratio can be expressed as

$$\begin{aligned}\mathrm{CR}(t) &= \frac{\lambda_2(t|T_1 = t_1)}{\lambda_2(t|T_1 > t_1)} \\ &= \frac{\mathcal{L}''(\Lambda_1(t_1) + \Lambda_2(t))}{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t))}\left(\frac{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t))}{\mathcal{L}(\Lambda_1(t_1) + \Lambda_2(t))}\right)^{-1}.\end{aligned}$$

Intuitively, if there is positive dependence between the two event times, $\mathrm{CR}(t) \geq 1$. Hougaard (2000) suggested that a more interpretable comparison would be to replace the denominator by $\lambda_2(t|T_1 > t)$, to compare the hazard given that "individual 1 died at time t_1" with "individual 1 is alive now". For $t > t_1$, the adjusted cross ratio is defined as

$$\begin{aligned}\mathrm{CR}_a(t) &= \frac{\lambda_2(t|T_1 = t_1)}{\lambda_2(t|T_1 > t)} \\ &= \frac{\mathcal{L}''(\Lambda_1(t_1) + \Lambda_2(t))}{\mathcal{L}'(\Lambda_1(t_1) + \Lambda_2(t))}\left(\frac{\mathcal{L}'(\Lambda_1(t) + \Lambda_2(t))}{\mathcal{L}(\Lambda_1(t) + \Lambda_2(t))}\right)^{-1}.\end{aligned}$$

For $t \leq t_1$, this quantity does not have a direct interpretation.

In Figure 17, we illustrate the unadjusted and adjusted cross-ratio functions for the gamma, inverse Gaussian and positive stable distributions. For comparison purposes, the distributions are matched by Kendall's tau rather than variance. Both unadjusted and adjusted cross-ratio functions show that the marginal hazard of individual 2 is larger if individual 1 has an event. For the unadjusted, the cross-ratio for the gamma frailty is constant, showing that the event of individual 1 affects the hazard in a "proportional" manner. The shape of $\mathrm{CR}(t)$ for the inverse Gaussian and positive stable frailties show that there is a strong immediate dependence that vanished in time.

The adjusted cross-ratio paints a slightly different picture. For the gamma, it implies that, if the partner dies, the hazard for the survivor would appear increasingly larger as compared to the scenario where the partner would still be alive. For the positive stable distribution, the surviving individual is at a perceived high risk right after the partner died, but the differences quickly decreases. This can be interpreted as a large correlation between the life times on the short term. As before, the inverse Gaussian is somewhere in the middle.

$\mathrm{CR}(t)$ may be interpreted as an "instantaneous odds ratio" (Anderson et al., 1992), and for bivariate survival data it may be used for selecting the frailty distribution (Duchateau

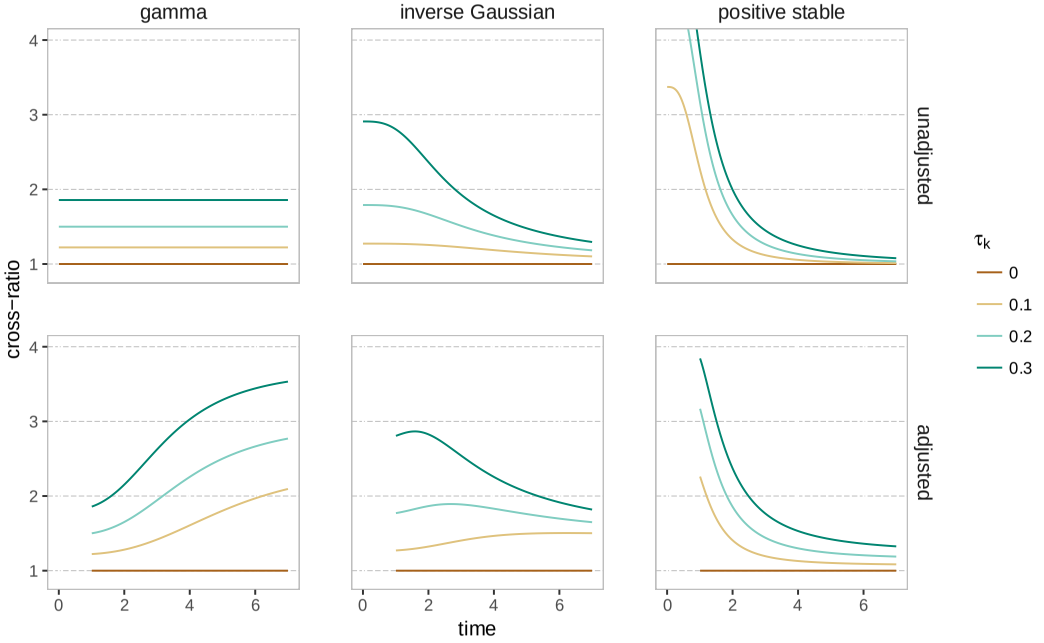

Figure 17: Cross-ratio and adjusted cross-ratio ($t_1 = 1$) for the gamma, inverse Gaussian and positive stable distributions, for different values of Kendall's tau. The individual hazard is taken as $\lambda(t) = t^2/20$.

and Janssen, 2007, ch. 4). One disadvantage is that CR depends on the conditional cumulative hazard; a scaled cross-ratio that overcomes this has been proposed by Paddy Farrington, Unkel, and Anaya-Izquierdo (2012).

The gamma frailty is said to induce "late dependence" (a high probability of events occurring close by at later time points), the positive stable frailty induces "early dependence" (a high probability of event occurring close by early in the follow-up) and the inverse Gaussian is somewhere in the middle. The timing of the dependence can be studied by analyzing the marginal joint distribution of T_1 and T_2 (Hougaard, 2000). A disadvantage of this approach is that the marginal distributions of T_1 and T_2 must be known separately, which is usually not possible.

1.3.3 Frailty model for recurrent events

Recurrent events are most commonly defined in the framework of counting processes. Each individual is described by a process $N(t)$ with the property that $N(t)$ "counts" the number of events experienced by the individual until time t.

The two common frameworks for modeling N are the Poisson process and the renewal process (Cook and Lawless, 2007). If unobserved individual heterogeneity is present, then there are two approaches that may be used in practice. One is the marginal approach, where the unobserved heterogeneity is treated as a nuisance (Cook and Lawless, 1997). In that case, the focus of analysis is the *rate* of N, which is defined as the probability of observing an increase in N in the small interval $(t, t + dt)$.

The second approach is to model the intensity of N. While the hazard is defined as the instantaneous probability of an event given that the individual is alive, the intensity is defined as the instantaneous probability of an event given *the whole previous event history* of the individual. Let Z be the frailty of the individual. The intensity of N can be specified as

$$\lambda(t|Z) = Y(t)Z \exp(\beta^\top \mathbf{x})\lambda_0(t), \tag{1.13}$$

where $Y(t)$ is an indicator function that is 1 if the individual is under observation at time t and 0 otherwise. Similarly to the univariate frailty, the variance of Z describes between-individual unobserved heterogeneity.

The marginal intensity is obtained by replacing Z by $\mathrm{E}[Z|\mathcal{F}_{t-}]$, with $\mathcal{F}_{t-}$ now representing the event history of the individual until just before time t. As in the case of univariate frailty in Section 1.2.4, the effect of the covariates in the marginal intensity is usually time dependent. Similar to the clustered failures scenario, $\mathrm{E}[Z|\mathcal{F}_{t-}]$ includes information on previous event times.

The intensity in (1.13), with t referring to "time since origin of the recurrent event process", is referred to as the *calendar time* or *Andersen-Gill* formulation. Conditional on Z, N is assumed to be a Poisson process, meaning that its intensity conditional on Z does not depend on the history $\mathcal{F}_{t-}$. Alternatively, in the gap-time scale, t refers to "time since the previous event". The intensity may then be expressed as $\lambda(t|Z) = \lambda(t - B(t)|Z)$, where $B(t)$ is the time of the event before time t. From a practical point of view, the gap

time scale has a very similar representation to (1.7), where $\lambda_{ij}(t|Z)$ interpreted as the hazard of the j-th event. Conditional on Z, N is then a renewal process. The resulting marginal intensities are said to define a *modulated* Poisson or renewal processes.

In the case of recurrent events, the frailty mainly expresses that, if two individuals with identical covariates were observed over the same period of time, the expected number of events is larger for the one with the higher frailty. The number of events carries the most information on the frailty (Hougaard, 2000, ch. 9). Therefore, the measures of dependence discussed in Section 1.3.2 are of little interest in this context.

Modeling recurrent events is a complex task and several types of models may be accommodated with counting processes (Therneau and Grambsch, 2000, ch. 8.5). Furthermore, time dependent covariates representing, for example, the number of previous events, may also be added in the model (Aalen, Borgan, and Gjessing, 2008, ch. 8), thus severely complicating the time dependence mechanisms. A comprehensive reference on recurrent event modeling may be found in Cook and Lawless (2007).

1.4 Frailty models in practice

1.4.1 Estimation and inference

Depending on the nature of λ_0, the models may be classified as semiparametric or parametric. In semiparametric models, no assumptions are made on the baseline hazard λ_0 and the maximum likelihood estimate of λ_0 has mass only at the event times, as is the case for the Breslow estimator (Breslow, 1972). In parametric models, λ_0 is determined by a small number of parameters, such as the exponential, Weibull or Gompertz models, or flexible parametric approaches employing spline-based estimators.

Likelihood and EM-based approaches

The likelihood construction for counting processes is detailed in most survival analysis textbooks (Aalen, Borgan, and Gjessing, 2008; Kalbfleisch and Prentice, 2002). To cover all the scenarios described in this chapter, assume that i denotes the cluster, (i, j) the j-th individual within the cluster i and t_{ijk} denotes the k-th event or censoring observed on individual (i, j). We define the event indicator δ_{ijk} as 1 if t_{ijk} is an event time and 0 otherwise. Suppose that the conditional hazard of subject (i, j), conditional on the frailty Z_i is given by $\lambda_{ij}(t|Z_i) = Z_i\lambda_{ij}(t)$ with $\lambda_{ij}(t) = \lambda_0(t)\exp(\beta^\top \mathbf{x}_{ij})$. Denote the at risk indicator of subject j in cluster i by $Y_{ij}(t)$ and let $\Lambda_{i\bullet} = \sum_j \int_0^\infty Y_{ij}(t)\lambda_{ij}(t)dt$ be the sum of conditional cumulative hazards of cluster i, as defined in equation (1.9).

Assuming that the frailties Z_i are observed, the *conditional likelihood* contribution of cluster i is given by

$$L_i(\beta, \lambda_0|Z_i) = \prod_{jk} \left(\lambda_{ij}(t_{ijk}|Z_i)\right)^{\delta_{ijk}} \times \exp\left(-Z_i\Lambda_{i\bullet}\right).$$

and the likelihood for all the individuals is a product of all L_is. The clustered failure data is represented by having only one time point per individual ($k \equiv 1$), while the recurrent events case is represented by having only one individual per cluster ($j \equiv 1$). An implicit assumption here is that censoring is independent. In terms of counting processes, the at-risk process $Y(t)$ is assumed to be independent of $N(t)$, given the covariates and event history up to time t.

In the first part of this expression, Z_i appears to the power N_i, the total number of events from the cluster i. The *marginal likelihood* contribution of cluster i is obtained as taking the expectation over Z_i:

$$L_i(\beta, h_0, \theta) = \prod_{j,k} \lambda_{ij}(t_{ijk})^{\delta_{ijk}} \, \mathrm{E}\left[Z_i^{N_{i\bullet}} \exp\left(-Z_i \Lambda_{i\bullet}\right) \right]. \tag{1.14}$$

For valid inference based on $L(\beta, h_0, \theta)$, the censoring or at-risk process must also not involve the frailty, for reasons outlined in Nielsen et al. (1992). This assumption is similar to that of regular Cox models, where dependent censoring arises, for example, if the censoring process depends on unobserved covariates.

The "posterior" distribution of Z_i, $Z_i|\mathcal{F}_\tau$, has the density kernel

$$f_{Z_i}(z|\mathcal{F}_\tau) \propto z^{N_{i\bullet}} \exp(-z\Lambda_{i\bullet}) f_{Z_i}(z).$$

This is available in closed form only for the gamma frailty. A consequence of this is that the expectation in (1.14) is typically difficult to calculate for other frailty distributions. The expectation of this distribution is also known as the *empirical Bayes* frailty estimate. It can be calculated via the Laplace transform, as discussed in Section 1.3.2. This may involve having to take many derivatives of the Laplace transform, if $N_{i\bullet}$ is large. Another difficulty arises in semiparametric models, where the dimension of λ_0 is usually equal to the total distinct event time points from the data. This prevents a direct maximization of the likelihood.

The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) has been proposed for semiparametric gamma frailty models (Nielsen et al., 1992; Klein, 1992), and can be easily extended to the PVF family of distributions (Hougaard, 2000, ch. 8). This involves iterating between two steps:

1. The "E" step, which involves calculating the expected log-likelihood,

$$\mathrm{E}\ell(\beta, \lambda_0, \theta) = \sum_i \mathrm{E}\left[\log L_i(\beta, \lambda_0|Z)\right].$$

   In practice, this involves calculating $\mathrm{E}[Z_i|\mathcal{F}_\tau]$ and $\mathrm{E}[\log Z_i|\mathcal{F}_\tau]$.

2. The "M" step, where β, h_0 and θ are updated, by maximizing $\mathrm{E}\ell(\beta, \lambda_0, \theta)$.

The advantage of this approach is that the M step may be calculated via Cox's partial likelihood (Cox, 1975), effectively eliminating the problem of the high-dimensional

λ_0. However, the E step still requires numerical integration for distributions except the gamma.

Of the two posterior expectations that are calculated in the E step, $\mathrm{E}[Z_i|\mathcal{F}_\tau]$ may be expressed as a ratio of derivatives of the Laplace transform. The calculation of $\mathrm{E}[\log Z_i|\mathcal{F}_\tau]$ can be avoided via a "profile EM" algorithm, which involves performing the EM algorithm described here for fixed values of θ, resulting in a two-stage maximization. Alternatively, the Monte Carlo EM algorithm may be employed, which involves a stochastic approximation of the E step (Vaida and Xu, 2000).

Alternative approaches

The penalized likelihood method (Ripatti and Palmgren, 2000; Therneau, Grambsch, and Pankratz, 2003) is a very popular way of estimating gamma and log-normal semiparametric frailty models. The basic idea behind it is that, for fixed θ, the $\log Z_i$'s may be treated as regular parameters (on the same scale with the regression coefficients β). Afterwards, a penalization of a specific form is imposed upon them. Depending on the penalization, the results are equivalent to those of a gamma or a log-normal distributional assumption. This approach is typically the fastest for semiparametric models. A downside is that it is not immediately possible to extend the estimation to other frailty distributions.

Other approaches include a pseudo-likelihood method (Gorfine, Zucker, and Hsu, 2006), which leads to consistent estimators and may be employed for a larger number of frailty distributions, and the h-likelihood method (Ha, Lee, and Song, 2001; Ha, Jeong, and Lee, 2017). This approach relies on maximizing the joint likelihood of the observed *and* unobserved data. It has been developed for the gamma and log-normal distributions.

Inference

For parametric models, the variance-covariance matrix is typically obtained directly, as the inverse of the numeric Hessian matrix. This is usually provided by directly by an optimization software.

For models estimated with the EM algorithm, Louis' formula may be used (Louis, 1982) to obtain standard errors of the estimates. It has been shown that the h_0 may be regarded, for practical purposes, as an ordinary finite dimensional parameter and the information matrix may be constructed from the matrix of second derivatives (Andersen, Klein, et al., 1997).

For the profile EM algorithm, the variance covariance matrix for (β, h_0) is obtained under the assumption of fixed θ. Similarly to the penalized likelihood methods, the variance covariance matrix for β, based on the partial likelihood, is also obtained under fixed θ. The complete variance-covariance matrix for β, h_0 or for β should then be adjusted for the variability of θ (Hougaard (2000), ch B3; Putter and Van Houwelingen (2015)), although this is often ignored in practice.

Inference regarding the frailty variance is more challenging. The limiting case, when the variance is 0, is a proportional hazards model without frailty. A likelihood ratio test based on a mixture of χ^2 distributions can be employed to test the difference between these two models (Self and Liang, 1987; Claeskens, Nguti, and Janssen, 2008). Another issue is that, since the variance must be positive, symmetric confidence intervals are not very meaningful. An alternative is to calculate likelihood based confidence intervals, as is illustrated in Therneau and Grambsch (2000, ch. 9).

1.4.2 Software

Support for frailty models exists in major statistical packages such as R (R Core Team, 2017), SAS (Inc., 2003) and Stata (StataCorp, 2017). The `PHREG` command in SAS implements the penalized likelihood method for the gamma and log-normal frailty models. The `streg` procedure in Stata implements parametric gamma and inverse Gaussian frailty models. In what follows we will focus on packages for R.

Semiparametric gamma and log-normal frailty models may be estimated via the penalized likelihood method in the **survival** package (Therneau and Grambsch, 2000; Therneau, 2015a). Semiparametric frailty models with the infinitely divisible class of frailty distributions discussed in Section 1.2.3 may be estimated via the profile EM algorithm with the **frailtyEM** package. Log-normal frailty models (including correlated frailties, discussed in Section 1.5) may be estimated with the **coxme** package (Therneau, 2015b). Similar models may be fitted with the Monte Carlo EM algorithm with the **phmm** R package (Donohue and Xu, 2013). Log-normal and gamma frailty models can also be estimated via h–likelihood with the **frailtyHL** package (Do Ha, Noh, and Lee, 2012). The pseudo-likelihood approach is implemented in the **frailtySurv** package (Monaco, Gorfine, and Hsu, 2017), supporting some of the infinitely divisible distributions from the PVF family.

Parametric and flexible parametric frailty models for the gamma and log-normal distributions are supported by the **frailtypack** package (Rondeau and Gonzalez, 2005; Rondeau, Mazroui, and Gonzalez, 2012) (including correlated random effects, nested random effects and numerous other scenarios). Parametric frailty models with support for some of the PVF family distributions are implemented in the **parfm** package (Munda, Rotolo, and Legrand, 2012).

1.4.3 Data representation

In R (R Core Team, 2017), the canonical resources for survival analysis are found in the **survival** package (Therneau, 2015a). Event histories corresponding to survival times or to recurrent events have a very similar representation, as is described in detail in Therneau and Grambsch (2000).

An event history is represented by a collection of *observations*, which are vectors $(t_L, t_R, \delta, \mathbf{x})$ where (t_L, t_R) are two time points that define an "at-risk" interval, δ is equal

to 1 if the interval ended with an event and 0 otherwise, and $\mathbf{x}$ is a vector of covariate values that are constant on this interval. In R, the tuple (t_L, t_R, δ) is referred to as (`tstart`, `tstop`, `status`). Univariate survival times and clustered failures are usually represented by having $t_L = 0$ and a simplified `(tstop, status)` notation. Furthermore, this notation may also be used to express:

- Recurrent events in calendar time (or "Andersen-Gill" representation). In this case, for an individual, t_R are event times and t_L is usually 0 or the time of the previous event. Usually, the last observation is censored with the last t_R being the end of follow-up.
- Recurrent events in gap time. In this case, $t_L = 0$ and t_R are observed gap times. The last observation may be censored, indicating an incomplete gap time at the end of follow-up.
- Left truncated survival times, where t_L is the time point after which the individual enters the study.
- Time dependent covariates. In this case, if the value $\mathbf{x}$ changes at time $\tilde{t} \in (t_L, t_R)$, this results in two observations corresponding to time intervals $(t_L, \tilde{t})$ and $(\tilde{t}, t_R)$, with the first one being artificially censored.

In the presence of frailty, an observation is interpreted as a contribution to the conditional likelihood of the form

$$L(\beta, \lambda_0 \mid Z; t_L, t_R, \delta, \mathbf{x}) = \left\{ Z\lambda_0(t_R)e^{\beta^\top \mathbf{x}} \right\}^{\delta} \cdot \exp\left(-Z(\Lambda_0(t_R) - \Lambda_0(t_L))e^{\beta^\top \mathbf{x}} \right).$$

For a collection of observations sharing the same frailty Z, the software maximizes

$$\mathrm{E}_Z \left[\prod_{\text{intervals}} L(\beta, \lambda_0 \mid Z; t_L, t_R, \delta, \mathbf{x}) \right],$$

which is the contribution of one cluster to the marginal likelihood (1.14). This is appropriate in the case of recurrent events and time dependent covariates, or for clustered survival times without left truncation.

For left truncated survival times however, this is generally incorrect. In the univariate case, the frailty distribution of a left truncated individual is $Z|T \geq t_L$, referred to as the distribution of survivors in Section 1.2.4.

In the case of clustered survival times, the event of observing the whole cluster must be taken into account (Erikson, Martinussen, and Scheike, 2015; Van den Berg and Drepper, 2011; Jensen et al., 2004). If the individuals from the same cluster have truncation times $t_{L,1}, t_{L,2}, \ldots t_{L,J}$ that are independent given Z, then the frailty distribution of the cluster is $Z|T_1 > t_{L,1}, \ldots T_J > t_{L,J}$.

More complicated selection schemes arise when the left truncation times are not independent, even conditional on the frailty (Rodríguez-Girondo et al., 2018). In the

case of recurrent events, such selection schemes may arise when individuals are included into the study only if they experience a certain number of events (Balan, Jonker, et al., 2016). Such scenarios usually require *ad-hoc* estimation procedures and are not generally supported by the main software packages.

In R, one of the reasons why the same notation is used to denote both recurrent events and left truncation is because they lead to the same likelihood in frailty-less models. In the case of frailty models, the treatment depends on the package used. For example, the **survival** package calculates the correct likelihood for the recurrent events case, **parfm** calculates the correct likelihood for the left truncation case. In **frailtypack** and **frailtyEM**, both scenarios are supported.

1.5 Extensions

In the models discussed in Sections 1.2 and 1.3, the frailty plays the role of a random intercept. In certain scenarios, particularly in studies on bivariate outcomes, correlated random effects have been proposed (Yashin, Vaupel, and Iachine, 1995; Yashin, Iachine, et al., 2001; Wienke, 2010). These address the limitation that shared frailty models may only be employed for positively correlated event times.

Furthermore, the random effect Z has been so far assumed to be time constant. This is consistent with the interpretation that Z accounts for individual specific or cluster specific characteristics that are fixed from the time origin, and have an effect that is constant in time. However, the unobserved heterogeneity might be time dependent, thus better explained by an unobserved random processes that unfolds in time. Several approaches based on this idea have been proposed. The frailty may be modeled with diffusion processes (Yashin and Manton, 1997; Aalen and Gjessing, 2004) or Levy processes (Gjessing, Aalen, and Hjort, 2003). More recently, an approach on birth-death Poisson processes has been proposed (Putter and Van Houwelingen, 2015). Simpler, piecewise constant, frailty models have also been considered (Paik, Tsai, and Ottman, 1994; Wintrebert et al., 2004). A limited implementation combining the birth-death processes and the piecewise constant frailty is implemented in the R package **dynfrail** (Balan, 2017). Related approaches include the constructions of auto-regressive frailty processes based on log-normal frailties (Yau and McGilchrist, 1998; Munda, Legrand, et al., 2016) or gamma frailties (Fiocco, Putter, and Van Houwelingen, 2008).

For the models presented in Sections 1.2 and 1.3 are intended for the analysis of one stochastic event process, it has been assumed that the censoring does not depend on the frailty. This assumption may be tested (Balan, Boonk, et al., 2016), or the event and censoring processes can be jointly modeled. An example is when the observation recurrent event process may be stopped by death (Liu, Wolfe, and Huang, 2004) or when the frailty is also associated with the censoring (Huang and Wolfe, 2002).

Moreover, we assumed that the time dependent covariate vector $\mathbf{x}$ is somewhat "external" to the event process, in the sense of (Kalbfleisch and Prentice, 2002). If $\mathbf{x}$ contains internal time dependent covariates, such as repeated individual measurements, the pro-

cesses should be jointly analyzed (Rizopoulos, 2012, ch. 2). In this case, the frailty is shared by the model for the time dependent covariate (or biomarker) and the model for the event process. Software for estimating joint models is also available in R (Rizopoulos, 2016).

1.6 Outline of the thesis

Frailty models account for unobserved individual or cluster characteristics. In the case of gap-time recurrent events and clustered failure times, they relax the usual independence of event times assumption to a conditional independence assumption. In the case of recurrent events in calendar time, the assumption of a Poisson process is relaxed.

In Chapter 2, the topic of identifiability of shared frailty models is analyzed by means of a simulation study. It has been shown that the univariate frailty model is identifiable, as long as the frailty has finite expectation and covariates are present (Elbers and Ridder, 1982). This result implies that, for univariate survival data, it is very difficult to distinguish between the effect of unobserved heterogeneity can and a possible time dependent effect of the covariates. We analyze how this problem extends to shared frailty models.

In Chapter 3, we study the situation where a recurrent event process may be associated with a terminal event, such as death, due to unobserved factors. Because the recurrent event cannot be observed any more after death, this is an example where the observation of the process is not independent of the process itself. We propose a score test for association between the recurrent event process and the terminal event. This test provides evidence against the usual assumption of independent observation.

In Chapter 4, we analyze the phenomenon of ascertainment of patients in observational studies on recurrent events data. More specifically, we study the case when individuals are included in the study only if at least one event is observed in a specific ascertainment time frame. We propose a solution for accounting for this selection scheme.

In Chapter 5, we discuss maximum likelihood estimation for frailty models and present the implementation from the **frailtyEM** package (Balan and Putter, 2017) in R (R Core Team, 2017). The package, which supports semiparametric estimation of frailty models with distributions from the PVF family, employs a profile expectation-maximization algorithm. Advantages and disadvantages of such approach are discussed, together with a practical demonstration of the software.

Chapter 2

Non-proportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference?

Abstract

Multivariate survival data are frequently encountered in biomedical applications in the form of clustered failures (or recurrent events data). A popular way of analyzing such data is by using shared frailty models, which assume that the proportional hazards assumption holds conditional on an unobserved cluster-specific random effect. Such models are often incorporated in more complicated joint models in survival analysis.

If the random effect distribution has finite expectation, then the conditional proportional hazards assumption does not carry over to the marginal models. It has been shown that, for univariate data, this makes it impossible to distinguish between the presence of unobserved heterogeneity (e.g. due to missing covariates) and marginal non-proportional hazards. We show that difficulties also arise when the data consists of small sized clusters, or individuals experience only a small number of recurrent events.

This chapter is currenlty under review for publication as: T.A. Balan and H. Putter (Forthcoming). Non-proportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference?

We carry out a simulation study to assess the behavior of test statistics and estimators for frailty models in such contexts. The gamma, inverse Gaussian and positive stable shared frailty models are contrasted using a novel software implementation for estimating semiparametric shared frailty models. Two main questions are addressed in the contexts of clustered failures and recurrent events: whether covariates with a time-dependent effect may appear as indication of unobserved heterogeneity, and whether the additional presence of unobserved heterogeneity can be detected in this case. Finally, the practical implications are illustrated in a real-world data analysis example.

2.1 Introduction

Multivariate survival data often arise in biomedical applications. Event times are correlated when individuals are grouped in clusters (e.g. families, patients in hospitals) or observations are clustered within individuals (e.g. recurrent event episodes). Several extensions of the Cox proportional hazards model (Cox, 1972) are used in these contexts (Therneau and Grambsch, 2000, ch. 8–9). A popular class of regression models employs random effects to account for the structure of the data. Shared frailty models commonly assume that the proportional hazards assumption holds conditional on an unobserved cluster specific random effect (Hougaard, 2000, ch. 7).

The frailty model was originally introduced in the context of demographics (Vaupel, Manton, and Stallard, 1979). In this case, an individual-specific random effect (or "frailty") is used to account for individual unobserved heterogeneity. Early research focused on how the frailty may explain different shapes of observed marginal (i.e. population) hazards (Vaupel and Yashin, 1985). The univariate frailty model with covariates and conditional proportional hazards has been shown to be identifiable if the random effect distribution has finite expectation (Elbers and Ridder, 1982). Distributions for which the moments are not well defined, such as the positive stable, are not usually identifiable with univariate data (Hougaard, 1986b).

In univariate frailty models, the marginal hazards and marginal covariate effects may differ from the conditional ones (Vaupel and Yashin, 1985; Aalen, 1994). In particular, under some regularity assumptions Elbers and Ridder, 1982, the marginal hazards are "dragged down" and the marginal hazard ratios are shrunk towards 1. The same effect is observed in the presence of unobserved heterogeneity due to missing covariates (Hougaard, 2000, ch. 2.4.6). In particular, the marginal covariate effects are time-dependent, and such models are not compatible with a proportional hazards assumption on the population hazards (Therneau and Grambsch, 2000, ch. 6.6). One implication of this is that, in practice, the frailty model with conditional proportional hazards and a Cox regression with a time-dependent covariate effect can not usually be distinguished on the basis of the data alone.

Another implication of the identifiability result Elbers and Ridder, 1982 is that frailty models for multivariate survival data are also identifiable under the same conditions. Shared frailties are used to model common unobserved risk, where observations within

cluster are independent conditional on the random effect and marginally dependent. Therefore, the estimated spread (e.g. variance) of the frailty distribution measures both the strength of dependence and between-cluster unobserved heterogeneity.

When the cluster size is small and covariates are present however, the regression parameters and the dependence structure may be confounded (Hougaard, 2000, ch. 7.2.7), since the frailty model is identifiable also by considering only one event time from each cluster. This is a well-known problem in twin studies, where more complicated random effect structures might be more appropriate (Yashin, Iachine, et al., 2001). Nevertheless, shared frailty models are commonly used in the context of twin studies without considering the possible impact of time-dependent covariate effects (Gharibvand and Liu, 2009; Gerster, Madsen, and Andersen, 2014; Dai et al., 2013). Conversely, in a twin study on depression (Kendler et al., 2009), the authors found covariate effects that decay over time and fitted a model for non-proportional hazards, which might be a by-product of unobserved common risk.

In this chapter, we study the degree to which the distinction between non-proportional covariate effects and the presence of unobserved heterogeneity can be made in practice. In particular, the behaviour of shared frailty models is assessed on data sets where a time-dependent covariate effect is present. The impact of cluster size and sample size is ascertained by means of a simulation study, in the context of both clustered failures and recurrent events.

This chapter is structured as follows. In Section 2.2, we discuss the theoretical background of proportional hazards models and frailty models, in Section 2.3 we present the results of a simulation study comprising a large number of scenarios, in Section 2.4 we review real life data analysis scenarios and we present the conclusions of this study and discussion in Section 2.5.

2.2 Models

2.2.1 Proportional hazards models

In Cox-type proportional hazards models, the hazard of individual j from cluster i is specified as

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_0(t)\exp(\mathbf{x}_{ij}^\top\beta), \tag{2.1}$$

where $Y_{ij}(t)$ is an indicator function which is 1 when individual (i, j) is at risk and 0 otherwise, $\lambda_0(t)$ is an unspecified "baseline" hazard, $\mathbf{x}_{ij}$ is a $p \times 1$ vector of observed covariates and β is a $p \times 1$ vector of unknown regression coefficients.

This formulation covers both the clustered failures and recurrent events scenarios in gap-time (in the latter, (i, j) symbolizes the j-th episode of individual i). For recurrent events in the Andersen-Gill or calendar time formulation, it is common to take $j \equiv 1$, and in this case λ_i represents the intensity (or "hazard process") of the recurrent event process. The case of univariate survival data may be seen as either that of clustered failures with only one individual per cluster, or that of recurrent events with at most

one event per individual. For simplicity, $\mathbf{x}$ is taken constant in time here, although time-dependent covariates are easily accommodated (Kalbfleisch and Prentice, 2002). It is assumed that the censoring is independent, given $\mathbf{x}$ and the event history.

When the proportional hazards assumption does not hold, the observed effect of the covariates is time-dependent. In this case, the hazard can be specified as

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_0(t)\exp(\mathbf{x}_{ij}^\top\beta(t)). \tag{2.2}$$

The assumption of proportional hazards can be visualized for a small number of covariates or tested using Schoenfeld residuals (Grambsch and Therneau, 1994).

2.2.2 Frailty models

In frailty models, the hazard is specified conditional on a cluster-specific random effect Z_i:

$$\lambda_{ij}(t|Z_i) = Y_{ij}(t)Z_i\exp(\mathbf{x}_{ij}^T\beta)\lambda_0(t). \tag{2.3}$$

Z_i is referred to as the "frailty" of cluster i. The Z_i's are taken as iid random variables with a distribution with positive support. In addition to the censoring assumptions of model (2.1), it is also assumed that the censoring does not depend on the frailty Z_i (Nielsen et al., 1992).

Denote the Laplace transform of Z as $\mathcal{L}(c) = \mathrm{E}[\exp(-cZ)]$ and its k-th derivative as $\mathcal{L}^{(k)}(c)$. A large family of infinitely divisible distributions is described in Hougaard, 2000, with the form

$$\mathcal{L}(c) = \exp(-\alpha\psi(c;\gamma)). \tag{2.4}$$

This so-called Power-Variance-Function (Hougaard, 1986b) family of distributions includes the gamma, inverse Gaussian, positive stable, and compound Poisson distributions. The parametrizations of the distributions used in the rest of this chapter are detailed in the Appendix.

The marginal hazard corresponding to (2.3) is given by

$$\bar{\lambda}_{ij}(t) = \mathrm{E}[Z_i|O_i(t_-)]\exp(\mathbf{x}_{ij}^\top\beta)\lambda_0(t) \tag{2.5}$$

where $O_i(t_-)$ is the observed event and covariate history of cluster i up to (but not including) time t and $E[Z_i|O_i(t_-)]$ is the "posterior" expectation of Z_i given $O_i(t_-)$. If $N_i(t)$ denotes the number of events observed in the cluster i by time t, then this expectation is equal to

$$\mathrm{E}[Z_i|O_i(t_-)] = -\frac{\mathcal{L}^{(N_i(t)+1)}(\Lambda_i(t))}{\mathcal{L}^{(N_i(t))}(\Lambda_i(t))} \tag{2.6}$$

where

$$\Lambda_i(t) = \sum_{j=1}^{J_i}\int_0^t Y_{ij}(s)\exp(\mathbf{x}_{ij}^\top\beta)\lambda_0(s)ds,$$

and $\mathcal{L}^{(p)}(c)$ denotes the p^{th} derivative of $\mathcal{L}$. Consider that $\mathbf{x}_{ij} \equiv x_{ij} \in \{0, 1\}$. The marginal survival curve for a group defined by a fixed value of x is given by

$$\bar{S}_x(t) = \mathrm{E}\left[\exp\left(-Z\int_0^t e^{\beta x}\lambda_0(s)ds\right)\right] = \mathcal{L}\left(e^{\beta x}\Lambda_0(t)\right).$$

The marginal cumulative intensity (or hazard) for a given x is then given by $\bar{\Lambda}_x(t) = -\log \bar{S}_x(t)$ and the marginal intensity (hazard) as $\bar{\lambda}_x(t) = d/dt\bar{\Lambda}_x(t)$. For a binary covariate x, the conditional hazard ratio e^{β} is then interpreted as the hazard ratio between two individuals with the same frailty. By contrast, the marginal hazard ratio $\bar{\lambda}_1(t)/\bar{\lambda}_0(t)$ is the observed (usually time-dependent) ratio of the hazards of the two groups.

2.2.3 Non-proportional hazards

Non-proportional hazards in univariate data The frailty model (2.3) represents a model where the proportional hazards assumption holds conditional on the Z_i. As a function of $\mathbf{x}_{ij}$, the marginal hazard (2.5) is in general a model of the type (2.2), where the marginal covariate effects are time-dependent. In Figure 21, we show, for different frailty distributions and degrees of dependence, the marginal hazard ratio between two groups of individuals that have a conditional hazard ratio of 5. The perceived attenuation of the hazard ratio reflects that the two groups become more homogeneous in time, as individuals with a higher frailty leave the data set sooner. However, from a practical point of view, the same hazard ratio might be explained by a true reduction in the effect of the covariate at the individual level (e.g. treatment effect decreasing in time).

In the case of univariate survival data, if Z has finite variance, the marginal hazards are not proportional (Aalen, 1994). The intuition behind the identifiability result (Elbers and Ridder, 1982) relies on the fact that this observed departure from proportional hazards is considered to be a product of unobserved heterogeneity. If the frailty distribution does not have finite expectation, then the model is not necessarily identifiable. An example is the positive stable distribution, which shows marginal proportional hazards, as seen in Figure 21. Therefore, in the univariate case, a time-dependent covariate effect may give the impression of unobserved heterogeneity.

Non-proportional hazards in multivariate data In the case of multivariate survival data, an unobserved cluster effect induces positive dependence between these observations. If no such dependence is observed, then the shared frailty model can not be a suitable model for the data. The presence of the within cluster correlation structure indicates that the (shared) frailty model does not appear to be confounded with a possible time-dependent covariate effect. In other words, the shared frailty model must also be compatible with the observed joint distribution of the event times.

However, there are cases when no real dependence structure is observed. An extreme example would be that of the analysis of lifetimes of fathers and daughters in the presence of a strong risk factor (Hougaard, 2000). Even if all daughters would be censored

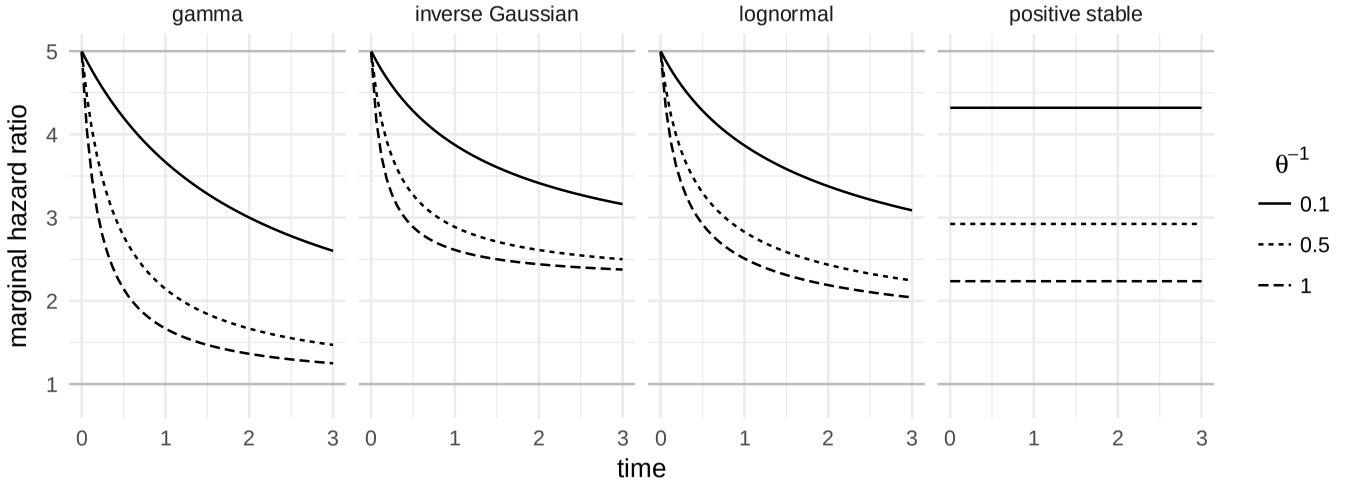

Figure 21: Marginal hazard ratio of survivors obtained a conditional hazard ratio of 5, for the gamma, inverse Gaussian, lognormal and positive stable distributions, where the baseline hazard is $\lambda_0(t) = 1$. The gamma, inverse Gaussian and lognormal have fixed $\mathrm{E}Z = 1$ and $\mathrm{Var}Z = \theta^{-1}$. For the positive stable, θ may still be used as a measure of association, although it is not comparable with the others. The parametrizations used here are detailed in the Appendix. Horizontal lines are added at $y = 5$ (corresponding to $\theta^{-1} = 0$) and at $y = 1$ (corresponding to no covariate effect).

and no relation between their lifetimes and the father's lifetimes can be inferred, the shared frailty model may be estimated. In particular, the model is identifiable, because of the observed covariate. Therefore, the *amount* of observed dependence is important in whether a time dependent marginal hazard ratio may be attributed to a common-risk frailty effect.

The main question posed by this observation is: how *much* of the dependence structure must be observed so that a time-dependent covariate effect does not appear as evidence in favor of the shared frailty model? This is studied in the following section, in the context of three scenarios: clustered failures where an observed covariate may vary within cluster, clustered failures where the observed covariate only varies between clusters, and recurrent events where the observed covariate varies between individuals.

2.3 Simulation study

2.3.1 General framework

We consider $x \sim \text{Bernoulli}(0.5)$ a binary covariate. First, data are simulated from a model without unobserved heterogeneity, but with a time-dependent effect of x. Specifically, this is a model of the type (2.2). On the simulated data sets, four models are estimated: a Cox proportional intensity model and frailty models with gamma, inverse Gaussian and positive stable distributions. The Commenges-Andersen test for heterogeneity (Commenges and Andersen, 1995) and, for the frailty models, the likelihood ratio test are evaluated. Furthermore, all estimates and confidence intervals are collected. A test for the proportional hazards assumption (Grambsch and Therneau, 1994) is also evaluated, to determine the degree of non-proportionality in each simulated data set. Second, this is repeated by having data simulated also with unobserved heterogeneity in addition to the time-dependent covariate effect.

Three main scenarios are analyzed. The first is that of clustered failures, with cluster sizes 1 (univariate survival), 2, 3, 5 and 10, and x simulated independently for each individual. The second is identical to the first scenario, with the exception that x is simulated independently for each cluster. Lastly, recurrent events in calendar time are simulated (Jahn-Eimermacher et al., 2015), with x simulated independently for each individual. In the recurrent events case, 1, 2, 3, 5 and 10 events are simulated for each individual.

Two distributions are considered to simulate data with time-varying covariate effects. The Weibull baseline with shape α and scale γ, where the covariate effect is taken to have an interaction with log time, leading to

$$\lambda_{ij}(t|Z_i;\alpha,\gamma) = Z_i\alpha\gamma t^{\alpha-1}\exp\left((\beta_0+\beta_1\log t)x_{ij}\right), \tag{2.7}$$

which is again a Weibull distribution with shape $\alpha+\beta_1 x_{ij}$ and scale

$$Z_i\alpha\gamma e^{\beta_0}(\alpha+\beta_1 x_{ij})^{-1}.$$

Both shape and scale parameters must be positive. In the case of clustered failures, this is the hazard while in the case of recurrent events this is taken as the intensity of the

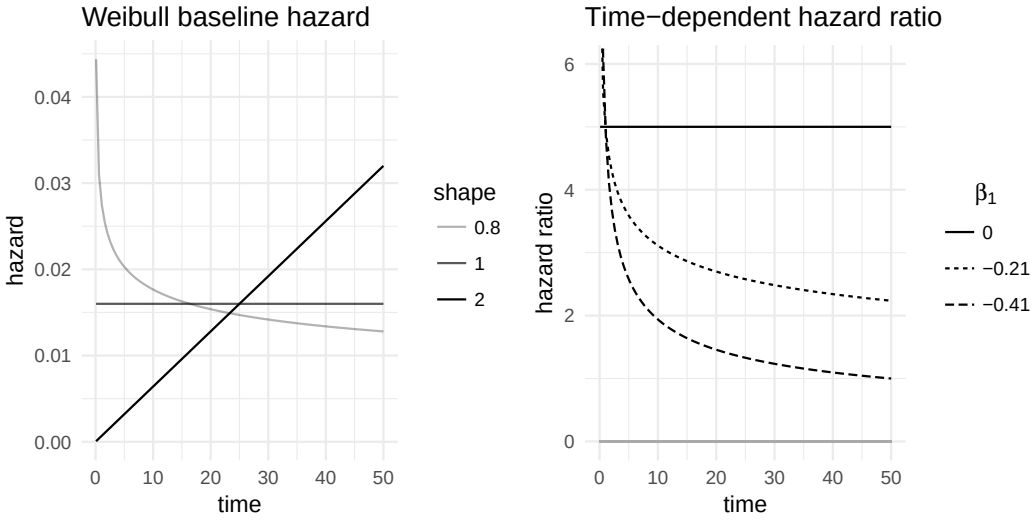

Figure 22: Left: Weibull baseline hazards used in the simulation, where the scale parameter is chosen so that the cumulative baseline hazard at 50 is 0.8. Right: time dependent hazard ratio used in the simulation and describe in equation (2.7), i.e. $5 \exp(\beta_1 \log t)$.

recurrent events process. The baseline intensity is a decreasing function of time if $\alpha < 1$, and decreasing for $\alpha > 1$. For $\alpha = 1$, the exponential distribution is obtained, where the hazard is constant.

The second distribution used in our simulations is the Gompertz distribution, using an interaction with time instead of log time. However, the Gompertz distribution has an increasing hazard regardless of the parameter choice. Henceforth, we only report results on the Weibull distribution.

The shape parameter of the Weibull distribution is taken as $\alpha \in \{0.8, 1, 2\}$, corresponding to a decreasing, constant and increasing intensity. For the clustered failures scenarios, the scale parameter is chosen so that the cumulative baseline intensity $\Lambda_0(50) = 0.8$. The different hazard shapes are shown in Figure 22. The covariate effects are defined as in (2.7), with $\beta_0 = \log(5)$, and 3 values for β_1, denoted as $\beta_1^{(0)}, \beta_1^{(1)}$ and $\beta_1^{(2)}$, corresponding to different degrees of time-dependent effect. $\beta_1^{(2)}$ is selected so that $\beta_0 + \beta_1^{(2)} \log 50 = 0$); $\beta_1^{(1)}$ is taken as the average of 0 and $\beta_1^{(2)}$, and $\beta_1^{(0)} = 0$ corresponds to the proportional hazards scenario. The corresponding hazard ratios for $\alpha = 0.8$ are visualized in Figure 22. To keep the results comparable across scenarios, for the recurrent events with j events for an individual, the scale parameter is chosen so that $\Lambda_0(50) = 0.8j$. Therefore, the average number of events can be compared to a cluster with j individuals.

Artificial censoring is imposed in each data set so that, on average, the earlier 70% events are observed. The censoring time is determined by simulation for each scenario

and combination of parameters. For the recurrent events, all individuals are censored at the 0.7 quantile of all (uncensored) event times. All calculations are performed in the R software (R Core Team, 2017), using the packages `survival` (Therneau, 2015a) and `frailtyEM` (Balan and Putter, 2017).

2.3.2 Likelihood Ratio Test

The likelihood ratio test (LRT) is usually used to test the null hypothesis of *no frailty*. For the gamma and inverse Gaussian, this is equivalent to testing $H_0 : \text{Var}[Z] = 0$ versus $H_A : \text{Var}[Z] > 0$, but similar considerations hold for the positive stable frailty model. The model under H_0 is equivalent to a Cox proportional intensity model assuming independent observations. It is common to approximate the distribution of the LRT statistic under H_0 by a mixture distribution $\left(\chi^2(1) + \chi^2(0)\right)/2$ (Zhi, Grambsch, and Eberly, 2005; Claeskens, Nguti, and Janssen, 2008). This result is provided by the `emfrail` function in the `frailtyEM` R package.

No frailty When no frailty is included in the simulation, the percentage of rejections of H_0 is shown in Figure 23, for the gamma frailty model and Weibull shape parameter is $\alpha = 0.8$. Alongside this is the percentage of rejections of the null hypothesis of the ZPH test for proportionality (Grambsch and Therneau, 1994).

When the data are indeed simulated with proportional hazards ($\beta_1 = 0$), the percentage of rejections for both tests is close to the nominal alpha level of 5% across all scenarios, regardless of cluster size. When the hazards are not proportional ($\beta_1 < 0$), the percentage of rejections grows with total sample size. For larger cluster sizes, the LRT shows a decreasing number of false positives. In particular, for smaller clusters, there is a visibly large proportion of rejections, even when the time-dependent covariate effect is moderate. The rate of rejections of the ZPH test does not appear to be strongly influenced by the cluster size. Whether the covariate varies within the cluster (the "clustered" case) or only between clusters ("clustered/common" case) does not make a practical difference. These observations carry over also for the recurrent events. The conclusion is that, the time-dependent covariate effect alone may appear as evidence in favor of the gamma frailty model, unless the cluster size is moderate to large. The results for the inverse Gaussian frailty are very similar to those of the gamma frailty and can be found in the supplementary material.

For the positive stable distribution, the corresponding results are shown in Figure 24. In the case of clustered events, the LRT shows around 5% rejections regardless of the degree of non-proportionality. However, when the covariate does not vary within cluster or in the case of recurrent events, where the covariate is constant for each individual, the large amount of non-proportionality may still be somewhat confounded with unobserved heterogeneity. This is explained by the fact that, in these cases, there is virtually no observed within-cluster heterogeneity. Therefore, the differences explained by x are essentially confounded with the differences that may be explained by cluster-specific

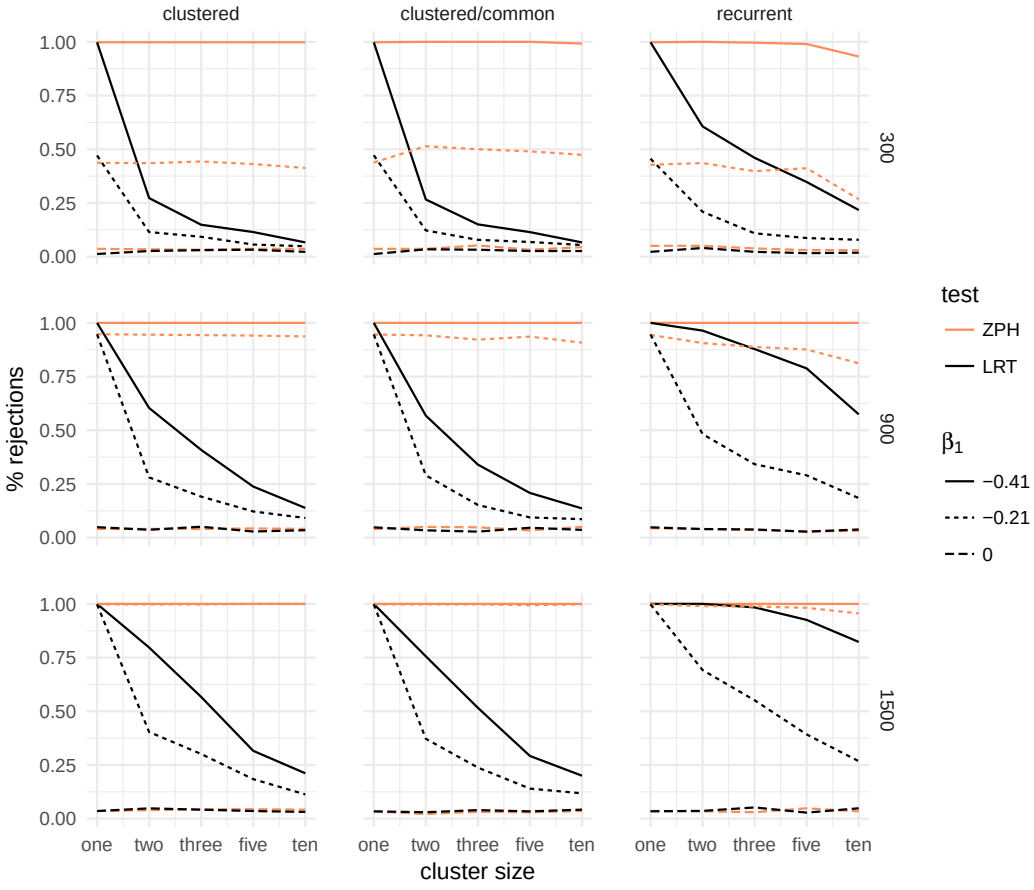

Figure 23: Percentage of rejections of the likelihood ratio test (LRT) between a gamma frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated without unobserved common risk and an increasing Weibull baseline hazard with shape $\alpha = 0.8$. The rows correspond to the total sample size (300, 900, 1500) and the columns to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

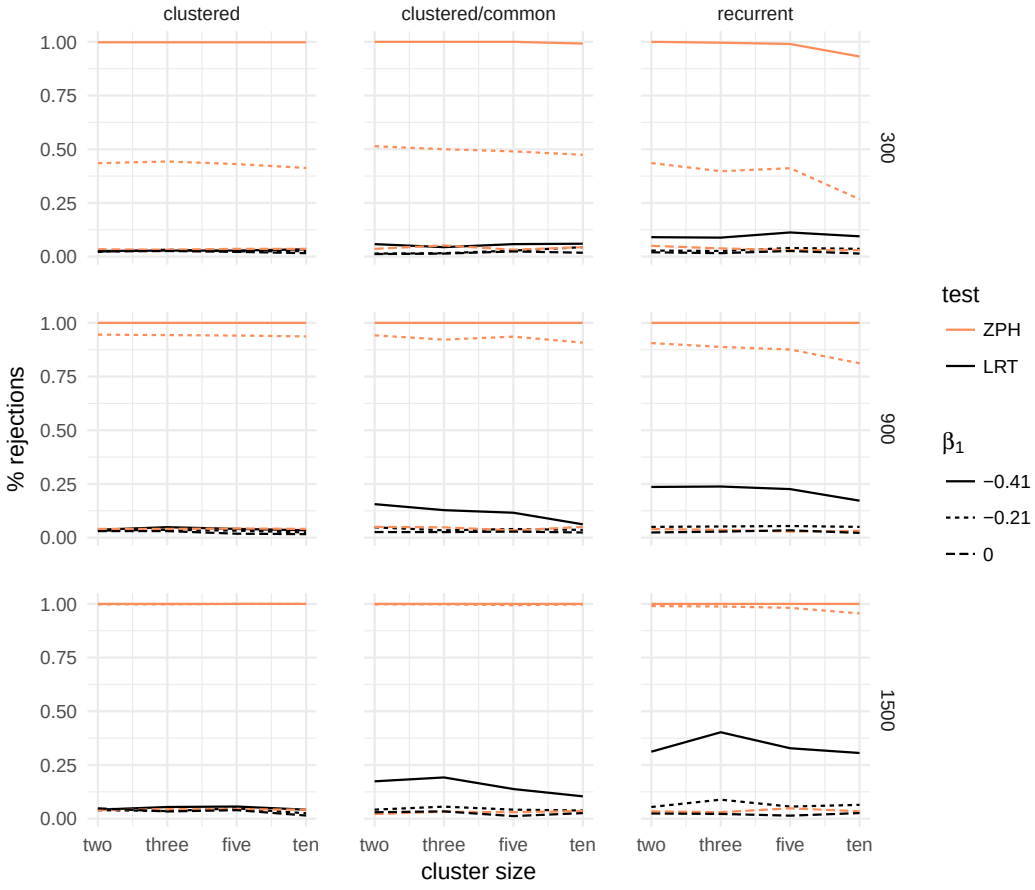

Figure 24: Percentage of rejections of the likelihood ratio test (LRT) between a positive stable frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated without unobserved common risk and an increasing Weibull baseline hazard with shape $\alpha = 0.8$. The rows correspond to the total sample size $(300, 900, 1500)$ and the columns to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

unobserved heterogeneity. The conclusion is that the positive stable distribution is not affected by the departures from proportionality as long as there is some within-cluster variation of the observed covariates.

Frailty When the data are simulated as before, but also with unobserved heterogeneity, the percentage of rejections of the LRT is larger, as expected, and the ZPH test rejects the null hypothesis more than 5% of the time. This is due to the fact that marginal non-proportionality arises both from the time-dependent covariate effect and from the frailty effect.

The results for the gamma frailty model are shown in Figure 25. Even under conditional proportional hazards ($\beta_1 = 0$), the LRT rejects the null hypothesis more than 5% of the times. In the scenarios where the covariate does not vary between clusters (including the recurrent events), the power of the ZPH test increases with cluster size. Therefore, presence of such a time-dependent covariate effect in addition to unobserved heterogeneity increases the power of the LRT.

The results for the positive stable frailty model are shown in Figure 26. In this case, a visible effect is that of the degree of non-proportionality. A stronger time-dependent effect of the covariate leads to a substantially larger proportion of rejections.

Although the data were simulated with unobserved heterogeneity, the difference in the rate of rejections when $\beta_1 < 0$ as compared to $\beta_1 = 0$ may be regarded as *rejecting the null hypothesis for the wrong reasons*.

In conclusion, time-dependent covariate effects may appear as evidence in favor of frailty models, even if unobserved heterogeneity does not actually exist. If that exists, then the non-proportionality of the covariate effect may lead to overestimating the evidence in favor of the frailty model. The results for other shapes of the baseline hazard (and for the inverse Gaussian distribution) are shown in the supplementary material. Similar conclusions apply in those cases as well, although the percentage of rejections is the largest for the decreasing baseline hazard (shown here). This is explained in part by the fact that, with a decreasing hazard, events occur earlier on in the follow-up, leading to earlier censoring. The resulting smaller window of observation makes the *observed* time-dependent hazard ratio more compatible with the one predicted by the frailty models shown in Figure 21.

2.3.3 Commenges-Andersen test

The Commenges-Andersen (CA) test for heterogeneity shows in general the same behaviour as the LRT from the gamma frailty or inverse Gaussian frailty models, albeit with slightly fewer rejections. This is not surprising, since it is a score test, which are generally less powerful than LRT's. For example, in Tables 21, 22 and 23 the CA, LRT and ZPH tests are shown side-by-side for varying cluster sizes for total sample size of 300 and Weibull shape parameter 1.

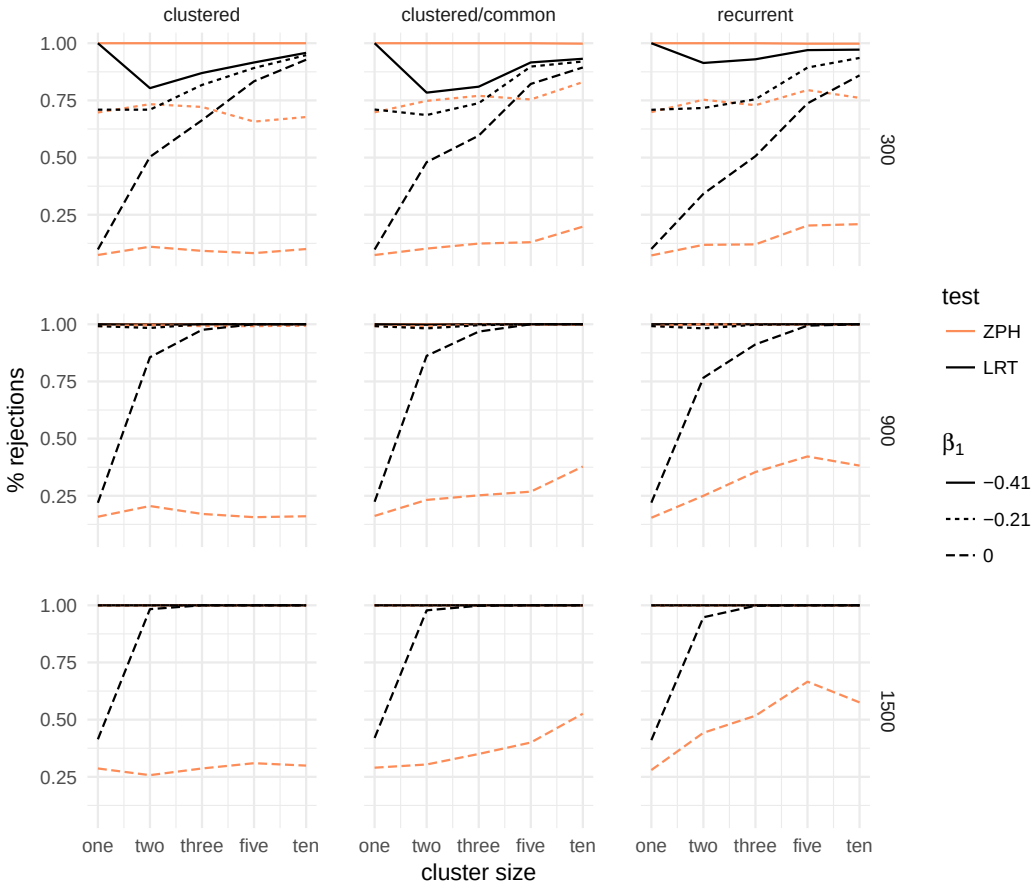

Figure 25: Percentage of rejections of the likelihood ratio test (LRT) between a gamma frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 1 and variance 0.25 and an increasing Weibull baseline hazard with shape $\alpha = 0.8$. The rows correspond to the total sample size (300, 900, 1500) and the columns to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

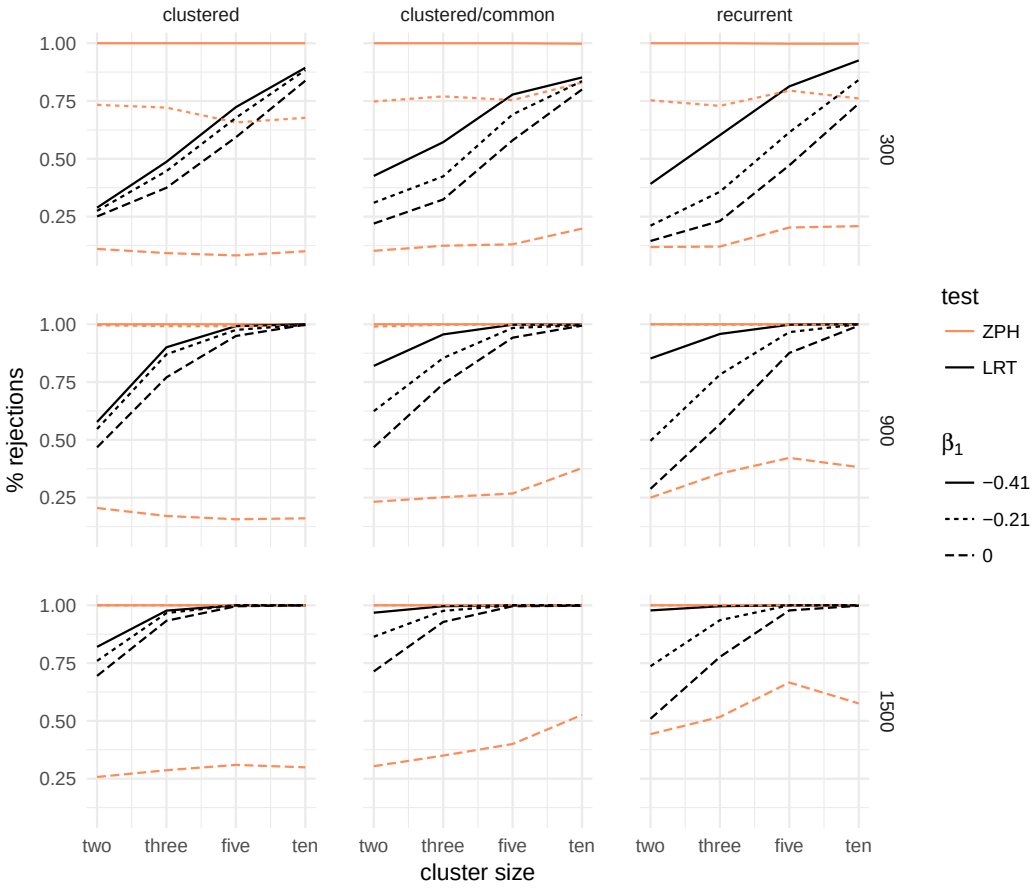

Figure 26: Percentage of rejections of the likelihood ratio test (LRT) between a positive stable frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 1 and variance 0.25 and an increasing Weibull baseline hazard with shape $\alpha = 0.8$. The rows correspond to the total sample size (300, 900, 1500) and the columns to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

2.3.4 Estimated frailty variance

In the case of the gamma frailty, the estimated frailty variance is often considered an indication of the strength of the frailty effect. For the univariate case, these estimates were very large under all scenarios of non-proportionality. In the data sets simulated without frailty, the estimates decrease towards 0 with increasing cluster size and are not influenced by the total sample size across all scenarios, while they are larger with increased departure from proportional hazards. When data sets were simulated with frailty, a similar phenomenon is observed, although the estimates approach a value close to 0.25, which is the variance of the lognormal simulated frailty. This is illustrated, for a total sample of 900 and for the decreasing and constant hazard shapes in Figure 27.

The coverage of the frailty variance estimates can be analyzed with the likelihood-based confidence intervals implemented in the `frailtyEM` package. There is a 1-1 correspondence between the lower bound of this confidence interval being 0 and the rejection of the LRT null hypothesis. As expected, in the univariate case, the coverage is almost 0 under non-proportionality, and it improves with larger cluster size. The degree of departure from proportionality, as in the case of the LRT, plays a large role in determining whether the confidence interval of the estimated frailty variance includes 0 or not. For a total sample of 900 and for the decreasing and constant hazard, this is shown in Figure 28.

2.3.5 Cumulative hazard

As shown in Section 2.2, the observed hazard ratio of the groups defined by the values of x can be determined by integrating out the frailty. In the case of no frailty and $\beta_1 = 0$, all methods estimate roughly the same cumulative marginal hazard at the end of follow-up. If $\beta_1 < 0$, the models also act similarly: the fitted cumulative hazard for $x = 0$ is larger and that for $x = 1$ is lower, resulting in the shrinkage phenomenon shown in Figure 21.

In the case when a frailty effect is also included in the simulation, the gamma and inverse Gaussian show similar results. The positive stable distribution is slightly closer to the marginal Cox model, since both models specify a marginal model where the hazards are proportional.

2.4 Application

Kidney Cathether Insertions

The kidney catheter data (McGilchrist and Aisbett, 1991) have often been used to illustrate the use of frailty models for recurrent events. Recurrent times to infection for 38 patients that use portable dialysis equipment were recorded. A gap time may be censored when the catheter is removed for a reason other than infection. At most two gap times are included for each individual. For 23 patients, there were two observed events, for 12 patients there was one observed event and one censored, while for 3 patients both

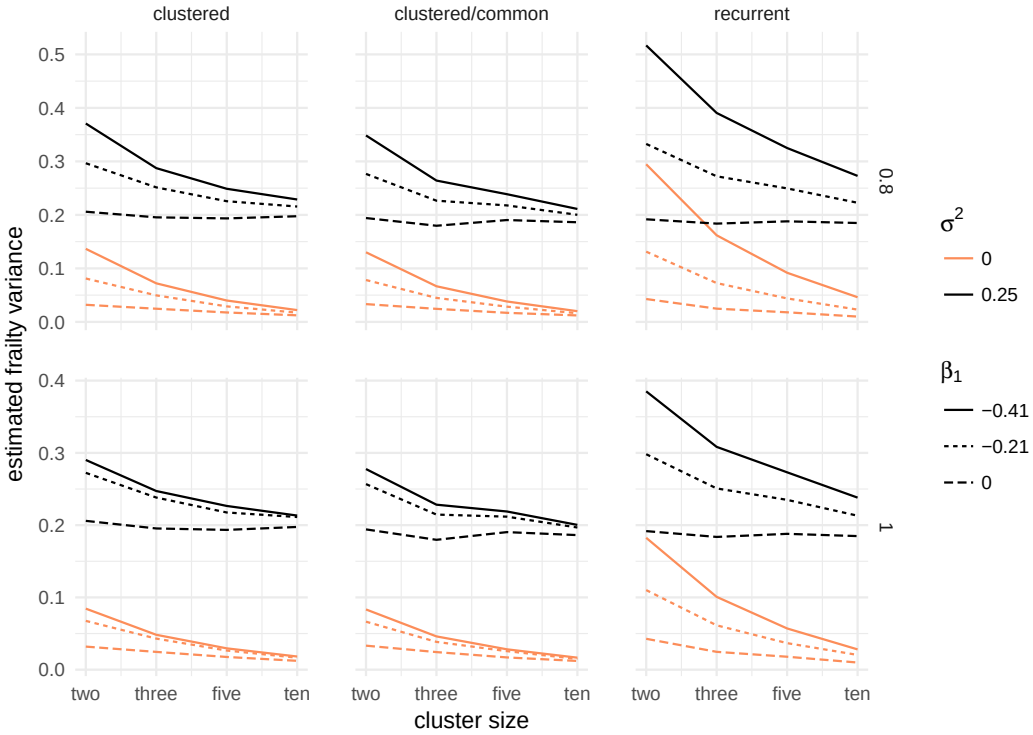

Figure 27: Estimated frailty variance for a gamma frailty model, when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 1 and variance $\sigma^2 \in \{0, 0.25\}$ and a total sample size of 300. The rows correspond to the Weibull baseline shape parameter, increasing for $\alpha = 0.8$ and constant for $\alpha = 1$. The columns correspond to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

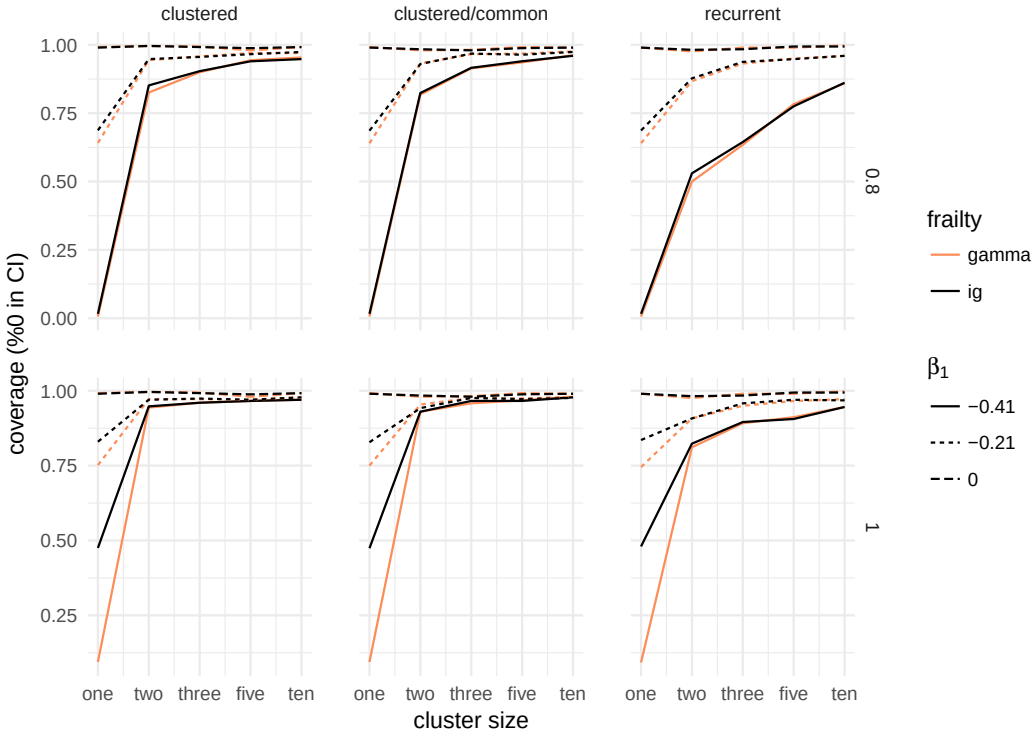

Figure 28: Coverage of the likelihood-based confidence interval for the gamma frailty variance for the gamma and inverse Gaussian distributions, when the data are simulated with no unobserved heterogeneity (true variance is 0) and a total sample size of 300. The rows correspond to the Weibull baseline shape parameter, increasing for $\alpha = 0.8$ and constant for $\alpha = 1$. The columns correspond to the three main simulation scenarios: clustered failures, clustered failures where the observed covariate only varies between clusters, and recurrent events. β_1 indicates the strength of the time-dependent covariate effect.

gap times were censored. The observed covariates consist of age, sex and disease type (4 level categorical variable).

The data set is included in the `survival` package (Therneau, 2015a) in the R statistical software (R Core Team, 2017). A gamma frailty model without any covariates leads to an estimated frailty variance of 0.177 with a 95% CI $[0, 0.985]$, which is not significant ($p = 0.259$ for the LRT, $p = 0.22$ for C-A). While the addition of age does not impact the model fit in an important way, the addition of sex leads to an estimated frailty variance of 0.388 with a 95% CI $[0.04, 1.01]$, which is significant ($p = 0.012$ for the LRT, $p = 0.002$ for the Commenges-Andersen test). The effect of sex is also highly significant, with $\beta = -1.55\,(0.49)$. With the removal of an outlier (a male with very long observed gap times), the evidence in favor of the frailty model disappears (Therneau and Grambsch, 2000, ch. 9.5), where the authors note that *with this subject in the model, it is a toss-up whether the disease or the frailty term will be credited with "significance"*. Nevertheless, it is remarkable that the frailty variance estimate increases with the addition of a covariate, which in principle should account for part of the heterogeneity in the data.

A Cox proportional hazards no-frailty model including age and sex as covariates show a reduced effect of sex with $\beta = -0.82\,(0.48)$, not significant. Furthermore, the effect of sex is highly non-proportional ($p < 0.01$). Plots of the Schoenfeld residuals from this model and a model with the logarithm of the posterior gamma frailty expectations included as an offset are shown in Figure 29. The departure from proportionality is represented by the departure of the fitted line from a horizontal line. It can be seen that the gamma frailty model "fixes" this by taking the marginal time-dependent effect as evidence for the effect of unobserved heterogeneity.

An ad-hoc way of modeling time-dependent effects is by fitting an extended model where an interaction between sex and time is also included. The interaction is highly significant with $\beta = -0.016\,(0.002)$ while the main effect of sex is of an opposite sign $\beta = 0.88(0.47)$. This implies a decreasing effect of sex with $\beta(t) = 0.88 - 0.016\,t$. At the median catheter survival time, the effect of sex is already negative with $\beta(78) = -0.37$. Since the effect of the usual frailty distributions leads to an attenuation of the marginal hazard ratio but not to a change of signs in $\beta(t)$ (as can be seen for example, in Figure 21), it is likely that there is a time-dependent effect of sex acting at the individual level.

A shared frailty model using a positive stable distribution for the random effect does not show a significant frailty. It was seen in the previous section that this distribution is less susceptible to rejecting the null hypothesis of no frailty because of time-dependent covariate effects.

Therefore, two competing explanations are plausible. The first is that there is unobserved heterogeneity and a time-constant effect of sex that appears time-dependent (as it does with the marginal model implied by the gamma frailty). The second is that the apparent unobserved heterogeneity is an artifact induced by a time-dependent effect of sex. Deciding between these two on the basis of these results alone is a difficult matter. This is in line with the explanation that non-proportional hazard effects and unobserved heterogeneity are confounded when the cluster size is small, as was shown in Section

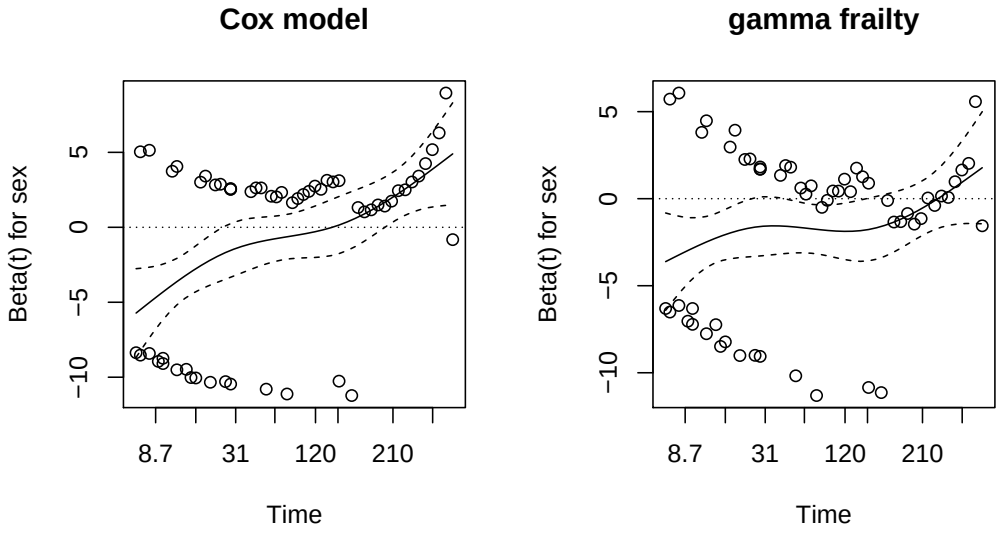

Figure 29: Plot of the Schoenfeld residuals for sex from a Cox marginal model and a gamma frailty model estimated on the kidney catheter insertions data.

2.3. Finally, we note that if the third variable (disease type) is included in the model, the evidence in favor of the frailty vanishes.

2.5 Conclusion

In univariate survival data, it is well known that a proportional hazards frailty model and a non-proportional hazards model (with a certain type of departure from proportionality) can not be distinguished on the basis of the data alone. We have studied how this problem extends to correlated survival data, such as clustered failures or recurrent events. The novelty of this chapter is that the confounding effect between marginal covariate effects and cluster effects was studied for different cluster sizes, and reasonable rates of false rejections are obtained only when the cluster size is large (e.g. 10 or more observations). Furthermore, the shape of the baseline hazard was shown to have a strong effect, with hazards that are large early on in the follow-up more likely to be influenced by the time-dependent effect of the covariates.

Although the simulation study in Section 2.3 aimed to cover a large number of scenarios, only a particular type of covariate effect was considered. In practice, this effect may be very different according to the true mechanism that generates the data. Nevertheless, this consideration should play an essential role in deciding whether the frailty

model is plausible or not. We found that the conclusions presented in Section 2.3 extend to a large number of scenarios, including a similar simulation study carried out with a Gompertz baseline hazard. However, a scenario worth further investigation is that when the frailty is present and a covariate has an increasingly protective effect. This would translate, in the terms of equation (2.7), as having $\beta_1 > 0$ and $\text{Var}[Z] > 0$. This may be seen as the time-dependent covariate effect offsetting the shrinking of the hazard ratio seen in Figure 21.

The frailty models attempt to recover an individual covariate effect. This may not be possible when the proportional hazards assumption does not hold conditional on the frailty, particularly when the cluster size is small.

All fitted models aim to accommodate the observable quantities according to different assumptions. The marginal hazards and marginal hazard ratios are somewhat more interpretable, as they "stick to this world" (Andersen and Keiding, 2012). Identifying the nature of what leads to the observable effects involves an additional number of assumptions that should be carefully considered in the problem being analyzed.

Supplementary material

The supplementary material referenced in this chapter is available online, at `https://github.com/tbalan/small_clusters`.

Appendix

Denote γ as the scale parameter and α as the shape parameter.

The Gamma(α, γ) distribution is described by the Laplace transform

$$\mathcal{L}_Z(c) = \left(\frac{\gamma}{\gamma + c}\right)^{\alpha}.$$

This is scaled by setting $\text{E}Z = 1$ and variance θ^{-1} by $\gamma = \alpha = \theta$.

The inverse Gaussian distribution IG(α, γ) is described by the Laplace transform

$$\mathcal{L}_Z(c) = \exp\left[-\alpha\left\{\left(\frac{\gamma + c}{\gamma}\right)^{1/2} - 1\right\}\right].$$

This is scaled by setting $\text{E}Z = 1$ and variance θ^{-1} by $\gamma = \theta/2$ and $\alpha = \theta$.

The positive stable distribution PS(α, γ) with $\gamma \in [0, 1]$is described by the Laplace transform

$$\mathcal{L}_Z(c) = \exp\left(-\alpha c^{\gamma}\right).$$

This is scaled with $\gamma = \frac{\theta}{\theta+1}$ and $\alpha = 1$. The expectation is infinite and the variance is not defined. Nevertheless, with $\theta = \infty$ ($\gamma = 1$) the case of no association is obtained and

the distribution only has mass at 1, while smaller values of θ indicate higher degrees of association.

For all the distributions above, the LRT tests the null hypothesis of $H_0 : \theta = \infty$, equivalent to no variability in the frailty distribution.

The lognormal distribution $LN(\mu, \sigma^2)$ is usually parametrized on the log scale, i.e. $\mathrm{E} \log Z = \mu$ and $\mathrm{Var} \log Z = \sigma^2$. In Section 2.3, the frailty was simulated by setting $\mathrm{E}Z = 1$ and $\mathrm{Var}Z = \theta^{-1}$, which is $LN(-1/2 \log(\theta + 1), \log(\theta + 1))$. The Laplace transform is not available in closed form. However, for Z a $LN(\mu, \sigma^2)$ a common approximation is

$$\mathcal{L}_Z(c) = (1 + W(e^{\mu}\sigma^2 c))^{-1/2} \exp\left(-\frac{W^2(e^{\mu}\sigma^2 c) + 2W(e^{\mu}\sigma^2 c)}{2\sigma^2}\right),$$

where $W(x)$ is the Lambert W function (Asmussen, Jensen, and Rojas-Nandayapa, 2016).

Table 21: Percentage of rejection of the null hypothesis for the Commenges-Andersen, ZPH and likelihood ratio tests for gamma (GA), inverse Gaussian (IG) and positive stable (PS) frailty models, for different cluster sizes (n). σ_1^2 is the variance of the lognormal frailty used in the simulation and β_1 represents the strength of the time-dependent part of the covariate effect as in equation (2.7). The results are shown for a total sample size of 300 and Weibull shape parameter $\alpha = 1$ and the clustered failures scenario.

	Test	$n = 2$	$n = 3$	$n = 5$	$n = 10$
$\sigma^2 = 0$					
$\beta_1 = 0$	CA	0.020	0.046	0.048	0.044
	ZPH	0.034	0.032	0.036	0.036
	LRT (GA)	0.026	0.030	0.032	0.022
	LRT (IG)	0.026	0.028	0.032	0.022
	LRT (PS)	0.024	0.026	0.022	0.016
$\beta_1 = -0.21$	CA	0.050	0.054	0.052	0.056
	ZPH	0.327	0.329	0.315	0.293
	LRT (GA)	0.078	0.066	0.042	0.044
	LRT (IG)	0.080	0.070	0.044	0.048
	LRT (PS)	0.024	0.032	0.026	0.024
$\beta_1 = -0.41$	CA	0.078	0.066	0.062	0.060
	ZPH	0.952	0.954	0.948	0.942
	LRT (GA)	0.120	0.090	0.062	0.052
	LRT (IG)	0.110	0.092	0.062	0.056
	LRT (PS)	0.026	0.028	0.028	0.030
$\sigma^2 = 0.25$					
$\beta_1 = 0$	CA	0.415	0.565	0.770	0.910
	ZPH	0.110	0.092	0.082	0.100
	LRT (GA)	0.503	0.663	0.834	0.928
	LRT (IG)	0.511	0.679	0.842	0.932
	LRT (PS)	0.251	0.375	0.593	0.838
$\beta_1 = -0.21$	CA	0.591	0.693	0.836	0.938
	ZPH	0.513	0.519	0.489	0.527
	LRT (GA)	0.667	0.776	0.880	0.952
	LRT (IG)	0.665	0.776	0.890	0.948
	LRT (PS)	0.273	0.429	0.669	0.874
$\beta_1 = -0.41$	CA	0.591	0.703	0.862	0.934
	ZPH	0.984	0.976	0.980	0.978
	LRT (GA)	0.667	0.776	0.888	0.940
	LRT (IG)	0.669	0.782	0.888	0.944
	LRT (PS)	0.255	0.451	0.683	0.876

Table 22: Percentage of rejection of the null hypothesis for the Commenges-Andersen, ZPH and likelihood ratio tests for gamma (GA), inverse Gaussian (IG) and positive stable (PS) frailty models, for different cluster sizes (n). σ_1^2 is the variance of the lognormal frailty used in the simulation and β_1 represents the strength of the time-dependent part of the covariate effect as in equation (2.7). The results are shown for a total sample size of 300 and Weibull shape parameter $\alpha = 1$ and the clustered failures covariate specific covariate scenario.

	Test	$n = 2$	$n = 3$	$n = 5$	$n = 10$
$\sigma^2 = 0$					
$\beta_1 = 0$	CA	0.062	0.064	0.042	0.060
	ZPH	0.036	0.052	0.032	0.044
	LRT (GA)	0.034	0.032	0.026	0.026
	LRT (IG)	0.032	0.032	0.028	0.026
	LRT (PS)	0.012	0.014	0.024	0.018
$\beta_1 = -0.21$	CA	0.074	0.044	0.062	0.054
	ZPH	0.382	0.328	0.358	0.294
	LRT (GA)	0.084	0.048	0.052	0.038
	LRT (IG)	0.090	0.044	0.052	0.038
	LRT (PS)	0.016	0.018	0.028	0.038
$\beta_1 = -0.41$	CA	0.100	0.064	0.068	0.050
	ZPH	0.960	0.964	0.952	0.942
	LRT (GA)	0.122	0.076	0.062	0.044
	LRT (IG)	0.118	0.070	0.066	0.046
	LRT (PS)	0.046	0.032	0.042	0.048
$\sigma^2 = 0.25$					
$\beta_1 = 0$	CA	0.404	0.526	0.772	0.876
	ZPH	0.102	0.124	0.130	0.198
	LRT (GA)	0.480	0.596	0.822	0.894
	LRT (IG)	0.492	0.604	0.832	0.902
	LRT (PS)	0.220	0.324	0.580	0.800
$\beta_1 = -0.21$	CA	0.570	0.644	0.868	0.894
	ZPH	0.576	0.576	0.622	0.668
	LRT (GA)	0.640	0.718	0.886	0.912
	LRT (IG)	0.642	0.716	0.890	0.920
	LRT (PS)	0.286	0.396	0.674	0.818
$\beta_1 = -0.41$	CA	0.570	0.664	0.848	0.906
	ZPH	0.998	0.986	0.990	0.990
	LRT (GA)	0.638	0.724	0.884	0.920
	LRT (IG)	0.640	0.724	0.890	0.924
	LRT (PS)	0.370	0.488	0.712	0.832

Table 23: Percentage of rejection of the null hypothesis for the Commenges-Andersen, ZPH and likelihood ratio tests for gamma (GA), inverse Gaussian (IG) and positive stable (PS) frailty models, for different cluster sizes (n). σ_1^2 is the variance of the lognormal frailty used in the simulation and β_1 represents the strength of the time-dependent part of the covariate effect as in equation (2.7). The results are shown for a total sample size of 300 and Weibull shape parameter $\alpha = 1$ and the recurrent events scenario.

	Test	$n = 2$	$n = 3$	$n = 5$	$n = 10$
$\sigma^2 = 0$					
$\beta_1 = 0$	CA	0.060	0.034	0.036	0.038
	ZPH	0.050	0.038	0.030	0.030
	LRT (GA)	0.040	0.022	0.016	0.018
	LRT (IG)	0.038	0.026	0.022	0.020
	LRT (PS)	0.020	0.016	0.026	0.014
$\beta_1 = -0.21$	CA	0.122	0.068	0.064	0.074
	ZPH	0.301	0.293	0.285	0.173
	LRT (GA)	0.155	0.082	0.066	0.070
	LRT (IG)	0.145	0.074	0.066	0.064
	LRT (PS)	0.026	0.022	0.032	0.028
$\beta_1 = -0.41$	CA	0.263	0.153	0.127	0.094
	ZPH	0.956	0.920	0.924	0.857
	LRT (GA)	0.313	0.197	0.151	0.096
	LRT (IG)	0.283	0.201	0.159	0.106
	LRT (PS)	0.054	0.058	0.062	0.068
$\sigma^2 = 0.25$					
$\beta_1 = 0$	CA	0.309	0.460	0.691	0.837
	ZPH	0.118	0.120	0.203	0.209
	LRT (GA)	0.341	0.506	0.737	0.859
	LRT (IG)	0.359	0.512	0.737	0.867
	LRT (PS)	0.145	0.231	0.472	0.739
$\beta_1 = -0.21$	CA	0.530	0.629	0.835	0.916
	ZPH	0.600	0.590	0.663	0.665
	LRT (GA)	0.590	0.669	0.855	0.918
	LRT (IG)	0.588	0.677	0.867	0.924
	LRT (PS)	0.209	0.323	0.580	0.827
$\beta_1 = -0.41$	CA	0.657	0.719	0.880	0.938
	ZPH	0.996	0.984	0.980	0.988
	LRT (GA)	0.715	0.767	0.906	0.944
	LRT (IG)	0.727	0.779	0.906	0.944
	LRT (PS)	0.295	0.452	0.711	0.880

Chapter 3

Score test for association between recurrent events and a terminal event

Abstract

The statistical analysis of recurrent events relies on the assumption of independent censoring. When random effects are used, this means, in addition, that the censoring cannot depend on the random effect. Whenever the recurrent event process is terminated by death, this assumption might not be satisfied. Joint models for recurrent and terminal events are often difficult to fit. Thus, clinicians rarely check whether they are preferred to separate models. In this chapter, we propose and compare simple, yet efficient ways of testing whether the terminal event and the recurrent events are associated or not. The proposed methods are evaluated in a simulation study and are illustrated through a data set consisting of repeated observations of skin tumors on T-cell lymphoma patients.

3.1 Introduction

Recurrent event data have become increasingly common in clinical studies, in reliability theory, and in other fields (Cook and Lawless, 2007). The shared frailty model (Nielsen et al., 1992) is a popular method for analyzing this type of data, because it retains a

This chapter has been published as: T.A. Balan, S.E. Boonk, M.H. Vermeer, H. Putter (2016). Score test for association between recurrent events and a terminal event. *Statistics in Medicine* 35(18), 3037-3048.

similar semiparametric specification with the well known Cox model, it is supported by asymptotic results (Murphy, 1995a; Parner, 1998) and is available in standard statistical software (Therneau and Grambsch, 2000). The *frailty* (Vaupel, Manton, and Stallard, 1979) is a random effect which accounts for heterogeneity that can not be explained by observable covariates. In other words, it describes whether a subject or a cluster of subjects is at a higher risk (large frailty) than others (small frailty). In the recurrent events framework, the frailty accounts for the dependence between the observations on the same individual. Conditional on the frailty, one hopes that the stochastic processes underlying the individuals are independent. Thus, frailty models allow an elegant and parsimonious explanation of the mechanism which generates the data.

In a clinical context, recurrent events are often a symptom of a medical condition which might lead to the end of follow-up in the form of dependent censoring by terminal event, such as death. In particular, a more frail subject might not only be associated with a higher recurrence rate, but also an increased or decreased risk of experiencing the terminal event, to a greater or lesser extent. If this is the case, the recurrences and the terminal event should be jointly modeled, allowing for the frailty to describe both the unaccounted differences in the risk for both recurrences and death. Such a model was introduced in Liu, Wolfe, and Huang (2004), who adapted a model for clustered failures with informative censoring (Huang and Wolfe, 2002). For estimation of a semiparametric joint frailty model, the Expectation-Maximization (EM) algorithm can be used, the method being very similar to the estimation of the shared frailty model (Nielsen et al., 1992; Klein, 1992).

There are however disadvantages of the joint model. It is notably easier to consider separate models for the recurrences and death, both in terms of difficulty of fitting and interpretation; a comparison between the estimation methods of the shared frailty model (Nielsen et al., 1992) and the joint model (Liu, Wolfe, and Huang, 2004) can attest to this. Furthermore, expressions for marginal features of the recurrent events or terminal event processes are not readily obtained, and the interpretation of features of interest, such as treatment effects, is not as straightforward as for the separate models. Although software for parametric models for recurrent and terminal events exists (Rondeau and Gonzalez, 2005), there is no method to check a priori whether separate models are similarly appropriate or not. This may lead to situations when clinical practitioners will ignore the dependence between the two event types.

In this chapter, we aim to develop a simple statistical test for association between the recurrent events and the terminal events, which does not require the estimation of a joint model. This provides an answer to a clinically relevant problem and it also indicates whether the joint modeling of the processes is more suitable. The idea that we follow is similar to a test for informative censoring (Huang, Wolfe, and Hu, 2004) and heterogeneity (Commenges and Andersen, 1995) in the context of clustered failures.

The outline of the article is as follows. In Section 3.2, we review a joint model closely related to that of Liu, Wolfe, and Huang (2004). In Section 3.3, we review possible tests for association and introduce the robust score test, and in Section 3.4 we discuss the

efficiency and validity of our approach in a simulation study. Finally, in Section 3.5 we illustrate the proposed methods on a data set of successive hospital readmissions.

3.2 Models

Let D_i and C_i denote the time of the terminal event and right censoring time respectively, both of which correspond to the end of followup. Also define $T_i = \min(D_i, C_i)$, and $Y_i(t) = 1(t \leq T_i)$ the "at risk" indicator. While $Y_i(t) = 1$, we observe two counting processes, $N_i^D(t) = 1(D_i \leq t)$ corresponding to the terminal event and $N_i^R(t)$ which is equal to the number of recurrences in $(0, t]$, or equivalently their increments $\Delta N_i^R(t)$ and $\Delta N_i^D(t)$, equal to the number of respective events in the small interval $(t, t + \Delta t]$. We can consider a $p \times 1$ vector of possibly time-dependent covariates $\{x_i(t) : t \geq 0\}$ and denote their path up to time t as $x_i^{(t)} = \{x_i(s) : 0 \leq s \leq t\}$. We require the time-dependent covariates to be external, in the sense of Kalbfleisch and Prentice (2002). The history up to time t is then

$$H_i(t) = \left\{ \left(N_i^R(s),\, N_i^D(s)\right) : 0 \leq s \leq t;\, x_i^{(t)} \right\}. \tag{3.1}$$

The intensities of N_i^R and N_i^D can be associated, meaning that the rate of recurrences and that of the terminal event can depend on elements of (3.1). It is, for example, plausible that a high rate of recurrent events is associated with a reduced survival. Often, this can be an indication of a "hidden" factor, such as a severe disease, which influences both intensities of N_i^R and N_i^D.

As in the model of Liu, Wolfe, and Huang (2004), we consider a frailty variable $\mathbf{Z} = (Z_1, ..., Z_n)$ with Z_i's i.i.d. with a distribution function $G(z; \theta)$, with mean 1 and variance θ. Conditional on $\mathbf{Z} = (z_1, ..., z_n)$, the intensities of N_i^R and N_i^D are:

$$\begin{aligned} r_i(t|z_i) &= \lim_{\Delta t \to 0} \frac{Pr\left\{\Delta N_i^R(t) = 1 | z_i,\, H_i(t-)\right\}}{\Delta t}, \\ \lambda_i(t|z_i) &= \lim_{\Delta t \to 0} \frac{Pr\left\{\Delta N_i^D(t) = 1 | z_i,\, H_i(t-)\right\}}{\Delta t}. \end{aligned}$$

Further, we assume that both N_i^R and N_i^D can not increase after D_i. Although a natural assumption for the terminal event, for the recurrent events death is an instance of potentially informative censoring. In particular, a violation of the assumptions of the classical shared frailty model (Nielsen et al., 1992) occurs if z_i can not be dropped from the expression of λ_i. Finally, we follow Liu, Wolfe, and Huang (2004) in choosing a multiplicative model for the intensities, so that r_i and λ_i can be expressed as

$$\begin{cases} r_i(t|z_i) &= z_i 1(D_i > t)\, e^{\beta' x_i^R(t)} r_0(t) \\ \lambda_i(t|z_i) &= z_i^\gamma 1(D_i > t)\, e^{\alpha' x_i^D(t)} \lambda_0(t) \end{cases}. \tag{3.2}$$

The baseline intensities r_0 and λ_0 are assumed for now to be continuous positive functions. The regression coefficients α and β have the dimensions of the corresponding covariates x_i^D and x_i^R.

The question of association between N_i^R and N_i^D is closely related to the parameter γ in (3.2), which describes the direction and magnitude at which the frailty influences the hazard λ_i. Thus, the interest lies in testing the hypothesis $H_0 : \gamma = 0$ against $H_A : \gamma \neq 0$. Under H_0, the expressions of λ_i and r_i do not share any parameters, and then both processes can be analyzed separately; in particular, the censoring of the recurrent event process by the terminal event is non-informative, in the sense of Nielsen et al. (1992).

Assume that the baseline intensities are fully described by some parameters ϕ_r and ϕ_d, i.e. $r_0(t) \equiv r_0(t;\, \phi_r)$ and $\lambda_0(t) \equiv \lambda_0(t;\, \phi_d)$. If ϕ_r and ϕ_d are finite dimensional, then the model is parametric; otherwise, the model is semi-parametric, as originally proposed by Liu, Wolfe, and Huang (2004). Nevertheless, we denote the nuisance parameter vector by $\eta = \{\beta,\, \alpha,\, \theta,\, \phi_r,\, \phi_d)\}$.

For subject i, we denote the observed data O_i as the event "n_i observed recurrent events at t_{i1}, ..., t_{in_i} over $[0, t_i]$ and $\delta_i = 1(D_i < C_i)$". Under the regularity conditions of Liu, Wolfe, and Huang (2004), the "conditional likelihood" based on $(H_i(\infty);\ i = 1...n;\ \mathbf{Z})$ is formed from the conditional probabilities

$$Pr(O_i|z_i) = \prod_j \left\{ r_i(t_{ij}|z_i) \right\} \exp\left\{ -\int_0^\tau Y_i(s) r_i(s|z_i) ds \right\} \lambda_i(t_i|z_i)^{\delta_i} \times \\ \times \exp\left\{ -\int_0^\tau Y_i(s) \lambda_i(s|z_i) ds \right\}.$$

Similarly, the "marginal likelihood" based on $H_i(\infty)$ alone is obtained from the marginal contributions to the likelihood $Pr(O_i) = \int_0^\infty Pr(O_i|z) dG(z;\, \theta)$. The marginal log-likelihood is then

$$\ell\,(\gamma, \eta) = \sum_i \left[\sum_{j=1}^{n_i} \left\{ \beta' x_i(t_{ij}) + \log r_0(t_{ij}) \right\} + \delta_i \left\{ \alpha' x_i(t_i) + \log \lambda_0(t_i) \right\} + \right. \\ \left. + \log \int_0^\infty K_i(z, t_i) f_\theta(z) dz \right] \quad (3.3)$$

where $K_i(z,t) f_\theta(z)$, is the kernel of the "posterior" distribution $Z_i|H_i(t)$ computed with the data available until time t. We denote the cumulative given $z_i = 1$ as $R_i(t) = \int_0^t Y_i(s) e^{\beta' x_i^R(s)} r_0(s) ds$ and $\Lambda_i(t) = \int Y_i(s) e^{\alpha' x_i^D(s)} \lambda_0(s) ds$, and then

$$K_i(z,t) = z^{N_i^R(t-) + \gamma N_i^D(t-)} \exp\left\{ -z R_i(t) - z^\gamma \Lambda_i(t) \right\}. \quad (3.4)$$

Under $\gamma = 0$, K_i is the kernel of a Gamma distribution, so a convenient choice for G is the Gamma distribution as well (Nielsen et al., 1992); also see Duchateau and Janssen (2007).

If $\gamma \neq 0$ the Expectation-Maximization algorithm must be employed to maximize the log-likelihood, using numerical methods to approximate integrals at every iteration (Liu, Wolfe, and Huang, 2004). The numerical approximations and the slow convergence of the EM algorithm result in an overall slow and complicated method.

One way out is to consider a parametric version of the joint model. At the expense of introducing assumptions about the functional form of r_0 and λ_0, one can obtain a numerically tractable form of the log-likelihood (3.3), which can be maximized with standard maximum likelihood methods (Rondeau, Mathoulin-Pelissier, et al., 2007). This approach is implemented in the R package **frailtypack** (Rondeau and Gonzalez, 2005; Rondeau, Mazroui, and Gonzalez, 2012), which also offers the option to choose flexible parametric specifications for r_0 and λ_0, such as piecewise constant or spline-approximated.

There are however reason not to employ the joint model. First, clinicians prefer more familiar models such as a frailty model for the recurrent events (available in e.g. R, SAS, Stata) or a Cox model for the terminal event (also available in SPSS), if there is no need of doing something more complicated. The parametric assumptions have their price as well. Splines, for example, require the specification of two "smoothing parameters", which may or may not be easy to obtain. We will return to considerations about computation in section 3.4. Thus, it would be useful to be able to see if there is evidence against H_0 even before the joint model is used. While the Likelihood Ratio Test (LRT) or the Wald test require the maximization of (3.3), the score test does not. If the null hypothesis is rejected, the shared frailty model is not appropriate and the terminal event should be jointly modeled (Liu, Wolfe, and Huang, 2004; Ye, Kalbfleisch, and Schaubel, 2007).

In the following section, we describe tests for H_0 based on (3.3), with a focus on those that do not require the maximization of (3.3).

3.3 Tests for independence

Our goal is to test $H_0 : \gamma = 0$, in the presence of the nuisance parameters η; a complete specification of the null hypothesis is $H_0 : (\gamma, \eta) = (0, \eta)$. Abiding by our purpose of developing a simple test for this hypothesis, we first focus on how this can be achieved while avoiding the direct maximization of (3.3). This can be done by considering the maximum likelihood estimate $\hat{\eta}_0$ under $\gamma = 0$ and measuring the variation of (3.3) around $\gamma = 0$. This forms the basis of the score test in section 3.3.1. Other approaches, for which estimation of the joint model is needed, are detailed in Section 3.3.2.

3.3.1 Score Test

The starting point for this is the score function for γ under H_0, defined as the derivative with respect to γ in (3.3):

$$U_\gamma(0, \eta) = \frac{\partial}{\partial \gamma} \ell(\gamma, \eta)\Big|_{\gamma=0}.$$

If we denote $\hat{\eta}_0$ the estimate of η under H_0, then

$$\frac{\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\}^2}{\text{Var}\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\}} \tag{3.5}$$

follows asymptotically a χ^2 distribution with 1 degree of freedom. The variance of the score is

$$\text{Var}\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\} = \left(I_{\gamma\gamma} - I_{\gamma\eta} I_{\eta\eta}^{-1} I_{\eta\gamma}\right)\Big|_{\gamma=0, \eta=\hat{\eta}_0}, \tag{3.6}$$

where the Is are obtained from the Fisher information matrix

$$I(\gamma, \eta) = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\eta} \\ I_{\eta\gamma} & I_{\eta\eta} \end{pmatrix}.$$

If the model is semi-parametric, the score function and information matrix of η are replaced by a score and an information operator (Rabinowitz, 2000; Kosorok, 2008). Although this does not lead to a closed form of (3.6), any "good" estimate of the variance of the score can be used. The first choice is to replace the denominator of (3.5) with $I_{\gamma\gamma}\big|_{\gamma=0}$, which is the variance of the score if η were *known* to be equal to $\hat{\eta}_0$. By this, the variance will be underestimated, thus leading to a conservative test statistic. We refer to this as the **naive score test** (NST).

Further insight can be obtained by calculating U_γ:

$$U_\gamma(\gamma, \eta) = \sum_i \frac{\int N_i^D(t_i) \log z - \Lambda_i(t_i|z) z^\gamma \log z K_i(z) f_\theta(z) dz}{\int K_i(z) f_\theta(z) dz}.$$

Setting $\gamma = 0$ and replacing η with $\hat{\eta}_0$, we obtain

$$\begin{aligned} U_\gamma(0, \hat{\eta}_0) &= \sum_i \frac{\int \left\{ N_i^D(t_i) - \widehat{\Lambda}_i(t_i|x_i,\ z)\right\} \log z \widehat{K_i}(z) f_\theta(z) dz}{\int K_i(z) f_\theta(z) dz} \\ &= \sum_i \widehat{M_i^D} \cdot \widehat{\log z_i}, \end{aligned} \tag{3.7}$$

where $\widehat{M_i^D}$ and $\widehat{\log z_i}$ are the estimates of $M_i^D = N_i^D(t_i) - \int_0^{t_i} Y_i(s)\lambda_i(s)ds$, the martingale residual of the terminal event, and of $E\left[\log Z_i | O_i(t_i)\right]$, where the expectation is taken with respect to the "posterior" distribution $\widehat{K_i}(z) f_\theta(z)$ of (3.4), with R_i and Λ_i replaced by their estimates under H_0.

A similar expression involving a correlation between martingale residuals and aspects of the posterior distribution of random effects was obtained in Jacqmin-Gadda et al. (2010) in the context of joint latent classes and survival models.

Both estimates in (3.7) are only asymptotically independent samples; in practice, there is a dependency between the estimates (Therneau and Grambsch, 2000). In particular, the martingale residuals $\widehat{M_i^D}$ are constrained to have mean 0 , therefore (3.7)

is proportional to the sample covariance of the martingale residuals and expected log-frailties, which is a measure of linear dependence. In fact, if an ordinary linear regression model is considered:

$$\widehat{M_i^D} = a + b\,\widehat{\log z_i} + \varepsilon_i,$$

then the departure of (3.7) from 0 is equivalent to the departure of the regression coefficient b from 0. Thus, for testing H_0, the regular t statistic can be used:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.8}$$

where $r = \text{Corr}(\widehat{M_i^D}, \widehat{\log z_i})$ and t follows asymptotically a t distribution with $n-2$ degrees of freedom under the null hypothesis under H_0. We refer to the test based on (3.8) as the **robust score test** (RST).

Heuristically, a justification for the RST can be derived by interpreting the quantities which appear in (3.7). The martingale residuals $\widehat{M_i^D}$ can be informally interpreted as an "observed - expected" quantity for the terminal event. For example, if $\widehat{M_i^D} > 0$, then the rate of the terminal event is *larger than expected, taking only the x_i into account*, and how much *larger* is determined by how large $\widehat{M_i^D}$ is. A large (log-)frailty estimate corresponds to a subject who is at high risk for recurrences. Hence, the larger the value of (3.7), the stronger the evidence for the association between recurrent and terminal events is. More frail subjects are more likely to experience the terminal event earlier if $r > 0$, or later if $r < 0$, so the sign of the RST statistic also indicated the direction of the association.

3.3.2 Alternative tests

The **likelihood ratio test** (LRT) can be computed by maximizing the likelihood (3.3) via the expectation-maximization algorithm, as described in Liu, Wolfe, and Huang (2004), and comparing it to the likelihood under H_0. If (3.3) is maximized in $(\widehat{\gamma}, \widehat{\eta})$, then the LRT statistic is

$$D = -2\log\left\{\frac{l(0, \hat{\eta}_0)}{l(\widehat{\gamma}, \widehat{\eta})}\right\}$$

and it asymptotically follows a χ^2 distribution with one degree of freedom under H_0.

The **efficient score test** (EST) is described by (3.5) and the efficient information (3.6), and as previously mentioned it can be computed numerically. As shown in Murphy and Vaart (2000), the efficient information can be obtained as minus the second derivative of the profile likelihood

$$\ell_{\text{prof}}(\gamma) = \sup_{\eta} \ell(\gamma, \eta). \tag{3.9}$$

In practice, we can approximate $\tilde{I}_\gamma\Big|_{\gamma=0} = -E\left(\frac{d^2}{d\gamma}\ell_{\text{prof}}(\gamma)\Big|_{\gamma=0}\right)$ with the numeric Hessian of (3.9) in $\gamma = 0$. This can be obtained from general purpose optimization software, such

Table 31: Average number of recurrent event in simulated data sets.

	θ		
γ	0.5	1	1.5
-0.5	2.36	2.44	2.52
-0.25	2.32	2.36	2.41
0	2.27	2.27	2.27
0.25	2.21	2.16	2.12
0.5	2.13	2.05	1.96

as the function `optim` in `R` or `S-Plus` or the package `numDeriv` in `R`. We comment in the Appendix on computational considerations regarding the EST and how it is related to the NST in this light.

Alternatively, the $\hat{\gamma}$ can be obtained from maximizing (3.9) with respect to γ. The variance of the estimate $\text{Var}(\hat{\gamma})$ can be obtained from the numeric Hessian, and then the **Wald test** statistic is

$$W = \frac{\hat{\gamma}}{\sqrt{\text{Var}(\hat{\gamma})}}$$

and it asymptotically follows a standard normal distribution under H_0. The sign of W also corresponds to the direction of the tested association.

3.4 Simulation

A simulation study has been conducted to assess the validity of the Robust Score Test (RST) and compare the small sample properties to those of the other tests described in Section 3.3. The simulations have been carried out in the following setting: data sets consist of $n \in 100, 200, 500$ subjects; for each subject the data is generated according to model (3.2), for scenarios pertaining to $\gamma \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The frailty is generated from a Gamma distribution with mean equal to 1 and variance $\theta \in \{0.5, 1, 1.5\}$. One binary covariate is generated from a Binom(n, 1/2) distribution with fixed regression coefficients $\beta = \alpha = 1$. An exponential baseline hazard is used, with $\lambda_0(t) = 1/2$ and $r_0(t) = 2$. The follow-up is ended by either the terminal event, or by an administrative censoring time $C_i = 1$, whichever occurs first. Note that the recurrent event rate and the terminal event rate are independent only under H_0. Every simulation cycle consist of 1000 replications under the same conditions.

In Table 31 we show an indication on the size of the simulated data sets. It can be seen that the number of recurrent events decreases with γ and with θ. The asymmetry is explained by the fact that when $\gamma < 0$ the recurrent events have a "protective" effect and subjects with many events exit the data set later. The degree to which there is more variance in the frailty amplifies this effect.

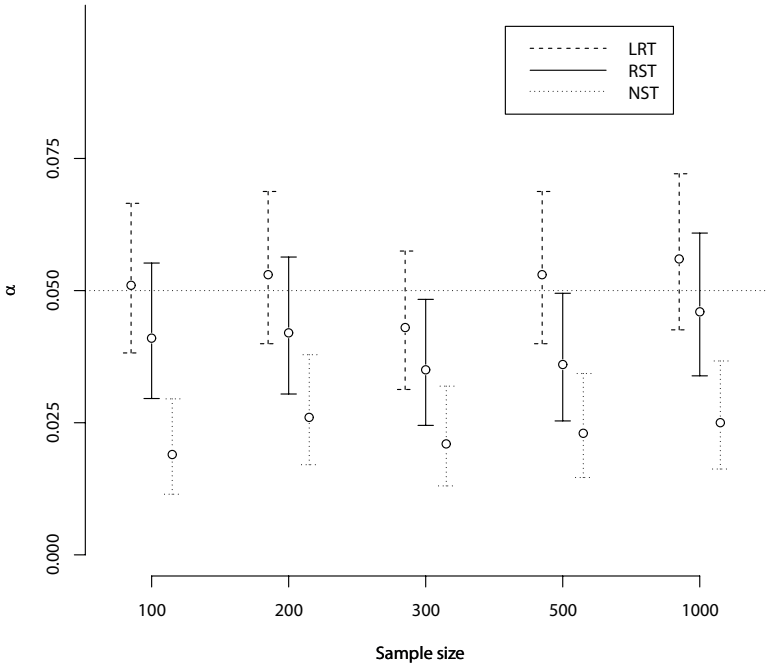

Figure 31: Estimated α levels with simulated data under H_0. Wald and EST (not shown here) are close to the LRT estimates. Binomial confidence intervals are also shown, where a "success" is a p-value smaller than 0.05

We use the abbreviations of the tests as described in Sections 3.3.1 and 3.3.2. Furthermore, we also consider the Wald test from a parametric model where the baseline intensities are considered piecewise constant with 3 intervals, from the R package `frailtypack`; this approach is described in Section 3.2, and we see it as an approximation to the semiparametric joint model.

Figure 1 compares the type 1 error (false rejections) of the LRT, RST and NST, as a function of n, under H_0, in the case $\theta = 1$. Although the estimated α level seems consistently lower for RST than for LRT, bionomial confidence intervals for the proportion of rejections have a notable overlap, and both seem to approach the desired 0.05 with a sufficiently large sample. In this comparison, it can also be seen that the naive score test (NST) is indeed over-conservative, as it is argued also in Appendix 3.6: even as the sample size becomes larger, the proportion of rejections is significantly lower then the nominal α level of 0.05. Finally, we note that the results for $\theta \in \{0.5, 1.5\}$ (not shown) are very similar.

To better illustrate the relation between the different tests, we plotted the p-values obtained in the case $\gamma = 0$, $\theta = 1$, $n = 500$ in Figure 32. Under the null hypothesis, one

would expect the p-values of a valid test to be approximately uniformly distributed on $[0, 1]$. The Wald test and EST are virtually indistinguishable from the LRT in this case. The figure indicates that RST approximates the LRT for small deviations as well. The parametric Wald (WaldPar) test is also shown in the plot; it can be seen that the p values can differ wildly from those of the semi-parametric Wald; this can be seen as a trade off for the parametric assumption. For other values of n or θ very similar figures were obtained.

Finally, we analyze the power of the aforementioned tests against the alternatives $\gamma \in \{-0.5, -0.25, 0.25, 0.5\}$, for $\theta \in \{0.5, 1, 1.5\}$. The results are summarized in Table 32. Two trends are visible regardless of sample size. First, the power of the tests grows with the frailty variance, meaning that it is more likely to reject the null hypothesis of no association in more heterogeneous data sets, if this association exists. Second, in particular for LRT, Wald and EST, the tests fare slightly better for alternatives with $\gamma < 0$, which can be explained by the asymmetric size of the simulated data sets showed in table 31.

As expected, the tests are more powerful when there is a higher number of individuals in the data set. The RST performs better than Wald for small sample sizes ($n = 100$), however there is no clear difference for others. Generally, the power of the RST is slightly lower but reasonably comparable with the other tests. In Figure 33 we compare the power of the tests for $\theta = 1$. It can be seen that, except for NST which is over-conservative, the LRT, Wald, EST and RST are quite similar. It looks like for small samples there is a slight advantage in power of LRT and EST, while the RST is closer to the Wald test.

Finally, we note that the computation time is much smaller for the RST, as compared to the other tests, including the parametric Wald test, WaldPar. Average computation times from the simulations are shown in Table 33.

3.5 Application

We illustrate our methods using data from a study on Mycosis Fungoides (MF). MF is the most common type of cutaneous T-cell lymphoma that generally presents with patches and plagues Doorn, Scheffer, and Willemze (2002). Over time a number of patients progress to tumor stage disease (stage IIB) and a minority develop extracutaneous localization of the disease. It is well known that there is considerable variability in the number of recurrent skin tumors and is believed that an increased number of recurrent skin tumors is associated with disease progression and survival. In addition, it has been reported that folliculotropism of neoplastic cells is associated with an adverse prognosis. In Boonk et al. (2014), 46 patients with stage IIB MF were selected from the cutaneous lymphoma database of the Dutch Cutaneous Lymphoma Group. During follow-up, data on recurrences of skin tumor and disease progression and survival were collected. We consider overall survival as the terminal event. Median follow-up was 88 months. Covariates considered in this application are age (median 69, range 39–90), gender (33

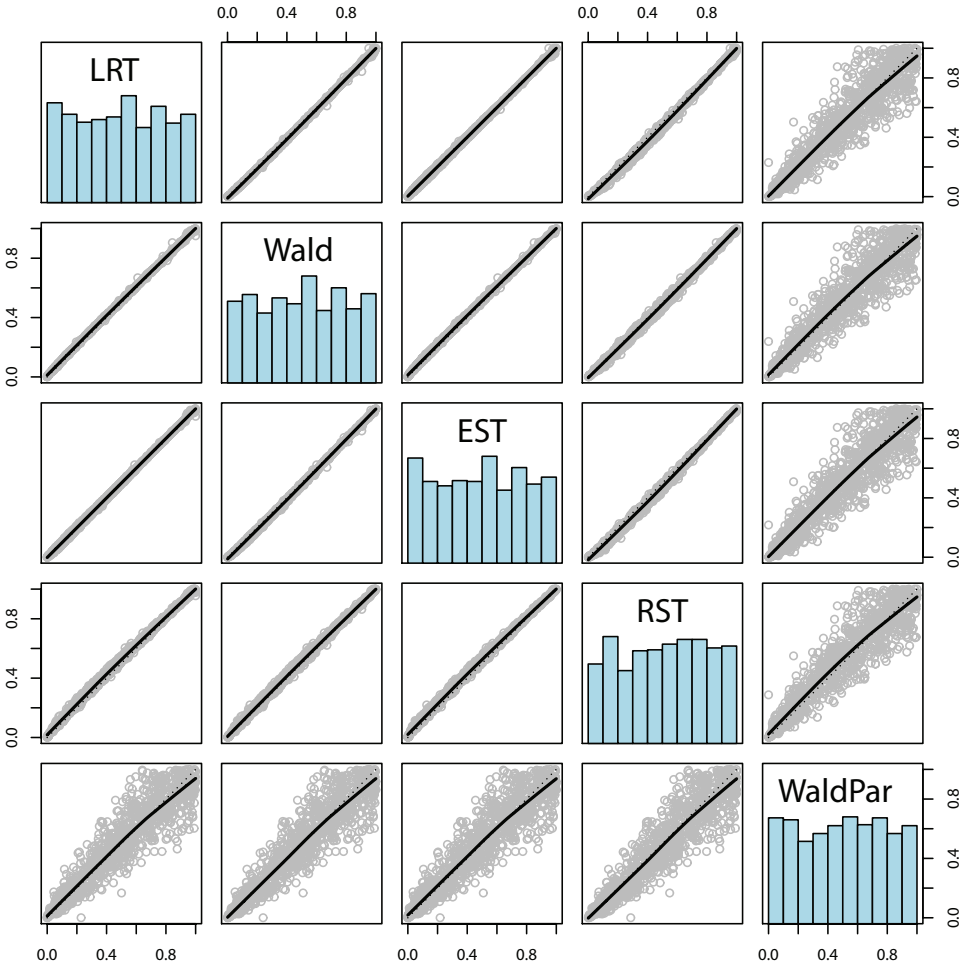

Figure 32: Histograms and scatterplots of p-values corresponding to 1000 datasets simulated under $H_0 : \gamma = 0$. Within the scatter plots, a straight line with equation $y = x$ has been added, as well as a dotted nonparametric smoother. The data sets follow the simulation scenarios of Section 3.5 with $n = 500$.

Table 32: Power against alternative hypotheses with varying sample size $n \in \{100, 200, 500\}$ and frailty variance $\theta \in \{0.5, 1, 1.5\}$

		LRT			Wald			EST			RST		
θ	γ	100	200	500	100	200	500	100	200	500	100	200	500
0.5													
	-0.5	0.37	0.60	0.94	0.25	0.56	0.94	0.42	0.63	0.95	0.32	0.54	0.92
	-0.25	0.13	0.19	0.49	0.08	0.19	0.47	0.18	0.28	0.52	0.11	0.18	0.42
	0.25	0.14	0.17	0.51	0.08	0.17	0.47	0.14	0.22	0.50	0.10	0.18	0.44
	0.5	0.33	0.54	0.96	0.21	0.54	0.93	0.33	0.6	0.94	0.27	0.54	0.92
1													
	-0.5	0.70	0.94	1.00	0.64	0.93	1.00	0.72	0.94	1.00	0.65	0.92	1.00
	-0.25	0.29	0.55	0.91	0.22	0.51	0.90	0.37	0.60	0.93	0.26	0.50	0.89
	0.25	0.30	0.49	0.89	0.20	0.42	0.87	0.28	0.46	0.88	0.25	0.44	0.88
	0.5	0.72	0.95	1.00	0.61	0.92	1.00	0.69	0.94	1.00	0.65	0.92	1.00
1.5													
	-0.5	0.85	0.99	1.00	0.82	0.99	1.00	0.75	0.92	0.99	0.82	0.98	1.00
	-0.25	0.47	0.80	0.99	0.39	0.76	0.98	0.56	0.83	0.99	0.44	0.77	0.98
	0.25	0.46	0.77	0.99	0.35	0.71	0.98	0.44	0.75	0.98	0.4	0.71	0.98
	0.5	0.88	0.99	1.00	0.82	0.99	1.00	0.87	0.99	1.00	0.86	0.98	1.00

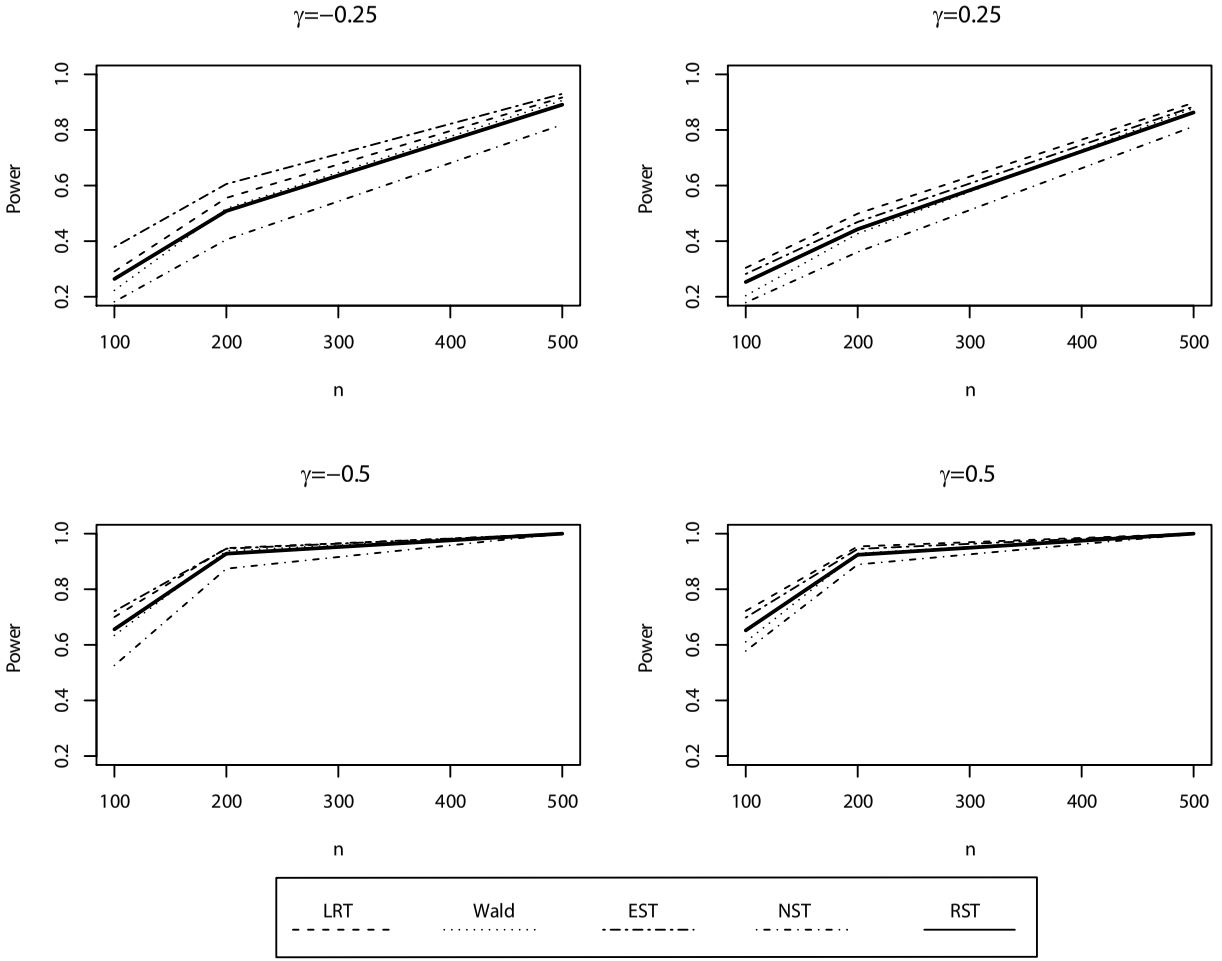

Figure 33: Power of LRT, Wald, EST, NST, and RST compared for $\theta = 1$

Table 33: Average computation time for different tests. For RST the standard `survival` package was used, for WaldPar the `frailtypack` package, and for EST and LRT or Wald a self-written algorithm was used, similar to that described in Liu, Wolfe, and Huang (2004).

	Computation time (s)		
	100	200	500
RST	0.04	0.09	0.31
EST	16.18	48.05	138.03
WaldPar	1.04	1.64	2.68
LRT/Wald	44.57	128.25	331.61

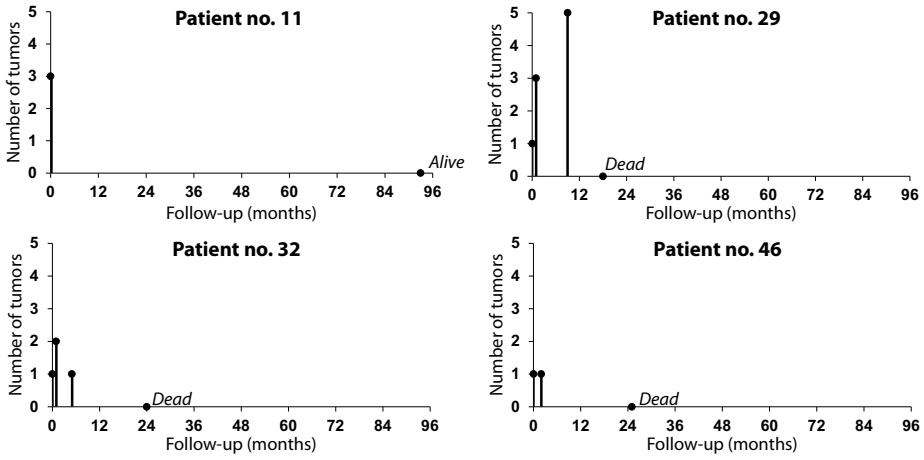

Figure 34: Recurrent event history and survival outcome of 4 patients

males, 13 females), and the presence of folliculotropic MF (26 absent, 20 present). Figure 34 shows examples of the variability in the number of tumors and time intervals between tumor recurrences. It can be seen that some patients experienced multiple recurrences at a single follow-up visit; the ties caused by these simultaneous recurrences were randomly broken. 11 patients (23.9%) experienced 0 recurrences, 5 (10.8%) 1 recurrence, 6 (13.0%) 2 recurrences, and 24 (52.1%) more than 2 recurrences. The maximum number of recurrences was 21. The original publication (Boonk et al., 2014) used the number of recurrent skin tumors in the first year as explanatory variable in a landmark Cox model at 1 year for overall survival, and showed that the number of recurrent skin tumors was highly prognostic for subsequent survival.

A gamma frailty model ignoring possible informative censoring due to the terminal event death, yielded the results shown in Table 34, under "Separate models". The frailty variance was estimated to be 1.574. The estimates of the Cox model for the terminal

Table 34: Estimated regression coefficients for recurrent events and terminal event, using separate models and the joint model.

	Separate models			Joint model		
	Beta	SE	p-value	Beta	SE	p-value
Recurrent events						
Male gender	0.230	0.687	0.74	0.286	0.476	0.54
Age	0.039	0.020	0.058	0.039	0.018	0.035
Folliculotropic MF	0.019	0.595	0.97	0.039	0.276	0.88
Frailty variance	1.574		< 0.0001	1.358	0.323	< 0.0001
Association parameter (γ)				0.778	0.276	0.004
Terminal event						
Male gender	0.616	0.486	0.20	0.747	0.648	0.24
Age	0.048	0.019	0.012	0.067	0.023	0.004
Folliculotropic MF	0.378	0.402	0.35	0.127	0.486	0.79

event, ignoring the recurrent events is also shown under "Separate models". Figure 35 shows a scatterplot of the posterior log frailties from the gamma frailty models against the martingale residuals of the Cox model for the terminal event. The correlation was estimated to be 0.488, and the p-value of the robust score test was 0.0006. The result of this quick test indicates that a joint model is really needed to reliably model the association between the recurrent skin tumors and death. The result of this joint model, using a self-written EM-algorithm, is shown in Table 34, under the "Joint model". The regression coefficients in the joint model are generally comparable with the ones from the separate models. The association parameter γ was estimated to be positive and highly significant, indicating an increased death rate for the subjects with a high propensity of recurrent events, in agreement with the findings in Boonk et al. (2014).

3.6 Discussion

We have shown that the estimated correlation between the martingale residual and the estimated log-frailties can be used as the basis for a test of association between recurrent events and a terminal event. The advantage of the robust score test is that it is easy to compute and does not require fitting the joint model. Thus, it can serve as a simple preliminary check whether models for the recurrent events and for the terminal events can be fitted separately or whether more complex joint models are needed to obtain reliable estimates.

We note that heterogeneity with respect to the recurrent events is required not only for the joint model to be estimated, but also for the implementation of the RST. This can be assessed via a likelihood ratio test (Nielsen et al., 1992; Therneau and Grambsch, 2000). In addition, we note that the model described in Section 3.2 leads to the interpretation

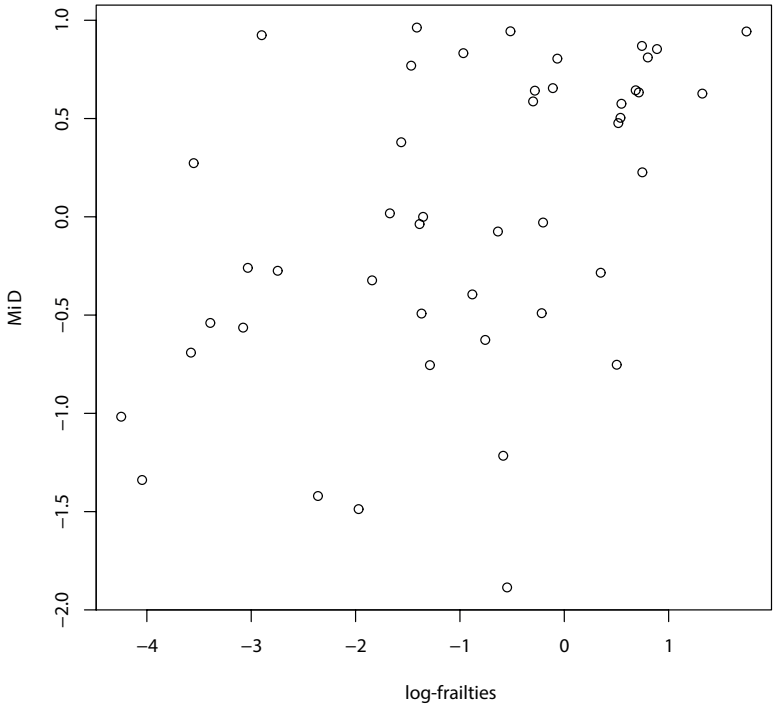

Figure 35: Martingale residuals of terminal event versus the posterior log-frailties estimated from the recurrent events

of a common hidden factor influencing both risks of experiencing recurrent events and the terminal event. The plausibility of this should be assessed separately, because more models can describe the type of data encountered in this chapter (Cook and Lawless, 2007, ch. 6.6) and the effects of internal time dependent covariates are often difficult to separate from that of the frailty (Aalen, Borgan, and Gjessing, 2008, ch. 8.5).

The fact that the martingale residuals and the estimates of the log-frailty are not samples coming from a bivariate normal distribution should also lead to a cautious interpretation of correlation coefficients and of the test statistic (3.8). In the simulations of Section 3.4 we did not notice any increase in the estimated α levels of the RST, but this might depend on the data set on which the method is employed. Finally, note that there is no closed form connection between the parameter which describes the association between recurrent events and terminal event γ and the correlation ρ used to calculate the RST statistic (3.8).

Although we have not explicitly stated that the frailty should follow a gamma distribution throughout Section 3.2, we still employed this assumption in Sections 3.4 and 3.5. The RST can accommodate any distribution for the frailty, including, for example,

a two-point mixture or a compound Poisson distribution, as long as the shared frailty model for recurrences can be estimated. It can be seen from (3.8) that the choice of the frailty distribution will affect only the estimation of $\widehat{\log z_i}$. We expect the RST to have the largest power if the true frailty distribution is used, however this was not checked in the simulation study.

The idea of a simple test, here in the form of RST, could be extended to more models which inherit the issues which would prevent practicians to use a more complicated joint model. Because a recurrent event data in the presence of a terminal event is a particular case of a multistate model with competing risks (Cook and Lawless, 2007, ch. 6.6), similar methods could be found by generalizing RST to multistate models with frailty (Putter and Houwelingen, 2015).

Appendix: Estimation via profile likelihood

In Sections 3.3.2 and 3.4 we used the profiling out of the nuisance parameters from the log-likelihood (3.3), in the sense shown by the definition (3.9). First, note that, if $(\hat{\gamma}, \hat{\eta})$ maximizes (3.3), then $\hat{\gamma}$ maximizes (3.9), and $\hat{\eta}$ is the estimate of η obtained by maximizing $\ell(\hat{\gamma}, \eta)$. It is clear that

$$\ell_{\text{prof}}(0) = \ell(0, \hat{\eta}_0)$$

and $\ell_{\text{prof}}(\gamma) \geq \ell(\gamma, \hat{\eta}_0)$ with equality only when $\gamma = 0$. It follows that $\ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \geq 0$. Thus,

$$\frac{d}{d\gamma} \left\{ \ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \right\} \Big|_{\gamma=0} = 0,$$

which shows that $U_\gamma(0, \hat{\eta}_0)$ from (3.7) is equal to the efficient score function, $U_\gamma(0) = \frac{d}{d\gamma} \ell_{prof}(\gamma) \Big|_{\gamma=0}$. This justifies why (3.7) is the correct score function for testing H_0. Further, because $\ell_{prof}(\gamma) - \ell(\gamma, \hat{\eta}_0)$ is always positive and it has a minimum, it follows that

$$\frac{d^2}{d\gamma^2} \left\{ \ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \right\} \geq 0$$

for any value of γ. This implies that $\frac{d^2}{d\gamma^2} \ell_{prof}(\gamma) > \frac{d^2}{d\gamma^2} \ell(\gamma, \hat{\eta}_0)$ for all values of γ, which is equivalent to

$$-\frac{d^2}{d\gamma^2} \ell_{\text{prof}}(\gamma) = I_\gamma \leq I_{\gamma\gamma} = -\frac{d^2}{d\gamma^2} \ell(\gamma, \hat{\eta}_0)$$

for all γ. We conclude that $\ell_{\text{prof}}(\gamma)$ and $\ell(\gamma, \hat{\eta}_0)$ have the same value and the first derivative in $\gamma = 0$, but the curvature of $\ell(\gamma, \hat{\eta}_0)$ is more pronounced. This is the intuition behind the reason why the likelihood $\ell(\gamma, \hat{\eta}_0)$ with fixed nuisance parameters can be used to obtain the correct score, but not the correct information.

Chapter 4

Ascertainment correction in frailty models for recurrent events data

Abstract

In retrospective studies involving recurrent events, it is common to select individuals based on their event history up to the time of selection. In this case, the ascertained subjects might not be representative for the target population, and the analysis should take the selection mechanism into account. The purpose of this chapter is two-fold. First, to study what happens when the data analysis is not adjusted for the selection, and second, to propose a corrected analysis. Under the Andersen-Gill and shared frailty regression models, we show that the estimators of covariate effects, incidence and frailty variance can be biased if the ascertainment is ignored, and we show that with a simple adjustment of the likelihood, unbiased and consistent estimators are obtained. The proposed method is assessed by a simulation study and is illustrated on a data set comprising recurrent pneumothoraces.

4.1 Introduction

In the study of recurrent events it is of interest to model the rate at which the events occur in time, along with estimating the effects of different factors which may influence

This chapter has been published as: T.A. Balan, M.A. Jonker, P.C. Johannesma, H. Putter (2016). Ascertainment correction in frailty models for recurrent events data. *Statistics in Medicine* 35(23), 4183-4201.

this rate, such as treatments or individual covariates. Usually, it is assumed that the process which generates the recurrent events starts at a time 0; this can be, for example, the diagnosis of a certain disease, a medical intervention, or birth, and that it is ended by some form of right-censoring. Ultimately, the research aims to extrapolate the conclusions from a sample of individuals to a larger "population at risk". The sample selection process is critical for the validity and interpretation of the results.

In a prospective study a random sample is drawn from the target population at time 0, and then followed up for the occurrence of events, whereas in a retrospective study design, the sample is selected at a time later than 0, with the data up to the time of selection being collected on the ascertained individuals. Prospective studies are desirable although they may require a long time to be conducted. Their main advantage is that all aspects of the data collection are under the control of the researcher. Retrospective studies are usually observational in nature. While cheaper and shorter than prospective studies, they are associated with less control on the sample selection process. Ideally, the sampling mechanism should lead to a sample that can be viewed as a random representation of the full population of interest at time 0.

When the sampling happens at a time point after 0, the probability for a subject to be included in the study may depend on the subject's event history. For example, registries are often kept only for patients who experienced some recurrent events, not on the whole population that is at risk to experience these events. Such a sample can not be regarded as representative for the target population. The necessity to adjust the analysis to take the selection mechanism into account has been underlined in the context of recurrent events in Cook and Lawless (2007, ch. 7.3), although most approaches for this problem are ad-hoc in nature.

In the motivating example of this chapter, only subjects who experienced at least one occurrence of the event of interest between 1990 and 2014 were registered. Hence, subjects who only experience events before 1990 or after 2014 are not included in the study. As a consequence, the individuals who have a higher rate of recurrent events are over-represented in the sample. If not adjusted for, the ascertainment can bias the estimation of model parameters in the statistical analysis of such data. Selection bias is a known problem in epidemiology (Hernán, Hernández-Díaz, and Robins, 2004). Several paradoxical results in studies involving recurrent events might be explained by a closely related "index bias" (Dahabreh and Kent, 2011).

The effects of the selection scheme are more difficult to disentangle when random effects are used to model additional heterogeneity or correlation structures present in the data. The frailty model (Vaupel, Manton, and Stallard, 1979) is commonly used for recurrent events or clustered failures data (Hougaard, 2000; Cook and Lawless, 2007). When the selection of individuals depends on the previous history of events, it might also depends on the value of the random effect, further complicating the estimation of frailty models.

Most of the literature on event-based selection in this context has focused on models for the waiting times (gaps) between the events. For example, Scheike, Petersen, and

Martinussen (1999) use an adjusted frailty model to analyze time to pregnancy. In their case, the selection scheme reflects itself in truncation of the observed gap times. Their approach can be used when the selection of individuals is directly related to the length of the waiting times, rather than a specified calendar time interval. It is worth mentioning that a closely related problem is that of frailty models for clustered survival data in the presence of left truncation. Jensen et al. (2004) proposed a corrected likelihood for family data when observations are collected only on the failure has not occured before some time. Several papers followed up on this work (Van den Berg and Drepper, 2011; Eriksson, Martinussen, and Scheike, 2015). Sun and Li (2004) used a frailty model to analyze clustered survival data, where random effects are used to describe a familial structure. In their work, a family is ascertained when at least two members have experience the failure before a certain age. They also provide an "ascertainment-adjused" likelihood, with the focus lying on estimating parameters which describe the latent structure. The selection scheme in the present motivating example is similar. However, we focus on quantities which are of more interest in the recurrent event context, such as covariate effects or the intensity of the recurrent event processes.

We show that the selection based on event history may lead to a sample which is not representative for the initial population at risk, even in very simple ascertainment scenarios, and that this may lead to biased estimates when not properly accounted for in the analysis. The novelty of this chapter lies in the fact that we analyze the effects of ignoring the selection process in the context of recurrent events, along with comparing the adjusted and unadjusted estimators. In Section 4.2, we review the Andersen-Gill and the shared frailty models, we discuss the general idea of constructing a likelihood for models which take the selection mechanism into account, and we propose estimation procedures for both parametric and semiparametric models. The proposed methods are evaluated through a simulation study in Section 4.3, where we investigate properties of the estimators of the baseline intensities, regression coefficients and frailty variances under several scenarios. Finally, we illustrate the considerations of this chapter on a data set on recurrent pneumothoraces in Section 4.4, and we lay out our concluding remarks in Section 4.5.

4.2 Methods

This section is outlined as follows: in 4.2.1 we review the Andersen-Gill and the shared frailty models, and in 4.2.2 we adapt these models to take the selection mechanism into account. In 4.2.3 we discuss the estimation of the proposed adjusted models for parametric specification and we introduce a novel approach for their semiparametric estimation.

4.2.1 Statistical models

The canonical framework for recurrent events is that of counting processes, and particularly that of Poisson processes (Cook and Lawless, 2007, ch. 2). The history of events

of an individual i is "counted" by a stochastic process $N_i(t)$ for $t \geq 0$, with an intensity function $\lambda_i(t)$.

In the Andersen-Gill (AG) model, N_i is assumed to follow the specifications of a non-homogeneous Poisson process with intensity

$$\lambda_i(t;\, \beta, \phi) = \lambda_0(t; \phi) \exp\left(\beta' \mathbf{x}_i(t)\right) \tag{4.1}$$

where $\mathbf{x}_i$ is a vector of possibly time dependent covariates, β is a vector of regression coefficients and ϕ is a vector of parameters which characterize the "baseline" intensity λ_0. In the shared frailty model, N_i is assumed to follow the specifications of a non-homogeneous Poisson process conditional on the unobserved "frailty" $Z_i = z_i$:

$$\lambda_i(t|z_i; \beta, \phi) = z_i \lambda_0(t; \phi) \exp\left(\beta' \mathbf{x}_i(t)\right) \tag{4.2}$$

where it is assumed that $Z_i > 0$ and that the Z_i's are i.i.d. distributed with some density f_θ. In both cases we assume that the censoring is independent given the covariates, and for the shared frailty model we also assume that it is non-informative for the frailty.

Although there is a wide variety of distributions that can be used for Z_i, the most common are the gamma distribution and the log-normal distribution. In the rest of this chapter, we will consider f_θ as the gamma density with expectation 1 and variance θ,

$$f_\theta(z) = \frac{1/\theta^{1/\theta}}{\Gamma(1/\theta)} z^{1/\theta - 1} \exp(-z/\theta), \tag{4.3}$$

with $\theta > 0$ and for $z > 0$. This choice is particularly convenient because the marginal features can be obtained in closed form; see Nielsen et al. (1992) and Murphy (1995a). The AG model can be seen as a limiting case of the shared frailty model when $\theta \longrightarrow 0$; indeed, it can be seen that in this case all z_i's are equal to 1 and (4.2) simplifies to (4.1).

For a subject i, let n_i be the number of observed events. We denote the follow-up time as t_i, and the observed recurrent event times as t_{ij} with $j \in 1...n_i$. Traditionally, the observed data of a certain individual i is denoted as O_i and it represents the probabilistic event "n_i events at $t_{i1} < ... < t_{in_i}$ over the observation time $(0, t_i)$". In the absence of any event-dependent sampling, the construction of likelihoods based on counting processes is detailed in Kalbfleisch and Prentice (2002, ch. 6). In the AG model, from λ_i defined as in (4.1), $P(O_i)$ can be written as

$$P(O_i; \beta, \phi) = \prod_{j=1}^{n_i} \lambda_i(t_{ij}; \beta, \phi) \exp\{-\Lambda_i(t_i; \beta, \phi)\} \tag{4.4}$$

where $\Lambda_i(t_i; \beta, \phi)$ is the cumulative intensity, i.e. $\Lambda_i(t_i; \beta, \phi) = \int_0^{t_i} \lambda_i(s; \beta, \phi) ds$. This leads to the log-likelihood

$$\ell^O(\beta, \phi) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \{\log \lambda_0(t_{ij}; \phi) + \beta' \mathbf{x}_i(t_{ij})\} - \sum_{i=1}^{n} \int_0^{t_i} \exp(\beta' \mathbf{x}_i(s)) \lambda_0(s) ds. \tag{4.5}$$

In the shared frailty model, a similar expression as (4.4) is obtained conditional on the frailty $Z_i = z_i$, by using the conditional intensity (4.2). The unconditional marginal probability is obtained by taking the expectation over the random effects:

$$\begin{aligned} P(O_i; \beta, \phi, \theta) &= \mathrm{E}_{Z_i} P(O_i | Z_i; \beta, \phi) \\ &= \int_0^\infty \prod_{j=1}^{n_i} \lambda_i(t_{ij}|z_i; \beta, \phi) \exp\{-\Lambda_i(t_i|z_i; \beta, \phi)\} f_\theta(z_i) dz_i. \end{aligned} \tag{4.6}$$

If the frailty follows the gamma distribution with density (4.3), this leads to the log-likelihood

$$\begin{aligned} \ell^O(\beta, \phi, \theta) = &\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ \log \lambda_0(t_{ij}; \phi) + \beta' \mathbf{x}_i(t_{ij}) \right\} + \\ &+ \sum_{i=1}^{n} \left[-(1/\theta + N_{i.}) \log \left\{ 1/\theta + \int_0^{t_i} \exp(\beta' \mathbf{x}_i(s)) \lambda_0(s) ds \right\} + g_i(\theta) \right], \end{aligned} \tag{4.7}$$

where $g_i(\theta) = 1/\theta \log(1/\theta) + \log \Gamma(1/\theta + N_{i.}) - \log \Gamma(1/\theta)$ and $N_{i.}$ represents the total number of events observed for subject i; see Nielsen et al. (1992) for a rigorous and more detailed derivation of this expression.

4.2.2 Ascertainment adjustment

Ascertainment schemes The specification of A_i depends on the design of the study. We introduce three examples to provide the intuition behind this concept.

1. (Left truncation) At the time of the selection, data for subject i is available only if no event occurrences were observed until the age t_{Ri}, In this case, A_i is the probabilistic event "no events occurred between $t_{Li} = 0$ and t_{Ri}", and $P(A_i) = P(N_i(t_{Ri}) = 0)$.

2. In the case of recurrent events, registry data is available on subjects who experienced at least one occurrence in the last k years before the sampling time. Denote the age of individual i at selection as t_i. In this case, A_i is the probabilistic event "at least one event occurred in (t_{Li}, t_{Ri})" where $t_{Li} = t_i - k$ and $t_{Ri} = t_i$, and $P(A_i) = P(N_i(t_i) - N_i(t_i - k) > 0)$.

3. A population is at risk for recurrent events, although only a fraction actually experience an occurrence during follow-up. After the first event, the subjects enter a database where all subsequent recurrences are collected. If data is collected retrospectively from this database, at a time point where subject i's age is t_i, then A_i is the probabilistic event "at least one event in (t_{Li}, t_{Ri})" with $t_{Li} = 0$ and $t_{Ri} = t_i$, and $P(A_i) = P(N_i(t_i) > 0)$.

As in the motivating example, it is more often the case that, in studies involving recurrent events, subjects must experience at least one event during a certain time period, which we refer to as "one-in" ascertainment. This is illustrated by scenarios 2 and 3 above. We define the ascertainment interval (t_{Li}, t_{Ri}) with $0 \leq t_{Li} < t_{Ri} \leq t_i$. Left truncation is a particular selection scenario that, for clarity purposes, we will not develop further, and focus instead on the "one-in" ascertainment.

The adjusted likelihood For now, denote the parameters of the chosen model (AG or shared frailty) as η. We define the event A_i as the ascertainment event, the sampling of subject i from the "population at-risk". The case of interest is when A_i depends on the event history of subject i, and implicitly on the intensity of the counting process N_i. In this case, A_i is a more general event than O_i, since an individual needs to be ascertained in order for O_i to be observed, therefore $O_i \subset A_i$. This implies that $A_i \cap O_i = O_i$ and the likelihood contribution of subject i is given by

$$P(O_i|A_i;\eta) = \frac{P(O_i \cap A_i;\eta)}{P(A_i;\eta)} = \frac{P(O_i;\eta)}{P(A_i;\eta)}. \tag{4.8}$$

Heuristically, the meaning of (4.8) is that each contribution is weighted so that subjects with a low chance of being ascertained (small $P(A_i)$) receive more weight, as they are representative for a part of the population of interest which is under-represented in the ascertained sample.

We define the (ascertainment) adjusted likelihood as the product over the individual contributions (4.8). The adjusted log-likelihood for n individuals is given by

$$\ell(\eta) = \sum_{i=1}^{n} \log P(O_i;\eta) - \sum_{i=1}^{n} \log P(A_i;\eta). \tag{4.9}$$

We will refer to $\ell^O = \sum_{i=1}^n \log P(O_i;\eta)$ as the unadjusted log-likelihood. For the AG and shared frailty mode, this is given by (4.5) and (4.7). We denote the remaining part, $\ell^A = \sum_{i=1}^n \log P(A_i;\eta)$, as the ascertainment adjustment, so that $\ell(\eta) = \ell^O(\eta) - \ell^A(\eta)$.

One-in ascertainment: Andersen-Gill We define

$$\Lambda_{Ai}(\beta,\phi) = \int_{t_{Li}}^{t_{Ri}} \lambda_i(s;\beta,\phi)ds \tag{4.10}$$

with λ_i as specified in (4.1). The probability of no events in (t_{Li}, t_{Ri}) is $P(A_i;\beta,\phi) = \exp(-\Lambda_{Ai}(\beta,\phi))$. Therefore, when subjects are ascertained only when they experience at least one event in (t_{Li}, t_{Ri}),

$$P(A_i;\ \beta,\phi) = 1 - \exp\left(-\Lambda_{Ai}(\beta,\phi)\right).$$

This yields the adjusted log-likelihood

$$\ell(\beta,\phi) = \ell^O(\beta,\phi) - \sum_{i=1}^{n} \log\{1 - \exp\left(-\Lambda_{Ai}(\beta,\phi)\right)\} \tag{4.11}$$

with ℓ^O as defined in (4.5). As the window of observation, or more precisely Λ_{Ai} becomes smaller, the ascertainment correction becomes larger in magnitude. The score functions provide insight into the effect of the ascertainment adjustment.

Denote $S_i^{(1)}(s,t) = \int_s^t \mathbf{x}_i(u)\exp(\beta'\mathbf{x}_i(u))\lambda_0(u)du$. The derivatives of (4.11) with respect to the components of β are

$$U_\beta(\beta,\phi) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}_i(t_{ij}) - \sum_i S_i^{(1)}(0,t_i) - \sum_{i=1}^{n}\frac{\exp\left(-\Lambda_{Ai}(\beta,\phi)\right)}{1-\exp\left(-\Lambda_{Ai}(\beta,\phi)\right)}S_i(t_{Li},t_{Ri})$$

where $S_i(s,t) = \int_s^t \mathbf{x}_i(u)\exp(\beta'\mathbf{x}_i(u))\lambda_0(u)du$. The last term in this expression arises as from the ascertainment adjustment, and omitting it would lead to a biased estimate of β. In principle, similar considerations apply also for ϕ, the parameters which describe the baseline intensity. For example, if we consider the Breslow estimator for λ_0, i.e. ϕ is a vector of elements $\phi_k = \lambda_0(s_k)$ where s_k is a time point at which an event is observed, the score vector for ϕ is composed of elements

$$U_{\phi_k}(\beta,\phi) = \frac{N_k}{\phi_k} - \sum_{i=1}^{n} Y_i(s_k)\exp(\beta'\mathbf{x}_i(s_k)) - \\ - \sum_{i=1}^{n} Y_i^A(s_k)\exp(\beta'\mathbf{x}_i(s_k))\frac{\exp\left(-\Lambda_{Ai}(\beta,\phi)\right)}{1-\exp\left(-\Lambda_{Ai}(\beta,\phi)\right)} \tag{4.12}$$

where $Y_i(t)$ is an indicator function which is 1 as long as subject i is at risk at t and 0 otherwise, and $Y_i^A(t)$ is an indicator function which is 1 as long as $t \in (t_{Li}, t_{Ri})$. The second sum term in (4.12) appears due to the ascertainment correction and omitting it would lead to a biased estimate of ϕ.

One-in ascertainment: shared frailty Here we use the same definition for Λ_{Ai} as in (4.10) which can be interpreted as the integrated intensity over the ascertainment interval with the frailty fixed to 1. Conditional on the frailty, the probability of no events in (t_{Li}, t_{Ri}) is $\exp\left(-z_i\Lambda_{Ai}(\beta,\phi)\right)$. The unconditional probability is obtained by integrating over the random effect,

$$E_{Z_i}\left\{\exp\left(-Z_i\Lambda_{Ai}(\beta,\phi)\right)\right\} = \int_0^\infty \exp\left(-z_i\Lambda_{Ai}(\beta,\phi)\right) f_\theta(z_i)dz_i \\ = \frac{1/\theta^{1/\theta}}{\left(1/\theta+\Lambda_{Ai}(\beta,\phi)\right)^{1/\theta}}.$$

Therefore, when subjects are ascertained only when they experience at least one event in (t_{Li}, t_{Ri}),

$$P(A_i;\beta,\phi,\theta) = 1 - \frac{1/\theta^{1/\theta}}{\left(1/\theta+\Lambda_{Ai}(\beta,\phi)\right)^{1/\theta}}$$

yielding the adjusted log-likelihood

$$\ell(\beta, \phi, \theta) = \ell^O(\beta, \phi, \theta) - \sum_{i=1}^{n} \log \left\{ 1 - \frac{1/\theta^{1/\theta}}{(1/\theta + \Lambda_{Ai}(\beta, \phi))^{1/\theta}} \right\}, \tag{4.13}$$

with ℓ^O as defined in (4.7).

Similarly to the Andersen-Gill case, the ascertainment adjustment gives rise to an extra term involving Λ_{Ai} in the score functions for β, and this can be seen to be true also in the score function for θ. The extent of the bias which appears if the ascertainment adjustment is ignored depends in this case also on θ, in addition to ϕ and β. If, again, we consider the Breslow estimator with $\phi_k = \lambda_0(s_k)$ for s_k event time points, we obtain

$$U_{\phi_k}(\beta, \phi, \theta) = \frac{N_k}{\phi_k} - \sum_{i=1}^{n} Y_i(s_k) \exp\left(\beta' \mathbf{x}_i(s_k)\right) h_1(\beta, \phi, \theta) - \\ - \sum_{i=1}^{n} Y_i^A(s_k) \exp\left(\beta' \mathbf{x}_i(s_k) h_2(\beta, \phi, \theta)\right) \tag{4.14}$$

with

$$h_1(\beta, \phi, \theta) = \frac{1/\theta + N_{i.}}{1/\theta + \Lambda_i(t_i; \beta, \phi)}$$

and

$$h_2(\beta, \phi, \theta) = \frac{1/\theta^{1/\theta+1}}{(1/\theta + \Lambda_{Ai}) \left\{ (1/\theta + \Lambda_{Ai})^{1/\theta} - 1/\theta^{1/\theta} \right\}}.$$

4.2.3 Estimation of λ_0

The "baseline" intensity λ_0 is seen as parametrized by a vector of parameters ϕ. Our intention is to cover two cases: fully parametric models (where ϕ is low-dimensional) and semiparametric models (where ϕ is infinite-dimensional).

Parametric models Parametric specifications of λ_0, such as exponential or Weibull, which lead to closed forms of the log-likelihood can be estimated with general purpose optimization software such as the function `optim` in `R`. Such software also provides an estimate of the Hessian matrix at the maximum likelihood estimate from which standard errors can be obtained in a straight-forward way. In this chapter we choose a flexible piecewise constant specification for λ_0, where we consider the baseline intensity to be constant on a small number of intervals which partition the follow-up time. The limits of these intervals are determined so that they contain a roughly equal number of events.

Semiparametric models A semiparametric estimator for λ_0 is, for example, the Breslow estimator, which is obtained by solving the score equations corresponding to the score functions (4.12) for AG and (4.14) for the shared frailty model. Let $\lambda_0(t) = \phi_t$ for

t a known event time and 0 otherwise. The baseline intensity is then parametrized by $\phi = (\phi_1, ..., \phi_N)$ where N is the number of distinct event time points in the data. The difficulties induced by this specification are that the dimension of ϕ may be large, and it becomes larger as there are more unique event time points in the data set. Therefore, direct maximization of the log-likelihood is not usually feasible. We propose a new two-step iterative algorithm to obtain maximum likelihood estimates for semiparametric models.

Without any ascertainment adjustment, the high-dimensional ϕ parameter can be profiled out directly in the AG model (Johansen, 1983), and indirectly, within an Expectation-Maximization algorithm, in the shared frailty model (Nielsen et al., 1992). With ascertainment adjustment, these methods are not available. We propose to alternate between maximizing the log-likelihood with respect to the low-dimensional parameters β and θ for fixed ϕ and updating the high-dimensional parameter by solving a set of "pseudo score equations", which we derive from the score functions (4.12) and (4.14). If we denote the parameters of the model as η, then the score function for ϕ_k takes the form $U_{\phi_k}(\eta) = \frac{N_k}{\phi_k} - h(\eta)$ where h depends on whether the AG or the shared frailty model is used, and whether the likelihood is adjusted for ascertainment. For the adjusted models this can be seen in (4.12) and (4.14). We define the pseudo-score function as

$$\widetilde{U}_{\phi_k}(\phi_k|\tilde{\eta}) = \frac{N_k}{\phi_k} - h(\tilde{\eta}) \tag{4.15}$$

where η is seen as fixed to $\tilde{\eta}$. Solving the equation $\widetilde{U}_{\phi_k} = 0$ with respect to $\phi_k$leads to

$$\hat{\phi}_k = \frac{N_k}{h(\tilde{\eta})},$$

increases the log-likelihood if η is regarded as fixed to $\tilde{\eta}$. Finally, the algorithm follows the following steps. First, choose initial values β^0, θ^0 and ϕ^0, and fix a small $\varepsilon > 0$ as the desired precision.

1. At the ith iteration, maximize $\ell(\beta, \theta|\phi^{(i-1)})$, with ϕ fixed to $\phi^{(i-1)}$, with a general optimization software, e.g. `optim` in R. Obtain the updated $\beta^{(i)}$ and $\theta^{(i)}$.
2. Denote $\tilde{\eta} = \left(\beta^{(i)}, \theta^{(i)}, \phi^{(i-1)}\right)$ and solve the pseudo-score equations (4.15). Obtain the updated $\phi^{(i)}$.
3. Repeat steps 1 and 2 until $\ell(\beta^{(i)}, \theta^{(i)}, \phi^{(i)}) - \ell(\beta^{(i-1)}, \theta^{(i-1)}, \phi^{(i-1)}) < \varepsilon$

The advantage of this procedure is that it can estimate any model with a semiparametric baseline intensity for which an explicit expression of the pseudo-score (4.15) exists. The initial values β^0, θ^0 and ϕ^0 can be obtained from a Cox model ignoring any possible dependence between observations or ascertainment correction. A similar algorithm was proposed for frailty models without ascertainment correction (Gorfine, Zucker, and Hsu, 2006). Simulations which we do not show here indicate that the log-likelihood increases

with each iteration. Furthermore, for the semiparametric AG and shared frailty models without ascertainment adjustment, the estimates of the proposed algorithm coincide with the estimates provided by the available standard software. In the remainder of this chapter we fix the convergence criterion to correspond to $\varepsilon = 10^{-6}$.

Standard error estimates can be obtained from the "non-parametric information matrix", obtained by taking the second derivatives of the log-likelihood $\ell(\eta)$ with respect to all the parameters, including ϕ. This has been shown to lead to valid standard error estimates for the shared frailty model without ascertainment correction (Andersen, Klein, et al., 1997). As long as the estimates in the ascertainment-adjusted model enjoy similar asymptotic properties as the ones in the shared frailty model, a similar reasoning may be applied for this case. Since the semiparametric model can be seen as a limiting case of a parametric model with a piecewise constant baseline, with the piecewise intervals becoming smaller, it is to be expected that inverting the non-parametric information matrix will lead to correct estimates of the standard errors.

4.3 Simulation study

4.3.1 Toy example

We first consider a basic example to illustrate the bias which arises by not taking the ascertainment into account. For this we consider a "full" data set and a "truncated" version of the same data set where the subjects who have not experienced any event are removed. This reflects a simple situation where the selection of subjects is based on a registry where only individuals with at least one occurrence are present, as described by case 3 in Section 4.2.2.

First, we simulate subjects under a scenario where 300 individuals have the same risk, with $\lambda_i(t) = \lambda_i = 1$, without covariates or frailty. The cumulative intensity is then $\Lambda_i(t) = t$ for all subjects. In Figure 41 (left) the estimates $\hat{\Lambda}_i$ based on 20 full and truncated data sets are shown, and with the black line the true value is plotted. The cluster around this line are estimates based on the AG model on the full data set and the other lines are the estimates from the truncated data sets. It can be seen that if the subjects which do not experience any events during follow-up are not part of the data set, the uncorrected estimates are biased upwards.

Next, we simulate data sets and truncate them as before, this time with a binary covariate from a Bernoulli(1/2) distribution, as in (4.1). We repeat this procedure 30 times for a grid of values of β, we estimate $\widehat{\beta}$, and we collect the bias $\widehat{\beta} - \beta$. For every value of β a boxplot of the bias is shown in Figure 41 (center). It can be seen that, for $\beta < 0$, the estimate has a positive bias and for $\beta > 0$ the bias is negative. The absolute value of the bias is larger as β is further away from 0, and for negative values this phenomenon is more severe.

Finally, we simulate a large data set of 3000 patients from the frailty model (4.2), with a gamma distributed random effect (4.3) with $\theta = 1$. We truncate this data set in

the same way as described before. The individual frailty value is unknown, and it can be inferred from the conditional distribution of Z_i given the data. A particularity of the gamma shared frailty model is that this "posterior" distribution is also gamma, however with mean $1/\hat{\theta} + N_{i.}$ and variance $1/\hat{\theta} + \hat{\Lambda}_i(t_i)$, see (Nielsen et al., 1992). From the full data set, we compute the "posterior" frailty estimates, equal to the expectation of this conditional distribution:

$$\widehat{z_i} = \frac{1/\hat{\theta} + N_{i.}}{1/\hat{\theta} + \hat{\Lambda}_i(t_i)}.$$

The logarithm of these estimates are shown in a histogram in Figure 41 (right, above). We show the estimated frailties of the subjects who are part of the truncated data set in Figure 41 (right, below), clearly indicating that the one-in ascertainment favors the selection of individuals with a high frailty value. This is because a high frailty value is associated with a higher rate of recurrent events, leading to an ascertained sample which is less heterogeneous and not representative of the population at risk.

4.3.2 Set up

The idea of the simulation study is to first simulate a random sample of M subjects, from which the estimates of baseline intensity, regression coefficients and eventually frailty variance are obtained. These are regarded as the "correct" estimates. Next, from this data set we obtain 3 different "ascertained" data sets, by selecting only a subset of the M subjects. On these "ascertained" data sets, we perform two analyses: one ignoring the ascertainment correction, in order to assess the extent of the bias induced by event-dependent selection, and one in which the correct ascertainment correction is used, to evaluate how the estimates obtained from the adjusted likelihood compare to the "correct" ones.

By S0 we will refer to the full-data scenario, comprising the M simulated individuals. The three "ascertained" data sets are obtained from the following scenarios:

- By S1 we refer to the situation where only subjects who experience at least one event during follow-up are selected in the sample.
- In S2 only the subjects who experience at least one event during an observation window at the end of the follow-up are kept in the sample.
- In S3 only subjects who experience at least one event in an observation window in the middle of the follow-up are kept; this pertains to a selection scheme similar to S2, where the ascertained subjects are also followed until the end of their follow-up.

In all three cases we assume that the whole follow-up, including the events outside the ascertainment window, is known for the subjects in the sample.

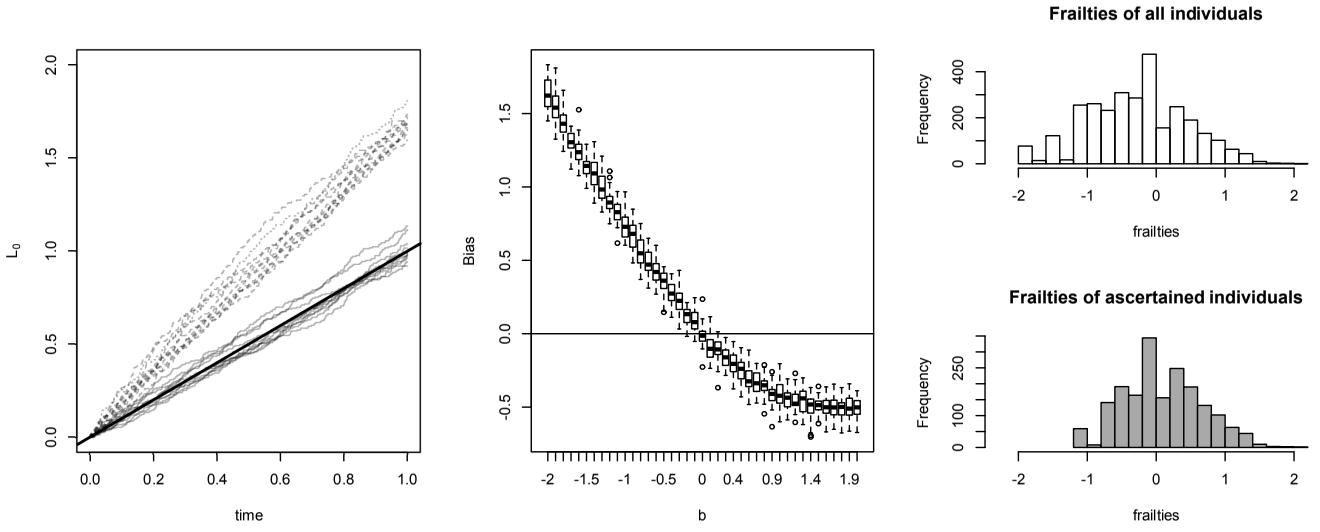

Figure 41: Left: estimates of Λ_0 based on the full data set clustered around the true value (black line) and from the truncated data set (the higher sloped lines); center: bias in log-hazard ratio for two groups when the ascertainment is ignored in the truncated data sets, from the model $\lambda_i(t) = e^{\beta' x_i}\lambda_0(t)$ with x_i an indicator variable for whether the subject belongs to the high-risk group, for values of β between -2 and 2; right: histograms of the posterior log-frailties estimates from the full data set analysis (above) and the frailties of the subjects which experience one event or more during follow-up (below).

The general set-up of the simulations is as follows: we take the baseline intensity $\lambda_0(t) \equiv \lambda_0 = 2$ for each individual. Two covariates are generated independently from a Bernoulli distribution with $P(X_{iq} = 1) = P(X_{iq} = 0) = 0.5$ for $i = 1...M$ and $q = 1, 2$. The regression coefficients used for the simulation are $\beta_1 = 1$ and $\beta_2 = -0.5$. The subjects are censored at time $t_C = 1$ or by a "drop-out process" determined by an exponential distribution with mean $2\exp(x_{i1})$, whichever comes first. For the frailty model (4.2), we simulate M gamma-distributed random variables according to (4.3) with $\theta = 0.5$, and subsequently with $\theta = 1$, and assume that the dropout does not depend on the frailty.

The simulations consist of 1000 replicated data sets, each with $M = 500$ counting processes (individuals) that are simulated from a Poisson process, with the intensities (4.1) for the no-frailty case and (4.2) for the frailty cases, i.e.

$$\lambda_i(t) = 2\exp(\beta_1 x_{i1} + \beta_2 x_{i2})$$

for the $\theta = 0$ (AG) case and

$$\lambda_i(t|z_i) = 2z_i\exp(\beta_1 x_{i1} + \beta_2 x_{i2})$$

for the $\theta = 0.5$ and $\theta = 1$ cases.

For scenario S2 the observation window is chosen as $(0.7, 1.0]$ and for scenario S3 as $(0.3, 0.5)$. Because the data sets used in S1, S2 and S3 are subsets of the full data set S0, they contain fewer individuals. The average sizes of the truncated data sets is shown in Table 41.

Table 41: Data set sizes for S0, S1, S2 and S3 in terms of average number of individuals (M) and average number of events per individual ($N./M$)

		S0	S1	S2	S3
$\theta = 0$					
	M	500	384.15	181.64	173.72
	$N./M$	2.34	3.05	4.08	4.00
$\theta = 0.5$					
	M	500	342.59	161.28	157.93
	$N./M$	2.35	3.43	4.87	4.87
$\theta = 1$					
	M	500	308.37	146.07	144.06
	$N./M$	2.34	3.80	5.56	5.64

For the regression parameters and the frailty variance, the estimates are evaluated according to systematic bias, root mean-squared error and . The systematic bias is defined as

$$\frac{1}{1000}\sum_{j=1}^{1000}(\widehat{\beta}_{qj} - \beta_q)$$

for $q \in 1, 2$, where $\widehat{\beta}_{qj}$ is the estimate of β_q from the j-th simulation and β_q is the true value of the parameter. The root mean-squared error (RMSE) is defined as

$$\frac{1}{1000}\sqrt{\sum_{j=1}^{1000}(\widehat{\beta}_{qj} - \beta_q)^2}$$

and the coverage is the percentage of times that the 95% confidence interval of the estimates contains the true value of the parameter. For the regression coefficients, the estimated standard error obtained from the maximization software is used to construct symmetric confidence intervals. A special case is represented by θ, which is restricted to positive values. Instead, an unrestricted estimate $\log\theta$ is obtained from the maximization of the likelihood together with a standard error $\text{se}(\widehat{\log\theta})$. A symmetric 95% confidence interval for $\log\theta$ can then be constructed on the log-scale as

$$\left[\widehat{\log\theta} - 1.96 * \text{se}(\widehat{\log\theta}),\ \widehat{\log\theta} + 1.96 * \text{se}(\widehat{\log\theta})\right].$$

After exponentiating these bounds, a 95% asymmetric confidence interval is obtained for $\widehat{\theta}$.

Both ascertainment unadjusted and adjusted estimates are obtained from a self-written algorithm which maximizes the likelihoods (4.11) and (4.13), using a piecewise constant parametric form for λ_0. Throughout the simulations, the time axis is split into 8 intervals, which are chosen so that each interval includes roughly the same number of events. This implies that the intervals themselves may vary from simulation to simulation.

4.3.3 Simulation results

Andersen-Gill The estimates of the baseline intensity λ_0 from the unadjusted and adjusted AG model are shown in Figure 42, the estimate from each simulation being represented by a piece-wise constant function. The 8 intervals are of roughly similar length across the simulations due to the Poisson specification in Section 4.3.2. It can be seen that scenario S1 induces an upward bias that seems to be reasonably constant throughout time. S2 and S3 seem to bias the analysis at all time points, to a similar extent as S1, however with peaks during the observation window. The adjusted estimates are unbiased; nevertheless, for the heavier ascertainment scenarios S2 and S3 the estimates seem to exhibit a noticeably larger variance.

Boxplots of the unadjusted and adjusted estimates of β_1 and β_2 are shown in Figure 43. It can be seen that the adjusted estimates are unbiased, although they exhibit a larger variance. Indeed, the ascertainment-adjusted estimates also have higher estimated standard errors (not shown here).

The unadjusted estimators are also associated with an underestimated standard error (not shown here). This reflects itself in the poor coverage properties of the estimators. The simulation results of the $\theta = 0$ (AG) case are summarized in Table 2. It can be

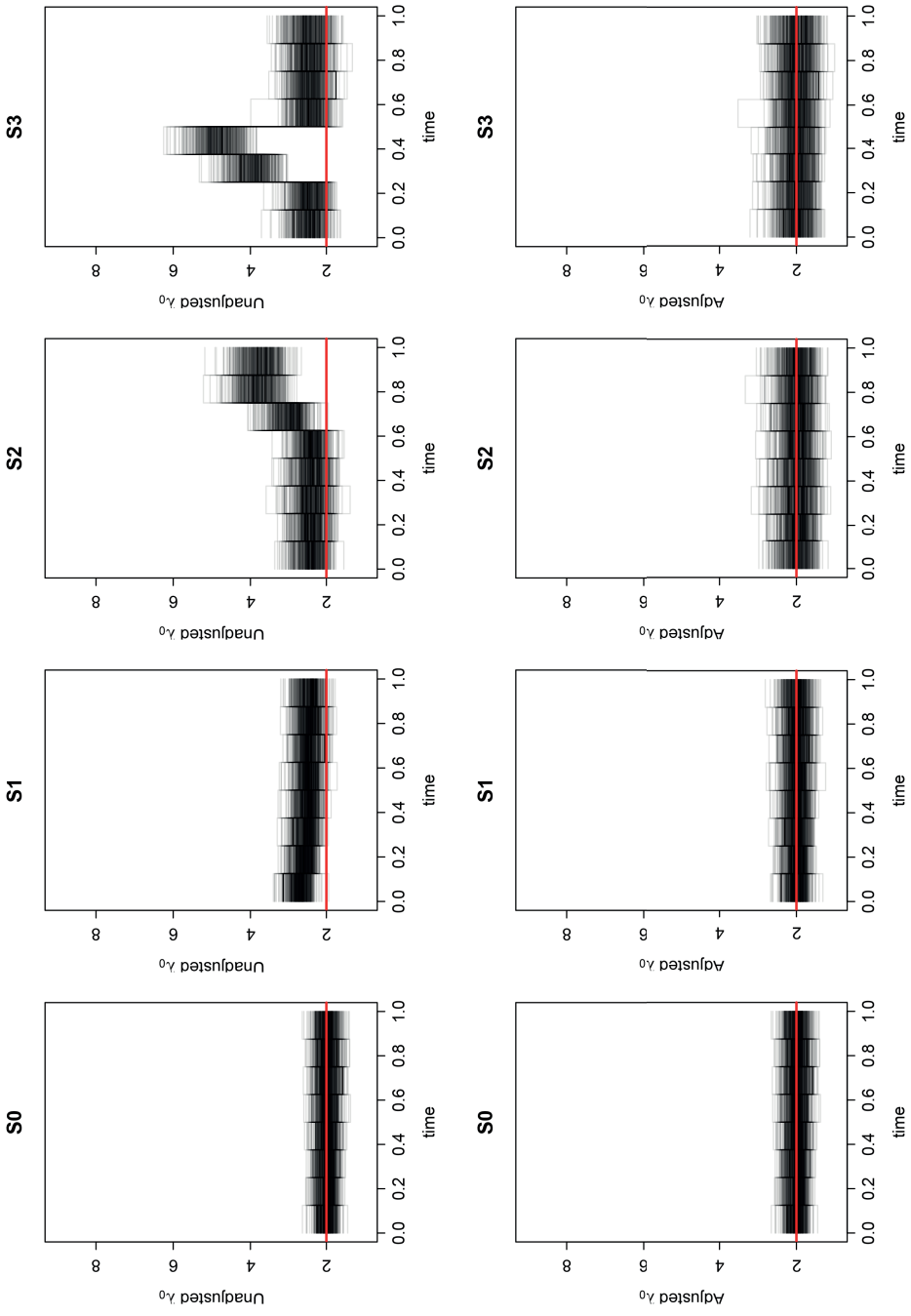

Figure 42: Baseline intensities, AG model; the horizontal line at $y = 2$ corresponds to the true $\lambda_0 = 2$.

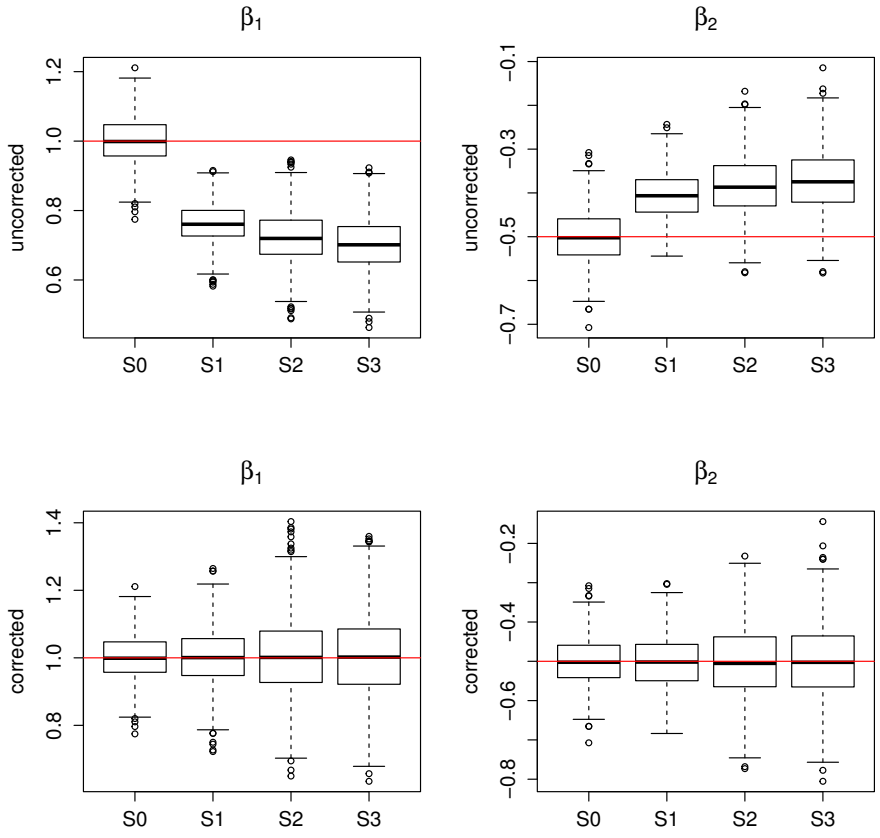

Figure 43: Point estimates for regression coefficients, AG model. The horizontal lines correspond to the true value of the parameters, $\beta_1 = 1.0$ and $\beta_2 = -0.5$.

observed that, if not corrected for ascertainment, the estimates of β_1 are affected more than those for β_2. This is due to an imbalance which is caused in the data set: because x_1 also influences the risk of censoring as specified in Section 4.3.2, it is less likely that the subjects experience events during the ascertainment window, simply because they have less time at risk. The correct model, which adjusts for the ascertainment, provides unbiased estimates and show a drastic reduction of the RMSE from their unadjusted counterparts.

Table 42: Simulation results, AG model

		Uncorrected			Corrected		
		S0	S1	S2	S3	S1	S2	S3
β_1								
	Bias	0.003	-0.232	-0.273	-0.293	0.010	0.012	0.011
	RMSE	0.062	0.238	0.283	0.303	0.085	0.117	0.127
	Coverage	0.966	0.036	0.124	0.106	0.950	0.956	0.952
β_2								
	Bias	0.000	0.096	0.119	0.127	0.000	0.002	0.001
	RMSE	0.057	0.109	0.136	0.146	0.067	0.089	0.091
	Coverage	0.968	0.678	0.690	0.690	0.952	0.956	0.962

Shared frailty We employ two scenarios for the frailty models, one of high heterogeneity ($\theta = 1$) and one of medium heterogeneity ($\theta = 0.5$). For the $\theta = 1$ scenario, the estimates of the baseline intensities are shown in Figure 44. By contrast with Figure 42, the 8 intervals are more different across simulations. This is due to the fact that the frailty-induced heterogeneity induces a more uneven spread of the events in time. Therefore, when determining the piecewise constant intervals as described in Section 4.3.2, the outcome can vary more than in the AG case. The larger heterogeneity, as represented by $\theta = 1$, also leads to more bias when the ascertainment is not adjusted for. However, the adjusted estimates are still unbiased and exhibit a larger variance, similarly with those shown in Figure 42. The baseline intensity estimates with $\theta = 0.5$ (not shown here) show a similar behavior.

The results are summarized in Tables 43 and 44, as well as in Figures 45 and 46. In terms of the regression coefficients, the bias is more severe in the higher heterogeneity scenario, even if only slightly so. As in the case of the AG model, the corrected estimates are unbiased. The major difference lies in terms of the estimate of θ. The unadjusted estimates show a large bias towards 0, notably more acute when $\theta = 1$. The large bias seems to lead to a very large variance of the adjusted estimators, as can be seen in Figure 45. Nevertheless, in the corrected estimators consistently provide major improvements in terms of RMSE and coverage over their unadjusted counterparts.

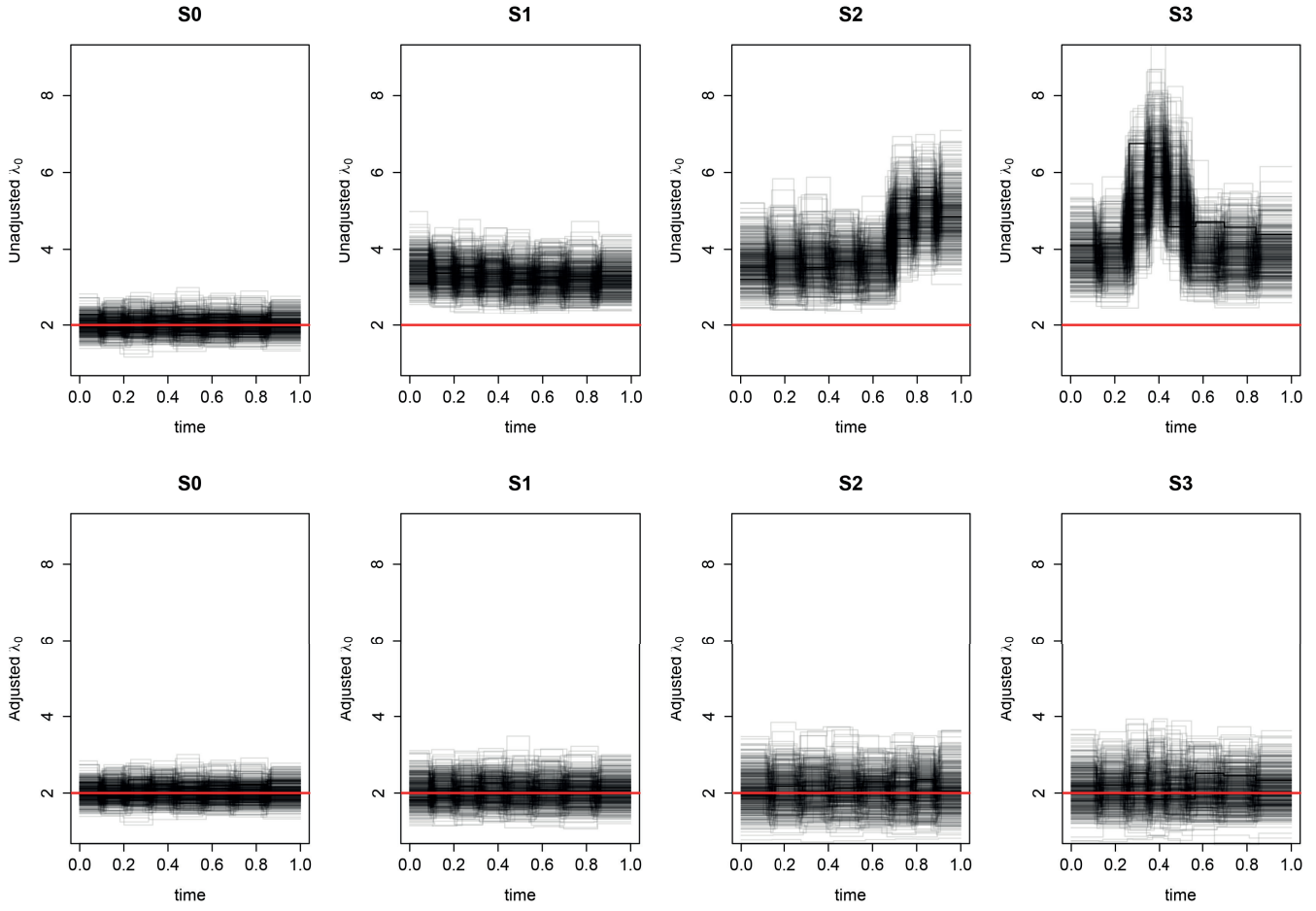

Figure 44: Baseline intensities, shared frailty model with $\theta = 1$; the horizontal line at $y = 2$ corresponds to the true $\lambda_0 = 2$.

Table 43: Simulation results, shared frailty model, $\theta = 1$

		Uncorrected				Corrected	
	S0	S1	S2	S3	S1	S2	S3
β_1							
Bias	0.004	-0.296	-0.302	-0.305	0.001	0.009	0.005
RMSE	0.116	0.311	0.327	0.332	0.148	0.190	0.196
Coverage	0.946	0.143	0.343	0.352	0.946	0.945	0.934
β_2							
Bias	0.002	0.140	0.147	0.145	0.001	0.001	-0.002
RMSE	0.115	0.172	0.193	0.193	0.139	0.175	0.180
Coverage	0.947	0.660	0.758	0.765	0.948	0.956	0.937
θ							
Bias	-0.007	-0.657	-0.695	-0.716	0.001	0.004	0.011
RMSE	0.108	0.659	0.697	0.718	0.237	0.344	0.404
Coverage	0.963	0.000	0.000	0.000	0.963	0.963	0.946

Table 44: Simulation results, shared frailty model, $\theta = 0.5$

		Uncorrected				Corrected	
	S0	S1	S2	S3	S1	S2	S3
β_1							
Bias	0.001	-0.286	-0.297	-0.301	0.000	0.003	0.003
RMSE	0.096	0.298	0.316	0.320	0.127	0.163	0.168
Coverage	0.937	0.069	0.240	0.241	0.935	0.941	0.934
β_2							
Bias	0.000	0.127	0.136	0.137	-0.003	-0.002	-0.005
RMSE	0.092	0.151	0.171	0.173	0.113	0.145	0.148
Coverage	0.941	0.630	0.717	0.707	0.954	0.954	0.946
θ							
Bias	-0.005	-0.304	-0.327	-0.34	-0.004	-0.007	-0.007
RMSE	0.066	0.306	0.330	0.343	0.109	0.154	0.169
Coverage	0.958	0.000	0.000	0.000	0.958	0.957	0.941

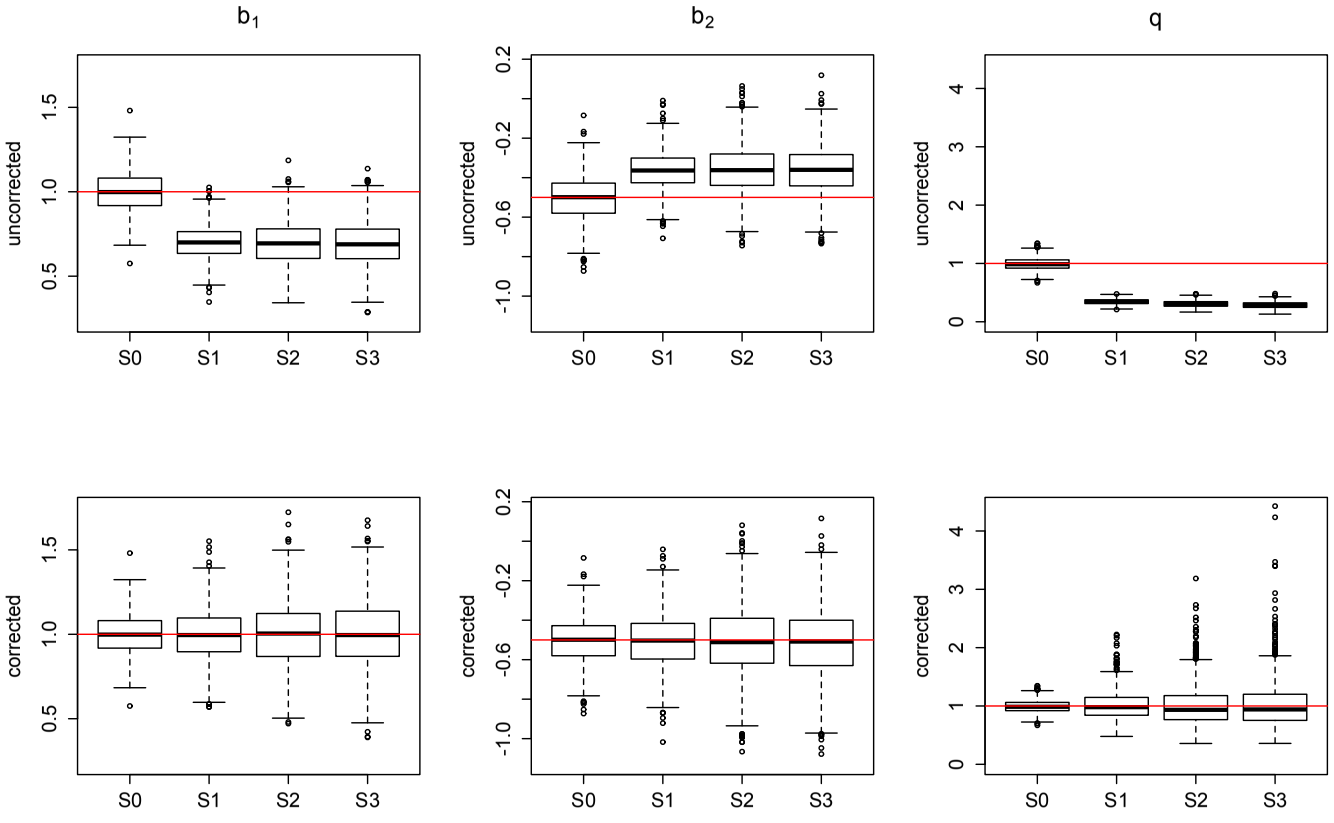

Figure 45: Point estimates, $\theta = 1$. The horizontal line corresponds to the true value of the parameters. Top: without ascertainment correction, bottom: with ascertainment correction.

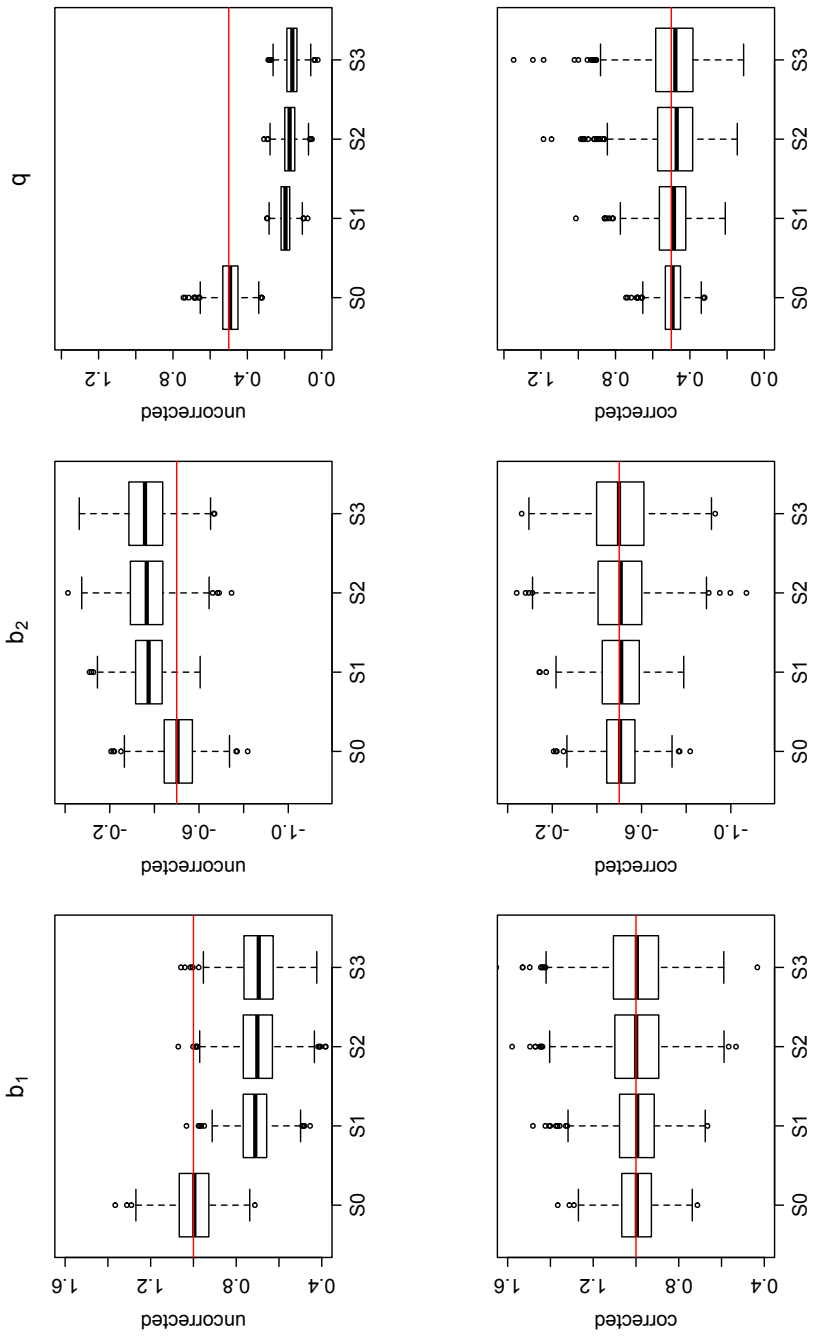

Figure 46: Point estimates, $\theta = 0.5$. The horizontal line corresponds to the true value of the parameters. Top: without ascertainment correction, bottom: with ascertainment correction.

4.3.4 Incomplete history

In the selection schemes described in Section 4.2, it is essential that, if an individual is selected for the study, the whole history of events outside the observation window is collected. As will be seen in the motivating example in Section 4.4, this might not always be the case. To assess the performance of the indicated adjusted models, we consider the scenario when the data on the history previous to the beginning of the observation period is (partly) missing. Using the same data sets that were simulated before obtained under selection schemes 2 and 3, we induce this incomplete character of the data in the following way. In the "mild incompleteness" scenario, 10% of the individuals are randomly selected from the ascertained data sets. For them, a "recollection time" is generated from a uniform distribution between 0 and the left time point of the observation window (0.7 in S2 and 0.3 in S3). The events before this time point are subsequently removed from the data set and the adjusted and unadjusted analyses are performed. In the "heavy incompleteness" scenario, the same is repeated with 50% of the individuals in the data sets. The results for $\theta = 1$ are summarized in Table 45. For $\theta = 0.5$, similar results were observed and are not shown here.

The corrected estimates of the frailty variance θ seem to be the most severely affected, particulary in scenario S2. The large bias (0.114 and 0.668), coupled with very wide confidence intervals (with coverages of 0.975 and 0.998) indicate that the standard errors are overestimated. In Figure 47 we plot the ascertainment-adjusted estimated baseline hazards for this case. By comparison with 44, it can be seen that the ascertainment adjustment does not work as well. The missing event history before time 0.7 leads to the underestimation of the intensity of the recurrent events process, mostly visible in the 50% missing case.

The general conclusion is that, when unadjusted for ascertainment, the incompleteness seems to slightly aggravate the problems illustrated in Tables 43 and 44. Nevertheless, the bias, RMSE and coverage remain comparable. The ascertainment adjustment seems to work well also with the incomplete data sets in terms of regression coefficients, at the price of a small increase in bias in and a slight increase in RMSE. In terms of the estimation of θ, we remark that the adjustment induces a positive bias to the estimates. In addition, the overly optimistic results of the coverage, in conjunction with the increase in RMSE and the bias results, reveals an over estimation of the standard errors. We conclude that, when the events outside the observation window are not completely collected, the ascertainment correction is robust in regards to the regression coefficients, however the frailty variance parameter should be interpreted with caution.

4.4 Data analysis

Data description The motivating data set comprises observations on primary spontaneous pneumothoraces (PSP). Risk factors for developing a PSP are male gender, smoking, and age, with a peak at 25-35 years of age; see Baumann and Noppen (2004) for an overview on the recurrent characteristics of PSPs.

Table 45: Simulation results, $\theta = 1$, with 10% and 50% missing data outside the observation window

		Uncorrected				Corrected			
		S2 10%	S2 50%	S3 10%	S3 50%	S2 10%	S2 50%	S3 10%	S3 50%
β_1									
	Bias	-0.310	-0.333	-0.308	-0.319	0.011	0.037	0.006	0.009
	RMSE	0.335	0.360	0.335	0.345	0.192	0.221	0.198	0.203
	Coverage	0.334	0.316	0.350	0.318	0.942	0.948	0.935	0.941
β_2									
	Bias	0.157	0.167	0.154	0.158	0.009	-0.003	0.007	0.003
	RMSE	0.205	0.215	0.200	0.204	0.188	0.206	0.183	0.189
	Coverage	0.719	0.715	0.744	0.738	0.944	0.959	0.929	0.929
θ									
	Bias	-0.689	-0.675	-0.716	-0.722	0.114	0.668	0.037	0.101
	RMSE	0.691	0.677	0.718	0.724	0.432	1.136	0.416	0.498
	Coverage	0.000	0.000	0.000	0.000	0.975	0.998	0.959	0.968

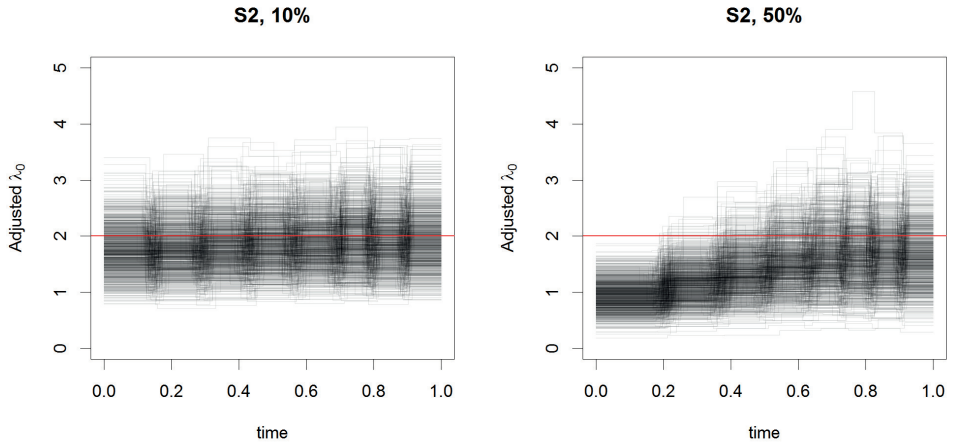

Figure 47: Ascertainment-adjusted estimates of baseline intensities, shared frailty model with $\theta = 1$, scenario S2, with 10% incompleteness (left) and 50% incompleteness (right) before the ascertainment window. the horizontal line at $y = 2$ corresponds to the true $\lambda_0 = 2$.

More recently, several genetic syndromes have been associated with an increased risk for (recurrent) episodes of spontaneous pneumothorax, like the Birt-Hogg-Dubé (BHD) syndrome, see Menko et al. (2009) and Johannesma et al. (2015). BHD is vastly under-diagnosed and usually patients with PSP do not receive a genetic test for this event, although (recurrent) PSP in BHD patients are caused by multiple cysts in the lower parts in the lung. By contrast, the non genetic PSP does not show these cysts at all on a thoracic CT-scan.

A variety of treatments are available for PSP, which we can divide into 3 categories: conservative (waiting, drainage, manual aspiration), sticking (pleurectomy, (chemical) pleurodesis) and cutting (lobectomy, bullectomy). Usually, the patients first receive a non-invasive (conservative) treatment, followed by a more invasive treatment for the next recurrent episodes.

The selection of the patients occurred as follows; between 2010 and 2014 a questionnaire was sent to the patients treated after 1990 for (recurrent) PSP. A number of respondents returned to the hospital and received a folliculin (*FLCN*) test to confirm BHD and then their PSP and treatment history was recorded. Information on the location of the PSP was not available, except for which lung the event took place in. The age of the patients at each event was recorded, approximated to the year. In total, the data set comprises 95 patients out of which 65 had PSP episodes only on one of the lungs, with a total of 220 episodes of PSP.

We define a tie as PSPs occurring in the same year in the same lung. This is observed in 26 of the 125 lungs with events in the data set. The presence of ties poses a difficulty

in analyzing the gap times, since 0-length gaps are not meaningful. Even if artificially spread out over a small interval of time, this might lead to a wrong impression on the length of the true gaps. This poses less of a problem if the intensity of the occurrences is treated as a non-homogeneous Poisson process, with age as time scale. This is the course that we follow in this analysis.

The data are shown in a Lexis diagram (Plummer and Carstensen, 2011) in Figure 48. The ascertainment process can be clearly seen. The general lack of events prior to 1990 casts some doubt on the completeness of the retrospective data collection, however this aspect is not further considered here. The effects of incomplete collection of the event history prior to the observation window were analyzed by simulation in Section 4.3.4.

Model construction The next step is the construction of a model. Each individual is represented by two counting processes corresponding to the two lungs, λ_i^L and λ_i^R. The intensities of these two processes can be influenced either by subject-level factors or by lung-specific history. A priori, there is no reason why one lung should be at a higher risk than the other. The general idea is that the subject-specific factors affect both lungs equally, while the difference in treatments between the lungs account for the observed differences between λ_i^L and λ_i^R.

We choose to model the events on the age time scale for which we take a baseline intensity common to all lungs from all subjects. The non-parametric estimate of the baseline intensity will have jumps only at event times, with the first event occurs at age 16. However, for the piecewise constant baseline a start of the recurrent events process must be explicitly defined. We choose this as the age of 15, since the risk of PSP before puberty is practically 0. We include BHD carrier indicator as a time-constant covariate in the model. At the lung level, we assume that the lungs may be in 4 states: "not under treatment", if there was no previous event, or under one of the 3 treatments: *conservative*, *sticking* or *cutting*. To account for differences between individuals, we include a gamma frailty which is shared for both λ_i^L and λ_i^R. These may be due to unmeasured covariates, such as shared environmental or behavioral variables.

In terms of treatments, *sticking* and *cutting* should be compared to *conservative*. To accommodate this, we assume that the intensity gets multiplied by $\exp(\beta_{\text{cons}})$, $\exp(\beta_{\text{cons}} + \beta_{\text{stick}})$ or $\exp(\beta_{\text{cons}} + \beta_{\text{cut}})$ according to which treatment the lung is under. In this case, $\exp(\beta_{\text{cons}})$ is the intensity ratio between one lung which had at least one event and is under conservative treatment and one lung which had no event and is not under treatment. This effect can also be interpreted as the intensity ratio between a lung which experienced PSPs and one which has not. On the other hand, $\exp(\beta_{\text{stick}})$ and $\exp(\beta_{\text{cut}})$ are intensity ratios between a lung which is under *sticking* or *cutting* and a lung under *conservative* treatment. For example, for individual i without BHD and with frailty z_i, which at time t has the left lung under *cutting* treatment and the right lung on the

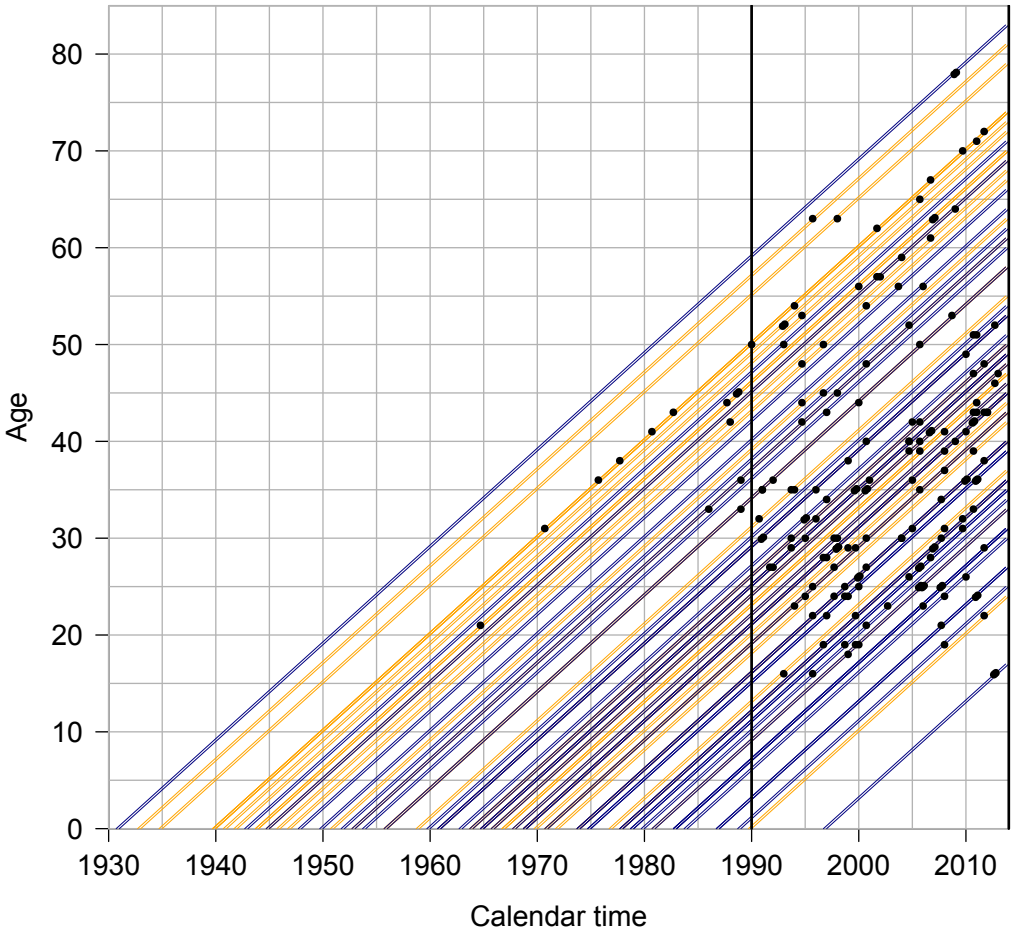

Figure 48: Lexis diagram of the PSP data. In blue the non-BHD patients. Dots represent observed events. The ascertainment window is marked between vertical lines.

conservative treatment, the intensities are

$$\lambda_i^L(t|z_i;\beta,\phi) = z_i \exp(\beta_{\text{cons}} + \beta_{\text{cut}})\lambda_0(t;\phi)$$
$$\lambda_i^R(t|z_i;\beta,\phi) = z_i \exp(\beta_{\text{cons}})\lambda_0(t;\phi)$$

A question of interest is whether the treatments perform equally for BHD and non-BHD patients. Finally, we also include interactions between these terms and BHD status.

Below, we show the results from a parametric baseline with 7 piecewise constant intervals and a semiparametric model.

Results The results are described in Table 46. The p-values for the regression coefficients are obtained from a Wald test statistic. It can be seen that adjusting for the ascertainment does not drastically influence the point estimates of the regression coefficients, which change at most by one standard error. Also, the estimated standard errors are larger in the adjusted model, in both the piecewise constant and in the semiparametric models. The noticeable effect of this is that the main effect of the conservative treatment loses siginificance, with the p-value increasing form 0.02 to about 0.3.

Other than that, we remark that statistical significance at the $\alpha = 0.05$ level was not observed for any of the variables in the model. Nevertheless, the adjusted model does give slightly different results. It can be seen that BHD patients are at a higher risk for PSPs as compared to non-BHD patients. For the non-BHD group, it can be seen that all treatments elevate the intensity of the event process. Among the 3 treatments, the sticking seems to perform better. For the BHD group, the cutting treatment seems to perform better, relative to conservative or sticking. If one of the lungs does not have any events, the intensity of the treated lung relative to the untreated one is modified multiplicatively by a factor of $\exp(0.419 + 0.075) = 1.63$. Conversely, for sticking this is $\exp(0.419 - 0.187 + 0.075 - 0.166) = 1.15$ and for cutting $\exp(0.419 + 0.595 + 0.075 - 0.934) = 1.16$.

In Figure 49, the unadjusted and adjusted baseline intensity estimates are shown, for both the semiparametric and the piecewise constant models. Similarly with the results of the simulation study in Section 4.3, the unadjusted baseline is overall larger than the adjusted estimate.

The p-values are missing in the case of the frailty variance θ, because the null hypothesis $H_0 : \theta = 0$ is at the border of the parameter space and a Wald test would not be valid in this case. A Likelihood Ratio Test for $H_0 : \theta = 0$ based on a χ^2 distribution with 1 degree of freedom can be constructed by contrasting the estimated frailty model versus the AG model, which is seen as the limiting case when $\theta \longrightarrow 0$, see Therneau and Grambsch (2000) and Nielsen et al. (1992). The LRT statistics for this hypothesis in the unadjusted / adjusted models are < 0.01 / 10.64 (piecewise constant) and < 0.01 / 8.57 (semiparametric model), corresponding to p-values of 0.98 / < 0.01 (piecewise constant) and 0.99 / < 0.01 (semiparametric model). The large differences in significance can be explained by noting that, without correcting for ascertainment, the effect of the frailty is not captured at all, as was seen in the simulation study in Section 4.3. It can however

Table 46: Data analysis results

		Unadjusted			Adjusted		
		Coef.	SE	p	Coef.	SE	p
Piecewise constant							
	BHD	0.073	0.19	0.71	0.212	0.35	0.55
	Cons	0.866	0.36	0.02	0.500	0.44	0.26
	Stick	-0.202	0.42	0.69	-0.220	0.51	0.67
	Cut	0.525	0.48	0.28	0.567	0.60	0.35
	BHD:Cons	0.126	0.44	0.78	-0.001	0.52	1.00
	BHD:Stick	-0.167	0.52	0.75	-0.087	0.62	0.89
	BHD:Cut	-0.707	0.67	0.29	-0.958	0.85	0.26
	θ	< 0.001	*NA*	-	1.302	4.18	-
Semiparametric							
	BHD	0.069	0.19	0.72	0.191	0.22	0.39
	Cons	0.871	0.38	0.02	0.419	0.42	0.33
	Stick	-0.221	0.42	0.61	-0.187	0.46	0.69
	Cut	0.491	0.49	0.32	0.595	0.55	0.28
	BHD:Cons	0.173	0.46	0.71	0.075	0.49	0.88
	BHD:Stick	-0.198	0.53	0.71	-0.166	0.58	0.77
	BHD:Cut	-0.646	0.67	0.34	-0.934	0.77	0.23
	θ	< 0.001	0.07	-	1.73	3.51	-

be seen that the standard errors corresponding to the estimator of θ are very large in the adjusted models, which was also the case in the simulation study in Section 4.3.4. This is in line with the suspicion that the history before 1990 was incompletely collected. The same phenomenon of large estimates and very large standard errors was observed in the same context in Section 4.3.4.

4.5 Discussion

We have shown in this chapter that event-based ascertainment may lead to biased results when unaccounted for. This bias can be severe and it may lead to very weak coverage properties, and the true effect of various factors might not be captured at all. We can correct for this bias with the methods proposed in Section 4.2. The merit of the approach used in this chapter is that unbiased results can be obtained if the event-dependent selection conditions are correctly accounted for in the estimation method. Furthermore, it was seen in Section 4.3 that the adjusted estimators only exhibit a small increase in the root mean-squared error as compared to the full-data scenario, as seen in Table 2, despite a smaller sample size. This suggests that the same results can be obtained from

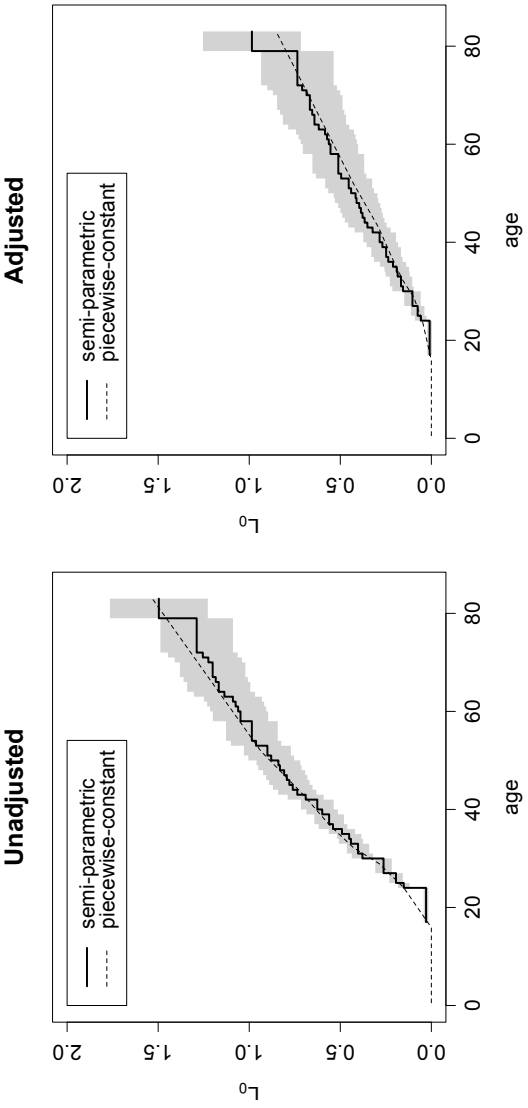

Figure 49: Adjusted and unadjusted estimates of the cumulative baseline intensity Λ_0. The gray band delimits the 95% confidence interval for the semiparametric estimate of Λ_0. With dotted lines, the parametric estimate with 7 piecewise constant intervals.

the two study designs, prospective study and retrospective study with event-dependent selection, as long as the ascertainment is correctly modeled. Hence, retrospective studies on recurrent events might prove to be a viable alternative to the prospective studies.

There are several limitations to the approach used throughout this chapter. First, we assumed that the whole event history of an individual can be collected at the time of the selection. This might not be true, especially when the event history has to be "remembered" by the patients. It can be seen in Figure 48 that very few events seem to happen before the start of the study (1990). It is possible that the subjects did not recall earlier events, or that registry data was not available for all the patients. However, the proposed methods showed promising results, as long as the complete history is collected for most individuals. The effects of incomplete collection of data outside the observation window are analyzed by simulation in Section 4.3.4.

Second, it is also common in the study of recurrent events that the subjects have to be alive at the time of the selection. If the rate of the recurrent events is associated with the terminal event, then joint models for recurrent and terminal events should be adopted; see, for example, Liu, Wolfe, and Huang (2004). In the data used in Section 4.4, it is reasonable to assume that death is not an event of interest, since the recurrent events in this case are not life-threatening. Nevertheless, Cook and Lawless (2007, ch. 7.3) provide some indication of possible strategies for this type of selection.

Third, as seen in Section 4.3, the adjusted estimators show a larger variance than their unadjusted counterparts, which is especially visible in the estimates of the frailty variance. This indicates that the frailty distribution itself is harder to identify than covariate effects in ascertainment-adjusted models.

Among the advantages of the approach presented in this chapter, is that several extensions can be obtained to accommodate more complicated models. First, in the framework outlined in Section 4.2, the distributional assumption for the frailty (gamma distribution) can be relaxed. Likelihoods can be constructed from (4.6) for a larger family of distributions, however these do not lead to closed form expressions; see Hougaard (2000).

Second, other similar models which lead to similar likelihood expressions as (4.4) or (4.6) could be accommodated with this approach. An example is a two-state Markov model for duration of recurrent episodes, where only subjects who have a first recurrence in an observation window are ascertained; see Cook and Lawless (2007, ch. 6.5).

Finally, we note that the framework introduced in Section 4.2 can be itself extended. As long as A_i is an event which is more general than O_i, in the sense of equation (4.8), a similar argumentation can be employed. This can be achieved by extending the definition of what amounts to the event history. Several examples can be found in Cook and Lawless (2007, ch. 7.3).

The promising results shown by the semiparametric estimation method proposed in Section 4.2.3 suggest that the properties of the algorithm should be further investigated, and completed by a proof of convergence. The `R` code used for the simulations in Section 4.3 is available upon request from the corresponding author. Future work, espe-

cially in terms of software development, would likely prove to be useful for clinicians. A focus of future research will be to provide an extension of R's `survival` package for ascertained and truncated data.

Chapter 5

frailtyEM: An R Package for Estimating Semiparametric Shared Frailty Models

Abstract

When analyzing correlated time to event data, shared frailty (random effect) models are particularly attractive. However, the estimation of such models has proved challenging. In semiparametric models, this is further complicated by the presence of the nonparametric baseline hazard. Although recent years have seen an increased availability of software for fitting frailty models, most software packages focus either on a small number of distributions of the random effect, or support only on a few data scenarios. **frailtyEM** is an R package that provides maximum likelihood estimation of semiparametric shared frailty models using the Expectation-Maximization algorithm. The implementation is consistent across several scenarios, including possibly left truncated clustered failures and recurrent events in both calendar time and gap time formulation. A large number of frailty distributions belonging to the Power Variance Function family are supported. Several methods facilitate access to predicted survival and cumulative hazard curves, both for an individual and on a population level. An extensive number of summary measures and statistical tests are also provided.

This chapter has been accepted for publication as T.A. Balan and H. Putter (2018). **frailtyEM**: an **R** package for estimating semiparametric shared frailty models. *Journal of Statistical Software*

5.1 Introduction

Time-to-event data are very common in medical applications. Often, these data are characterized by incomplete observations. For example, the phenomenon of right censoring occurs when the actual event time is not observed, but the only thing that is known is that the event has not taken place by the end of follow-up. Sometimes, individuals enter the data set only if they have not experienced the event before a certain time point. This is known as left truncation, which, if not accounted for correctly, leads to bias. Regression models for such data have been developed in the field of survival analysis. The most popular is the Cox proportional hazards model (Cox, 1972), which is semiparametric in nature: the effect of the covariates is assumed to be time-constant and fully parametric, while the time-dependent probability of observing an event arises from the nonparametric baseline hazard. Cox regression has been the standard in survival analysis for a few reasons. First, it does not require any a priori assumptions about the baseline hazard. Second, under the proportional hazards assumption, maximum likelihood estimation can be carried out efficiently using Cox's partial likelihood. Nowadays, such models may be estimated with most statistical software, such as R (R Core Team, 2016) Stata (StataCorp, 2017), SAS (Inc., 2003) or SPSS (IBM Corp, 2016).

When individuals belong to clusters, or may experience recurrent events, the observations are correlated. In this case the Cox model is not appropriate for modeling individual risk. A natural extension is represented by random effect "shared frailty" models. Originating from the field of demographics (Vaupel, Manton, and Stallard, 1979), these models traditionally assume that the proportional hazards model holds conditional on the frailty, a random effect that acts multiplicatively on the hazard. The variance of the frailty is usually indicative of the degree of heterogeneity in the data. This makes the choice of the random effect distribution relevant. However, the simplicity that made the Cox model so popular does not carry over to such models.

Arguably the most popular way of fitting semiparametric shared frailty models is via the penalized likelihood method (Therneau, Grambsch, and Pankratz, 2003), available for the gamma and log-normal frailty distributions. This is the standard in the **survival** package (Therneau and Grambsch, 2000; Therneau, 2015a) in R, in the `PHREG` command in SAS and the `streg` procedure in Stata. This method has the advantage that it is generally fast and the Cox model is contained as a limiting case when the variance of the frailty is 0. However, this algorithm can not be used for estimating other frailty distributions or left-truncated data, and the provided standard errors are presented under the assumption that the estimated parameters of the frailty distribution are fixed. Log-normal frailty models may also be estimated in R via Laplace approximation in **coxme** (Therneau, 2015b), h-likelihood in **frailtyHL** (Do Ha, Noh, and Lee, 2012) or Monte Carlo Expectation-Maximization **phmm** (Donohue and Xu, 2013; Vaida and Xu, 2000; Donohue, Overholser, et al., 2011). Parametric and spline based shared frailty models are implemented for the gamma and log-normal distributions in the **frailtypack** package (Rondeau, Mazroui, and Gonzalez, 2012; Rondeau and Gonzalez, 2005).

In Hougaard, 2000, the Power Variance Function (PVF) family was proposed for mod-

eling the frailty distribution. This family of frailty distributions includes the gamma, positive stable (PS), inverse Gaussian (IG) and compound Poisson distributions with mass at 0. Each choice of the distribution for the frailty implies a different marginal model, with some emphasizing early dependence of the observations (IG) and others late dependence (gamma). Of particular interest is the PS distribution: with assumed proportional hazards conditional on the frailty, the PS implies proportional hazards also unconditional on the frailty. This is unlike the other distributions which imply non-proportional hazards at the marginal level. Therefore, this is the only distribution where the potential violation of the proportional hazards is not confounded with a frailty effect.

The software implementation of the the PVF family of distributions so far been limited. At this time, two R packages incorporate a larger number of distributions from this family: the **frailtySurv** package (Monaco, Gorfine, and Hsu, 2017; Gorfine, Zucker, and Hsu, 2006) implements the above mentioned distributions except the PS via a pseudo full likelihood approach and the **parfm** package (Munda, Rotolo, and Legrand, 2012) estimates fully parametric gamma, IG, PS and log-normal frailty models.

In this chapter we present **frailtyEM** (Balan and Putter, 2017), an R package which uses the general Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) for fitting semiparametric shared frailty models. This implementation comes to complete the landscape of packages that may be used for such models, with support for the whole PVF family of distributions for the scenarios of clustered failures, clustered failures with left truncation and recurrent events data. In the latter case, different time scales are supported, such as calendar time (time since origin of the recurrent event process) and gap time (time since previous recurrent event). Point estimates for regression coefficients are provided with confidence intervals that take into account the estimation of the frailty distribution parameters, and plotting methods facilitate the visualization of both conditional and marginal survival or cumulative hazard curves with 95% confidence bands, marginal covariate effects, and empirical Bayes estimates of the random effects. A comparison with respect to functionality between **frailtyEM** and other R packages is provided in Table 51.

The rest of this chapter is structured as follows. In Section 5.2 we present a brief overview the semiparametric shared frailty model, and the implications of left truncation. In Section 5.3 we discuss the estimation method and its implementation. In Section 5.4 we illustrate the usage of the functions from the **frailtyEM** package on three classical data sets available in R.

5.2 Model

5.2.1 Shared frailty models

In **frailtyEM**, the general framework is of I clusters with J_i individuals within cluster i, $i = 1, \dots, I$. The event history of individual j from cluster i is represented by a counting process N_{ij}, with $N_{ij}(t)$ representing the number of events observed until time t. The

	frailtyEM	**survival**	**coxme**	**frailtySurv**	**frailtyHL**	**frailtypack**	**parfm**	**phmm**
Distributions								
Gamma	yes	yes	no	yes	no	yes	yes	no
Log-normal	no	yes	yes	yes	yes	yes	yes	yes
PS	yes	no	no	no	no	no	yes	no
IG	yes	no	no	yes	no	no	yes	no
Compound Poisson	yes	no	no	no	no	no	no	no
PVF	yes	no	no	yes	no	no	no	no
Data								
Clustered failures	yes	yes	yes	yes	yes	yes	yes	yes
Recurrent events (AG)	yes	yes	yes	no	no	yes	no	no
Left truncation	yes	no	no	no	no	yes	yes	no
Correlated structure	no	no	yes	no	no	yes	no	yes
Estimation								
Semiparametric	yes	yes	yes	yes	yes	no	no	yes
Posterior frailties	yes	yes	no	no	no	yes	no	no
Conditional Λ_0, S_0	yes	limited	no	yes	no	yes	yes	no
Marginal Λ_0, S_0	yes	no	no	no	no	no	no	no

Table 51: Comparison of R packages for frailty models. Versions: **frailtyEM** 0.8.3, **survival** 2.40-1, **coxme** 2.2-5, **frailtyHL** 1.1, **frailtypack** 2.10.5, **parfm** 2.7.1, **phmm** 0.7-5.

"at-risk" process $Y_{ij}(t)$ is defined as 1 when individual (ij) is under observation and 0 otherwise, and a vector of possibly time dependent covariates is denoted as $\mathbf{x}_{ij}(t)$.

The clustered failures scenario is represented when the $N_{ij}(t) \leq 1$ and $Y_{ij}(t) = 0$ after an event or right censoring. The data in cluster i consists of J_i possibly right censored survival times. If $N_{ij}(t)$ exceeds 1, the case of recurrent events is obtained. In this scenario, it is considered that each cluster contains only one individual ($J_i \equiv 1$, with the corresponding counting process N_i). Calendar time (also known as Andersen-Gill) models, when the time scale is "time since origin" and gap time models, where the time scale is "time since the previous event" are commonly employed (Cook and Lawless, 2007). When subject i is no longer under observation, the last time point is typically considered right censored.

The intensity of N_{ij} (or hazard, in the clustered failure scenarios) is specified as

$$\lambda_{ij}(t|Z_i) = Y_{ij}(t)Z_i \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(t) \tag{5.1}$$

where Z_i is an unobserved random effect common to all observations from cluster i (the "shared frailty"), β a vector of unknown regression coefficients and $\lambda_0(t) \geq 0$ an unspecified baseline intensity function. It is assumed that the Z_i are iid random variables with a distribution referred to as Z, and that event times are independent given Z_i. A stratified model (5.1) may also be specified by specifying different baseline intensities for different groups of observations. In this case, if individual (i, j) belongs to strata s, $\lambda_0(t)$ is replaced by $\lambda_{0s}(t)$.

We consider the general case where the Z follows a distribution with positive support from the infinitely divisible family, i.e., they are i.i.d. realizations of a random variable described by the Laplace transform

$$\mathcal{L}_Z(c; \alpha, \gamma) \equiv \mathrm{E}\left[\exp(-Zc)\right] = \exp(-\alpha\psi(c; \gamma)) \tag{5.2}$$

with $\alpha > 0$ and $\gamma > 0$. This formulation includes several distributions, such as the gamma, positive stable, inverse Gaussian and compound Poisson distributions. This so-called power-variance-function (PVF) family of distributions have been extensively studied in Hougaard, 2000. As detailed in Appendix A1, we assume that an identifiability constraint is imposed on the parameters α and γ and that the distribution of Z is indexed by a scalar parameter θ.

5.2.2 Likelihood

Henceforth, we consider the problem of estimating β, λ_0 and θ via maximum likelihood. This is achieved by maximizing the marginal likelihood, based on the observed data and obtained by integrating over the random effect. For simplicity, we omit potential strata in this section. From model (5.1), the marginal likelihood is obtained as the product over

clusters of expected marginal contributions, i.e.,

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \mathrm{E}_\theta \left[\prod_j \int_0^\infty \left\{ Y_{ij}(t) Z \exp(\beta^\top \mathbf{x}_{ij}(t) \lambda_0(t) \right\}^{dN_{ij}(t)} \times \exp\left(- \sum_j \int_0^\infty Y_{ij}(t) Z \exp(\beta^\top \mathbf{x}_{ij}(t)) \lambda_0(t) dt \right) \right]$$

The first part reduces to a product of contributions from the observed event times of the counting processes from cluster i. Denote the k-th observed time corresponding to the counting process N_{ij} as t_{ijk} and $\delta_{ijk} = 1$ if t_{ijk} is an event time and 0 otherwise. Let $\tilde{\Lambda}_i = \sum_j \int_0^\infty Y_{ij}(t) \exp(\beta^\top \mathbf{x}_{ij}(t)) \lambda_0(t) dt$ and $n_i = \sum_j \int_0^\infty Y_{ij}(t) dN_{ij}(t)$ the number of observed events in cluster i. The marginal likelihood can be written as

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \left[\prod_j \prod_k \left\{ \exp(\beta^\top \mathbf{x}_{ij}(t_{ijk})) \lambda_0(t_{ijk}) \right\}^{\delta_{ijk}} \right] \mathrm{E}_\theta \left[Z^{n_i} \exp(-Z\tilde{\Lambda}_i) \right]. \tag{5.3}$$

By using (5.2), the last term may be expressed in terms of the n_i-th derivative of the Laplace transform, i.e.

$$\mathrm{E}_\theta \left[Z^{n_i} \exp(-Z\tilde{\Lambda}_i) \right] = (-1)^{n_i} \mathcal{L}_Z^{(n_i)}(\tilde{\Lambda}_i).$$

In **frailtyEM**, the Breslow estimator is employed for the baseline hazard, i.e., $\lambda_0(t) \equiv \lambda_{0t}$ for t an event time, and 0 otherwise. This is equivalent with estimating $\int_0^t \lambda_0(s) ds$ as a step function with "jumps" of size λ_{0t} at event times.

5.2.3 Ascertainment and left truncation

The problem of ascertainment with random effect time-to-event data is usually difficult. If Z_i is the distribution of the frailty of cluster i and A_i denotes the event of selecting the observations in cluster i, the random effect distribution of cluster i given the ascertainment is of the form $Z_i|A_i$. The Laplace transform of $Z_i|A_i$ follows from Bayes' rule as

$$\mathcal{L}_{Z_i|A_i}(c) = \frac{\mathrm{E}\left[\mathsf{P}(A_i|Z_i) \exp(-cZ_i)\right]}{\mathrm{E}\left[\mathsf{P}(A_i|Z_i)\right]}. \tag{5.4}$$

Expressing $\mathsf{P}(A_i|Z_i)$ depends on the type of the study at hand and on the way the data were collected.

In **frailtyEM** an option is included to deal with the scenario of left truncation for clustered failures. Consider that from a cluster of size $\tilde{J}_i$, $J_i \leq \tilde{J}_i$ individuals are selected and A_i is the event "selecting J_i individuals with left truncation times $\mathbf{t}_{L,i} =$

$\left\{ t_{L,i1} \dots t_{L,iJ_i} \right\}$". Then A_i can be expressed as

$$\mathsf{P}(A_i|Z_i) = \mathsf{P}(T_{i1} > t_{L,i1}, T_{i2} > t_{L,i2} \dots T_{J_i} > t_{L,iJ_i}|Z_i).$$

A hidden assumption here is that the true cluster size $\tilde{J}_i$ does not depend on the frailty. For example, if a high frailty is associated with both a high rate of events and smaller cluster sizes, then the distribution of $\tilde{J}_i|Z$ must also be considered (Jensen et al., 2004).

Assume that, given Z_i, the left truncation times $\mathbf{t}_{L,i}$ are independent. In this case,

$$\mathsf{P}(A_i|Z_i) = \prod_{j=1}^{J_i} \exp\left(-Z_i \int_0^{t_{L,ij}} \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(t)dt \right). \tag{5.5}$$

A difficulty here is that the values of the covariate vector and of the baseline intensity must be known prior to the entry time in the study. Therefore, only cases when $\mathbf{x}_i$ is time constant are considered.

Denote $\tilde{\Lambda}_{L,i} = \sum_j \int_0^{t_{L,ij}} \exp(\beta^\top \mathbf{x}_{ij})\lambda_0(t)dt$. The marginal likelihood may be obtained from (5.3), (5.4) and (5.5) as

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \left[\prod_j \prod_k \left\{ \exp(\beta^\top \mathbf{x}_{ij}(t_{ijk}))\lambda_0(t_{ijk}) \right\}^{\delta_{ijk}} \right] \times \\ \times \frac{\mathrm{E}_\theta \left[Z^{n_i} \exp\left(-Z(\tilde{\Lambda}_{L,i} + \tilde{\Lambda}_i) \right) \right]}{\mathrm{E}_\theta \left[\exp(-Z\tilde{\Lambda}_{L,i}) \right]}.$$

It can also be seen that, if the frailty distribution is degenerate and has no variability (i.e. E_θ may be removed), then the contribution of $\tilde{\Lambda}_{L,i}$ cancels out. In particular, under left truncation, the Laplace distribution of $Z|A_i$ is given by

$$\mathcal{L}_{Z|A}(c) = \frac{\mathcal{L}(c + \tilde{\Lambda}_{L,i})}{\mathcal{L}(\tilde{\Lambda}_{L,i})}. \tag{5.6}$$

This distribution is often referred to as the frailty distribution of the survivors (Hougaard, 2000). If Z is from the PVF family, it can be shown that $Z|A$ is also in the PVF family. As a result, if Z is gamma distributed, then also $Z|A$ is gamma distributed.

Note that, in general, the ascertainment scheme does not have a simple description and $P(A_i|Z_i)$ may or may not be available in closed form. For example, in family studies, the families may be selected only when a number of individuals live long enough (Rodríguez-Girondo et al., 2018). In this case, (5.5) does not hold. In the case of registry data on recurrent events, individuals (clusters) may be selected only if they have at least one event during a certain time window (Balan, Jonker, et al., 2016). These specific cases are not currently accommodated by **frailtyEM**.

5.2.4 Analysis and quantities of interest

Inference

In **frailtyEM**, inference from the likelihood (5.3) is based on the non-parametric information matrix. This is obtained by treating each $\lambda_0(t) \equiv \lambda_{0t}$ as a finite-dimensional parameter. Even though its dimension grows with the number of event time points in the data, this has been shown to lead to consistent variance estimators (Andersen, Klein, et al., 1997).

For assessing whether the frailty model is a better fit than the Cox proportional hazards model, the likelihood ratio test may be used. With the parametrizations described in Appendix A1, this is a problem of testing on the edge of the parameter space, and the test statistic under the null hypothesis follows asymptotically a mixture of $\chi^2(0)$ and $\chi^2(1)$ distribution (Zhi, Grambsch, and Eberly, 2005). This test is provided as standard output in **frailtyEM**.

The Commenges-Andersen score test for heterogeneity Commenges and Andersen, 1995 is implemented in **frailtyEM**. It may be applied to a proportional hazards model as fitted by the `coxph` function or automatically calculated when estimating a frailty model. If the null hypothesis of no unobserved heterogeneity is not rejected, it might be preferable to employ simpler Cox-type models.

Marginal and conditional quantities

Several quantities are of interest in the context of frailty models. For a group of individuals with covariate vector $\mathbf{x}_{ij}(t)$ and frailty Z_i, the cumulative intensity (hazard) is defined as

$$\Lambda_{ij}(t|Z_i) = Z_i \int_0^t \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(s)ds. \tag{5.7}$$

The survival function for such individual is given by $S_{ij}(t|Z_i) = \exp\left(-\Lambda_{ij}(t|Z_i)\right)$. These quantities are *conditional* on the random effect Z_i.

The population-level, or *marginal* quantities may be obtained by integrating out the frailty from the conditional ones. The marginal survival is given by

$$S_{ij}(t) = \mathrm{E}_\theta\left[\exp(-\Lambda_{ij}(t|Z_i))\right] = \mathcal{L}_Z\left(\int_0^t \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(s)ds\right). \tag{5.8}$$

The marginal cumulative intensity is then given by $\Lambda_{ij}(t) = -\log S_{ij}(t)$. The "baseline" intensities or survival refer to an individual with $\mathbf{x}_{ij}(t) \equiv 0$.

In the simple case of only one binary covariate, we assume that there are two groups, the baseline with $x = 0$ and "treatment" group with $x = 1$. In this case, the estimated β may be interpreted as the *conditional* intensity ratio (hazard ratio) between two individuals with the same frailty. Under a frailty model, the observed hazard ratio between these two groups is typically attenuated in time (Aalen, Borgan, and Gjessing, 2008, ch. 6).

This *marginal* intensity ratio is calculated as the ratio of the corresponding marginal cumulative intensities $\Lambda_{ij}(t)$.

Several measures of dependence are implemented in **frailtyEM**. The first is the variance of the estimated frailty distribution Z, which is useful for the gamma and the PVF family. The variance of $\log Z$ is also useful for the positive stable distribution for which the variance is infinite. Other measures of association include Kendall's τ and the median concordance. A thorough discussion and comparison of these measures can be found in Hougaard, 2000.

5.2.5 Goodness of fit

Given a large choice of distributions for the frailty, the question comes in selecting the most suitable one. A comparison of the PVF family of frailty distributions can be found in Hougaard (2000, ch. 7.8). In **frailtyEM**, all the frailty distributions depend on a positive parameter θ (see Appendix A1). Given that all the distributions are part of the same family (with gamma and positive stable being limiting cases in the PVF family), the likelihood of different models is comparable across distributions. This argument suggests that it makes sense, within the PVF family, to select the model with the distribution that has the highest likelihood.

An explicit assumption of model (5.1) is that the censoring is non-informative on the frailty. This assumption is usually difficult to test. In **frailtyEM**, a correlation score test is implemented for the gamma distribution, following Balan, Boonk, et al., 2016. This can also be used, for example, for testing whether a recurrent event event process and a terminal event are associated.

Martingale residuals have been used to assess goodness of fit in terms on functional form of the covariates (Therneau, Grambsch, and Fleming, 1990; Lin, Wei, and Ying, 1993). These are provided by the `residuals()` function. For Cox models, there are several methods for assessing the proportional hazards assumption (Therneau and Grambsch, 2000, ch. 6). Graphical methods involve plotting estimated survival or cumulative intensity curves. The plotting capabilities of **frailtyEM** are discussed in Section 5.3.4. A second method is based on Schoenfeld residuals (Grambsch and Therneau, 1994). In R, this is implemented for Cox models in the `cox.zph` function from the **survival** package. In **frailtyEM**, this is provided as part of the output and may be used to test whether the conditional proportional hazards model (5.1) holds. This is detailed in Section 5.3.5.

5.3 Estimation and implementation

5.3.1 Syntax

```
R> library("frailtyEM")
```

The main model fitting function in **frailtyEM** is `emfrail`:

```
R> emfrail(formula, data, distribution, control, ...)
```

The `formula` argument contains a `Surv` object as left hand side and a `+cluster()` statement on the right hand side, specifying the column of `data` that defines the different clusters (this is common to other packages such as **frailtypack**). This formulation, that is common to most survival analysis packages, allows for the representation of clustered failures with left truncation, recurrent events in both calendar time and gap time, time dependent covariates and discontinuous intervals at risk (Therneau and Grambsch, 2000, ch. 3.7, ch. 8). Two other statements may be used in the right hand side: `+strata()` for defining a column with a stratifying variable, and `+terminal()` for defining an event status column for dependent censoring (e.g. a terminal event in the case of recurrent events; this triggers the score test for dependent censoring described Section 5.2.5).

The `distribution` argument determines the frailty distribution. It may be generated by the `emfrail_dist()`:

```
R> str(emfrail_dist(dist = "gamma", theta = 2))
```

```
List of 4
 $ dist           : chr "gamma"
 $ theta          : num 2
 $ pvfm           : num -0.5
 $ left_truncation: logi FALSE
 - attr(*, "class")= chr "emfrail_dist"
```

where `dist` may be one of `"gamma"`, `"stable"` or `"pvf"`. For `"pvf"`, the `m` parameter determines the precise distribution: for $m = -1/2$ for the IG, $m \in (-1, 0)$ for the so-called Hougaard distribution and $m > 0$ a compound Poisson distribution with mass at 0. The `theta` parameter determines the starting value of the optimization. The `left_truncation` argument, if `TRUE`, leads to the calculation described in Section 5.2.3. The `control` argument may be generated by the `emfrail_control()` function and regulates parameters regarding to the estimation.

5.3.2 Profile EM algorithm

In **frailtyEM**, a general full-likelihood estimation procedure is implemented for the gamma, positive stable and PVF frailty models, using a semi-parametric Breslow estimator for the baseline intensity. The goal is to find $\theta, \beta, \lambda_0(\cdot)$ that maximize $L(\theta, \beta, \lambda_0(\cdot))$ (5.3). This can be achieved in two steps, as

$$\max_{\theta,\beta,\lambda_0} L(\theta, \beta, \lambda_0) = \max_{\theta} \left\{ \max_{\beta,\lambda_0} L(\beta, \lambda_0 | \theta) \right\}$$

where $\hat{L}(\theta) = \max_{\beta,\lambda_0} L(\beta, \lambda_0|\theta)$ is the profile likelihood of θ. The profile EM algorithm refers to using a two-stage maximization procedure: the "inner problem" which involves

calculating $\hat{L}(\theta)$ (maximizing $L(\beta, \lambda_0|\theta)$ for fixed θ with the EM algorithm), and the "outer problem", maximizing the profile likelihood $\hat{L}(\theta)$ over θ.

The inner problem Maximizing the likelihood for fixed θ has been proposed for the gamma frailty in Nielsen et al., 1992 and Klein, 1992, and generalizations are discussed in Hougaard, 2000. The crucial observation is that the E step involves calculating the empirical Bayes estimates of the frailties $\hat{z}_i = \mathrm{E}[Z_i|data]$. This expectation is taken with respect to the "posterior" distribution of the random effect. This is detailed in Appendix A2. The M step involves estimating a proportional hazards model with the log $\hat{z}_i$ as offset for each cluster. This is done via the `agreg.fit()` function in the **survival** package, which obtains estimates of β via Cox's partial likelihood. Subsequently, λ_0 and $\tilde{\Lambda}_i$ (and $\tilde{\Lambda}_{L,i}$, in the case of left truncation) are calculated.

The EM algorithm is guaranteed to increase $L(\beta, \lambda_0|\theta)$ with every iteration and to converge to a local maximum. Convergence is achieved when the difference in $L(\beta, \lambda_0|\theta)$ between two consecutive iterations is smaller than ε.

The outer problem The "outer" problem involves maximizing $\hat{L}(\theta)$. For this, a general purpose Newton-type algorithm is employed (`nlm` from the **stats** package).

5.3.3 Standard errors and confidence intervals

The non-parametric information matrix is not directly obtained by the estimation procedure described in Section 5.3.2. From the inner problem, the standard error of the estimates for β and $\lambda_0(\cdot)$ are calculated with Louis' formula (Louis, 1982), under the assumption that θ is fixed to the maximum likelihood estimate. The standard errors obtained in this way are included in the output as `se` and are comparable to the ones from other semi-parametric frailty models (**survival** or **coxme** packages) that assume that θ is fixed. However, this leads to an underestimate of the variability of β and $\lambda_0(\cdot)$.

In **frailtyEM**, adjusted standard errors, presented in the column `adj se`, are calculated by "propagating" the uncertainty from the estimation of θ to $\beta, \lambda_0(\cdot)$. This is described in more detail in Appendix A3.

From the outer problem, standard errors for θ (more precisely, of $\log\theta$, since the maximization takes place on the log-scale for numerical stability) are directly obtained from the numeric Hessian calculated by `nlm`. The delta method, as implemented in the **msm** package (Jackson, 2011), is employed for calculating the standard errors for θ and the measures of dependence that are detailed in Appendix A1.

Two types of confidence intervals for θ (and for the frailty variance, which, in the cases where it exists, is $1/\theta$) are provided. The first are derived from symmetric confidence intervals on the log-scale. The resulting asymmetric confidence interval has been shown to provide good coverage (Balan, Jonker, et al., 2016). The second, more computationally intensive, are referred to as "likelihood-based confidence intervals". Under the null hypothesis, the likelihood ratio test statistic follows a $\chi^2(0) + \chi^2(1)$ distribution.

The critical value associated with this test statistic is approximately 1.92. Based on $\hat{L}(\theta)$, a one-dimensional search is performed to find the confidence interval around the maximum likelihood estimate $\widehat{\theta}$ within which $\log \hat{L}(\theta) \geq \log \hat{L}(\widehat{\theta}) - 1.92$. The advantage of this type of confidence interval is that it is transformation invariant (with the same coverage for all derived dependence measures) and it has a 1-1 correspondence with the likelihood ratio test.

5.3.4 Methods

The `emfrail` function returns an object of class `emfrail` that is documented in `?emfrail`. Usual methods are associated with this class of objects: `print()`, `coef()`, `vcov()`, `residuals()`, `model.matrix()`, `model.frame()`, `logLik()`.

The `summary()` method returns an object of class `emfrail_summary()`, the printing of which contains general fit information, covariate estimates and distribution-specific measures of dependence and goodness of fit, discussed in Section 5.2.5. Arguments to `summary()` may be used to show confidence intervals based on either the likelihood function or the delta method, as described in Section 5.3.3. Other arguments control the amount of information that is printed and may be used when less output is desirable.

The method for prediction of survival curves and cumulative intensity curves is implemented in `predict()`. Both conditional and marginal curves defined in Section 5.2.4 may be produced. The prediction is made for individuals with covariate values specified in a data frame (via the `newdata` argument) or for a fixed linear predictor (via the `lp` argument). For stratified models, the strata must also be specified. By default, the `predict` function creates predictions for each row of `newdata` or for each value of `lp` separately. With the `individual` argument, predicted curves may be produced for individuals with specific at-risk patterns (for example, if an individual is not at risk during a certain time frame), or for individuals with time dependent covariates.

After $\mathbf{x}_{ij}(t)$ is specified to `predict()`, $\Lambda_{ij}(t|Z = 1)$ is calculated as in (5.7) and from this the other quantities are derived, including the conditional survival, the marginal survival (5.8) and the marginal cumulative intensity. Confidence bands are based on the asymptotic normality of the estimated λ_0, and are produced both adjusted and unadjusted for the uncertainty of θ.

5.3.5 Plotting and additional features

Two plot methods are provided based on both **graphics** package via `plot()` and the **ggplot2** package, via `autoplot()`, both with identical syntax. Behind the scenes, they use calls to `predict()`. The `type` argument determines the type of plot:

- `type = "hist"` for a histogram of the posterior estimates of the frailties;
- `type = "pred"` for plotting marginal and conditional cumulative hazard or survival curves;

- `type = "hr"` for plotting marginal or conditional estimated hazard ratios between two groups of individuals. The marginal hazard ratio is determined as the ratio of the marginal intensities, as described in Section 5.2.4;
- `type = "frail"` for a scatter plot of the ordered posterior estimates of the frailties (also called a "caterpillar plot"). For the gamma distribution, quantiles of the posterior distribution are also included. Only available with the `autoplot()` method.

The Commenges-Andersen score test for heterogeneity is by default calculated every time `emfrail` is called and is part of the standard output. A separate function `ca_test()` is also provided, that may be used independently on Cox models produced by `coxph()` from the **survival** package.

While martingale residuals may be obtained with the `residuals()` method, the test for conditional proportional hazards, based on Schoenfeld residuals described in Section 5.2.5 may be accessed in the `$zph` field of the fit. This is an object of class `cox.zph` borrowed from the **survival** package and equivalent to calling `cox.zph` on a Cox model with the estimated log-frailties as offset. The structure and plot methods are described in `?cox.zph`.

An additional function is provided to calculate the marginal log-likelihood for a vector of values of θ, `emfrail_pll()`, without actually performing the outer optimization. This may be useful for visualizing the profile log-likelihood or when debugging (e.g., to see if the maximum likelihood estimate of θ lies on the boundary).

5.4 Illustration

The features of the package will now be illustrated with three well-known data sets available in R: The `CGD` data set (recurrent events, calendar time), the `kidney` data set (recurrent events, gap time) and the `rats` data set (clustered failures).

5.4.1 CGD

The data are from a placebo controlled trial of gamma interferon in chronic granulotomous disease (CGD) and is available in the **survival** package. It contains the time to recurrence of serious infections observed, from randomization until end of study for each patient (i.e. the time scale is calendar time). For the purpose of illustration, we will use `treat` (treatment or placebo) and `sex` (female or male) as covariates, although a larger number of variables are recorded in the data set.

```
R> data("cgd")
R> cgd <- cgd[c("tstart", "tstop", "status", "id", "sex", "treat")]
R> head(cgd)
```

```
  tstart tstop status id    sex   treat
1      0   219      1  1 female  rIFN-g
2    219   373      1  1 female  rIFN-g
3    373   414      0  1 female  rIFN-g
4      0     8      1  2   male placebo
5      8    26      1  2   male placebo
6     26   152      1  2   male placebo
```

A basic gamma frailty model can be fitted like this:

```
R> gam <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd)
R> summary(gam)

Call:
emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    cluster(id), data = cgd)

Regression coefficients:
              coef exp(coef) se(coef) adj. se      z    p
sexfemale   -0.227     0.797    0.396   0.396 -0.575 0.57
treatrIFN-g -1.052     0.349    0.310   0.310 -3.389 0.00
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00172
no-frailty Log-likelihood: -331.997
Log-likelihood: -326.619
LRT: 1/2 * pchisq(10.8), p-val 0.00052

Frailty summary:
frailty variance = 0.821 / 95% CI: [0.231, 1.854]
Kendall's tau: 0.291 / 95% CI: [0.104, 0.481]
Median concordance: 0.289 / 95% CI: [0.101, 0.491]
E[log Z]: -0.464 / 95% CI: [-1.164, -0.12]
Var[log Z]: 1.241 / 95% CI: [0.26, 4.341]
theta = 1.218 (0.59) / 95% CI: [0.539, 4.326]
Confidence intervals based on the likelihood function
```

The first two parts of this output, about regression coefficients and fit summary, exist regardless of the frailty distributions. The last part, "frailty summary", provides a different output according to the distribution.

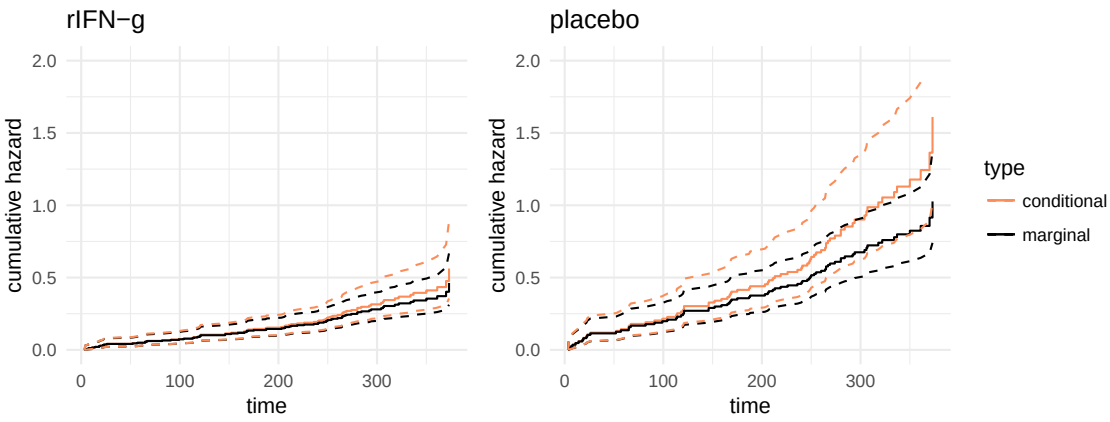

Figure 51: Predicted conditional and marginal cumulative hazards for males, one from the treatment arm and one from the placebo arm, as produced by `autplot()` with `type = "pred"`.

Both the Commenges-Andersen test for heterogeneity and the one-sided likelihood ratio test deems the random effect highly significant. This is also suggested by the confidence interval for the frailty variance, which does not contain 0.

To illustrate the predicted cumulative hazard curves we take two individuals, one from the treatment arm and one from the placebo arm, both males:

```
R> library("ggplot2")
R> p1 <- autoplot(gam, type = "pred",
+   newdata = data.frame(sex = "male", treat = "rIFN-g")) +
+   ggtitle("rIFN-g") + ylim(c(0, 2)) + theme_minimal()
R> p2 <- autoplot(gam, type = "pred",
+   newdata = data.frame(sex = "male", treat = "placebo")) +
+   ggtitle("placebo") + ylim(c(0, 2)) + theme_minimal()
```

The two plots are shown in Figure 51.

The cumulative hazard in this case can be interpreted as the expected number of events at a certain time. It can be seen that the frailty "drags down" the marginal hazard. This is a well-known effect observed in frailty models, as described in Aalen, Borgan, and Gjessing (2008, ch. 7). All prediction results could also be obtained directly:

```
R> dat_pred <- data.frame(sex = c("male", "male"),
+   treat = c("rIFN-g", "placebo"))
R> predict(gam, dat_pred)
```

For a hypothetical individual that changes treatment from placebo to rIFN-g at time 200, predictions may also be obtained:

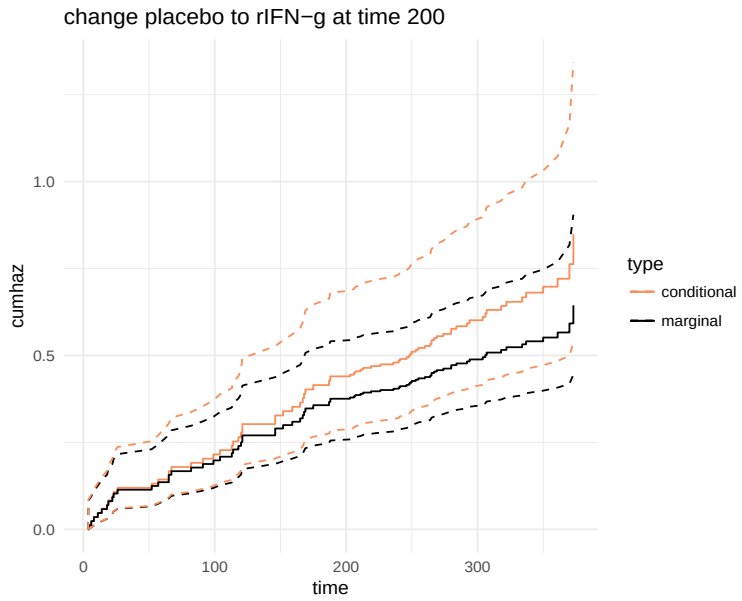

Figure 52: Predicted conditional and marginal cumulative hazards for a male that switches treatment from placebo to rIFN-g at time 200 as produced by `autoplot()` with `type = "pred"`

```
R> dat_pred_b <- data.frame(sex = c("male", "male"),
+   treat = c("placebo", "rIFN-g"),
+   tstart = c(0, 200), tstop = c(200, Inf))
R> p <- autoplot(gam, type = "pred",
+                newdata = dat_pred_b,
+                individual = TRUE) +
+   ggtitle("change placebo to rIFN-g at time 200") + theme_minimal()
```

This plot is shown in Figure 52.

A positive stable frailty model can also be fitted by specifying the `distribution` argument.

```
R> stab <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd,
+   distribution = emfrail_dist(dist = "stable"))
R> summary(stab)

Call:
emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
```

```
    cluster(id), data = cgd, distribution = emfrail_dist(dist = "stable"))

Regression coefficients:
              coef exp(coef) se(coef) adj. se      z    p
sexfemale   -0.137     0.872    0.407   0.407 -0.337 0.74
treatrIFN-g -1.085     0.338    0.332   0.336 -3.230 0.00
Estimated distribution: stable / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00172
no-frailty Log-likelihood: -331.997
Log-likelihood: -329.39
LRT: 1/2 * pchisq(5.21), p-val 0.0112

Frailty summary:
Kendall's tau: 0.104 / 95% CI: [0.011, 0.236]
Median concordance: 0.102 / 95% CI: [0.011, 0.233]
E[log Z]: 0.067 / 95% CI: [0.006, 0.179]
Var[log Z]: 0.406 / 95% CI: [0.037, 1.176]
Attenuation factor: 0.896 / 95% CI: [0.764, 0.989]
theta = 8.572 (5.41) / 95% CI: [3.232, 90.316]
Confidence intervals based on the likelihood function
```

The coefficient estimates are similar to those of the gamma frailty fit. The "Frailty summary" part is quite different. For the positive stable distribution, the variance is not defined. However, Kendall's τ is easily obtained, and in this case it is smaller than in the gamma frailty model. Unlike the gamma or PVF distributions, the positive stable frailty predicts a marginal model with proportional hazards where the marginal hazard ratios are an attenuated version of the conditional hazard ratios shown in the output. The calculations are detailed in Appendix A1.

The conditional and marginal hazard ratios from different distributions can also be visualized easily. We also fitted an IG frailty model on the same data, and plots of the hazard ratio between two males from different treatment arms created below are shown in Figure 53.

```
R> ig <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd,
+   distribution = emfrail_dist(dist = "pvf"))
R> newdata <- data.frame(treat = c("placebo", "rIFN-g"),
+   sex = c("male", "male"))
R> pl1 <- autoplot(gam, type = "hr", newdata = newdata) +
```

```
+   ggtitle("gamma") + theme_minimal()
R> pl2 <- autoplot(stab, type = "hr", newdata = newdata) +
+   ggtitle("PS") + theme_minimal()
R> pl3 <- autoplot(ig, type = "hr", newdata = newdata) +
+   ggtitle("IG") + theme_minimal()
```

While all models shrink the hazard ratio towards 1, it can be seen that this effect is slightly more pronounced for the gamma than for the IG, while the PS exhibits a constant "average" shrinkage. This type of behaviour from the PS is often seen as a strength of the model (Hougaard, 2000).

5.4.2 Kidney

The `kidney` data set is also available in the **survival** package. The data, presented originally in McGilchrist and Aisbett, 1991, contains the time to infection for kidney patients using a portable dialysis equipment. The infection may occur at the insertion of the catheter and at that point, the catheter must be removed, the infection cleared up, and the catheter reinserted. Each of the 38 patients has exactly 2 observations, representing recurrence times from insertion until the next infection (i.e. the time scale is gap time). There are 3 covariates: sex, age and disease (a factor with 4 levels). A data analysis based on frailty models is described in Therneau and Grambsch (2000, ch. 9.5.2). For the purpose of illustration, we do not include the `disease` variable here.

```
R> data("kidney")
R> kidney <- kidney[c("time", "status", "id", "age", "sex" )]
R> kidney$sex <- ifelse(kidney$sex == 1, "male", "female")
R> head(kidney)
```

```
  time status id age    sex
1    8      1  1  28   male
2   16      1  1  28   male
3   23      1  2  48 female
4   13      0  2  48 female
5   22      1  3  32   male
6   28      1  3  32   male
```

```
R> zph_t = emfrail_control(zph = TRUE)
R> m_gam <- emfrail(Surv(time, status) ~ age + sex + cluster(id),
+   data = kidney, control = zph_t)
R> m_ps <- emfrail(Surv(time, status) ~ age + sex + cluster(id),
+   data = kidney,
+   distribution = emfrail_dist("stable"),
+   control = zph_t)
```

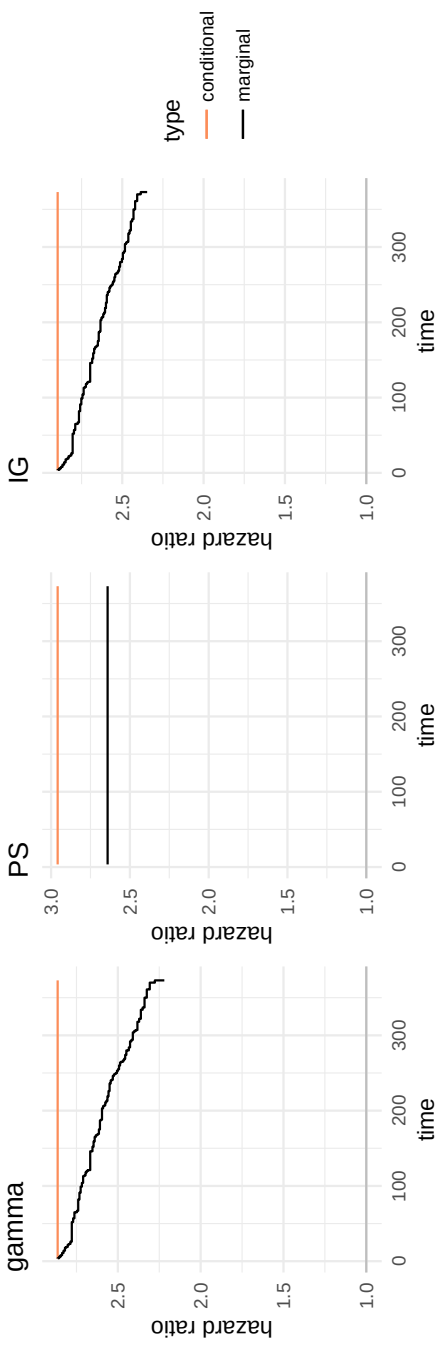

Figure 53: Conditional and marginal hazard ratio between two males from the placebo and rIFN-g treatment arms from the gamma, PS and IG frailty models as produced by `autoplot()` with `type = "hr"`.

Therneau and Grambsch discuss the gamma fit conclude that an outlier case is at the source of the frailty effect. We omit the frailty part of the output; the estimated frailty variance is 0.397 with a 95% likelihood based confidence interval of (0.04, 1.03) and therefore significantly different from 0.

```
R> summary(m_gam, print_opts = list(frailty = FALSE))

Call:
emfrail(formula = Surv(time, status) ~ age + sex + cluster(id),
    data = kidney, control = zph_t)

Regression coefficients:
           coef exp(coef) se(coef) adj. se       z    p
age     0.00544   1.00545  0.01158 0.01170 0.46481 0.64
sexmale 1.55284   4.72487  0.44518 0.49952 3.10868 0.00
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00238
no-frailty Log-likelihood: -184.657
Log-likelihood: -182.053
LRT: 1/2 * pchisq(5.21), p-val 0.0112
```

However, the LRT is not significant for the positive stable frailty model (which does not have a defined frailty variance, for comparison). Furthermore, the estimated regression coefficients are different.

```
R> summary(m_ps, print_opts = list(frailty = FALSE))

Call:
emfrail(formula = Surv(time, status) ~ age + sex + cluster(id),
    data = kidney, distribution = emfrail_dist("stable"), control = zph_t)

Regression coefficients:
           coef exp(coef) se(coef)       z    p
age     0.00218   1.00218  0.00922 0.23649 0.81
sexmale 0.82100   2.27278  0.29873 2.74831 0.01
Estimated distribution: stable / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00238
```

```
no-frailty Log-likelihood: -184.657
Log-likelihood: -184.657
LRT: 1/2 * pchisq(0), p-val>0.5
```

The test for proportional hazards described in Section 5.2.5 reveals an insight into how the two models work. The gamma frailty model specifies conditional proportional hazards and marginal non-proportional hazards, while the positive stable model specifies proportional hazards at both levels.

```
R> m_gam$zph
```

```
            rho  chisq     p
age      0.0368 0.0764 0.782
sexmale -0.2207 2.4923 0.114
GLOBAL       NA 2.5445 0.280
```

```
R> m_ps$zph
```

```
            rho  chisq        p
age      0.0841  0.477 0.489990
sexmale -0.4364 11.392 0.000738
GLOBAL       NA 11.480 0.003215
```

Therefore, the gamma frailty model appears to explain the marginal non-proportionality, while the positive stable frailty model does not. Such a phenomenon may be observed if, for example, the PS marginal model is a bad fit for the data. Further research is being carried out on this topic (Balan and Putter, Forthcoming).

5.4.3 Rats data

These is an example of clustered failure data from Mantel, Bohidar, and Ciminera, 1977 Three rats were chosen from each of 100 litters, one of which was treated with a drug (`rx = 1`) and the rest with placebo (`rx = 0`), and then all followed for tumor incidence. The data are also available in the **survival** package.

```
R> data("rats")
R> head(rats)
```

```
  litter rx time status sex
1      1  1  101      0   f
2      1  0   49      1   f
3      1  0  104      0   f
4      2  1   91      0   m
5      2  0  104      0   m
6      2  0  102      0   m
```

While often used to illustrate frailty models, the gamma frailty fit shows a relatively large, yet not significant frailty variance

```
R> summary(emfrail(Surv(time, status) ~ rx + sex + cluster(litter),
+                  data = rats))

Call:
emfrail(formula = Surv(time, status) ~ rx + sex + cluster(litter),
    data = rats)

Regression coefficients:
        coef exp(coef) se(coef) adj. se       z    p
rx    0.7873    2.1974   0.3135  0.3135  2.5112 0.01
sexm -3.1341    0.0435   0.7385  0.7409 -4.2298 0.00
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.201
no-frailty Log-likelihood: -200.426
Log-likelihood: -199.73
LRT: 1/2 * pchisq(1.39), p-val 0.119

Frailty summary:
frailty variance = 0.445 / 95% CI: [0, 1.678]
Kendall's tau: 0.182 / 95% CI: [0, 0.456]
Median concordance: 0.179 / 95% CI: [0, 0.464]
E[log Z]: -0.239 / 95% CI: [-1.038, 0]
Var[log Z]: 0.559 / 95% CI: [0, 3.678]
theta = 2.245 (2.28) / 95% CI: [0.596, Inf]
Confidence intervals based on the likelihood function
```

The `Surv` object takes two arguments here: time of event and status. This implicitly assumes that each row of the data (in this case, each rat) is under follow-up from time 0 to `time`. This is very similar to the representation of the recurrent events in gap-time, where each recurrent event episode is "at risk" from time 0 (time since the previous event).

We artificially simulated left truncation from an exponential distribution with mean 50, which is now an entry time to the study. The rats with a follow-up smaller than the entry time are removed.

```
R> set.seed(1)
R> rats$tstart <- rexp(nrow(rats), rate = 1/50)
R> rats_lt <- rats[rats$tstart < rats$time, ]
```

The first model, m1, is what happens if left truncation is completely ignored. Each rat is assumed to have been at risk from time 0, which is not the case.

```
R> m1 <-
+   emfrail(Surv(time, status) ~ rx  + cluster(litter),
+           data = rats_lt)
```

The second model, m2, is what happens when the at-risk indicator is correctly adjusted for, with the entry time also present. Refering back to Section5.2.3, this is equivalent to considering $P(Z)$ instead of $P(Z|A)$.

```
R> m2 <-
+   emfrail(Surv(tstart, time, status) ~ rx + sex + cluster(litter),
+           data = rats_lt)
```

As may be seen from equation (5.6), this is correct only if there is in fact no left truncation, or if there is no variability in Z (i.e. Z is degenerate at 1). Therefore, this formulation is correct, for example, when the Surv object represents recurrent events in calendar time, as is the case in Section 5.4.1. This is, for example, what is returned by the frailty models in the **survival** package.

The third model, m3, specifies the correct time at risk but also the fact that the distribution of the frailty must be taken conditional on the entry time. Under this (artificial) left truncation problem, this would be the correct way of analyzing this data.

```
R> m3 <-
+   emfrail(Surv(tstart, time, status) ~ rx + sex + cluster(litter),
+           data = rats_lt,
+           distribution = emfrail_dist(left_truncation = TRUE))
```

In this case, the output shows little difference between models. This is because the frailty, even in the complete data set, is not significant. In this case, the frailty distribution is also not significant in either m2 or m3 and they lead to estimates very close to each other. In a limited unpublished simulation study, we have seen that applying the correction in m3 leads to approximately unbiased estimates of the regression coefficients, unlike m1 or m2.

5.5 Conclusion

In the current landscape for modeling random effects in survival analysis, **frailtyEM** is a contribution that focuses on implementing classical methodology in an efficient way with a wide variety of frailty distributions. We have shown that the EM based approach has certain advantages in the context of frailty models. First of all, it is semiparametric, which means that it is a direct extension of the Cox proportional hazards model. In this way, classical results from semiparametric frailty models (for example, based on the

data sets in Section 5.4) can be replicated and further insight may be obtained by fitting models with different frailty distributions. Until now, the Commenges-Andersen test, positive stable and PVF family, have not all been implemented in a consistent way in an R package. Another advantage of the EM algorithm is that, by its nature, it is a full maximum likelihood approach, and the estimators have well known desirable asymptotic properties.

To our knowledge, no other statistical package provides similar capabilities for visualizing conditional and marginal survival curves, or the marginal effect of covariates. Since this is implemented across a large number of distributions, this might come to the aid of both applied and theoretical research into shared frailty models. While the question of model selection with different random effect distributions is still an open one, the functions included **frailtyEM** may be useful for further research in this direction.

Evaluating goodness of fit for shared frailty models is still a complicated issue, particularly in semiparametric models. However, tests based on martingale residuals, such as that of Commenges and Rondeau, 2000, should be now possible by extrating the necessary quantities from an `emfrail` fit.

Regarding the left truncation implementation in **frailtyEM**, it is very similar to that from the **parfm** package. However, performing of a larger simulation study to assess the effects of left truncation in clustered failure data with semiparametric frailty models is now possible. In a limited simulation study we have seen that correctly accounting for this phenomenon leads to unbiased estimates. The scenario of time dependent covariates and left truncation is not supported at this time. This is because this would require also specifying values of these covariates from time 0 to the left truncation time, which would likely involve some speculation.

Technically, extending the package to other distributions is possible, as long as their Laplace transform and the corresponding derivatives may be specified in closed form. An interesting extension would be to choose discrete distributions from the infinitely divisible family for the random effect, such as the Poisson distribution. The newest features will be implemented in the development version of the package at https://github.com/tbalan/frailtyEM.

Appendix A1: Results for the Laplace transforms

We consider distributions from the infinitely divisible family Ash, 1972, ch 8.5 with the Laplace transform

$$\mathcal{L}_Y(c) = \exp(-\alpha\psi(c;\gamma)).$$

We now consider how α and γ can be represented as a function of a positive parameter θ.

The gamma distribution For Y a gamma distributed random variable, $\psi(c;\gamma) = \log(\gamma + c) - \log(\gamma)$, the derivatives of which are

$$\psi^{(k)}(c;\gamma) = (-1)^{k-1}(k-1)!(\gamma + c)^{-k}.$$

For identifiability, the restriction $\mathrm{E}Y = 1$ is imposed; this leads to $\alpha = \gamma$. The distribution is parametrized with $\theta > 0$, $\theta = \alpha = \gamma$. The variance of Y is $\operatorname{var} Y = \theta^{-1}$. Kendall's τ is then $\tau = \frac{1}{1+2\theta}$ and the median concordance is $\kappa = 4\left(2^{1+1/\theta} - 1\right)^{-\theta} - 1$. Furthermore, $\mathrm{E}\log Y = \psi(\theta) - \log\theta$ and $\operatorname{var}\log Y = \psi'(\theta)$ where ψ and ψ' are the digamma and trigamma functions.

The positive stable distribution For Y a positive stable random variable, $\psi(c;\gamma) = c^{\gamma}$ with $\gamma \in (0,1)$, the derivatives of which are

$$\psi^{(k)}(c;\gamma) = \frac{\Gamma(k-\beta)}{\Gamma(1-\gamma)}(-1)^{k-1}c^{\gamma-1}.$$

For identifiability, the restriction $\alpha = 1$ is made; $\mathrm{E}Y$ is undefined and $\operatorname{var} Y = \infty$. The distribution is parametrized with $\theta > 0$, $\gamma = \frac{\theta}{\theta+1}$.

Kendall's τ is then $\tau = 1 - \frac{\theta}{\theta+1}$ and the median concordance is $\kappa = 2^{2-2\frac{\theta}{\theta+1}} - 1$. Furthermore,

$$\mathrm{E}\log Y = -\left(\left\{\frac{\theta}{1+\theta}\right\}^{-1} - 1\right)\psi(1)$$

and

$$\operatorname{var}\log Y = \left(\left\{\frac{\theta}{1+\theta}\right\}^{-2} - 1\right)\psi'(1).$$

In the case of the PS distribution, the marginal hazard ratio is an attenuated version of the conditional hazard ratio. If the conditional log-hazard ratio is β, the marginal hazard ratio is equal to $\beta\frac{\theta}{\theta+1}$.

The PVF distributions For Y a PVF distribution with fixed parameter $m \in \mathbb{R}$, $m > -1$ and $m \neq 0$,

$$\psi(c;\gamma) = \operatorname{sign}(m)(1 - \gamma^m(\gamma + c)^{-m})$$

where $\operatorname{sign}(\cdot)$ denotes the sign. This is the same parametrizaion as in Aalen, Borgan, and Gjessing, 2008. The derivatives of ψ are

$$\psi^{(k)}(c;\gamma) = \operatorname{sign}(m)(-\gamma)^m(\gamma + c)^{-m-k}(-1)^{k+1}\frac{\Gamma(m+k)}{\Gamma(m)}.$$

The expectation of this distribution can be calculated as minus the first derivative of the Laplace transform calculated in 0, i.e.,

$$\mathrm{E}Y = \alpha\psi'(0;\gamma)\mathcal{L}(0;\alpha,\gamma) = \frac{\alpha}{\gamma}m.$$

The second moment of the distribution can be calculated as the second derivative of the Laplace transform at 0,

$$\mathrm{E}Y^2 = \alpha^2\psi'^2(0) - \alpha\psi''(0) = \frac{\alpha^2}{\gamma^2}m^2 + \frac{\alpha}{\gamma^2}m(m+1).$$

For identifiability, we set $\mathrm{E}Y = 1$. The distribution is parametrized through a parameter $\theta > 0$ which is determined by $\gamma = (m+1)\theta$ and $\alpha = \mathrm{sign}(m)\frac{m+1}{m}\theta$. This results in $\mathrm{var}Y = \theta^{-1}$.

A slightly different parametrization is presented in Hougaard, 2000, dependent on the parameter η_H. The correspondence is obtained by setting $\eta_H = (m+1)\theta$.

The PVF family of distributions includes the gamma as a limiting case when $m \longrightarrow 0$. When $\gamma \longrightarrow 0$ the positive stable distribution is obtained. When $m = -1$ the distribution is degenerate, and with $m = 1$ a non-central gamma distribution is obtained. Of special interest is the case $m = -0.5$, when the inverse Gaussian distribution is obtained. With $m > 0$, the distribution is compound Poisson with mass at 0. In this case, $P(Y = 0) = \exp(-\frac{m+1}{m}\theta)$.

For $m < 0$, closed forms for Kendall's τ and median concordance are given in Hougaard (2000, Section 7.5).

Left truncation

To determine the Laplace transform under left truncation, we determine $\tilde{\psi}$ from (5.4) and (5.5).

For the gamma distribution, we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = \log(\gamma + \Lambda_L + c) - \log(\gamma + \Lambda_L)$$

which implies that the frailty of the survivors is still gamma distributed, but with a change in the parameter γ.

For the positive stable we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = (c + \Lambda_L)^\gamma - \Lambda_L^\gamma,$$

which is not a positive stable distribution any more.

For the PVF distributions, we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = \mathrm{sign}(m)\left(\gamma^m(\gamma + \Lambda_L)^{-m} - (\gamma + \Lambda_L)^m(\gamma + \Lambda_L + c)^{-m}\right),$$

which is not PVF any more (however, it stays in the same "infinitely divisible" family.

Closed forms

The gamma distribution leads to a Laplace transform for which the derivatives can be calculated in closed form. It can be seen that

$$\mathcal{L}(c;\alpha,\gamma) = \gamma^\alpha(\gamma + c)^{-\alpha}.$$

The k-th derivative of this expression is

$$\mathcal{L}^{(k)}(c;\alpha,\gamma) = \gamma^{\alpha}(\gamma + c)^{-\gamma-k}\frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}.$$

This can be exploited also in the case of left truncation, since the gamma frailty is preserved, as shown in the previous section.

The inverse gaussian distribution is obtained when the PVF parameter is $m = -\frac{1}{2}$. Under the current parametrization, we have $\beta = \theta/2$ and $\alpha = \theta$. In this case, the Laplace transform is

$$\mathcal{L}(c;\theta) = \exp\left\{\theta\left(1 - \sqrt{1 + 2c/\theta}\right)\right\}.$$

The k-th derivative of this can be written as

$$\mathcal{L}^{(k)}(c;\theta) = (-1)^k\left(\frac{2}{\theta}c + 1\right)^{-k/2}\frac{\mathcal{K}_{k-1/2}\left(\sqrt{2\theta\left(c + \frac{\theta}{2}\right)}\right)}{\mathcal{K}_{1/2}\left(\sqrt{2\theta\left(c + \frac{\theta}{2}\right)}\right)}$$

where $\mathcal{K}$ is the modified Bessel function of the second kind.

The `emfrail()` uses the closed form formulas when possible, by default.

Appendix A2: The E step

For the E step β and λ_0 are fixed, either at their initial values or at the values from the previous M step. Let $n_i = \sum_{j,k}\delta_{ijk}$ be the number of events in cluster i. The conditional distribution of Z_i given the data is described by the Laplace transform

$$\mathcal{L}(c) = \frac{\mathrm{E}\left[Z_i^{n_i}\exp(-Z_i\tilde{\Lambda}_i)\exp(-Z_i c)\right]}{\mathrm{E}\left[Z_i^{n_i}\exp(-Z_i\tilde{\Lambda}_i)\right]} = \frac{\mathcal{L}^{(n_i)}(c + \tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{5.9}$$

The E step reduces to calculating the expectation of this distribution, i.e. the derivative of (5.9) in 0:

$$\hat{z}_i = -\frac{\mathcal{L}^{(n_i+1)}(\tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{5.10}$$

The marginal (log-)likelihood is also calculated at this point to keep track of convergence of the EM algorithm. It can be seen that (5.3) involves the denominator of (5.9) in addition to a straight-forward expression of β and λ_0.

The E step is generally the expensive operation of the EM algorithm. In a few scenarios (5.10) may be expressed in a closed form: for the gamma and the inverse gaussian distributions. In these scenarios, the E step is calculated with the `fast_estep()` routine. For all other cases, the E step is calculated via a recursive algorithm with an internal routine which is described here. For easing the computational burden, this is implemented in C++ and is interfaced with R via the **Rcpp** library (Eddelbuettel and François, 2011; Eddelbuettel, 2013).

As shown in (5.9), the calculation of the E step for the general case involves taking derivatives of Laplace transforms of the form

$$\mathcal{L}(c) = \exp(g(c))$$

where for simplicity we denote $g(c) = -\alpha\psi(c;\gamma)$. The expression for the k-th derivative of $\mathcal{L}(c)$ can be obtained with a classical calculus result, di Bruno's formula, i.e.,

$$\mathcal{L}^{(n)}(c) = \sum_{\mathbf{m}\in\mathcal{M}_n} \frac{n!}{m_1!m_2!\ldots m_n!} \prod_{j=1}^{n} \left(\frac{g^{(j)}(c)}{j!}\right)^{m_j} \mathcal{L}(c), \tag{5.11}$$

where $\mathcal{M}_n = \{(m_1,\ldots,m_n) | \sum_{j=1}^{n} j \times m_j = n\}$. For example, for $n = 3$,

$$\mathcal{M}_3 = \{(3,0,0),\ (1,1,0),\ (0,0,1)\}\,.$$

This corresponds to the "partitions of the integer" 3, i.e., all the integers that sum up to 3:

$$\{(1,1,1),\ (1,2,0), (3,0,0)\}\,.$$

We implemented a recursive algorithm in C++ which resides in the `emfrail_estep.cpp` which loops through these partitions, calculates the corresponding derivatives of ψ and the coefficients.

Appendix A3: Standard errors

Considering the vector of parameters $\eta = (\beta, \lambda_0(\cdot))$, and consider that for a given θ, η_θ is the maximizer of the "inner problem" described in Section (5.3.2), i.e. $\eta(\theta) = \text{argmax}_\eta L(\eta|\theta)$. Further, for a given θ, the variance-covariance matrix $\text{var}(\eta(\theta))$ is obtained with Louis' formula (Louis, 1982). The restulting standard errors for η are underestimated because they do not factor in the uncertainty in estimating θ, as is noted also in Therneau and Grambsch (2000, sec. 9.5). Below is the sketch of how this is addressed in **frailtyEM**, following Hougaard (2000, Appendix B.3).

Let $\hat{\theta}$ be the maximum likelihood estimate with variance $\text{var}(\hat{\theta})$ and standard error $\text{se}(\hat{\theta})$, which are given by the maximizer from the "outer problem". The correct information matrix for inference on η is a "perturbed" version of $\text{var}(\eta(\hat{\theta}))$, namely

$$\text{var}(\eta(\hat{\theta})) + \left(\frac{d\eta}{d\theta}\right) \text{var}(\hat{\theta}) \left(\frac{d\eta}{d\theta}\right)^{\top}.$$

Here, $d\eta/d\theta$ may be approximated as $(\eta^+ - \eta^-)/\text{se}(\hat{\theta})$ where $\eta^+ = \eta(\hat{\theta} + \text{se}(\hat{\theta})/2)$ and $\eta^- = \eta(\hat{\theta} - \text{se}(\hat{\theta})/2)$. In `emfrail`, this whole calculation takes place for $\log\theta$ for computational stability, and to avoid the edge problem when θ is close to 0.

Confidence intervals for the conditional cumulative hazard are obtained from the part of the variance-covariance matrix corresponding to $\lambda_0(\cdot)$, and confidence intervals for $\Lambda_0(t) = \sum_{s\le t} \lambda_0(t)$ are obtained with the usual formula. For confidence intervals, the delta method is used to calculate a symmetric confidence interval for $\log\Lambda_0(t)$ for all t, which is then exponentiated.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* 7 (11), pp. 1121–1137.

— (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research* 3 (3), pp. 227–243.

Aalen, O. O., R. J. Cook, and K. Røysland (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* 21 (4), pp. 579–593.

Aalen, O. O. and H. K. Gjessing (2004). Survival models based on the Ornstein-Uhlenbeck process. *Lifetime Data Analysis* 10 (4), pp. 407–423.

Aalen, O. O., M. Valberg, T. Grotmol, and S. Tretli (2014). Understanding variation in disease risk: the elusive concept of frailty. *International Journal of Epidemiology* 44 (4), pp. 1408–1421.

Aalen, O., O. Borgan, and H. Gjessing (2008). *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag New York. DOI: 10.1007/978-0-387-68560-1.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Bodels Based on Counting Processes*. Springer Science & Business Media.

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pp. 1100–1120.

Andersen, P. K. and N. Keiding (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine* 31 (11-12), pp. 1074–1088.

Andersen, P. K., J. P. Klein, K. M. Knudsen, and R. T. y Palacios (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, pp. 1475–1484.

Anderson, J. E., T. A. Louis, N. V. Holm, and B. Harvald (1992). Time-dependent association measures for bivariate survival distributions. *Journal of the American Statistical Association* 87 (419), pp. 641–650.

Ash, R. P. (1972). *Real Analysis and Probability*. Academic press.

Asmussen, S., J. L. Jensen, and L. Rojas-Nandayapa (2016). On the Laplace transform of the lognormal distribution. *Methodology and Computing in Applied Probability* 18 (2), pp. 441–458.

Balan, T. A., M. A. Jonker, P. C. Johannesma, and H. Putter (2016). Ascertainment Correction in Frailty Models for Recurrent Events Data. *Statistics in Medicine* 35 (23), pp. 4183–4201. DOI: 10.1002/sim.6968.

Balan, T. A. (2017). *dynfrail: Fitting Dynamic Frailty Models with the EM Algorithm*. R package version 0.5.2.

Balan, T. A. and H. Putter (Forthcoming). *Non-proportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference?*

— (2017). ***frailtyEM**: Fitting Frailty Models with the EM Algorithm*. R package version 0.8.2.

Balan, T.-A., S. E. Boonk, M. H. Vermeer, and H. Putter (2016). Score Test for Association Between Recurrent Events and a Terminal Event. *Statistics in Medicine* 35 (18), pp. 3037–3048. DOI: 10.1002/sim.6913.

Baumann, M. H. and M. Noppen (2004). Pneumothorax. *Respirology* 9 (2), pp. 157–164.

Boonk, S., H. Putter, L. Koolhof, R. Willemze, and M. Vermeer (2014). Quantitation of tumour development correlates with prognosis in tumour stage (stage IIB) mycosis fungoides. *British Journal of Dermatology* 170 (5), pp. 1080–1086.

Breslow, N. E. (1972). Contribution to discussion of papeer by DR Cox. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, pp. 216–217.

Claeskens, G., R. Nguti, and P. Janssen (2008). One-sided tests in shared frailty models. *Test* 17 (1), pp. 69–82.

Commenges, D. and P. K. Andersen (1995). Score Test of Homogeneity for Survival Data. *Lifetime Data Analysis* 1 (2), pp. 145–156. DOI: 10.1007/BF00985764.

Commenges, D. and V. Rondeau (2000). Standardized Martingale Residuals Applied to Grouped Left Truncated Observations of Dementia Cases. *Lifetime Data Analysis* 6 (3), pp. 229–235.

Cook, R. J. and J. F. Lawless (1997). Marginal Analysis of Recurrent Events and a Terminating Event. *Statistics in Medicine* 16, pp. 911–924.

Cook, R. J. and J. Lawless (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, pp. 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* 62 (2), pp. 269–276.

Dahabreh, I. J. and D. M. Kent (2011). Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA* 305 (8), pp. 822–823.

Dai, J., R. E. Krasnow, L. Liu, S. G. Sawada, and T. Reed (2013). The association between postload plasma glucose levels and 38-year mortality risk of coronary heart disease: the prospective NHLBI Twin Study. *PloS one* 8 (7), e69332.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, pp. 1–38.

Do Ha, I., M. Noh, and Y. Lee (2012). **frailtyHL**: A Package for Fitting Frailty Models with h-likelihood. *R Journal* 4 (2), pp. 28–36.

Donohue, M. C., R. Overholser, R. Xu, and V. Florin (2011). Conditional Akaike Information under Generalized Linear and Proportional Hazards Mixed Models. *Biometrika* (98, 3), pp. 685–700. DOI: 10.1093/biomet/asr023.

Donohue, M. C. and R. Xu (2013). ***phmm**: Proportional Hazards Mixed-effects Models*. R package version 0.7-5.

Doorn, R. van, E. Scheffer, and R. Willemze (2002). Follicular mycosis fungoides, a distinct disease entity with or without associated follicular mucinosis: a clinicopathologic and follow-up study of 51 patients. *Archives of Dermatology* 138 (2), pp. 191–198.

Duchateau, L. and P. Janssen (2007). *The Frailty Model*. Springer.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. ISBN 978-1-4614-6867-7. Springer-Verlag New York. DOI: 10.1007/978-1-4614-6868-4.

Eddelbuettel, D. and R. François (2011). **Rcpp**: Seamless R and C++ Integration. *Journal of Statistical Software* 40 (8), pp. 1–18. DOI: 10.18637/jss.v040.i08.

Elbers, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies* 49 (3), pp. 403–409.

Erikson, F., T. Martinussen, and T. H. Scheike (2015). Clustered Survival Data with Left-truncation. *Scandinavian Journal of Statistics*. DOI: 10.1111/sjos.12157.

Fiocco, M., H. Putter, and J. Van Houwelingen (2008). A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics* 10 (2), pp. 245–257.

Gerster, M., M. Madsen, and P. K. Andersen (2014). Matched survival data in a co-twin control design. *Lifetime Data Analysis* 20 (1), pp. 38–50.

Gharibvand, L. and L. Liu (2009). Analysis of survival data with clustered events. *SAS Global Forum* 2009 (237), pp. 1–11.

Gjessing, H. K., O. O. Aalen, and N. L. Hjort (2003). Frailty models based on Lévy processes. *Advances in Applied Probability* 35 (2), pp. 532–550.

Gorfine, M., D. M. Zucker, and L. Hsu (2006). Prospective Survival Analysis with a General Semiparametric Shared Frailty Model: A Pseudo Full Likelihood Approach. *Biometrika*, pp. 735–741.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81 (3), pp. 515–526.

Ha, I. D., J.-H. Jeong, and Y. Lee (2017). *Statistical Modelling of Survival Data with Random Effects*. Springer.

Ha, I. D., Y. Lee, and J.-k. Song (2001). Hierarchical likelihood approach for frailty models. *Biometrika* 88 (1), pp. 233–233.

Heckman, J. and B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pp. 271–320.

Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004). A structural approach to selection bias. *Epidemiology* 15 (5), pp. 615–625.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika* 71 (1), pp. 75–83.

— (1986a). A class of multivanate failure time distributions. *Biometrika* 73 (3), pp. 671–678.

— (1986b). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73 (2), pp. 387–396.

— (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York. DOI: 10.1007/978-1-4612-1304-8.

Huang, X. and R. A. Wolfe (2002). A Frailty Model for Informative Censoring. *Biometrics* 58, pp. 510–520.

Huang, X., R. A. Wolfe, and C. Hu (2004). A test for informative censoring in clustered survival data. *Statistics in Medicine* 23, pp. 2089–2107.

IBM Corp (2016). *IBM SPSS Statistics for Windows, Version 24.0*. IBM Corp. Armonk, NY.

Inc., S. I. (2003). *SAS/STAT Software, Version 9.4*. Cary, NC.

Jackson, C. H. (2011). Multi-State Models for Panel Data: The **msm** Package for R. *Journal of Statistical Software* 38 (8), pp. 1–29. DOI: 10.18637/jss.v038.i08.

Jacqmin-Gadda, H., C. Proust-Lima, J. M. Taylor, and D. Commenges (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics* 66 (1), pp. 11–19.

Jahn-Eimermacher, A., K. Ingel, A.-K. Ozga, S. Preussler, and H. Binder (2015). Simulating recurrent event data with hazard functions defined on a total time scale. *BMC medical research methodology* 15 (1), p. 16.

Jensen, H., R. Brookmeyer, P. Aaby, and P. K. Andersen (2004). *Shared frailty model for left-truncated multivariate survival data*. Department of Biostatistics, University of Copenhagen.

Johannesma, P. C., R. Reinhard, Y. Kon, J. D. Sriram, H. J. Smit, R. J. A. van Moorselaar, F. H. Menko, and P. E. Postmus (2015). Prevalence of Birt–Hogg–Dubé syndrome in patients with apparently primary spontaneous pneumothorax. *European Respiratory Journal* 45 (4), pp. 1191–1194.

Johansen, S. (1983). An Extension of Cox's Regression Model. *International Statistical Review* 51 (2), pp. 165–174.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. Ed. by Wiley-Interscience. Second edition. Wiley.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53 (282), pp. 457–481.

Kendler, K. S., C. O. Gardner, A. Fiske, and M. Gatz (2009). Major depression and coronary artery disease in the Swedish twin registry: phenotypic, genetic, and environmental sources of comorbidity. *Archives of General Psychiatry* 66 (8), pp. 857–863.

Klein, J. P. (1992). Semiparametric Estimation of Random Effects using the Cox Model based on the EM Algorithm. *Biometrics* 48 (3), pp. 795–806.

Klein, J. P. and M. L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Kosorok, M. (2008). *Introduction to Empirical Process and Semiparametric Inference*. Springer.

Lin, D. Y. and L.-J. Wei (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 84 (408), pp. 1074–1078.

Lin, D. Y., L.-J. Wei, and Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80 (3), pp. 557–572.

Liu, L., R. A. Wolfe, and X. Huang (2004). Shared Frailty Models for Recurrent Events and a Terminal Event. *Biometrics* 60 (3), pp. 747–756.

Louis, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society B*, pp. 226–233.

Mantel, N., N. R. Bohidar, and J. L. Ciminera (1977). Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research* 37 (11), pp. 3863–3868.

McGilchrist, C. and C. Aisbett (1991). Regression with Frailty in Survival Analysis. *Biometrics*, pp. 461–466. DOI: 10.2307/2532138.

Menko, F. H., M. A. van Steensel, S. Giraud, L. Friis-Hansen, S. Richard, S. Ungari, M. Nordenskjöld, T. vO Hansen, J. Solly, E. R. Maher, et al. (2009). Birt-Hogg-Dubé syndrome: diagnosis and management. *The Lancet Oncology* 10 (12), pp. 1199–1206.

Monaco, J. V., M. Gorfine, and L. Hsu (2017). ***frailtySurv**: General Semiparametric Shared Frailty Model*. R package version 1.3.2.

Munda, M., C. Legrand, L. Duchateau, and P. Janssen (2016). Testing for decreasing heterogeneity in a new time-varying frailty model. *Test* 25 (4), pp. 591–606.

Munda, M., F. Rotolo, and C. Legrand (2012). **parfm**: Parametric Frailty Models in R. *Journal of Statistical Software* 51 (1), pp. 1–20. DOI: 10.18637/jss.v051.i11.

Murphy, S. A. (1995a). Asymptotic Theory for the Frailty Model. *The Annals of Statistics* 23 (1), pp. 182–198.

Murphy, S. A. and A. W. van der Vaart (2000). On Profile Likelihood. *Journal of The American Statistical Association* 95 (450), pp. 449–465.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, pp. 712–731.

— (1995b). Asymptotic theory for the frailty model. *The Annals of Statistics*, pp. 182–198.

Nielsen, G. G., R. D. Gill, P. K. Andersen, and T. I. Sørensen (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* 19 (1), pp. 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84 (406), pp. 487–493.

Paddy Farrington, C., S. Unkel, and K. Anaya-Izquierdo (2012). The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (4), pp. 673–696.

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

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* 26 (1), pp. 183–214.

Plummer, M. and B. Carstensen (2011). **Lexis**: An R Class for Epidemiological Studies with Long-Term Follow-Up. *Journal of Statistical Software* 38 (5), pp. 1–12.

Putter, H. and H. C. van Houwelingen (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical methods in medical research* 24 (6), pp. 675–692.

Putter, H. and H. C. Van Houwelingen (2015). Dynamic Frailty Models Based on Compound Birth–Death Processes. *Biostatistics* 16 (3), pp. 550–564. DOI: 10.1093/biostatistics/kxv002.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

— (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rabinowitz, D. (2000). Computing the Efficient Score in Semi-parametric Problems. *Statistica Sinica* 10, pp. 265–280.

Ripatti, S. and J. Palmgren (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 56 (4), pp. 1016–1022.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. Chapman and Hall.

— (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software, Articles* 72 (7), pp. 1–46. ISSN: 1548-7660. DOI: 10.18637/jss.v072.i07.

Rodríguez-Girondo, M., J. Deelen, E. P. Slagboom, and J. J. Houwing-Duistermaat (2018). Survival analysis with delayed entry in selected families with application to human longevity. *Statistical Methods in Medical Research* 27 (3). PMID: 27177884, pp. 933–954. DOI: 10.1177/0962280216648356.

Rondeau, V. and J. R. Gonzalez (2005). **frailtypack**: A Computer Program for the Analysis of Correlated Failure Time Data Using Penalized Likelihood Estimation. *Computer Methods and Programs in Biomedicine* 80 (2), pp. 154–164. DOI: 10.1016/j.cmpb.2005.06.010.

Rondeau, V., S. Mathoulin-Pelissier, H. Jacqmion-Gadda, V. Brouste, and P. Soubeyran (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* 8 (4), pp. 708–721.

Rondeau, V., Y. Mazroui, and J. R. Gonzalez (2012). **frailtypack**: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal of Statistical Software* 47 (4), pp. 1–28. DOI: 10.18637/jss.v047.i04.

Scheike, T. H., J. H. Petersen, and T. Martinussen (1999). Retrospective ascertainment of recurrent events: an application to time to pregnancy. *Journal of the American Statistical Association* 94 (447), pp. 713–725.

Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82 (398), pp. 605–610.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC. College Station, TX.

Sun, W. and H. Li (2004). Ascertainment-adjusted maximum likelihood estimation for the additive genetic gamma frailty model. *Lifetime Data Analysis* 10 (3), pp. 229–245.

Therneau, T. M. (2015a). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. (2015b). ***coxme**: Mixed Effects Cox Models*. R package version 2.2-5.

Therneau, T. M., P. M. Grambsch, and V. S. Pankratz (2003). Penalized Survival Models and Frailty. *Journal of Computational and Graphical Statistics* 12 (1), pp. 156–175. ISSN: 10618600. DOI: 10.2307/1391074.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag, New York. ISBN: 0-387-98784-3. DOI: 10.1007/978-1-4757-3294-8.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika* 77 (1), pp. 147–160.

Vaida, F. and R. Xu (2000). Proportional Hazards Model with Random Effects. *Statistics in Medicine* (19), pp. 3309–3324.

Van den Berg, G. J. and B. Drepper (2011). Inference for shared-frailty survival models with left-truncated data. *IZA Discussion Paper No. 6031*.

Van Noorden, R., B. Maher, and R. Nuzzo (2014). The top 100 papers. *Nature News* 514 (7524), p. 550.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography* 16 (3), pp. 439–454. DOI: 10.2307/2061224.

Vaupel, J. W. and A. I. Yashin (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician* 39 (3), pp. 176–185.

Wienke, A. (2010). *Frailty Models in Survival Analysis*. CRC Press.

Wintrebert, C., H. Putter, A. H. Zwinderman, and J. Van Houwelingen (2004). Centre-effect on Survival after Bone Marrow Transplantation: Application of Time-dependent Frailty Models. *Biometrical Journal* 46 (5), pp. 512–525.

Woodbury, M. A. and K. G. Manton (1977). A random-walk model of human mortality and aging. *Theoretical Population Biology* 11 (1), pp. 37–48.

Yashin, A. I., I. A. Iachine, A. Z. Begun, and J. W. Vaupel (2001). Hidden frailty: myths and reality. *Document de Travail* 34.

Yashin, A. I. and K. G. Manton (1997). Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies. *Statistical Science*, pp. 20–34.

Yashin, A. I., J. W. Vaupel, and I. A. Iachine (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 5 (2), pp. 145–159.

Yau, K. and C. McGilchrist (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* 17 (11), pp. 1201–1213.

Ye, Y., J. D. Kalbfleisch, and D. E. Schaubel (2007). Semiparametric Analysis of Correlated Recurrent and Terminal Events. *Biometrics* 63, pp. 78–87.

Zhi, X., P. M. Grambsch, and L. E. Eberly (2005). Likelihood Ratio Test for the Variance Component in a Semi-Parametric Shared Gamma Frailty Model. *Research Report 2005-5*.

Summary

Survival analysis is the study of time to event data, and it is a major topic in statistics. A prominent type of time to event data is represented by life times, which motivates much of the terminology in the field. As a convention, it is common to refer to the event of interest as *death* or *failure*. An individual that is at risk for *dying* is said to be *alive*. Probably the most distinctive feature of survival data is that the event of interest is not always observed. Rather, the only information available is that the individual had not died before a certain time point. This phenomenon is known as *right censoring* and has motivated the development of special statistical methods for this kind of data.

The probability of being alive at a given time point is given by the *survival* function. The most popular way of estimating this in the presence of right censoring is the "product-limit" estimator, better known as the Kaplan-Meier estimator (Kaplan and Meier, 1958). Their seminal paper, *Nonparametric estimation from incomplete observations*, was found to be the most cited paper in statistics in a recent article in *Nature* (Van Noorden, Maher, and Nuzzo, 2014).

The instantaneous probability of dying at a given time point, given that the individual has not died before, is known as the *hazard* function. In demographics, it is also referred to as the "instantaneous mortality rate". In survival analysis, it is more common to work with the hazard rather than the probability density function. The most popular regression model for survival data is the "proportional hazards" model, commonly referred to as the Cox model (Cox, 1972). The paper that introduced this, titled *Regression Models and Life-Tables*, is the second most cited paper in statistics, according to the same Nature article.

In *The impact of heterogeneity in individual frailty on the dynamics of mortality* (Vaupel, Manton, and Stallard, 1979), the authors refer to the effect of unobserved heterogeneity on mortality as *frailty*. The authors state that "mortality rates for individual may increase faster with age than observed mortality rates for cohorts". This implies that there is a distinction between the individual hazard ("mortality rate") and the population hazard ("mortality rate for cohorts"). Most importantly, Vaupel et al. recognize that the individual hazard cannot be directly observed in the presence of unobserved heterogeneity.

The subtle aspect of the hazard is that, by definition, it refers to the individuals still alive at a certain time point. As individuals with a high frailty tend to die faster, it is likely that individuals who survived longer are less frail, on average, as compared to

the whole sample at the start of follow-up. Frailty models, which aim to model the unobserved heterogeneity with random effects, are discussed in most survival analysis monographs (Andersen, Borgan, et al., 1993; Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2005; Aalen, Borgan, and Gjessing, 2008). Several books offer an exhaustive treatment of such models (Hougaard, 2000; Duchateau and Janssen, 2007; Wienke, 2010).

This dissertation describes new statistical methodology that aims to provide more insight into different aspects of frailty models. Both theoretical properties and practical problems are addressed. Of special interest are the "shared frailty" models, that are employed when the frailty is "shared" between several observations. This is usually the case when an individual may experience more events (recurrent events) or when individuals are related (clustered survival data). In Chapter 1 we focus on the frailty effects on observable quantities in Cox models. In Chapter 2, we present a simulation study that focuses on the properties of shared frailty models for clustered survival data, when the size of the clusters is small. In Chapter 3, we discuss a proposed score test for association between a recurrent event process and a terminal event, when the frailty is shared by both processes. In Chapter 4, we discuss selection bias in the context of recurrent events, where the selection depends on the outcome and on the underlying frailty. In Chapter 5, we present the estimation procedure implemented in the **frailtyEM** R package. In what follows, we show a more detailed summary for each chapter.

Chapter 1 is the introduction to this dissertation. It follows the structure of a tutorial, providing an overview of theory and practice surrounding frailty models. In Section 1.2, we address to *univariate frailty* models. These are related to the original formulation of Vaupel, Manton, and Stallard (1979), where the outcome of interest is a singular event for individuals (death), and the individual event times are assumed to be independent of each other. Via simulated examples, we illustrate two phenomena specific to Cox models. First, the *selection* process, that describes the distribution of risk factors in the population of survivors. Second, the observed *marginal* covariate effect in the Cox model, when important explanatory variables are omitted. The same phenomena are then studied in detail with frailty models, for different frailty distributions. The chapter concludes with a discussion of the identifiability properties of frailty models in univariate survival data.

In Section 1.3, we illustrate via a simulated data example how marginal correlation between event times may arise, when covariates "shared" by related individuals are missing. This is further studied with *shared frailty* models, wherein the random effect is assumed to be shared between different individuals. We study how different correlation patterns arise from different frailty distributions and we discuss how shared frailty models may be used for modeling recurrent events. In Section 1.4 we address practical issues surrounding the estimation of frailty models. We discuss different procedures for semiparametric and parametric models, we review the available software and describe how different data types can be accommodated by software packages. Finally, in Section 1.5 we discuss several proposed extensions of the frailty model.

In **Chapter 2** we analyze situations where it is difficult to tell the difference between non-proportional hazards and unobserved heterogeneity. This chapter builds on the results discussed in Chapter 1, especially those regarding the identifiability of frailty models. A well known result is that the frailty model is identifiable if covariates are present and the frailty distribution has finite moments. We argue that this is problematic, because the frailty may falsely explain a time dependent covariate effect as evidence for unobserved heterogeneity. While generally thought that this is not a problem for shared frailty models, we show that it may be, especially if the cluster size is small.

In Section 2.2, we review the proportional hazards models and the conditional proportional hazards assumption commonly made for frailty models. Next, we discuss how marginal non-proportional hazards may arise from different frailty models. In Section 2.3, we present the simulation study. We study the effect of the cluster size (in fact, how "multivariate" the outcome is) on detecting frailty models, when there is no real unobserved heterogeneity. We analyze the results for different quantities of interest: the likelihood ratio test, the score test for heterogeneity and estimated parameters. Our main conclusion is that time dependent covariate effects may falsely appear as evidence for frailty, when the path of the effect is somewhat similar to the marginal hazard ratio implied by the frailty model. Although this problem is mitigated with larger sample sizes, when the cluster size is small (e.g. 2, 3) the distinction between unobserved heterogeneity and time-dependent covariate effects is subtle. The results are extended to recurrent events, and a combination of time dependent covariate effects in the presence of frailty. Finally, the phenomenons analyzed in this chapter are illustrated with a data analysis of a well known data set on recurrent kidney infections.

In **Chapter 3**, we introduce a score test for association between recurrent events and a terminal event. If frailty is present and high frailty individuals are associated both with a higher rate of recurrent events and a higher mortality, then the two event processes must be jointly analyzed. This is complicated in practice, especially for semiparametric models. We propose a simple score test for association testing the null hypothesis that the two models are independent. If this is not rejected, simpler separate analyses may be carried out.

In Section 3.2, a joint model for recurrent events and a terminal event is introduced, employing a gamma distributed frailty. This model includes an association parameter that may be estimated, for which different inference methods are compared. In Section 3.3, the "robust score test" is introduced, together with other well known statistical tests, for the null hypothesis of no association. In Section 3.4, we show via a simulation study that the proposed test behaves well and, in terms of power, is comparable to more complicated alternatives. In Section 3.5, the proposed methodology is illustrated on a data set comprising recurrent skin tumors.

In **Chapter 4**, the problem of selection bias (or "ascertainment" bias) in recurrent events is analyzed. The motivating example is a data set comprising recurrent pneu-

mothoraces. The data was collected only for individuals that had at least one recorded event during a certain accrual time window. For the selected individuals, the whole event history was collected. The problem is that, by design, individuals with a higher rate of events will be over represented in this sample. If unobserved heterogeneity is present, high frailty patients are over represented. In this chapter, we study the estimation of frailty parameters and covariate effects in this type of scenarios.

In Section 4.2, several selection schemes and a general adjusted likelihood approach are introduced. We discuss the effects of the ascertainment on the estimates from a model without frailty and from a model with frailty. For the latter, a pseudo maximum likelihood estimation algorithm is presented. In Section 4.3, the performance of the adjusted likelihood approach is studied for different selection scenarios, and it is shown to work well in general. Finally, in Section 4.4, the proposed methodology is illustrated on the original motivating data set.

In **Chapter 5**, we study the estimation of semiparametric shared frailty models in practice, with a focus on the **frailtyEM** package (Balan and Putter, 2017) for the R programming language. This software is meant to combine the flexibility of semiparametric models with a large choice of frailty distributions. A major motivation behind writing this package was to provide well documented user level features. In Section 5.1, we present an overview of the currently available software for the estimation of frailty models.

In Section 5.2, the likelihood construction and the effect of left truncation and ascertainment are discussed in the context of frailty models. Next, we make an overview of related results regarding practical problems: hypothesis testing, marginal and conditional quantities and goodness of fit. In Section 5.3, the software implementation of a profile expectation maximization algorithm is discussed. The proposed estimation method and the calculations required to obtain standard errors are presented. From a practical point of view, the functions provided by the package are presented, together with their corresponding syntax. Finally, the features of the package are illustrated with examples involving three well known data sets, covering three important scenarios: recurrent events in calendar time, recurrent events in gap time and clustered failures.

Samenvatting

De overlevingsanalyse behelst de studie van de tijdsduur tot een gebeurtenis, wat een belangrijk onderwerp binnen de statistiek is. Een prominent type data over de tijdsduur tot een gebeurtenis is de levensduur, waar veel van de terminologie aan wordt ontleend. Het is gebruikelijk om naar de gebeurtenis waar de interesse naar uitgaat te refereren als *overlijden* of *falen*. Een individu dat risico loopt om te *overlijden* wordt in *leven* genoemd. Waarschijnlijk is de meest karakteristieke eigenschap van overlevingsdata dat de gebeurtenis waar de interesse naar uitgaat niet altijd wordt geobserveerd. De enige informatie die dan beschikbaar is, is dat het individu niet voor een bepaald tijdstip overleden is. Dit fenomeen staat bekend als *rechtscensurering* en is de drijfveer geweest voor de ontwikkeling van speciale statistische methoden voor dit soort data.

De kans om op een bepaald tijdstip in leven te zijn wordt gegeven door de *overlevingsfunctie*. De meest populaire manier om deze te schatten in geval er rechtsgecensureerde waarnemingen zijn is de “product-limit” schatter, die beter bekend is als de “Kaplan-Meier” schatter (Kaplan en Meier, 1958). Hun belangrijke paper, *Nonparametric estimation from incomplete observations*, bleek in een recent artikel in *Nature* (Van Noorden, Maher en Nuzzo, 2014) het meest geciteerde statistiek-artikel te zijn.

De instantane kans om te overlijden op een bepaald tijdstip, gegeven dat het individu niet al eerder overleden is, staat bekend als de hazardfunctie. In demografische studies wordt dit ook wel de ïnstantane mortaliteitsgraad"genoemd. In de overlevingsanalyse is het gebruikelijker om met de hazard te werken dan met de dichtheidsfunctie. Het meest populaire regressiemodel voor overlevingsdata is het "proportionele hazardsmodel, waar vaak naar verwezen wordt als het Cox-model. Het artikel waarin dit model werd geïntroduceerd, genaamd Regression Models and Life-Tables, is het op één na meest geciteerde statistiekartikel, volgens hetzelfde artikel in Nature.

In *The impact of heterogeneity in individual frailty on the dynamics of mortality* (Vaupel, Manton en Stallard, 1979), verwijzen de auteurs naar het effect van ongeobserveerde heterogeniteit op mortaliteit als fragiliteit (*frailty*). De auteurs zeggen dat “de mortali-

teit van een individu kan sneller toenemen als de leeftijd toeneemt dan de waargenomen mortaliteit in cohorten". Dit impliceert dat er een onderscheid is tussen de individuele hazard ("mortaliteit") en de populatiehazard ("mortaliteit voor cohorten"). Belangrijk hierbij is dat Vaupel et al. inzien dat de individuele hazard niet direct kan worden waargenomen wanneer er sprake is van ongeobserveerde ongelijksoortigheid.

Het subtiele kenmerk van de hazard is dat deze per definitie de individuen betreft die op een bepaald tijdstip nog in leven zijn. Aangezien individuen met een hoge fragiliteit geneigd zijn eerder te overleden, is het aannemelijk dat individuen die langer overleefd hebben, gemiddeld gezien minder fragiel zijn, in vergelijking met de hele steekproef aan het begin van de studie. Fragiliteitsmodellen, die als doel hebben om ongeobserveerde ongelijksoortigheid te modelleren met behulp van zogeheten "random effects", worden behandeld in de meeste boeken over overlevingsanalyse (Andersen, Borgan e.a., 1993; Kalbfleisch en Prentice, 2002; Klein en Moeschberger, 2005; Aalen, Borgan en Gjessing, 2008). Meerdere boeken bieden een uitvoerige uiteenzetting van zulke modellen (Hougaard, 2000; Duchateau en Janssen, 2007; Wienke, 2010).

In dit proefschrift wordt nieuwe statistische methodologie beschreven, die als doel heeft om meer inzicht in verschillende aspecten van fragiliteitsmodellen te bieden. Zowel theoretische eigenschappen als praktische problemen worden behandeld. Speciale aandacht gaat uit naar "gedeelde fragiliteit" modellen, die gebruikt worden wanneer de fragiliteit "gedeeld" wordt onder meerdere waarnemingen. Dit is meestal het geval wanneer een individu meerdere gebeurtenissen kunnen overkomen (recurrente gebeurtenissen) of wanneer individuen verwant zijn aan elkaar (geclusterde overlevingsdata). In Hoofdstuk 1 ligt de nadruk op fragiliteitseffecten op waarneembare grootheden in Cox-modellen. In Hoofdstuk 2 presenteren we een simulatiestudie die gericht is op eigenschappen van gedeelde fragiliteitsmodellen voor geclusterde overlevingsdata, wanneer de clusters klein zijn. In Hoofdstuk 3 bespreken we een voorgestelde score toets voor associatie tussen een recurrent gebeurtenissenproces en een terminale gebeurtenis, wanneer de frailty wordt gedeeld door beide processen. In Hoofdstuk 4 bespreken we selectiebias in de context van recurrente gebeurtenissen, waar de selectie afhangt van de uitkomst en de onderliggende fragiliteit. In Hoofdstuk 5 presenteren we de schattingsprocedure geïmplementeerd in de **frailtyEM** R software. Hieronder volgt een meer gedetailleerde samenvatting van elk hoofdstuk.

Hoofdstuk 1 is de inleiding van dit proefschrift. Het heeft de structuur van een tutorial, en geeft een overzicht van de theorie en praktijk rondom fragiliteitsmodellen. In Sectie 1.2 bespreken we univariate fragiliteitsmodellen. Deze worden gerelateerd aan de originele formulering van Vaupel, Manton en Stallard (1979), waar de uitkomst waar de interesse naar uitgaat een enkelvoudige gebeurtenissen is for individuen (overlijden), en waarbij wordt aangenomen dat de individuele gebeurtenistijdstippen onafhankelijk zijn van elkaar. Met behulp van gesimuleerde voorbeelden illustreren we twee fenomenen die specifiek voor Cox-modellen zijn. Ten eerste het selectieproces, dat de verdeling van de risicofactoren in de overlevendenpopulatie beschrijft. Ten tweede, het geobser-

veerde marginale covariateneffect in het Cox-model, wanneer belangrijke verklarende variabelen worden weggelaten. Dezelfde fenomenen worden dan in detail binnen fragiliteitsmodellen bestudeerd, voor verschillende fragiliteitsverdelingen. De sectie wordt afgesloten met een discussie van de identificeerbaarheidseigenschappen van fragiliteitsmodellen in univariate overlevingsdata.

In Sectie 1.3 illustreren we door middel van simulatie hoe marginale correlatie tussen gebeurtenistijdstippen kan ontstaan, wanneer door verwante individuen "gedeelde" covariaten ontbreken. Dit wordt verder bestudeerd aan de hand van gedeelde fragiliteitsmodellen, waarin wordt aangenomen dat een "random effect" wordt gedeeld door meerdere individuen. We bestuderen hoe verschillende correlatiepatronen ontstaan bij verschillende fragiliteitsverdelingen, en we bespreken hoe gedeelde fragiliteitsmodellen gebruikt kunnen worden voor het modelleren van recurrente gebeurtenissen. In Sectie 1.4 gaan we in op praktische zaken rondom het schatten van fragiliteitsmodellen. We bespreken verschillende procedures voor semiparametrische en parametrische modellen, geven een overzicht van de beschikbare software en beschrijven hoe verschillende soorten data kunnen worden geanalyseerd in softwarepakketten. Tot slot bespreken we in Sectie 1.5 verscheidene voorgestelde uitbreidingen van het fragiliteitsmodel.

In **Hoofdstuk 2** analyseren we situaties waarin het moeilijk is om het verschil te zien tussen hazards en ongeobserveerde ongelijksoortigheid. Dit hoofdstuk bouwt voort op de resultaten uit Hoofdstuk 1, in het bijzonder degene over de identificeerbaarheid van fragiliteitsmodellen. Een zeer bekend resultaat is dat het fragiliteitsmodel identificeerbaar is als er covariaten zijn en de fragiliteitsverdeling eindige momenten heeft. We beargumenteren dat dit problematisch is, omdat de fragiliteit onterecht het effect van een tijdsafhankelijke covariaat kan toewijzen aan ongeobserveerde ongelijksoortigheid. Terwijl over het algemeen gedacht wordt dat dit geen probleem is voor gedeelde fragiliteitsmodellen, laten we zien dat het dat toch kan zijn, vooral als de clusters klein zijn.

In Sectie 2.2 beschouwen we het proportionele hazards-model en de conditionele proportionele hazards-aanname die vaak gemaakt wordt voor fragiliteitsmodellen. Vervolgens bespreken we hoe marginale hazards kunnen ontstaan van verschillende fragiliteitsmodellen. In Sectie 2.3 presenteren we de simulatiestudie. We bestuderen het effect van clustergrootte (in feite hoe "multivariaat" de uitkomst is) op het detecteren van fragiliteitsmodellen, wanneer er in werkelijkheid geen ongeobserveerde ongelijksoortigheid is. We analyseren de uitkomsten voor meerdere grootheden waar de interesse naar uitgaat: de likelihood ratio-toets, de score toets voor ongelijksoortigheid en geschatte parameters. Onze belangrijkste conclusie is dat effecten van tijdsafhankelijke covariaten onterecht kunnen worden opgevat als bewijs voor fragiliteit, wanneer het tijdsverloop van het effect enigszins lijkt op de marginale hazardratio geassocieerd met het fragiliteitsmodel. Alhoewel dit probleem minder sterk is bij grotere steekproefgroottes, is het onderscheid tussen ongeobserveerde ongelijksoortigheid en covariaateffecten subtiel wanneer de clusters klein zijn (e.g. 2,3). De resultaten worden uitgebreid naar

recurrente gebeurtenissen, en een combinatie van tijdsafhankelijke covariaateffecten in de aanwezigheid van fragiliteit. Ter afsluiting worden de fenomenen die geanalyseerd worden in dit hoofdstuk geïllustreerd aan de hand van een data-analyse van een bekende dataset over recurrente nierinfecties.

In **Hoofdstuk 3**, introduceren we een score toets voor associatie tussen recurrente gebeurtenissen en een terminale gebeurtenis. Als er fragiliteit aanwezig is en zeer fragiele individuen een associatie hebben met zowel een hoger aantal recurrente gebeurtenissen als een hogere mortaliteit, dan moeten beide gebeurtenisprocessen gezamenlijk worden geanalyseerd. Dit is ingewikkeld in de praktijk, vooral met semiparametrische modellen. We stellen een eenvoudige score toets voor associatie voor, die de nulhypothese dat de twee modellen onafhankelijk zijn toetst. Als deze niet verworpen wordt, kunnen eenvoudigere analyses worden uitgevoerd.

In Sectie 3.2 wordt een gezamenlijk model voor recurrente gebeurtenissen en een terminale gebeurtenis geïntroduceerd, met een fragiliteit die een gammaverdeling heeft. Dit model bevat een associatieparameter die geschat kan worden, waarvoor verschillende inferentiemethoden worden vergeleken. In Sectie 3.3 wordt de "robuuste score-toets"geïntroduceerd, samen met andere bekende statistische toetsen, voor de nulhypothese dat er geen associatie is. In Sectie 3.4 laten we met een simulatiestudie zien dat de voorgestelde toets goed werkt en qua onderscheidend vermogen vergelijkbaar is met gecompliceerde alternatieven. In Sectie 3.5 wordt de voorgestelde methodologie geïllustreerd met een toepassing op data over recurrente huidtumoren.

In **Hoofdstuk 4**, wordt het probleem van selectiebias (ook wel “toerekeningsbias”) bij recurrente gebeurtenissen geanalyseerd. Het begeleidende voorbeeld is een dataset bestaande uit recurrente klaplongen. De data is alleen verzameld voor individuen met tenminste één geregistreerde gebeurtenis gedurende een zekere aanwasperiode. Voor de geselecteerde individuen is de gehele gebeurtenissengeschiedenis verzameld. Het probleem is dat door deze opzet individuen met een hoger aantal gebeurtenissen overgepresenteerd zullen zijn in deze steekproef. Als ongeobserveerde gelijksoortigheid aanwezig is, zullen zeer fragiele patiënten overgerepresenteerd zijn. In dit hoofdstuk bestuderen we het schatten van fragiliteitsparameters en covariaateffecten in dit soort scenarios.

In Sectie 4.2 worden een aantal selectiemanieren en een algemene geadjusteerde likelihoodbenadering geïntroduceerd. We bespreken de effecten van de selectie op de schattingen van een model zonder fragiliteit en van een model met fragiliteit. Voor laatstgenoemd model wordt een pseudo-maximum-likelihood schattingsalgoritme gepresenteerd. In Sectie 4.3 worden de prestaties van de geadjusteerde likelihoodbenadering bestudeerd voor verschillende selectiescenarios, en wordt getoond dat deze over het algemeen goed werkt. Tot slot wordt in Sectie 4.4 de voorgestelde methodologie geïllustreerd door toepassing op de begeleidende dataset.

In **Hoofdstuk 5**, bestuderen we het schatten van semiparametrische gedeelde fragiliteitsmodellen in de praktijk, met de nadruk op de **frailtyEM** software (Balan en Putter, 2017) voor de programmeertaal R. Deze software is bedoeld om de flexibiliteit van semiparametrische modellen te combineren met een ruime keuze aan fragiliteitsverdelingen. Een belangrijke reden om deze software te schrijven was om te voorzien in goedgedocumenteerde mogelijkheden op gebruikersniveau. In Sectie 5.1 geven we een overzicht van op het moment beschikbare software voor het schatten van fragiliteitsmodellen.

In Sectie 5.2 worden de constructie van de likelihood en het effect van linkstruncatie en selectie besproken in de context van fragiliteitsmodellen. Daarna geven we een overzicht van gerelateerde resultaten voor praktische problemen: hypothese toetsen, marginale en conditionele grootheden en kwaliteit van de fit. In Sectie 5.3 wordt de softwareimplementatie van een geprofileerd verwachtingsmaximalisatiealgoritme besproken. De voorgestelde schattingsmethode en de berekeningen die nodig zijn om standaardfouten te verkrijgen worden gepresenteerd. Uit praktisch oogpunt worden de functies uit de software gepresenteerd, samen met de bijbehorende syntax. Tot slot worden de mogelijkheden van de software geïllustreerd met voorbeelden waar drie bekende datasets in voorkomen, en die drie belangrijke scenarios omvatten: recurrente gebeurtenissen in kalendertijd, recurrente gebeurtenissen in tussenliggende tijd, en geclusterde overleveingsgegevens.

List of Publications

T.A. Balan, S.E. Boonk, M.H. Vermeer, H. Putter (2016). Score test for association between recurrent events and a terminal event. *Statistics in Medicine* 35(12), 3037–3048.

T.A. Balan, M.A. Jonker, P.C. Johannesma, H. Putter (2016). Ascertainment correction in frailty models for recurrent events data. *Statistics in Medicine* 35(23), 4183–4201

T.A. Balan, H. Putter (2018). **frailtyEM**: an R package for estimating semiparametric shared frailty models. *Journal of Statistical Software, manuscript accepted for publication.*

T.A. Balan, H. Putter (2018). Non-proportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference?. *Manuscript submitted for publication.*

T.A. Balan, H. Putter (2018). A tutorial in frailty models. *Manuscript in preparation.*

Acknowledgements

I would like hereby to express a few words of thank you to the people who have been with me on this journey.

I would like to extend my deepest gratitude to Prof.dr. Hein Putter, whom I have been working with for over five years. Thank you for your generosity, for the trust you put in me and for being a role model. I am deeply honored to call you my scientific mentor.

During the research that led to this dissertation, I had the pleasure to work with a number of inspiring people. These include Marianne, David, Nan, Erik and Ewout. I am also especially thankful to my colleagues from the *survival analysis* group, including Mia, Marta, Liesbeth, Mar and Irene. I have learned a lot from our discussions and collaborations.

Indirectly, however, many more colleagues contributed in one way or another to this end result. I would like to especially thank Stéphanie for the help with composing the Dutch summary of this thesis. To my past and present fellow PhD students, I would like to say that it was a pleasure sharing this experience with you. I would also like to show my appreciation to my colleagues Saskia, Roula, Bart, Stefan, Jelle, Mirko, Szymon, Jeanine and Theo. And of course to Bruna, Alexia and Giorgos - I will always remember fondly our conversations.

I am grateful for the good friends that I have made in Leiden over the years: Thanos, Olga, Rafat, Abdelrahman, Viola, Polykarpos, Irini and Anthippi. Thank you all for your support. A special place will always be reserved for the statisticians that I used to call colleagues, and now I am happy to call friends: Katerina, Manos, Razieh, Shane and Flavio. In one way or another, you motivated me to go forward with the PhD studies. I would like to thank Alexandru for the wonderful work of designing the cover of this thesis, and for his priceless friendship over the years. To my closest Romanian friends, Dragoș, Simona, Matei and Sabina. Words cannot express what you all mean to me.

Last but not least, I would have not been able to dream of pursuing studies abroad without the everlasting support of my family. I will never be able to show how grateful I am for your encouragement. I dedicate this thesis to my parents, Luminița and Corneliu. To Miruna and Cristian. Thank you all for being with me at every step. Time and distance seem like nothing, in the face of your unconditional love.

Curriculum Vitae

Theodor Adrian Bălan was born on the 8th of March 1989 in Bucharest, Romania. He completed his secondary studies in 2008 at *Colegiul Național Sfântul Sava* (Saint Sava National College).

In 2011, he earned a bachelor's degree in Applied Mathematics from the University of Bucharest, with a dissertation titled *Factor Analysis and Applications*. Afterwards, he continued his education at Leiden University within the Statistical Science for Life & Behavioural Sciences programme. From 2012 to 2013 he was president of the International Student Network Leiden organization. In 2013, he graduated with a master's degree *cum laude* in Mathematics, with the dissertation *Joint Modeling of Recurrent and Terminal Events: A Simulation Study* written under the supervision of Prof. dr. Hein Putter.

In 2013, he started his PhD research at the department of Medical Statistics and Bioinformatics (now Biomedical Data Sciences) at Leiden University Medical Center under the supervision of Prof. dr. Hein Putter. His work focused on extending methodology regarding random effect models for time to event data, also known as *frailty models*. The results of this research are presented in this thesis. During this time, he was awarded three travel grants (2014, University of Milano-Biccoca; 2016, Leiden University Funds; 2017, International Biometric Society).

He is the author and maintainer of the R packages **frailtyEM** and **dynfrail**. He co-wrote and taught the course *Frailty Models: Theory and Practice* in Prague in 2017, for the Czech National Group of the International Society for Clinical Biostatistics, and as part of an invited *Statistics in Practice* session at the International Biometrical Conference in Barcelona in 2018. He is a reviewer for Biostatistics, Statistics in Medicine and Statistical Methods in Medical Research and Biometrical Journal.