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

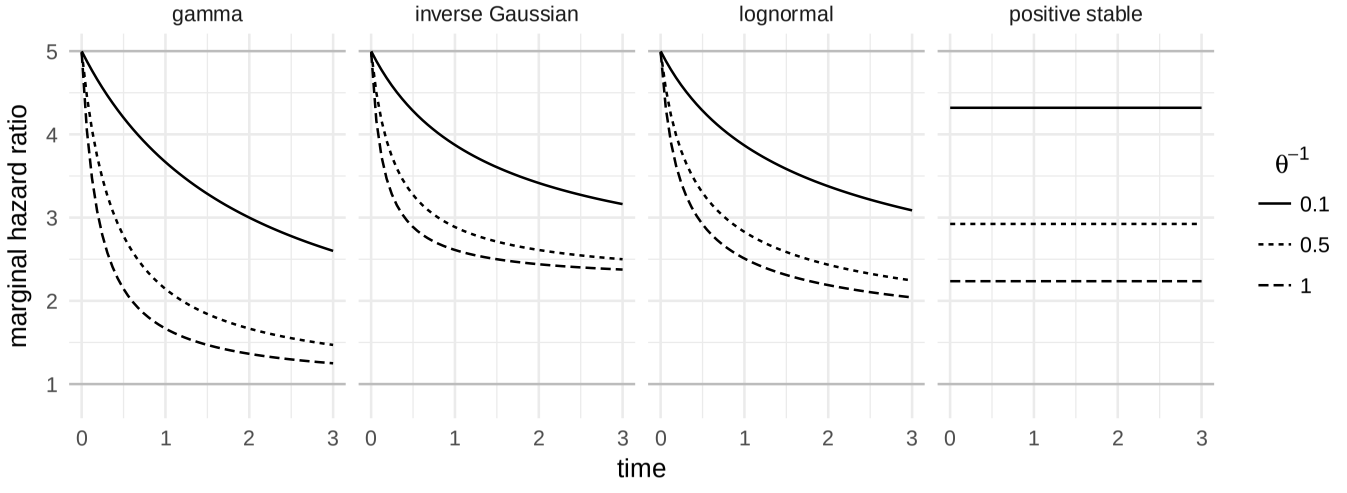

Figure 21: Marginal hazard ratio of survivors obtained a conditional hazard ratio of 5, for the gamma, inverse Gaussian, lognormal and positive stable distributions, where the baseline hazard is $\lambda_0(t) = 1$. The gamma, inverse Gaussian and lognormal have fixed $\mathrm{E}Z = 1$ and $\mathrm{Var}Z = \theta^{-1}$. 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

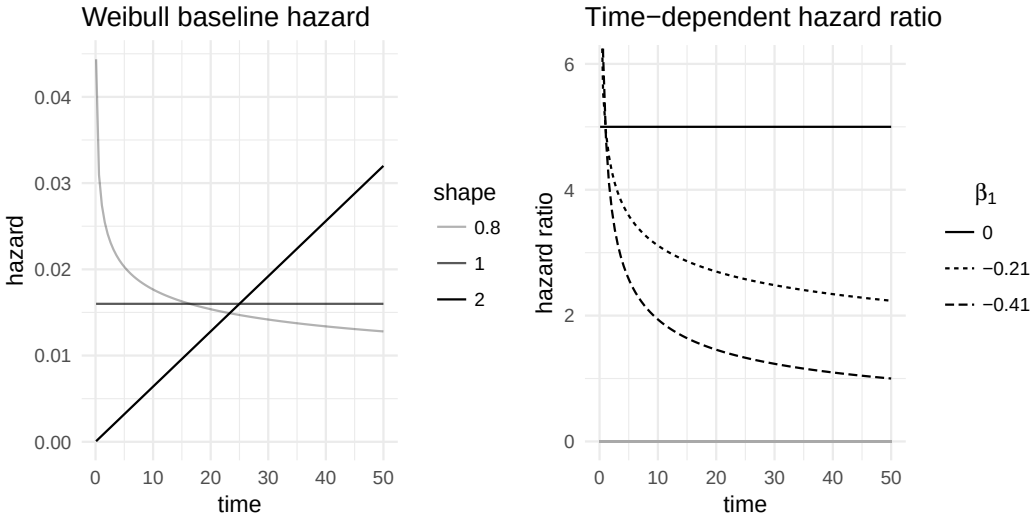

Figure 22: Left: Weibull baseline hazards used in the simulation, where the scale parameter is chosen so that the cumulative baseline hazard at 50 is 0.8. Right: time dependent hazard ratio used in the simulation and describe in equation (2.7), i.e. $5 \exp(\beta_1 \log t)$.

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

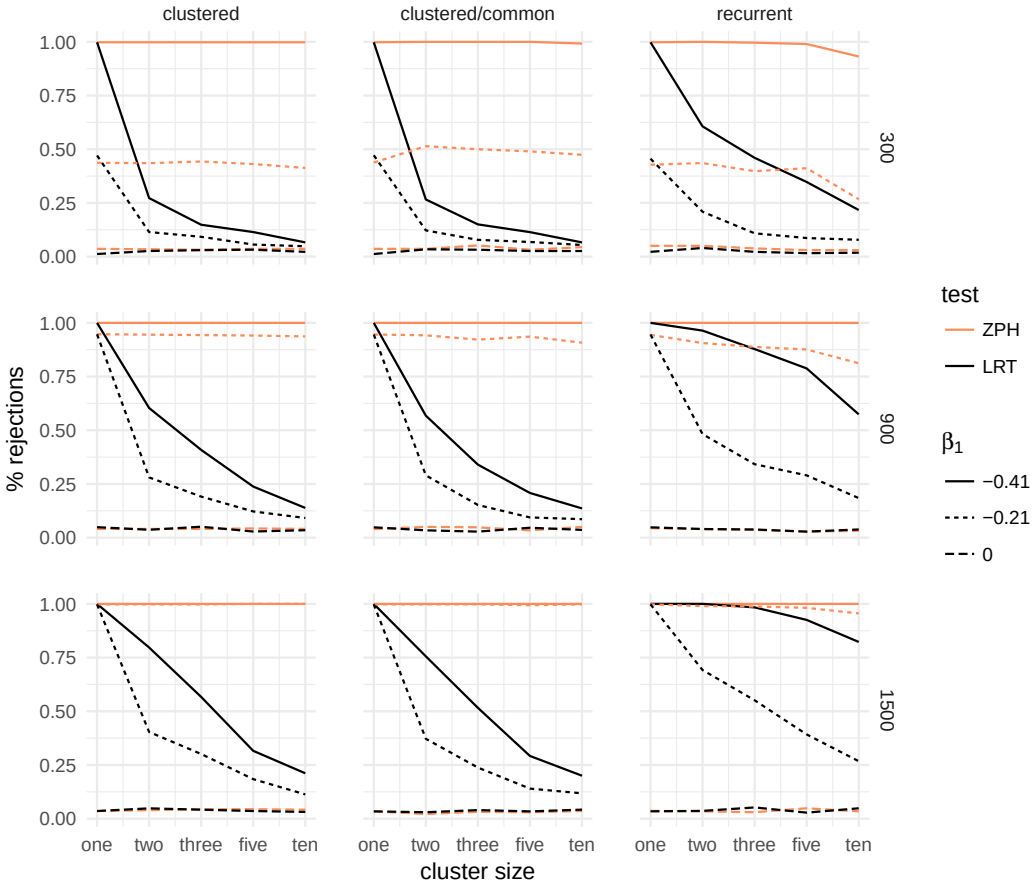

Figure 23: Percentage of rejections of the likelihood ratio test (LRT) between a gamma frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated without unobserved common risk and an increasing Weibull baseline hazard with shape $\alpha = 0.8$. 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.

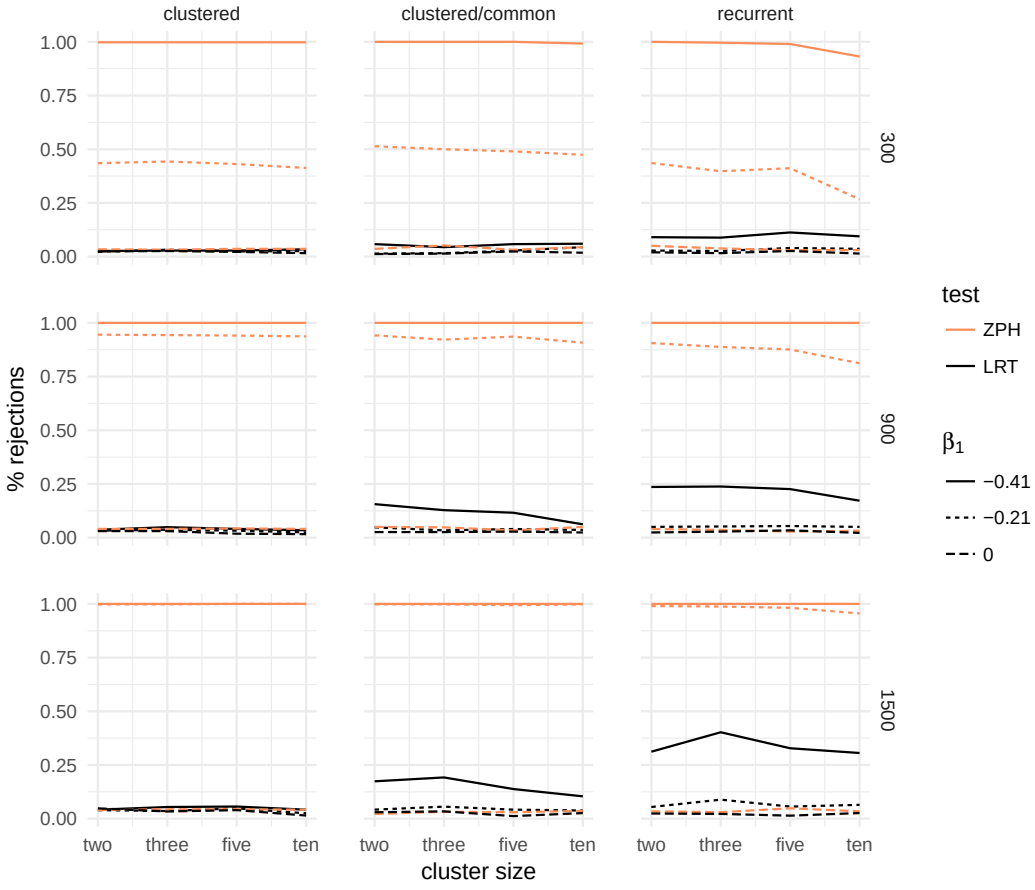

Figure 24: Percentage of rejections of the likelihood ratio test (LRT) between a positive stable frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated without unobserved common risk and an increasing Weibull baseline hazard 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.

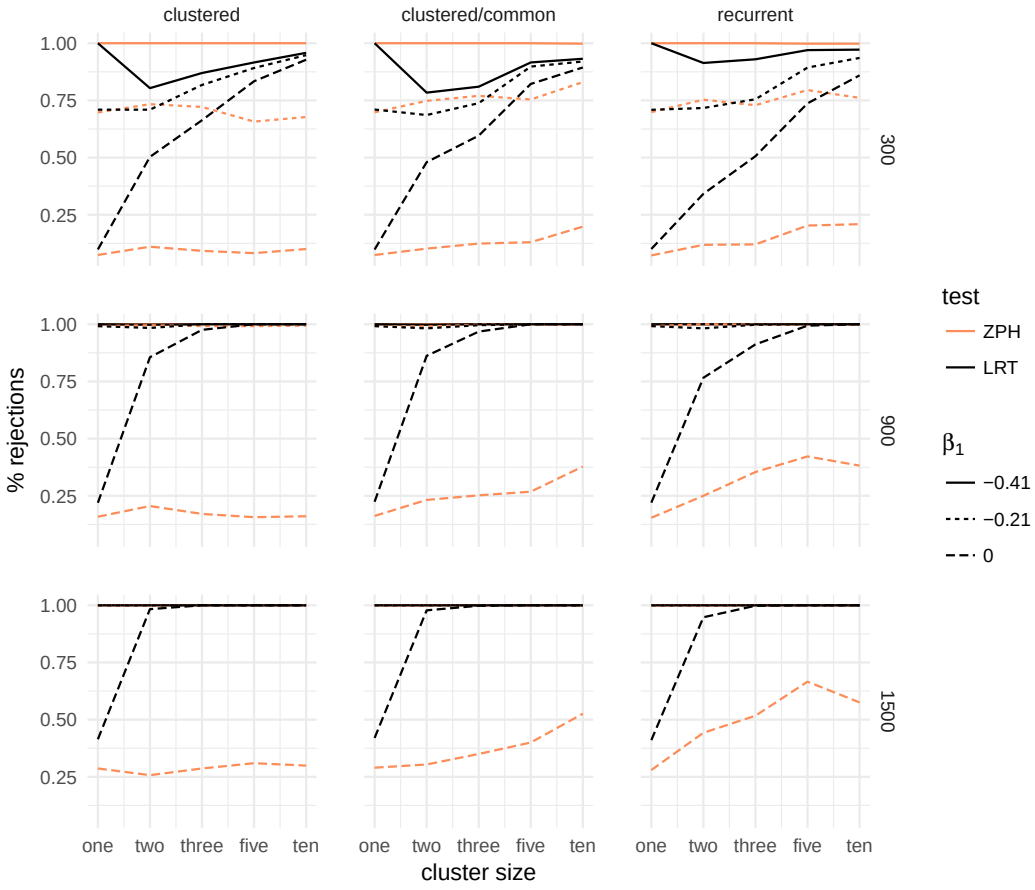

Figure 25: Percentage of rejections of the likelihood ratio test (LRT) between a gamma frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 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.

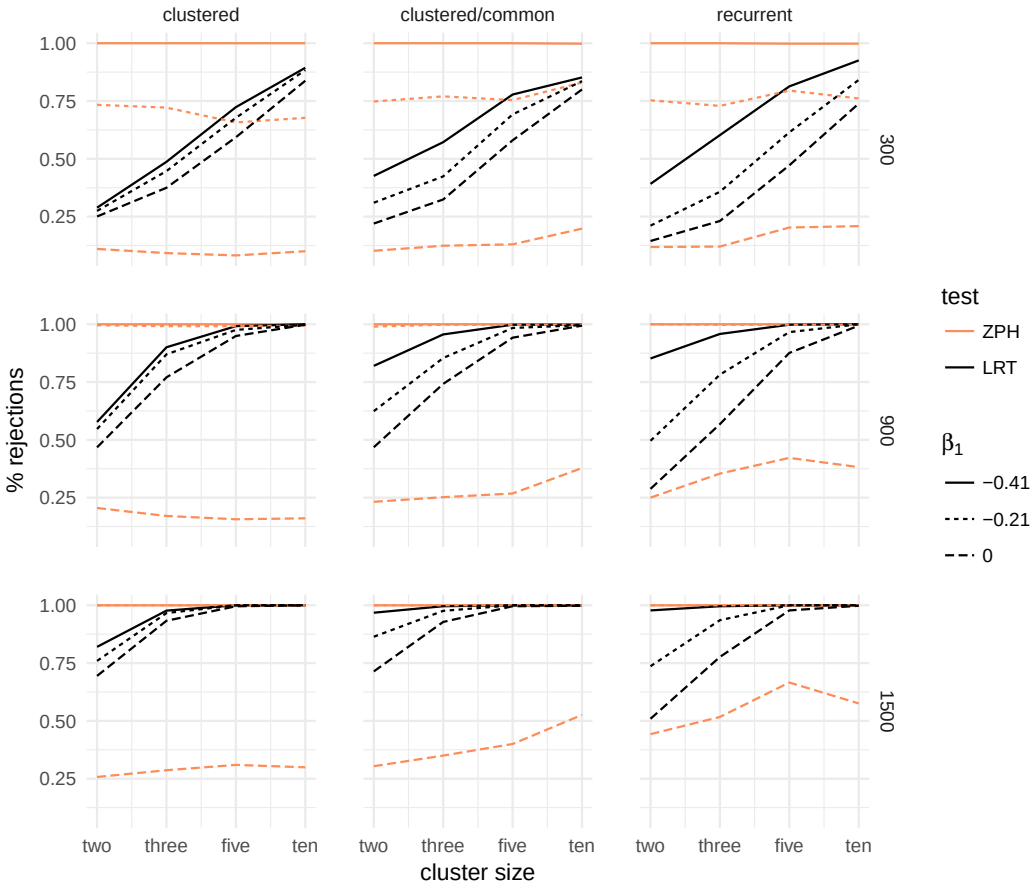

Figure 26: Percentage of rejections of the likelihood ratio test (LRT) between a positive stable frailty model and a proportional hazard model compared to the test for non-proportional hazards (ZPH), when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 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

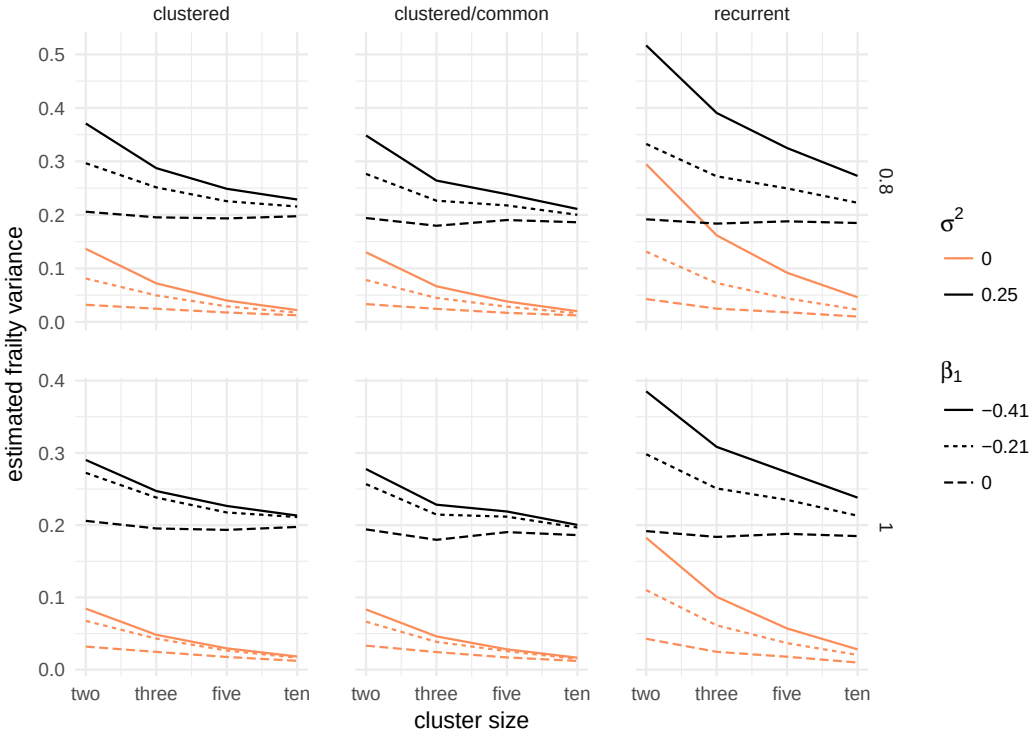

Figure 27: Estimated frailty variance for a gamma frailty model, when the data are simulated with an unobserved common risk following a lognormal distribution with expectation 1 and variance $\sigma^2 \in \{0, 0.25\}$ and a total sample size of 300. The rows correspond to the Weibull baseline shape parameter, 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.

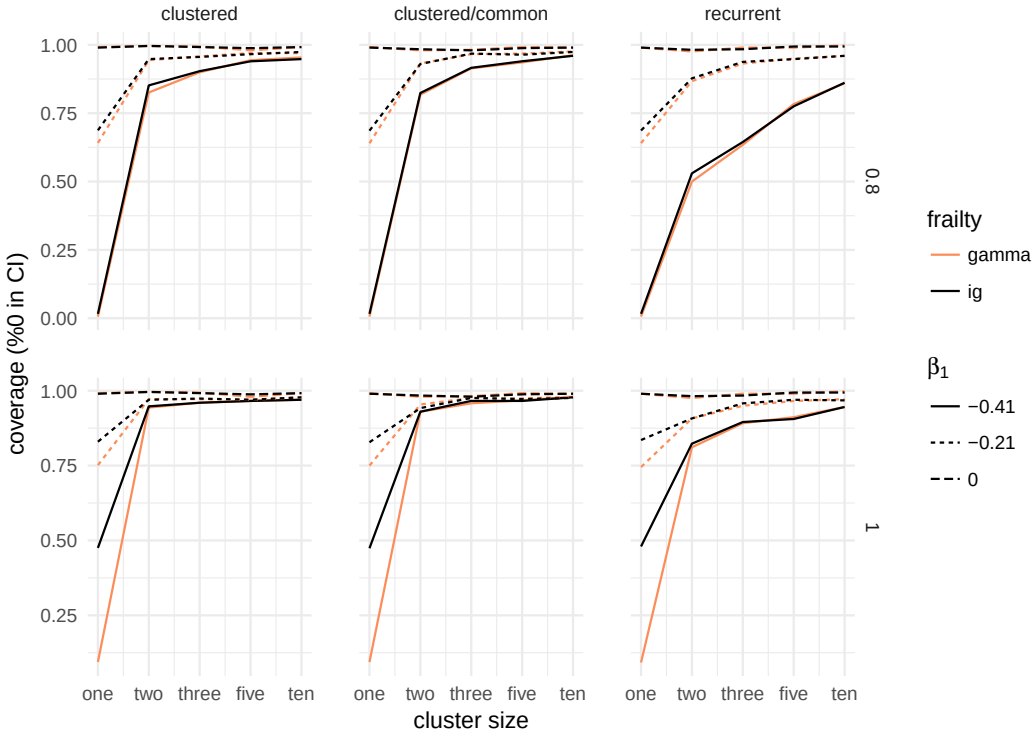

Figure 28: Coverage of the likelihood-based confidence interval for the gamma frailty variance for the gamma and inverse Gaussian distributions, when the data are simulated with no unobserved heterogeneity (true variance is 0) and a total sample size of 300. The rows correspond to the Weibull baseline 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

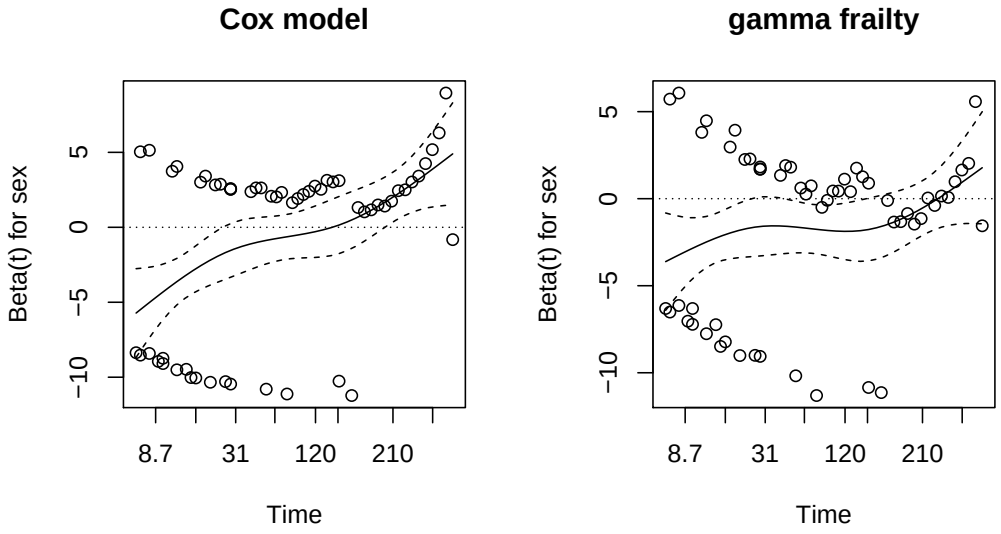

Figure 29: Plot of the Schoenfeld residuals for sex from a Cox marginal model and a gamma frailty model estimated on the kidney catheter insertions data.

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