I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/55

Testing for Decreasing Heterogeneity Between Hospitals in Time to Death from Chronic Myeloid Leukemia

MUNDA, M., LEGRAND, C., DUCHATEAU, L. and P. JANSSENS

Testing for Decreasing Heterogeneity Between Hospitals in Time to Death from Chronic Myeloid Leukemia

Marco Munda*, Catherine Legrand, Luc Duchateau, Paul Janssen

October 2013

Abstract

Frailty models adjust for between-cluster variability in survival data by including a cluster-specific random factor, the frailty term, in the Cox model. The frailty term acts on the hazard rate in a multiplicative way and is assumed to be constant over time. This assumption is questionable in some particular settings, e.g., in cancer clinical trials on chronic myeloid leukemia. We therefore relax this assumption and consider frailty models with time-varying frailties. Instead of working with hazard models, we rather model the log cumulative hazard function, making use of the mixed model framework, and introduce a time-varying random effect at that level. Simulations demonstrate that the proposed method has acceptable size and power to detect time-dependent clustering. The method is applied to data from a multicentre clinical trial in patients with chronic myeloid leukemia.

1 Introduction

The Cox model is one of the basic models for survival data, e.g. time to death in cancer clinical trials. The model, however, requires independent (homogeneous) data up to measured covariates. For clustered survival data, we can use shared frailty models which adjust for the cluster-to-cluster variability by including a random factor, the frailty term, in the Cox model. Frailty models are therefore well suited for the analysis of large-scale clinical trials conducted at multiple sites (Legrand et al., 2006).

*marco.munda@uclouvain.be

In the standard frailty model, the frailty term is assumed to be constant over time. For some particular studies, this assumption is known to be questionable. For example, the cancer clinical trial on chronic myeloid leukemia (CML) analysed in Wintrebert et al. (2004) is one such study. Patients first receive bone marrow transplantation, whence they are at risk of death due to transplant-related causes. Within that period, substantial heterogeneity between transplant centres can be foreseen. Thereafter, heterogeneity is expected to decrease.

Following Massonnet et al. (2008), Teodorescu et al. (2010), and López-de-Ullibarri et al. (2012), survival models can be based on the logarithm of the cumulative hazard. The key idea is that, on the logarithmic scale, the cumulative hazard function has a linear model structure. Random effects can then be introduced at that level, leading to a linear mixed effects model. While the counterpart of the frailty model in the mixed model world is the simple random intercept model, more complex data structures can be studied in this framework, provided that enough data information is available. Furthermore, linear mixed effects models can easily be handled by standard software, e.g. the `proc mixed` in SAS.

To model decreasing (or, more generally, time-varying) heterogeneity between clusters on a time-to-event endpoint, we therefore propose to model the logarithm of the estimated cumulative hazard as a linear mixed effects model with time-varying random effects.

In Section 2, we introduce the time-varying frailty model and we show how to estimate the fixed effects and how to predict the time-varying random effects using the mixed model methodology. The way the proposed model can be used to test for declining variability between clusters is illustrated in Section 3 for the CML data. In Section 4, we run simulations to assess the performance of the method. Concluding remarks and given in Section 5.

2 Time-varying frailty model

We consider the following data structure. Observations are partitioned into s clusters, and each cluster is divided into two groups by a dichotomous covariate x (e.g., a treatment indicator or a prognostic biomarker whose value is used to partition the population into two risk groups). Let $h_{ij}(t)$ denote the hazard rate at time t for individual j of cluster i ($j = 1, \ldots, n_i$; $i = 1, \ldots, s$). The frailty model is defined as

$$h_{ij}(t) = h_0(t) \exp(w_i + x_{ij}\beta) \tag{1}$$

where $h_0(\cdot)$ is a baseline hazard function, x_{ij} is the value of the covariate (0 or 1), β is the unknown fixed effect parameter, and w_i is the log-frailty,

hereafter called random effect, of cluster i. An alternative formulation for model (1) in terms of the cumulative hazard function $H_{ij}(t) = \int_0^t h_{ij}(v)\,\mathrm{d}v$ is

$$H_{ij}(t) = H_0(t)\exp(w_i + x_{ij}\beta)$$

with $H_0(t) = \int_0^t h_0(v)\,\mathrm{d}v$. Since we use mixed models, a convenient choice for the distribution of the w_i's is the normal distribution with mean 0 and variance θ.

The extension that we propose is of form

$$H_{ij}(t) = H_0(t)\exp(w_i(t) + x_{ij}\beta) \tag{2a}$$

That is, the w_i's are time-dependent at the cumulative hazard level. Using a logarithmic transformation, we have

$$\log\left(H_{ij}(t)\right) = \log\left(H_0(t)\right) + w_i(t) + x_{ij}\beta \tag{2b}$$

Model (2b) provides the bridge between the survival model (2a) and the linear mixed effects model (see Section 2.1).

2.1 Towards the linear mixed effects model

Let T_{ij} be the event time of individual j from cluster i, possibly right-censored by a non-negative random variable C_{ij}, where C_{ij} is independent of T_{ij} ($j = 1, \ldots, n_i$; $i = 1, \ldots, s$). For each i and j, the random variables that we observe are $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\Delta_{ij} = I(T_{ij} \leq C_{ij})$, together with the covariate $x_{ij} \in \{0, 1\}$. We denote by $H_i^{(k)}(\cdot)$ the cumulative hazard function common to all individuals of cluster i with $x_{ij} = k$ ($k = 0, 1$).

To arrive at a linear mixed effects model, we need "pseudo-responses" which will serve as responses in equation (2b). In group k of cluster i, the pseudo-responses $\hat{\phi}_{ik,\ell}$ are obtained by estimating $\phi_{ik,\ell} \coloneqq \log(H_i^{(k)}(t_\ell))$ on a fixed grid of time points t_ℓ; $\ell = 1, \ldots, L$. Using the data from cluster i, $\{(y_{ij}, \delta_{ij}, x_{ij}) \,|\, j = 1, \ldots, n_i\}$, we take $\hat{\phi}_{ik,\ell} = \log(\hat{H}_i^{(0)}(t_\ell)\exp(k\hat{\gamma}_i))$, with $\hat{H}_i^{(0)}(\cdot)$ an estimator of $H_i^{(0)}(\cdot)$ and $\hat{\gamma}_i$ the fixed effect parameter estimate obtained from the Cox model fitted to the data of cluster i. For $\hat{H}_i^{(0)}(\cdot)$, we propose a modification of the Breslow cumulative baseline hazard curve using a parametric fit to estimate $H_i^{(0)}(t_\ell)$ for t_ℓ below the first or above the last event time of cluster i (Moeschberger & Klein, 1985); cf. Appendix A.

In terms of the pseudo-responses, we have

$$\hat{\phi}_{ik,\ell} = \beta_{0,\ell} + w_{i,\ell} + k\beta + e_{ik,\ell} \tag{3}$$

with $\beta_{0,\ell} \coloneqq \log(H_0(t_\ell))$, $w_{i,\ell} \coloneqq w_i(t_\ell)$, and $e_{ik,\ell} \coloneqq \hat{\phi}_{ik,\ell} - \phi_{ik,\ell}$. Model (3) is a linear mixed effects model with the following assumptions:

$$\begin{aligned}
&\boldsymbol{w} = (w_{1,1} \ldots w_{1,L} \ldots \ldots w_{s,1} \ldots w_{s,L})' \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}) \\
&\boldsymbol{e} = (e_{10,1}\, e_{11,1} \ldots e_{10,L}\, e_{11,L} \ldots \ldots e_{s0,1}\, e_{s1,1} \ldots e_{s0,L}\, e_{s1,L})' \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{R}) \\
&\mathrm{Cov}(\boldsymbol{w}, \boldsymbol{e}) = \boldsymbol{0}
\end{aligned}$$

where the forms of the $\boldsymbol{G}$ and $\boldsymbol{R}$ matrices are specified in Section 2.2 and in Section 2.3, respectively. For more details on the model structure, we refer the reader to Appendix B where we give the matrix formulation for model (3).

2.2 **Random effects covariance structure**

Random effects pertaining to different clusters are assumed to be independent so that

$$\text{Cov}(w_{i,\ell}, w_{i',\ell'}) = 0 \quad \text{if} \quad i \neq i'$$

Consequently, the $\boldsymbol{G}$ matrix has a block diagonal structure, with s blocks of size $L \times L$. We further assume that the blocks are identical. The common block matrix specifies the covariance structure between the L random effects within the same cluster. Many different choices can be made (Littell et al., 2000). When the t_ℓ's are equally spaced, a popular choice is the first-order autoregressive structure (AR(1)), i.e.,

$$\text{Cov}(w_{i,\ell}, w_{i,\ell'}) = \theta\, \rho^{|\ell-\ell'|} \qquad (\theta > 0,\, 0 < \rho < 1)$$

In words, the between-cluster variation is the same at any point ($\text{Var}(w_{i,\ell}) = \theta$ for $\ell = 1, \ldots, L$) while association between pairs of random effects within the same cluster declines with increasing distance in time ($\text{Cor}(w_{i,\ell}, w_{i,\ell'}) = \rho^{|\ell-\ell'|}$). However, the fact that the variance remains constant over time makes it impossible to assess whether the cluster-to-cluster variability is changing over time. For that purpose, an autoregressive structure with heterogeneous variances (ARH(1)) is more appropriate, i.e.,

$$\text{Cov}(w_{i,\ell}, w_{i,\ell'}) = \sqrt{\theta_\ell}\,\sqrt{\theta_{\ell'}}\,\rho^{|\ell-\ell'|}$$

2.3 **Residual covariance structure**

In model (3), the error term $e_{ik,\ell}$ adjusts for the fact that we have substituted an estimate for $\phi_{ik,\ell}$. The $\boldsymbol{R}$ matrix specifies the covariance structure of the error terms. By construction, however, this information cannot be recovered from the pseudo-data (the latter do not contain any replicate for given i, k, and ℓ). In addition to the pseudo-responses, therefore, the raw data also have to provide an estimate of $\boldsymbol{R}$. This can be done by bootstrap resampling.

The fact that $\hat{\phi}_{i0,\ell}$ and $\hat{\phi}_{i1,\ell}$ ($\ell = 1, \ldots, L$) are constructed using only the data from cluster i results in a block diagonal structure for $\boldsymbol{R}$, with s blocks of size $2L \times 2L$. In addition, block i differs from block i' because clusters are generally not identical in terms of sample size and event rate. We obtain an estimate of $\text{Cov}(e_{ik,\ell}, e_{ik',\ell'})$ as follows. First, we draw B bootstrap samples from cluster i of the original data (cf. Appendix C). For each bootstrap sample, we calculate the pseudo-responses $\hat{\phi}^\star_{ik,\ell}$ and $\hat{\phi}^\star_{ik',\ell'}$. An estimate of $\text{Cov}(e_{ik,\ell}, e_{ik',\ell'})$ then follows from the empirical covariance between the $\hat{\phi}^\star_{ik,\ell}$'s and the $\hat{\phi}^\star_{ik',\ell'}$'s.

2.4 Fitting the model

Model (3) can be fitted in SAS by means of `proc mixed` where the `parms` statement is used to fix the residual variance components. A SAS program (including R code via `proc iml`) implementing the procedure is available as supplementary material.

2.5 Testing for decreasing heterogeneity

To see whether "a time-constant cluster-to-cluster variability" is a reasonable model assumption, we can test the null hypothesis

$$\mathrm{H}_0\colon\ \theta_1 = \cdots = \theta_L$$

against the alternative

$$\mathrm{H}_1\colon\ \exists \ell, \ell' \in \{1, \ldots, L\} : \theta_\ell \neq \theta_{\ell'}$$

Under H_0, each block of $\boldsymbol{G}$ has a homogeneous AR(1) structure. The AR(1) structure has 2 parameters (θ and ρ) as compared to $L + 1$ parameters ($\theta_1, \ldots, \theta_L$, and ρ) for the ARH(1) structure. Model comparison can be done via a (restricted) likelihood ratio test (Littell et al., 2006, Section A1.6).

For the CML study considered in Section 3, however, we are mostly interested in decreasing heterogeneity among clusters. In that case, we rather consider the ordered alternative

$$\mathrm{H}_1\colon\ \theta_1 > \cdots > \theta_L$$

To account for the ordering specified under H_1, we rely on the ordered heterogeneity family of tests (Rice & Gaines, 1994a,b). An ordered heterogeneity test can be used to convert almost any non-directional test into a directional one when a specific ordered test is not available or is though to implement. The test statistic consists of combining a measure of evidence against H_0 with the independent ordering information specified under H_1. In our case, the test statistic becomes $T_O = r_\mathrm{S}(1 - p_\mathrm{LRT})$, with p_LRT the p-value obtained from the (non-directional) likelihood ratio test, and r_S the Spearman correlation between $\{\theta_1, \ldots, \theta_L\}$ and $\{\hat{\theta}_1, \ldots, \hat{\theta}_L\}$. The test statistic T_O becomes increasingly large as the data increasingly refute the null hypothesis in the direction of the alternative hypothesis (Rice & Gaines, 1994a). Critical values are tabulated in Rice & Gaines (1994b).

3 Example

We consider the CML data used in Wintrebert et al. (2004). The data consists of 4177 CML patients treated at 215 transplant centres among which

1610 deaths ($\approx$ 40%) were recorded. To obtain good pseudo-responses, sufficient data information is required within each cluster. In the CML data, however, many centres are quite small (number of patients per centre: min = 1, 1$^{\text{st}}$ quartile = 3, median = 10, mean = 19.43, 3$^{\text{rd}}$ quartile = 23.5, max = 280) and substantial data information is censored ($\approx$ 60%). We thus drop the centres that do not contribute much information (< 50 patients). In total, our analysis is based on 1767 CML patients from 19 centres among which 706 deaths were recorded (censoring rate $\approx$ 60%).

We account for the effects of five known risk factors (patient age, disease stage, time interval from diagnosis to transplant, donor type, and donor-recipient sex combination). To avoid the "curse of dimensionality", we incorporate the covariate information into the EBMT risk score (Gratwohl, 2012), a validated prognostic index ranging from 0 to 7 points, from which we identify a "low-risk group" (EBMT risk score $= 0, 1, 2, 3$) and a "high-risk group" (EBMT risk score $= 4, 5, 6, 7$). Table 1 shows the repartition of the patients in the CML data. Fitting the standard frailty model (1), we find a hazard ratio of $\widehat{\text{HR}} = \exp(\hat{\beta}) = 2.140$ (95% CI: $[1.827, 2.507]$) and $\hat{\theta} = 0.025$.

To investigate whether the centre-to-centre variability decreases over follow-up time, we fit model (3). Pseudo-responses are calculated on a grid of $L = 5$ time points equally spaced between $t_1 = 1$ month and $t_5 = 9$ months, which is approximately the time it takes for a patient to recover and to produce normal blood cell levels (The Leukemia & Lymphoma Society, 2013). To obtain an estimate of the residual covariance matrix, we draw $B = 500$ bootstrap samples (cf. Section 2.3).

Regarding the fixed effect obtained from model (3), we find $\widehat{\text{HR}} = 2.141$ (95% CI: $[1.726, 2.656]$), similar to what we have obtained from the standard frailty model.

Figure 1 shows the empirical best linear unbiased predictions (EBLUP's) of the random effects under the ARH(1) specification for the $\boldsymbol{G}$ matrix (cf. Section 2.2). The values of the variance at the different time points (i.e. the diagonal elements of each block of $\boldsymbol{G}$) are given at the bottom of Figure 1.

The likelihood ratio statistic (AR(1) versus ARH(1)) equals 15.907. The asymptotic sampling distribution under H_0 is a chi-squared distribution with $L - 1 = 4$ degrees of freedom so that $p_{\text{LRT}} = 0.003$. The correlation coefficient between the observed ordering $(5, 4, 3, 2, 1)$ and the expected ordering $(5, 4, 3, 2, 1)$ equals $r_{\text{S}} = 1$. Thus, $T_O = 0.997$. For $L = 5$, the critical region C at the 5% significance level is $C = \{T_O > 0.509\}$. We therefore reject H_0 and we conclude that the data display declining centre-to-centre variability. Figure 1 indicates that there is heterogeneity between centres during the first few months following transplantation, but not much afterwards.

Table 1 – Repartition of the patients in the CML data set. In parentheses are the number of events.

centre	low-risk $x = 0$	high-risk $x = 1$	total
1	31 (8)	19 (7)	50 (15)
2	41 (13)	11 (3)	52 (16)
3	52 (12)	6 (6)	58 (18)
4	48 (15)	11 (4)	59 (19)
5	53 (9)	7 (12)	60 (21)
6	49 (17)	12 (6)	61 (23)
7	52 (13)	10 (11)	62 (24)
8	53 (16)	14 (8)	67 (24)
9	52 (16)	16 (9)	68 (25)
10	41 (23)	27 (3)	68 (26)
11	55 (20)	15 (8)	70 (28)
12	60 (20)	12 (8)	72 (28)
13	51 (21)	21 (9)	72 (30)
14	66 (16)	7 (16)	73 (32)
15	67 (27)	14 (9)	81 (36)
16	85 (29)	21 (10)	106 (39)
17	121 (42)	64 (41)	185 (83)
18	145 (61)	78 (32)	223 (93)
19	206 (82)	74 (44)	280 (126)
total	1328 (460)	439 (246)	1767 (706)

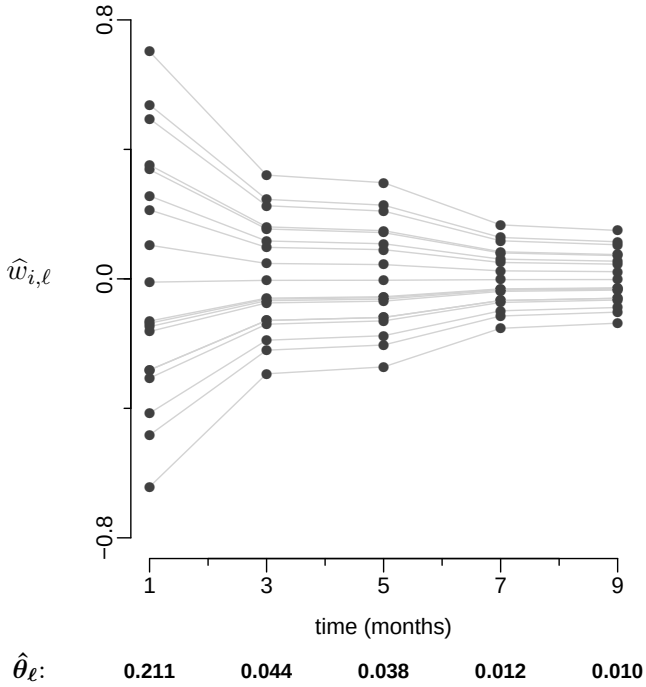


Figure 1 – Empirical best linear unbiased predictions (EBLUP's) as a function of time. The evolution of the variance over time is shown below the x-axis.

4 Simulation study

To provide further insight into the method, we present a small simulation study that we have conducted in a context close to the CML study.

4.1 Data generation

We consider $s = 19$ clusters with the same number of observations per cluster as in the CML data (cf. Table 1). Observations within a cluster are randomly allocated to two groups with an allocation ratio of 3:1 to reflect the structure of Table 1. Two scenarios are examined. In the first scenario (time-constant frailties), event times are generated from model (1). In the second scenario (time-varying frailties), event times are generated from model (2a). In both cases, censoring times are generated from an exponential distribution whose rate parameter is chosen to control the amount of censoring in the data, either 60% (like in the CML data) or 20%. We take $\beta = \log(2)$, and we use an exponential distribution at baseline (so that $h_0(t) = \lambda$ for all $t > 0$) with a low event rate of $\lambda = 0.05$.

4.1.1 Scenario 1 (time-constant frailties)

To generate t_{ij}, the j^{th} event time of cluster i, from model (1), we use the fact that, given w_i randomly drawn from a $\mathrm{N}(0, \theta)$ distribution, t_{ij} has an exponential distribution with rate $\lambda \exp(w_i + x_{ij}\beta)$. We take $\theta = 0.22$ to mimic the initial between-centre variability in the CML data (cf. Figure 1).

4.1.2 Scenario 2 (time-varying frailties)

To generate the j^{th} event time of cluster i from model (2a), we use the fact that $S_{ij}(T) \sim \mathrm{U}(0, 1)$, with

$$\begin{aligned} S_{ij}(t) &= \exp\{-H_{ij}(t)\} \\ &= \exp\{-\exp(\log(\lambda t) + w_i(t) + x_{ij}\beta)\} \end{aligned}$$

and we solve $S_{ij}(t_{ij}) = u$ for t_{ij}, where u is a uniform variate.

The specific time-dependency in the random effects that we consider is depicted in Figure 2. The time-dependency dies out in time ($w_i(t) \to 0$ as $t \to \infty$). The actual value of the random effect in cluster i at time $t = 0$, $w_i(0)$, is sampled from a $\mathrm{N}(0, \theta)$ distribution. We take $\theta = 0.22$, as above. Given the random start, the way the time-dependency dies out is deterministic. For details, we refer the reader to the web-based supplementary material where we explain how Figure 2 is obtained.

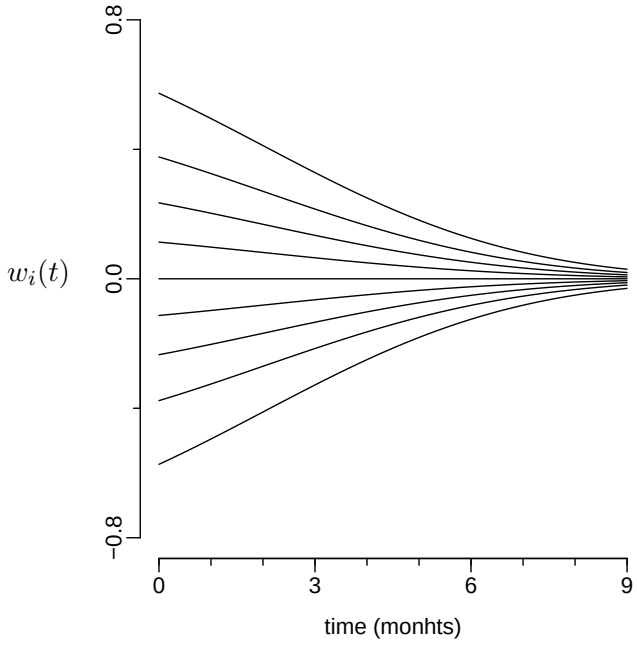


Figure 2 – Declining heterogeneity between clusters as used in the simulation study. In this picture, the initial values correspond to the deciles of the $\mathrm{N}(0, \theta)$ distribution. In the simulations, the initial values are randomly drawn from that distribution.

4.2 Results

For each of 500 simulated data sets, we fitted model (3) at $t_\ell = 1, 3, 5, 7, 9$ with both the AR(1) and the ARH(1) specifications for the $\boldsymbol{G}$ matrix. For the latter, we recorded the EBLUP's of the random effects along with the estimated variances over time. Figure 3 displays averages at each time point for the 60% censoring case (the results are similar in the 20% censoring case). We also obtained the estimates of the fixed effect parameter β.

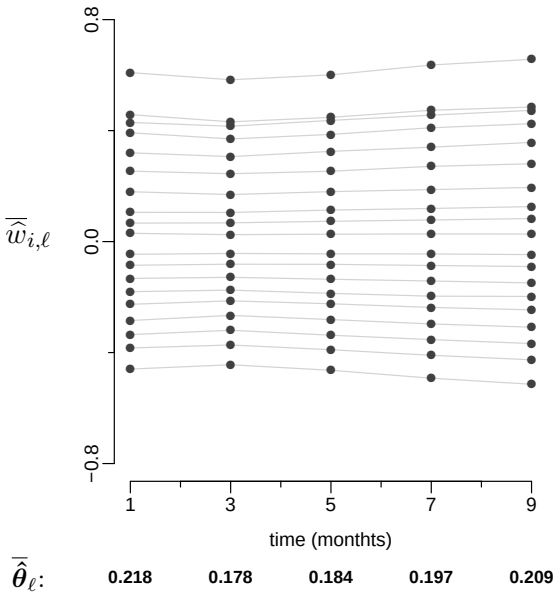
In the first scenario (time-constant frailties), the predictions of the random effects do not, on average, show any systematic trend in time (cf. Figure 3a). The means of the estimated variances fluctuate around 0.197 (resp. 0.209 in the 20% censoring case), and the estimates of β have a mean value of 0.699 (resp. 0.691) and a standard deviation of 0.080 (resp. 0.061). Regarding the ordered heterogeneity test, we have observed, in the context of these simulations, slow convergence of the likelihood ratio statistic to the limiting chi-squared distribution. For this reason, we have determined the null sampling distribution using supplementary data sets. Based on the empirical distribution, the rejection rate in the 500 simulated data sets for the ordered heterogeneity test under the null hypothesis of constant heterogeneity, i.e. the size of the test, corresponds to 0.064 (resp. 0.041).

In the second scenario (time-varying frailties), the predictions of the random effects do follow, on average, the decreasing pattern of Figure 2 (cf. Figure 3b). The estimates of β have a mean value of 0.694 (resp. 0.695) and a standard deviation of 0.081 (resp. 0.062). The rejection rate in the 500 simulated data sets for the ordered heterogeneity test under the alternative hypothesis of scenario 2, i.e. the power of the test, corresponds to 0.740 (resp. 0.840).

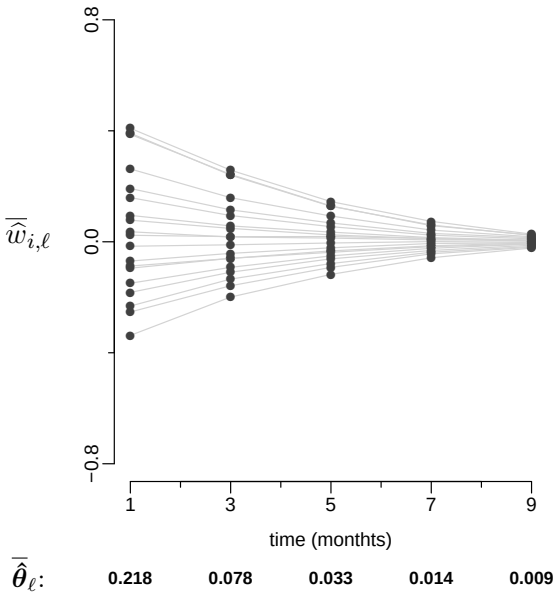
5 Discussion

We proposed a method to test for decreasing heterogeneity in clustered survival data by extending the frailty model to a time-varying frailty model. Starting from the cumulative hazard representation, we embedded the frailty model into the linear mixed model world via a logarithmic transformation and we derived pseudo-responses for the linear mixed effects model by estimating the logarithm of the cumulative hazard on a grid of time points.

The proposed method depends on a wise choice of the time points at which the pseudo-responses are calculated. Often, the grid can be chosen based on clinical or biological grounds. Further, note that, within each cluster, sufficient data information is needed for accurate estimation of the pseudo-responses. The slow convergence rate of the likelihood ratio statistic that we observed in the simulations (where the grid was the same for each



(a) scenario 1: time-constant frailties



(b) scenario 2: time-varying frailties

Figure 3 – Simulation results in the 60% censoring case.

simulated data set) appears to be related to the latter point.

We illustrated the method with the same CML data as in Wintrebert et al. (2004) and we confirm the conclusion that heterogeneity between transplant centres declines over time. Note that the way in which we extend the frailty model is different from extensions given in Wintrebert et al. (2004) where the random effects are time-dependent at the hazard level. Compared to the latter, the method we proposed is very flexible in specifying the correlation structure between the random effects and it avoids complex likelihood functions that are difficult to maximise.

Appendices

A

A hybrid estimator of the cumulative baseline hazard function

Given a sample of independent survival data (up to a vector of measured covariates $\boldsymbol{x}$), the Breslow estimator of the cumulative baseline hazard function is defined by

$$\hat{H}_{0B}(t) = \sum_{\tilde{y}_\ell \leq t} \frac{d_\ell}{\sum_{j \in R(\tilde{y}_\ell)} \exp(\boldsymbol{x}_j' \hat{\boldsymbol{\beta}})}$$

with $\tilde{y}_1 < \cdots < \tilde{y}_r$ the ordered distinct event times, d_ℓ the number of events at time $\tilde{y}_\ell$, and $R(\tilde{y}_\ell)$ the set containing those individuals still under observation just prior to $\tilde{y}_\ell$. This estimator suffers from two limitations: (i) it returns zero below $\tilde{y}_1$ (regarding the logarithmic transformation, this is a problem), and (ii) it remains constant beyond $\tilde{y}_r$ (estimates are thus not reliable in the right tail). We therefore propose a hybrid estimator by completing $\hat{H}_{0B}(\cdot)$ using a parametric distribution to estimate the tails, namely,

$$\hat{H}_0(t) = \begin{cases} \hat{\lambda} t^{\hat{\rho}} & \text{if } t < \tilde{y}_1 \\ \hat{H}_{0B}(t) & \text{if } \tilde{y}_1 \leq t \leq \tilde{y}_r \\ \hat{H}_{0B}(\tilde{y}_r) + \hat{\lambda}\left(t^{\hat{\rho}} - \tilde{y}_r^{\hat{\rho}}\right) & \text{if } t > \tilde{y}_r \end{cases}$$

with $\hat{\lambda}$ and $\hat{\rho}$ obtained by fitting a Weibull distribution to the data where $\boldsymbol{x} = \boldsymbol{0}$ under the constraint that $\hat{\lambda}\tilde{y}_1^{\hat{\rho}} = \hat{H}_{0B}(\tilde{y}_1)$.

B

Model (3) in matrix notation

In matrix form, the linear mixed effects model is written as

$$\left| \begin{array}{l} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{w} + \boldsymbol{e} \\ \boldsymbol{w} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}) \\ \boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{R}) \\ \mathrm{Cov}(\boldsymbol{w}, \boldsymbol{e}) = \boldsymbol{0} \end{array} \right.$$

where $\boldsymbol{Y}$ denotes the vector of responses, $\boldsymbol{\beta}$ and $\boldsymbol{w}$ are respectively the vectors of fixed effects and random effects parameters, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the corresponding design matrices, and $\boldsymbol{e}$ is the vector of random errors.

Below is the matrix form of model (3) for the particular case where $s = 3$

and $L = 2$:

$$\begin{pmatrix} \hat{\phi}_{10,1} \\ \hat{\phi}_{11,1} \\ \hat{\phi}_{10,2} \\ \hat{\phi}_{11,2} \\ \hdashline \hat{\phi}_{20,1} \\ \hat{\phi}_{21,1} \\ \hat{\phi}_{20,2} \\ \hat{\phi}_{21,2} \\ \hdashline \hat{\phi}_{30,1} \\ \hat{\phi}_{31,1} \\ \hat{\phi}_{30,2} \\ \hat{\phi}_{31,2} \end{pmatrix} = \left(\begin{array}{cc:c} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ \hdashline 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ \hdashline 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{array}\right) \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \\ \hdashline \beta \end{pmatrix} + \left(\begin{array}{cccc:cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \begin{pmatrix} w_{1,1} \\ w_{1,2} \\ \hdashline w_{2,1} \\ w_{2,2} \\ \hdashline w_{3,1} \\ w_{3,2} \end{pmatrix} + \begin{pmatrix} e_{10,1} \\ e_{11,1} \\ e_{10,2} \\ e_{11,2} \\ \hdashline e_{20,1} \\ e_{21,1} \\ e_{20,2} \\ e_{21,2} \\ \hdashline e_{30,1} \\ e_{31,1} \\ e_{30,2} \\ e_{31,2} \end{pmatrix}$$

$$\underbrace{\phantom{\hat{\phi}_{10,1}\\ \hat{\phi}_{11,1}\\ \hat{\phi}_{10,2}\\ \hat{\phi}_{11,2}\\ \hat{\phi}_{20,1}\\ \hat{\phi}_{21,1}\\ \hat{\phi}_{20,2}\\ \hat{\phi}_{21,2}\\ \hat{\phi}_{30,1}\\ \hat{\phi}_{31,1}\\ \hat{\phi}_{30,2}\\ \hat{\phi}_{31,2}}}_{\boldsymbol{Y}} = \underbrace{}_{\boldsymbol{X}} \underbrace{\phantom{\beta_{0,1}\\ \beta_{0,2}\\ \beta}}_{\boldsymbol{\beta}} + \underbrace{}_{\boldsymbol{Z}} \underbrace{\phantom{w_{1,1}\\ w_{1,2}\\ w_{2,1}\\ w_{2,2}\\ w_{3,1}\\ w_{3,2}}}_{\boldsymbol{w}} + \underbrace{\phantom{e_{10,1}\\ e_{11,1}\\ e_{10,2}\\ e_{11,2}\\ e_{20,1}\\ e_{21,1}\\ e_{20,2}\\ e_{21,2}\\ e_{30,1}\\ e_{31,1}\\ e_{30,2}\\ e_{31,2}}}_{\boldsymbol{e}}$$

In general, we have

$$\boldsymbol{X} = \left(\begin{array}{c:c} \mathbb{1}_s \otimes \boldsymbol{I}_L \otimes \mathbb{1}_2 & \mathbb{1}_s \otimes \mathbb{1}_L \otimes \binom{0}{1} \end{array}\right)$$

and

$$\boldsymbol{Z} = \boldsymbol{I}_s \otimes \boldsymbol{I}_L \otimes \mathbb{1}_2$$

with $\boldsymbol{I}_n$ the $n \times n$ identity matrix and $\mathbb{1}_n$ the vector of length n with all entries equal to 1.

Continuing the particular case where $s = 3$ and $L = 2$, $\boldsymbol{G} = \mathrm{Var}(\boldsymbol{w})$ and $\boldsymbol{R} = \mathrm{Var}(\boldsymbol{e}) = \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{w})$ are block diagonal matrices,

$$\boldsymbol{G} = \mathrm{diag}(\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3) \quad \text{and} \quad \boldsymbol{R} = \mathrm{diag}(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3)$$

with $\boldsymbol{G}_i$ and $\boldsymbol{R}_i$ the symmetric matrices

$$\boldsymbol{G}_i = \begin{pmatrix} \mathrm{Var}(w_{i,1}) & \mathrm{Cov}(w_{i,1}, w_{i,2}) \\ & \mathrm{Var}(w_{i,2}) \end{pmatrix}$$

and

$$\boldsymbol{R}_i = \begin{pmatrix} \mathrm{Var}(e_{i0,1}) & \mathrm{Cov}(e_{i0,1}, e_{i1,1}) & \mathrm{Cov}(e_{i0,1}, e_{i0,2}) & \mathrm{Cov}(e_{i0,1}, e_{i1,2}) \\ & \mathrm{Var}(e_{i1,1}) & \mathrm{Cov}(e_{i1,1}, e_{i0,2}) & \mathrm{Cov}(e_{i1,1}, e_{i1,2}) \\ & & \mathrm{Var}(e_{i0,2}) & \mathrm{Cov}(e_{i0,2}, e_{i1,2}) \\ & & & \mathrm{Var}(e_{i1,2}) \end{pmatrix}$$

The marginal variance of the pseudo-responses, given by $\boldsymbol{ZGZ}' + \boldsymbol{R}$, is thus a block diagonal matrix whose i^{th} block is given by

$$\operatorname{Var}\begin{pmatrix}\hat{\phi}_{i0,1}\\ \hat{\phi}_{i1,1}\\ \hat{\phi}_{i0,2}\\ \hat{\phi}_{i1,2}\end{pmatrix} = \begin{pmatrix}\theta_1 & \theta_1 & \sqrt{\theta_1}\sqrt{\theta_2}\rho & \sqrt{\theta_1}\sqrt{\theta_2}\rho\\ \theta_1 & \theta_1 & \sqrt{\theta_1}\sqrt{\theta_2}\rho & \sqrt{\theta_1}\sqrt{\theta_2}\rho\\ \sqrt{\theta_1}\sqrt{\theta_2}\rho & \sqrt{\theta_1}\sqrt{\theta_2}\rho & \theta_2 & \theta_2\\ \sqrt{\theta_1}\sqrt{\theta_2}\rho & \sqrt{\theta_1}\sqrt{\theta_2}\rho & \theta_2 & \theta_2\end{pmatrix} + \boldsymbol{R}_i$$

C

Bootstrap for right-censored survival data

Given a sample of independent survival data (up to a vector of measured covariates $\boldsymbol{x}$),

$$\{(y_j, \delta_j, \boldsymbol{x}_j) \,|\, j = 1, \ldots, n\}$$

where y_j is the time to event or censoring, whichever comes first, and δ_j indicates whether y_j is observed ($\delta_j = 1$) or censored ($\delta_j = 0$), a bootstrap sample

$$\{(y_j^\star, \delta_j^\star, \boldsymbol{x}_j) \,|\, j = 1, \ldots, n\}$$

can be drawn as follows (Davison & Hinkley, 1997, Algorithm 7.2, page 351).

For $j = 1, \ldots, n$,

1. Generate $t_j^\star$ from the estimated event time survival function given by $\hat{S}_0(t)^{\exp(\boldsymbol{x}_j'\hat{\boldsymbol{\beta}})}$.
2. If $\delta_j = 0$, set $c_j^\star = y_j$; otherwise, generate $c_j^\star$ from an estimate of the censoring time survival function conditional on $C_j > y_j$, i.e. from $\hat{G}(\cdot)/\hat{G}(y_j)$ with $\hat{G}(\cdot)$ a Kaplan-Meier estimator of the censoring time survival function.
3. Set $y_j^\star = \min(t_j^\star, c_j^\star)$ and $\delta_j^\star = I(t_j^\star \leq c_j^\star)$.

D

Time-varying random effects generation

For data generation purposes, we assume that the differential equation

$$\left|\begin{array}{l}\dfrac{\mathrm{d}w_i(t)}{\mathrm{d}t} = k_{1i}\exp\left\{-\left(\dfrac{t-k_2}{k_3}\right)^2\right\}\\ w_i(0) = w_{i0}\end{array}\right. \tag{4}$$

describes the evolution in time of $w_i(t)$, where k_{1i} is a parameter that governs the increasing/decreasing rate, and where k_2, k_3 are two additional tuning constants related to "where variation takes place and how long it lasts". The solution of (4) is given by

$$w_i(t) = w_{0i} + k_{1i}k_3\sqrt{\pi}\left[\Phi\left(\sqrt{2}\,\frac{t-k_2}{k_3}\right) + \Phi\left(\sqrt{2}\,\frac{k_2}{k_3}\right) - 1\right]$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable. Taking

$$k_{1i} = -\frac{w_{0i}}{k_3\sqrt{\pi}\Phi\left(\sqrt{2}\,\frac{k_2}{k_3}\right)}$$

we obtain that the cluster-to-cluster variability dies out in time ($w_i(t) \to 0$ as $t \to \infty$).

We now need to ensure that, for this choice of $w_i(t)$,

$$H_{ij}(t) = H_0(t)\exp(w_i(t) + x_{ij}\beta)$$

does result in a cumulative hazard curve, i.e.,

$$\left|\begin{array}{ll} 1. & \lim_{t\to 0} H_{ij}(t) = 0 \\ 2. & \lim_{t\to\infty} H_{ij}(t) = \infty \\ 3. & \frac{\mathrm{d}H_{ij}(t)}{\mathrm{d}t} \geq 0 \quad \text{for all } t \end{array}\right.$$

The first two conditions are satisfied. Assuming a Weibull baseline hazard ($h_0(t) = \lambda\rho t^{\rho-1}$; $\lambda > 0$, $\rho > 0$), condition 3 can be rewritten as

$$\lambda t^{\rho}\exp(w_i(t) + x_{ij}\beta)\left(\frac{\rho}{t} + \frac{\mathrm{d}w_i(t)}{\mathrm{d}t}\right) \geq 0$$

from which it follows that we must have

$$\frac{\rho}{t} \geq -\frac{\mathrm{d}w_i(t)}{\mathrm{d}t} \tag{5}$$

for all $t > 0$ and $i \in \{1, \ldots, s\}$. If $w_i(t)$ is increasing, then constraint (5) always holds. Otherwise, this condition is most easily interpreted from model (2b),

$$\log\left(H_{ij}(t)\right) = \log\left(H_0(t)\right) + w_i(t) + x_{ij}\beta$$

In that case, constraint (5) means that the rate at which $\log(H_0(t))$ (whose derivative is $\frac{\rho}{t}$) increases must be larger than the rate at which $w_i(t)$ decreases in order for $\log(H_{ij}(t))$, or equivalently for $H_{ij}(t)$, to be increasing.

References

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Gratwohl, A. (2012). The EBMT risk score. *Bone Marrow Transplantation* **47**, 749–756.

Legrand, C., Duchateau, L., Sylvester, R., Janssen, P., van der Hage, J. A., van de Velde, C. J. & Therasse, P. (2006). Heterogeneity in disease free survival between centers: lessons learned from an EORTC breast cancer trial. *Clinical Trials* **3**, 10–18.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. & Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Publishing, 2nd ed.

Littell, R. C., Pendergast, J. & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* **19**, 1793–1819.

López-de-Ullibarri, I., Janssen, P. & Cao, R. (2012). Continuous covariate frailty models for censored and truncated clustered data. *Journal of Statistical Planning and Inference* **142**, 1864–1877.

Massonnet, G., Janssen, P. & Burzykowski, T. (2008). Fitting conditional survival models to meta-analytic data by using a transformation toward mixed-effects models. *Biometrics* **64**, 834–842.

Moeschberger, M. L. & Klein, J. P. (1985). A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics* **41**, 253–259.

Rice, W. R. & Gaines, S. D. (1994a). Extending nondirectional heterogeneity tests to evaluate simply ordered alternative hypotheses. *Proceedings of the National Academy of Sciences* **91**, 225–226.

Rice, W. R. & Gaines, S. D. (1994b). The ordered-heterogeneity family of tests. *Biometrics* **50**, 746–752.

Teodorescu, B., Van Keilegom, I. & Cao, R. (2010). Generalized time-dependent conditional linear models under left truncation and right censoring. *Annals of the Institute of Statistical Mathematics* **62**, 465–485.

The Leukemia & Lymphoma Society (2013). http://www.lls.org/resourcecenter/freeeducationmaterials/treatment/.

Wintrebert, C. M. A., Putter, H., Zwinderman, A. H. & van Houwelingen, H. C. (2004). Centre-effect on survival after bone marrow transplantation: Application of time-dependent frailty models. *Biometrical Journal* **46**, 512–525.