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

2013/36

Adjusting for centre differences in multicentre clinical trials: a simulation-based investigation for time-to-event outcomes

MUNDA, M. and C. LEGRAND

Adjusting for centre differences in multicentre clinical trials: a simulation-based investigation for time-to-event outcomes

Marco Munda* & Catherine Legrand

June 2013

Abstract

Multicentre clinical trials allow the required number of participants to be recruited within a fairly short time frame. Conducting a clinical trial at more than one centre also deepens the generalisability of its findings. But the other side of the coin is that this raises the issue of whether and how to adjust for centre effects in the statistical analysis. In this paper, we discuss a number of strategies for time-to-event outcomes. We balance the conventional fixed effects and stratified Cox models against the frailty modelling approach with the objective to provide a clear picture of the weaknesses of the current practice, and to encourage the use of frailty models. A special attention is paid to the problem of misspecification of the frailty distribution. Based on simulations, we illustrate the performances of the frailty model over its competitors, and we argue that it can generally be used advantageously to draw inferences about the treatment effect, even if the frailty distribution is misspecified.

Keywords: Multicenter clinical trial; Time-to-event outcome; Shared frailty model; Misspecification

1 Introduction

In a multicentre clinical trial, more than one centre participates in the study so that patients are recruited from a broad population within a fairly short time frame. The validity of such a multicentre study is based on an agreement between all participating centres to follow the same trial protocol.

*marco.munda@uclouvain.be

However, and despite all precautions taken to standardise the way the trial is conducted in each centre, it is generally agreed that inherent differences persist. The result is that the data to be analysed are clustered by design. In this paper, we consider the problem of adjusting for centre heterogeneity when assessing a treatment effect that we assume homogeneous across centres.

Even though the ICH guidance document "statistical principles for clinical trials" (ICH E9, `http://www.ich.org/products/guidelines.html`) clearly states that "*The main treatment effect may be investigated first using a model which allows for centre differences, but does not include a term for treatment-by-centre interaction.*", this is still rarely done in practice. In particular, for multicentre clinical trials with a time-to-event endpoint, recommendations on how to adjust for centre heterogeneity are still limited. Such type of endpoint is, however, frequently used in medical research. For example, time to death, or the progression-free survival (time from randomisation to disease progression or death from any cause) is common to assess the effect of new therapies in cancer studies.

In Andersen et al. (1999), it is argued that centre differences have to be taken into account to arrive at valid conclusions about the treatment effect. However, possible ways to do this are not discussed in detail. Probably the most natural option is to include additional fixed centre effects parameters. Alternatively, accounting for centre differences can also be achieved by stratification. The frailty model is another approach that has gained in popularity in recent years. It is essentially a proportional hazards model with a random effect, called frailty, that accounts for the residual centre to centre variability. In Section 2, we briefly review the basics of these modelling strategies. Statistical aspects are clearly discussed in Glidden & Vittinghoff (2004) where the authors found advantages in using the frailty model.

The frailty modelling approach requires choosing a probability distribution for the frailties. This is a difficult problem, which to date has only been sparsely addressed. Small simulation studies conducted in the context of multicentre clinical trials already suggest that inference on the treatment effect is not affected by misspecification of the frailty distribution (Pickles & Crouchley, 1995; O'Quigley & Stare, 2002; Glidden & Vittinghoff, 2004). However, we will alert that misspecification can sometimes be an issue, therefore putting existing findings in broader perspectives.

In this paper, we build on the work done in Glidden & Vittinghoff (2004). Our first objective is to put emphasis on the pros and the cons of each model. This is done in Section 3, with numerical illustrations. Our second objective, covered in Section 4, is to further investigate the performances of the frailty model over its competitors when the frailty distribution is misspecified. These issues are important for the applied statistician and our aim is to provide pragmatic guidelines. Finally, Section 5 contains some concluding remarks and further considerations.

2

Modelling clustered time-to-event data

2.1 The (unadjusted) Cox model in a nutshell

We start with non-clustered time-to-event data for which the observed information consists of

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

with $y_j = \min(t_j, c_j)$ the time to event or censoring, whichever occurs first, $\delta_j = I(t_j \leq c_j)$ the event indicator, and $\boldsymbol{x}_j$ a vector of covariates. Throughout this paper, we assume that the event times (the t_j's) and the censoring times (the c_j's) are independent given the covariate information (independent censoring), and that the censoring distribution has no common parameter with the event time distribution (non-informative censoring). These are standard working assumptions.

Let $h_j(t)$ denote the hazard rate for individual j at time t. Modelling non-clustered time-to-event data is most often done by assuming a proportional hazards structure,

$$h_j(t) = h_0(t) \exp(\boldsymbol{x}_j' \boldsymbol{\beta}) \tag{1}$$

with $h_0(\cdot)$ a baseline hazard function and $\boldsymbol{\beta}$ a vector of fixed effects parameters. Following Cox (1972), the tradition is to treat $h_0(\cdot)$ as a nuisance function. That is, $h_0(\cdot)$ is left unspecified, leading to the well-known Cox proportional hazards model. An estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is then obtained by maximising a partial likelihood given by (assuming no ties between event times)

$$L(\boldsymbol{\beta}; \boldsymbol{z}) = \prod_{j=1}^{N} \left(\frac{\exp(\boldsymbol{x}_j' \boldsymbol{\beta})}{\sum_{\ell \in R(y_j)} \exp(\boldsymbol{x}_\ell' \boldsymbol{\beta})} \right)^{\delta_j}$$

with $R(y_j)$ the risk set at time y_j containing all individuals still susceptible to the event at y_j. Estimates of variances can be found on the diagonal of the negative second derivative of $\log(L(\boldsymbol{\beta}; \boldsymbol{z}))$ evaluated at $\hat{\boldsymbol{\beta}}$. Even though L is not a genuine likelihood, it has been shown in Gill (1984) that consistency and asymptotic normality properties for the underlying estimator of $\boldsymbol{\beta}$ are preserved.

2.2 Adjusting for centre differences

With model (1), the population under study is regarded as homogeneous in terms of survival, except for measured covariates. Typically, though, this does not apply to a multicentre clinical trial where there is likely to be residual heterogeneity across centres. This is because patients from the

same centre have a number of risk factors in common like, for example, centre type (e.g., general, specialised, or university hospitals) and catchment area, or medical staff and facilities. To account for this, centre effects must somehow be included in the statistical model used for the analysis.

The fixed effects approach Centre effects can enter model (1) as additional fixed effects parameters,

$$h_{ij}(t) = h_0(t) \exp(\boldsymbol{c}_i'\boldsymbol{\alpha} + \boldsymbol{x}_{ij}'\boldsymbol{\beta}) \tag{2}$$

where we now use two indices, $i \in \{1, \ldots, s\}$ for the centre and $j \in \{1, \ldots, n_i\}$ for the n_i patients in centre i, to reflect the hierarchical structure of the data (the vector of observations $\boldsymbol{z}$ is changed accordingly). In model (2), $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{s-1})'$ contains the fixed centre effects, and $\boldsymbol{c}_i = (c_{i,1}, \ldots, c_{i,s-1})'$ denotes the all-zeros vector except for a 1 in the i^{th} position. The last centre does not need an indicator because we know that an observation belongs to that centre when $c_{i,1} = \cdots = c_{i,s-1} = 0$. If we had included an additional indicator for the last centre, then the model would have been overparametrised. One may note that choosing one particular centre (here, the last one) as reference is consistent with the interpretation of $h_0(\cdot)$ as being the hazard rate for patients with covariate values all equal to zero. However, this choice is arbitrary and any centre can play the role of the reference centre.

The stratified approach Instead of entering the centre variable in model (1) as additional fixed effects parameters, the baseline hazard can rather be stratified on that variable to indicate that different subpopulations are exposed to different baseline risks, i.e.,

$$h_{ij}(t) = h_{0i}(t) \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \tag{3}$$

where $h_{01}(\cdot), \ldots, h_{0s}(\cdot)$ are s unspecified and unrelated baseline hazard functions. The partial likelihood principle is readily adapted by multiplying the partial likelihoods specific to each stratum (Glidden & Vittinghoff, 2004).

The frailty modelling approach Participating centres could also be viewed as one possible sample from a broader population of centres. In that case, centre i has a random effect on the hazard rate, which we call frailty and give the notation u_i; $i = 1, \ldots, s$. The (shared) frailty model is then defined as (Legrand et al., 2006; Duchateau & Janssen, 2008; Wienke, 2010)

$$h_{ij}(t) = h_0(t) u_i \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \tag{4}$$

The repartition of the unobserved u_i's is described by an assumed probability distribution, called the frailty distribution. The most customary choice is

the one-parameter gamma. We refer to Cortiñas Abrahantes et al. (2007) for a detailed overview of several model estimation procedures which have to take into account the latent nature of the frailties. Some of them are readily available in standard software (see Hirsch & Wienke (2012), Munda et al. (2012), and Rondeau et al. (2012)).

3 Pros and cons of the various modelling approaches

In this section, we discuss the strengths and the weaknesses of models (2)–(4) to adjust for centre differences and we illustrate this discussion via simulations. For the sake of completeness, model (1) is also considered. The issue of misspecifying the frailty distribution in model (4) will be specifically addressed in the next section.

3.1 Simulation settings

We consider two opposite situations—6 centres of size 48, and 48 centres of size 6—as well as an intermediate situation—8 centres of size 18 plus 24 centres of size 6—thus keeping the total number of patients constant ($N = 288$). We mimic a 1:1 (resp. 2:1) allocation ratio in each centre by randomly selecting $N/2$ (resp. $2N/3$) patients for the treatment arm ($x = 1$) and the other $N/2$ (resp. $N/3$) for the control arm ($x = 0$).

Frailties $u_1, \ldots, u_s$ are drawn from the one-parameter gamma distribution with variance θ, and the event times are generated according to model (4) assuming a constant baseline yearly hazard rate, $h_0(t) = \lambda$ for all $t > 0$. Conditional on u_i, the event time t_{ij} has an exponential distribution with rate $\lambda u_i \exp(x_{ij}\beta)$. We set $\theta = 2/3$, $\lambda = 0.5$, and $\beta = -0.7$; the between-centre heterogeneity induced by this parameter setting is shown by the spread in the median times to event from centre to centre in Figure 1, which is obtained following Duchateau & Janssen (2005). The censoring times are also generated from an exponential distribution whose rate parameter controls the amount of censoring; either 0%, 30%, or 60%.

In each of these 18 settings, models (1)–(4) are fitted to $K = 10000$ simulated data sets by means of the `coxph()` function in R (Therneau & Grambsch, 2000, Chapter 9). For model (4), we use the (correctly specified) gamma frailty distribution (impact of misspecification is studied in the next section). We report the mean of the $\hat{\beta}_k$'s and their standard deviation (SD). We also give the empirical coverage rate of the asymptotic 95% confidence interval (CI cov.) based on the normal approximation, i.e. the proportion of such confidence intervals that cover the true value of β.

For these simulations, the number of replications has been determined as follows (Burton et al., 2006). Let p_c be the true coverage probability and X the number of times that the confidence interval covers β out of K

replications; then $X \sim \text{Bin}(K, p_c)$. The empirical estimator $\hat{p}_c = {}^{X}/_{K}$ has an asymptotic normal distribution with mean p_c and variance ${}^{p_c(1-p_c)}/_{K}$ so that the width of its 95% confidence interval is approximately $2\sqrt{{}^{p_c(1-p_c)}/_{K}}$, which is bounded from above by $\sqrt{{}^{1}/_{K}}$. With $K = 10000$, the width of that confidence interval equals 0.01. Empirical coverage probabilities below 0.945 (resp. beyond 0.955) therefore correspond to under-coverage (resp. over-coverage).

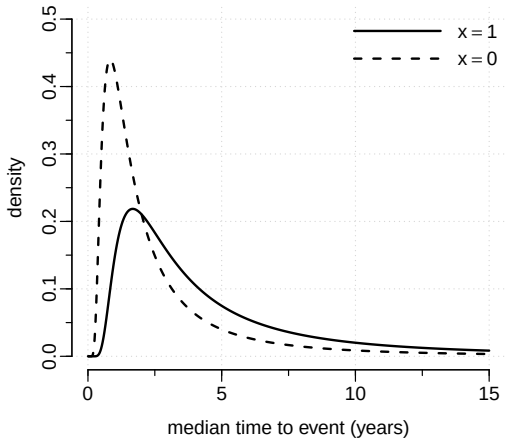

Figure 1 – Density function of the median time to event over centres for $\theta = {}^{2}/_{3}$, $\lambda = 0.5$, and $\beta = -0.7$.

3.2 Simulation results and guidelines

Only the results under 30% of censoring are displayed (Table 1). Conclusions are similar under the other censoring fractions and results are given in A (Table 5 and Table 6).

The unadjusted approach Model (1) makes no attempt to account for clustering. Now, failure to account for clustering alters the way the treatment effect has to be interpreted. Indeed, the treatment effect, as measured by $\exp(\beta)$, does not have the same meaning in an unadjusted model (marginal model) and in the adjusted models that we consider (conditional models). On the one hand, in model (1), $\exp(\beta)$ compares the hazard rates of two patients randomly drawn from the population under study, one treated and one untreated, but regardless of where they come from (population-averaged interpretation). On the other hand, in conditional models (and in

Table 1 – Simulation results under the **moderate censoring** fraction (30%) for the unadjusted Cox model (1), the fixed effects Cox model (2), the stratified Cox model (3), and the semi-parametric gamma frailty model (4). The true value of β is -0.700.

sample size	stat.	model			
		(1)	(2)	(3)	(4)
1:1					
6×48	mean	−0.530	−0.713	−0.704	−0.700
	SD	0.142	0.151	0.153	0.149
	CI cov.	0.780	0.949	0.953	0.953
8×18	mean	−0.501	−0.771	−0.701	−0.698
$+ 24 \times 6$	SD	0.125	0.172	0.169	0.153
	CI cov.	0.738	0.907	0.951	0.949
48×6	mean	−0.496	−0.820	−0.702	−0.696
	SD	0.120	0.186	0.174	0.155
	CI cov.	0.734	0.857	0.951	0.951
2:1					
6×48	mean	−0.530	−0.714	−0.703	−0.701
	SD	0.147	0.155	0.156	0.152
	CI cov.	0.791	0.948	0.952	0.953
8×18	mean	−0.504	−0.773	−0.704	−0.701
$+ 24 \times 6$	SD	0.130	0.178	0.172	0.158
	CI cov.	0.763	0.905	0.952	0.949
48×6	mean	−0.499	−0.823	−0.706	−0.701
	SD	0.125	0.191	0.180	0.160
	CI cov.	0.760	0.855	0.949	0.947

particular in model (4) used to simulate the data), $\exp(\beta)$ compares the hazard rates of two patients, one treated and one untreated, randomly drawn from the same centre (centre-specific interpretation). Therefore, in our simulations, the unadjusted model estimates a quantity that is different from the target, explaining the results seen for model (1) in Table 1. It is worth noting that both the marginal and the conditional interpretations coincide when there is no centre effect.

The fixed effects approach Model (2) requires maximisation over a $(p + s - 1)$-parameter space, with p the number of parameters in $\boldsymbol{\beta}$ (here, $p = 1$). This is numerically challenging whenever the number of centres, s, is large relative to the total sample size. The fixed effects approach therefore performs poorly when $s = 8 + 24$ or $s = 48$. It produces estimates that are biased away from the true β and the coverage of the confidence interval is below 95%. In the context of multicentre clinical trials, the fixed effects approach further shows additional limitations. (i) It implicitly assumes that centres participating in the trial have been designedly chosen and are by themselves of interest. Inference is to be made for those centres only and conclusions are thus restricted in scope. (ii) It provides neither a summary measure of heterogeneity between centres, nor a convenient framework to test for the presence of centre effects (Andersen et al., 1999). While it is sometimes of interest to assess whether a covariate explains heterogeneity in outcome between centres (Legrand et al., 2006), it is unfeasible in this model to include a covariate whose values only change at the centre level. (iii) Precision in centre effects estimates is dependent upon the centre size, and interpreting them can therefore be misleading. A related problem is that the centre effects estimates (and their interpretation) also depend on the choice of the reference centre, which is generally arbitrary.

The stratified approach Model (3) performs well, with good point estimates and good coverage probabilities. However, no between-strata comparisons are made by the stratified approach which therefore fails to make good use of all the information at hand. This explains why standard deviations deteriorate when the centre size decreases. Besides, and similar to the fixed effects approach, (i) interpretation of the treatment effect is restricted to participating centres, (ii) no heterogeneity measure is returned, and centre-specific covariates cannot be investigated because strata in which patients have the same covariate information do not contribute at all to the likelihood.

The frailty modelling approach Model (4) has good performance in every investigated setting with virtually no bias and good coverage probabilities. Unlike the stratified model, it also makes use of between-centre

comparisons to gather information on the treatment effect; this explains why standard deviations are smaller for the frailty model. This superiority is even more pronounced under high levels of censoring. The frailty modelling approach further provides a rich framework for the analysis of multicentre clinical trials. (i) Due to their random nature, the actual levels of the frailty term (i.e., the centre effects for those centres participating in the trial) are not of intrinsic interest and the conclusions of the study are intended to be generalised more broadly to all hospitals represented by the sample at hand. (ii) The variance of the gamma frailty distribution, θ, is a key parameter: it provides information on how much the baseline risk within each centre deviates from $h_0(\cdot)$. To help interpretation, this parameter can further be translated into clinically relevant quantities like the spread in the median times to event (as done in Figure 1), or in the five-year disease-free percentages from centre to centre (Duchateau & Janssen, 2005). Alternatively, the θ parameter can be transformed into the Kendall's τ, which measures the degree of association within a centre (Hougaard, 2000, Section 4.2 and Section 7.2.5). For gamma frailties, in particular, $\tau = {}^{\theta}/_{(\theta+2)}$. Considering the u_i's as random effects parameters also offers the possibility to study whether the inclusion of a centre-specific covariate explains/reduces heterogeneity between centres (Legrand et al., 2006).

4 Robustness against misspecification of the frailty distribution

One potential issue with the frailty model, which may discourage from using it, is that we have to make a choice regarding the frailty distribution. The usual assumption is to give the frailties a gamma distribution, the main reasons being mathematical convenience and software availability rather than clinical or empirical (data-driven) evidence. The most common form of misspecification is thus one in which the gamma frailty distribution is applied while frailties actually follow another distribution. Alternative distributions that have received interest include the positive stable and the inverse Gaussian. Work on diagnostic checks to assess appropriate form of the frailty distribution is underway, but not yet largely available (particularly in software) and research is still needed in this area. In the meanwhile, it is important to investigate robustness against misspecification of the frailty distribution in a variety of real-world applications.

To observe the result of misspecification of the frailty distribution on the inference for the treatment effect (and more generally for the fixed effects parameters included in the model), we generate data according to model (4) with inverse Gaussian and positive stable frailties, and we fit the misspecified gamma frailty model. Several cluster sizes, amounts of heterogeneity, and degrees of imbalance between the treatment arms are investigated to cover

a wide panel of clinical trial settings.

4.1 Simulation settings

Data are generated as in Section 3.1 with 30% of censoring. We consider various amounts of heterogeneity as quantified by various values for the Kendall's τ. We set $\tau = 0.1$, $\tau = 0.25$, and $\tau = 0.5$. This last value is however not possible under the inverse Gaussian frailty distribution, and this setting is therefore not considered. We also investigate the situation where treatment allocation is done at the centre level (cluster randomised trials): $s/2$ centres are randomly chosen to be assigned to the treatment arm and the other half is allocated to the control arm. In that case, the fixed effects and the stratified Cox models cannot be fitted as previously noted. As above, we run 10000 replications.

4.2 Simulation results

Table 2 and Table 3 display the results obtained when randomisation is done at the patient level. For small or moderate amounts of heterogeneity, misspecifying the frailty distribution has no effect on the inference for β. Mistaking positive stable frailties with gamma frailties when there is more heterogeneity, on the other hand, results in a treatment effect that is slightly overestimated when $s = 8+24$ or $s = 48$. The bias is however clinically irrelevant. Moreover, even in such cases the misspecified frailty model appears superior to the stratified model in terms of mean squared error.

Results under centre randomisation are shown in Table 4. In that case, the estimates of β are much more variable and we observe some undercoverage, even when the frailty distribution is correctly specified. This phenomenon is even more severe under high heterogeneity. When $\tau = 0.5$, furthermore, `coxph()` sometimes fails or experiences numerical difficulties. Nonetheless, in practice we rarely expect such a high within-centre association. Our results indicate that the coverage of the 95% confidence interval can be greatly improved by increasing the number of centres. Under a misspecified frailty distribution, though, the number of centres needed might be very large. It is worth noting that mistaking inverse Gaussian frailties appears to be of less consequence that mistaking positive stable frailties.

In order to investigate further this behaviour, we examine the shape of the marginal log-likelihood obtained by integrating out the frailties from the gamma frailty model, assuming a constant baseline hazard rate (see, e.g., (Duchateau & Janssen, 2008, Section 2.2)). In Figure 2a, the left panel is obtained by generating a data set from the gamma frailty model with $s = 6$ centres of size $n = 48$, $\theta = 2/3$, and with a 1:1 allocation ratio. The log-likelihood surface appears concave, irrespective of the value taken by θ. This suggests that the optimisation algorithm will reliably converge

Table 2 – Simulation results when randomisation is done at the **patient level with a 1:1 allocation ratio** in each centre. The actual frailties either follow the gamma (Gam), the inverse Gaussian (IG), or the positive stable (PS) frailty distribution. Both the stratified Cox model (3) and the semi-parametric gamma frailty model (4) are fitted. The true value of β is -0.700.

		true frailty distribution					
		Gam		IG		PS	
sample size	stat.	model (3)	(4)	(3)	(4)	(3)	(4)
$\tau = 0.10$							
6×48	mean	−0.706	−0.702	−0.704	−0.700	−0.704	−0.701
	SD	0.152	0.147	0.153	0.149	0.153	0.148
	CI cov.	0.950	0.949	0.950	0.950	0.949	0.950
8×18	mean	−0.704	−0.700	−0.703	−0.698	−0.704	−0.693
$+ 24 \times 6$	SD	0.168	0.151	0.168	0.153	0.167	0.151
	CI cov.	0.950	0.949	0.952	0.950	0.952	0.950
48×6	mean	−0.704	−0.699	−0.702	−0.696	−0.703	−0.694
	SD	0.173	0.152	0.176	0.155	0.174	0.152
	CI cov.	0.954	0.950	0.950	0.945	0.955	0.949
$\tau = 0.25$							
6×48	mean	−0.701	−0.698	−0.705	−0.701	−0.703	−0.700
	SD	0.154	0.149	0.155	0.150	0.153	0.148
	CI cov.	0.945	0.946	0.949	0.950	0.950	0.951
8×18	mean	−0.704	−0.701	−0.701	−0.693	−0.704	−0.692
$+ 24 \times 6$	SD	0.168	0.153	0.169	0.153	0.166	0.150
	CI cov.	0.952	0.950	0.947	0.947	0.953	0.950
48×6	mean	−0.705	−0.700	−0.705	−0.694	−0.703	−0.686
	SD	0.177	0.156	0.175	0.154	0.175	0.154
	CI cov.	0.949	0.949	0.956	0.952	0.951	0.950
$\tau = 0.50$							
6×48	mean	−0.704	−0.701			−0.703	−0.698
	SD	0.156	0.152			0.154	0.150
	CI cov.	0.951	0.952			0.950	0.950
8×18	mean	−0.701	−0.698			−0.703	−0.685
$+ 24 \times 6$	SD	0.172	0.157			0.169	0.152
	CI cov.	0.950	0.947			0.950	0.952
48×6	mean	−0.707	−0.701			−0.702	−0.674
	SD	0.176	0.157			0.175	0.154
	CI cov.	0.955	0.951			0.953	0.950

Table 3 – Simulation results when randomisation is done at the **patient level with a 2:1 allocation ratio** in each centre. The actual frailties either follow the gamma (Gam), the inverse Gaussian (IG), or the positive stable (PS) frailty distribution. Both the stratified Cox model (3) and the semi-parametric gamma frailty model (4) are fitted. The true value of β is -0.700.

		true frailty distribution					
		Gam		IG		PS	
sample size	stat.	model					
		(3)	(4)	(3)	(4)	(3)	(4)
$\tau = 0.10$							
6×48	mean	−0.708	−0.705	−0.706	−0.704	−0.702	−0.700
	SD	0.156	0.153	0.156	0.152	0.156	0.152
	CI cov.	0.950	0.948	0.950	0.950	0.950	0.951
8×18	mean	−0.705	−0.701	−0.705	−0.699	−0.705	−0.693
$+ 24 \times 6$	SD	0.173	0.156	0.172	0.157	0.171	0.154
	CI cov.	0.951	0.950	0.951	0.945	0.951	0.954
48×6	mean	−0.708	−0.703	−0.704	−0.699	−0.704	−0.693
	SD	0.179	0.158	0.176	0.155	0.181	0.158
	CI cov.	0.953	0.947	0.955	0.949	0.947	0.947
$\tau = 0.25$							
6×48	mean	−0.704	−0.701	−0.705	−0.702	−0.706	−0.702
	SD	0.155	0.152	0.157	0.153	0.157	0.153
	CI cov.	0.951	0.951	0.952	0.952	0.951	0.952
8×18	mean	−0.702	−0.699	−0.704	−0.697	−0.705	−0.692
$+ 24 \times 6$	SD	0.172	0.158	0.173	0.159	0.173	0.157
	CI cov.	0.949	0.946	0.948	0.949	0.949	0.951
48×6	mean	−0.707	−0.701	−0.706	−0.695	−0.704	−0.684
	SD	0.180	0.161	0.180	0.161	0.180	0.159
	CI cov.	0.950	0.949	0.949	0.945	0.950	0.950
$\tau = 0.50$							
6×48	mean	−0.706	−0.703			−0.706	−0.701
	SD	0.160	0.157			0.160	0.156
	CI cov.	0.951	0.950			0.948	0.949
8×18	mean	−0.703	−0.701			−0.702	−0.685
$+ 24 \times 6$	SD	0.174	0.161			0.176	0.160
	CI cov.	0.953	0.951			0.946	0.949
48×6	mean	−0.708	−0.702			−0.705	−0.675
	SD	0.184	0.166			0.183	0.161
	CI cov.	0.950	0.947			0.950	0.950

Table 4 – Simulation results when randomisation is done at the **centre level**. The actual frailties either follow the gamma (Gam), the inverse Gaussian (IG), or the positive stable (PS) frailty distribution. Only the frailty model (4) can be fitted. The true value of β is -0.700.

		true frailty distribution		
sample size	stat.	Gam	IG	PS
$\tau = 0.10$				
6×48	mean	-0.700	-0.705	-0.703
	SD	0.422	0.443	0.845
	CI cov.	0.827	0.814	0.747
8×18	mean	-0.710	-0.708	-0.694
$+ 24 \times 6$	SD	0.242	0.244	0.343
	CI cov.	0.926	0.927	0.894
48×6	mean	-0.699	-0.694	-0.690
	SD	0.206	0.209	0.234
	CI cov.	0.939	0.934	0.919
$\tau = 0.25$				
6×48	mean	-0.698	-0.699	-0.703
	SD	0.724	0.857	1.522
	CI cov.	0.820	0.761	0.641
8×18	mean	-0.704	-0.698	-0.659
$+ 24 \times 6$	SD	0.315	0.365	0.630
	CI cov.	0.941	0.918	0.839
48×6	mean	-0.692	-0.682	-0.647
	SD	0.290	0.324	0.478
	CI cov.	0.942	0.919	0.837
$\tau = 0.50$				
6×48	mean	-0.697		-0.732
	SD	1.385		2.830
	CI cov.	0.807		0.521
8×18	mean	-0.704		-0.590
$+ 24 \times 6$	SD	0.446		1.220
	CI cov.	0.955		0.752
48×6	mean	-0.695		-0.561
	SD	0.460		1.158
	CI cov.	0.950		0.811

to the maximum likelihood estimate for β. In the right panel, obtained under centre randomisation, on the other hand, the curvature of the log-likelihood surface appears to depend upon the value taken by θ. This affects our uncertainty about $\hat{\beta}$, and hence the coverage probabilities in Table 4. Figure 2b shows that increasing the number of clusters markedly improves the shape of the log-likelihood.

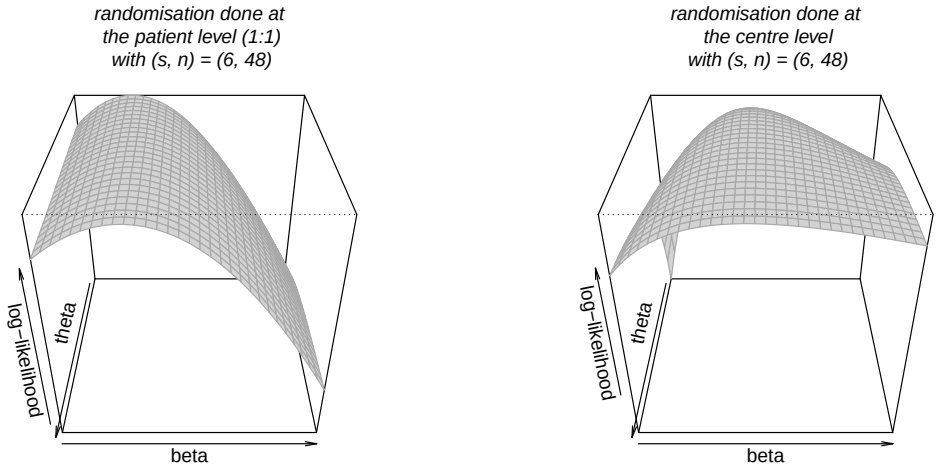

Figure 2 – Log-likelihood surfaces obtained from the gamma frailty model, with $\boldsymbol{s = 6}$ **centres of size** $\boldsymbol{n = 48}$, assuming a constant baseline hazard rate ($h_0(t) = \lambda = 0.5$ for all $t > 0$). The parameter β ranges from -1.5 to 1.5 and θ from 0.5 to 3.

Our conclusion is that the estimate of θ needs to be close to its true value to draw good inferences about the treatment effect when randomisation is done at the centre level. Additional simulations (results not shown) confirm that the coverage probabilities from Table 4 improve when θ is fixed to its true value. Obtaining a good estimate of θ, however, seems to be difficult, especially when the true θ is large, and/or the number of centres is limited, and/or the frailty distribution greatly departs from the correct specification. Similar trends are observed in Duchateau et al. (2002) and in (Duchateau & Janssen, 2008, Section 5.2.3 and Section 5.2.4).

5 Discussion

Models (2)–(4) are conditional models, in the sense that $\exp(\beta)$ compares risks within centre, as opposed to a marginal modelling approach for which

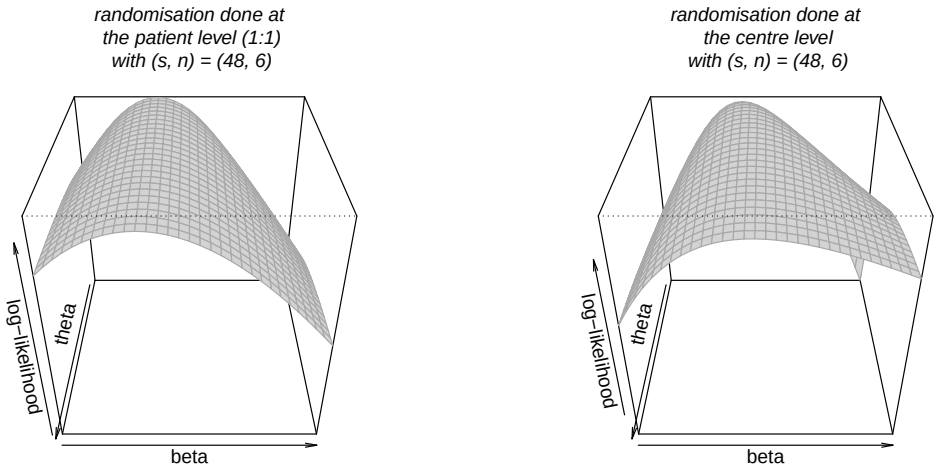

Figure 2 – Log-likelihood surfaces obtained from the gamma frailty model, with $s = 48$ **centres of size** $n = 6$, assuming a constant baseline hazard rate ($h_0(t) = \lambda = 0.5$ for all $t > 0$). The parameter β ranges from -1.5 to 1.5 and θ from 0.5 to 3.

$\exp(\beta)$ would rather have a population-averaged interpretation. It is also possible to tackle the problem of dependence due to clustering within the marginal framework (see, e.g., Duchateau & Janssen, 2008, Section 3.4). However, we have confined ourselves to the conditional approach which we find more relevant in this context as it compares "like-for-like" when assessing the treatment effect.

The fact that the frailty modelling approach requires a frailty distribution, a guess whose validity is difficult to assess due to the random nature of the frailties, might make the trial statistician quite reluctant to adopt the frailty model and rather be inclined to apply an alternative like the fixed effects or the stratified Cox model. The frailty model, however, allows a more natural interpretation of the treatment effect, makes a better use of the information contained in the data, and provides us with additional information on heterogeneity in outcome between centres (cf. Section 3). Furthermore, our simulations indicate that, most of the time, misspecification is not a matter of great concern (cf. Section 4).

Nonetheless, we pointed out a very specific clinical setting where caution is needed, namely when randomisation is done at the centre level. Such a randomisation scheme is used when, for logistical reasons in particular, it is easier to deliver interventions at the cluster level rather than at the individual level (Murray et al., 2004). To compensate the effect of misspecification, a large number of centres is needed, especially under high levels of hetero-

geneity. It is worth reminding here that one can rely neither on the fixed effects nor on the stratified Cox model in that case.

Also to be noted is that, as they stand, models (2)–(4) do not allow for treatment-by-centre interaction; $\exp(\beta)$ is an overall treatment effect assumed to remain unchanged across centres. It is also interesting to study heterogeneity in treatment effect over centres. To allow different treatment effects, additional fixed parameters could enter both model (2) and model (3). Such a strategy would, however, display the same weaknesses as the fixed effects approach discussed in Section 3.2, and would even be more problematic due to the large number of additional parameters. Alternatively, an extra random interaction term can be included in the frailty model, at a very limited cost in terms of additional parameters. This is particularly common in meta-analyses (Massonnet et al., 2008; Rondeau et al., 2008). In the context of multicentre clinical trials, a Bayesian approach using Laplacian integration is proposed in Legrand et al. (2005). For a more comprehensive review on how to handle random treatment-by-centre interactions, see (Wienke, 2010, Section 7.1).

Besides the "arbitrary" choice of the frailty distribution, another common reluctance of trial statisticians to use frailty models is the lack of software to fit them. Even though this has been true for a long time, this is no longer the case today. With the recent apparition of flexible and user-friendly implementations in standard software (such as R and SAS), frailty models are now readily available whenever needed.

References

Andersen, P. K., Klein, J. P. & Zhang, M.-J. (1999). Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* **18**, 1489–1500.

Burton, A., Altman, D. G., Royston, P. & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine* **25**, 4279–4292.

Cortiñas Abrahantes, J., Legrand, C., Burzykowski, T., Janssen, P., Ducrocq, V. & Duchateau, L. (2007). Comparison of different estimation procedures for proportional hazards model with random effects. *Computational Statistics & Data Analysis* **51**, 3913–3930.

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

Duchateau, L. & Janssen, P. (2005). Understanding heterogeneity in generalized mixed and frailty models. *The American Statistician* **59**, 143–146.

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

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R. & Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis* **40**, 603–620.

Gill, R. D. (1984). Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association* **79**, 441–447.

Glidden, D. V. & Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine* **23**, 369–388.

Hirsch, K. & Wienke, A. (2012). Software for semiparametric shared gamma and log-normal frailty models: An overview. *Computer Methods and Programs in Biomedicine* **107**, 582–597.

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

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.

Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R. & Duchateau, L. (2005). A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. *Statistics in Medicine* **24**, 3789–3804.

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.

Munda, M., Rotolo, F. & Legrand, L. (2012). parfm: Parametric frailty models in R. *Journal of Statistical Software* **51**, ??–??

Murray, D. M., Varnell, S. P. & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health* **94**, 423–432.

O'Quigley, J. & Stare, J. (2002). Proportional hazards models with frailties and random effects. *Statistics in Medicine* **21**, 3219–3233.

Pickles, A. & Crouchley, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine* **14**, 1447–1461.

Rondeau, V., Mazroui, Y. & Gonzalez, J. R. (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**, 1–28.

Rondeau, V., Michiels, S., Liquet, B. & Pignon, J. P. (2008). Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine* **27**, 1894–1910.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

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

Appendices

A

Simulation results (continued)

Table 5 – Simulation results under **no censoring** for the unadjusted Cox model (1), the fixed effects Cox model (2), the stratified Cox model (3), and the semi-parametric gamma frailty model (4). The true value of β is -0.700.

sample size	stat.	model			
		(1)	(2)	(3)	(4)
1:1					
6×48	mean	−0.505	−0.711	−0.703	−0.700
	SD	0.128	0.127	0.128	0.125
	CI cov.	0.606	0.947	0.950	0.950
8×18	mean	−0.450	−0.750	−0.695	−0.690
$+ 24 \times 6$	SD	0.106	0.142	0.140	0.129
	CI cov.	0.433	0.918	0.950	0.950
48×6	mean	−0.440	−0.807	−0.703	−0.698
	SD	0.097	0.153	0.147	0.132
	CI cov.	0.380	0.854	0.950	0.949
2:1					
6×48	mean	−0.507	−0.713	−0.704	−0.701
	SD	0.135	0.134	0.135	0.132
	CI cov.	0.643	0.947	0.949	0.951
8×18	mean	−0.458	−0.753	−0.697	−0.694
$+ 24 \times 6$	SD	0.113	0.149	0.145	0.136
	CI cov.	0.515	0.915	0.952	0.950
48×6	mean	−0.448	−0.808	−0.706	−0.702
	SD	0.105	0.160	0.153	0.139
	CI cov.	0.471	0.867	0.951	0.949

Table 6 – Simulation results under the **high censoring** fraction (60%) for the unadjusted Cox model (1), the fixed effects Cox model (2), the stratified Cox model (3), and the semi-parametric gamma frailty model (4). The true value of β is -0.700.

sample size	stat.	model			
		(1)	(2)	(3)	(4)
1:1					
6×48	mean	−0.590	−0.716	−0.705	−0.702
	SD	0.185	0.202	0.204	0.198
	CI cov.	0.922	0.954	0.956	0.956
8×18	mean	−0.574	−0.793	−0.706	−0.702
$+ 24 \times 6$	SD	0.179	0.238	0.230	0.204
	CI cov.	0.916	0.903	0.949	0.949
48×6	mean	−0.570	−0.840	−0.707	−0.699
	SD	0.173	0.253	0.232	0.203
	CI cov.	0.920	0.871	0.954	0.954
2:1					
6×48	mean	−0.586	−0.715	−0.704	−0.701
	SD	0.185	0.204	0.205	0.199
	CI cov.	0.925	0.948	0.951	0.950
8×18	mean	−0.568	−0.793	−0.703	−0.702
$+ 24 \times 6$	SD	0.175	0.237	0.227	0.203
	CI cov.	0.921	0.899	0.951	0.951
48×6	mean	−0.566	−0.842	−0.706	−0.699
	SD	0.174	0.257	0.237	0.207
	CI cov.	0.920	0.861	0.948	0.949

We performed additional simulation runs with 30% of censoring under a 1:1 randomisation scheme in the absence of treatment effect ($\beta = 0$). Table 7 records the results. They show that the treatment effect is now well estimated by model (2). However, we still observe some under-coverage, indicating that the null hypothesis of no treatment effect, $H_0 : \beta = 0$, is too often rejected.

Table 7 – Simulation results under the moderate censoring fraction (30%) for the unadjusted Cox model (1), the fixed effects Cox model (2), the stratified Cox model (3), and the semi-parametric gamma frailty model (4). The true value of β is 0.

sample size	stat.	model			
		(1)	(2)	(3)	(4)
6×48	mean	0.000	0.000	0.001	0.000
	SD	0.120	0.146	0.147	0.143
	CI cov.	0.978	0.947	0.952	0.952
8×18	mean	−0.001	0.000	−0.001	0.000
$+ 24 \times 6$	SD	0.118	0.165	0.162	0.147
	CI cov.	0.980	0.928	0.949	0.951
48×6	mean	−0.001	0.001	0.000	0.000
	SD	0.115	0.179	0.167	0.149
	CI cov.	0.982	0.915	0.950	0.951