MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1612 Published online 12 February 2014 in Wiley Online Library

Adjusting for centre heterogeneity in multicentre clinical trials with a time-to-event outcome

Marco Munda* and Catherine Legrand

Conducting a clinical trial at multiple study centres raises the issue of whether and how to adjust for centre heterogeneity in the statistical analysis. In this paper, we address this issue for multicentre clinical trials with a time-to-event outcome. Based on simulations, we show that the current practice of ignoring centre heterogeneity can be seriously misleading, and we illustrate the performances of the frailty modelling approach over competing methods. A special attention is paid to the problem of misspecification of the frailty distribution. The appendix provides sample codes in R and in SAS to perform the analyses in this paper. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: multicentre clinical trial; time-to-event outcome; frailty model; frailty distribution misspecification

1. INTRODUCTION

Clinical trials are conducted at multiple centres for two main reasons: to accrue the required number of patients within a short period of time and to broaden the scope of the trial results. Considerable efforts are made to standardise the way the trial is conducted in each centre according to the study protocol. However, patients from different centres are likely to have different prognoses due to, for example, differences in disease diagnosis, differences in referral patterns and differences in indications for background therapies. Hence, variability in outcome between patients within an individual centre tends to be lower than variability in outcome between patients at different centres.

The International Conference on Harmonisation (ICH) guidance document E9 'statistical principles for clinical trials' (www.ich.org/products/guidelines/efficacy/article/efficacy-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 the treatment-to-centre interaction.' For multicentre clinical trials with a time-to-event endpoint (e.g. time until tumour progression in cancer studies), however, recommendations on how to adjust for centre heterogeneity are limited. We address this problem in this paper.

Probably the most natural option is to enter additional fixed centre effects parameters into the Cox model. Alternatively, modelling heterogeneity between centres can be performed via stratification of the baseline hazard function. The frailty model is another approach that has gained in popularity in recent years. The frailty model is a proportional hazards model that includes a random factor, the frailty term, to account for the centre-to-centre variability. In Section 2, we briefly review the basics of these modelling strategies. In Section 3, a real data example is analysed using each method. Statistical aspects are clearly discussed in [1], where the authors found advantages in using the frailty approach.

The frailty approach requires specification of a distributional form for the frailty distribution. This is a difficult issue due to the latent nature of the frailty term. Therefore, it is of interest to investigate whether the frailty approach is the strategy to be recommended, considering the fact that the frailty distribution might be misspecified.

In this paper, it is our aim to provide pragmatic guidelines for the practising statistician in the pharmaceutical industry. Our first objective, covered in Section 4, is to highlight the limitations of the current practice of ignoring centre heterogeneity as well as the pros and the cons of the aforementioned modelling strategies to adjust for centre heterogeneity. Our second objective, covered in Section 5, is to further investigate the performances of the frailty model over its competitors when the frailty distribution is misspecified. Section 6 summarises the conclusions and presents our recommendations.

2. MODELLING CLUSTERED TIME-TO-EVENT DATA

2.1. The (unadjusted) Cox model

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

$$\mathcal{Z} = \{(y_j, \delta_j, \boldsymbol{x}_j) \mid j = 1, \ldots, N\}$$

where $y_j = \min(t_j, c_j)$ is the time to event or censoring, whichever comes first; $\delta_j = I(t_j \leqslant c_j)$ indicates whether an observation corresponds to an event ($\delta_j = 1$) or is censored ($\delta_j = 0$); and $\boldsymbol{x}_j$ is a vector of covariates. We make the standard assumptions that the event times (the t_j's) and the censoring times (the c_j's)

Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium

**Correspondence to: Marco Munda, Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium. E-mail: marco.munda@uclouvain.be*

are independent given the covariate information (independent censoring) and that the censoring distribution has no common parameter with the event time distribution (noninformative censoring).

Let $h_j(t)$ denote the hazard rate of subject j at time t. The Cox model specifies the way the explanatory variables act on the hazard rate, but lets its time dependence unspecified,

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

with $h_0(\cdot)$ a (nonspecified) baseline hazard function and $\boldsymbol{\beta} = (\beta_1 \ldots \beta_p)'$ a vector of fixed effects parameters. Owing to its semi-parametric nature, the Cox model has become routine in studies of time-to-event outcomes.

The ratio of the hazard functions for two subjects with different covariate information, say $\boldsymbol{x}_{j_1}$ and $\boldsymbol{x}_{j_2}$, is

$$\frac{h_{j_1}(t)}{h_{j_2}(t)} = \exp\left((\boldsymbol{x}_{j_1} - \boldsymbol{x}_{j_2})'\boldsymbol{\beta}\right)$$

A one-unit change in one of the explanatory variables (while all other are kept fixed) results in a proportional change in the hazard function. The parameter β_k in model (1) is thus interpreted as a conditional hazard ratio.

An estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by maximising a partial log likelihood given by (assuming no ties in the event times)

$$\ell(\boldsymbol{\beta}; \mathcal{Z}) = \sum_{j=1}^{N} \delta_j \left[\boldsymbol{x}_j'\boldsymbol{\beta} - \log\left(\sum_{\ell \in R(y_j)} \exp\left(\boldsymbol{x}_\ell'\boldsymbol{\beta}\right)\right)\right]$$

with $R(y_j)$ the risk set at time y_j containing all subjects still under observation just prior to y_j. Approximate standard errors are given by the square roots of the diagonal entries of the negative inverse matrix of second derivatives of $\ell(\cdot; \mathcal{Z})$ evaluated at $\hat{\boldsymbol{\beta}}$. Even though $\ell(\cdot; \mathcal{Z})$ is not a genuine log likelihood, it has been shown that consistency and asymptotic normality properties for the estimator of $\boldsymbol{\beta}$ are preserved [2].

2.2. Adjusting for centre heterogeneity

Model (1) requires independent (homogeneous) data up to measured covariates. In multicentre clinical trial data, however, there is likely to be heterogeneity across centres. To account for this, centre effects must somehow be included in the statistical model used for the analysis.

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

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

where we now use two indices, $i \in \{1, \ldots, s\}$ for the s centres 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 $\mathcal{Z}$ is changed accordingly). In model (2), $\boldsymbol{\alpha} = (\alpha_1 \ldots \alpha_{s-1})'$ contains the fixed centre effects, and $\boldsymbol{c}_i$ denotes the vector with a 1 in the i^{th} position and 0's elsewhere ($i = 1, \ldots, s-1$). The last centre does not need an indicator because an observation is known to belong to that centre when $\boldsymbol{c}_i = (0 \ldots 0)'$. If we had included an additional indicator for the last centre, then the model would have been overparametrised. Choosing one particular centre as reference is consistent with the interpretation of $h_0(\cdot)$ as being the hazard rate for subjects with covariate values all equal to 0. However, this choice is arbitrary, and any centre can play the role of the reference centre.

2.2.2. The stratified approach. Instead of entering the centre variable as additional fixed effects parameters, the baseline hazard can be stratified on that variable to indicate that different subpopulations are exposed to different baseline risks, that is,

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

where $h_{01}(\cdot), \ldots, h_{0s}(\cdot)$ are unspecified and unrelated baseline hazard functions. The partial likelihood approach is readily adapted by multiplying the partial likelihoods specific to each stratum [1].

2.2.3. The frailty approach. Participating centres may also be viewed as one possible sample from a broader population of centres. In that case, centre i has a random effect, called frailty and denoted by u_i, on the hazard rate. The frailty term reflects different levels of risk across centres. The (shared) frailty model is defined as [3,4]

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

The u_i's are the actual values of a random variable with probability density $f(\cdot)$, called the frailty distribution. The (one-parameter) gamma distribution, with density

$$f(u) = \frac{(1/\theta)^{1/\theta}u^{(1/\theta)-1}\exp(-(1/\theta)u)}{\Gamma(1/\theta)}$$

is the most commonly used. Note that $\mathrm{E}(U) = 1$ and that $\mathrm{Var}(U) = \theta$. The variance of the frailty term determines the degree of heterogeneity between centres.

In the frequentist approach, which we follow in this paper, the frailty model is fitted by maximising the marginal likelihood (also called the observed likelihood). The marginal likelihood is obtained by integrating out the u_i's from the joint likelihood of $\mathcal{Z}$ and $\boldsymbol{u} = (u_1 \ldots u_s)'$. For the gamma frailty distribution, the integration can be done in closed form; see, for example, [3, Section 2.2], leading to the following log-likelihood function

$$\begin{aligned}\ell_{\text{marg}}(h_0(\cdot), \boldsymbol{\beta}, \theta; \mathcal{Z}) &= \sum_{i=1}^{s}\left[\left(\sum_{j=1}^{n_i}\delta_{ij}\left(\log(h_0(y_{ij})) + \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right)\right. \\ &\quad - \left(D_i + \frac{1}{\theta}\right)\log\left(1 + \theta\sum_{j=1}^{n_i}H_0(y_{ij})\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right) \\ &\quad \left. + I(D_i > 0)\sum_{\ell=0}^{D_i-1}\log(1 + \ell\theta)\right]\end{aligned}$$

with D_i the number of events in cluster i and $I(D_i > 0)$ the indicator function that takes value 1 if $D_i > 0$ and 0 otherwise.

In ℓ_{marg}, the baseline hazard function $h_0(\cdot)$ can either be modelled using a parametric distribution (e.g. Weibull) or it can be modelled in a nonparametric way. The parametric approach results in a fully parametric log likelihood that can be maximised by means of an optimisation routine (e.g. a Newton-type algorithm). Alternatively, if the form of the baseline hazard is left unspecified (semi-parametric approach), then it has to be eliminated using partial likelihood ideas. For a detailed

overview of the estimation techniques that are available in the semi-parametric case, see [5].

Parametric frailty models can be fitted in R by means of the `parfm()` function (part of the `parfm` library); see [6]. For the semi-parametric approach, an overview of the available software is provided in [7]. In this paper, we use the `coxph()` function in R (part of the `survival` library) to fit the semi-parametric gamma frailty model (as well as models (1)–(3)). For a detailed description of the proper use of `coxph()` for frailty models, see [8, Chapter 9]. Generic sample codes are provided in the Appendix.

3. EXAMPLE

The European Organisation for Research and Treatment of Cancer (EORTC) trial 10854 is a randomised phase III breast cancer trial comparing surgery alone versus surgery followed by one course of perioperative chemotherapy. A total of 2793 women with early breast cancer (1395 in the control arm and 1398 in the perioperative chemotherapy arm) recruited at 14 centres (median number of patients per centre: 69) were followed up (median follow-up time: 8.8 years) for overall survival (primary endpoint), progression-free survival and locoregional recurrence (secondary endpoints). The study design and results have been published previously [9,10].

The results of fitting models (1)–(4) to the primary endpoint are shown in Table I. All models virtually lead to the same hazard ratio of 0.91, which does not differ significantly from 1. The largest standard error of $\hat{\beta}$ is found for model (3). Model (4) returns $\hat{\theta} = 0.049$, suggesting that the centre heterogeneity is quite low. In the next section, we will see that, in other settings (higher heterogeneity, smaller centres, etc.), models (1)–(3) have serious drawbacks compared with model (4). In particular, interpreting model (1) can be seriously misleading, the hazard ratio obtained from model (2) is usually biased and model (3) typically lacks power because of the large standard error.

4. COMPARISON OF THE MODELLING APPROACHES

In this section, we discuss the strengths and the weaknesses of models (2)–(4) to adjust for centre heterogeneity, and we illustrate this discussion with simulations. Because the unadjusted model (1) is often used in practice, we consider it as well.

4.1. Simulation setting

We consider two opposite situations with six centres of size 48 ($N = 6 \times 48$) and 48 centres of size 6 ($N = 48 \times 6$) as well as an intermediate situation with eight centres of size 18 plus 24 centres of size 6 ($N = 8 \times 18 + 24 \times 6$), thus keeping the total sample size fixed at $N = 288$. We mimic a 1:1 (respectively 2:1) allocation ratio in each centre by selecting $N/2$ (respectively $2N/3$) patients for the treatment arm ($x = 1$) and the remaining $N/2$ (respectively $N/3$) patients for the control arm ($x = 0$).

Table I. Results for the primary endpoint (overall survival) in the perioperative breast cancer clinical trial.

Model	$\widehat{\text{HR}}$	$\text{SE}(\hat{\beta}) \times 10^2$	95% CI($\hat{\beta}$)
Unadjusted (1)	0.9078	7.7448	0.7798–1.0565
Fixed effects (2)	0.9083	7.7504	0.7803–1.0573
Stratified (3)	0.9123	7.7550	0.7837–1.0621
Frailty (4)	0.9086	7.7488	0.7806–1.0577

Frailties $u_1, \ldots, u_s$ are randomly drawn from the one-parameter gamma distribution with variance θ. The event time for each patient is generated from model (4). By assuming a Weibull baseline hazard function ($h_0(t) = \lambda\rho t^{\rho-1}$ with scale $\lambda > 0$ and shape $\rho > 0$), the event time t_{ij} has, conditional on u_i, a Weibull distribution with scale $\lambda u_i \exp(x_{ij}\beta)$ and shape ρ. We take $\theta = 0.5$, $\lambda = 0.7$, $\rho = 1.5$, and $\beta = \log(2/3) \approx -0.4$ or $\beta = 0$. The between-centre heterogeneity induced by this parameter setting is shown in Figure 1 by the spread in the median time to event from centre to centre [11]. The censoring time for each patient is generated from an exponential distribution whose rate parameter is chosen so that 30% of the observations are censored. Additional simulations with 50% censoring are given in the supplementary material (available online as supporting information).

Note that by disregarding the clustering under this parameter setting (with 30% censoring), one will expect to have an 80% chance of declaring a hazard ratio of HR $= 2/3$ to be significant at the 5% level [12, Chapter 10].

For each setting, we fit models (1)–(4) to $K = 10000$ simulated data sets by means of `coxph()` in R. For model (4), we use the correctly specified gamma frailty distribution (the impact of misspecification is addressed separately in Section 5). We report

- $\overline{\text{HR}}$: the average hazard ratio;
- %bias: the percentage bias defined as $(\frac{1}{K}\sum_k \hat{\beta}_k - \beta)/\beta \times 100$ (it is undefined, and hence not reported, for $\beta = 0$);
- SD: the standard deviation of the $\hat{\beta}_k$'s;
- CI cov: the empirical coverage of the asymptotic 95% confidence interval based on the normal approximation, that is, the proportion of such confidence intervals that cover the true value of β;
- power/size: the empirical rejection rate for the null hypothesis of no treatment effect (H_0: $\beta = 0$) – under H_0, it equals 1 minus the empirical coverage probability.

4.1.1. Assessment of coverage. Let p_c be the true coverage probability and X the number of times the confidence interval covers

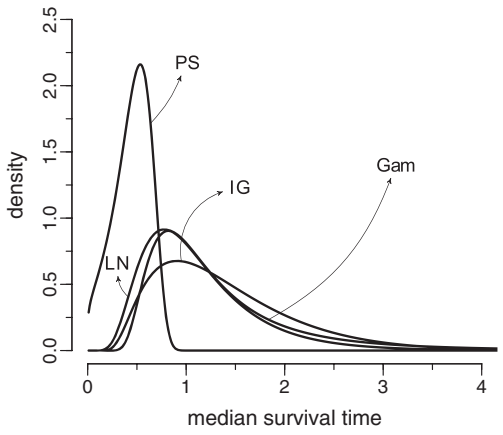

Figure 1. Density functions of the median time to event from centre to centre in the control group for different frailty distributions (Kendall's $\tau = 0.20$). Gam: gamma distribution; IG: inverse Gaussian distribution; LN: log-normal distribution; PS: positive stable distribution.

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 therefore approximately equals 0.01. Hence, empirical coverage probabilities below 0.945 correspond to undercoverage while empirical coverage probabilities above 0.955 correspond to overcoverage [13].

4.2. Results and guidelines

The results are displayed in Tables II ($\beta = \log(2/3) \approx -0.4$) and III ($\beta = 0$).

4.2.1. The unadjusted approach. Model (1) makes no attempt to account for clustering. This alters the way the treatment effect (HR $= \exp(\beta)$) has to be interpreted. Indeed, HR has different meanings in model (1) (marginal model) and in models (2)–(4) (conditional models). In model (1), HR compares the hazard rates of two subjects, one treated and one untreated, randomly drawn from the population under study, regardless of where they come from (population-averaged interpretation). On the other hand, in conditional models (and in particular in model (4) used to generate the data), HR compares the hazard rates of two subjects, 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. In Table II, we

Table II. Simulation results (30% censoring; Kendall's $\tau = 0.20$; $\beta = \log(2/3)$) under correct specification of the frailty distribution.

		Model			
Sample size	Statistic	(1)	(2)	(3)	(4)
1:1					
6×48	$\overline{\text{HR}}$	0.731	0.668	0.673	0.674
	% bias	−20.45	2.039	0.433	−0.021
	SD	0.133	0.148	0.149	0.145
	CI cov	0.929	0.948	0.954	0.953
	Power	0.617	0.809	0.784	0.798
8×18	$\overline{\text{HR}}$	0.742	0.645	0.674	0.674
$+24 \times 6$	% bias	−24.41	11.90	0.630	0.028
	SD	0.127	0.173	0.168	0.153
	CI cov	0.918	0.910	0.944	0.944
	Power	0.582	0.814	0.701	0.772
48×6	$\overline{\text{HR}}$	0.743	0.627	0.675	0.675
	% bias	−25.05	19.44	0.521	−0.205
	SD	0.123	0.184	0.173	0.150
	CI cov	0.925	0.890	0.954	0.955
	Power	0.572	0.825	0.656	0.764
2:1					
6×48	$\overline{\text{HR}}$	0.730	0.668	0.673	0.674
	% bias	−19.97	2.450	0.641	0.370
	SD	0.141	0.158	0.159	0.154
	CI cov	0.939	0.945	0.948	0.950
	Power	0.580	0.768	0.736	0.755
8×18	$\overline{\text{HR}}$	0.740	0.644	0.674	0.674
$+24 \times 6$	% bias	−23.50	12.22	0.916	0.403
	SD	0.134	0.177	0.172	0.158
	CI cov	0.935	0.917	0.954	0.952
	Power	0.545	0.780	0.665	0.732
48×6	$\overline{\text{HR}}$	0.742	0.628	0.675	0.675
	% bias	−24.41	19.35	0.938	0.133
	SD	0.132	0.195	0.182	0.160
	CI cov	0.933	0.893	0.951	0.950
	Power	0.535	0.781	0.620	0.719

Model (1): unadjusted Cox model; model (2): fixed effects Cox model; model (3): stratified Cox model; model (4): semi-parametric gamma frailty model.

Table III. Simulation results (30% censoring; Kendall's $\tau = 0.20$; $\beta = \log(1)$) under correct specification of the frailty distribution.

		Model			
Sample size	Statistic	(1)	(2)	(3)	(4)
1:1					
6×48	$\overline{HR}$	1.007	1.009	1.010	1.009
	SD	0.128	0.149	0.151	0.146
	Size	0.031	0.055	0.052	0.052
8×18	$\overline{HR}$	1.010	1.016	1.015	1.013
$+24 \times 6$	SD	0.124	0.170	0.164	0.150
	Size	0.024	0.074	0.050	0.047
48×6	$\overline{HR}$	1.011	1.020	1.017	1.014
	SD	0.123	0.187	0.173	0.152
	Size	0.022	0.090	0.049	0.051
2:1					
6×48	$\overline{HR}$	1.008	1.011	1.012	1.011
	SD	0.134	0.157	0.158	0.154
	Size	0.028	0.053	0.049	0.048
8×18	$\overline{HR}$	1.007	1.014	1.013	1.010
$+24 \times 6$	SD	0.129	0.178	0.171	0.157
	Size	0.022	0.074	0.050	0.048
48×6	$\overline{HR}$	1.006	1.017	1.016	1.011
	SD	0.130	0.198	0.183	0.162
	Size	0.024	0.094	0.050	0.053

Model (1): unadjusted Cox model; model (2): fixed effects Cox model; model (3): stratified Cox model; model (4): semi-parametric gamma frailty model.

observe that the population-averaged effect is attenuated compared with the centre-specific effect. Under the null hypothesis of no treatment effect (Table III), HR is well estimated (as there is no room for attenuation), but it can be seen from the type I error rate that ignoring the clustering leads to results that are too conservative. For more general results regarding the omission of important risk factors from nonlinear regression models, see [14].

4.2.2. 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 for $s = 8 + 24$ and for $s = 48$. It produces estimates that are biased away from the true β, and the coverage of the confidence interval (respectively the type I error rate) is below 95% (respectively above 5%). Regarding multicentre clinical trials, the fixed effects approach further shows additional limitations. (i) It implicitly assumes that the centres participating in the trial 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 [15]. (iii) It might be of interest to assess whether a covariate explains heterogeneity in outcome between centres [16]. It is, however, unfeasible in this model to include a covariate whose values only change at the centre level. (iv) Precision in centre effects estimates is dependent upon the centre size. Interpretation can therefore be misleading. A related problem is that the centre effects estimates (and their interpretation) depend on the choice of the reference centre, which is generally arbitrary.

4.2.3. The stratified approach. Model (3) performs well, with good point estimates and good coverage probabilities. However, no between-centre comparisons are made by the stratified approach which, therefore, does not make optimal use of all the information at hand (only within-centre comparisons are made). This explains why both the standard deviation inflates and the power deteriorates when the centre size decreases. Besides, similar to the fixed effects approach, (i) interpretation of the treatment effect is restricted to participating centres, (ii) no heterogeneity measure is returned, and (iii) centre-specific covariates cannot be investigated because no between-centre comparisons are made by the stratified approach.

4.2.4. The frailty approach. Model (4) shows good performances in every investigated setting with virtually no bias and good cov-

erage probabilities. Unlike the stratified model, the frailty model also makes use of between-centre comparisons to gather information on the treatment effect. This explains why both the standard deviation is smaller and the power is better for the frailty model than for the stratified model. The frailty modelling approach further provides a rich framework for the analysis of multicentre clinical trials. (i) Because of their random nature, the actual values 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 that determines the degree of heterogeneity between centres. To help interpretation, this parameter can further be translated into clinically relevant quantities like the spread in the median time to event (as we did in Figure 1) or in the 5-year survival rate from centre to centre [11]. Alternatively, the θ parameter can be transformed into the Kendall's τ that measures the degree of association between outcomes within the same centre [17, Section 4.2 and Section 7.2.5]. For gamma frailties, Kendall's τ is $\tau = \theta/(\theta + 2)$. (iii) Considering the u_i's as random effects parameters also makes it possible to study whether the inclusion of a centrespecific covariate explains/reduces heterogeneity between centres [16].

5. ROBUSTNESS AGAINST MISSPECIFICATION OF THE FRAILTY DISTRIBUTION

Different distributions can be used to model the frailty term. Diagnostic checks to assess the frailty distribution are not yet widely available (particularly in software), and research is still needed in this area. In the meantime, it is important to investigate robustness properties against misspecification of the frailty distribution via simulations.

The most common assumption, mainly made for mathematical convenience and software availability rather than for clinical or empirical (data-driven) evidence, is that the frailties have a gamma distribution. Therefore, the most common form of misspecification is that of using the gamma distribution while the frailties actually follow another distribution. Alternative distributions that have received interest to model the frailty term include the inverse Gaussian, log-normal, and positive stable distributions [3, Chapter 4].

To observe the impact of misspecifying the frailty distribution on the inferences for the treatment effect (and more generally for the fixed effects parameters included in the model), we simulate data from model (4) (cf. Section 4.1) using the inverse Gaussian, log-normal, and positive stable distributions to generate the frailties, and we fit the misspecified gamma frailty model. For each frailty distribution, the heterogeneity parameter is chosen to yield a Kendall's tau of $\tau = 0.20$, as earlier. Additional simulations with $\tau = 0.40$ are given in the supplementary material (available online as supporting information).

By comparing the results obtained under misspecification in Table IV (respectively Table V) with those obtained under correct specification in Table II (respectively Table III), it appears that inferences on the fixed effect parameter β are robust against misspecification of the frailty distribution. In particular, the frailty approach performs better than the competing stratified approach in terms of power in either misspecified situation.

Table IV. Simulation results (30% censoring; 1:1; Kendall's $\tau = 0.20$; $\beta = \log(2/3)$) under misspecification of the frailty distribution.

		True frailty distribution					
		IG		LN		PS	
		Model					
Sample size	Statistic	(3)	(4)	(3)	(4)	(3)	(4)
6×48	$\overline{HR}$	0.672	0.673	0.672	0.673	0.674	0.675
	% bias	0.900	0.341	0.800	0.358	0.018	−0.440
	SD	0.151	0.146	0.151	0.147	0.150	0.146
	CI cov	0.947	0.948	0.952	0.953	0.952	0.950
	Power	0.784	0.806	0.778	0.793	0.778	0.793
8×18	$\overline{HR}$	0.673	0.674	0.673	0.674	0.675	0.676
$+24 \times 6$	% bias	1.250	0.113	1.106	0.174	0.526	−0.677
	SD	0.166	0.152	0.165	0.151	0.167	0.151
	CI cov	0.950	0.948	0.951	0.950	0.952	0.950
	Power	0.710	0.774	0.707	0.778	0.695	0.765
48×6	$\overline{HR}$	0.674	0.675	0.674	0.676	0.675	0.676
	% bias	1.033	−0.201	1.005	−0.400	0.812	−0.687
	SD	0.172	0.152	0.176	0.154	0.174	0.152
	CI cov	0.954	0.950	0.950	0.948	0.951	0.950
	Power	0.660	0.758	0.654	0.756	0.654	0.757

Model (3): stratified Cox model; model (4): semi-parametric gamma frailty model.

Table V. Simulation results (30% censoring; 1:1; Kendall's $\tau = 0.20$; $\beta = \log(1)$) under misspecification of the frailty distribution.

		True frailty distribution					
		IG		LN		PS	
		Model					
Sample size	Statistic	(3)	(4)	(3)	(4)	(3)	(4)
6×48	$\overline{HR}$	1.010	1.009	1.012	1.011	1.010	1.010
	SD	0.150	0.145	0.149	0.144	0.148	0.144
	Size	0.054	0.050	0.048	0.044	0.051	0.050
8×18	$\overline{HR}$	1.012	1.010	1.018	1.015	1.014	1.012
$+24 \times 6$	SD	0.166	0.151	0.165	0.150	0.162	0.149
	Size	0.055	0.052	0.054	0.049	0.049	0.048
48×6	$\overline{HR}$	1.015	1.012	1.015	1.011	1.014	1.012
	SD	0.172	0.152	0.172	0.153	0.170	0.152
	Size	0.049	0.051	0.047	0.051	0.045	0.049

Model (3): stratified Cox model; model (4): semi-parametric gamma frailty model.

6. CONCLUSIONS

In clinical trials with a time-to-event outcome, the primary analysis is commonly based on model (1) with a single covariate for the treatment group or on the equivalent log-rank test. When the trial is conducted at multiple centres, the treatment effect obtained from model (1) has a population-averaged (marginal) interpretation, the effect being averaged over all centres, rather than a centre-specific (conditional) interpretation. Of note, model (1) leads to a consistent estimate of the population hazard ratio, but the standard error is not consistent and a robust estimator that copes with the clustering should be used (cf. the `cluster()` function in `coxph()`) [3, Section 3.4]. The centre-specific treatment effect, that is, the ratio of the hazard rate of a treated subject versus an untreated subject from the same centre, is particularly relevant in the context of clinical trials as it compares 'like-for-like'. Models (2)–(4) allow a centre-specific interpretation of the treatment effect.

Important conclusions from our simulations are as follows:

- The population-averaged effect is attenuated compared with the centre-specific effect.
- Ignoring the clustering leads to results that are too conservative.
- The centre-specific treatment effect is usually biased when it is estimated from the fixed effects Cox model (2).
- Power is lost when fitting the stratified Cox model (3) compared with the frailty model (4).
- Inferences on the centre-specific treatment effect obtained from the frailty model (4) are robust against misspecification of the frailty distribution in many settings.

In the light of these results, we recommend to use the frailty model, which is now readily available in standard software (e.g. R and SAS), to adjust for centre heterogeneity in multicentre clinical trials with a time-to-event outcome.

Acknowledgements

We are grateful to the associate editor and to the two reviewers for their valuable insights and helpful comments. M. Munda is supported by a F. R. I. A. fellowship. C. Legrand is supported by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

REFERENCES

[1] Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine* 2004; **23**:369–388.
[2] Gill RD. Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association* 1984; **79**:441–447.
[3] Duchateau L, Janssen P. *The frailty model*. Springer: New York, 2008.
[4] Wienke A. *Frailty models in survival analysis*. Chapman and Hall/CRC: Boca Raton, 2010.
[5] Cortiñas Abrahantes J, Legrand C, Burzykowski T, Janssen P, Ducrocq V, Duchateau L. Comparison of different estimation procedures for proportional hazards model with random effects. *Computational Statistics & Data Analysis* 2007; **51**:3913–3930.
[6] Munda M, Rotolo F, Legrand C. parfm: parametric frailty models in R. *Journal of Statistical Software* 2012; **51**:1–20.
[7] Hirsch K, Wienke A. Software for semiparametric shared gamma and log-normal frailty models: an overview. *Computer Methods and Programs in Biomedicine* 2012; **107**:582–597.
[8] Therneau TM, Grambsch PM. *Modeling survival data: extending the cox model*. Springer: New York, 2000.
[9] Clahsen PC, van de Velde CJ, Julien JP, Floiras JL, Delozier T, Mignolet FY, Sahmoud TM. Improved local control and disease-free survival after perioperative chemotherapy for early-stage breast cancer. A European Organization for Research and Treatment of Cancer Breast Cancer Cooperative Group Study. *Journal of Clinical Oncology* 1996; **14**:745–753.
[10] van der Hage JA, van de Velde CJ, Julien JP, Floiras JL, Delozier T, Vandervelden C, Duchateau L. Improved survival after one course of perioperative chemotherapy in early breast cancer patients: long-term results from the European Organization for Research and

Treatment of Cancer (EORTC) Trial 10854. *European Journal of Cancer* 2001; **37**:2184–2193.
[11] Duchateau L, Janssen P. Understanding heterogeneity in generalized mixed and frailty models. *The American Statistician* 2005; **59**:143–146.
[12] Collett D. *Modelling survival data in medical research*. Chapman and Hall/CRC: London, 2003.
[13] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**:4279–4292.
[14] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* 1998; **19**:249–256.
[15] Andersen PK, Klein JP, Zhang MJ. Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* 1999; **18**:1489–1500.
[16] Legrand C, Duchateau L, Sylvester R, Janssen P, van der Hage JA, van de Velde CJ, Therasse P. Heterogeneity in disease free survival between centers: lessons learned from an EORTC breast cancer trial. *Clinical Trials* 2006; **3**:10–18.
[17] Hougaard P. *Analysis of multivariate survival data*. Springer: New York, 2000.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

APPENDIX

Models (1)–(4) can be fitted in R by means of `coxph()` (part of the `survival` library) and in SAS by means of `proc phreg`. In the generic sample codes below, `data` has the following columns:

- `cluster`: cluster identification number;
- `time`: minimum between the actual event time and the censoring time;
- `status`: 1 if the observation is an event, 0 if it is right-censored;
- `x`: treatment group indicator (0 or 1).

R CODES

```
# unadjusted Cox model (1)
coxph(Surv(time, status) ~ x, data=data)

# fixed effects Cox model (2)
coxph(Surv(time, status) ~ x + factor(cluster),
data=data)

# stratified Cox model (3)
coxph(Surv(time, status) ~ x + strata(cluster),
data=data)

# semi-parametric gamma frailty model (4)
coxph(Surv(time, status) ~ x +
    frailty.gamma(x=cluster, eps=1e-11),
   outer.max=50, data=data)
```

SAS CODES

```
/* unadjusted Cox model (1) */
 proc phreg data=data; class x(ref="0");
 model time*status(0) = x / ties=efron;
 run;

/* fixed effects Cox model (2) */
 proc phreg data=data;
 class x(ref="0") cluster(ref="1");
 model time*status(0) = x cluster / ties=efron;
 run;

/* stratified Cox model (3) */
 proc phreg data=data; class x(ref="0");
 model time*status(0) = x / ties=efron;
 strata cluster;
 run;

/* semi-parametric log-normal frailty model (4) */
/* !!! gamma frailty dist not yet available !!! */
 proc phreg data=data;
 class x(ref="0") cluster;
 model time*status(0) = x / ties=efron;
 random cluster / method=REML;
 run;
```