Resampling Plans for Frailty Models

Non Peer-reviewed author version

MASSONNET, Goele; BURZYKOWSKI, Tomasz & JANSSEN, Paul (2006) Resampling Plans for Frailty Models. In: COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION, 35(2). p. 497-514.

DOI: 10.1080/03610910600591586
Handle: http://hdl.handle.net/1942/1793

Resampling plans for frailty models

Goele Massonnet *, Tomasz Burzykowski, Paul Janssen

Hasselt University, Center for Statistics, Agoralaan, Diepenbeek, Belgium.

Abstract

Obtaining the standard error of the estimated heterogeneity in shared frailty models is in general difficult. Klein and Moeschberger (1997) show that the use of the observed information matrix is often not feasible because of its high dimension. Therneau and Grambsch (2000) use a nonparametric bootstrap algorithm to obtain standard errors for the estimated parameters in a shared frailty model. For parametric shared frailty models we define two model-based resampling schemes and use them to obtain standard errors. Based on a simulation study, we show that model-based resampling compares favourable to nonparametric resampling and that for all resampling schemes robustness is an issue of concern.

Key Words: Shared frailty model; variance estimation; model-based bootstrap.

*Correspondance: Goele Massonnet, Hasselt University, Center for Statistics, Agoralaan, Building D, B-3590 Diepenbeek, Belgium; Tel: +32-11-268244; Fax:+32-11-268299; E-mail: goele.massonnet@uhasselt.be.

1 Introduction

The shared frailty model is used in order to model correlated survival times. The unobserved risk factor that is common for all the observations in the same cluster is called the frailty. A commonly used estimation procedure in frailty models is the EM algorithm (Klein, 1992). The EM algorithm provides estimates for the fixed effects and for the variance of the frailty density, but does not automatically provide estimates for the variances of these estimates. Klein and Moeschberger (1997, p.413) show how the standard errors of the estimates for the gamma frailty model can be obtained from the inverse of the observed information matrix. This information matrix has rank equal to the number of distinct event times plus the number of covariates plus one (for the heterogeneity parameter). For large data sets, this procedure is not appropriate because of the high dimensionality.

For the gamma frailty model, Therneau and Grambsch (2000, p.254) proved that the estimates obtained from the penalized partial likelihood maximization coincide with the estimates obtained from the EM algorithm for any fixed value of the heterogeneity parameter. Hence we can use the fast algorithm for the penalized partial likelihood procedure available in S-Plus. However, the standard error estimates reported in S-Plus are computed under the assumption of fixed θ. Since θ needs to be estimated, the given standard errors are too small (Therneau and Grambsch, 2000, p.249).

Thus, the issue of estimating the standard errors of the parameter estimates requires further investigation. A useful tool might be the bootstrap. The results developed for resampling in linear mixed models show that resampling schemes need to be chosen in a careful way (Davison and Hinkley, 1997, p.100-102; Morris, 2002). Therneau and Grambsch (2000, p.249) proposed a nonparametric bootstrap algorithm to obtain standard error estimates. For parametric frailty

models, model-based resampling schemes might be preferred above nonparametric resampling plans. In this paper we propose two model-based resampling plans that can be used to find standard errors of the estimated parameters (Section 4). We compare the two model-based bootstrap algorithms to the nonparametric resampling algorithm of Therneau and Grambsch (2000). The comparison is based on a simulation study (Section 5). The results indicate that one of the proposed algorithms provides precise assessment of the empirical variability of the parameter estimates, even if the model is misspecified. Another important finding is that the empirical variability of the heterogeneity parameter can be much different for the correct and the misspecified model. This provides evidence that robustness in terms of the heterogeneity parameter is not guaranteed for the bootstrap algorithms (including the nonparametric bootstrap); but robustness holds for the fixed effects. Prior to the discussion on resampling schemes we give a short review on frailty models (Section 2) and on estimation methods for frailty models (Section 3). In Section 6 we collect main conclusions and further research questions.

2 The shared frailty model

Assume we have a total of N individuals that come from K different groups, group i having n_i individuals ($N = \sum_{i=1}^{K} n_i$). Each subject is observed from a time zero to a failure time T_{ij}^0 or to a potential right censoring time C_{ij}. Let $T_{ij} = \min(T_{ij}^0, C_{ij})$ be the observed time and δ_{ij} be the censoring indicator which is equal to 1 if $T_{ij} = T_{ij}^0$ and 0 otherwise. Hence the observed data available for the jth individual in the ith group is $y_{ij} = (T_{ij}, \delta_{ij})$, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$. The number of observed events in group i is $D_i = \sum_{j=1}^{n_i} \delta_{ij}$.

The frailty model is given by

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

where $h_{ij}(t)$ is the hazard rate at time t for individual j from group i, $h_0(t)$ is the baseline hazard at time t, x_{ij} is the vector of p covariates recorded for the individual and w_i is the random effect for group i. In this model $h_0(t)$ can be left unspecified or it may be assumed to have some specific parametric form. The w_i's, $i = 1, \ldots, K$, are a sample (independent and identically distributed) from a density $f_W(.)$.

Model (1) can be rewritten as:

$$h_{ij}(t) = h_0(t)u_i \exp(x_{ij}^T \beta).$$

The factor $u_i = \exp(w_i)$ is termed the frailty for the ith group. The following choices for the frailty density will be considered:

(a) The one-parameter gamma density of the form

$$f_U(u) = \frac{u^{(1/\theta)-1} \exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}, \quad \theta > 0.$$

The corresponding density for W is

$$f_W(w) = \frac{\{\exp(w)\}^{1/\theta} \exp\{-\exp(w)/\theta\}}{\theta^{1/\theta}\Gamma(1/\theta)}.$$

For the gamma density $\mathrm{E}(U) = 1$. Typically $\mathrm{Var}(U) = \theta$ is used to describe heterogeneity.

(b) The one-parameter normal density for W with $\mathrm{E}(W) = -\sigma^2/2$ and $\mathrm{Var}(W) = \sigma^2$.

The corresponding density of U is

$$f_U(u) = \frac{1}{u\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(\log u + \frac{\sigma^2}{2}\right)^2}{2\sigma^2} \right\},$$

with $\mathrm{E}(U) = 1$ and $\mathrm{Var}(U) = e^{\sigma^2} - 1$. In the further discussion, σ^2 is chosen so that $\mathrm{Var}(U) = \theta$, i.e., $\sigma^2 = \log(\theta + 1)$.

We will use $\mathrm{Var}(U) = \theta$ to describe heterogeneity for both frailty distributions.

3 Methods of estimation for the shared frailty model

For the gamma frailty model, Klein (1992) shows that the observable (marginal) likelihood is given by

$$\begin{aligned} l_{obs}(\beta,\theta,h_0(.)) &= \sum_{i=1}^{K}\Bigg[D_i\log\theta-\log\Gamma(\frac{1}{\theta})+\log\Gamma(\frac{1}{\theta}+D_i) \\ &\quad -\left(\frac{1}{\theta}+D_i\right)\log\{1+\theta\sum_{j=1}^{n_i}H_0(t_{ij})\exp(x_{ij}^T\beta)\} \\ &\quad +\sum_{j=1}^{n_i}\delta_{ij}\{x_{ij}^T\beta+\log h_0(t_{ij})\}\Bigg], \end{aligned} \tag{2}$$

where $H_0(t)=\int_0^t h_0(u)du$ is the cumulative baseline hazard.

As noted in the previous section, the baseline hazard $h_0(t)$ in the frailty model can be specified explicitly or left unspecified. Under the parametric assumption, the parameters in the resulting model can be estimated using maximum likelihood estimation procedures. For example, for $h_0(t)\equiv h_0$ constant, the parameters β, θ and h_0 can be estimated by maximizing the observable log likelihood $l_{obs}(\beta,\theta,h_0)$. If $h_0(t)$ is left unspecified, the EM algorithm (Klein, 1992) and the penalized partial likelihood approach (Therneau and Grambsch, 2000) can be used to estimate the unknown parameters in (2). The latter can also be used to estimate the parameters of the lognormal frailty model.

The EM algorithm for the gamma frailty

To estimate $\zeta=(\theta,\beta)$, we would like to base the likelihood maximization on the observable log likelihood (2). However, this likelihood is difficult to maximize as it contains, apart from ζ, also the unspecified baseline hazard. We therefore rely on the EM algorithm to estimate ζ (for details see, e.g., Duchateau et al., 2002).

It is worth noting that Therneau and Grambsch (2000, p.254) have shown that for any fixed θ, the EM algorithm and the penalized partial likelihood maximization have the same solution for the gamma frailty case. Since S-Plus contains a fast algorithm for the penalized partial likelihood approach, this property is very important from a practical point of view.

The penalized partial likelihood for shared frailty models

An alternative proposal for the likelihood to use for the estimation of $\zeta = (\theta, \beta)$ is the penalized partial likelihood

$$l_{ppl}(\zeta, w) = l_{part}(\zeta, w) - l_{pen}(\zeta, w),$$

where

$$l_{part}(\zeta, w) = \sum_{l=1}^{r} \left[\sum_{t_{ij}=t_{(l)}} \eta_{ij} - N_{(l)} \log \left\{ \sum_{t_{qs} \geq t_{(l)}} \exp(\eta_{qs}) \right\} \right],$$

with $\eta_{ij} = x_{ij}^T \beta + w_i$, r denoting the number of different event times, $t_{(1)} \leq \ldots \leq t_{(r)}$ being the ordered event times, $N_{(l)}$ denoting the number of events at time $t_{(l)}$, $l = 1, \ldots, r$ and

$$l_{pen}(\theta, w) = -\sum_{i=1}^{K} \log f_W(w_i).$$

For random effects w_i, $i = 1, \ldots, K$, with corresponding one-parameter gamma density for the frailties, we have

$$l_{pen}(\theta, w) = -\sum_{i=1}^{K} \left\{ \frac{w_i - \exp(w_i)}{\theta} \right\} - K \left\{ \frac{\log \theta}{\theta} - \log \Gamma \left(\frac{1}{\theta} \right) \right\}.$$

The maximization of the penalized log likelihood consists of an inner and an outer loop. In the inner loop the Newton-Raphson procedure is used to maximize, for a provisional value of θ, $l_{ppl}(\zeta, w)$ for β and w. In the outer loop, a likelihood similar to (2) is maximized for θ as in the case of the EM algorithm. The process is iterated until convergence (for details see, e.g., Duchateau et al., 2002).

For random effects w_i, $i = 1, \ldots, K$, having a normal density, we have

$$l_{pen}(\sigma^2, w) = \frac{1}{2} \sum_{i=1}^{K} \left\{ \frac{(w_i - \mu)^2}{\sigma^2} + \log(2\pi\sigma^2) \right\}.$$

This term penalizes random effects that are far away from the mean value by reducing the penalized partial likelihood. The maximization of the penalized log likelihood consists of an inner and an outer loop. The inner loop is identical to the one described for gamma frailty parameters. In the outer loop, the restricted maximum likelihood estimator for σ^2 is obtained using BLUPs. The process is iterated until convergence.

4 Bootstrap : Resampling schemes

The EM algorithm does not provide estimates for the variances of the estimates in the frailty model. Klein and Moeschberger (1997) determine the standard errors of the estimates of β and θ from the inverse of the observed information matrix of the observable likelihood. The information matrix is a square matrix of size $r + p + 1$. For large data sets, this approach is not appropriate because of the high dimensionality. On the other hand, the standard error estimates of $\hat{\beta}$ reported by S-Plus are computed under the assumption of θ known (Therneau and Grambsch, 2000, p.249). In many situations, this assumption is not true and the estimated standard errors are too small. An alternative approach for finding variance estimates might be provided by the bootstrap.

Therneau and Grambsch (2000, p.249) proposed the following nonparametric bootstrap technique to obtain standard error estimates:

1. Choose K groups by sampling with replacement from the K groups in the study.

2. The bootstrap sample contains the subjects from the selected groups.

3. Fit a gamma or lognormal frailty model with covariates to this bootstrap sample.

This procedure is repeated a number of times. The estimates of the coefficients $\hat{\beta}^*$ and the estimates of the heterogeneity parameter $\hat{\theta}^*$ are stored for each bootstrap sample. The standard errors of the estimated parameters $\hat{\beta}$ and $\hat{\theta}$ are calculated based on the variability of $\hat{\beta}^*$ and $\hat{\theta}^*$. If a parametric model is appropriate, we might prefer model-based resampling techniques above the nonparametric resampling plan. We therefore propose two model-based resampling schemes. We rely on a resampling plan for a simple random effects model with a balanced design, proposed by Davison and Hinkley (1997, p.102). A random effects model can be written as

$$y_{ij} = x_i + z_{ij}, \quad j = 1, \ldots, n_i = n, \quad i = 1, \ldots, K,$$

where K is the number of groups, $n_i = n$ is the number of subjects per group, the x_i's are randomly sampled from F_x and independent of the z_{ij}'s, which are randomly sampled from F_z with $E(Z) = 0$ to force uniqueness of the model.

In the "naive" version of their algorithm, Davison and Hinkley (1997, p.102) define

$$\hat{x}_i = \bar{y}_i \quad \text{and} \quad \hat{z}_{ij} = y_{ij} - \bar{y}_i \quad .$$

The resampled data set is then obtained in the following way

1. Choose $x_1^*, \ldots, x_K^*$ by randomly sampling with replacement from $\hat{x}_1, \ldots, \hat{x}_K$;
2. Choose $z_{i1}^*, \ldots, z_{in}^*$ randomly with replacement from one group of residuals $\hat{z}_{k1}, \ldots, \hat{z}_{kn}$, either from a randomly selected group or the group corresponding to x_i^*;
3. Set $y_{ij}^* = x_i^* + z_{ij}^*, \quad j = 1, \ldots, n, \quad i = 1, \ldots, K.$

To construct a resampling plan for frailty models, we can argue that sampling from the means of the groups in the case of the random effects model is like sampling from the frailty estimates

in the case of the frailty model. However, in the situation of frailty models, we do not have any residuals to resample from. We therefore propose a new resampling scheme that extends a resampling algorithm for independent survival times, proposed by Hjort (1985) (see also Davison and Hinkley, 1997, p.351).

Model-based bootstrap, algorithm 1:

For $j = 1, \ldots, n_i$, $i = 1, \ldots, K$,

1. Fit the model; obtain the estimate $\hat{\beta}$ and the estimates of the frailties $\hat{u}_1, \ldots, \hat{u}_K$.

2. Choose $u_1^*, \ldots, u_K^*$ by sampling with replacement from $\hat{u}_1, \ldots, \hat{u}_K$.

3. Generate the true failure time T_{ij}^* from the estimated failure time survivor function $\hat{S}_{ij}(t) = \{\hat{S}_0(t)\}^{u_i^* \exp(x_{ij}^{T*}\hat{\beta})}$, where x_{ij}^{T*} is the vector of covariates recorded for the jth individual from the cluster that corresponds to u_i^*.

4. Let $\tilde{\delta}_{ij}^*$ and $\tilde{T}_{ij}^{0*}$ be the censoring indicator and the observed time for the jth individual from the cluster that corresponds to u_i^*. If $\tilde{\delta}_{ij}^* = 0$, set $C_{ij}^* = \tilde{T}_{ij}^{0*}$, and if $\tilde{\delta}_{ij}^* = 1$, generate C_{ij}^* from the conditional censoring distribution given that $C_{ij} > \tilde{T}_{ij}^{0*}$, namely

$$\frac{\hat{G}(t) - \hat{G}(\tilde{T}_{ij}^{0*})}{1 - \hat{G}(\tilde{T}_{ij}^{0*})},$$

   where $\hat{G}$ is an estimate (e.g., Kaplan-Meier) of the common censoring distribution G. Assume that G is independent of the covariates.

5. Set $T_{ij}^{0*} = \min(T_{ij}^*, C_{ij}^*)$, with $\delta_{ij}^* = 1$ if $T_{ij}^{0*} = T_{ij}^*$ and zero otherwise.

Steps 3, 4 and 5 are the adaption of the algorithm proposed by Hjort (1985) (see also Davison and Hinkley, 1997, p.351).

For a semi-parametric model, the true failure times in step 3 are generated from the estimated failure time survival function

$$\hat{S}_{ij}(t) = \{\hat{S}_0(t)\}^{u_i^* \exp(x_{ij}^{T*}\hat{\beta})},$$

where $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$ is the estimated baseline survival function, with

$$\hat{H}_0(t) = \sum_{t_{(l)} \leq t} \hat{h}_{l0},$$

where $\hat{H}_0(t)$ is the estimated baseline cumulative hazard at time t and

$$\hat{h}_{l0} = \frac{N_{(l)}}{\sum_{t_{sq} \geq t_{(l)}} u_s^* \ \exp(x_{sq}^{T*}\hat{\beta})}.$$

For a parametric model, the true failure times are generated under the parametric assumption.

For mixed models it has been demonstrated (Morris, 2002) that the variances of the BLUP's are biased downwards as estimators of the variance components. Due to this bias, bootstrapping BLUP's results in underestimation of the variation in the data, causing standard error estimates biased downwards. The above-mentioned model-based resampling algorithm may suffer from this problem. Therefore, we propose a second resampling scheme, where resampled frailty parameters are obtained by sampling the appropriate frailty distribution with variance $\hat{\theta}$. We again assume that censoring is independent of the covariates.

Model-based bootstrap, algorithm 2:

For $j = 1, \ldots, n_i$, $i = 1, \ldots, K$,

1. Fit the model; obtain the estimates $\hat{\beta},\ \hat{\theta}$.
2. Sample $u_1^*, \ldots, u_K^*$ from a gamma or lognormal distribution with mean 1 and variance $\hat{\theta}$.
3. Generate the true failure time T_{ij}^* from the estimated failure time survivor function $\hat{S}_{ij}(t) = \{\hat{S}_0(t)\}^{u_i^* \exp(x_{ij}^T\hat{\beta})}$.

4. If $\delta_{ij} = 0$, set $C^*_{ij} = T^0_{ij}$, and if $\delta_{ij} = 1$, generate C^*_{ij} from the conditional censoring distribution given that $C_{ij} > T^0_{ij}$, namely

$$\frac{\hat{G}(t) - \hat{G}(T^0_{ij})}{1 - \hat{G}(T^0_{ij})}.$$

5. Set $T^{0*}_{ij} = \min(T^*_{ij}, C^*_{ij})$, with $\delta^*_{ij} = 1$ if $T^{0*}_{ij} = T^*_{ij}$ and zero otherwise.

5 Simulations

5.1 Motivation

Based on simulations we compare the two model-based resampling plans and the nonparametric resampling plan. As simulation model we consider the setting of a multicenter clinical trial. The following issues will be discussed:

(i) The comparison of the nonparametric and the model-based resampling schemes assuming that the model is correct.

(ii) The effect of the size of the multicenter clinical trial on the precision of the variance estimation. Note that the size of a trial is determined by K, the number of centers, and by the number of patients per center (which we assume to be equal over the centers for simplicity).

(iii) The effect of the size of θ, the heterogeneity parameter, and $h_0(t)$, the event rate (assumed to be constant in time for simplicity) on the precision of the variance estimation.

(iv) The robustness of the resampling plans to misspecification of the model.

5.2 The simulation setting

For each parameter setting $(K, n, h_0, \theta, \beta)$, with β the treatment effect parameter, 100 data sets are generated. Given a particular parameter setting, the observations for each data set are generated in the following way. First, K frailties $u_1, \ldots, u_K$ are generated from a gamma or lognormal frailty distribution with mean one and variance θ. The time to event for the jth patient from center i is randomly generated from an exponential distribution with parameter $h_{ij} = h_0 u_i \exp(x_{ij}^T \beta)$, where x_{ij} is generated from a Bernoulli distribution with success parameter 0.5. The censoring time for each patient is randomly generated from a uniform distribution so that approximately 30% censoring is obtained.

For each simulated data set, two model assumptions are considered to investigate the performance of the bootstrap algorithms under the correct and misspecified models. First, we assume that the frailties are gamma distributed. For each simulated data set, $R = 100$ bootstrap samples are taken by using the nonparametric bootstrap and the two model-based resampling plans under the assumption of gamma distributed frailties. Next, the same procedure is followed under the assumption of lognormal distributed frailties.

Under both assumptions of the frailty distribution, a semi-parametric frailty model is considered to estimate the treatment effect and the heterogeneity parameter in the nonparametric and the two model-based resampling plans. The penalized partial likelihood approach is used to obtain the parameter estimates (Therneau and Grambsch, 2000). In the model-based resampling plans, we also consider a parametric frailty model with a constant baseline hazard if the frailty parameters are assumed to be gamma distributed. For the parametric gamma frailty model, the model-based resampling schemes assume that the time to event follows an exponential distribution with parameter h_{ij}. Under this assumption, the parameters β, θ and

h_0 can be estimated by maximizing the observable log likelihood $l_{obs}(\beta, \theta, h_0)$, given in (2), using the Newton-Raphson method.

5.3 Choice of the parameters

For the concrete simulation, the number of centers is taken equal to 15 or 30 centers, with 20 or 40 patients per center. For 'true' frailties that are gamma distributed, we additionally consider 15 or 30 centers, with 5 patients per center. The parameter values h_0, β and θ are chosen in such a way that a different magnitude of spread in the median time to event from center to center is induced. The median time to event T_{M_0} (for $x_{ij} = 0$) and T_{M_1} (for $x_{ij} = 1$) is the solution of $\exp(-h_0 U T_{M_0}) = 0.5$ and $\exp(-h_0 U \exp(\beta) T_{M_1}) = 0.5$, with U one-parameter gamma distributed, i.e. $T_{M_0} = \frac{\log 2}{h_0 U}$ and $T_{M_1} = \frac{\log 2}{h_0 U \exp(\beta)}$. The magnitude of spread in the median time to event from center to center was determined by computing the density functions of T_{M_0} and T_{M_1} (Figure 1). It can be shown that, for $x_{ij} = 1$ and for a gamma frailty density, the density function $f_{T_{M_1}}(t)$ is given by

$$f_{T_{M_1}}(t) = \left(\frac{\log 2}{\theta h_0 \exp(\beta)}\right)^{\frac{1}{\theta}} \frac{1}{\Gamma(1/\theta)} \left(\frac{1}{t}\right)^{1+1/\theta} \exp\left(-\frac{\log 2}{\theta t h_0 \exp(\beta)}\right).$$

For the treatment effect, we use $\beta = 0.25$. As true values for the event rate, we take $h_0 = 0.1$ and $h_0 = 0.5$. The heterogeneity parameter is set at $\theta = 0.1$ and $\theta = 0.6$.

For the settings (θ, h_0) = (0.6, 0.5) and (0.1, 0.5), there is little spread in the median time to event over the centers, with a bigger spread for $\theta = 0.6$. For the settings (θ, h_0) = (0.6, 0.1) and (0.1, 0.1), there is much spread in the median time to event over the centers. Furthermore, Figure 2 clearly explains our motivation for choosing $\theta = 0.1$ and $\theta = 0.6$. For $\theta = 0.1$ we have a situation where the gamma and the lognormal density functions are close, whereas for $\theta = 0.6$ these densities are more apart.

5.4 Results

By performing the bootstrap, we obtain for each simulated data set a bootstrap estimate of the standard error of the treatment effect and the heterogeneity parameter. The mean of these 100 estimated standard errors is denoted by mean(SE^B). The values of mean(SE^B) for each resampling scheme are compared to the empirical standard error of $\hat{\beta}$ and $\hat{\theta}$, denoted by SE^E. In the following discussion we will focus on the standard error estimates of the heterogeneity. For completeness, the results for the treatment effect are given in Table 2. In all settings studied, the estimated standard error of the heterogeneity parameter obtained by the first model-based resampling plan underestimates the standard error, as compared to SE^E. Since the estimates obtained by the second model-based resampling plan are in most cases more precise than those obtained by the first model-based bootstrap algorithm, only the results of the second model-based resampling plan are shown.

5.4.1 Nonparametric versus model-based resampling

Figures 3 and 4 are used to compare the nonparametric and the model-based resampling plan assuming that the model is correct. In Figure 3 we consider 'true' frailties that are gamma and for the resampling scheme we rely on penalized partial likelihood with gamma frailties (gam., s.-par. in Table 1). Figure 4 is the equivalent of Figure 3 for 'true' frailties that are lognormal (logn., s.-par. in Table 1). The resampling schemes are compared in terms of the absolute relative bias. Take, e.g., Figure 3 for the setting $(\theta, h_0) = (0.6, 0.5)$. In that picture we plot for the settings $(K, n) = (15, 5), (15, 20), (15, 40), (30, 5), (30, 20)$ and $(30, 40)$ the points (RB_N, RB_{MB}) where RB_N, resp. RB_{MB}, is the absolute relative bias $|\text{mean}(SE^B) - SE^E|/SE^E$ for the nonparametric, resp. model-based, resampling scheme. The actual value for, e.g., $(K, n) = (15, 40)$

and $(30, 40)$ can be obtained from Table 1. In the picture we add the bisector. For 'true' frailties that are lognormal (Figure 4), we do not consider the settings $(K, n) = (15, 5)$ and $(30, 5)$.

There is no single consistent pattern for all settings in the results. When $(\theta, h_0) = (0.1, 0.1)$, model-based resampling has a smaller relative bias compared to the nonparametric resampling plan (i.e., most of the points (RB_N, RB_{MB}) are below the bisector), even if the cluster size is small $(n = 5)$. For $(\theta, h_0) = (0.6, 0.5)$ and $(0.1, 0.5)$ the general conclusion from Figures 3 and 4 is that, unless the cluster size is small $(n = 5)$, the performance of model-based resampling is often better than that of nonparametric resampling. In situations where the nonparametric resampling is better (points above the bisector) the performance of the model-based plan is almost as good as that of the nonparametric resampling plan. More deviation from this is seen for $(\theta, h_0) = (0.6, 0.1)$ where the nonparametric resampling scheme performs clearly better for some settings (K, n).

Based on the bootstrap estimates $\hat{\theta}^*$, we can construct bootstrap confidence intervals. In Table 3 we illustrate this idea. For a nominal coverage of 95%, we give the coverage proportions of the percentile and bias-corrected and accelerated (BCa) intervals for θ when $(\theta, h_0) = (0.6, 0.5)$. To obtain the BCa interval for a bootstrap sample, the acceleration is computed in terms of the jacknife values of $\hat{\theta}$. For clustered data, the jacknife is performed by leaving out one cluster instead of deleting one observation. We see that the coverage proportion of the percentile intervals is not satisfactory (see also Efron and Tibshirani, 1993, p.178) whereas the BCa intervals have a smaller coverage error, especially for the model-based resampling plan. For illustrative purposes, the results are shown for 100 bootstrap samples. A more extensive simulation study of the confidence intervals is a topic for further research.

5.4.2 Effect of the number of clusters and patients on the precision of the variance estimation

To study the effect of the number of clusters and the number of patients per cluster on the standard error we look at Figure 5 where, for the semi-parametric gamma model, we plot for $(K, n) = (15, 5), (15, 20), (15, 40), (30, 5), (30, 20)$ and $(30, 40)$, SE^E, mean(SE^B) for nonparametric resampling and mean(SE^B) for model-based resampling. The empirical standand error SE^E is considered as the reference point. The general conclusion, also based on pictures similar to Figure 5 for the semi-parametric lognormal model and for the parametric gamma model (pictures not shown), is that for both resampling plans the number of clusters is important to obtain accurate standard errors. We also see that, if the number of clusters is large enough (e.g., K=30) we can only improve the accuracy of the standard errors in a moderate way by increasing the number of patients.

5.4.3 Effect of heterogeneity and event rate on the precision of the variance estimation

To study the effect of the heterogeneity and the event rate on the estimated standard error we look at Figure 6 where, for the semi-parametric gamma model, we plot for $(\theta, h_0) = (0.6, 0.5), (0.6, 0.1), (0.1, 0.5)$ and $(0.1, 0.1)$, SE^E, mean(SE^B) for nonparametric resampling and mean(SE^B) for model-based resampling. The empirical standard error SE^E is considered as the reference point. The general conclusion, also based on pictures similar to Figure 6 for the semi-parametric lognormal model and for the parametric gamma model (pictures not shown), is that the bootstrap standard error obtained by both resampling plans are more accurate for small θ, i.e., $\theta = 0.1$. When h_0 increases, the accuracy of the standard errors is improved in a

moderate way, keeping θ constant.

5.4.4 Robustness

In all settings studied, the point estimates of the fixed effect in the correct and the misspecified model are close to each other (Table 2). Also the estimated standard errors of the fixed effect obtained by the nonparametric and the second model-based resampling scheme are similar, even if the model is misspecified. This means that there is robustness in terms of estimation of the fixed effects. This is in agreement with results in, e.g., Pickles and Crouchley (1995).

When $\theta = 0.6$, the point estimates of the heterogeneity parameter in the misspecified model are biased (Table 1). The empirical variability is also quite different for the correct and the misspecified model. For $\theta = 0.1$, the bias of the point estimates is smaller and the difference in variability is less pronounced. This can be explained since there is only little difference in shape between the gamma distribution and the lognormal distribution when $\theta = 0.1$, whereas there is more difference when $\theta = 0.6$ (Figure 2). The relative bias, compared to the empirical standard error, indicates that the estimated standard error of the heterogeneity obtained by the nonparametric and the model-based resampling schemes are close to the corresponding empirical standard error, both for the correct and the misspecified model. So, bootstrap is useful to estimate the standard error of the heterogeneity parameter. However, since the empirical variability of the heterogeneity parameter is rather different for the correct and misspecified model, lack of robustness is an issue when fitting frailty models.

6 Conclusions

In this paper, the use of bootstrap for the estimation of the standard errors of the parameter estimates in a frailty model is proposed. To complement the existing nonparametric resampling plan, we propose two model-based bootstrap algorithms. The comparison between the nonparametric and model-based resampling schemes and the robustness of the schemes to the model assumptions was studied by simulation. The results indicate that the first model-based resampling plan, based on resampling of the estimated frailties, underestimates the empirical variability of the parameter estimates for all settings studied. This corresponds to the conclusion drawn by Morris (2002) for linear mixed models. On the other hand, the second model-based algorithm, based on resampling from the estimated frailty distribution, provides relatively precise estimates, compared to the corresponding empirical variability. In general, unless the cluster size is small, the second model-based resampling plan gives estimates of the standard error of the heterogeneity estimator that are better or almost a good as those obtained by the nonparametric resampling plan. However, the empirical variability of the heterogeneity parameter is rather different for the correct and misspecified models. So, the results indicate that the proposed resampling schemes may offer a useful approach to obtain standard errors for the estimates of the model parameters (the treatment effect and the heterogeneity parameter) but one has to be careful with the variance estimation of the heterogeneity estimator if the model is misspecified. Further investigation of the properties of the resampling plans will be necessary. For instance, in the model-based resampling schemes we have made the assumption that censoring is independent of the covariates. In principle, it should be possible to extend the schemes to the more general situation where the censoring distribution depends on the covariates, using the approach developed by Davison and Hinkley (1997, p.351). Furthermore, it

also would be of interest to consider frailty densities other than gamma and lognormal. These are important topics for further research.

Acknowledgment

The authors gratefully acknowledge the financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

[1] Davison, A.C., Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

[2] 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 clinical trials. *Computational Statistics and Data Analysis* 40:603-620.

[3] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

[4] Hjort, N.L. (1985). Bootstrapping Cox's regression model. Technical Report NSF-241, Department of Statistics, Stanford University.

[5] Johnson, N. L., Kotz, S. (1970). *Continuous univariate distributions-1.* Houghton Mifflin Company: Boston.

[6] Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model-based on the EM algorithm. *Biometrics* 48:795-806.

[7] Klein, J.P., Moeschberger, M.L. (1997). *Survival Analysis, Techniques for Censored and Truncated Data.* Springer, New York.

[8] Morris, J.S. (2002). The BLUPs are not "best" when it comes to bootstrapping. *Statistics & Probability Letters* 56:425-430.

[9] Pickles, A., Crouchley, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine* 14:1447-1461.

[10] Therneau, T.M., Grambsch, P.M. (2000). *Modeling Survival Data, Extending the Cox Model.* Springer, New York.

Figure 1: Density function of the median time to event over centers ($\beta = 0.25$) .

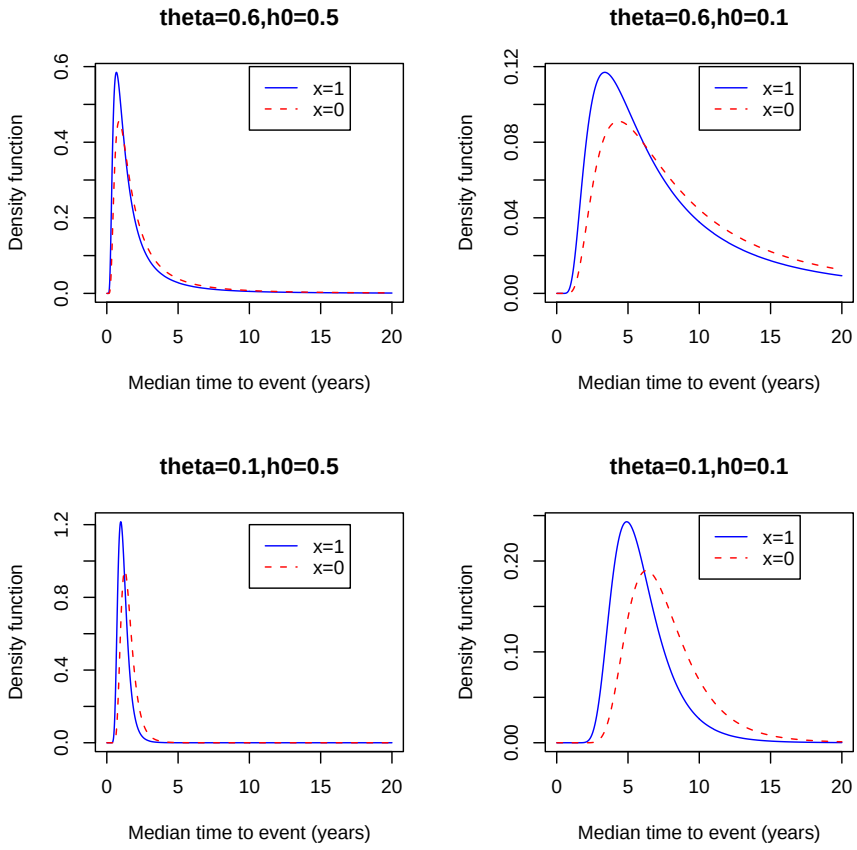

Figure 2: Density function for the lognormal and the gamma distribution .

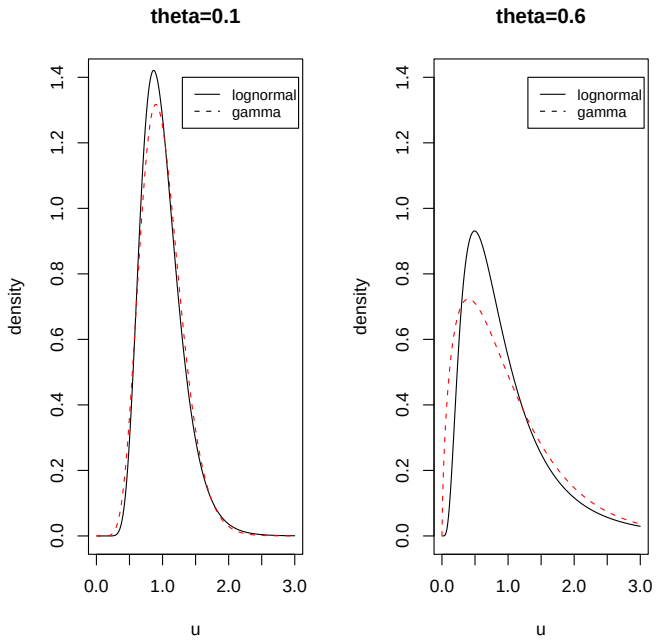

Figure 3: Absolute relative bias for the estimated standard error of the heterogeneity parameter (gamma frailties); $\circ = (15, 5)$, $\circ = (15, 20)$, $\bigcirc = (15, 40)$, $\mathrm{x} = (30, 5)$, $\mathrm{X} = (30, 20)$ $\mathrm{X} = (30, 40)$.

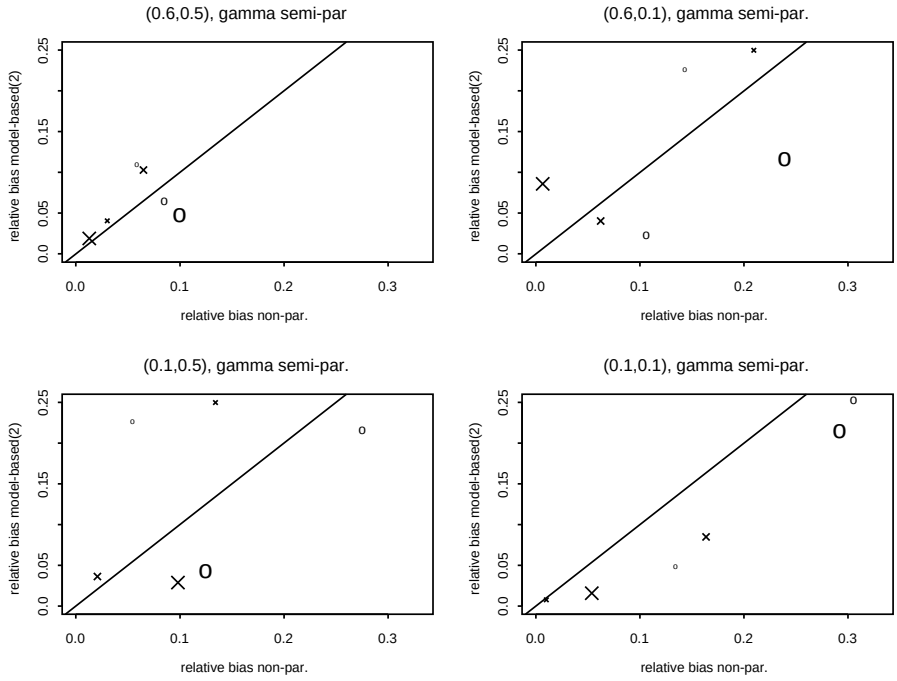

Figure 4: Absolute relative bias for the estimated standard error of the heterogeneity parameter (lognormal frailties); $\circ = (15, 20)$, $\bigcirc = (15, 40)$, $\mathrm{x} = (30, 20)$, $\mathrm{X} = (30, 40)$.

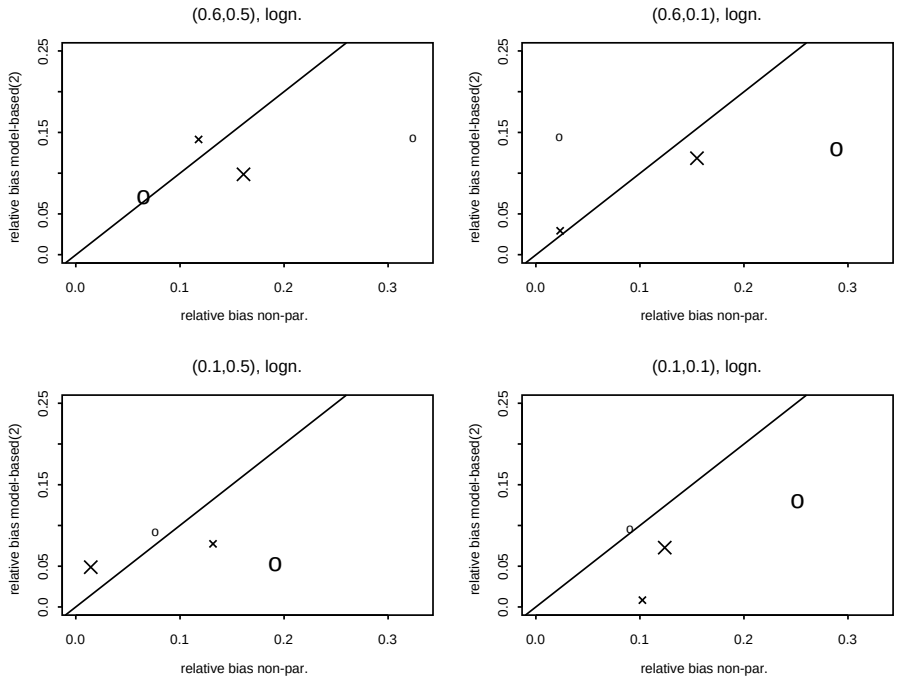

Figure 5: Effect of the number of clusters and patients on the mean estimated standard error, semi-parametric gamma model; ∘ = empirical, △ = nonparametric, □ = model-based (2).

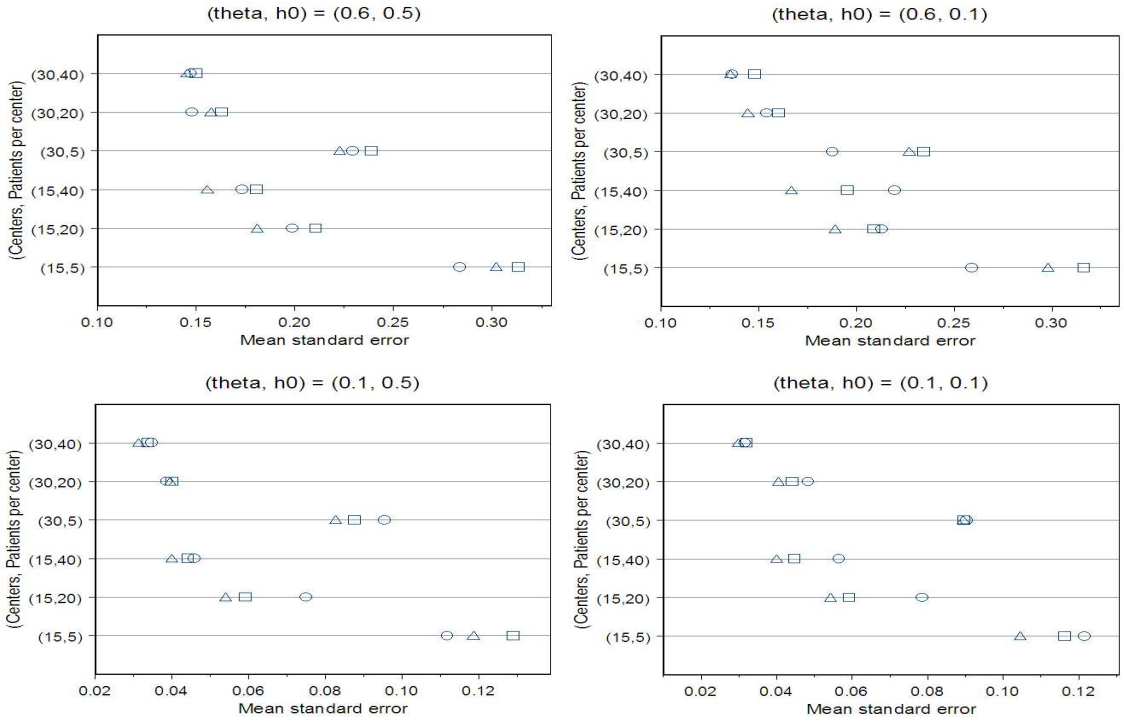

Figure 6: Effect of heterogeneity and event rate on the mean estimated standard error, semi-parametric gamma model; ∘ = empirical, △ = nonparametric, □ = model-based (2).

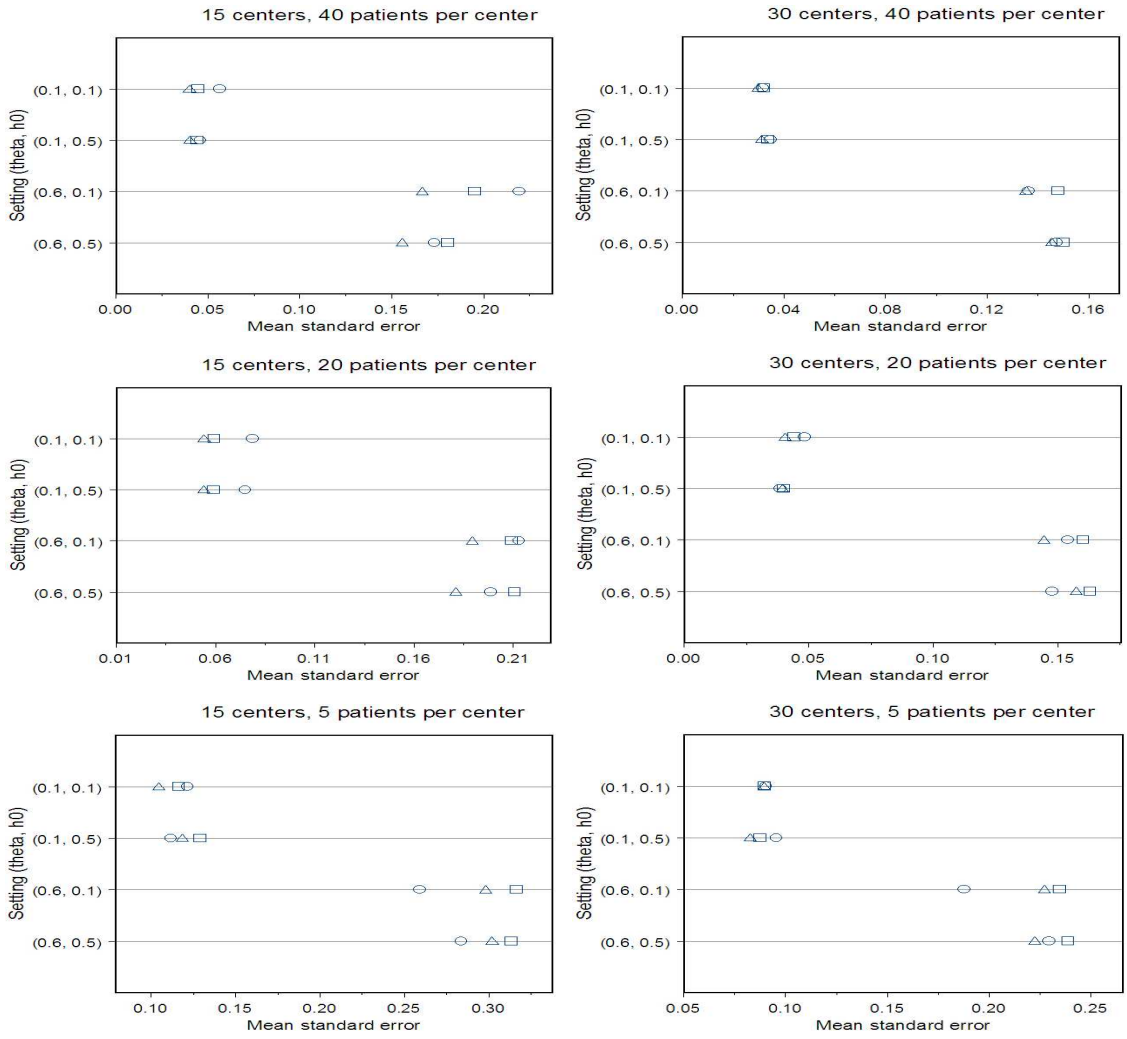

Table 1: Estimated standard error for heterogeneity parameter estimate; for each setting: 40 patients per center, first line for 15 centers, second line for 30 centers.

True Setting (θ, h_0)	Bootstrap assumption	mean $(\hat{\theta})$	SE^E	non-par. mean(SE^B)	model-based(2) mean(SE^B)
(0.6, 0.5)	Gam., par.	0.4129	0.1606		0.1460
		0.3997	0.1128		0.1020
Logn.	Gam., s.-par.	0.4038	0.1562	0.1332	0.1479
		0.3943	0.1116	0.0971	0.1048
	Logn., s.-par.	0.5947	0.3111	0.2903	0.3314
		0.5563	0.2247	0.1885	0.2025
(0.6, 0.5)	Gam., par.	0.4994	0.1722		0.1760
		0.5850	0.1441		0.1436
Gamma	Gam., s.-par.	0.4930	0.1732	0.1556	0.1805
		0.5846	0.1472	0.1453	0.1500
	Logn., s.-par.	0.9120	0.5794	0.5729	0.5940
		1.1672	0.5536	0.6653	0.5112
(0.6, 0.1)	Gam., par.	0.4214	0.1904		0.1508
		0.3988	0.1195		0.1016
Logn.	Gam., s.-par.	0.4144	0.1913	0.1270	0.1542
		0.3934	0.1205	0.0991	0.1049
	Logn., s.-par.	0.6287	0.4116	0.2918	0.4629
		0.5567	0.2313	0.1955	0.2039
(0.6, 0.1)	Gam., par.	0.5370	0.2138		0.1903
		0.5804	0.1373		0.1424
Gam.	Gam., s.-par.	0.5370	0.2195	0.1666	0.1953
		0.5749	0.1362	0.1353	0.1479
	Logn., s.-par.	1.1026	0.8512	0.7907	0.7719
		1.0862	0.4510	0.5260	0.4541
(0.1, 0.5)	Gam., par.	0.0840	0.0358		0.0362
		0.0914	0.0237		0.0260
Logn.	Gam., s.-par.	0.0782	0.0434	0.0341	0.0404
		0.0876	0.0304	0.0296	0.0314
	Logn., s.-par.	0.0883	0.0511	0.0412	0.0487
		0.0957	0.0348	0.0343	0.0365
(0.1, 0.5)	Gam., par.	0.0934	0.0375		0.0392
		0.0987	0.0274		0.0277
Gamma	Gam., s.-par.	0.0903	0.0458	0.0400	0.0441
		0.0969	0.0347	0.0313	0.0337
	Logn., s.-par.	0.1020	0.0547	0.0480	0.0544
		0.1069	0.0411	0.0374	0.0400
(0.1, 0.1)	Gam., par.	0.0885	0.0404		0.0371
		0.0937	0.0275		0.0266
Logn.	Gam., s.-par.	0.0840	0.0496	0.0369	0.0421
		0.0890	0.0347	0.0303	0.0318
	Logn., s.-par.	0.0941	0.0588	0.0439	0.0515
		0.0984	0.0412	0.0361	0.0382
(0.1, 0.1)	Gam., par.	0.0969	0.0487		0.0410
		0.0940	0.0245		0.0275
Gam.	Gam., s.-par.	0.0924	0.0565	0.0399	0.0447
		0.0910	0.0316	0.0299	0.0321
	Logn., s.-par.	0.1066	0.0704	0.0506	0.0567
		0.1005	0.0371	0.0354	0.0378

Table 2: Estimated standard error for estimate of treatment effect; for each setting: 40 patients per center, first line for 15 centers, second line for 30 centers.

True Setting (θ, h_0)	Bootstrap assumption	mean $(\hat{\beta})$	SE^E	non-par. mean(SE^B)	model-based(2) mean(SE^B)
(0.6, 0.5)	Gam., par.	0.2547	0.1143		0.0992
		0.2523	0.0651		0.0700
Logn.	Gam., s.-par.	0.2530	0.1159	0.0997	0.1047
		0.2518	0.0649	0.0688	0.0746
	Logn., s.-par.	0.2530	0.1165	0.0996	0.1064
		0.2521	0.0649	0.0688	0.0741
(0.6, 0.5)	Gam., par.	0.2538	0.0922		0.0985
		0.2464	0.0633		0.0703
Gamma	Gam., s.-par.	0.2530	0.0915	0.0931	0.1062
		0.2469	0.0640	0.0681	0.0754
	Logn., s.-par.	0.2529	0.0915	0.0930	0.1071
		0.2469	0.0639	0.0680	0.0777
(0.6, 0.1)	Gam., par.	0.2474	0.0938		0.0999
		0.2443	0.0609		0.0703
Logn.	Gam., s.-par.	0.2471	0.0944	0.0950	0.1050
		0.2441	0.0604	0.0690	0.0726
	Logn., s.-par.	0.2469	0.0943	0.1051	0.1052
		0.2443	0.0601	0.0689	0.0742
(0.6, 0.1)	Gam., par.	0.2523	0.1008		0.1007
		0.2490	0.0758		0.0709
Gam.	Gam., s.-par.	0.2537	0.1021	0.0975	0.1070
		0.2481	0.0758	0.0670	0.0761
	Logn., s.-par.	0.2537	0.1019	0.0973	0.1096
		0.2483	0.0757	0.0670	0.0764
(0.1, 0.5)	Gam., par.	0.2719	0.1112		0.0999
		0.2626	0.0609		0.0703
Logn.	Gam., s.-par.	0.2712	0.1124	0.0946	0.1013
		0.2627	0.0609	0.0688	0.0715
	Logn., s.-par.	0.2714	0.1124	0.0940	0.1002
		0.2627	0.0608	0.0688	0.0714
(0.1, 0.5)	Gam., par.	0.2370	0.1008		0.0983
		0.2501	0.0721		0.0698
Gamma	Gam., s.-par.	0.2372	0.1015	0.0947	0.1014
		0.2492	0.0713	0.0700	0.0701
	Logn., s.-par.	0.2373	0.1013	0.0943	0.1004
		0.2492	0.0714	0.0700	0.0712
(0.1, 0.1)	Gam., par.	0.2411	0.1006		0.0985
		0.2440	0.0780		0.0703
Logn.	Gam., s.-par.	0.2413	0.1007	0.0934	0.1013
		0.2431	0.0774	0.0682	0.0709
	Logn., s.-par.	0.2415	0.1007	0.0933	0.1001
		0.2432	0.0773	0.0682	0.0719
(0.1, 0.1)	Gam., par.	0.2541	0.1038		0.0991
		0.2512	0.0710		0.0699
Gam.	Gam., s.-par.	0.2533	0.1056	0.0921	0.1002
		0.2513	0.0709	0.0680	0.0701
	Logn., s.-par.	0.2531	0.1056	0.0916	0.1009
		0.2513	0.0709	0.0680	0.0719

Table 3: Coverage proportion of percentile and bias-corrected accelerated intervals for θ for a nominal coverage of 95%; $(\theta, h_0) = (0.6, 0.5)$, gamma distributed frailties; 40 patients per center, first line for 15 centers, second line for 30 centers.

Bootstrap confidence interval	Bootstrap assumption	non-par	model-based(2)
Percentile	Gam., par.		0.79
			0.92
	Gam., s.-par.	0.71	0.76
		0.86	0.90
BCa	Gam., par.		0.90
			0.93
	Gam., s.-par.	0.79	0.91
		0.86	0.94