Supplementary material to ‘Exploring the clustering effect of the frailty survival model by means of the Brier score.’

Robin Van Oirbeek¹, Emmanuel Lesaffre^{1,2}

¹ Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KU Leuven,
Belgium

² Department of Biostatistics, Erasmus Medical Center Rotterdam, The Netherlands

robin.vanoirbeek@gmail.com

Key Words: Brier score, Clustering, Frailty model, Predictive ability, Survival analysis.

Abstract

An adaptation of the Brier score to frailty models is proposed, resulting in a conditional Brier score $BS_C(t)$ and a marginal Brier score $BS_M(t)$. In this document, the interrelation of $BS_C(t)$ and $BS_M(t)$ is shown as well as the procedure of calculating a credible/confidence interval based on the BC_a non-parametric bootstrap technique. Further, a simulation study and its results are described in which the properties of the Bayesian/likelihood point estimates and the percentile and BC_a credible/confidence intervals of $BS_C(t)$ and $BS_M(t)$ for the parametric and the semi-parametric survival model are investigated. Finally, a simulation study is shown which investigates the effect of misspecifying the frailty distribution and the censoring model on the estimation properties of both measures for the parametric survival model. Brier score; Clustering; Frailty model; Predictive ability; Survival analysis.

A: Relationship of $BS_C(t)$ and $BS_M(t)$

$BS_M(t)$ can be decomposed as:

$$\begin{aligned}
BS_M(t) &= E_{\mathbf{X},T,W}\{Y(t) - S_C(t|W, \mathbf{X}) + S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X})\}^2 \\
&= BS_C(t) + 2E_{\mathbf{X},T,W}\{(Y(t) - S_C(t|W, \mathbf{X}))(S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X}))\} \\
&\quad + E_{\mathbf{X},T,W}\{S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X})\}^2 \\
&= BS_C(t) + 2E_{\mathbf{X},W}\{(S_C(t|w, \mathbf{x}) - S_M(t|\mathbf{x}))E_T\{(Y(t) - S_C(t|W, \mathbf{X}))|\mathbf{X} = \mathbf{x}, W = w\}\} \\
&\quad + E_{\mathbf{X},T,W}\{S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X})\}^2 \\
&= BS_C(t) + E_{\mathbf{X},T,W}\{S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X})\}^2,
\end{aligned}$$

such that $BS_M(t) \geq BS_C(t)$ since $E_{\mathbf{X},T,W}\{S_C(t|W, \mathbf{X}) - S_M(t|\mathbf{X})\}^2 \geq 0$. Note that this implies that $BS_M(t) = BS_C(t)$ when there is no intra-cluster correlation.

B: Estimation procedures with credible/confidence interval

In the main article, we propose to compute all measures using a Bayesian and a likelihood procedure, combined with the percentile non-parametric bootstrap method to construct a credible interval or a confidence interval respectively. Here we show how credible/confidence intervals can be constructed by means of the BC_a non-parametric bootstrap method (Efron and Tibshirani, 1993). Note that both bootstrapping techniques have been adapted to the clustered data setting, i.e. resampling by cluster and always selecting the same number of clusters in each bootstrap sample (Field and Welsh, 2007). In the remainder of this section, the measure of interest is denoted as $M(t)$.

1 Procedure using the BC_a method

The BC_a method is a more refined version of the percentile method for which the lower and upper bound of the 95% credible/confidence interval of the measure of interest $M(t)$ correspond to the $\alpha_l(t)$ and $\alpha_u(t)$ percentile of the bootstrap point estimates $\widehat{M}^{(k)}(t)$ with $k = 1, \dots, K$. $\alpha_l(t)$ and $\alpha_u(t)$ are estimated as:

$$\hat{\alpha}_l(t) = \Phi \left(\hat{z}_0(t) + \frac{\hat{z}_0(t) + z^{(.025)}}{1 - \hat{a}(t)(\hat{z}_0(t) + z^{(.025)})} \right),$$

and,

$$\hat{\alpha}_u(t) = \Phi \left(\hat{z}_0(t) + \frac{\hat{z}_0(t) + z^{(.975)}}{1 - \hat{a}(t)(\hat{z}_0(t) + z^{(.975)})} \right),$$

with $\Phi(\cdot)$ the standard normal cumulative distribution function, $z_0(t)$ the bias-correction, $a(t)$ the acceleration and $z^{(.025)}$ ($z^{(.975)}$) the 2.5 (97.5) percentile of the standard normal distribution. The bias-correction $z_0(t)$ is estimated as:

$$\hat{z}_0(t) = \Phi \left(\frac{\sum_{k=1}^K I(\widehat{M}(t) > \widehat{M}^{(k)}(t))}{K} \right),$$

and the acceleration $a(t)$ is estimated as:

$$\hat{a}(t) = \frac{\sum_{i=1}^n (\widehat{M}^{(\cdot)}(t) - \widehat{M}^{(-i)}(t))^3}{6(\sum_{i=1}^n (\widehat{M}^{(\cdot)}(t) - \widehat{M}^{(-i)}(t))^2)^{3/2}},$$

with $\widehat{M}^{(-i)}(t)$ the jackknife value of $\widehat{M}(t)$, i.e. $\widehat{M}(t)$ with the i th subject removed from the data set, and $\widehat{M}^{(\cdot)}(t) = \sum_{i=1}^n \widehat{M}^{(-i)}(t)/n$. Note that the calculation of the acceleration $a(t)$ can be computationally intensive. For the Bayesian and the likelihood method, the procedure corresponds to:

1.1 Bayesian procedure

a), b), c), d) of the Bayesian procedure using the percentile bootstrap method

e) Calculate $\widehat{M}^{(-i)}(t)$ for each subject i and estimate $a(t)$, $z_0(t)$, $\alpha_l(t)$ and $\alpha_u(t)$.

- f) Choose the $\alpha_l(t)$ percentile and the $\alpha_u(t)$ percentile of the bootstrap point estimates $\widehat{M}^{(k)}(t)$ ($k = 1, \dots, K$) as the lower bound and upper bound of the 95% credible interval of $M(t)$.

1.2 Likelihood procedure

- a), b), c), d) of the likelihood procedure using the percentile bootstrap method.
- e) Calculate $\widehat{M}^{(-i)}(t)$ for each subject i and estimate $a(t)$, $z_0(t)$, $\alpha_l(t)$ and $\alpha_u(t)$.
- f) Choose the $\alpha_l(t)$ percentile and $\alpha_u(t)$ percentile of the bootstrap point estimates $\widehat{M}^{(k)}(t)$ ($k = 1, \dots, K$) as the lower bound and upper bound of the 95% confidence interval of $M(t)$.

C: Simulation study

By means of a simulation study, the properties of the Bayesian and likelihood estimates of $BS_C(t)$ and $BS_M(t)$, calculated by the procedures described in Section 3.3 of the paper, are investigated. After describing the design and the fitting procedures of the simulation study, the properties of the Bayesian/likelihood point estimates and the percentile and BC_a credible/confidence intervals of $BS_C(t)$ and $BS_M(t)$ for the parametric and the semi-parametric survival model (i.e. a survival model in which the baseline hazard is estimated non-parametrically) are presented. Additional simulations are performed for the parametric survival model only. These additional investigations entail the effect of changing the intra-cluster correlation, of changing the integration interval and of increasing the sample size on the Bayesian/likelihood point estimates of $BS_C(t)$ and $BS_M(t)$.

1 Design of the simulation study

1.1 Definition of the simulation scenarios

A gamma frailty proportional hazard (PH) model with a Weibull baseline hazard (shape $\rho = 2$, scale $\lambda = 2.236$) was chosen as the failure time distribution. Two covariates were considered, each covariate was sampled from a standard univariate normal distribution assigning different covariate values to each subject of the same cluster. The β parameters were taken as $(-2.45, 2.4)$, representing a population with strong covariate effects. All clusters were equal in size with cluster sizes 2, 5, 10 and 20. The variance parameter of the gamma frailty distribution equals 1 representing a population with a relatively large unobserved heterogeneity. This results in 4 different populations and for each of these populations censoring times were generated from a uniform distribution $U(0, \tau)$ with τ chosen to obtain a censoring percentage of 0%, 25%, 50% or 80%. This leads to 16 scenarios for which the performance of IBS_C and IBS_M (integrated over the interval $[0,1]$) is evaluated. For each of the scenarios, 100 (500) data sets of size 400 were necessary for the estimation of the bias and MSE (coverage probability) to obtain stable results. All population parameters were chosen such that the covariates have an appreciable effect on the failure time distribution.

In order to evaluate the properties of IBS_C and IBS_M , the true values IBS_C^T and IBS_M^T are computed empirically in the absence of censoring by sampling a very large data set ($N = 20,000$) from each of the 4 populations and by calculating the IBS_C and IBS_M using the true values of the parameters of the failure time distribution. The fitted failure time model of the parametric model corresponds to a PH model with a Weibull baseline hazards. The fitted failure time model of the semi-parametric model corresponds to a classical Cox PH model with a partial likelihood estimation procedure (Cox, 1972) for the likelihood procedure and to a gamma independent increments model for the baseline hazards and a PH specification for the covariates (Kalbfleisch, 1978) for the Bayesian procedure. Since no covariates influence the censoring process, the Kaplan-Meier model was chosen as the censoring model for the likelihood procedure (Kaplan and Meier, 1958) and the gamma-

independent increments model for the Bayesian procedure (Kalbfleisch, 1978). The empirical bias was calculated as the averaged difference between the true and the estimated Brier scores and the MSE as the mean of the squared differences between the true and estimated Brier scores. A positive (negative) empirical bias was defined as any estimate greater (smaller) than the true value of IBS_C^T and IBS_M^T , respectively.

For the additional simulations of the parametric survival model, the same simulation design as described above is used, except for the following modifications of the population model of the survival model.

- the degree of the intra-cluster correlation which is changed by varying the variance parameter of the frailty distribution to 0.001 and 0.3 representing a population with a weak and moderate intra-cluster correlation, respectively;
- the size of data sets which is increased to 800;
- the integration interval which is now taken as $[0, 2]$ and $[0, 4]$.

The above modifications were introduced to the original design of the simulation study one at a time. Note that the coverage probabilities for the BC_a non-parametric bootstrapping method were estimated for the unmodified design only.

1.2 Fitting procedures

On each data set, a Bayesian and a likelihood approach was applied to estimate the model parameters using a parametric and a semi-parametric survival model. All the fitting procedures are explained in detail now.

Bayesian procedure: The gamma frailty PH model with a Weibull baseline hazard was chosen as the parametric model, while the gamma independent increments model was selected as the semi-parametric model. For both models, the covariate effects β were given a diffuse normal prior (mean = 0, sd = 1,000) and the variance parameter of the one-parameter gamma frailty distribution ζ was given the half- t distribution prior distribution (Gelman, 2006). For

the parametric model, a diffuse gamma distribution (shape = 1, scale = 1,000) was selected for the shape parameter ρ of the Weibull baseline hazard and the half- t distribution prior distribution for the scale parameter λ of the Weibull baseline hazard (Gelman, 2006). For the semi-parametric model, the time grid of the independent increment process consisted of every 5th percentile of the observed failure time distribution. Furthermore, the Breslow estimator of the cumulative baseline hazard of the likelihood gamma frailty PH model with a non-parametric baseline hazard was chosen as the prior guess for the cumulative baseline hazard process. The scale of the gamma independent increments process represents the confidence in the prior guess of the cumulative baseline hazard process. Here we fixed it to 0.1 to indicate a weak prior belief. Model fitting was done in JAGS. By means of the R package 'R2jags', the results were transferred from JAGS to R to calculate the Brier scores in R. Posterior samples of the model parameters were obtained from samplers of the base and bugs JAGS modules. Preliminary investigations showed that convergence was attained quickly, i.e. within the first 1000 iterations, such that a posterior sample of size 5000 was generated for each data set, discarding the first 1000 samples as burn-in.

Likelihood procedure: The gamma frailty PH model with a Weibull baseline hazard and a non-parametric baseline hazard were chosen as the parametric and semi-parametric model. They were fitted using a marginal likelihood and a penalized partial likelihood approach respectively (Duchateau and Janssen, 2008). This provides estimates for the covariate effects β , the shape ρ and scale parameter λ of the Weibull baseline hazard and the variance parameter ζ of the frailty distribution. The non-parametric baseline hazard was estimated by the Breslow estimator. The frailty terms w were estimated using an Empirical Bayes approach (Therneau and Grambsch, 2000). All calculations were performed in R.

1.3 Simulation results for the parametric model

Bayesian procedure: For the parametric model, no substantial empirical bias is seen over all the different settings for both measures. The empirical MSE of IBS_C and IBS_M increases as the cluster size increases and slightly as the censoring percentage increases (Figure 1).

For both measures, the estimated 95% coverage probability approaches the nominal level for all scenarios and the width of the credible interval increases as the cluster size and/or the censoring percentage increases (Table 1). For both measures, the estimated 95% coverage probability of the BC_a bootstrapping method approaches the nominal level in all scenarios and the width of the credible interval increases as the cluster size and/or the censoring percentage increases (Table 2).

Likelihood procedure: For the parametric model, IBS_C shows a negative empirical bias which increases as the cluster size decreases and the censoring percentage increases. IBS_M shows no substantial empirical bias over the different settings. The empirical MSE of both Brier scores increases as the cluster size increases, and slightly as the censoring percentage increases (Figure 2). The estimated coverage probability approaches the nominal value for all scenarios of which the point estimates are well estimated. The width of the confidence interval increases as the cluster size increases and/or the censoring percentage increases (Table 1). For both measures, the estimated 95% coverage probability of the BC_a bootstrapping method approaches the nominal level in all scenarios and the width of the credible interval increases as the cluster size and/or the censoring percentage increases (Table 2).

1.4 Simulation results for the semi-parametric model

Bayesian procedure: For IBS_C and for a cluster size of 2, a positive empirical bias is seen for the censoring percentages 0%, 25% and 50%. Further, the empirical bias of all cluster sizes decreases for a censoring percentage of 80% as compared to the other censoring percentages, resulting in a negative empirical bias especially for cluster size 5, 10 and 20. IBS_M shows very little empirical bias, except for a negative empirical bias for a censoring percentage of 80%. The empirical MSE of all Brier scores increases mildly as the cluster size and/or the censoring percentage increases (Figure 3).

Likelihood procedure: IBS_C shows a negative empirical bias which increases as the cluster size decreases and the censoring percentage increases. IBS_M shows no substantial empirical bias over the different settings. The empirical MSE of both Brier scores increases as the

cluster size increases, and is virtually unaffected by censoring (Figure 4).

1.5 Additional simulation results for the parametric model

Bayesian procedure: No substantial empirical bias is seen when the intra-cluster correlation, the sample size or the integration interval are varied. The empirical MSE of both measures is virtually unaffected by the intra-cluster correlation and the length of the integration interval, but decreases as the sample size increases.

Likelihood procedure: The empirical bias of IBS_C increases slightly as the intra-cluster correlation increases, but increases considerably as the integration interval becomes longer. For IBS_M however, the empirical bias remains unaffected by changing the intra-cluster correlation or the integration interval. For both measures, the sample size has little effect on the empirical bias. The empirical MSE of both measures is virtually unaffected by the intra-cluster correlation and decreases as the sample size increases. For IBS_C , the empirical MSE increases as the integration interval increases while it is unchanged for IBS_M .

2 Summary of the simulation results

From these investigations we learn that IBS_C and IBS_M are well estimated by the Bayesian method for the parametric survival model. For the likelihood method, IBS_M is well estimated under all investigated scenarios, while IBS_C suffers from a negative empirical bias as the censoring percentage increases and/or the cluster size decreases. The estimation of IBS_C by the Bayesian method and the estimation of IBS_M by the Bayesian and likelihood method are not influenced by the length of the integration interval, nor by the size of the intra-cluster correlation and the sample size. The quality of the credible/confidence intervals of the percentile and the BC_a bootstrap method are found to be very similar. Note that the estimate of the acceleration $\hat{a}(t)$ of the BC_a non-parametric bootstrap method is found to be very small in each scenario, such that we recommend to fix the acceleration to zero.

Slightly different results were found for the semi-parametric model. For the Bayesian

method, the estimation of IBS_C and ISB_M show a negative empirical bias when the censoring percentage is high. Moreover, IBS_C is also not well estimated when the cluster size is small, since under these conditions the frailty variance term is grossly underestimated. The semi-parametric model of the likelihood method shows a similar behavior as the parametric model of the likelihood method.

Note that the MSE estimates of the likelihood and the Bayesian procedure were found to be very similar throughout the simulation study, given that bias patterns of the two types of procedures did not differ too strongly. This is due the variance part of the MSE estimate which in general dominated the bias part and which was found to be very similar for both procedures.

D: The effect of misspecification on the estimation of $BS_C(t)$ and $BS_M(t)$

In this section, the effect of misspecifying the frailty distribution and the effect of misspecifying the censoring model on the estimation properties of $BS_C(t)$ and $BS_M(t)$ are investigated for the parametric survival model.

1 Misspecification of the frailty distribution

1.1 Design of the simulation study

The same design as the basic design of Section C of the supplementary material is used for this simulation study, except for the frailty distribution of the population survival model. Following the notation in Duchateau and Janssen (2008), a lognormal distribution (with $\gamma = 0.5$), a 50-50 mixture of a gamma (with $\theta = 1$) and a positive stable distribution (with $\alpha = 0.75$), and a positive stable distribution (with $\alpha = 0.75$) were considered as frailty distributions. The performance of IBS_C and IBS_M were investigated for the intervals $[0, 0.5]$

and $[0, 1]$.

1.2 Simulation results for the Bayesian procedure

For the lognormal model and the mixture model, a small positive empirical bias is found for IBS_C , especially for the scenario with a cluster size of 2, whereas a slight to moderate negative empirical bias is found for IBS_M , especially for the integration interval $[0, 1]$ (Figure 5 - Figure 8). For IBS_C of the positive stable model, no substantial empirical bias is seen for the scenarios with a cluster size of 10 and 20. The scenarios with a cluster size of 5 and 2 however show a slight to large positive empirical bias, which slightly decreases as the censoring percentage increases. IBS_M of the positive stable model shows a large negative empirical bias for all scenarios, which increases as the length of the integration interval increases (Figure 9 - Figure 10). Note that this negative empirical bias is not strongly affected by the censoring percentage. For the lognormal and the mixture (positive stable) models, the empirical MSE of IBS_C increases (decreases) as the cluster size increases, indicating that the empirical MSE for the positive stable models are dominated by their corresponding empirical bias patterns. For IBS_M , the empirical MSE increases as the cluster size increases.

1.3 Simulation results for the likelihood procedure

The empirical bias and empirical MSE patterns of the likelihood procedure are very similar to the corresponding patterns of the Bayesian procedure, except for an additional negative empirical bias for IBS_C . This negative empirical bias becomes more pronounced as the censoring percentage increases and it affects the empirical bias and empirical MSE patterns of the integration interval $[0, 1]$ more strongly than the empirical bias patterns of the integration interval $[0, 0.5]$. Because of this, the empirical MSE of IBS_C of the lognormal model and the mixture model evaluated over the integration interval $[0, 1]$ now even decreases as the cluster size increases (Figure 11 - Figure 16).

1.4 Interpretation of the simulation results

Given that the cluster size is sufficiently large (larger than 10 in our simulation study), $BS_C(t)$ is virtually unaffected by misspecification of the frailty distribution. $BS_M(t)$, however, is in general affected by misspecification, especially when the frailty distribution is strongly misspecified. The results of the Bayesian procedure were found to be very similar to the result of the likelihood procedure except for an additional negative bias for $BS_C(t)$ which increases as the cluster size and/or censoring percentage and/or degree of misspecification increases.

An in-depth investigation of the results revealed that the scale parameter λ of the Weibull baseline hazard is poorly estimated for all investigated scenarios. Moreover, the empirical distribution of the estimated frailty terms $\hat{\mathbf{w}}$ showed less variability than the empirical distribution of the true frailty terms $\mathbf{w}$, but the empirical distribution of the product of $\hat{\lambda}$ and $\hat{\mathbf{w}}$ was very similar to empirical distribution of the product of λ and $\mathbf{w}$, especially when no empirical bias was detected for IBS_C for that scenario. Since both $\mathbf{w}$ and λ operate as a product in $S_C(t|\mathbf{x}_{ji}, w_j) = \exp(-\lambda w_j \exp(\mathbf{x}_{ji}\boldsymbol{\beta})t^\rho)$, λ and therefore the baseline hazard seems to ‘compensate’ for the misspecification of the frailty distribution. The conditional survival curves therefore remain of good quality. The effect of the misspecification of λ and the variance parameter of the frailty distribution $\boldsymbol{\zeta}$ resulted in marginal survival curves that descend faster than their homologues of the correct model, resulting in overoptimistic IBS_M estimates. Of course, the direction of the bias of IBS_M depends entirely on how the misspecified model parameters affect the marginal survival curves, and for another population and model choice this bias could either be negative or positive.

We therefore recommend to use $BS_C(t)$ for model selection and $BS_M(t)$ to check how the clustering effect improves upon the predictive ability of the covariates for the model at hand. Again, this improvement in predictive ability does not necessarily correspond to the one of the correct model. In case of a minor misspecification and a not so high censoring percentage (0%-50%) however, the marginal Brier score was seen to be correctly estimated. This suggests that the predictive improvement due to clustering will be evaluated correctly

when the marginal Brier score is evaluated for a carefully selected model.

2 Misspecification of the censoring model

2.1 Design of the simulation study

In this section we use a slightly different simulation design as before, namely we base our simulation design on the one of Gerds and Schumacher (2006). A gamma frailty PH model with a exponential baseline hazard was chosen as the population model of the failure time distribution. One covariate X (with a covariate effect of -1) was considered, sampled from a standard univariate normal distribution and assigning different covariate values to each subject of the same cluster. All clusters are of size 20 and the variance of the gamma frailty distribution equals 1. Censoring times were generated from an exponential model using the same covariate $X \equiv Z$ (with a covariate effect of 1) as in the failure time model, meaning that the covariates of the failure time model and the censoring model are not independent. The rate parameter of the censoring model was fixed in such a way that a censoring percentage of 0%, 25%, 50% or 80% was obtained. This leads to 4 scenarios to evaluate the performance of IBS_C and IBS_M (integrated over the interval $[0, 1]$) and for each of the scenarios, 100 data sets of size 400 were sampled.

Three censoring models were tested on this population: a null model, an exponential model and a lognormal model. The latter two models both contain the correct covariate Z and for both models the effect of adding a covariate that has no influence on the censoring model (sampled from a standard normal model and attributing a different value for each subject) and the effect of functionally misspecifying the covariate effect (introducing the covariate as Z^2 instead of Z) was investigated. The Kaplan-Meier model was chosen as the null model model for the likelihood procedure (Kaplan and Meier, 1958) and the gamma-independent increments model for the Bayesian procedure (Kalbfleisch, 1978).

One extra population was considered in which the covariate of the censoring model Z was sampled from a standard normal distribution, hereby differing from the covariate of

the failure time model X . In this case the covariate of the failure time model is different and independent from the covariate of the censoring model. For this population only the null censoring model was tested, which corresponds to a missing at random scenario, also considered by Gerds and Schumacher (2006).

2.2 Simulation results for the Bayesian procedure

For both measures and for a censoring percentage of 50% and 80% only, a slight to moderate positive empirical bias was observed for the null model, the lognormal model, the lognormal model with an additional covariate and the exponential and lognormal model with a functional misspecification of the covariate effect. All the other scenarios show no significant empirical bias. For each scenario, the empirical MSE increases as the censoring percentages increases, attaining in general a much higher empirical MSE value for a censoring percentage of 80% (Figure 17 - Figure 19). The results of the missing at random scenario (Figure 20) is similar to the ones of the null model (Figure 17), the positive empirical bias of the former scenario being smaller at a censoring percentage of 80% than for the latter scenario.

2.3 Simulation results for the likelihood procedure

The empirical bias and empirical MSE patterns of the likelihood procedure are very similar to the ones of the Bayesian procedure, except for an additional negative empirical bias for IBS_C . This negative empirical bias becomes more pronounced as the censoring percentage increases. The results of the missing at random scenario is similar to the ones of the null model, the positive empirical bias of the former scenario being smaller at a censoring percentage of 80% than for the latter scenario.

2.4 Interpretation of the simulation results

A functionally misspecified covariate effect in the censoring model and/or a completely misspecified censoring model will result in a positive bias, especially for high censoring percent-

ages (50% and 80% in our simulation study). It is, however, easier to detect misspecifications of the censoring model as the censoring percentage increases. Covariates that have no impact on the censoring process will not influence the quality of the estimation when included in the censoring model. Omitting important variables, however, does result in a positive bias. This positive bias was found to be more pronounced for X equal to Z , than when they are statistically independent. As a result, we recommend to use a saturated censoring model and to select the censoring model that results in the lowest estimates for $BS_C(t)$ and $BS_M(t)$. This agrees with the results of Gerds and Schumacher (2006). In practice, several equally performing censoring models can be proposed. This is however not a problem since the censoring model is merely used as a working model. Good estimates for the explained variance $\widehat{R}(t)$ (based on either $\widehat{BS}_C(t)$ or $\widehat{BS}_M(t)$) are obtained even when the censoring model is misspecified. Indeed, when the Brier score is used for comparative purposes only, it is only necessary that the same censoring model is used for all survival models (Gerds and Schumacher, 2006).

References

- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 187–220.
- Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. New York: Springer Science + Business Media.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction To The Bootstrap*, Chapman & Hall/CRC.
- Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B*, **69**, 369–390.
- Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score

- in general survival models with right-censored event times. *Biometrical Journal*, **48**, 1029–1040.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.
- Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.
- Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data*. New York: Springer Verlag.

Figure 1: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method. Each plot represents a set of scenarios sharing the same cluster size ('2', '5', '10' or '20'). The X-axis refers to the censoring percentage ('0', '25', '50' and '80') and the Y-axis to the empirical bias (upper panels) or to the empirical MSE (lower panels).

Figure 2: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method. For further information, see the legend of Figure 1

Figure 3: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method for the semi-parametric model. For further information, see the legend of Figure 1.

Figure 4: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method for the semi-parametric model. For further information, see the legend of Figure 1.

Figure 5: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 0.5]$ for the lognormal frailty distribution. For further information, see the legend of Figure 1.

Figure 6: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 1]$ for the lognormal frailty distribution. For further information, see the legend of Figure 1.

Figure 7: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 0.5]$ for the mixture frailty distribution. For further information, see the legend of Figure 1.

Figure 8: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 1]$ for the mixture frailty distribution. For further information, see the legend of Figure 1.

Figure 9: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 0.5]$ for the positive stable frailty distribution. For further information, see the legend of Figure 1.

Figure 10: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method over the interval $[0, 1]$ for the positive stable frailty distribution. For further information, see the legend of Figure 1.

Figure 11: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 0.5]$ for the lognormal frailty distribution. For further information, see the legend of Figure 1.

Figure 12: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 1]$ for the lognormal frailty distribution. For further information, see the legend of Figure 1.

Figure 13: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 0.5]$ for the mixture frailty distribution. For further information, see the legend of Figure 1.

Figure 14: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 1]$ for the mixture frailty distribution. For further information, see the legend of Figure 1.

Figure 15: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 0.5]$ for the positive stable frailty distribution. For further information, see the legend of Figure 1.

Figure 16: Simulation study misspecification frailty distribution: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the likelihood method over the interval $[0, 1]$ for the positive stable frailty distribution. For further information, see the legend of Figure 1.

Figure 17: Simulation study misspecification censoring model: empirical bias (left panel) or empirical MSE (right panel) in estimating IBS_C and IBS_M by the Bayesian method for the null censoring model using the same covariate in the failure time and the censoring population definition

. For further information, see the legend of Figure 1.

Figure 18: Simulation study misspecification censoring model: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method for the exponential censoring model, either with an extra covariate that has no influence on the censoring process ('Extra') or with a functional misspecified covariate effect ('Quadr.'). For further information, see the legend of Figure 1.

Figure 19: Simulation study misspecification censoring model: empirical bias (upper panels) or empirical MSE (lower panels) in estimating IBS_C and IBS_M by the Bayesian method for the lognormal censoring model ('LN'), either with an extra covariate that has no influence on the censoring process ('Extra') or with a functional misspecified covariate effect ('Quadr.'). For further information, see the legend of Figure 1.

Figure 20: Simulation study misspecification censoring model: empirical bias (left panel) or empirical MSE (right panel) in estimating IBS_C and IBS_M by the Bayesian method for the null censoring model using a different covariate in the failure time and the censoring population definition. This corresponds to the missing at random scenario. For further information, see the legend of Figure 1.

Table 1: Simulation study: coverage probability (left) and width (right) of the credible/confidence intervals of IBS_M and IBS_C using the percentile method estimated by the Bayesian and likelihood method, investigated for varying degrees of censoring ('0%', '25%', '50%' and '80%') and cluster size ('2', '5', '10' or '20').

	IBS_M				IBS_C				IBS_M				IBS_C			
	0%	25%	50%	80%	0%	25%	50%	80%	0%	25%	50%	80%	0%	25%	50%	80%
<i>Bayesian</i>																
2	97.6	97.4	97.4	98.0	98.2	97.4	98.6	97.0	0.033	0.033	0.033	0.035	0.029	0.029	0.032	0.036
5	96.2	94.8	94.8	96.4	97.4	96.8	93.6	96.0	0.036	0.035	0.037	0.040	0.031	0.032	0.032	0.036
10	94.6	95.2	95.8	94.8	95.8	96.0	94.6	94.2	0.041	0.040	0.041	0.044	0.035	0.034	0.036	0.039
20	96.2	93.8	96.6	94.6	94.4	94.2	94.8	95.0	0.046	0.046	0.046	0.050	0.041	0.041	0.040	0.044
<i>Likelihood</i>																
2	97.0	97.2	95.6	98.8	92.8	92.4	92.8	91.0	0.032	0.033	0.032	0.035	0.027	0.027	0.027	0.039
5	95.6	96.4	95.8	96.6	94.4	95.8	96.0	94.0	0.035	0.035	0.035	0.039	0.029	0.029	0.031	0.034
10	95.2	95.0	94.2	97.0	94.8	95.2	93.4	93.4	0.038	0.040	0.041	0.041	0.033	0.035	0.035	0.037
20	97.0	96.2	94.8	93.8	94.6	95.0	94.2	94.8	0.046	0.045	0.048	0.049	0.040	0.040	0.041	0.043

Table 2: Simulation study: coverage probability (left) and width (right) of the credible/confidence intervals of IBS_M and IBS_C using the BC_a method estimated by the Bayesian and likelihood method, investigated for varying degrees of censoring ('0%', '25%', '50%' and '80%') and cluster size ('2', '5', '10' or '20').

		IBS_M				IBS_C				IBS_M				IBS_C			
		0%	25%	50%	80%	0%	25%	50%	80%	0%	25%	50%	80%	0%	25%	50%	80%
<i>Bayesian</i>																	
38	2	95.2	96.6	97.4	98.0	95.8	97.2	98.6	94.2	0.033	0.032	0.034	0.035	0.029	0.029	0.032	0.035
	5	96.6	97.6	94.4	94.8	97.4	96.8	93.8	94.8	0.036	0.035	0.036	0.039	0.031	0.032	0.033	0.037
	10	94.8	94.6	95.8	93.6	96.8	97.8	93.2	94.0	0.042	0.040	0.041	0.043	0.036	0.035	0.035	0.039
	20	93.8	93.0	96.4	92.8	93.4	92.6	94.6	96.4	0.048	0.047	0.046	0.050	0.042	0.040	0.040	0.045
<i>Likelihood</i>																	
	2	94.2	96.8	95.6	98.0	93.2	93.2	92.4	92.4	0.032	0.032	0.032	0.035	0.027	0.027	0.028	0.039
	5	96.0	96.2	95.2	96.4	92.6	96.6	94.8	94.4	0.035	0.035	0.035	0.038	0.029	0.029	0.030	0.034
	10	94.6	96.6	93.6	95.2	94.0	95.0	96.2	93.4	0.038	0.040	0.042	0.041	0.033	0.034	0.035	0.037
	20	96.8	93.4	96.8	96.8	93.4	95.0	92.8	95.6	0.045	0.045	0.047	0.050	0.040	0.040	0.040	0.043