Exploring the Clustering Effect of the Frailty Survival Model by Means of the Brier Score

ROBIN VAN OIRBEEK¹ AND EMMANUEL LESAFFRE¹,²

¹Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KU Leuven, Leuven, Belgium
²Department of Biostatistics, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands

In this article, the Brier score is used to investigate the importance of clustering for the frailty survival model. For this purpose, two versions of the Brier score are constructed, i.e., a “conditional Brier score” and a “marginal Brier score.” Both versions of the Brier score show how the clustering effects and the covariate effects affect the predictive ability of the frailty model separately. Using a Bayesian and a likelihood approach, point estimates and 95% credible/confidence intervals are computed. The estimation properties of both procedures are evaluated in an extensive simulation study for both versions of the Brier score. Further, a validation strategy is developed to calculate an internally validated point estimate and credible/confidence interval. The ensemble of the developments is applied to a dental dataset.

Keywords Brier score; Clustering; Frailty model; Predictive ability; Survival analysis

Mathematics Subject Classification 62N99

1. Introduction

For frailty models, the importance of clustering typically is assessed by means of the variance of the frailty distribution only: when the variance is large, clustering is important and vice versa. In this article, we would like to investigate the importance of clustering in more detail. To this end, we propose to assess how clustering affects the quality of the model predictions.

In the course of the last three decades, many different measures of the predictive ability have been proposed for univariate survival data (O’Quigley et al., 2005; Schemper and Stare, 1996). These measures either measure discrimination, calibration, or a combination thereof, also called the overall performance (Steyerberg et al., 2010). Until now only the concordance probability, which is a discrimination measure, has been extended to the clustered data setting (Van Oirbeek and Lesaffre, 2010). In practice, many researchers have argued that a measure of the overall performance should be evaluated on top of a discrimination measure (Steyerberg et al., 2010). We therefore extended the Brier score in this article, which is a measure of the overall prediction performance of a regression model,
Exploring Clustering Effect of Frailty Model 3295
to the clustered setting (Graf et al., 1999; Steyerberg et al., 2010). This extension results in
two different versions of the Brier score, namely a conditional Brier score, which captures
the predictive ability of the covariate and clustering effects jointly, and a marginal Brier
score, which captures the predictive ability of the covariate effects only. By comparing
both versions of the Brier score, one can evaluate explicitly the additional predictive gain
of the clustering effects besides that of the covariate effects. This information cannot
be obtained from a marginal model since it does not gauge the intra-cluster correlation
explicitly. In order to check whether this predictive gain is important, we developed a 95%
credible/confidence interval procedure as well as an internal validation procedure for both
versions of the Brier score.

We first review the Brier score and its estimator for univariate survival models
(Section 2). In Section 3, two versions of the Brier score are developed for the frailty
model using a Bayesian and a likelihood approach. Their properties have been investigated
in an extensive simulation study, detailed in the supplementary materials (available online)
and summarized in Section 4. An internal validation scheme is presented in Section 5. We
apply all our developments to a dental dataset in Section 6, followed by a discussion.

2. Brier Score for Univariate Survival Data

Consider a randomly selected subject $i$ with a true failure time $T_i$, a stochastic covariate
vector $X_i$, and the predicted survival probability $S(t|X_i)$ at a given time point $t > 0$. The
time-dependent squared residual $BS_i(t)$ for the $i$th subject at time point $t$ is defined as:
$$BS_i(t) = (Y_i(t) - S(t|X_i))^2,$$
where $Y_i(t) = I\{T_i > t\}$ corresponds to the event status of subject $i$ at time point $t$. The
Brier Score $BS(t)$ is the population version of this measure and corresponds to the average
squared distance between the event status and the predicted survival curve at $t$, or
$$BS(t) = E_{X,T}\{Y(t) - S(t|X))^2,$$
where the expectation is taken over the failure time distribution and the covariate distri-
bution. This measure ranges between 0 and 1 whereby a value close to zero indicates a
well-predicting survival model. $BS(t)$ can be used to construct a time-dependent measure
of the explained variance $R^2(t)$ by scaling the Brier score of the model of interest to the
Brier score under a noninformative model $BS_0(t)$, i.e., $R^2(t) = 1 - \frac{BS(t)}{BS_0(t)}$. If we assume
that the model of interest has a predictive ability that is superior or equal to one of the null
model, then $R^2(t)$ ranges by definition between 0 and 1. In the likelihood framework the
Kaplan-Meier model (Kaplan and Meier, 1958), and in the Bayesian framework the gamma
independent increments model (Kalbfleisch, 1978) can be used as a noninformative model.
An integrated version of the Brier score ($IBS$) can be constructed on a time interval $[\tau_1, \tau_2]$ as
$$IBS = \int_{\tau_1}^{\tau_2} BS(t)dw(t),$$
with $\tau_1$ mostly equal to 0 and the weight function $W(t)$ mostly
equal to $t/(\tau_2 - \tau_1)$. $IBS$ therefore expresses the average behavior of the Brier score over
the time interval $[\tau_1, \tau_2]$. Other weight functions $W(t)$ can be proposed, for instance when
one wishes to focus on the predictive ability of the model in certain parts of the considered
time interval (Graf et al., 1999).

For univariate data, a uniformly consistent estimator has been provided for the Brier
score, when censoring and failure times are conditionally independent given the covariates
(Gerds and Schumacher, 2006). Let $t_i$ be the observed failure time of subject $i$ such that
\( t_i = \min(T_i, T_{c,i}) \) corresponds to the observed failure time, with \( T_{c,i} \) the right censoring time and \( T_i \) the true failure time. If \( t_i = T_i \), then \( t_i \) is a true failure time and the censoring indicator \( \delta_i \) equals 1. If \( t_i = T_{c,i} \), then \( t_i \) is a censoring time and the censoring indicator \( \delta_i \) equals 0. For a sample of size \( n \), Gerds and Schumacher (2006) proposed to estimate the Brier score as:

\[
\hat{BS}(t) = \frac{1}{n} \sum_{i=1}^{n} \{ y_i(t) - \hat{S}(t|x_i) \} \phi(t, \hat{G}(t|z_i)),
\]  

where \( y_i(t) \) is the observed event status, \( x_i \) (\( z_i \)) the observed covariate vector of the survival (censoring) model, \( \phi(t, \hat{G}(t|z_i)) \) the time-dependent censoring weight, and \( \hat{S}(t|x_i) \) (\( \hat{G}(t|z_i) \)) the estimator of the conditional survival function of the survival (censoring) model of subject \( i \). Since the estimated model parameters are directly used in the estimation of the Brier score, they are part of both the selected failure time model and the estimated Brier score. The censoring model is fitted to the data by interchanging the role of event time and censoring time. The censoring model (and its covariates \( Z \)) and the survival model (and its covariates \( X \)) are usually not the same. For a subject \( i \), the estimated censoring weight \( \phi_i(t) \) equals:

\[
\phi_i(t) \equiv \phi(t, \hat{G}(t|z_i)) = \frac{I\{t < t_i\}}{\hat{G}(t|z_i)} + \frac{I\{t \geq t_i\} \delta_i}{\hat{G}(t_i - |z_i})
\]  

with \( t_i \) – the time point just before \( t_i \). \( \hat{BS}(t) \) is a consistent estimator of the Brier score under conditional independence of the censoring mechanism and the failure time model and if the censoring model is correctly specified. Hence, because \( \phi(\cdot) \) only depends on the censoring model, estimator (1) is robust against misspecification of the failure time model (Gerds and Schumacher, 2006). Since the event status \( y_i(t) \) is unknown for a censored subject \( i \) at a time point \( t \geq t_i \), Graf et al. (1999) defined that \( \phi(\cdot) \) equals 0 for this subject when \( t \geq t_i \). Consistency of this estimator was proven by Gerds and Schumacher (2006).

Based on an extensive simulation study, the latter authors also stated that inverse probability of censoring weighted (IPCW) estimates based on the Kaplan-Meier model (Kaplan and Meier, 1958) can be used, in case that covariates have no or little impact on the censoring process.

3. Brier Score for Clustered Survival Data

3.1. The Two Proposals

In the presence of clustering, a frailty model can be used to model the survival times. This model accounts for clustering by introducing a frailty term \( w_j \) to each separate cluster \( j \) of the dataset. The frailty term \( w_j \) is a positive random variable with a frailty distribution \( f_{W}(w|\zeta) \), depending on parameters \( \zeta \). A parametric distribution such as the gamma distribution is typically chosen.

Two model formulations can be proposed for a frailty model: a conditional frailty model that explicitly contains the frailty terms in its formulation and a marginal frailty model that is constructed by integrating out the frailty terms from the conditional frailty model (Duchateau and Janssen, 2008, chap. 2). Hence, a conditional survival curve \( S_C(t|X, W) \) and a marginal survival curve \( S_M(t|X) \equiv \int S_C(t|X, w) f_{W}(w|\zeta) dw \) can be constructed. To exemplify the difference between the two survival functions, take
for example the proportional hazards (PH) frailty model. For this model, it holds that
\( S_C(\tau | X, W) = \Lambda_0(\tau) \exp(-W \exp(\beta^T X)) \) with \( \Lambda_0(\tau) \) the cumulative baseline hazard at time point \( \tau \) and \( \beta \) a vector of regression coefficients. For PH frailty models of the power variance family, \( S_M(\tau | X) = \exp(\nu/(\theta(1 - v))(1 - (\theta \Lambda_0(\tau) \exp(\beta^T X))/\nu)^{1-v}) \) with \( \mu, \theta > 0, 0 < v \leq 1 \) clearly \( S_M(\tau | X) \) depends on the choice of the frailty distribution, while \( S_C(\tau | X, W) \) does not.

Two versions of \( BS(t) \) can therefore be defined for frailty survival models. The conditional Brier score \( BS_C(t) \) compares for each subject at time point \( t \) the binary event status with the conditional survival curve, or:
\[
BS_C(t) = E_{X,T,W}\{Y(t) - S_C(t|X, W)\}^2 = E_W\{E_{X,T}\{Y(t) - S_C(t|X, w)\}^2|W = w\}
= E_W\{BS_{C,w}(t)\},
\]
with \( BS_{C,w}(t) \) the conditional Brier score of the cluster characterized by \( w \). \( BS_C(t) \) can be seen as a weighted average cluster conditional Brier score. The marginal Brier score \( BS_M(t) \) compares at time point \( t \) the binary event status with the marginal survival curve, or:
\[
BS_M(t) = E_{X,T}\{Y(t) - S_M(t|X)\}^2.
\]
In summary, two different versions of the Brier score can be defined for the frailty model: \( BS_C(t) \) evaluates the predictive ability of the chosen frailty model while accounting for the covariate effects and the clustering effect, while \( BS_M(t) \) accounts for the covariate effects of the chosen model only. As such, the comparison of \( BS_M(t) \) with \( BS_C(t) \) expresses the added (predictive) value of considering the clustering effects explicitly on top of the covariate effects for the chosen model. As proven in Section A of the supplementary material (available online), \( BS_C(t) \) is always lower than or equal to \( BS_M(t) \). This means that, on average, the conditional survival curve follows the event status at least as closely as the marginal survival curve.

3.2. Estimating \( BS_C(t) \) and \( BS_M(t) \) for Frailty Models

\( BS_C(t) \) and \( BS_M(t) \) can be estimated by plugging in the estimates of the conditional \( S_C(t|X, w) \) and the marginal survival curve \( S_M(t|X) \) respectively in (1). Consider a sample of size \( n \) with \( J \) clusters of size \( n_1, n_2, \ldots, n_J \) and \( \sum_{j=1}^J n_j = n \). Both Brier scores can then be estimated as:
\[
\widehat{BS}_C(t) = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji}(t) - \hat{S}_C(t|x_{ji}, \hat{w}_{ji}))^2 \phi(t, \hat{G}(t|z_{ji})) = \sum_{j=1}^J \frac{n_j}{n} \widehat{BS}_{C,j}(t),
\]
and,
\[
\widehat{BS}_M(t) = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji}(t) - \hat{S}_M(t|x_{ji}))^2 \phi(t, \hat{G}(t|z_{ji})),
\]
with \( x_{ji}, z_{ji}, \) and \( y_{ji} \) the covariate vector of the survival model, the covariate vector of the censoring model, and the event status for subject \( i \) of cluster \( j \), respectively, and with \( \widehat{BS}_{C,j}(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ji}(t) - \hat{S}_C(t|x_{ji}, \hat{w}_{ji}))^2 \phi(t, \hat{G}(t|z_{ji})) \) the estimated conditional Brier score of cluster \( j \) at time point \( t \). Clearly, \( \widehat{BS}_C(t) \) is a weighted average of all cluster conditional Brier scores of the dataset. The censoring weight \( \phi_i(t) \) of a given subject \( i \)
is estimated as in (2). When an analytical expression for the marginal survival curve is lacking for the frailty model, \( \hat{SM}(t|x_{ji}) \) can be estimated as \( 1/K \sum_{k=1}^{K} \hat{SC}(t|x_{ji}, \hat{w}_k) \) with \( \hat{w}_k \) sampled from \( f_W(w|\hat{\xi}) \) and \( K \) large (Molenberghs and Verbeke, 2005, chaps. 7 and 11). Note that \( \hat{BS}_C(t) \) and \( \hat{BS}_M(t) \) are consistently estimated for the considered survival model only if the censoring model is correctly specified (Gerds and Schumacher, 2006).

3.3. Point and Interval Estimates of the Brier Scores

Both Brier scores can be computed using a Bayesian or a likelihood procedure, combined with the percentile nonparametric bootstrap method to construct a credible interval or a confidence interval, respectively. In these procedures, the estimation of the failure time model as well as the estimation of the Brier score, which includes the estimation of the censoring model, are both entirely performed using either Bayesian approaches or likelihood approaches. The bootstrap technique has 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). The Bayesian procedure also necessitates this bootstrap technique to ensure that the resulting 95% credible intervals also capture the uncertainty of the estimation of the model parameters while using future values of the observed covariates. The description of both procedures using the \( BC_a \) method instead of the percentile method can be found in Section B of the supplementary material (available online). In the remainder of this section, the measure of interest is denoted as \( M(t) \). For the Bayesian procedure, this amounts to:

(a) After removing the burn-in part of the Markov chain, let the Markov chain Monte Carlo (MCMC) procedure run for an extra \( m \) iterations. At iteration \( p \) with \( p = 1, \ldots, m \) of the MCMC procedure, determine the posterior estimate for each model parameter. For the PH frailty model, posterior estimates for the parameters of the baseline hazard function \( \Lambda_0(t) \), the covariate effects \( \beta \), and the frailty terms \( w \) are obtained in this manner (Duchateau and Janssen, 2008, chap. 2). This step needs to be applied to the failure time model and the censoring model separately.

(b) For each subject \( i \), obtain the posterior sample of the censoring weight \( \phi_i(t) \) based on the posterior sample of the censoring model parameters.

(c) Compute \( M^{(p)}(t) \) at each iteration \( p \) with \( p = 1, \ldots, m \) based on the posterior sample of the survival model parameters and the censoring weights \( \phi(t) \) at iteration \( p \). The median of the posterior values of \( M(t) \) could be used as the point estimate \( \hat{M}(t) \).

(d) Apply the Bayesian bootstrap technique (Rubin, 1981) to sample \( k = 1, \ldots, K \) new bootstrap datasets. Calculate the point estimate of each \( k \)th bootstrap dataset \( \hat{M}^{(k)}(t) \) based on the posterior samples \( M^{(p,k)}(t) \) with \( p = 1, \ldots, m \).

(e) Compute the 95% credible interval (CI) by means of the percentile method (Efron and Tibshirani, 1993, chap. 13) using the point estimates of the \( K \) bootstrap datasets \( \hat{M}^{(k)}(t) \). As such, the 2.5 and the 97.5 percentile of the bootstrap point estimates \( \hat{M}^{(k)}(t) \) constitute the lower and the upper bounds of the 95% credible interval of \( M(t) \).

The likelihood procedure corresponds to:

(a) Determine the point estimates of the model parameters, which are obtained for the PH frailty model by maximizing the marginal likelihood for the parameters of the baseline hazard function \( \Lambda_0(t) \) and the covariate effects \( \beta \) after which empirical
Bayes estimates are calculated for the frailty terms $\mathbf{w}$ (Duchateau and Janssen, 2008, chap. 2). This step needs to be applied to the failure time model and the censoring model separately.

(b) For each subject $i$, obtain a point estimate of the censoring weight $\phi_i(t)$ based on the estimates of the censoring model parameters.

(c) Compute the point estimate $\hat{M}(t)$ based on the estimates of the survival model parameters and point estimates of the censoring weights $\phi(t)$.

(d) Apply Efron’s bootstrap technique (Efron and Tibshirani, 1993, chaps. 6 and 13) to sample $k = 1, \ldots, K$ new bootstrap datasets. Calculate the point estimate $\hat{M}^{(k)}(t)$ of each $k$th bootstrap dataset.

(e) Compute the 95% confidence interval (CI) by means of the percentile method (Efron and Tibshirani, 1993, chap. 13) using the point estimates of the $K$ bootstrap datasets $\hat{M}^{(k)}(t)$. As such, the 2.5 and the 97.5 percentile of the bootstrap point estimates $\hat{M}^{(k)}(t)$ constitute the lower and the upper bounds of the 95% credible interval of $M(t)$.

4. Simulation Study

The estimation properties of the likelihood and Bayesian estimation procedures (as described in Section 3.3) are checked for the parametric and semi-parametric survival model in an extensive simulation study, in which factors such as the censoring percentage and the cluster size are varied. Apart from an overoptimistic empirical bias of $IBS_C$ and $IBS_M$ when the censoring percentage is very high (80%) and an underoptimistic empirical bias of $IBS_C$ when the cluster size is very low (2), the semi-parametric likelihood approach shows a similar behavior as the parametric likelihood approach for the Bayesian method. For the likelihood method, the estimation properties of $IBS_C$ and $IBS_M$ using either the parametric or semi-parametric survival model are similar under all investigated scenarios.

In general, $BS_M(t)$ is well estimated by both the Bayesian and likelihood methods. $BS_C(t)$, however, suffers from an overoptimistic bias for the likelihood method, especially when the cluster size is small and when the censoring percentage is high, while good estimation properties are observed for the Bayesian method. This overoptimistic bias is caused by the empirical Bayes estimates of the frailty terms: in a small simulation study, we saw that the overoptimistic bias of $\hat{BS}_C(t)$ disappeared once the frailty terms were replaced by their population values. Similar results were found elsewhere (Van Oirbeek and Lesaffre, 2010, 2012). The 95% credible/confidence intervals based on the percentile or $BC_a$ method possess good coverage properties.

Sensitivity against misspecification of the frailty distribution and the censoring model were investigated as well for the parametric model. The estimation properties of the likelihood procedure are very similar to the corresponding properties of the Bayesian procedure, except for an additional overoptimistic empirical bias for $IBS_C$.

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 strongly affected by misspecification. We observed that the conditional survival curves remain of good quality and that the baseline hazard “compensates” the misspecification of the frailty distribution. As a consequence, the quality of the marginal survival curve may be poor. 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. Note that this improvement in predictive ability does not necessarily correspond to one of the correct model.
A functionally misspecified covariate effect or a misspecified censoring model will result in an overoptimistic bias, especially for high censoring percentages (50% and 80% in our simulation study). Note, however, that it is 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 an overoptimistic bias. As a result, we recommend to use a saturated censoring model and to select the censoring model that results in the lowest estimates for $BSC(t)$ and $BSM(t)$. Note that this agrees with the results of Gerds and Schumacher (2006). Good estimates for the explained variance $\hat{R}(t)$ (based on either $\hat{BSC}(t)$ or $\hat{BSM}(t)$) 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). For more details on the setup of the simulation study and its results, we refer to Sections C and D of the supplementary material (available online).

5. Internal Validation

Since external data are often lacking, the same data are usually used for both model building and measuring the predictive ability. As a result, overoptimistic estimates of the predictive ability are often acquired. One can correct for this by applying an adaptation to the clustered setting of the .632 and .632+ bootstrap cross-validation procedures of the Brier score for univariate survival models (Gerds and Schumacher, 2007). Let $M(t)$ be the measure that one wants to validate and $\hat{M}(t)$ the overoptimistic point estimate of the measure. The procedure then corresponds to:

(a) Split the sample randomly in a training and a validation set. Make sure that the training (validation) set contains around 63.2% (36.8%) of the subjects of the original sample.

(b) Repeat the same modeling steps in the training sample as in the original dataset. Obtain estimates for the model parameters,

(c) Calculate a point estimate of the measure $\hat{M}_{\text{boot}}(t)$ on subjects of the validation set only.

(d) Repeat the first three steps of this procedure $B$ times with $B$ reasonably large.

Despite that step (a) is different from the procedure proposed by Gerds and Schumacher (2007), the properties of the upper procedure are the same as the properties of their procedure since for both procedures the training set is supported on 63.2% of the original data points (Efron and Tibshirani, 1997). The internally validated estimate corresponds to $(1 - \psi(t))\hat{M}(t) + \psi(t)\frac{\sum_{b=1}^{B} \hat{M}_{\text{boot},b}(t)}{B}$, with $\psi(t) = 0.632$ for the .632 bootstrap cross-validation estimate $\hat{M}_{362}(t)$ and $\psi(t) = \frac{0.632}{1 - 0.632 R_{\text{ror}}(t)}$ for the .632+ bootstrap cross-validation estimate $\hat{M}_{362+}(t)$ (Gerds and Schumacher, 2007). In the latter expression, $R_{\text{ror}}(t)$ is the relative overfitting rate that corresponds to:

$$
\hat{R}_{\text{ror}}(t) = \frac{\frac{\sum_{b=1}^{B} \hat{M}_{\text{boot},b}(t)}{B} - \hat{M}(t)}{\hat{M}_{\text{NoInf}}(t) - \hat{M}(t)},
$$

where $\hat{M}_{\text{NoInf}}(t)$ corresponds to the no-information error rate or the Brier score for which the event status is independent from the covariates $x$. As such, for each subject $i$ of cluster
j, its contribution to $\hat{M}_{\text{NoInf}}(t)$ is computed by averaging over all observed $x$ values of the dataset:

$$\hat{M}_{\text{NoInf}}(t) = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left[ \sum_{k=1}^{J} \sum_{l=1}^{n_k} \{ y_{ji}(t) - \hat{S}(t|x_{kl}) \}^2 \phi(t, \hat{G}(t|z_{ji})) \right].$$

When $\hat{R}_{\text{ror}}(t)$ is lower than 0 (higher than 1), $\hat{R}_{\text{ror}}(t)$ is fixed to 0 (1). The .632+ bootstrap cross-validation estimate is found to perform better than the .632 bootstrap cross-validation estimate (Efron and Tibshirani, 1997).

In order to compute $\hat{BS}_C(t)$, an estimate of the random effect of each of the clusters of the dataset is needed. However, in the above procedure it might happen that some clusters of the original dataset are not represented in the training set of step (a) such that in the validation set of step (c) the cluster-specific estimate of $BS_C(t)$ cannot be calculated for these clusters. In that case we propose to calculate $\hat{BS}_C(t)$ in the validation set of step (c) only for those clusters that were present in the training set of step (a). It is assumed that $BS_C(t)$ cannot be calculated for a minority of the clusters of the validation dataset only such that the 63.2/36.8 proportion of the training and validation set is not too strongly disrupted. A too strong disruption of this proportion invalidates the above internal validation scheme.

A credible/confidence interval can be constructed around the internally validated point estimate by combining the above internal validation procedure and the bootstrap procedures of Section 3. More specifically, we start by sampling $K$ bootstrap datasets. For each of these bootstrap datasets, the above procedure is repeated resulting in $K$ validated measures $\hat{M}_{.632}^{(k)}(t)$ and $\hat{M}_{.632+}^{(k)}(t)$ with $k = 1, \ldots, K$. These validated measures $\hat{M}_{.632}^{(k)}(t)$ and $\hat{M}_{.632+}^{(k)}(t)$ are then used to construct a 95% credible/confidence interval by applying step (e) of the percentile method (see Section 3).

6. Application to the Amalgam Dataset

6.1. Introduction

Up until two decades ago, amalgam restorations were commonly used as a material for dental fillings. To this end, a clinical study was set up to investigate the effect of different factors on the longevity of amalgam restorations. The 12 covariates and 15 derived variables of the study are the four cavity wall treatments “CSA,” “CWF,” “Copalite,” and “Silver Suspension,” the alloy of the amalgam (levels: “NTD,” “Tytin,” and “CNG”), the caries status, the type of the restoration (levels: “MO/DO” and “MOD”), the type of tooth (levels: “premolar” and “molar”), the position of the tooth in the mouth (levels: “upper left,” “upper right,” “lower left,” and “lower right”), the operator (levels: “operator 1,” “operator 2,” and “operator 3”), gender, and age. The amalgam dataset is composed of three clinical trials and in each clinical trial the four cavity wall treatment effects and the three investigated amalgam alloys were combined into four treatment modalities assigning each modality randomly to four (or 8 or 12) teeth in a 2 x 2 (factorial) design within patient (Kreulen et al., 1998). Note that this study was mainly designed to evaluate significance of the four cavity wall treatment effects and the three investigated amalgam alloys. In this article, we will evaluate the predictive potential of all the measured covariates.

Since the early failures were suspected to be of a different nature than the late failures and since there were only two early failures, we restrict our analysis to the late failures only, i.e., failures that occurred at least nine years after enrollment in the study. This results in
174 patients contributing with 1,347 amalgam restorations. The clustered data structure is unbalanced, with 4, 8, and 12 restorations seen in 36, 89, and 33 patients, respectively. Note that 16 patients contributed to the study with a number of teeth differently from 4, 8, or 12. The median follow-up time is 14.48 and the true failure time is observed for 187 amalgam restorations only, leading to a censoring percentage of 86.2%. In Kreulen et al. (1998), it is mentioned that, in addition to these 187 failures (called true failures in their paper), 20 false failures were observed. A competing risk approach might therefore be needed here. However, we limited our analysis to the true failures only, since the 20 false failures were not identified in the dataset, but more importantly the current analysis merely serves to illustrate our proposed measures.

Based on the simulation results described in Section C of the supplementary material (available online), we have chosen to use the Bayesian gamma frailty generalized gamma model as the survival model. Different censoring models for the IPCW weights were tested with the Bayesian generalized gamma model consisting of all 12 covariates resulting in the lowest Brier score estimates. For the computation of $\hat{IBS}_C$, $\hat{IBS}_M$, and $\hat{IBS}_0$, $\hat{BS}_C(t)$, and $\hat{BS}_M(t)$ were integrated from 9 to 15 years, since, as recommended by Graf et al. (1999), censoring is not too heavy at this time point. The .632+ bootstrap cross-validation estimators were chosen for the internal validation of $BS_C(t)$ and $BS_M(t)$. The R code that is used to perform the below analysis can be found in the supplementary material (available online). For reasons of confidentiality, we could not provide the amalgam dataset such that a fictive dataset is supplied with this code.

**6.2. The Analysis**

6.2.1. **Model selection.** The objective of this analysis was to find the best predicting survival model. A model selection step was therefore performed to find the covariate combination that results in this best predicting model. To this end, the all-possible-regression procedure was applied, meaning that the predictive ability of the corresponding survival model is evaluated for each of the $2^{12} - 1$ different subsets of the 12 covariates. Based on the simulation results of Section D of the online supplementary material ($IBS_C$ is robust against misspecification of the frailty distribution, but $IBS_M$ not), we have chosen $IBS_C$ as selection criterion. To avoid the effect of overfitting, the $IBS_C$ values were internally validated. We chose the all-possible-regression procedure, though other methods could be applied as well. From Fig. 1, one can see that the best predicting model is a model with five covariates, containing “CSA,” “CWF,” and “Copalite” cavity wall treatments, caries status, and gender as covariates.

6.2.2. **Predictive ability of the final model.** The predictive ability of the selected model will be investigated in more detail now. The apparent and validated point estimate and 95% percentile credible interval of the Brier scores ($B = 100, K = 100$) are shown in Table 1. We see that $IBS_M$ improves upon the null model and that $IBS_C$ improves upon both the null model and $IBS_M$. Indeed, even after internal validation, $IBS_M$ attains an explained variance of 17.7% and $IBS_C$ an explained variance of 31.2%. This means that the marginal (conditional) survival curves are on average 17.7% (31.2%) closer to the observed event status than the survival curve of the null model. The high predictive ability of the model can thus be attributed to a combination of strong covariate effects and a strong clustering effect. Therefore, if one would succeed in a future study to fully capture the clustering effect in covariates and if also the same covariates were measured as the ones included in this model, the explained variance of $IBS_M$ of this future study can potentially attain 31.2%.
Exploring Clustering Effect of Frailty Model

Figure 1. Amalgam data analysis: results of the all-possible-regression procedure using $IBS_C$ as a criterion. The X-axis refers to the number of variables of the considered model and the Y-axis to the $IBS_C$ value. Jittering was added to visualize the performance of all 4,095 models. The best performing model (at the bottom of the figure) is indicated with an asterisk.

Note that all the apparent and validated 95% percentile credible intervals show a strong overlap and that the validated 95% percentile credible intervals are wider than the apparent 95% percentile credible intervals. Also, the drop in explained variance before and after internal validation is more pronounced for $IBS_C$ than for $IBS_M$. Due to a low mean of the relative overfitting rate for both $IBS_C$ and $IBS_M$ (33.4% and 3.0% respectively), the .632 estimate of both $IBS_C$ and $IBS_M$ only differ mildly from their corresponding .632+ estimate ($\hat{IBS}_{C,.623} = 0.0191$ versus $\hat{IBS}_{C,.623+} = 0.0194$ and $\hat{IBS}_{M,.623} = 0.0231$ versus $\hat{IBS}_{M,.623+} = 0.0232$).

In the left panel of Fig. 2, the apparent and validated time-varying estimates of the Brier scores are shown. Up until 14 years after the start of follow-up, all curves are very similar and barely differ from zero. From 14 years onwards, the apparent and validated $BS_C(t)$ are clearly lower than the apparent and validated $BS_M(t)$ and the apparent $\bar{BS}_0(t)$.

Table 1

|           | $IBS_0$             | $IBS_M$             | $IBS_C$             |
|-----------|---------------------|---------------------|---------------------|
| Apparent  | 0.0282 (22.3%) [0.0144,0.0349] | 0.0168 (40.4%) [0.0072,0.0244] |                      |
| Validated | 0.0232 (17.7%) [0.0152,0.0465] | 0.0194 (31.2%) [0.0113,0.0394] |                      |

Note: The explained variance of $IBS_C$ and $IBS_M$ is added in parentheses.
Figure 2. Amalgam data analysis: the left panel visualizes the point estimates of the apparent time-varying patterns of $BS_0(t)$, $BS_M(t)$, and $BS_C(t)$ and internally validated time-varying patterns of $BS_M(t)$ and $BS_C(t)$. The right panel shows the Kaplan-Meier estimate of the failure time distribution with 95% confidence limits, which are corrected for the presence of clustering (Williams, 1995). Ticks are added to visualize the observed failure times.

and the apparent and validated $BS_M(t)$ are clearly lower than the apparent $\hat{BS}_0(t)$. From the right panel of Fig. 2, it is clear that these time patterns are due to the shape of the observed failure time distribution. Indeed, the majority of the observed failures occur after 14 years after the start of follow-up and since the survival curves of the null model and the considered survival model remain close to 1 up until 14 years after the start of follow-up, the Brier scores result in low values.

7. Discussion

In this article, two different versions of the Brier score have been proposed to investigate the importance of clustering for the frailty model, resulting in a “conditional Brier score” and a “marginal Brier score.” In an extensive simulation study, the likelihood estimation procedure showed difficulties in estimating unbiasedly the conditional Brier score, while the Bayesian estimation procedure resulted in good estimation properties for both versions of the Brier score. We also developed a procedure to calculate the 95% percentile bootstrap credible/confidence interval and a .632+ internal validation procedure.

We have chosen to estimate the Brier score by means of the IPCW method, which has the disadvantage that the censoring time distribution needs to be carefully modeled. Despite of this possible shortcoming, it is up to now, to our knowledge, the best method to estimate the Brier score. Other very useful measures (O’Quigley et al., 2005; Schemper and Stare, 1996) can be extended to the frailty model in a similar way as proposed in this article. We chose the Brier score for its complementarity with discrimination measures, its easy interpretation, and because it can be calculated for every type of survival model (Steyerberg et al., 2010).

An external validation of $BS_C(t)$ may not be straightforward in practice. Indeed, unless the same clusters are used in both external and training dataset, no estimates of the frailty
terms of the clusters of the external dataset are available, such that $BSC(t)$ cannot be calculated for that cluster. This underlines the importance of a good internal validation procedure for this measure. A topic of future research might be to estimate a frailty term of such a new cluster.

In this article we extended the Brier score for frailty models with a two-level clustering structure, since this is the most common clustering structure in survival analysis (Duchateau and Janssen, 2008, chaps. 2 and 6). In practice, however, more complicated clustering structures may be encountered. In Van Oirbeek and Lesaffre (2012) for example, it is shown how the concordance probability and the Brier score can be extended for nested two-level and three-level multilevel binary regression models.

Acknowledgments

The authors thank the referees and the associate editor for their helpful comments and Cees Kreulen for the permission to use the amalgam dataset (Kreulen et al., 1998) for this research.

Funding

For the computations, we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government, Department EWI. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

Conflict of Interest

None declared.

Supplemental Material

Supplementary data for this article can be accessed on the publisher’s website.

References

Duchateau, L., Janssen, P. (2008). *The Frailty Model*. New York: Springer.
Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.
Efron, B., Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* 92:548–560.
Field, C. A., Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B* 69:369–390.
Gerds, T. A., Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* 48:1029–1040.
Gerds, T. A., Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* 63:1283–1287.
Graf, E., Schmoor, C., Sauerbrei, W., Schumacher M., (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18:2529–2545.
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., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53:457–481.
Korn, E. L., Simon, R. (1991). Explained residual variation, explained risk and goodness of fit. The American Statistician 45:201–206.

Kreulen, C. M., Tobi, H., Gruythuysen, R. J. M., van Amerongen, W. E., Borgmeijer, P. J. (1998). Replacement risk of amalgam treatment modalities: 15-year results. Journal of Dentistry 26:627–632.

Molenberghs, G., Verbeke, G. (2005). Models for Discrete Longitudinal Data. New-York: Springer-Verlag.

O’Quigley, J., Xu, R., Stare, J. (2005). Explained randomness in proportional hazards models. Statistics in Medicine 24:479–489.

Rubin, D. B. (1981). The Bayesian bootstrap. The Annals of Statistics 9:130–134.

Schemper, M., Stare, J. (1996). Explained variation in survival analysis. Statistics in Medicine 15:1999–2012.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T. A., Gönen, M., Obuchowski, N., Pencina, M., Kattan, M. W. (2010). Assessing the performance on prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138.

Van Oirbeek, R., Lesaffre, E. (2010). An application of Harrell’s C-index to PH frailty models. Statistics in Medicine 29:3160–3171.

Van Oirbeek, R., Lesaffre, E. (2012). Assessing the predictive ability of a multilevel binary regression model. Computational Statistics & Data Analysis 56:1966–1980.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. Lifetime Data Analysis 1:171–186.