Calibration plots for risk prediction models in the presence of competing risks

Thomas A. Gerds
Per K. Andersen
Michael W. Kattan

Research Report 13/09

Department of Biostatistics
University of Copenhagen
Calibration plots for risk prediction models in the presence of competing risks

Thomas A. Gerds and Per K. Andersen

Department of Biostatistics, University of Copenhagen, Denmark

Michael W. Kattan

Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, USA

July 2, 2013

Abstract. A predicted risk of 17% can be called reliable if it can be expected that the event will occur to about 17 out of 100 patients who all received a predicted risk of 17%. Statistical models can predict the absolute risk of an event such as cardiovascular death in the presence of competing risks such as death due to other causes. For personalized medicine and patient counseling it is necessary to check that the model is calibrated in the sense that it provides reliable predictions for all subjects. There are three often encountered practical problems when the aim is to display or test if a risk prediction model is well calibrated. The first is lack of independent validation data, the second is right censoring, and the third is that when the risk scale is continuous the estimation problem is as difficult as density estimation. To deal with these problems, we propose a new approach to estimate calibration curves for competing risks models based on jackknife pseudo-values which are combined with a nearest neighborhood smoother and a cross-validation approach to deal with all three problems.

Keywords: Calibration plots; Kernel smoothing; Risk models; Pseudo values; Reliability diagram

1 Introduction

In order to be useful for personalized medicine and patient counseling, a statistical model that predicts the absolute risk of an event should be calibrated. A calibration plot displays how well observed and predicted event status connect on the absolute probability scale. In this note we show how calibration plots can be obtained and summarized in survival analysis with competing risks and right censored data.
To define calibration, first pick a time origin at which it is of interest to predict the future status of a patient. Until time \( t \) after the time origin three things can happen:

1. the event has occurred
2. a competing event has occurred
3. the patient is alive and event-free.

A risk prediction model uses patient characteristics and other measurements that are available at the time origin to predict the probability that the event will occur within the next \( t \)-years. The model is said to be calibrated if the expected percentage of the sub-population who receives a predicted risk of \( p \) is \( p \) for all predicted probabilities \( p \). The calibration plot, also known as the reliability diagram, is a method to visually inspect calibration; it shows predicted against expected probabilities.

If some patients are lost to followup before time \( t \), their event time is right censored, and what happened between their lost to followup time and time \( t \) remains unknown. To consistently estimate the expected probabilities, and hence the calibration plot, one has to deal with the right censored data. Thus, cases that are lost to followup within the predicted time point need to be imputed, to account for the possibility that they would have indeed have the event if followed to the predicted time point. Our approach for this imputation is to replace the status indicator at time \( t \) with a jackknife pseudo-value (Andersen et al., 2003). Importantly, in order to get this to work we have to impute the status for all patients not just those censored before time \( t \).

2 Calibration curves based on pseudo-values

For patient \( i \) let \( T_i \) be the time from the time origin until one of \( \kappa \) mutually exclusive events occurs (competing risks) and \( \eta_i \) the type of the event. In all what follows \( \eta_i = 1 \) is the event of interest and the remaining \( \kappa - 1 \) events are competing risks. Observed are the right censored time \( \bar{T}_i = \min(T_i, C_i) \) and the right censored event type \( \bar{\eta}_i = \Delta_i \eta_i \) where \( \Delta_i = 1 \) if \( T_i \leq C_i \) and \( \Delta_i = 0 \) otherwise. Here \( C_i \) denotes the time of censoring. For any future time point \( t \), a risk prediction model \( r \) responds to patient specific covariates \( Z_i \in \mathbb{R}^d \) with a probability of the event: \( r(t|Z_i) \). It may happen that the model predicts the same probability for two or more different patients. However, if at least one of the covariates is continuous then typically the predicted risks for patients with different covariate values are different.

To define the calibration curve formally, we introduce for patient \( i \) the event status indicator variable \( Y_i(t) = 1\{T_i \leq t, \eta_i = 1\} \), i.e., \( Y_i(t) = 1 \) if event 1 occurred to subject \( i \) by time \( t \), and \( Y_i(t) = 0 \), otherwise. For each probability \( p \in [0, 1] \) we also define \( G_p(t;p) = \{z \in \mathbb{R}^d : r(t|z) = p\} \), the set of patients characteristics for which the model predicts that the event occurs with probability \( p \).
p until time t. The calibration curve of the model at time t is defined as the graph of the mapping

\[ p \mapsto C(p, t, r) = E_{Y,Z} \{ Y(t) | Z \in G_r(t; p) \} = E_{Y,Z} \{ Y(t) | r(t|Z) = p \}. \] (1)

To obtain the graph we need to estimate the expectation on the right hand side of (1). Three often encountered practical problems arise:

- Continuity: the size of the sets \( G_r(t; p) \) may be small and it may happen that a set includes only a single patient.
- Right censoring: if patient \( i \) is not followed until time \( t \), the status \( Y_i(t) \) is unknown.
- Generalizability: we would like to know if the model will be reliable for new patients, not just in the data set which was used to specify and estimate the model.

For the remainder of this section we hide the generalizability issue by assuming an independent external validation data set, which includes right censored event times and patient characteristics from \( n \) patients. In the next section we also discuss internal validation where the same data set is used for fitting the model and for estimating the calibration curve.

To overcome the right censoring problem, we first construct a pseudo-value for all patients. For patient \( i \) the pseudo-value is given by:

\[ \tilde{Y}_i(t) = n\hat{F}_n(t) - (n - 1)\hat{F}^{(i)}_n(t). \]

Here, \( \hat{F}_n \) is the Aalen-Johansen estimate of the cumulative incidence function for event 1, \( F(t) = E(Y_i(t)) \) based on all patients and \( \hat{F}^{(i)}_n \) that based on the subset where the data of the \( i \)th patient are removed. See [Andersen and Perme (2010)] for a comprehensive introduction to the pseudo-value approach for censored survival analysis in medical statistics.

Suppose first that the model is discrete in the sense that it assigns all patients to one of \( L \) different risk groups, \( G_r(t; p_1), \ldots, G_r(t; p_L) \). Risk group models occur naturally when all predictor variables are class variables, but can also be obtained by discretizing the continuous predictor variables [Breiman et al., 1984]. This makes most sense when there are different medical or health care actions corresponding to different risk groups. The calibration curve of a risk group model consists only of \( L \) points which can be estimated by averaging the pseudo-values of the patients whose predicted probability is in this group:

\[ \hat{C}(p_l, t, r) = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i(t) I\{ Z_i \in G_r(t; p_l) \}. \] (2)

Now suppose the model is continuous in the sense that it predicts a probability exactly \( p \) with probability zero. One way to overcome this problem is to group the predicted risks into \( K \) different groups. For no good reason \( K = 10 \) is a popular
choice which is often used in connection with the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1989). The disadvantages of using a fixed number of groups or a fixed group size were illustrated in (Le Cessie and Van Houwelingen, 1991) where the authors propose nonparametric smoothing as an alternative.

A simple smoother is a moving average where at each probability \( p \) the calibration of the model is estimated by averaging the pseudo-values of the patients that have received a predicted probability within a pre-specified interval around \( p \). We here consider a nearest neighborhood smoother (Yang, 1981; Dabrowska, 1989; Akritas, 1994) to obtain a smooth estimate of the calibration curve of a continuous model \( r \):

\[
\hat{C}_{an}(t,p,r) = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i(t)K_{an}(p,r(t|Z_i)).
\]  

(3)

Here, \( K_{an}(p,q) = \mathcal{I}\{H_n(p) - H_n(q) \leq a_n\}/a_n \) is a rectangular kernel function and \( H_n(p) = 1/n \sum_{i=1}^{n} \mathcal{I}\{r(t|Z_i) \leq p\} \) denotes the empirical distribution function of the predicted risks. The bandwidth \( a_n \) defines the size of the neighborhoods and is chosen data adaptively. If \( a_n \) converges towards 0 then \( \hat{C}_{an}(t,p,r) \) converges to \( C(t,p,r) \) in uncensored data. In right censored data, if the censoring mechanism is independent of the event time and the covariates, the results of Graw et al. (2009) imply that the estimates in (2) and (3) consistently estimate the calibration curve. If the censoring mechanism depends on the covariates the pseudo-values can be constructed based on a working Cox model for the censoring probabilities to avoid bias (Binder et al., 2013).

As an alternative to the pseudo-value approach (a smoothed version of) the Aalen-Johansen estimate can be applied directly to estimate \( C(p,t,r) \). Similar results are expected, but the advantage of the pseudo-value approach is that the pseudo-values can be shown together with the calibration graph.

3 Internal validation

Internal validation is what one can do in the absence of external validation data. Internal validation works by repeated splitting and cross-validation of the one and only data set. From now on we call the model built in all data the final model. The apparent calibration plot uses the predictions of the final model in its own learning data set. For data intensive modelling strategies the apparent performance typically overestimates the prediction accuracy and the apparent calibration plot is systematically closer to the diagonal than the expected calibration plot of the final model in new patients. Many different but similar versions of cross-validation were proposed and the terminology can be quite confusing. We therefore define here the proposed procedure explicitly.

Suppose we have competing risks data from \( n \) patients. These are randomly split \( B \) times into learning sets \( L_1, \ldots, L_B \) and corresponding validation sets \( V_1, \ldots, V_B \), such that \( L_b \) and \( V_b \) are non-overlapping for all \( b \). The special case where \( B < n \) and also the learning sets are non-overlapping is called \( B \)-fold cross-validation, and leave-one-out cross-validation when \( B = n - 1 \). A variant is to draw the learning
sets from the data either of size \( n \) with replacement (regular bootstrap) or of size \( m < n \) without replacement (subsampling bootstrap). For the regular bootstrap the validation sets include on average 36.8% of the data. The procedure which then averages the validation sample performance is called bootstrap cross-validation or leave-one-out bootstrap (Efron and Tibshirani 1997; Mogensen et al., 2012).

To define the cross-validated calibration curve we let \( R \) denote the strategy which when applied to a data set performs all model specification steps and parameter estimations needed to eventually return a prediction model. Specifically we apply \( R \) separately to each learning data set. This yields a series of prediction models \( R(V_b) = \hat{R}_b, b = 1, \ldots, B \). Predicted event probabilities \( \hat{R}_b(t|X_i) \) are then obtained for all patients \( i \in V_b \). We now describe our proposed approach to combine cross-validation with pseudo-values.

The idea is simply to collect the predictions and pseudo-values from the \( B \) validation sets into a new data set which is then used to calculate the leave-one-out bootstrap calibration curve at \( p \):

\[
\tilde{C}_{\tilde{a},n,B}(t,p,R) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{b,i \in V_b} \tilde{Y}_i(t)K_{\tilde{a}}(p,\hat{R}_b(t|Z_i)).
\]

Here it is practical to choose \( \tilde{a}_n \) as the optimal bandwidth for smoothing the predictions of the final model for the full data set (Sheather and Jones 1991). Note also that we compute the pseudo-values only once based on all data. Alternatively, one could replace \( \tilde{Y}_i \) by pseudo-values \( \tilde{Y}_{i,b} \) which are calculated using the information in the \( b \)th validation set.

A different approach would be to calculate one calibration curve for each split and then somehow average across calibration curves. However, an average of smoothed functions seems not very attractive.

4 Summarizing predictive accuracy

The accuracy of a risk prediction model can be decomposed into calibration and discrimination. The discrimination ability of a competing risks model can be assessed by a time-truncated version of the concordance index (Gerds et al., 2013; Wolbers et al., 2013) or by a version of the time-dependent area under the ROC curve (Saha and Heagerty, 2010).

The calibration of a competing risks model can be visualized by the calibration plots described above. The closer the calibration curve of the model is to the diagonal the better. For the predicted probability at \( Z \) the bias of the model \( \hat{R}_n \) is given by

\[
E_Y(Y(t)|\hat{R}_n(t|Z)) - \hat{R}_n(t|Z)
\]

and a useful summary of the calibration curve is given by the the squared bias:

\[
\text{Bias}^2 = E_Z((E_Y(Y(t)|\hat{R}_n(t|Z)) - \hat{R}_n(t|Z))^2).
\]

Further, an estimate of the squared bias equals the average squared distance between the observed calibration curve and the diagonal.
The population level Brier score for event 1 at time $t$ is defined as the expected squared distance between the observed status at time $t$ and the predicted probability. It can be decomposed into the squared bias term and a variance term:

$$BS(t, \hat{R}_n) = E_{Y,Z} \{ Y(t) - \hat{R}_n(t|Z) \}^2 = \text{Bias}^2 + \text{Var}$$

where $\text{Var} = E_Z \{ Y(t) - E_Y(Y(t)|\hat{R}_n(t|Z)) \}^2$. The Brier score measures discrimination and calibration at the same time. To illustrate this consider first a constant model which predicts a constant event risk of 50% for all subjects. As any constant model this model has zero discrimination ability, i.e. AUC($t$) = c-index = 50%, and the Brier score is 25%. Among the constant models there is one perfectly calibrated model. It predicts to every patient the true population level event prevalence at time $t$ (in practice the prevalence at time $t$ can be estimated by the Aalen-Johansen estimate). It is easy to see that the Brier score of the perfectly calibrated constant model is smaller than that of the 50% risk model (with equality only in the case where the event prevalence equals 50%). This shows that the Brier score can see a difference between two models which have equal discrimination ability. Now consider a model which distinguishes between men and women and predicts the true event prevalence among males for men and the true event prevalence among females for women. This stratified model is also perfectly calibrated but it has positive discrimination ability and hence a lower Brier score than the perfectly calibrated constant model (with equality only in the case where the population prevalence is the same for women and men).

5 Illustration

PBC3 was a multi-centre randomized clinical trial conducted in six European hospitals [Lombard et al. (1993)]. Between 1 Jan. 1983 and 1 Jan. 1987, 349 patients with the liver disease primary biliary cirrhosis (PBC) were randomized to either treatment with Cyclosporin A or placebo. The purpose of the trial was to study the effect of treatment on the survival time. However, during the course of the trial the use of liver transplantations for patients with this disease increased considerably, and transplantation will therefore be regarded as a competing cause of failure of medical treatment. Patients were followed from randomization until treatment failure, drop-out or 1 Jan, 1989; 61 patients died, another 29 were transplanted and 4 patients were lost to followup before 1 Jan. 1989. At entry a number of clinical, biochemical and histological variables, including serum bilirubin, sex, age and standard liver function measures such as aspartate transaminase and alkaline phosphatase were recorded. Data are available from the home page of the textbook Andersen and Skovgaard (2010).

5.1 External validation

The PBC3 study is a multi-center study which combines data from patients treated in the following 6 clinical centres: Hvidovre-Denmark (n=23), London-England (n=150), Copenhagen-Denmark (n=46), Barcelona-Spain (n=79), Munich-Germany
(n=23), Lyon-France (n=28). All analyses are for the sole purpose of illustrating the pseudo-value based calibration plots for competing risks. To achieve an external validation setting we use the London data only to build 4 different models which all predict the probability of dying within 1000 days after the randomization, handling a liver transplant as a competing risk. The first two models are based on Fine-Gray regression analyses [Fine and Gray 1999] where in the first model the variables treatment, age, bilirubin, aspartate transaminase and alkaline phosphatase enter in linear form into the linear predictor whereas for the second model the variables bilirubin, aspartate transaminase and alkaline phosphatase are log-transformed. Note that the Fine-Gray model directly specifies the covariate specific cumulative incidence at 1000 days. It, however, requires a model for the censoring distribution and was performed under the working hypothesis that the censoring is independent of the covariates and the event times. Models three and four are both obtained by combining two cause-specific Cox regression models, one for death and one for liver transplant, into a model for the cumulative incidence at 1000 days. The values of bilirubin, aspartate transaminase and alkaline phosphatase enter model three in raw form and model four after log-transformation. Using the same strategies further 4 models were obtained that predict the risk of liver transplantation handling death as a competing risk.

Table 1 shows summaries of prediction accuracy for the 4 models. The results show that log-transformation of the variables bilirubin, aspartate transaminase and alkaline improves the prediction accuracy (Brier score) and also the discrimination ability (c-index). A log-transformation also has a positive effect on calibration of the models (Figure 1). The figure also indicates that the Fine-Gray models for predicting the risk of liver transplantation are not well calibrated and that none of the models are well calibrated for the predicted probabilities above 50%. However, few predicted probabilities are above 75%. The log-transformations are seen to improve the results. It is worth noting that the Brier score of the models without log-transformations is larger than the Brier score of the reference model. This confirms the miscalibration of these models. However, the concordance index, being a rank statistic, does not assess calibration. In this example the concordance index indicates that the two miscalibrated models are better than the null model.

Table 1 about here.

Figure 1 about here.

5.1.1 Comparison with grouped predictions

A standard approach to calibration curves is to group the predicted risk and then to compute the observed risk of the event by applying a non-parametric estimate (e.g., Kaplan-Meier for survival and Aalen-Johansen for competing risks) locally in each risk group. Our non-parametric estimate is the average of the pseudo-values in the risk groups. Figure 2 compares the calibration curves obtained by non-overlapping groups based on 5, 10 and 20 groups obtained by risk quantiles to the smooth calibration curves obtained by the nearest neighborhood method.
The dotted lines are estimated by the Aalen-Johansen estimate and the solid lines by average pseudo-values. The curves show roughly identical patterns though that based on 20 groups appears very ragged and none of the curves based on separate Aalen-Johansen estimates suggest any problems with the calibration for large predicted probabilities.

5.1.2 Effect of time-point

The pseudo-values can have values below zero and above one. Their value depends on the censoring percentage. In the beginning where few patients are right censored, almost all of the pseudo-values are very close to one or zero. As time proceeds the amount of censoring increases and the range of the predicted gets larger (Figure 3).

5.2 Internal validation

To illustrate our internal validation approach, we applied 1000 bootstrap cross-validation steps to the full Pbc3 data set and compared the resulting calibration curve to the apparent calibration curve. Figure 4 shows the results for the prediction model for the risk of death obtained by combining cause-specific Cox regressions for death and liver transplantation where in both models the variables bilirubin, aspartate transaminase and alkaline phosphatase entered log-transformed. In each bootstrap sample the parameters of the cause-specific Cox regression models were estimated and no variable selection was performed. The apparent calibration curve is obtained by using the full data set both for fitting the model and for estimating the calibration curve. The two bootstrap procedures are seen to provide quite similar results and differ somewhat from the apparent calibration curve (Figure 4). Table 2 summarizes the results. We get a quantification of the over-optimism associated with the apparent calibration curve. Further, the improvement of results by log-transforming the biochemical variables is clearly seen.

6 Discussion

There are numerous published statistical models that predict the probability of an event within a given time frame (e.g., the probability of cardiovascular death within 10 years). It can be difficult to choose between alternative risk models on the basis of an accuracy measure which reflects only discrimination and not calibration. Calibration, or how closely predicted probabilities and observed proportions match, has often generally been restricted to graphical assessment rather than quantification.
The main difficulty with assessing the calibration of a prognostic model lies in the fact that, almost always, the test data used for assessing calibration involve censoring. The presence of censoring causes that some standard procedures for assessing calibration may lead to conclusions that are not consistent with the data. Competing risks constitute another hurdle for the estimation of calibration and also for the interpretation.

The estimation of a calibration plot has similar problems as density estimation. For the nearest neighbor method we need to select a bandwidth and for the grouped risk version of the calibration plot we need to define the groups. Some studies have based a graphical assessment of a prognostic model on a grouping of the predicted event probabilities. Then, the observed outcome is usually reflected by the subgroup Kaplan-Meier estimate. In the presence of competing risks [Wolbers et al. (2009)] used the subgroup Aalen-Johansen estimate to estimate the observed risk of the event of interest. However, this calculation requires arbitrary grouping of patients (e.g., deciles of predicted risk) and this grouping can affect the visual interpretation of calibration. Similarly, by selecting different bandwidths the user can affect the result of the nearest neighbor approach. To achieve somewhat unbiased results it is recommended to group risk only if there are different medical actions are associated with the risk groups. For the nearest neighbor method we recommend to use an optimal bandwidth (e.g. Sheather and Jones [1991]).

Many measures of discrimination are in use that do not (and are not intended to) reflect calibration. For example, the concordance index or time-dependent AUC, each are useful for specific discrimination assessments. Because they do not reflect calibration, they generally require supplementation with a calibration plot to determine (and not just visualize) calibration accuracy. The Brier score summarizes both calibration and discrimination at the same time. It is also quite interpretable and intuitive: the square root of the Brier score (root mean squared error) is the expected distance between the observation and the prediction on the probability scale. The relation of the Brier score of a prediction model to that of the reference model which ignores the covariates can further be used to construct a time-dependent $R^2$ measure [Gerds et al. (2008)].

An important limitation of our study is that we are assuming equal misclassification costs in the prognostic model setting. In addition, we are not addressing the use of a prognostic model for a specific decision.

Commonly these procedures are underestimating the performance of the final model because during cross-validation the models are built on less information. Several estimates have been proposed to reduce this bias by a linear combination of the optimistic apparent performance and the pessimistic cross-validation performance, e.g. bootstrap validation [Steyerberg [2009] chapter 17], the bias corrected bootstrap [Harrell [2001] p. 494] and the .632+ bootstrap [Efron and Tibshirani [1997]]. While these approaches may work for grouped predictions for most modelling strategies, it is not clear how to combine these approaches with smoothing.
References


## List of Figures

1. Pbc3 trial: Pseudo-values and smoothed calibration curves for prediction models built in the data of one center (London) and estimated using the data of the other centers (Hvidovre, Copenhagen, Barcelona, Munich, Lyon). The variables bilirubin, aspartate transaminase and alkaline phosphatase are log-transformed in the lower panel but not in the upper panel. All lines are obtained with a nearest neighborhood smoother.  
   ![Figure 1](image1.png)  
   
2. Pbc3 trial: Comparison of smoothed calibration curve with differently coarse classifications of the risk predicted by the cause-specific Cox regression model for death within 1000 days. 
   
3. Pbc3 trial: Smoothed calibration curves for the risk of death predicted by the cause-specific Cox regression model evaluated at different time points. 
   
4. Pbc3 trial: Smoothed calibration curves obtained via internal validation of the cause-specific Cox regression model for predicting the risk of death within 1000 days. 
   
---

13  
14  
15  
16
Figure 1: Pbc3 trial: Pseudo-values and smoothed calibration curves for prediction models built in the data of one center (London) and estimated using the data of the other centers (Hvidovre, Copenhagen, Barcelona, Munich, Lyon). The variables bilirubin, aspartate transaminase and alkaline phosphatase are log-transformed in the lower panel but not in the upper panel. All lines are obtained with a nearest neighborhood smoother.
Figure 2: Pbc3 trial: Comparison of smoothed calibration curve with differently coarse classifications of the risk predicted by the cause-specific Cox regression model for death within 1000 days.
Figure 3: Pbc3 trial: Smoothed calibration curves for the risk of death predicted by the cause-specific Cox regression model evaluated at different time points.
Figure 4: Pbc3 trial: Smoothed calibration curves obtained via internal validation of the cause-specific Cox regression model for predicting the risk of death within 1000 days.
List of Tables

1. Pbc3 trial: External validation results. C-index truncated at 1000 days and Brier scores evaluated at 1000 days for prediction models built in the data of one center (London). The calibration curves are estimated using the data of the other centers (Hvidovre, Copenhagen, Barcelona, Munich, Lyon). AJ: Aalen-Johansen estimate; FGR: Fine-Gray regression; CSC: Cause-specific Cox regression. . . . 18

2. Pbc3 trial: Internal validation results based on 1000 bootstrap cross-validation steps. C-index truncated at 1000 days and Brier scores evaluated at 1000 days. AJ: Aalen-Johansen estimate; FGR: Fine-Gray regression; CSC: Cause-specific Cox regression. . . . . . . . . . 19

17
<table>
<thead>
<tr>
<th>Model Type</th>
<th>Risk of liver transplantation</th>
<th>C-index</th>
<th>Brier score</th>
<th>Risk of death</th>
<th>C-index</th>
<th>Brier score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJ Reference</td>
<td>50.00</td>
<td>3.18</td>
<td>50.00</td>
<td>12.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGR Model 1</td>
<td>84.98</td>
<td>3.23</td>
<td>81.63</td>
<td>12.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGR (log) Model 2</td>
<td>92.78</td>
<td>2.49</td>
<td>86.27</td>
<td>8.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSC Model 3</td>
<td>61.31</td>
<td>3.34</td>
<td>82.00</td>
<td>11.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSC (log) Model 4</td>
<td>90.29</td>
<td>2.82</td>
<td>85.58</td>
<td>8.86</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Pbc3 trial: Internal validation results based on 1000 bootstrap cross-validation steps. C-index truncated at 1000 days and Brier scores evaluated at 1000 days. AJ: Aalen-Johansen estimate; FGR: Fine-Gray regression; CSC: Cause-specific Cox regression.

<table>
<thead>
<tr>
<th>Risk of liver transplantation</th>
<th>Risk of death</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-index</td>
<td>Brier score</td>
</tr>
<tr>
<td>AJ Reference</td>
<td>50</td>
</tr>
<tr>
<td>FGR Model 1</td>
<td>80.21</td>
</tr>
<tr>
<td>FGR (log) Model 2</td>
<td>85.5</td>
</tr>
<tr>
<td>CSC Model 3</td>
<td>81.91</td>
</tr>
<tr>
<td>CSC (log) Model 4</td>
<td>86.4</td>
</tr>
</tbody>
</table>

Bootstrap cross-validation (B=1000,1000 days)

<table>
<thead>
<tr>
<th>Risk of liver transplantation</th>
<th>Risk of death</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-index</td>
<td>Brier score</td>
</tr>
<tr>
<td>AJ Reference</td>
<td>50</td>
</tr>
<tr>
<td>FGR Model 1</td>
<td>77.91</td>
</tr>
<tr>
<td>FGR (log) Model 2</td>
<td>83.77</td>
</tr>
<tr>
<td>CSC Model 3</td>
<td>78.45</td>
</tr>
<tr>
<td>CSC (log) Model 4</td>
<td>85</td>
</tr>
</tbody>
</table>
Research Reports available from Department of Biostatistics
http://www.pubhealth.ku.dk/bs/publikationer

Department of Biostatistics
University of Copenhagen
Øster Farimagsgade 5
P.O. Box 2099
1014 Copenhagen K
Denmark

11/1 Kreiner, S. Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment.

11/2 Andersen, P.K. Competing risks in epidemiology: Possibilities and pitfalls.

11/3 Holst, K.K. Model diagnostics based on cumulative residuals: The R-package gof.


11/5 Holst, K.K., Budtz-Jørgensen, E. & Knudsen, G.M. A latent variable model with mixed binary and continuous response variables.


12/01 Keiding, N., Hansen, O.K.H., Sørensen, D.N. & Slama, R. The current duration approach to estimating time to pregnancy.

12/02 Andersen, P.K. A note on the decomposition of number of life years lost according to causes of death.
12/03 Martinussen, T. & Vansteelandt, S. A note on collapsibility and confounding bias in Cox and Aalen regression models.


12/05 Martinussen, T. & Pipper, C.B. Estimation of Causal Odds of Concordance based on the Aalen additive model.

12/06 Binder, N., Gerds, T.A. & Andersen, P.K. Pseudo-observations for competing risks with covariate dependent censoring.

12/07 Keiding, N. & Clayton, D. Standardization and control for confounding in observational studies: a historical perspective.

12/08 Hilden, J. & Gerds, T.A. Evaluating the impact of novel biomarkers: Do not rely on IDI and NRI.


13/01 Forman, J.L. & Sørensen M. A new approach to multi-modal diffusions with applications to protein folding.


13/03 Wolbers, M., Koller, M.T., Witteman, J.C.M. & Gerds, T.A. Concordance for prognostic models with competing risks.

13/04 Andersen, P.K., Canudas-Romo, V. & Keiding, N. Cause-specific measures of life lost.

13/05 Keiding, N. Event history analysis.


13/07 Hansen, S., Andersen, P.K. & Parner, E. Events per variable for risk differences and relative risks using pseudo-observations.