Incorporating model selection uncertainty into prognostic factor model predictions

Nicole H. Augustin*†, Willi Sauerbrei*
and Martin Schumacher*

Universität Freiburg i. Br.

Nr. 76 February 2002

† Zentrum für Datenanalyse und Modellbildung, Universität Freiburg
Eckerstrasse 1, D–79104 Freiburg im Breisgau, Germany
nicole@fdm.uni-freiburg.de

*Institut für Medizinische Biometrie und Medizinische Informatik,
Universitaetsklinikum Freiburg,
Stefan-Meier-Str. 26, D-79104 Freiburg, Germany
wfs@imbi.uni-freiburg.de, ms@imbi.uni-freiburg.de

SUMMARY. Prediction of disease outcome in prognostic factor models are usually based on one selected model. However, often several models fit the data equally well, but these models might differ substantially in terms of included factors and might lead to different predictions for individual patients. We give an overview of approaches for accounting model selection uncertainty and present two specific adaptations for the use with proportional hazard Cox models. The first one is Bayesian model averaging (BMA) adapted for the proportional hazard model (Volinsky et al., 1997). As a second approach we propose a method which does not incorporate model uncertainty within a strictly Bayesian framework. This is done by averaging over the set of possible models using weights estimated from bootstrap resampling. To reduce the set of models we perform an initial screening of factors based on the inclusion probability of each factor. In two data examples we compare predictions of the two approaches with "conventional" predictions from a single model. Comparisons are based on the partial predictive score and the Brier score. In one example we also compare confidence intervals of linear predictors.

KEY WORDS: bootstrap; model selection uncertainty; model averaging; prognostic factor models; survival analysis.

1. Introduction

Prognostic studies are about identifying factors influencing disease outcome. Identifying and ranking these factors involves statistical model building, usually of a regression model of some form, e.g. in medicine this is often the logistic regression model or the Cox proportional hazards model (Cox, 1972). The resulting models are the basis for predictions of disease outcome, which are of potential value for therapeutic and other clinical decisions. Also, risk groups based on the prognostic factors can be defined. These are useful for risk-adapted therapy strategies or for selecting groups of patients for clinical trials.

For many diseases studied there is a vast number of potential factors and there is often a lot of uncertainty about the best model for prediction. For example in breast cancer the effects of more than 160 prognostic factors are currently discussed. Even if only 10 variables are investigated in one study, this can lead to a large set of possible models. Not surprisingly it is often the case that several models provide similar fits to the data, but these models differ in terms of included factors and predicted prognostic indices for individual patients. The differences are often substantial if these predictions are made at extreme values of the range of the data. Also, finding the variables with the highest prognostic value is a difficult task, mainly because selected models are often unstable (Steyerberg et al., 2001; Sauerbrei, 1999; Breiman, 1996b; Sauerbrei and Schumacher, 1992; Altman and Andersen, 1989; Chen and George, 1985).

Standard statistical methodology involved in making the predictions and variance estimation is based on the assumption that the model used was pre-specified without any data-dependent modelling and that it is the *true model*. Ignoring the uncertainty about the selection process may lead to unjustifiable confidence in parameter estimates and classification schemes derived from the selected model (Holländer et al., 2001). For example, confidence intervals may not yield their nominal coverage. One possible way to incorporate model selection uncertainty is to average over the set of possible models. In cases where the aim is to find the most likely model, it is useful to state the uncertainty attached to the selected model along with the results of other good approximating models. In this paper we describe some of the currently existing methods for handling model selection uncertainty. These are mainly Bayesian approaches (Chatfield, 1995; Draper, 1995; Kass and Raftery, 1995; Raftery et al., 1997; Hoeting et al., 1999) and methods based on bootstrap resampling (Buckland et al., 1997; Burnham and Anderson, 1998). Most of the applications of these methods have been in the context of mark-recapture (Hook and Regal, 1997; Stanley and Burnham, 1998; Madigan and York, 1997; Brooks et al., 2000). Viallefont et al. (2001) apply Bayesian model averaging in a case-control study.

Here we are specifically interested in applications of the above approaches to survival data. The Bayesian model averaging (BMA) method was adapted to survival data by Volinsky et al. (1997); see also Hoeting et al. (1999) for an application. We propose as an alternative a *two step* approach where first a subset of variables is selected and then model averaging is applied using the bootstrap. Then we investigate whether these two methods can improve predictions for prognostic models of cancer (glioma) and a liver disease (primary biliary cirrhosis) relative to three alternatives: a single model selected using backward elimination (BE), the full model and predictions using the Kaplan-Meier estimate.

2. Data

Example 1. **Glioma**

Data are available from a randomised clinical trial comparing two chemotherapies of 411 patients with malignant glioma. The median follow-up time was 712 days. Besides therapy, 15 variables are considered to be potential prognostic factors including patient characteristics, e.g. age, gender and clinical characteristics, e.g. presence of aphasia, epilepsy. A Cox model is fitted to the survival data with 274 deaths. The data have been analysed in the past (Ulm et al., 1989; Sauerbrei and Schumacher, 1992; Sauerbrei, 1999). In the most recent investigation the full model was considered and three other possible models were selected using BE with different significance levels (α =0.157, 0.05 and 0.01). Resulting models had nine, five and four variables respectively (Table 1). All models had the variables *mali1*, *age*, *Karn1* and *surg1* in common. Although the individual prognostic indices, i.e. the linear predictors, from the three different models are highly correlated, with the highest pairwise Pearson's correlation coefficient of 0.99 and the lowest 0.94 (full model and BE(0.01)) (Table 1), the sizes of standard errors of the prognostic indices for the four models differ substantially, with the full model having the highest standard errors. Judged from the log-likelihood after the full model, the model based on BE(0.157) with 9 variables marginally yields the best fit to the observed data. The two complex models may be overparameterised and we do not know whether they are useful for describing data of a similar study collected under slightly different conditions, e.g. at a different time point, or data from different clinics but under the same protocol. In the following we will describe methods for attaching levels of certainty to the different possible models.

Example 2. **Primary Biliary Cirrhosis**

Our second example are survival times of 312 patients with the fatal chronic liver disease primary biliary cirrhosis (PBC). The data are from a double-blind randomised clinical trial to compare the drug DPCA with a placebo and have median follow-up time of 2504 days. More details about the study can be found in Dickson et al. (1985), Markus et al. (1989), Fleming and Harrington (1991). There are 131 events. The clinical, biochemical, serological and histological parameters listed in Table 5 are possible prognostic factors for the disease. Fleming and Harrington (1991) used a Cox regression model for the data and selected five of the variables (age, l.albu, l.bili, edema and l.prothr). Hoeting et al. (1999) and Raftery et al. (1996) analysed these data using BMA.

3. Some background on survival models

The most popular model fitted to survival data from prognostic studies is the Cox model where the hazard λ at time t of the event is modelled as a function of the explanatory variables $\boldsymbol{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$, which are assumed to have been recorded at the start of the study. Below we explain the basic structure of the Cox model; this will be useful later for understanding to which quantity model averaging should be applied and how the predictive performance is best assessed. The model is

$$\lambda_i(t|\boldsymbol{X}) = \lambda_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients of the explanatory variables. The parameter $\lambda_0(t)$ is the baseline hazard, which is the hazard function when all explanatory variables take the value zero. $\lambda_0(t)$ is estimated in an extra step of the iterative scoring algorithm. For the estimation of the vector $\boldsymbol{\beta}$, assuming no ties are present, the following partial likelihood is maximised

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{l\in R(t_i)}\exp(\mathbf{x}_l'\boldsymbol{\beta})}\right)^{\delta_i}, \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, t_i is the observed time to the i'th event from one of the $1, ..., n$ patients. The indicator δ_i is zero if t_i is right-censored, unity otherwise. $R(t_i)$ denotes the set of individuals at risk at time t_i. If the model is used for prognosis the survivor function will be estimated by:

$$\hat{S}_i(t) = [\hat{S}_0(t)]^{\exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})},$$

where $S_0(t)$ is the baseline survivor function which is estimated in an extra step in each iteration of the scoring algorithm used to maximise $L(\boldsymbol{\beta})$.

As apparent from the two data examples there is typically a high number of possible prognostic factors, say above 10. Even without fitting interactions and considering only the in- and exclusion of variables this results in $2^{10} = 1024$ models. In the glioma example we have $2^{15} = 32768$ and in the PBC example there are 2^{13} possible models. If prior information does not lead to a smaller set of candidate variables with a potential effect, the model space needs to be reduced. The two approaches we describe below achieve this (a) by directly choosing a subset of models or (b) indirectly by choosing a smaller subset of explanatory variables.

4. Quantifying model selection uncertainty

In any modelling process several implicit assumptions are made before the actual formal model selection process starts. These assumptions include data interpretation, e.g. variable definitions, coding and dealing with missing values. For the analysis of a case-control study Blettner and Sauerbrei (1993) give some background on these pre-model selection decision processes. Also a model class is chosen, for instance whether a Cox model or a fully parametric model is fitted to survival data. Then a decision on the appropriate criterion for defining the "best" model for model selection is needed. These include combinations of measures on the discrepancy of fit (the deviance function), the prediction error, model stability or criteria like model complexity. Although decisions on these aspects can have a strong influence on the prognostic index, the methods we present here are restricted to the kind of model selection carried out after these decisions have been made. We also restrict ourselves to only fitting linear terms in our model.

4.1 *Bayes approaches*

Bayesian model averaging (BMA) is one approach to incorporating model uncertainty into inference. We give a short summary of the method; more details are given in Chatfield, 1995; Draper, 1995; Kass and Raftery, 1995; Hoeting et al., 1999. Consider the possible set of models $M_1, ..., M_j, ..., M_J$ with associated prior probabilities of M_j being the true model. The set $M_1, ..., M_J$ is often called the model space. Let θ be the parameter or quantity of interest, e.g. the linear predictor (prognostic index). Given the data D the posterior distribution $\mathrm{pr}(\theta|M_j, D)$ of the estimated parameter θ under a specific model j is averaged over the different models considered, weighted by the respective $\mathrm{pr}(M_j|D)$:

$$\mathrm{pr}(\theta|D) = \sum_j^J \mathrm{pr}(\theta|M_j, D)\mathrm{pr}(M_j|D) \tag{3}$$

Both terms in equation 3 require solving integrals, for the general linear model these are available in closed form; this is not the case for the Cox model. The posterior summary statistics are evaluated by incorporating the model posterior probabilities, e.g. the posterior mean is

$$E(\theta|D) = \sum_j^J E(\theta_j|M_j, D)\mathrm{pr}(M_j|D). \tag{4}$$

The specification of prior model probabilities $\mathrm{pr}(M_j)$ assumes that one of the models considered is true, since all other models which do not belong to the set $M_1, ..., M_J$ have a prior probability of zero. Defining suitable priors $\mathrm{pr}(M_j)$ for the different models appears to be an unresolved problem with this approach, especially with a large set of possible models. The simplest suggested solution is to give each model equal weight when no prior information is available. With respect to the priors of the model parameters Fernández et al. (1998) have developed methodology for choosing "benchmark" priors. With benchmark priors they mean an "automatic" prior structure, which is partly non-informative. They show that in certain simulated scenarios their benchmark priors have little influence on posteriors when BMA is applied in linear regression. Since heir benchmark priors are specific for linear regression, prior definition is still a problem for other models. Another problem of the BMA approach is that large summations and integrals are necessary especially if the model space is large, and these are often hard to compute. One approach to handle the large model space is to restrict it to models with a posterior probability above a certain threshold, e.g. by using Occam's algorithm proposed by Madigan and Raftery (1994). In this algorithm all models with a posterior probability greater than a certain proportion of the model with the highest posterior are included (Occam's window). Another approach uses Markov chain Monte Carlo (MCMC) to approximate the model space. Different MCMC sampler that jump between parameter subspaces of differing dimensions have been suggested (George and McCullagh, 1993; Madigan and York, 1995; Madigan and York, 1997; Green, 1995; Dellaportas and Forster, 1999; King and Brooks, 2001). The reversible jump algorithm of Green (1995) is one of these. To our knowledge MCMC has not been used in the context of censored survival data.

4.2 *Bayesian model averaging for survival data (approx. BMA)*

For carrying out BMA within the Cox model, the integrals of both components in equation 3 are not available in closed form. Volinsky et al. (1997) suggest several approximations (see also Hoeting et al., 1999). Instead of integrating the posterior distribution of the quantity of interest Volinsky et al. (1997) use the partial maximum likelihood approximation. If the quantity of interest is the prognostic index, this so called predictive distribution is the partial likelihood in equation 2 with the maximum likelihood estimator $\hat{\theta}_j$ inserted. For deriving $\mathrm{pr}(M_j|D)$, the posterior model probabilities in equation 3, the log of the integrated likelihood is approximated by the Bayesian information criterion (BIC) (Schwarz, 1978).

To find the subset of models the steps of the Occam's window algorithm are followed. All models where the posterior probability is greater than a specified constant, e.g. $1/C$ of the model with the highest posterior are included (Occam's window). Instead of calculating all posterior model probabilities an altered version of the leaps and bounds algorithm, also used for all subset selection (Furnival and Wilson, 1974; Lawless and Singhal, 1978; Kuk, 1984), is used to narrow down the choice of models. Volinsky et al. (1997) claim that by choosing a large enough set of "best" models with the leaps and bounds algorithm, all models within Occam's window are found. For these pre-selected models the $\mathrm{pr}(M_j|D)$ is calculated using the BIC approximation and models which do not meet the Occam's window requirement are excluded. Posterior mean estimates and variance of θ are calculated using weighted averages from the separate models with weights equal to $\mathrm{pr}(M_j|D)$, with summation over models within Occam's window.

Volinsky et al. (1997) apply this approach to Cox models for assessing the risk of stroke from a cardiovascular health study. As a measure for the importance of the different variables, posterior effect probabilities $\mathrm{pr}(\beta_l \neq 0|D)$ for the coefficients associated with each variable, are given. This is the sum over all $\mathrm{pr}(M_j|D)$ of all models containing variable x_l. In the following we will call their approach *approx. BMA*. To our knowledge it is currently not implemented in any of the mainstream statistics packages or Bayesian specialist software such as BUGS, but the authors have made some of their code publicly available under `http://www.research.att.com/~volinsky/bma.html`.

4.3 *Non-Bayesian approaches*

Whereas in Bayesian statistics there have been extensive developments for model selection uncertainty, the efforts in non-Bayesian approaches have been more tentative. Buckland et al. (1997) have suggested several more or less *ad hoc* approaches to accounting for model uncertainty. In their approach an information criterion (I), e.g. the Akaike information criterion, is used to estimate model weights for parameter averaging and variance estimation. These weights have the following form

$$w_j = \frac{\exp(-I_j/2)}{\sum_{i=1}^{J} \exp(-I_i/2)}, j = 1, ...J,$$

where I_j is the information criterion for model j. The expected value of the averaged parameter is then

$$\hat{\theta} = \sum_{j}^{J} \hat{\theta}_j \; w_j, \tag{5}$$

where $\hat{\theta}_j$ is the estimated parameter or quantity from the observed data. Setting I equal to the Bayes information criterion, these weights are almost equal to the model posterior $\mathrm{pr}(M_j|D)$ as it is approximated by Volinsky et al. (1997) (see above). The only difference is that in $\mathrm{pr}(M_j|D)$ the normalising constant is not the sum over all possible models, but a sum over a smaller subset of best models as identified by the leaps and bounds algorithm and Occam's window.

As an alternative way of obtaining model weights w_j Buckland et al. (1997) suggest to apply the bootstrap to model selection. Then the weight of model j is estimated by the proportion of bootstrap samples in which model j was selected. The idea of bootstrapping the modelling process has been suggested in the past mainly for investigating model selection stability issues (e.g. Efron and Gong, 1983; Chen and George, 1985; Dijkstra and Veldkamp, 1994; Altman and Andersen, 1989; Sauerbrei and Schumacher, 1992; Hjorth, 1994), but not many investigations on utilising the results for model averaging exist. Candolo et al. (2002) show that in the linear regression model the above approach reduces the mean squared error if effects of covariates are relatively weak.

A method that appears to be very similar to model averaging is bootstrap aggregating or bagging (Breiman, 1996a). In bagging the bootstrap is applied to model selection and the parameter of interest, e.g. the prognostic index, is an average of the B bootstrap estimates:

$$\hat{\theta} = \frac{1}{B} \sum_{j}^{B} \hat{\theta}_j^b,$$

where $\hat{\theta}_j^b$ is the j'th bootstrap estimate with $j = 1, ..., B$.

4.4 *Screening of variables and bootstrap model averaging (bootstrap MA)*

We use the approach of Buckland et al. (1997) but reduce in a first step single variables thought to have no influence before we apply model averaging. This is in contrast to the approx. BMA approach where the set of models is reduced. For the *first step* we use parts of a strategy suggested by Sauerbrei and Schumacher (1992) based on bootstrap resampling. Model selection is carried out in the bootstrap samples. In order to make our selection procedure comparable to approx. BMA we use BE with a selection level corresponding to the BIC criterion where the penality for sample size is $\log(n)$, n is here the number of events in the data. We denote this selection algorithm by BE_{BIC}. Then the number of times each covariate is included in the model, i.e. the inclusion frequency $h(x_i)$, is recorded. Instead of the BIC criterion another criterion, e.g. the AIC can be used. Now a subset of variables is selected only containing variables with an inclusion frequency above a minimum value. According to Sauerbrei and Schumacher (1992) variables with no or hardly any prognostic influence will have very low inclusion frequencies and will therefore be eliminated in this first step. In a *second step* we carry out another bootstrap resampling procedure and apply BE_{BIC} to the subset of remaining variables from the *first step*. The set of possible models is substantially reduced after the first step, e.g. from 2^{15} to 2^7 as in our first example. Now we record the number of times each model is selected and use these inclusion frequencies $h(M_j)$ as weights in equation 5 for model averaging. The *second step* in our procedure makes model selection in each bootstrap sample much faster and considers only models in which every factor has probably some influence and therefore gives larger weights to the remaining models. The precision of relative model inclusion frequencies $h(M_j)$ can be increased by a larger number of bootstrap replicates.

5. Assessing the predictive performance

5.1 *Criteria*

We compare the predictive performance of the different approaches using (1) the partial predictive score (PPS) and (2) the Brier score (Brier, 1950; Graf et al., 1999).

(1) Partial predictive log-score (PPS): Volinsky et al. (1997) suggested using the partial predictive log-score (PPS). This is a possible criteria if the quantity of interest is the prognostic index. Let D^{build} be the data used for the model building process and D^{test} the data used for validation. The PPS is based on the logarithmic scoring rule, the summed logarithm of the observed ordinate of the predictive density distribution for the observations in D^{test}:

$$\sum_{d \in D^{test}} \log \mathrm{pr}(d|M_j, D^{build})$$

and for BMA the PPS is then

$$\sum_{d \in D^{test}} \log \left\{ \sum_{j=1}^{J} \mathrm{pr}(d|M_j, D^{build}) \ \mathrm{pr}(M_j|D^{build}) \right\}.$$

Since for the Cox model the predictive density $\mathrm{pr}(d|M_j, D^{build})$ is not available Volinsky et al. (1997) suggest to use the partial likelihood (PL) instead, with the maximum likelihood estimator $\hat{\theta}_j$ from D^{build} inserted.

$$\mathrm{PPS} = \sum_{d \in D^{test}} \log \left\{ \sum_{j=1}^{J} PL(\hat{\boldsymbol{\beta}}) \mathrm{pr}(M_j|D^{build}) \right\},$$

then the PPS is the sum of the partial log-likelihood evaluated for each observation in D^{test}. For evaluating the PPS under bootstrap MA the weight

$\mathrm{pr}(M_j|D^{build})$ is replaced by bootstrap inclusion frequency $h(M_j)$ for model j. The higher the PPS the better the predictive performance. Since observations in D^{test} yielding high ordinates in the predictive density distribution as derived from the model fitted to D^{build}, the model is also plausible for D^{test}.

(2) Brier score: If the quantity of interest is the predicted survivor function at a given time t^*, which is a function of the linear predictor and the baseline hazard function, a criterion evaluating the predicted survivor function directly is possibly most appropriate. The Brier score (Graf et al., 1999) is such a criterion. The prediction of the survivor function involves estimation of the prognostic index and of the baseline hazard $\lambda_0(t_0)$ in equation 1. The Brier score is the expected prediction error at a given time point t^*:

$$BS(t^*) = E[(I(T > t^*) - \hat{S}(t^*|X))^2],$$

where $I(.)$ is an indicator function. It is estimated as follows

$$\widehat{BS(t^*)} = \frac{1}{n}\sum_{i=1}^{n} w_i(I(t_i > t^*) - \hat{S}(t^*|x_i))^2,$$

where the weights w_i are equal to one if no censoring is present. With censoring weights for observations with an event are equal to the inverse of the Kaplan-Meier estimate of the censoring distribution at the time of the event. Individuals without an event up to t^* are weighted according to the Kaplan-Meier estimate of the censoring distribution at time t^*. Observations censored before time t^* do not contribute to the Brier score, i.e. $w_i = 0$. The lower the Brier score the better the predictive performance.

With model averaging the event probability $S(t^*|X)$ is a weighted average of the estimated survivor functions under the different models. In the situation where one data set D^{build} is used for the model building process and another data set D^{test} is used for validation, the parameters $\boldsymbol{\beta}$ and the baseline hazard λ_0 are estimated from D^{build}.

The Brier score from the pooled Kaplan-Meier estimate of the survivor function is a benchmark value for a prediction without taking explanatory information into account. For measuring the gain from explanatory information in the model we use the percentage change in Brier score compared to a prediction from the Kaplan-Meier estimate (Graf et al., 1999): $\widehat{R^2} = 1 - B\widehat{S(t^*)}_{model}/B\widehat{S(t^*)}_{K-M}$. R^2 is thus one (of many) analogous to the explained variation commonly quoted for general linear models. Useful as a "worst case" benchmark of the BS is 0.25 from a naive prediction with constant survival probability of 0.5.

5.2 *Cross-validation*

We use *K-fold cross-validation* (Davison and Hinkley, 1997) to estimate the predictive criteria. The data are split into K disjoint sets of similar size. A rule of thumb is to use either $K = \sqrt{n}$ or $K = 10$ in the case of large data sets, where choosing $K = \sqrt{n}$ becomes very computer intensive (Davison and Hinkley, 1997). These K sets define the test set D^{test} for assessing the predictive performance and the build set D^{build} used to fit the model. In turn, one of the K sets is the test set and the remaining data is the build set. This procedure ensures one prediction per observation. The observations in the K sets are stratified by event and censored cases.

Then the following cross-validation procedure is carried out:

1. Perform the model building procedure (backwards elimination (BE_{BIC}), bootstrap MA or approx. BMA) on the *build* data to obtain

(a) the estimates $\hat{\boldsymbol{\beta}}_b$ and $\hat{\lambda}_0$ under BE_{BIC}, ($\hat{\boldsymbol{\beta}}_b$ is a vector and coefficients of variables which are not selected are set to zero.);

(b) respectively $\hat{\boldsymbol{\beta}}_{b1}, ..., \hat{\boldsymbol{\beta}}_{bJ}$ and $\hat{\lambda}_{01}, ..., \hat{\lambda}_{0J}$ for the different J selected models under bootstrap MA or approx. BMA.

2. Predict the quantity of interest, e.g. $\hat{S}(t^*|X)$ or the linear predictor $\boldsymbol{\theta}$ for the test data, with covariate matrix $\boldsymbol{X}^{test}$.

3. Calculate the assessment criteria, e.g. the Brier score or the PPS.

4. repeat steps 1 - 3 for each of the K data splits and sum the criteria obtained from the 1 to K data splits.

When D^{test} is too small, say below 10 observations, to use its censoring distribution to calculate the w_i's in the Brier score, we use D^{build} for this purpose. If D^{test} consists only of one observation and D^{build} contains the remaining data, which is leave-one-out cross-validation, the PPS criterion is the same as the cross-validated log-likelihood introduced by Van Houwelingen and LeCessie (1990).

6. Examples

6.1 *Glioma*

We applied our bootstrap MA approach to the glioma data as described above using 500 bootstrap resamples in all steps involved.

In the first step of bootstrap MA the inclusion frequencies $h(x_i)$ range between 4% and 100% (Table 2 and Figure 1). The variables *mali1*, *age* and *surg1* have the highest inclusion frequencies, all are above 75%. The variable *Karn2* has the lowest inclusion frequency of 4%. In the second step of the bootstrap MA procedures only variables with $h(x_i) > 30\%$ are included and the bootstrap procedure is applied again to the remaining subset with *mali1*, *mali2*, *age*, *Karn1*, *surg1*, *epil* and *cort*. In total there were 27 different models selected out of 2^7 possible models; the top ten are shown with model inclusion frequencies $h(M_j)$ in Table 2. The model including variables *mali1*, *age*, *Karn1* and *surg1* has the highest inclusion frequency (17%). Repeating the bootstrap MA procedure with different selection levels $h(x_i) > 40\%, 20\%, 10\%$ and 0% in the first screening step always yields the same model with the highest $h(M_j)$, but the distributions of the respective $h(M_j)$ differ of course (Table 2). For the relatively high selection level of 40% in the screening step a set of 7 models were selected. For selection levels 20%, 10% and 0% the resulting model sets are much larger. There where 107, 184 and 226 models in the respective sets. This reduced the $h(M_j)$ of the strongest models.

When applying approx. BMA to the glioma data we obtain fairly similar answers. As in the example of Volinsky et al. (1997) all models with a posterior probability greater $1/C$ with $C = 20$, were included. In total 41 models were selected into the set. Table 3 shows the first 10 models with the highest posterior model probabilities. The model representing the highest part (16%) of the total posterior model probability is the same as was selected by BE_{BIC} and under bootstrap MA. The variables *mali1*, *age* and *surg1* have the highest posterior effect probabilities $\mathrm{pr}(\beta_l \neq 0|D)$ of 100%. These variables also have the highest inclusion frequency under bootstrap MA. Figure 2 shows the scatter plot of bootstrap inclusion probabilities from the bootstrap MA approach versus the posterior effect probabilities $\mathrm{pr}(\beta_l \neq$

$0|D)$. With *BMA* variables with strong effects appear more "important" than with bootstrap MA. Some of these differences are due to the fact that the $\mathrm{pr}(\beta \neq 0|D)$ are derived after the model space has been reduced, whereas the $h(x_i)$ are estimated before the model space reduction. The $\mathrm{pr}(\beta \neq 0|D)$ resulting from the *BMA* procedure without reducing the model space using Occam's window yield a closer agreement with the $h(M_j)$ for most variables (also shown in Figure 2).

We compare the predictive performance of the model averaging methods with the performance of predictions using the Kaplan-Meier estimate, the full Cox model and the single model as chosen by BE_{BIC}. We predict the probability of survival at a relatively early time point after diagnosis, at 310 days, and after two years at 730 days. These time-points where chosen by looking at the overall Kaplan-Meier curve, at day 310 the estimated survival probability is about 0.5. Day 730 is after the median follow-up time of 712 days and the estimated survival probability is about 0.2. We estimate the Brier score using 20-fold cross-validation. The results shown in Table 4 do not indicate any major differences. There is some benefit in introducing explanatory information: the percentage reduction of the Brier score (R^2) is around 15% for the four model based predictions at day 310 and at day 730 it is around 28%. The full model yields the highest R^2 at day 310 and at day 730 the lowest. Differences in predictions of the prognostic indices with the four model-based approaches are also small, the estimated partial predictive scores (PPS) from the 20-fold cross-validation are all similar (Table 4).

6.2 *Primary Biliary Cirrhosis*

In the PBC example our pre-model selection decisions are as follows. We apply BE_{BIC} to bootstrap samples in the bootstrap MA approach. For the 123 events this is in correspondence to the BIC with a penalty of log(123). Fleming and Harrington (1991) showed that the drug DPCA did only have a negligible effect on prognosis, hence we do not include it. Following Hoeting et al. (1999) we exclude 8 observations due to missing values in the covariates. In addition we exclude the binary variable ascites (swelling in the abdominal cavity) due to the fact that less than 8% of cases had ascites present and this caused convergence problems in some of the bootstrap samples. In the full model the effect of ascites effect is not strong. Like Hoeting et al. (1999) we used the log of bilirubin, albumin, SGOT, alkaline phosphates. In contrast to Hoeting et al. (1999) we do not use the log of prothrombin time. From previous analysis (Fleming and Harrington, 1991) it is not clear whether transforming this variable improves the model fit. Also, using the log leads to a very high expected odds ratio in the full model and there are convergence problems in the bootstrap samples.

We compare bootstrap confidence intervals of linear predictors from three observations for the non-Bayes methods with a small, medium and large value of the predictor (Figure 3). For all three observations BE_{BIC} yields the shortest 95% confidence interval. Confidence intervals of the full model and bootstrap MA have intervals of similar width, but do not cover the same range. In contrast to the full model, bootstrap MA does not require information from all variables.

The results given in Table 5 show that the model selection uncertainty in this example is much higher than in the glioma example. 65 different models were selected in the *second step* of bootstrap MA, of which inclusion frequencies are all below 12%. In Figure 4 we see that the relation between the bootstrap inclusion frequencies and posterior effect probabilities is similar as in the glioma example. The variables l.albu, age, prothr, l.urcop and edema with high posterior probabilities appear to be more important under approx. BMA than under two step bootstrap MA. Variables with low posterior probabilities appear more "important" under two step bootstrap MA.

We compare predictive performance of our averaging approaches with the predictions using the Kaplan-Meier estimate, the full Cox model, and the selected model from BE_{BIC} (Table 7). The Brier score is estimated, at given time points $t^* =1000$ and 3000 days, approximately 3 and 8 years; with an estimated survival probability of about 0.8 and 0.6 respectively using the Kaplan-Meier method. At $t^* =1000$ the percentage reduction of the Brier score (R^2) is around 35% for the four model-based predictions; the full model yields the highest $R^2 = 36\%$ and there is no difference between bootstrap MA and approx. BMA. At $t^* = 3000$ the bootstrap MA procedure leads the highest $R^2 = 31\%$. Differences in predictions of the prognostic indices with the four model-based approaches are small, again the estimated partial predictive scores (PPS) is lowest for the full model.

7. Discussion

Our two examples show situations where there is a potential gain from using the proposed techniques. For the glioma example predictions from the model averaging methods did not differ substantially from predictions of conventional methods (BE_{BIC}, full model). Given the results, this is not surprising: the actual reduction of the Brier score of all model-based predictions is relatively small with a maximum of R^2 of 17% at day 310 and 30% at day 730. The sample size in relation to the 15 possible explanatory variables is reasonable (274 events) and 33% of the cases are censored. Also, the model including variables *age, surg1, mali1* und *Karn1* is the clear winner. With both methods, BMA and bootstrap MA, this model has a relative high "certainty" attached to it. The posterior model probability is 16% and the model inclusion frequency as estimated by the bootstrap resampling is 17%. The model with the second highest certainty attached to it includes one additional variable (*epil*) and has a posterior probability of 10% and an inclusion frequency of 10%. Repeating bootstrap MA with different cut off values for $h(x_j)$ always resulted in the same "winner model" (Table 2).

In the PBC example there is less information in the data for model selection compared to the glioma example. 60% of the cases are censored and the number of events (123) in relation to explanatory variables is small. The bootstrap inclusion frequencies and the posterior model probabilities do not point to an obvious winner model. For bootstrap MA the first 50% of model inclusion frequency here is shared among the first 6 models.

It is also interesting to see that in both examples the full model performs marginally best in terms of prediction. But compared to bootstrap MA, the full model uses the information from all explanatory variables.

These findings show that the described procedures do not necessarily improve predictions, but are useful for obtaining a better understanding of the data and judge the certainty attached to a selected model. We have looked at confidence intervals of estimated prognostic indices from three observations in the PBC example (Figure 3). It is likely that confidence intervals from BE_{BIC} selected models are too short, because uncertainty of the model selection process is not included in the variance estimation. Confidence intervals from the full model may be too wide, because many variables without influence are included and the variables are correlated. To make any further conclusions on the unbiasedness and accuracy of estimates and their variance, a simulation study assessing the coverage of confidence intervals is necessary. Model averaging may improve the coverage of confidence intervals as has been demonstrated by Buckland et al. (1997) for a simple survival model example. Whether this is the case in prognostic models like ours deserves further investigation.

In both model averaging approaches more or less arbitrary decisions on settings for the model space reduction are made. For bootstrap MA this is the cut-off value of $h(x_i)$ above which variables are included in the subset after the first screening step. In approx. BMA it is the setting of C in the Occam's window algorithm. Only models with a posterior probability greater $1/C$ are included. In the glioma example we have shown how the number of selected models changes depending on the cut-off value of bootstrap MA. To what extend results are sensitive to these settings deserves further investigation.

In bootstrap MA results may also be sensitive to the choice of the selection level applied in the backward elimination procedure. We have used the BIC criterion to make our procedure comparable to the approx. BMA procedure. Results of the bootstrap MA approach are also very much driven by whether the selection algorithm uses an exact likelihood ratio test or not. We use function stepAIC() (Venables and Ripley, 1997) as implemented in the R-software, which uses the exact likelihood ratio test.

In the glioma example the posterior effect probabilities $\mathrm{pr}(\beta \neq 0|D)$ from approx. BMA (when no Occam's window was applied) attached a higher "certainty of inclusion" to the variables *mali1* and *Karn1* than the inclusion probabilities of the bootstrap MA. This difference could be caused by the fact that the calculation of posterior probabilities does not take multicollinearity into account. In *approx. BMA* different models are considered to be equally likely a priori. If we have the case where two models M_1 and M_2 are almost equivalent, giving them equal weight a priori amounts to giving the single model of which M_1 and M_2 are slightly different versions twice as much actual prior weight. The relatively high posterior effect probabilities of *mali1* and *Karn1* in comparison to bootstrap MA could be caused by this. Both variables are correlated to their complimentary dummy variables *mali2* and *Karn2*, with a correlation coefficient of 0.67 und 0.35 respectively. Using the bootstrap for estimating model weights and inclusion probabilities avoids this problem, since only one of the two models is selected.

We have described existing methods for incorporating model selection uncertainty and have suggested a new method (bootstrap MA). These kind of approaches are becoming increasingly important with the increasing number and diversity of prognostic factors available in medicine.

Acknowledgements

This research was partly funded by the Deutsche Forschungsgemeinschaft. We also thank our colleagues at the fdm for helpful comments.

References

Altman, D. G. and Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* **8**, 771–783.

Blettner, M. and Sauerbrei, W. (1993). Influence of model-building strategies on the results of a case-control study. *Statistics in Medicine* **12**, 1325–1338.

Breiman, L. (1996a). Bagging predictors. *Machine Learning* **24**, 123–140.

Breiman, L. (1996b). Heuristics of instability and stabilisation in model selection. *The Annals of Statistics* **24**, 2350–2383.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Brooks, S. P., Catchpole, E. A. and Morgan, B. J. T. (2000). Bayesian animal survival estimation. *Statistical Science* **15(4)**, 357–376.

Buckland, S., Burnham, K. P. and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603–618.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference. A Practical Information-Theoretic Approach.* Springer Verlag, New York.

Candolo, C., Davison, A. and Demétrio, C. (2002). A note on model uncertainty in linear regression. *submitted* .

Chatfield, C. (1995). Model selection uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society , Series A* **158**, 419–466.

Chen, C. and George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine* **4**, 39–46.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Jour-*

nal of the Royal Statistical Society , Series B **34**, 187–220.

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

Dellaportas, P. and Forster, J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.

Dickson, E., Fleming, T., Wiesner, R., Baldus, W., Fleming, C. and Ludwig, J. (1985). Trial of penicillamine in advanced primary biliary cirrhosis. *New England Journal of Medicine* **312**, 1011–1015.

Dijkstra, T. K. and Veldkamp, J. H. (1994). Data-driven selection of regressors and the bootstrap. In Beckmann, M. and Krelle, W., editors, *Lecture Notes in Economics and Mathematical Systems. On Model Selection Uncertainty and its Statistical Implications. Proceedings of a Workshop, held in Groningen, The Netherlands, September 25–26*, pages 17–38. Springer-Verlag, Heidelberg.

Draper, D. (1995). Assessment and propagation of model selection uncertainty (with) discussion. *Journal of the Royal Statistical Society , Series B* **57**, 45–97.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* **37**, 36–48.

Fernández, C., Ley, E. and Steel, M. F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Department of Econometrics, Tilburg University, The Netherlands, available at http://www.stats.bris.ac.uk:81/ mamcfl.

Fleming, R. R. and Harrington, D. H. (1991). *Counting Processes and Sur-*

vival Analysis. Wiley, New York.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499–511.

George, E. I. and McCullagh, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods. Validation, Model Selection and Bootstrap.* Chapman & Hall, London.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* **14**, 382–417.

Holländer, N., Sauerbrei, W. and Schumacher, M. (2001). Estimating the effect of a prognostic or risk facto after selection of an 'optimal' cutpoint. Technical report, Freiburg Centre for Data Analysis and Modelling, University Freiburg im Breisgau.

Hook, E. B. and Regal, R. R. (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology* **145(12)**, 1138–1144.

Kass, R. and Raftery, A. (1995). Bayes factors. *JASA* **90**, 773–795.

King, R. and Brooks, S. P. (2001). On the Bayesian analysis of population size. *Biometrika* **88(2)**, 317–336.

Kuk, A. Y. C. (1984). All subsets regression in a proportional hazards model. *Biometrika* **71**, 587–592.

Lawless, J. and Singhal, K. (1978). Efficient screening on nonnormal regression models. *Biometrics* **34**, 318–327.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model selection uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535–1546.

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84**, 19–31.

Madigan, Y. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Markus, B., Dickson, E., Grambsch, P., Fleming, T., Mazzaferro, V., Klintmalm, G., Wiesner, R., Thiel, D. V. and STarzl, T. (1989). Efficacy of liver transplantation in patients with primary biliary cirrhosis. *New England Journal of Medicine* **320**, 1709–1713.

Raftery, A., Madigan, D. and Volinsky, C. (1996). *Bayesian Statistics 5*, chapter Accounting for model uncertainty in survival analysis improves predictive performance. Oxford University Press.

Raftery, A. E., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92(437)**, 179–191.

Sauerbrei, W. (1999). The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society, Series C* **48**, 313–329.

Sauerbrei, W. and Schumacher, M. (1992). A boostrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* **11**, 2093–2109.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Stanley, T. R. and Burnham, K. P. (1998). Information-theoretic model selection and model averaging for closed-population capture-recapture studies. *Biometrical Journal* **40(4)**, 475–494.

Steyerberg, E. W., Eijkemans, M. J. C. and Habbema, J. D. F. (2001). Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica* **55(1)**, 76–88.

Ulm, K., Schmoor, C., Sauerbrei, W., Kemmler, G., Aydemir, U., Müller, B. and Schmuacher, M. (1989). Strategien zur Auswertung einer Therapiestudie mit der überlebenszeit als Zielkriterium. *Biometrie und Informatik in Medizin und Biologie* **20**, 171–205.

Van Houwelingen, J. C. and LeCessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine* **9**, 1303–1325.

Venables, W. and Ripley, B. (1997). *Modern Applied Statistics with S-PLUS*. Springer Verlag, Heidelberg.

Viallefont, V., Raftery, A. E. and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* **20**, 3215–3230.

Volinsky, C. T., Madigan, D., Raftery, A. E. and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society, Series C* **46**, 433–448.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

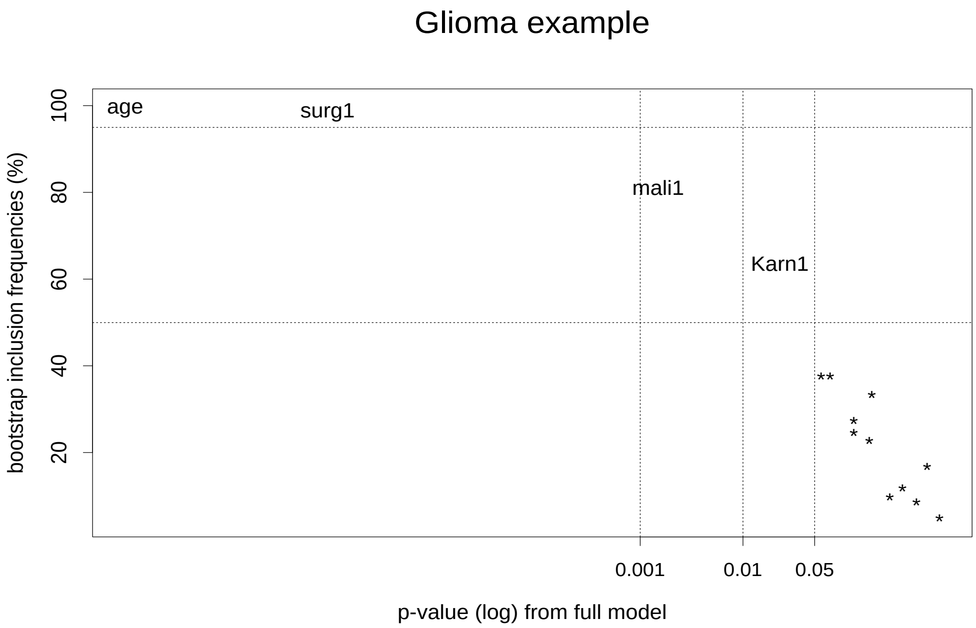

Figure 1. The bootstrap inclusion frequencies versus p-values from full model for the Glioma data.

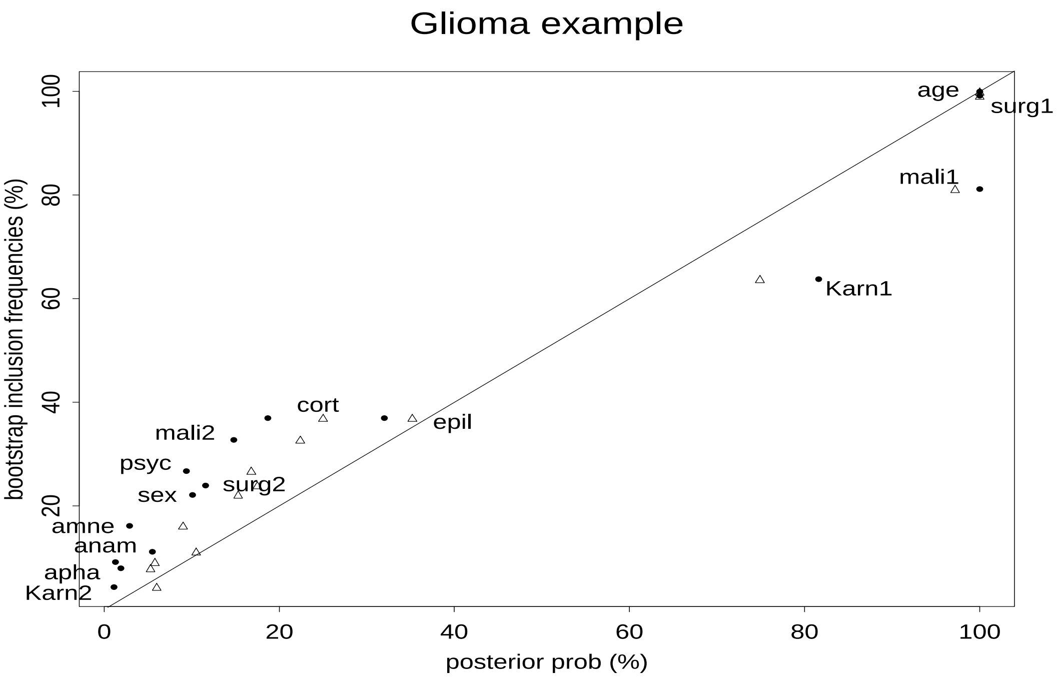

Figure 2. The bootstrap inclusion frequencies versus the posterior effect probabilities $\mathrm{pr}(\beta_l \neq 0|D)$ for the Glioma data marked with '*'. Posterior effect probabilities $\mathrm{pr}(\beta_l \neq 0|D)$ derived without Occam's window model space reduction are marked with '$\triangle$'.

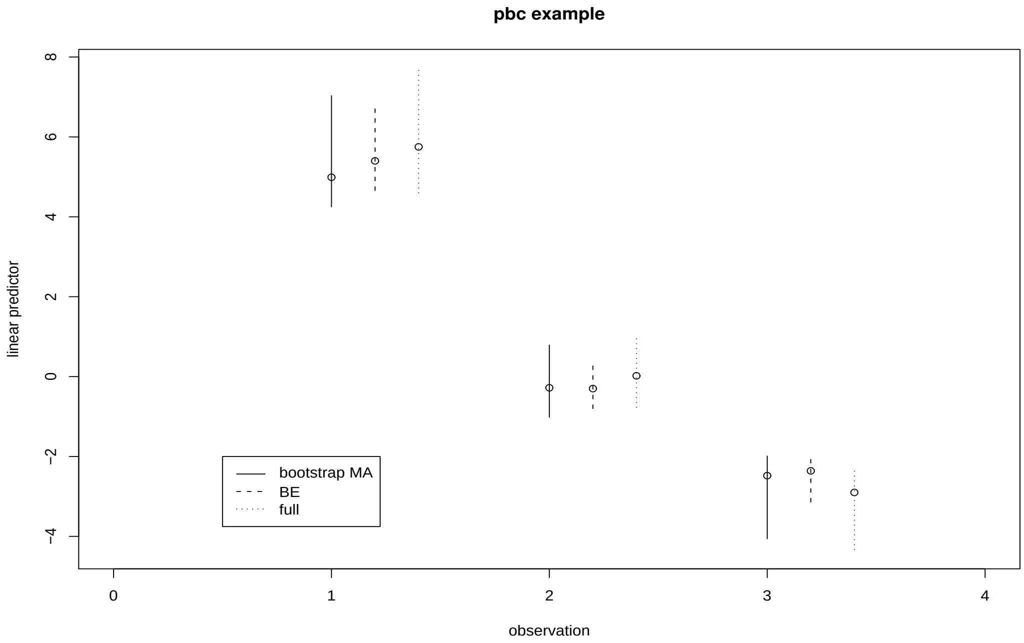

Figure 3. PBC example. Bootstrap confidence intervals with point estimates for three observations for bootstrap MA, BE and the full model.

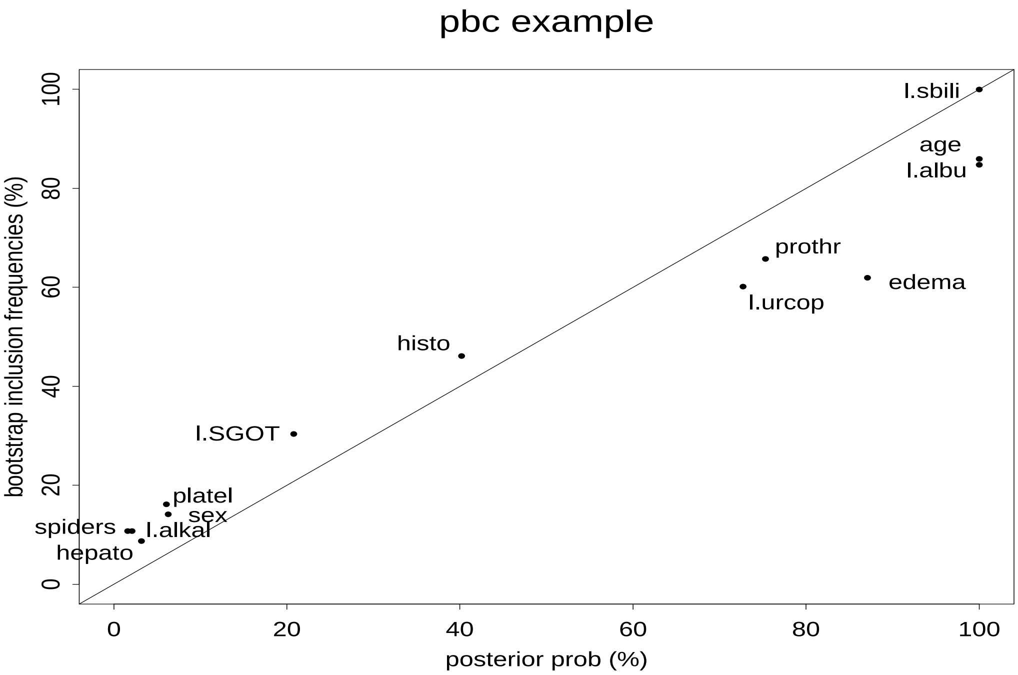

Figure 4. The bootstrap inclusion frequencies versus the posterior effect probabilities $\text{pr}(\beta_l \neq 0|D)$ for the PBC data.

Table 1

Brain tumour (Glioma) example: Results and selected explanatory variables of the full model and different selected models using backward elimination (BE) with different significance levels are shown.

	full model	BE(0.157)	BE(0.05)	BE(0.01)
log-likelihood	-1331.8	-1335.2	-1340.0	-1342.3
correlation in linear predictors with full model	-	0.99	0.95	0.94
sex	x			
anam: anamnese	x			
mali1: grade of malignancy 2-3	x	x	x	x
mali2: grade of malignancy 3	x	x		
age: age	x	x	x	x
Karn1: Karnofsky index >= 70	x	x	x	x
Karn2: Karnofsky index > 80	x			
surg1: surgical resection (0 = biopsy, 1 = partial or total)	x	x	x	x
surg2: surgical resection (total)	x	x		
conv: convulsiva	x			
cort: cortisone	x	x		
epil: epilepsy	x	x	x	
amne: amnesia	x			
psyc: organic psychosyndrome	x	x		
apha: aphasia	x			

Table 2

Glioma example: Results of the bootstrap MA approach. The inclusion frequencies $h(x_i)$ are based on the models selected in 500 bootstrap samples with BE_{BIC}. The 27 models are the result of step two based on 7 explanatory variables with $h(x_i) > 30\%$ selected in step one. The corresponding $h(M_j)$ resulting from different selection levels in step one are also given.

explan. variables	*step one*: $h(x_i)\%$	*step two*: 10 models with the highest $h(M_j)\%$ out of 27									
sex	22	-	-	-	-	-	-	-	-	-	-
anam	11	-	-	-	-	-	-	-	-	-	-
mali1*	81	x	x	x	x	x	x	-	x	-	-
mali2*	33	-	-	-	-	-	x	x	-	x	x
age*	100	x	x	x	x	x	x	x	x	x	x
Karn1*	64	x	x	x	-	-	x	x	-	x	x
Karn2	4	-	-	-	-	-	-	-	-	-	-
surg1*	99	x	x	x	x	x	x	x	x	x	x
surg2	24	-	-	-	-	-	-	-	-	-	-
conv	8	-	-	-	-	-	-	-	-	-	-
cort*	37	-	-	x	-	-	-	-	x	-	x
epil*	37	-	x	-	-	x	-	x	-	-	-
amne	16	-	-	-	-	-	-	-	-	-	-
psyc	27	-	-	-	-	-	-	-	-	-	-
apha	9	-	-	-	-	-	-	-	-	-	-
$h(M_j)\%$,	$h(x_i) > 0.3$	17	10	10	7	6	6	5	4	4	4
$h(M_j)\%$,	$h(x_i) > 0.4$	42	0	21	23	0	0	0	13	0	0
$h(M_j)\%$,	$h(x_i) > 0.2$	8	3	5	2	5	1	3	1	2	3
$h(M_j)\%$,	$h(x_i) > 0.1$	5	3	3	1	2	2	1	1	1	2
$h(M_j)\%$,	$h(x_i) > 0$	4	2	2	3	2	2	2	0	1	1

* Variables with $h(x_i) > 30\%$ and part of the subset of explanatory variables after *step one*.

Table 3

Glioma example: Results of approx BMA. The second column shows the posterior effect probabilities pr($\beta_l \neq 0|D$). In the right half models with the highest pr($M_j|D$) are shown.

explan. variables	$Pr(\beta_l \neq 0\|D)$ (%)	10 models with the highest pr($M_j\|D$) out of 41									
sex	10	-	-	-	-	-	-	-	x	-	x
anam	6	-	-	-	-	-	-	-	-	-	-
mali1	100	x	x	x	x	x	x	x	x	x	x
mali2	15	-	-	-	x	-	-	-	-	-	-
age	100	x	x	x	x	x	x	x	x	x	x
Karn1	82	x	x	x	x	-	x	-	x	x	x
Karn2	1	-	-	-	-	-	-	-	-	-	-
surg1	100	x	x	x	x	x	x	x	x	x	x
surg2	12	-	-	-	-	-	x	-	-	-	-
conv	2	-	-	-	-	-	-	-	-	-	-
cort	19	-	-	x	-	-	-	-	-	-	-
epil	32	-	x	-	-	-	-	x	-	-	x
amne	3	-	-	-	-	-	-	-	-	-	-
psyc	9	-	-	-	-	-	-	-	-	x	-
apha	1	-	-	-	-	-	-	-	-	-	-
pr($M_j\|D$)(%)	-	16	10	6	5	5	4	4	4	3	3

Table 4

Glioma example 20-fold cross-validation results: Partial predictive log-score (PPS) and Brier score estimated using 20-fold cross-validation.

	methods used for prediction				
criterion	Kaplan-Meier	*bootstrap MA*	*approx. BMA*	BE_{BIC}	full model
PPS	n.a.	-554.870	-553.621	-553.694	-549.675
$BS(t^* = 310)$	0.247	0.212	0.212	0.211	0.206
R^2 (%)	0	14.30	14.50	14.83	16.77
$BS(t^* = 730)$	0.144	0.104	0.103	0.101	0.105
R^2 (%)	0	27.87	28.76	30.07	27.55

Table 5

PBC example: Results of the bootstrap MA approach. The inclusion frequencies $h(x_i)$ are based on the models selected in 500 bootstrap samples with BE_{BIC}. The 65 models are the result of step two *based on 7 explanatory variables with $h(x_i) > 30\%$ selected in* step one, *of which 10 are shown. For these models the corresponding $h(M_j)$ resulting from different selection levels in* step one *are also given.*

explan. variables	*step one:* $h(x_i)$ %	*step two:* 10 models with the highest $h(M_j)$% out of 65									
age*†	86	x	x	x	x	x	x	x	x	x	x
sex	14	-	-	-	-	-	-	-	-	-	-
hepato: hepatomegaly	9	-	-	-	-	-	-	-	-	-	-
spiders	11	-	-	-	-	-	-	-	-	-	-
edema*†	62	x	-	-	x	x	x	x	-	x	x
l.sbili*†: log(bilirubin)	100	x	x	x	x	x	x	x	x	x	x
l.albu*†: log(albumin)	85	x	x	x	x	x	x	x	x	x	x
l.urcop†:log(urine copper)	60	x	x	-	-	-	-	x	x	x	x
l.alkal: log(alkaline phosphates)	11	-	-	-	-	-	-	-	-	-	-
l.SGOT: log(SGOT)	30	-	-	-	-	x	-	-	-	-	x
platel: platelets	16	-	-	-	-	-	-	-	-	-	-
prothr*†: prothrombin time	66	x	x	x	-	x	x	-	x	-	x
histo: histologic stage	46	-	-	-	x	-	-	x	x	-	-
$h(M_j)$ %	-	11	7	4	4	3	3	3	3	3	3

* Model selected by Fleming and Harrington (1991).

† Model selected using backward elimination with the BIC criterion (BE_{BIC}).

Table 6

PBC example: Results of approx BA applied. The second column shows the posterior effect probabilities pr($\beta_l \neq 0|D$). 10 of 33 models with the highest posterior probability are shown.

explan. variables	$Pr(\beta_l \neq 0\|D)$ %	10 models with the highest pr($M_j\|D$) out of 32									
age	100	x	x	x	x	x	x	x	x	x	x
sex	6	-	-	-	-	-	-	-	-	-	-
hepato	3	-	-	-	-	-	-	-	-	-	-
spiders	2	-	-	-	-	-	-	-	-	-	-
edema	87	x	x	x	x	-	x	x	x	x	x
l.sbili	100	x	x	x	x	x	x	x	x	x	x
l.albu	100	x	x	x	x	x	x	x	x	x	x
l.urcop	73	x	x	x	x	x	-	x	-	x	-
l.alkal	2	-	-	-	-	-	-	-	-	-	-
l.SGOT	21	-	-	-	-	-	-	x	x	x	x
platel	6	-	-	-	-	-	-	-	-	-	-
prothr	75	x	x	-	-	x	x	x	x	x	x
histo	40	-	x	x	-	-	-	-	x	x	-
pr($M_j\|D$)(%)	-	18	8	7	6	5	5	4	4	3	3

Table 7

PBC example: Partial predictive log-score (PPS) and Brier score at day 1000 and 3000 estimated using 20-fold cross-validation.

criterion	Kaplan-Meier	bootstrap MA	approx. BMA	BE_{BIC}	full model
	methods used for prediction				
PPS	n.a.	-183.1757	-189.8428	-184.4830	-180.9653
$BS(t^* = 1000)$	0.144	0.094	0.095	0.096	0.092
R^2 (%)	0	34.49	34.26	33.50	36.02
$BS(t^* = 3000)$	0.238	0.163	0.167	0.170	0.169
R^2 (%)	0	31.29	29.83	28.43	29.06