LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Riccardo De Bin, Silke Janitza, Willi Sauerbrei, Anne-Laure Boulesteix

Subsampling versus bootstrapping in resampling-based model selection for multivariable regression

Technical Report Number 171, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Subsampling versus bootstrapping in resampling-based model selection for multivariable regression

Riccardo De Bin*† Silke Janitza† Willi Sauerbrei‡
Anne-Laure Boulesteix†

Abstract

In the last few years, increasing attention has been devoted to the problem of the stability of multivariable regression models, understood as the resistance of the model to small changes in the data on which it has been fitted. Resampling techniques, mainly based on the bootstrap, have been developed to address this issue. In particular, the approaches based on the idea of "inclusion frequency" consider the repeated implementation of a variable selection procedure, for example backward elimination, on several bootstrap samples. The analysis of the variables selected in each iteration provides useful information on the model stability and on the variables' importance. Recent findings, nevertheless, show possible pitfalls in the use of the bootstrap, and alternatives such as subsampling have started to be taken into consideration in the literature. Based on model selection frequencies and variable inclusion frequencies, we aim to empirically compare these two different resampling techniques, investigating the effect of their use in a model selection procedure for multivariable regression. We conduct our investigations by analyzing two real data examples and by performing a simulation study. Our results reveal some advantages in using a subsampling technique rather than the bootstrap in this context.

Keywords: bootstrap; model selection; model stability; subsampling.

*corresponding author: debin@ibe.med.uni-muenchen.de

†Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Germany

‡Department of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

1 Introduction

In statistical practice, the analyst often faces the problem of choosing which variables should be included in the final model from the numerous potentially important variables collected in the study. Often, variable selection procedures such as backward elimination, stepwise regression or all-subset approaches are used, although it is well known that they have several shortcomings, such as high instability and a possible bias in parameter estimates (see e.g. Copas and Long, 1991; Miller, 2002). In this context, with "instability" we are referring to the sensitivity of a model to small changes in the data, which may modify the set of selected variables (Gifi, 1990). The selection criterion, usually the significance level related to a test on the parameters or an information criterion such as the AIC (Akaike, 1973) or the BIC (Schwarz, 1978), plays a central role. For the sake of various methodological issues it is important to distinguish between models for prediction and for explanation (Sauerbrei et al., 2014). Here we are mainly interested in the latter. In order to investigate model stability and to provide better insight into the variable selection procedure, methods based on bootstrap resampling have been presented in the literature (see, for example, Gong, 1982; Chen and George, 1985; Altman and Andersen, 1989; Sauerbrei and Schumacher, 1992). By using the bootstrap technique (Efron, 1979), it is possible to generate pseudo-samples which can be seen as perturbed versions of the original data. The possible differences among the models obtained by applying a stepwise selection procedure to the different pseudo-samples provide useful information on the stability of model selection. Please note that any selection procedure can be used within this framework: for example, Sauerbrei and Schumacher (1992) perform this analysis using backward elimination. In their paper, they focus on the frequency of inclusion of the variables in models derived from the pseudo-samples, which allows a better feeling for the final model, the importance of the different variables and their interrelationship.

Recent studies, however, have highlighted some issues related to the use of bootstrap pseudo-samples, in particular the tendency to select too many variables (see Janitza et al., 2014, for an overview). Alternatives such as subsampling (Hartigan, 1969) have been taken into consideration, and profitably applied in the context of model stability (Meinshausen and Bühlmann, 2006, 2010). The aim of this paper is to provide a detailed comparison between bootstrapping and subsampling in the context of model selection for multivariable regression based on inclusion frequencies, as first proposed by Gong (1982) and later extended by Sauerbrei and Schumacher (1992) to take into considerations the interrelationships. In particular, the use of subsampling

in this framework has not been extensively investigated and contrasted with the original bootstrap approach. We start our investigation from the same dataset used in Sauerbrei and Schumacher (1992), comparing the variable inclusion frequencies obtained for the different resampling approaches and some characteristics of the selected models. We extend the analysis by considering a second dataset, which is used both as an additional descriptive example and as the basis for a simulation study. In contrast to the former dataset, which contains survival data, the latter has a normally distributed response variable. Moreover, the analysis of simulated data drawn from a known distribution allows a suitable quantitative comparison of the performances of bootstrapping and subsampling in terms of identification of the relevant variables.

The paper is structured as follows: in Section 2 we briefly describe the two datasets, named "Glioma data" and "Ozone data", and we present the simulation design. The methods are described in Section 3: we introduce the concept of inclusion frequency and the statistical tools used in our analysis, including the model selection procedure and the resampling approaches. The results obtained from the two real datasets and from the simulation study are analyzed and reported in Section 4. Finally, some remarks and conclusions are provided in Section 5.

2 Data and simulation design

2.1 Glioma data

The Glioma dataset includes 411 patients with malignant glioma (an aggressive type of brain tumor) who took part in a randomized controlled trial for comparing two kinds of chemotherapy. Of these 411 patients, 276 (67.2%) died. In addition to the form of chemotherapy, 12 variables are considered, including sex, age, time from first symptoms to diagnosis (binary: either long or short) and information on health-related conditions (malignancy grade, Karnofsky index, presence/absence of convulsions, epilepsy, amnesia, organic psycho-syndrome, aphasia) and on treatment history (resection type, use of cortisone). Three variables that were originally measured on 3-value ordinal scales (malignancy grade, Karnofsky index, resection type) are coded by two dummy variables according to the split-coding scheme (see, e.g., Tutz, 2012, page 17). More details on the Glioma data can be found in Ulm et al. (1989) and Sauerbrei and Schumacher (1992). Please note that in these two studies 36 further observations were available and different sample sizes were used in the paper. Treatment will not be considered here. The data

used have no missing data and are publicly available at http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book.

Table A.1 in the Web Appendix shows the Cox model fitted by including all the available variables. Significant associations are present for the variables *age* (hazard ratio (HR): 1.04, $p < 0.0001$), *gradd1* (HR: 2.22, $p = 0.0015$), *kard1* (HR: 0.73, $p = 0.0230$) and *surgd1* (HR: 0.35, $p < 0.0001$). From a descriptive point of view, besides the high positive correlations (Spearman rank) between the dummy variables related to the same categorical variable (in particular between *gradd1* and *gradd2*, $\rho = 0.672$), we note a moderate positive correlation between the variables *amnesia* and *ops* (0.343), *convul* and *epi* (0.265) and between *age* and *gradd2* (0.215). Moreover, the variables *time* and *gradd2* show a non-negligible negative correlation (-0.233). All the other correlations are, in absolute value, below 0.200.

2.2 Ozone data

In a study by Ihorst et al. (2004) the long- and medium-term effects of ozone on the forced vital capacity and on the forced expiratory volume of 2153 school children are investigated. Forced vital capacity is the total amount of exhaled breath, and higher values indicate a better functionality of the lung. A well-defined subset of the data is used in Buchholz et al. (2008) in a paper on a two-step bootstrap model averaging approach and recently in a study on model stability (Sauerbrei et al., 2014). We use the same data, which feature 496 children and 24 variables potentially affecting the (continuous) outcome, "forced vital capacity in autumn 1997". For more details see Ihorst et al. (2004) and Buchholz et al. (2008). In the Web Appendix (Table A.2) we present the full model, which includes all 24 variables. For one variable (*f03h24*) a fractional polynomial of degree 2 was significantly better than the linear function, but the functional form was not much different from linearity (Royston and Sauerbrei, 2008). As in the aforementioned papers, we consider the linearity assumption acceptable for all variables.

From the analysis of the full model, we note that variables *sex*, *flgew* and *flgross* yield highly significant influence ($p < 0.0001$). Significant associations are also present for *hochozon* ($p = 0.0120$), *fnoh24* ($p = 0.0047$), and for *fspfei* ($p = 0.0283$). A moderate or strong Spearman correlation is present between pairs of variables *fsatem* and *fspei* (correlation: 0.553), *flgross* and *flgew* (0.716) and *fo3h24* and *fteh24* (0.860). There are strong positive correlations (up to 0.842) among different allergies (i.e., variables *adheu*, *fmilb*, *ftier*, *fpoll* *fspt*), and among coughing and breathing problems (*fsnight*, *fshlauf*, *fspfei*, *fsatem*). In summary, a relatively complex structure.

2.3 Simulated data

The analyses performed on the two real data examples should be considered only descriptive, because the true model is unknown. In particular, we do not know which variables are actually related to the outcome and which are pure noise, i.e. which variables should be selected for and which should be excluded from the final model. This prevents us from properly evaluating the quality of the inclusion frequencies for the available variables. To tackle this issue, we perform a simulation study, which allows for a more objective assessment of the inclusion frequencies obtained for the bootstrap and for subsampling. In order to attain a scenario which reflects realistic associations between explanatory variables and the response, we keep the data structure of the Ozone data. The idea is to generate a new outcome that depends only on some selected variables, in order to have a set of known relevant variables and a set of noise variables. We proceed as follows:

- we studentize the continuous variables of the Ozone data, to have comparable effects;
- we fit a full regression model (containing both the studentized and the binary variables);
- based on the estimates of the regression coefficients we define:
  - the variables with high effect, i.e. those with an estimate in absolute value larger than 0.15: here *flgross* and *sex*;
  - the variables with low effect, i.e. those with an estimate, in absolute value, between 0.06 and 0.15: here *flgew*, *hochozon*, *fsatem* and *fspfei*;
  - the noise variables, i.e. those with effect in absolute value smaller than 0.06;
- we generate a further noise variable from a standard Gaussian distribution, uncorrelated to all other variables.

Please note that the first six variables are related to the response, while the other nineteen are not. We use these six variables to generate 1,000 artificial outcomes, drawing from a Gaussian distribution with mean $2.5 + 0.2flgross + 0.1flgew - 0.2sex - 0.1hochozon + 0.1fsatem + 0.1fspfei$ and standard deviation 3.5. Both the values of the intercept and of the standard deviation are approximations of their estimates in the original data. Note that, in order to preserve the data structure, the signs of the regression coefficients are kept as they were in the original estimates. For presentation

clarity, we reorder and rename the variables. The true mean, then, is codified as $2.5 + 0.2x_1 - 0.2x_2 + 0.1x_3 - 0.1x_4 + 0.1x_5 + 0.1x_6$. Combining the new response vectors with the original explanatory variables, we finally obtain 1,000 artificial datasets, for which we know the true model. Note that the average R^2 of the full models fit on the artificial datasets is 0.476, smaller than the R^2 of the full model fit on the original data (0.648). It is worth noting that x_2, x_4, x_5 and x_6 are binary: the latter two, in particular, are strongly unbalanced, containing only 26 (5.34% of the total) non-zero values. This characteristic affects the variability of their regression coefficient, which in the simulated data may be far from the nominal 0.1 in some replications (see Table A.3 in the Appendix). Note that, due to the correlation structure inherited from the Ozone data, the variables x_5 and x_6 are strongly correlated with each other ($\rho = 0.553$), and with other variables (e.g., both have a correlation larger than 1/3 with x_{17} and x_{24}). Noticeable correlation involving at least one relevant variable is also present between pairs x_1 and x_3 ($\rho = 0.716$) and x_4 and x_9 ($\rho = -0.519$).

Summarizing, the artificial datasets have the following characteristics:

- 25 explanatory variables, of which 2 have a strong effect on the response, 4 have a weak effect and 19 are noise variables (no effect);
- the explanatory variables are correlated to each other as in the Ozone data, but the last variable that is totally uncorrelated with the others;
- the sample size is 496, as in the original Ozone data;
- the X matrix (values of the explanatory variables) is the same in each artificial dataset;
- the Gaussian distributed response vector is different for each artificial dataset.

For each of these 1,000 datasets, we perform our analyses by generating B = 1,000 pseudo-samples with each of the different resampling techniques and use backward elimination as variable selection strategy. Therefore, the results for the resampling approaches are based on 1,000,000 replications.

3 Methods

3.1 Variable selection

Variable selection is a crucial part of the model building process. A good model should include as few variables as possible, in order to avoid overfitting

and favor its interpretability, but without discarding any relevant variables, in order to not face the serious problem of underfitting (Sauerbrei et al., 2014). The literature on the variable selection issue is boundless and it is outside the scope of this paper to provide an overview. Here we use backward elimination, without reinclusion of previously excluded variables. Arguments in favor of backward elimination can be found in Mantel (1970), while for a brief comparison of the stepwise approaches for variable selection we refer the reader to Royston and Sauerbrei (2008, Section 2.7). More precisely, in our analysis we apply a numerically stable version of fast backward selection based on an algorithm described in Lawless and Singhal (1978). The method selects the variables through an approximation of the Wald statistic, computed using the conditional (restricted) maximum likelihood estimates under the hypothesis of multivariate normality; it is implemented in the R package *rms* (Harrell, 2013).

A key aspect of the variable selection procedure is the choice of the inclusion criterion. Although several alternatives are possible, for example choosing the total number of variables to include in the model by cross-validation (as, for example, in De Bin et al., 2014), the most common approach is to consider a significance level α for a statistical test on the regression coefficients, or a related quantity, for example an information criterion such as the AIC or the BIC. In the backward elimination procedure, at each step a variable, generally that corresponding to the highest p-value, is removed from the model if its p-value is larger than α. The procedure ends when all the p-values associated with the significance tests are smaller than α. The choice of the significance level greatly impacts the stability and the complexity of the final model (Royston and Sauerbrei, 2008). In this paper we use three different significance levels (namely 0.05, 0.10 and 0.157, with the last related to the Akaike information criterion), but throughout the paper we will only report the results for $\alpha = 0.05$. For $\alpha = 0.10$ and $\alpha = 0.157$ we give some results in the Web Appendix.

3.2 Resampling

3.2.1 Inclusion frequencies and models selected

The use of resampling techniques in the model building process is related to the stability issues mentioned in the introduction. The idea is to generate several pseudo-samples containing small perturbations of the original data. For each pseudo-sample, a model selection procedure, in our case backward selection, is then applied, leading to different models due to the small changes in the data. By analyzing the inclusion/exclusion of the variables in these

models, we can distinguish between the relevant variables, i.e. those useful for explaining the outcome, and the noise variables, which are not associated with the outcome. We expect, indeed, that the relevant variables are always (or almost always) included in the models, while the others are selected in only few cases, corresponding to particular configurations of the pseudo-sample. We define the proportion of times in which a variable is included in the models as the "inclusion frequency", which can range from 0 (never included) to 1 (always included). In the ideal case, the relevant variables have inclusion frequencies equal to 1 and the others 0, or, in terms of models, the same model (the one including only the relevant variables) is selected every time. Unfortunately, this does not occur in reality. Firstly, some variables have a "weak" effect and their inclusion may depend on chance: in earlier analyses inclusion frequencies between about 20% and 60% have often been observed (Sauerbrei and Schumacher, 1992; Buchholz et al., 2008). Secondly, variables without any effect are sometimes included because of type I errors. More critically, in the case of two highly correlated variables, it may happen that they are alternately selected for the models. For example, if both are relevant, we may obtain, instead of a theoretical value of 1, an inclusion frequency around 0.50 for both. Details on this issue can be found in Sauerbrei and Schumacher (1992). In real data, this "alternate selection" issue is even more relevant, due to complex and higher dimensional relationships (i.e., three-way correlation) among the variables.

3.2.2 Resampling strategies

In order to generate the pseudo-samples for our analyses, we need to chose a resampling technique. The literature provides several options: we mentioned in the introduction that the early studies on model building based on the variable inclusion frequencies (Gong, 1982; Chen and George, 1985; Altman and Andersen, 1989; Sauerbrei and Schumacher, 1992) use the bootstrap approach introduced by Efron (1979). This is likely the most popular resampling technique in statistical practice. It consists of drawing with replacement n observations from the original data, where n denotes the sample size of the original data. Sampling with replacement allows the possible replication of some observations, forcing the exclusion of others: on average, in a bootstrap pseudo-sample there are $0.632n$ unique observations. The approach just described is also known as nonparametric bootstrap, in order to distinguish it from the parametric bootstrap. In this latter approach, the pseudo-samples are instead randomly generated from a parametric model, in which the parameters are estimated using the original data. Several other modifications, such as the wild bootstrap (Wu, 1986; Mammen, 1993) or

the sequential bootstrap (Rao et al., 1997) are available in the literature (see,e.g., Chernick, 2011). In this paper, however, we focus on the classical nonparametric bootstrap.

The asymptotic properties of bootstrap procedures have been studied deeply in recent years, starting from Bickel and Freedman (1981), as have counterexamples where their consistency is not achieved (see, e.g., Mammen, 1992; Bickel et al., 1997). For this reason, alternative methods have been taken into consideration, especially those based on resampling fewer than n observations (Bickel et al., 1997). Among these alternatives, the subsampling technique (also known as delete-d jackknife, see Wu, 1986) has been intensively investigated (Shao and Wu, 1989; Politis and Romano, 1994; Politis et al., 1999), showing its asymptotic consistency even in cases where the classical bootstrap fails (Davison et al., 2003; Chernick, 2011). Subsampling consists of generating pseudo-samples by drawing without replacement $m < n$ observations from the original data. In this paper, we choose m equal to $[0.632n]$ (i.e., the nearest integer to $0.632n$), in order to have a number of observations in the subsample equal to the average number of unique observations in the bootstrap pseudo-samples. The optimal choice of this parameter is delicate (Davison et al., 2003), and it is not treated here. For more information on this specific issue, see Bickel and Sakov (2008).

In order to have a comparison between bootstrapping and subsampling based on the same sample size, in this paper we also explore the m out of n bootstrap, which consists of drawing with replacement m observations from the original data. As with subsampling, we set $m = [0.632n]$. Focusing on the m out of n version of the bootstrap, we can avoid possible differences caused by the different powers of the tests on the significance of the regression coefficients. The power of a test computed on a single pseudo-sample, indeed, is strictly related to the number of pseudo-observations: conversely to the classical bootstrap, subsampling and m out of n bootstrap here share the same sample size, and are thus directly comparable in our study. The difference between these two approaches lies only in the presence or absence of duplicated observations in the pseudo-samples and therefore their properties are often discussed together (see, e.g., Bickel and Sakov, 2008; del Barrio et al., 2009). In particular, the latter study highlights some relevant characteristics of the procedures based on resampling less than n observations; for example, their possible robustness against outliers. For a recent review on the properties of the bootstrap, subsampling and m out of n bootstrap, refer to Chernick (2011) and Mammen and Nandi (2012).

To summarize, in our study we use the following resampling schemes:

- *classical bootstrap:* n observations drawn from the original data with

replacement;

- *m out of n bootstrap:* $m = [0.632n]$ observations drawn from the original data with replacement;
- *subsampling:* $m = [0.632n]$ observations drawn from the original data without replacement.

Hereafter, we denote the three approaches by *bootstrap(n)*, *bootstrap(m)* and *subsample(m)*, respectively.

When dealing with time-to-event data, as, for example, in the Glioma dataset, some complications occur due the presence of censored observations. By directly applying the resampling technique, indeed, we obtain pseudo-samples with different effective sizes (number of events). In order to tackle this problem, it would be possible to sample events and censored observations separately. However, we do not see the randomness of the effective sample size as critical for our purposes, and therefore we perform the simpler alternative. We note that the number of events in our sample is relatively large, and therefore we do not face the issue of obtaining pseudo-samples with only a small number of events. In any case, at least for bootstrap(n), studies such as Burr (1994) seem to suggest that for censored data more elaborate schemes (e.g., in our case, stratified resampling) do not necessarily outperform the simpler ones. Other reasons not to sample events and censored observations separately, especially under the proportional hazards assumption, can be found in Zelterman et al. (1996).

3.2.3 Criteria to compare results

Our comparison focuses on the different variable inclusion frequencies obtained for bootstrap(n), bootstrap(m) and subsample(m). Dealing with real data, i.e. ignoring the true model, we can provide only descriptive analyses. Nevertheless, we know that the inclusion frequencies should allow us to recognize the importance of the variables, and to include in the final model only the relevant ones. For this reason, in the real data examples we enhance our investigation with a description of the effects of the different variable inclusion frequencies on the models: we investigate the average number of variables in the models, the number of unique models selected and the model selection frequencies. A small study on the prediction accuracy of the selected models allows us to draw some heuristic conclusions on their prediction performances and, indirectly, on the appropriateness of the variable inclusion frequencies derived using the three different resampling approaches. To compute the prediction ability, we follow a cross-validation procedure: we split the data

into 10 folds and we predict in turn the values of one fold with the model obtained by applying backward selection on a pseudo-sample generated from the observations belonging to the other 9 folds. We then measure the discrepancies between the observed and the predicted values with a quadratic score. For the Glioma dataset, in which we deal with time-to-event data, this is performed using the integrated Brier score (IBS) (Graf et al., 1999), which is the area under the prediction error curves based on the difference between the predicted survival probability and the true survival status (e.g., alive/dead) of each observation at time t. To avoid the issues related to small numbers of patients at risk for large values of t, we compute the IBS only up to the median follow-up calculated from the original data. For the Ozone datatset, in which we apply classical linear regression, instead, the prediction ability of the models is computed using the sum of squared residuals. For both datasets, we repeat the cross-validation 10,000 times for each resampling approach in order to reduce the influence of a specific split, and we provide the average value of the prediction accuracy measure.

With the results obtained in the simulation study, instead, we can directly assess the quality of the inclusion frequencies obtained via bootstrap(n), bootstrap(m) and subsample(m). The knowledge of the true model, indeed, allows us to compare the values of the observed inclusion frequencies with the expected ones (close to 1 for the high effect variables, close to 0.05 for the noise variables, between these two values for these with low effect). Moreover, we can compute a measure which quantifies how well the inclusion frequencies by an arbitrary resampling approach can be used to discriminate between the relevant and the noise variables. To do this, we compute the relative frequency of noise variables with lower inclusion frequencies than that of a relevant variable, in turn for all pairs of relevant $(x_1, \ldots, x_6)$ and noise $(x_7, \ldots, x_{25})$ variables. Then, we average these values, obtaining an estimate of the area under the curve (AUC). Please note that the AUC becomes 1 for a perfect discrimination and 0.5 for a discrimination which is not better than random. We compute this measure for the inclusion frequencies obtained with all the three resampling approaches and we compare the results in terms of distribution of the AUC over the 1,000 simulated datasets.

Note that with this approach we do not evaluate the appropriateness of the models with respect to the inclusion of all relevant variables since we completely ignore the models but only consider the variable inclusion frequencies. Moreover, we note that with our approach we give the same importance to the inclusion of the relevant variables and the exclusion of the noise variables. This is a result of our focus on explanatory models, otherwise different weighting schemes would be preferable.

4 Results

4.1 Results for the real data examples

4.1.1 Variable inclusion frequencies

Figures 1 and 2 show the inclusion frequencies for the variables of the Glioma and Ozone data, respectively. In both datasets we identify three variables with high inclusion frequencies, namely *gradd1*, *age* and *surgd1* (Glioma data) and *sex*, *flgross* and *flgew* (Ozone data): these variables seem to have strong effects, and for this reason we will refer to them as "core variables". With regard to the Glioma data (Figure 1), we note the ability of subsample(m) to achieve large inclusion frequencies for the three core variables, with values comparable to those obtained for bootstrap(n), despite the lower power of the significance tests due to $m < n$. The values obtained for bootstrap(m), instead, are smaller, likely indicating poor performance. For the Ozone data (Figure 2), this situation is less pronounced, due to the very strong effects of the three core variables, whose inclusion frequencies are close to 1 for all the three resampling approaches.

If we consider the least included variables, instead, subsample(m) provides smaller inclusion frequencies than the two bootstrap approaches for both the Glioma and the Ozone data. In the former dataset (Figure 1), this is evident for *time*, *convul*, *amnesia* and *aph*. For *kard2*, the inclusion frequency obtained for subsample(m) seems to be even too small. It is worth noting, indeed, that for an uncorrelated noise variable, we expect an inclusion frequency equal to the value of the type I error, here 0.05. The inclusion frequency of *kard2* is in fact influenced by the high correlation between this variable and *kard1*: the inclusion frequencies of both variables are probably underestimated due to the "alternate selection" problem described in Section 3.2.1. Although less pronounced, the same phenomenon seems to occur also between *convul* and *epi* and between *gradd1* and *gradd2*. It is worth noting that these correlation issues would have been completely missed had the backward selection been simply applied to the original data, without analyzing the variable inclusion frequencies.

Several variables have inclusion frequencies far from both 0.05 and 1. These variables may have low effect or their inclusion frequencies may be influenced by the inclusion frequencies of other variables. As per the strategies described in Sauerbrei and Schumacher (1992), further investigations are necessary to decide whether these variables should be included or excluded from the final model. Interestingly, *kard1*'s inclusion frequency for subsample(m), as well as those for the three core variables, is higher than that for

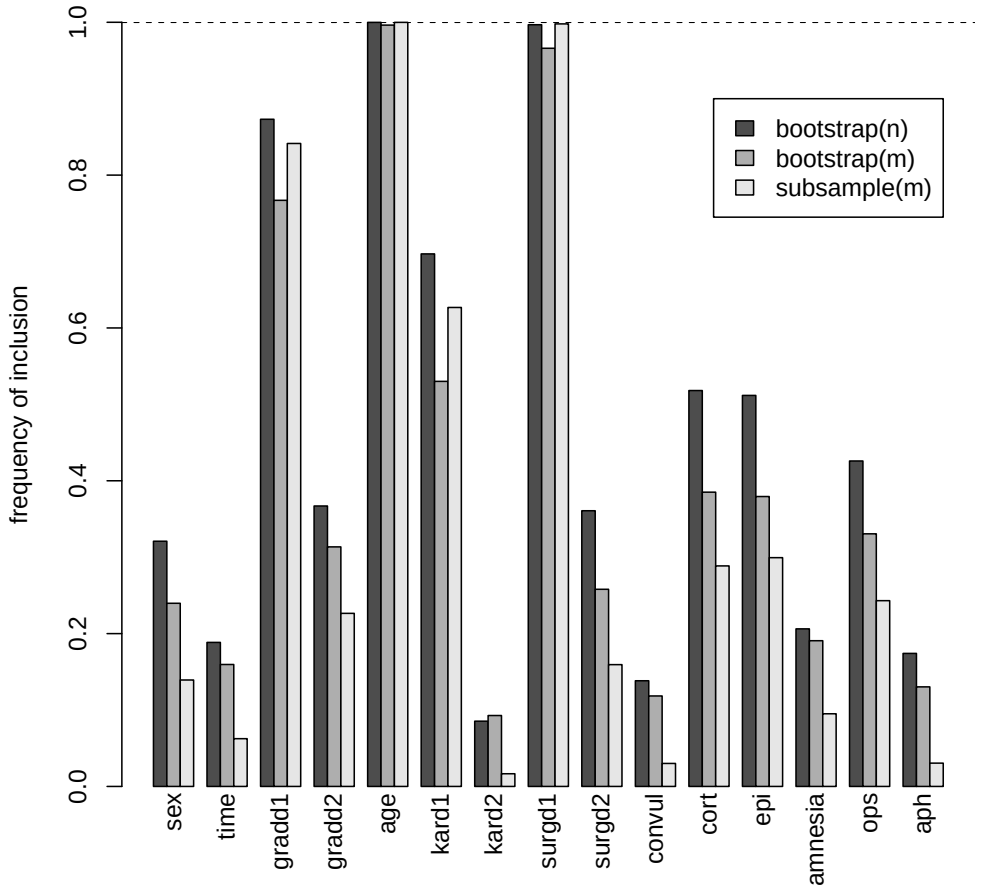

Figure 1: Glioma data: inclusion frequencies, based on 10,000 pseudo-samples, for all the 15 available variables. The results refer to the case $\alpha = 0.05$.

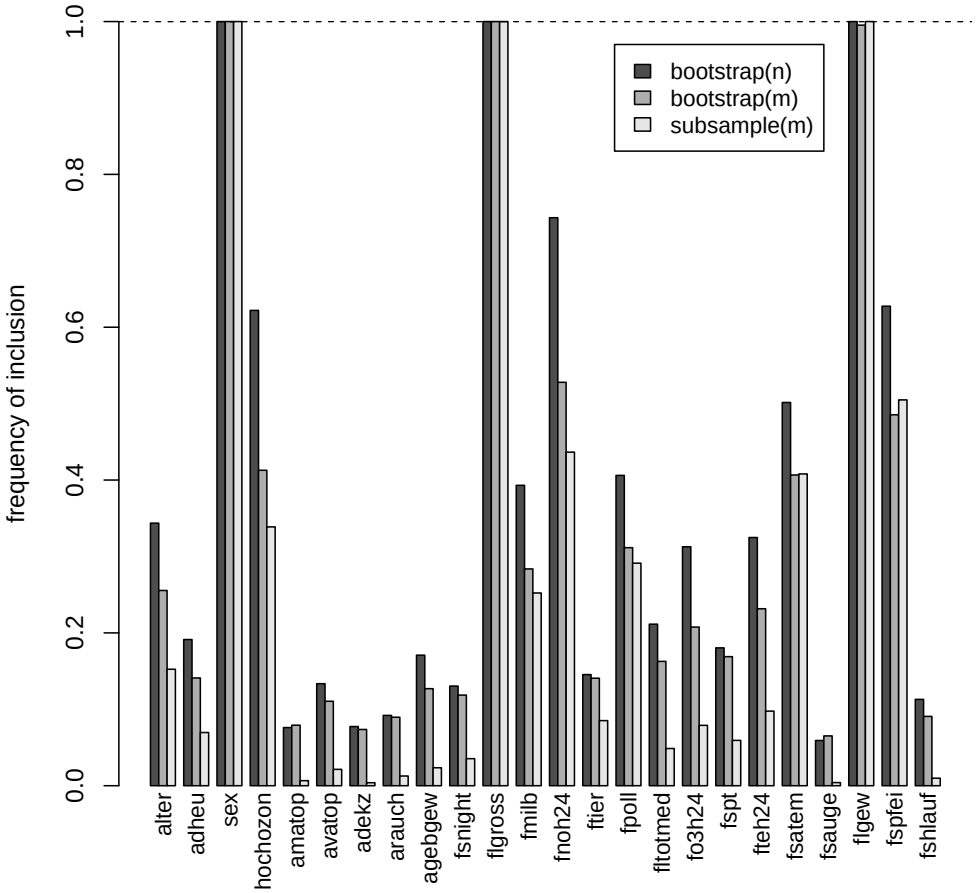

Figure 2: Ozone data: inclusion frequencies, based on 10,000 pseudo-samples, for all the 24 available variables. The results refer to the case $\alpha = 0.05$.

bootstrap(m), while for all the other variables the opposite is true. It seems that subsample(m) provides results in which the variables with high/medium effect and the variables with low/no effect are more distinctly separated than in the results for bootstrap(n) and bootstrap(m). It is worth noting, however, that the three resampling methods rank the variables in the same order. Similar observations apply to the Ozone data.

For the Glioma data, we also performed the analyses using 0.10 and 0.157 as significance levels, obtaining very similar results (see Table A.4 in the Appendix).

4.1.2 Number of unique models

The characteristics of the variable inclusion frequencies for the different resampling techniques described above have an effect on the number of different models selected in the 10,000 iterations: in the Ozone data, using subsample(m) we select 580 unique models, versus 1,787 and 1,829 for bootstrap(n) and bootstrap(m), respectively. From a practical point of view, this ability of subsample(m) to focus on few different models can be advantageous. The results for different significant levels (Table A.5 in the Web Appendix) confirm that this property is not related to the specific significance level used. We obtain an even more extreme result for the Ozone data, probably due to the larger number of variables available in this dataset: subsample(m) leads to 927 unique models, versus 4,650 for bootstrap(m) and 5,154 for bootstrap(n).

4.1.3 Model selection frequencies

In addition to the number of unique models, we compute the selection frequencies for the selected models for the three resampling approaches. The results for the Glioma data are reported in Table 1. We note the ability of subsample(m) to highlight a distinctly favorite model, namely that with the three core variables and *kard1*. Noticeably, it has a selection frequency almost 4 times larger than the second top ranked model (which includes the three core variables and *epi*). We see a similar situation for bootstrap(m) as well, but not as extreme: the selection frequency for the top ranked model is definitely smaller (326 versus 1,615 for subsample(m)) and the second top ranked model (in this case with *sex* instead of *epi*) is less separated. The results for bootstrap(n), instead, do not allow us to recognize a favored model, because the top ten models all have similar (and relatively small) selection frequencies. Finally, it is worth noting the higher selection frequencies of the sparsest models for subsample(m): for example, the model with only the core variables (denoted by "basic" in Table 1), is the third most selected (with

Table 1: Glioma data: selection frequencies of the 10 top ranked models for bootstrap(n), bootstrap(m) and subsample(m), based on 10,000 pseudo-samples for $\alpha = 0.05$ and presented in decreasing sum of the three selection frequencies.

	bootstrap(n)		bootstrap(m)		subsample(m)	
model	rank	freq.	rank	freq.	rank	freq.
basic+kard1	2	124	1	326	1	1615
basic+kard1+epi	8	93	7	128	2	417
basic+kard1+surgd2	6	103	3	163	4	352
basic+kard1+sex	3	108	2	187	6	290
basic	140	15	8	123	3	398
basic+kard1+cort	5	106	4	148	5	298
basic+kard1+sex+epi	1	156	6	140	9	225
basic+cort+ops	22	62	4	148	7	264
basic+epi	54	33	12	104	8	242
basic+ops	101	20	9	121	12	189
basic*	717	2	10	117	10	205
basic+kard1+cort+ops	7	97	15	93	15	134
basic+gradd2+kard1+cort	8	93	43	40	23	84
basic+gradd2+kard1+cort+ops	3	108	55	33	55	35
basic+kard1+surgd2+sex+epi	10	89	52	35	67	27

basic=intercept+gradd1+age+surgd1; basic*=intercept+gradd2+age+surgd1

selection frequency 398), while for bootstrap(m) it is the eighth (123) and only the 140th for bootstrap(n), selected only 15 times. This result is confirmed by the analysis of the structure of the models: we note the prevalence of the structure "3 core + 2 additional" variables for models selected for subsample(m), "3 core + 3 additional" for bootstrap(m) and "3 core + 4 additional" for bootstrap(n). This is even more clear if we consider *gradd2* and *surgd2* exchangeable with *gradd1* and *surgd1*, respectively (they are dummy codifications of the same original categorical variables). See Table A.7 for the details.

The results for the Ozone data are reported in Table A.8 in the Web Appendix. In this case the evidence for a favored model is less strong, probably due to the high correlation among the variables, which makes them interchangeable and, as a consequence, several models are competitive. Nevertheless, in this case as well the selection frequencies of the ten top models are larger for subsample(m) than for bootstrap(m) and bootstrap(n). Table A.9 in the Web Appendix reports the analysis of the models' structures for

this example as well.

4.1.4 Average number of variables in the models

We saw in the previous subsection that subsample(m) tends to favor sparser models. As a consequence, the average number of variables included in the selected models is smaller for subsample(m) (5.057 for the Glioma data, 5.941 for the Ozone data) than for bootstrap(m) (5.856 and 7.486) and bootstrap(n) (6.864 and 8.856). We obtain again a very similar tendency for $\alpha = 0.10$ and for $\alpha = 0.157$ (see Table A.6 in the Web Appendix).

4.1.5 Prediction accuracy

We stated before that in the real examples we are ignorant of the true model and, therefore, we do not know if the exclusion of the least included variables, and the consequent disfavor of the more complex models, is positive. Before considering the simulated data, we compare the performances of the three resampling approaches by looking at the cross-validated prediction accuracy of the selected models.

For the Glioma data, we obtained an estimate of the integrated Brier score of 0.157 for bootstrap(n), 0.160 for the bootstrap(m) and 0.156 for subsample(m), where larger IBS values correspond to worse prediction ability. For the Ozone data, the sums of squared residuals are 2.411 for bootstrap(n), 2.467 for bootstrap(m) and 2.358 for subsample(m), where again larger values correspond to worse predictions. In both the examples we note very similar results for the three resampling approaches, suggesting that the additional variables included in the models derived using bootstrap(m) and bootstrap(n) do not have added predictive value. The (very) slightly better results obtained for subsample(m), moreover, seem to suggest that the inclusion of additional variables can even worsen the prediction abilities of the models, likely due to overfitting.

Please note that for the Glioma data we compute the integrated Brier score up to 712 days, which represents the median follow-up (computed via reverse Kaplan-Meier). In our computations we experienced some cases in which a resampling procedure on the 9 folds produced a singular X matrix (usually because all the pseudo-observations have *surgd1* equal to 1): we discarded them, leading to a number of repetitions slightly smaller than 10,000.

4.2 Results for the simulation study

Figure 3 shows the variable inclusion frequencies obtained from the simulated data for the three resampling approaches. We recall that the first two variables have strong effects (0.2) while the third, fourth, fifth and sixth have low effects (0.1). All the others have no effect. We immediately note that for the two bootstrap approaches the variables with no effect are selected too many times: their inclusion frequencies, indeed, are noticeably higher than the theoretical value of 0.05 (type I error), the significance level used in the Wald tests during the backward elimination procedure. It is worth noting that for subsample(m) the inclusion frequencies of these variables are also slightly higher than the nominal value 0.05. This may partly be a consequence of the multiple testing problem associated with the backward elimination procedure and of the correlation among the variables. The effect of the former issue can be seen for variable x_{25}, whose inclusion frequency is 0.058 and therefore slightly larger than 0.05 although it is uncorrelated to the other variables. This agrees with results from previous simulation studies (Sauerbrei, 1992, 1993); see also some discussions in Royston and Sauerbrei (2008, Chapter 2). Therefore, multiple testing does not explain the high values obtained for the two bootstrap approaches.

About the correlation issue, we can see its effect on the inclusion frequency of x_9, which is noticeably larger than 0.05 for all three resampling approaches. In addition to the aforementioned multiple testing problem, this variable suffers the effect of its high correlation with x_4 ($\rho = -0.519$). We can analyze the effect of the correlation on the inclusion frequencies of x_4 and x_9 following the approach introduced by Sauerbrei and Schumacher (1992). It consists of displaying the inclusions/exclusions of the two variables in a 2×2 table: it is then straightforward to understand whether the inclusion of one variable influences the inclusion of the other. This is the case of variables x_4 and x_9: the inclusion of the former decreases the chances that the latter is selected by the backward elimination procedure, and vice versa. As a consequence, the inclusion frequencies of both variables are lower than those that we would have obtained in the uncorrelated case. Table A.10 in the Web Appendix reports the analysis.

If we consider the relevant variables, we note that their inclusion frequencies for subsample(m) are higher than those for bootstrap(m) (x_2, x_3 and x_4, while x_1's inclusion frequency is 1 for all approaches), behavior further recommending subsample(m) and seeming to validate its performance in the real data example. The only case in which the bootstrap approaches perform better than subsample(m) is for x_5 and x_6. As remarked in Section 2.3, these two variables are binary and strongly unbalanced: as a con-

sequence, the power of the tests decreases due to their variances, leading to lower inclusion frequencies. Moreover, the algorithm here used to perform backward elimination excludes a binary variable from the model when it takes the same value (either 0 or 1) for all observations in the pseudo-sample (see also Harrell, 2013), thus further decreasing the frequency of selection of strongly unbalanced binary variables. Finally, x_5 and x_6 are highly correlated ($\rho = 0.553$), and therefore their inclusion frequencies may be affected by the problem of "alternate selection" mentioned in Section 3.2.1 and previously observed between x_4 and x_9. An analysis based on the results of Sauerbrei and Schumacher (1992), similar to that performed for x_4 and x_9, confirms the presence of this issue (see also Table A.11 in the Web Appendix): also in this case, the inclusion of one variable seems to lead to the exclusion of the other. Furthermore, the variables x_{17} and x_{24} are correlated with x_5 and x_6 as well ($\rho = 0.368$ and 0.407, respectively, for x_{17}, $\rho = 0.336$ and 0.373 for x_{24}), and may also slightly contribute to the low inclusion frequencies of x_5 and x_6 (in any case the effect is minimum, being the inclusion frequencies of x_{17} and x_{24} not so far from 0.05). In any case, the difference between the inclusion frequencies of these two variables and the noise variables is greater for subsample(m) than for the two bootstrap approaches. Using inclusion frequencies as a criterion, we note that all three approaches rank the variables in the same way.

As described in Section 3.2.3, we base our considerations on the ability of the resampling approaches to identify the relevant variables on the computation of the area under the curve (AUC). The distributions of the AUC for bootstrap(n), bootstrap(m) and subsample(m), computed in the 1,000 datasets generated in our simulation study, are reported in Figure 4. We note a better performance for subsample(m) compared to the two bootstrap approaches, with bootstrap(m) slightly better than bootstrap(n). The reason for this result mainly lies in the tendency of the bootstrap approaches to include noise variables in the model. Bootstrap(n), indeed, has the worst AUC even though the inclusion frequencies for the relevant variables are higher than those obtained with bootstrap(m) and, to a lesser extent, than those obtained with subsample(m) (see Figure 3). The analysis of the AUC, therefore, also suggests that subsample(m) is preferable to the bootstrap in this context.

If we look at the models obtained in the analysis (Table 2), we note that the true model is selected only a few times, no matter which resampling approach is used. For subsample(m), its selection frequency is slightly better than for bootstrap(n) (1041 vs 780) and bootstrap(m) (774), but it appears lower in the ranking (128th, while it is 65th for bootstrap(n) and 105th for bootstrap(m)). This situation seems to be related to the unbalanced nature of

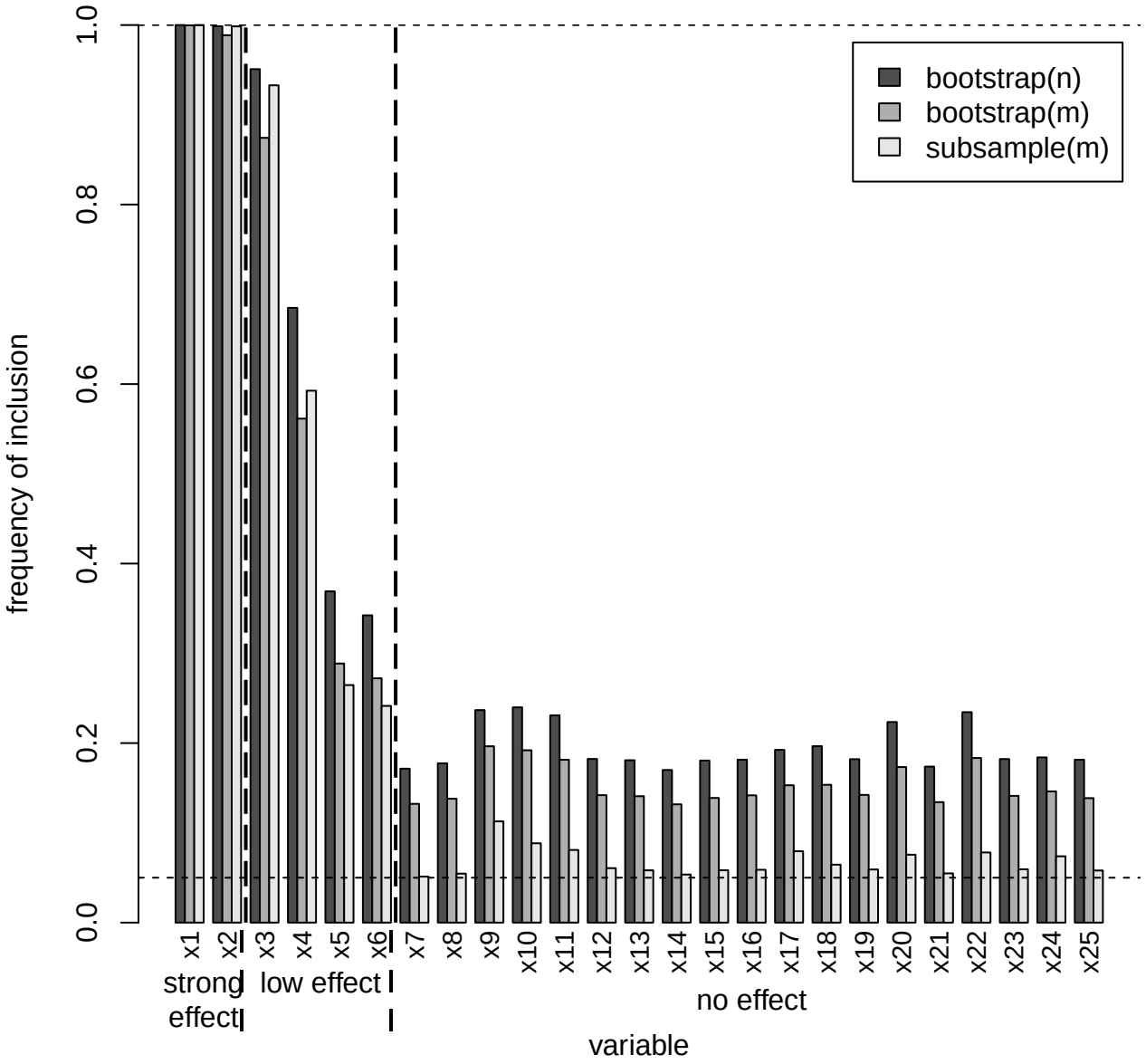

Figure 3: Simulated data: inclusion frequencies of the variables based on 1,000,000 pseudo-samples, 1,000 for each dataset.

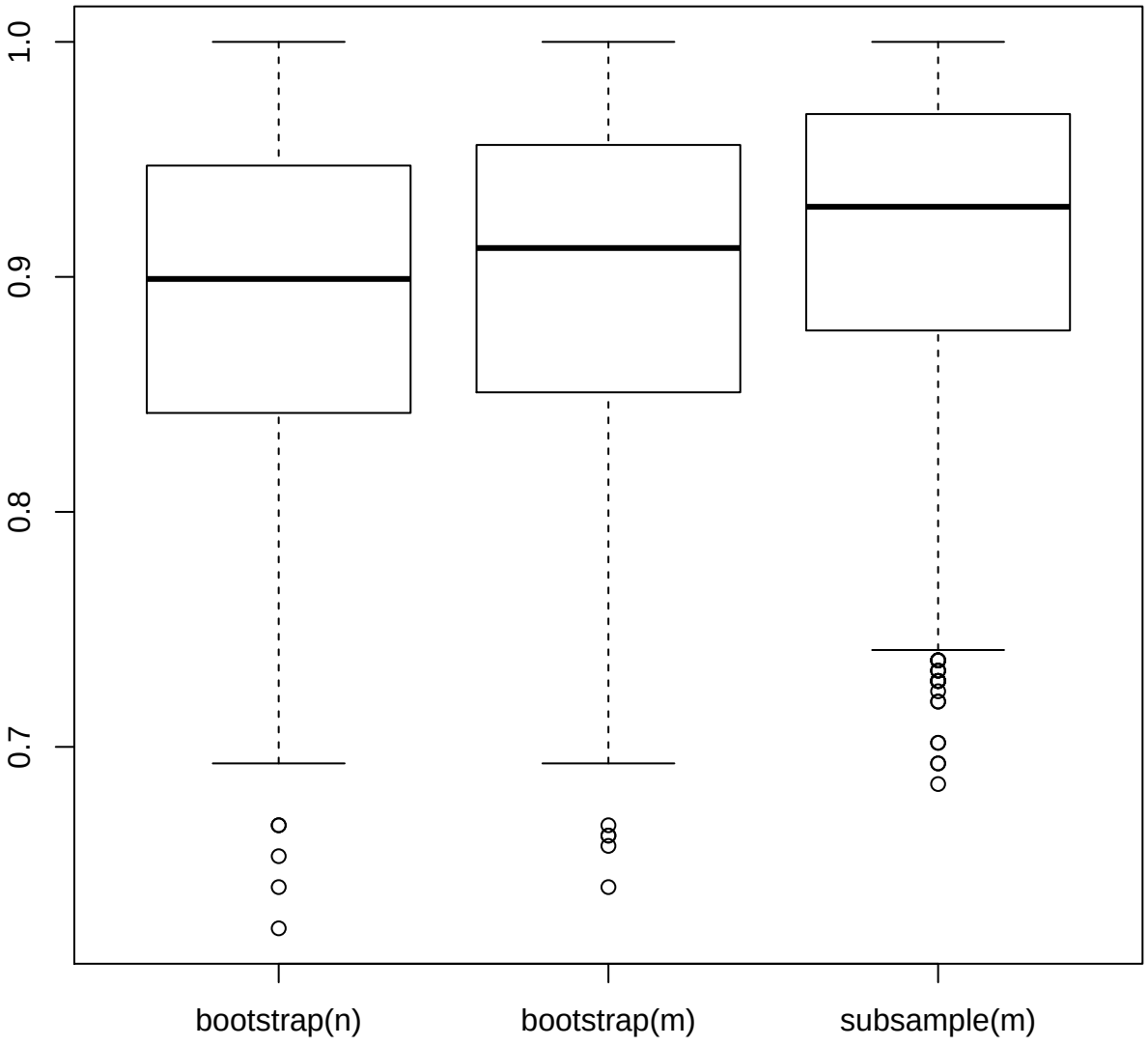

Figure 4: Simulated data: distribution of the AUC computed on 1,000 pseudo-datasets for bootstrap(n), bootstrap(m) and subsample(m).

Table 2: Simulated data: selection frequencies of the 10 top ranked models for bootstrap(n), bootstrap(m) and subsample(m), based on 1,000,000 pseudo-samples for $\alpha = 0.05$ and presented in decreasing sum of the three selection frequencies. The true model is included in the table although it is not selected in the top 10 positions for any resampling technique.

	bootstrap(n)		bootstrap(m)		subsample(m)	
model	rank	freq.	rank	freq.	rank	freq.
basic+x_3+x_4	3	5089	1	14317	1	77518
basic+x_3	5	2047	2	11621	2	56363
basic+x_3+x_4+x_5	1	7172	3	10312	3	55180
basic+x_3+x_4+x_6	2	6434	4	9850	4	53321
basic+x_3+x_5	4	2119	5	6281	5	28902
basic+x_3+x_6	8	1656	6	5054	6	21646
basic+x_3+x_4+x_{17}	6	1849	7	3132	7	10736
basic+x_3+x_9	46	959	9	3041	8	10450
basic+x_3+x_4+x_{24}	7	1772	8	3064	9	9996
basic+x_3+x_{17}	68	756	10	2541	10	8597
basic+x_3+x_4+x_5+x_{25}	9	1537	74	1508	20	3464
basic+x_3+x_4+x_5+x_{16}	10	1444	76	1443	90	3009
basic+x_3+x_4+x_5+x_6	65	780	105	774	128	1041

basic=intercept+x_1+x_2

x_5 and x_6 and to their correlations (between them and with other variables).

Due to the large number of variables available and the complex correlation structure, we note a high dispersion of selection frequencies. For example, for bootstrap(n) the most selected model has a selection frequency smaller than 1% (0.71%). The situation is slightly better for bootstrap(m) (selection frequency for the best model around 1.43%) and for subsample(m) (7.75%). The high dispersion is more evident when we look at the number of unique models selected in the analysis: for bootstrap(n) we obtain 244,392 different models, 181,258 for bootstrap(m) and 36,552 for subsample(m).

Finally, we also report the average number of variables included in the models for the simulation study. As seen in Table 2 and in agreement with the results of the real studies, we again note the preference of subsample(m) for sparse models. The average number of variables per model, indeed, is 5.311, slightly smaller than the true value (6). Conversely, the two bootstrap techniques tend to select too many variables, an average of 6.886 for bootstrap(m) and 8.048 for bootstrap(n).

5 Discussion

In this paper we compared the subsampling and the bootstrap approaches in a model building procedure for multivariable regression using backward elimination. From our study, subsampling emerged as a valid alternative to the bootstrap. Our simulation study, in particular, shows how the bootstrap approaches lead to high inclusion frequencies for noise variables, considerably larger than the theoretical value of 0.05 (see also Rospleszcz et al., 2014). As a consequence, subsampling provides better results in terms of variable inclusion frequencies, with consequences on the ability of recognizing the relevant variables (they are more separated from the noise ones) and, consequently, on the selection of useful models. This is confirmed by the analysis of the AUC, which summarizes the ability of separating relevant and noise variables using all possible thresholds.

The results of the simulation study cast new light on the least included variables in the real data studies. Those variables have inclusion frequencies close to 0.05, just as the noise variables in the simulation study do. Therefore, we may safely suppose that they do not have any effect, or, at most, a very weak one, and that the inclusion frequencies for these variables obtained in the real data examples with bootstrap(n) and bootstrap(m) are too high. Our conclusion here is confirmed by the results obtained by analyzing the prediction abilities of the models: we saw no improvement in the greater inclusion of these variables conferred by the two bootstrap approaches.

In the future, we would like to investigate the reasons for this behavior: as one possibility, the higher number of variables included in a model derived from a bootstrap sample is surely related to the incorrect significance level for a test based on a bootstrap sample, which is larger than the nominal (see, for example, Janitza et al., 2014, and references therein). However, other characteristics of the pseudo-samples generated via bootstrap may also play an important role: for example, the replication of possible influential points (or even outliers) due to the resampling with replacement may contribute to the selection of noise variables. An analysis based, for example, on the work of Sauerbrei et al. (2014) may help to clarify this point.

A possible issue related to the use of subsample(m) is the correct selection of the low-effect variables. We saw in the simulation studies that these variables may have inclusion frequencies which are too low, partially due to the correlation structure and partially to the decrease in the power of significance tests due to $m < n$. This may lead to the construction of models which are too sparse (as seen in the analysis of the average number of variables included in the models) and, eventually, to underfitting issues.

An important choice that may be related to this issue and that we did

not consider in this paper is that of m and its effect on the results. We used a value of $0.632n$ for m to set the size of the pseudo-samples generated via subsample(m) equal to the average number of unique observations for bootstrap(n). A larger value of m may improve the performance of subsample(m), increasing the inclusion frequencies of the low-effect variables (as an effect of the increased power of the significance tests). If m increases too much, however, we do not investigate instability anymore since the pseudo-samples are too similar to each other. A smaller value for m, instead, may decrease the too high inclusion frequencies of the noise variables for bootstrap(m). However, in decreasing m, the probable simultaneous decrease of the inclusion frequencies for the relevant variables may lead to serious problems of underfitting for both bootstrap(m) and subsample(m).

Acknowledgments

RDB was financed by grant BO3139/4-1 to ALB, SJ was financed by grant BO3139/2-2 to ALB and WS was supported by grant SA580/8-1. All the three grants are from the German Science Foundation (DFG). The authors wish to thank Rory Wilson for help with linguistic improvements.

Appendix

Table A.1: Glioma data, effect estimates (log hazard ratios), standard error and p-values for the Cox model including all variables.

variable	estimate	std error	p-value
sex	-0.175	0.129	0.17460
time	-0.128	0.140	0.36274
gradd1	0.798	0.251	0.00151
gradd2	0.257	0.190	0.17585
age	0.038	0.007	9×10^{-7}
kard1	-0.317	0.139	0.02305
kard2	-0.039	0.172	0.82208
surgd1	-1.046	0.213	9×10^{-9}
surgd2	-0.216	0.139	0.11962
convul	0.095	0.138	0.49361
cort	0.264	0.139	0.05755
epi	-0.270	0.150	0.07148
amnesia	0.097	0.198	0.62390
ops	0.253	0.164	0.12328
aph	-0.119	0.137	0.27478

Table A.2: Ozone data, effect estimates, standard error and p-values for the linear model including all variables.

variable	estimate	std error	p-value
intercept	-1.721	0.264	2×10^{-10}
alter	0.025	0.017	0.15708
adheu	-0.038	0.043	0.37135
sex	-0.197	0.020	1×10^{-16}
hochozon	-0.069	0.027	0.01202
amatop	-0.003	0.023	0.87883
avatop	-0.017	0.024	0.48672
adekz	0.009	0.025	0.70635
arauch	0.007	0.022	0.75821
agebgew	2×10^{-5}	2×10^{-5}	0.33302
fsnight	0.026	0035.	0.44492
flgross	0.026	0.002	1×10^{-16}
fmilb	-0.057	0.037	0.12073
fnoh24	-0.002	0.001	0.00468
ftier	-0.013	0.037	0.71378
fpoll	-0.060	0.045	0.18902
fltotmed	-0.054	0.028	0.05463
fo3h24	0.001	0.001	0.11463
fspt	0.032	0.049	0.51448
fteh24	-0.005	0.003	0.12744
fsatem	0.102	0.054	0.06102
fsauge	0.010	0.032	0.76082
flgew	0.012	0.002	3×10^{-9}
fspfei	0.122	0.055	0.02825
fshlauf	-0.032	0.043	0.45219

Table A.3: regression coefficients in the full models fitted for the first 10 simulated datasets.

	simulated datatset									
variable	1	2	3	4	5	6	7	8	9	10
intercept	2.473	2.564	2.533	2.487	2.493	2.546	2.470	2.439	2.537	2.435
x_1	0.182	0.222	0.175	0.187	0.190	0.208	0.231	0.174	0.168	0.211
x_2	-0.138	-0.210	-0.195	-0.226	-0.215	-0.159	-0.165	-0.194	-0.213	-0.152
x_3	0.123	0.086	0.120	0.124	0.130	0.125	0.083	0.091	0.099	0.068
x_4	-0.052	-0.118	-0.110	-0.094	-0.119	-0.146	-0.083	-0.040	-0.134	-0.077
x_5	-0.043	0.193	-0.040	0.203	0.096	0.131	0.185	0.162	0.045	0.125
x_6	0.136	-0.088	-0.026	0.090	0.134	0.133	0.037	0.053	0.012	-0.004
x_7	0.009	-0.023	-0.004	0.005	0.023	-0.016	-0.000	0.023	0.009	0.002
x_8	-0.018	0.047	0.001	0.006	-0.027	-0.017	0.029	-0.009	-0.031	0.021
x_9	0.029	-0.065	-0.018	0.025	-0.004	-0.017	0.039	0.035	-0.040	-0.024
x_{10}	-0.097	0.001	0.012	0.003	-0.026	0.046	0.018	-0.008	0.002	-0.036
x_{11}	0.102	0.049	-0.041	-0.004	0.027	-0.045	-0.036	-0.002	0.009	0.018
x_{12}	-0.033	-0.127	-0.032	0.083	-0.041	0.007	-0.016	-0.056	0.018	-0.072
x_{13}	-0.031	-0.003	0.009	0.012	0.033	0.044	0.060	-0.017	0.019	0.066
x_{14}	0.037	-0.011	0.018	-0.041	0.055	-0.007	-0.004	-0.021	-0.023	-0.017
x_{15}	0.010	0.008	-0.024	0.028	-0.047	-0.057	-0.006	0.068	-0.043	0.027
x_{16}	0.026	0.001	-0.012	0.038	0.018	-0.062	0.008	0.009	-0.030	0.048
x_{17}	-0.024	-0.117	0.052	-0.039	-0.072	0.017	0.003	0.037	0.080	-0.133
x_{18}	0.080	-0.017	0.059	-0.066	-0.058	0.023	-0.112	0.055	0.047	-0.016
x_{19}	0.042	-0.023	0.044	0.028	0.027	-0.078	0.046	-0.034	-0.048	-0.040
x_{20}	0.013	0.172	-0.102	0.046	0.041	0.060	0.055	0.202	-0.027	0.103
x_{21}	-0.025	0.032	0.044	-0.039	0.062	-0.083	0.010	0.009	0.050	0.065
x_{22}	-0.089	-0.133	0.006	-0.027	0.009	-0.112	-0.038	-0.177	-0.029	-0.067
x_{23}	-0.094	0.062	-0.006	0.020	0.029	0.013	0.048	-0.024	0.043	-0.032
x_{24}	-0.036	-0.045	0.150	0.002	0.052	-0.130	0.053	-0.125	0.023	0.185
x_{25}	0.000	0.009	-0.020	-0.009	-0.020	0.009	0.009	0.033	0.022	0.030

Table A.4: Glioma data, inclusion frequencies of the variables (based on 10,000 pseudo-samples).

	$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.157$		
	bootstrap (n)	bootstrap (m)	subsample (m)	bootstrap (n)	bootstrap (m)	subsample (m)	bootstrap (n)	bootstrap (m)	subsample (m)
sex	0.32	0.14	0.24	0.43	0.25	0.33	0.50	0.35	0.42
time	0.19	0.06	0.16	0.27	0.12	0.23	0.34	0.19	0.30
gradd1	0.87	0.84	0.77	0.92	0.89	0.81	0.94	0.94	0.85
gradd2	0.37	0.23	0.31	0.45	0.32	0.38	0.52	0.41	0.44
age	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
kard1	0.70	0.63	0.53	0.76	0.73	0.62	0.80	0.79	0.67
kard2	0.09	0.02	0.09	0.14	0.04	0.15	0.19	0.06	0.20
surgd1	1.00	1.00	0.97	1.00	1.00	0.98	1.00	1.00	0.99
surgd2	0.36	0.16	0.26	0.47	0.29	0.36	0.55	0.41	0.44
convul	0.14	0.03	0.12	0.22	0.07	0.20	0.28	0.12	0.26
cort	0.52	0.29	0.39	0.62	0.46	0.50	0.71	0.59	0.58
epi	0.51	0.30	0.38	0.62	0.45	0.49	0.69	0.57	0.57
amnesia	0.21	0.10	0.19	0.27	0.14	0.26	0.33	0.19	0.32
ops	0.43	0.24	0.33	0.52	0.37	0.43	0.60	0.48	0.50
aph	0.17	0.03	0.13	0.27	0.10	0.21	0.36	0.17	0.29

Table A.5: Glioma data, number of unique models based on 10,000 pseudo-samples for three significance level.

resampling approach	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.157$
bootstrap(n)	1787	2296	2450
bootstrap(m)	1829	2573	3100
subsample(m)	580	1047	1461

Table A.6: Glioma data, average number of included variable for three significance levels.

resampling approach	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.157$
bootstrap(n)	6.864	7.957	8.832
bootstrap(m)	5.857	6.931	7.829
subsample(m)	5.057	6.242	7.278

Table A.7: Glioma data, cluster of the selected models with respect to their structure (based on 10,000 pseudo-samples). The term "core" denotes the variable *gradd*1, *age* and *surgd*1. The term "core*" means that *gradd*1 and *surgd*1 can be replaced by *gradd*2 and *surgd*2, respectively.

Variable	bootstrap (n)	bootstrap (m)	subsample (m)	Variable	bootstrap (n)	bootstrap (m)	subsample (m)
Only the 3 core	15	123	398	Only the 3 core*	17	178	473
3 core + 1 additional	247	878	2432	3 core* + 1 additional	304	1213	2841
3 core + 2 additional	1030	1923	2786	3 core* + 2 additional	1272	2590	3441
3 core + 3 additional	2071	2123	1956	3 core* + 3 additional	2472	2772	2309
3 core + 4 additional	2505	1451	653	3 core* + 4 additional	2832	1825	730
3 core + 5 additional	1742	676	155	3 core* + 5 additional	1904	803	163
3 core + >5 additional	1103	275	27	3 core* + >5 additional	1180	321	27
Without at least 1 core	1287	2551	1593	Without at least 1 core*	19	298	16

Table A.8: Ozone data, selection frequencies of the 10 top ranked models for bootstrap(n), bootstrap(m) and subsample(m), based on 10,000 pseudo-samples for $\alpha = 0.05$. The order of presentation depends on the sum of the three selection frequencies (decreasing).

	bootstrap(n)		bootstrap(m)		subsample(m)	
model	rank	freq.	rank	freq.	rank	freq.
basic+fspfei+fpoll	25	24	1	80	1	416
basic+fsatem	94	10	2	73	2	371
basic+fspfei	94	10	4	68	3	340
basic+fsatem+fmilb	16	29	6	56	4	318
basic+fsatem+fpoll	22	25	3	69	5	312
basic+fspfei+fmilb+hochozon+fnoh24	1	72	5	63	6	295
basic+fspfei+fpoll+hochozon+fnoh24	3	59	10	48	7	269
basic+fspfei+fmilb	39	18	22	31	8	233
basic+fsatem+fmilb+hochozon+fnoh24	4	56	14	41	9	221
basic	244	5	7	51	10	206
basic+fsatem+hochozon+fnoh24	49	17	7	51	17	126
basic+fspfei+fmilb+hochozon+fnoh24+fo3h24+fteh24	2	60	9	49	31	70
basic+fspfei+fpoll+hochozon+fnoh24+fo3h24+fteh24+fltotmed	5	54	22	31	40	53
basic+fspfei+fpoll+fsatem+hochozon+fnoh24+fltotmed	6	46	49	19	64	30
basic+fspfei+hochozon+fnoh24+fo3h24+fteh24	7	42	19	32	24	79
basic+fspfei+fmilb+fsatem+hochozon+fnoh24+fltotmed	8	38	68	15	75	24
basic+fspfei+fpoll+fsatem+hochozon+fnoh24+fo3h24+fteh24+fltotmed	8	38	129	9	275	3
basic+fspfei+fpoll+hochozon+fnoh24+fltotmed	10	37	56	18	24	79

basic=intercept+sex+flgross+flgew

Table A.9: Ozone data, cluster of the selected models with respect to their structure (based on 10,000 pseudo-samples)

	bootstrap(n)	bootstrap(m)	subsample(m)
basic	5	51	206
basic + 1 additional	31	309	953
basic + 2 additional	217	921	2639
basic + 3 additional	643	1666	2333
basic + 4 additional	1333	1989	2167
basic + 5 additional	1648	1830	796
basic + $>$ 5 additional	6123	3189	906
Without at least 1 core	0	45	0

basic=intercept+sex+flgross+flgew

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281.

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

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **7,** 1196–1217.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica* **7,** 1–31.

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. *Statistica Sinica* **18,** 967–985.

Buchholz, A., Holländer, N., and Sauerbrei, W. (2008). On properties of predictors derived with a two-step bootstrap model averaging approach: a simulation study in the linear regression model. *Computational Statistics & Data Analysis* **52,** 2778–2793.

Burr, D. (1994). A comparison of certain bootstrap confidence intervals in the Cox model. *Journal of the American Statistical Association* **89,** 1290–1302.

Table A.10: Simulated data, contingency table with the inclusion/exclusion of variables x_4 and x_9

			x_9		
bootstrap(m)	x_4		in	out	Sum
		in	0.075	0.486	0.561
		out	0.121	0.317	0.438
		Sum	0.196	0.803	1.000

			x_9		
bootstrap(n)	x_4		in	out	Sum
		in	0.112	0.573	0.685
		out	0.125	0.190	0.315
		Sum	0.237	0.763	1.000

			x_9		
subsampling(m)	x_4		in	out	Sum
		in	0.030	0.563	0.593
		out	0.083	0.324	0.407
		Sum	0.113	0.887	1.000

Chen, C.-H. 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.

Chernick, M. R. (2011). *Bootstrap Methods: a guide for practitioners and researchers*. Wiley.

Copas, J. B. and Long, T. (1991). Estimating the residual variance in orthogonal regression with variable selection. *The Statistician* **40,** 51–59.

Davison, A. C., Hinkley, D. V., and Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science* **18,** 141–157.

De Bin, R., Sauerbrei, W., and Boulesteix, A. L. (2014). Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in Medicine* **DOI: 10.1002/sim.6246,**.

del Barrio, E., Janssen, A., and Matrán, C. (2009). Resampling schemes with low resampling intensity and their applications in testing hypotheses. *Journal of Statistical Planning and Inference* **139,** 184–202.

Table A.11: Simulated data, contingency table with the inclusion/exclusion of variables x_5 and x_6

			x_6		
			in	out	Sum
bootstrap(m)	x_5	in	0.032	0.256	0.288
		out	0.240	0.472	0.712
		Sum	0.272	0.728	1.000

			x_6		
			in	out	Sum
bootstrap(n)	x_5	in	0.056	0.313	0.369
		out	0.286	0.345	0.631
		Sum	0.342	0.658	1.000

			x_6		
			in	out	Sum
subsampling(m)	x_5	in	0.088	0.256	0.264
		out	0.233	0.503	0.736
		Sum	0.241	0.759	1.000

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7,** 1–26.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley.

Gong, G. (1982). Some ideas on using the bootstrap in assessing model variability. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 169–173. Springer.

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.

Harrell, F. E. (2013). *rms: Regression Modeling Strategies*. R package version 3.6-3.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association* **64,** 1303–1317.

Ihorst, G., Frischer, T., Horak, F., Schumacher, M., Kopp, M., Forster, J., Mattes, J., and Kuehr, J. (2004). Long-and medium-term ozone effects on

lung growth including a broad spectrum of exposure. *European Respiratory Journal* **23,** 292–299.

Janitza, S., Binder, H., and Boulesteix, A.-L. (2014). Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications. Technical Report 163, Department of Statistics, University of Munich.

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

Mammen, E. (1992). *When Does Bootstrap Work?* Springer.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21,** 255–285.

Mammen, E. and Nandi, S. (2012). Bootstrap and resampling. In *Handbook of Computational Statistics*, pages 499–527. Springer.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics* **12,** 621–625.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34,** 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72,** 417–473.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press.

Politis, D., Romano, J., and Wolf, M. (1999). *Subsampling*. Springer.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* **22,** 2031–2050.

Rao, C. R., Pathak, P., and Koltchinskii, V. (1997). Bootstrap by sequential resampling. *Journal of Statistical Planning and Inference* **64,** 257–281.

Rospleszcz, S., Janitza, S., and Boulesteix, A.-L. (2014). The effects of bootstrapping on model selection for multiple regression. Technical Report 164, Department of Statistics, University of Munich.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. Wiley.

Sauerbrei, W. (1992). *Variablenselektion in Regressionsmodellen unter besonderer Berücksichtigung medizinischer Fragestellungen.* PhD thesis, University of Dortmund.

Sauerbrei, W. (1993). Comparison of variable selection procedures in regression models – a simulation study and practical examples. In *Europäische Perspectiven der Medizinischen Informatik, Biometrie und Epidemologie*, pages 108–113. MMV Medizin.

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., and Binder, H. (2014). On stability issues in deriving multivariable regression models. *Biometrical Journal* **to appear,**.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap 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. *The Annals of Statistics* **6,** 461–464.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics* **17,** 1176–1197.

Tutz, G. (2012). *Regression for categorical data.* Cambridge University Press.

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

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14,** 1261–1295.

Zelterman, D., Le, C. T., and Louis, T. A. (1996). Bootstrap techniques for proportional hazards models with censored observations. *Statistics and Computing* **6,** 191–199.