Biometrics 72, 272–280
March 2016

DOI: 10.1111/biom.12381

Subsampling versus Bootstrapping in Resampling-Based Model Selection for Multivariable Regression

Riccardo De Bin,[1,*] Silke Janitza,[1,] Willi Sauerbrei,[2,***] and Anne-Laure Boulesteix[1,****]**

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany
[2]Department of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79106 Freiburg im Breisgau, Germany
*email: debin@ibe.med.uni-muenchen.de
**email: janitza@ibe.med.uni-muenchen.de
***email: wfs@imbi.uni-freiburg.de
****email: boulesteix@ibe.med.uni-muenchen.de

Summary. In recent 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 begun to be taken into consideration in the literature. Using model selection frequencies and variable inclusion frequencies, we empirically compare these two different resampling techniques, investigating the effect of their use in selected classical model selection procedures 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.

Key words: Bootstrap; Model selection; Model stability; Subsampling.

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., 2015). Here, we are mainly interested in the latter. In order to investigate model stability and variable selection procedures, 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). Using the bootstrap technique (Efron, 1979), one generates pseudo-samples which can be seen as perturbed versions of the original data. The 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 article, 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., 2015, 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 article 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 bootstrap. 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 characteristics of the selected models. We extend the analysis by considering a second dataset, 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 article is structured as follows: in Section 2 we describe the two datasets and 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 procedures and the resampling approaches. The results obtained from the two real datasets and from the simulation study are 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 who took part in a randomized controlled trial for comparing two kinds of chemotherapy. The outcome of interest is the survival time of these 411 patients, 276 (67.2%) of whom died. In addition to the form of chemotherapy, 12 variables are considered, including sex, age, time from first symptoms to diagnosis (binary: long/short), and information on health-related conditions (malignancy grade, Karnofsky index, presence of convulsions, epilepsy, amnesia, organic psycho-syndrome, aphasia) and on treatment history (resection type, use of cortisone). Three variables that were originally measured on three-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); these dummy variables are then treated as separate variables. More details on the Glioma data can be found in Sauerbrei and Schumacher (1992). The data used here have no missing data and are publicly available at http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book.

Web Table A.1 in the Web Appendix A 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, $p < 0.0001$). In addition to the high correlation present by construction between the dummy variables coding the same ordinal variable, 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 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. A well-defined subset of the data is used in Buchholz et al. (2008) and Sauerbrei et al. (2015). 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 A (Web 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 articles, 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 do not allow us the drawing of any conclusions regarding the reliability of the resulting variable rankings, as we do not know which variables are actually related to the outcome, i.e., which variables should—in the ideal situation—be included in the final model. This prevents us from properly evaluating the quality of the inclusion frequencies for the available variables. To handle 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 strong effect, i.e., those with an estimate in absolute value larger than 0.15: here *flgross* and *sex*;
  - the variables with weak 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 19 are not. We use these six variables to generate 1000 artificial outcomes, drawing from a Gaussian distribution with mean $2.5 + 0.2\,flgross - 0.2sex + 0.1\,flgew - 0.1hochozon + 0.1\,fsatem + 0.1\,fspfei$ 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 1000 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) nonzero 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 Web Table A.3 in the Web Appendix A). 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 25 explanatory variables, of which two have a strong effect on the response, four have a weak effect, and 19 are noise variables (no effect). The sample size is 496, as in the original Ozone data.

For each of these 1000 datasets, we perform our analyses by generating B = 1000 pseudo-samples with each of the different resampling techniques and variable selection strategies. 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, because, often, more variables than those necessary are included in a given study. There are several reasons to reduce the number of variables in a model, the most obvious being that some variables may not be related to the outcome and should be removed following the principle of Occam's razor (ontological sparsity). Moreover, even if all variables are somehow related to the outcome, it may be advantageous to remove some of those with small effect, either to increase the interpretability of the final model (epistemic sparsity) or to obtain, by reducing the variance, a model with better predictive ability (predictive sparsity). In this study, we are mainly interested in the ontological sparsity.

The literature on variable selection is boundless and it is outside the scope of this article to provide an overview. Here, we use the three classical procedures backward elimination, forward selection, and stepwise selection (see, e.g., Royston and Sauerbrei, 2008, Chapters 1 and 2, for an overview). Stepwise selection allows both inclusion and exclusion of the variables during the procedure, either starting from the null model (forward selection with possible re-exclusion) or from the full model (backward elimination with possible reinclusion). Since in our analyses the differences between these two approaches are not noteworthy (in the two real examples we even obtain exactly the same values), we report the results only for backward elimination with possible reinclusion. Results may differ more with more complex correlation structures, in particular in the case of small sample sizes.

We would like to stress the relevance of the choice of the selection criterion, which greatly impacts the stability and the complexity of the final model (Royston and Sauerbrei, 2008, Chapter 2). In this article, we use an approach based on the significance level α for a likelihood ratio test on the regression coefficients (α taking values 0.01, 0.05, 0.10, and 0.157) or based on an information criterion, here AIC and BIC. Throughout the article, we only report the results for backward elimination and forward selection with $\alpha = 0.05$. The results related to the other significance levels and to the information criteria, as well as those obtained for stepwise selection, are given in the Web Appendices. Note that the results for stepwise selection do not differ substantially from those for backward elimination.

3.2. *Resampling*

3.2.1. Inclusion frequencies and selected models. 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 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, i.e., those 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.

In addition, the analysis of the selected models themselves is of interest, as it may also provide further insights into the model building process.

3.2.2. Resampling strategies. In order to generate the pseudo-samples for our analyses, we need to choose 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 known as nonparametric bootstrap: several other versions of the bootstrap are available in the literature (see, among others, Chernick, 2011).

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 article, 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 article 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). 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 in the Glioma dataset, some complications occur due to censored observations. By directly applying the resampling technique, indeed, we obtain pseudo-samples with different effective sizes (number of events). In order to handle 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. Arguments 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). For the real data, we describe the results and give some ideas on the effects of the different resampling approaches on the selected models, in terms of number of unique models, model selection frequency, and model sparsity (average number of variables in the model).

With the results obtained in the simulation study, instead, we directly assess the quality of the inclusion frequencies obtained via the three resampling techniques. Knowing the true model, indeed, allows us to compare the values of the observed inclusion frequencies with the desired ones (close to 1 for the strong effect variables, close to α for the noise variables, between these two values for those with weak effect). Moreover, we compute a measure which quantifies how well the inclusion frequencies obtained through an arbitrary resampling approach can be used to discriminate between the relevant and the noise variables. To do this, we compute the proportion of noise variables $(x_7, \ldots, x_{25})$ which have a lower inclusion frequency than that of a given relevant variable. Averaging this proportion over all relevant variables $(x_1, \ldots, x_6)$, we obtain an estimate of the area under the curve (AUC). Note that the AUC is 1 for perfect discrimination and 0.5 for discrimination which is not better than random. We compute this measure for the inclusion frequencies obtained for all three resampling approaches and we compare the results in terms of distribution of the AUC over the 1000 simulated datasets. For further information about this measure, see Pepe (2003) and Janitza et al. (2013).

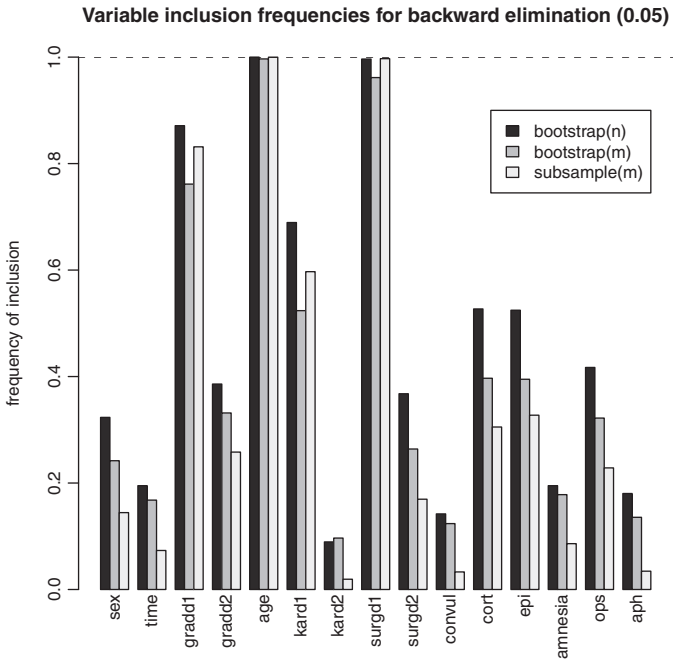

Figure 1. Glioma data: inclusion frequencies, based on 10,000 pseudo-samples, for all 15 variables, obtained with backward elimination and $\alpha = 0.05$.

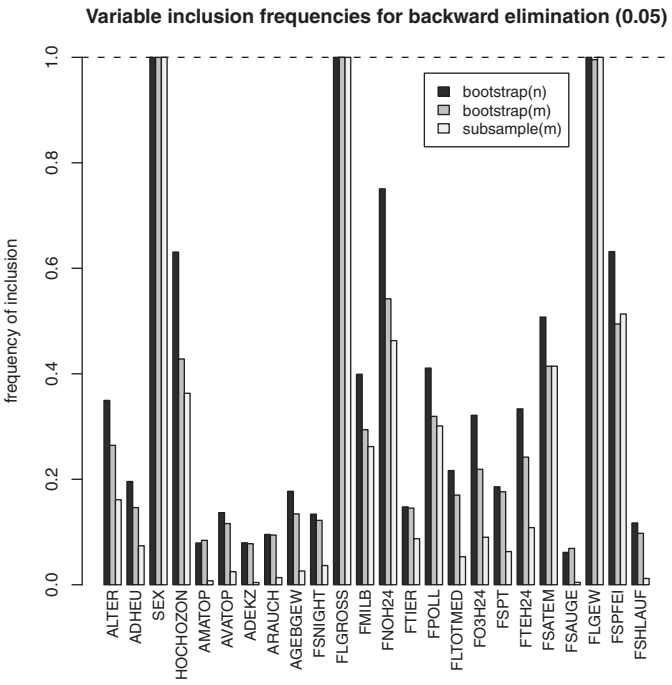

Figure 2. Ozone data: inclusion frequencies, based on 10,000 pseudo-samples, for all 24 variables, obtained with backward elimination and $\alpha = 0.05$.

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 ignore the models themselves and only consider the variable inclusion frequencies. Moreover, we give the same importance to the inclusion of the relevant variables as to the exclusion of the noise variables. As we focus on explanatory models, we do not consider alternative weighting schemes (for example, in the context of prediction models it would be preferable to assign more weight to the correct inclusion of the relevant variables than to the correct exclusion of the noise ones).

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, obtained by applying backward elimination with $\alpha = 0.05$ to 10,000 pseudo-samples. 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), the situation is similar, although less pronounced due to the very strong effects of the three core variables, whose inclusion frequencies are close to 1 for all 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 elimination 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 weak effect or their inclusion frequencies may be influenced by the inclusion frequencies of other variables; for a detailed investigation of these issues, we refer the reader to Sauerbrei and Schumacher (1992). Interestingly, *kard1*'s inclusion frequency for subsample(m), as well as those for the three core variables, is higher than that for bootstrap(m), while for all the other variables the opposite is true. It seems that subsample(m) provides re-

sults in which the variables with strong/medium effect and the variables with weak/no effect are more distinctly separated than in the results for bootstrap(n) and bootstrap(m) (see Figure 1). It is worth noting, however, that the three resampling methods rank the variables in the same order. Similar observations apply to the Ozone data.

Concerning the differences among the resampling procedures, note that we obtain very similar results for forward and stepwise selection and with different selection criteria (Web Tables A.4–A.6 in Web Appendix A and Web Figures B.1–B.18 in Web Appendix B for the Glioma data, Web Tables A.7–A.9 in Web Appendix A and Web Figures B.19–B.36 in Web Appendix B for the Ozone data).

4.1.2. Selected models. Web Appendix C contains the results of our investigations of the real data in terms of the selected models. For both Glioma and Ozone examples, we note that subsample(m) tends to select sparser models than the two bootstrap approaches (see Web Tables C.1–C.2 and C.7–C.8), with a resulting strong influence on the number of unique models obtained (Web Tables C.3 and C.9) and their selection frequencies (Web Tables C.4–C.6 and C.10–C.12), regardless of which variable selection technique and which selection criterion are implemented. While this result is completely expected for bootstrap(n), due to the increased power derived from the larger pseudo-sample size (and consequent inclusion of more weak effect variables), it is interesting to note that for bootstrap(m) we obtain values (in terms of number of unique models, model selection frequencies, and average number of variables in the models) more similar to those of bootstrap(n) than subsample(m). See Web Tables C.1–C.6 for the Glioma data and Web Tables C.7–C.12 for the Ozone data.

Although our main goal is to compare the abilities of resampling-based model selection procedures in identifying relevant variables, we note that the tendency of subsampling(m) to select sparse models may be advantageous from a predictive point of view: a small cross-validation-based study on the predictive abilities of the models selected using the three different resampling techniques—in combination with backward elimination and $\alpha = 0.05$—revealed that the models selected using subsample(m) have an average prediction ability slightly better than those selected using the two bootstrap techniques, both for the Glioma and the Ozone data (data not shown).

4.2. *Results for the Simulation Study*

4.2.1. Variable inclusion frequencies. Figures 3 and 4 show the variable inclusion frequencies obtained from the simulated data for the three resampling approaches, using backward elimination and forward selection with $\alpha = 0.05$. Concerning the differences among the resampling procedures, the results obtained for stepwise selection and for the other selection criteria (see Web Tables A.10–A.12 in Web Appendix A and Web Figures B.37–B.54 in Web Appendix B) are very similar; thus, the comments below also apply to those cases.

Recall that in our simulation study the first two variables have strong effects (0.2) while the third, fourth, fifth, and sixth have weak effects (0.1). All others have no effect. We immediately note that for the two bootstrap approaches the

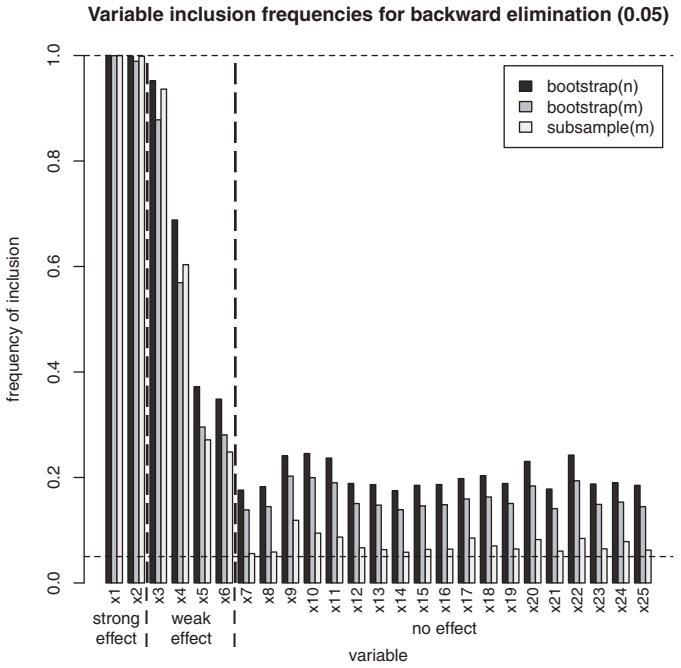

Figure 3. Simulated data: inclusion frequencies of the variables based on 1,000,000 pseudo-samples, 1000 for each dataset. Results are for backward elimination with $\alpha = 0.05$.

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 coverages obtained with subsample(m), instead, are much better, and around the

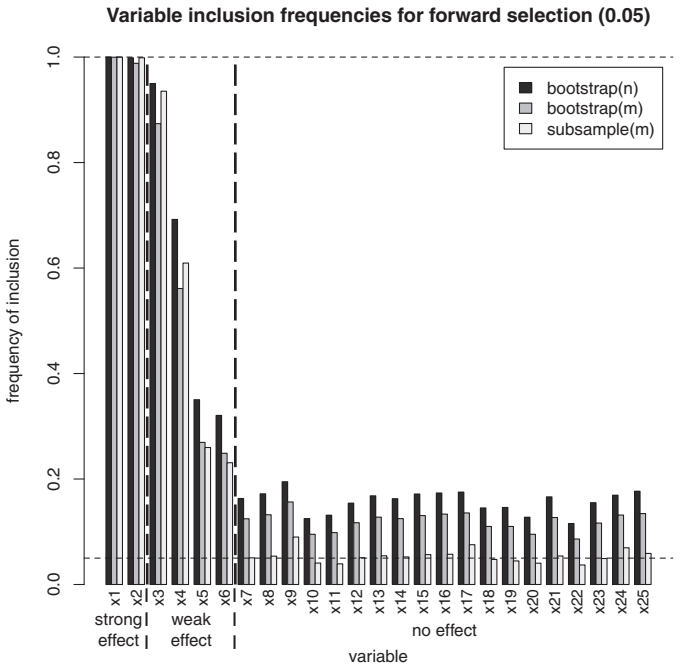

Figure 4. Simulated data: inclusion frequencies of the variables based on 1,000,000 pseudo-samples, 1000 for each dataset. Results are for forward selection with $\alpha = 0.05$.

nominal value. We note a small difference between the results of backward elimination and forward selection: in the case of highly correlated noise variables, such as x_{10} and x_{11} ($\rho = 0.860$), the inclusion frequencies obtained using the former variable selection technique are larger than 0.05, while using the latter technique we obtain values slightly smaller than 0.05 (Figures 3 and 4). We can also see the effect of this correlation issue on the inclusion frequency of x_9 ($\rho = -0.519$ with relevant variable x_4), which is noticeably larger than 0.05 for all resampling approaches.

It is worth noting that the multiple testing problem associated with the variable selection techniques may also play a role here: likely as a consequence of this issue, the inclusion frequency of the variable x_{25}, which is uncorrelated to the other variables, is slightly larger than the nominal 0.05. However, we know from previous simulation studies (Sauerbrei, 1992, 1993) that this multiple testing effect is minimal (see also Royston and Sauerbrei, 2008, Chapter 2) and therefore it does not explain the high inclusion frequencies obtained for the two bootstrap approaches.

Examining 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). This result is a strong argument in favor of the use of subsample(m) and seems to validate its performance in the real data example. The only variables for which the bootstrap approaches perform better than subsample(m) are x_5 and x_6. As remarked in Section 2.3, these two variables are binary and strongly unbalanced: as a consequence, the power of the tests decreases due to their variances, leading to lower inclusion frequencies, with correlation ($\rho = 0.553$) also likely playing a role. Regardless, 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. Note that the ranking of the variables by inclusion frequency is identical for the three subsampling methods.

4.2.2. Discrimination ability by inclusion frequency. As described in Section 3.2.3, we use the area under the curve (AUC) to evaluate the ability of each resampling approach to distinguish relevant variables. The distributions of the AUC for bootstrap(n), bootstrap(m), and subsample(m), taken over the 1000 simulated datasets, are reported in Figures 5 and 6. Again, we report the results for $\alpha = 0.05$ using backward elimination and forward selection; the results for stepwise selection and for the other selection criteria can be found in Web Appendices A (Web Table A.13) and B (Web Figures B.55–B.72).

The figures indicate slightly superior performance for subsample(m) over 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 more frequently in the models. Bootstrap(n), indeed, has the worst AUC even though its inclusion frequencies for the relevant variables are higher than those obtained with bootstrap(m) and, to a lesser extent, those obtained with subsample(m) (see Figures 3 and 4). The analysis of the AUC, therefore, also suggests that subsample(m) is preferable to the bootstrap in this regard.

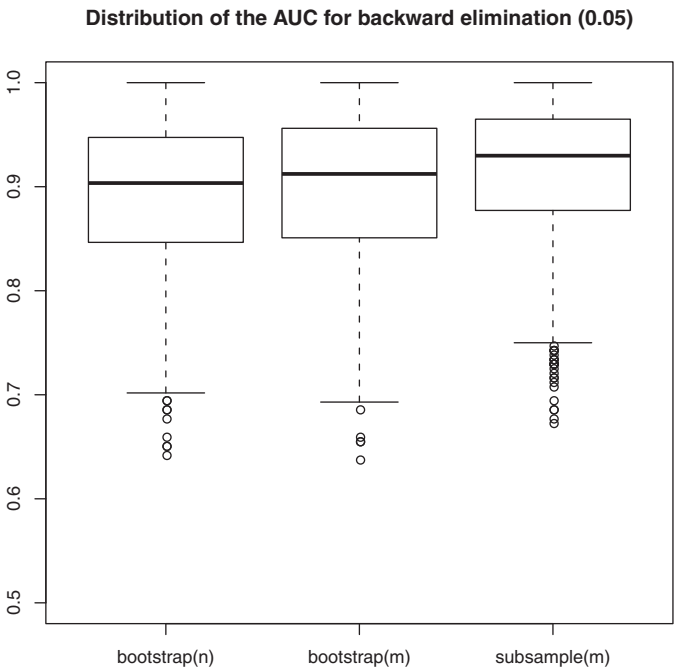

Figure 5. Simulated data: distribution of the AUC computed on 1000 pseudo-datasets for bootstrap(n), bootstrap(m), and subsample(m) using backward elimination with significance level 0.05.

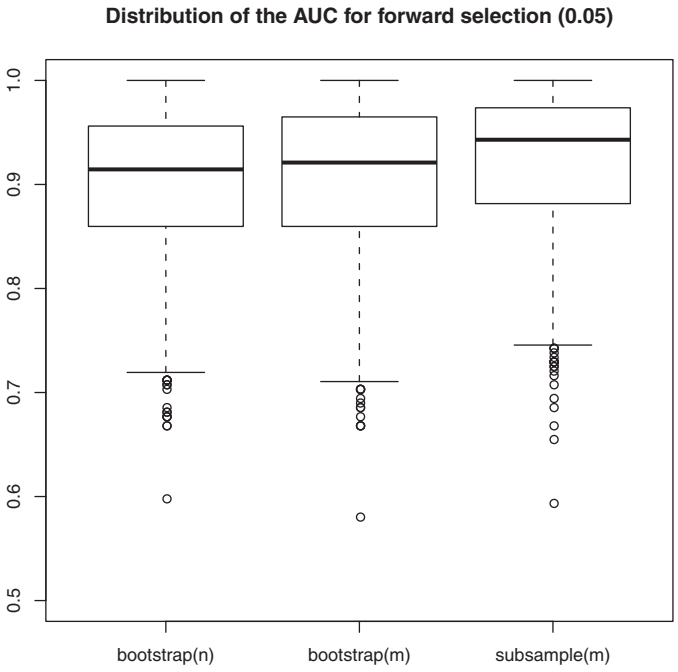

Figure 6. Simulated data: distribution of the AUC computed on 1000 pseudo-datasets for bootstrap(n), bootstrap(m), and subsample(m) using forward selection with significance level 0.05.

Note that there is no substantial difference in the relationships between the AUCs of the three resampling approaches when we vary the variable selection technique or the selection criterion.

4.2.3. Selected models. Web Appendix C reports the analyses on the relation between variable inclusion frequencies and selected models for the simulation study. It is again clear that subsample(m) tends to favor sparser models (Web Tables C.13–C.17). The latter three tables also display how frequently the true model is selected: regardless of resampling approach used, this occurs only at a rate of about 0.001%. This low frequency is likely a result of the combination of the weak effects of, and strong correlation between, the variables x_5 and x_6: in this type of situation, often only one of the two variables is selected (Sauerbrei and Schumacher, 1992). If we consider the models which include only one of the two variables (and the other relevant variables) also as "true," the selection frequency for subsample(m) is then much better than those for the two bootstrap approaches (approximatively 0.11% for subsample(m), 0.02% for bootstrap(n) and bootstrap(m)). The more frequent inclusion of one or more variables without effect is the main reason for the worse performance of the two bootstrap approaches. Surprisingly, the magnitude of the selection frequencies of the "true" model for bootstrap(n) and bootstrap(m) is similar, despite the larger pseudo-sample size of bootstrap(n).

5. Discussion

In this article, we compared the subsampling and the bootstrap approaches in a model building procedure for multivariable regression using three classical variable selection procedures. From our study, subsampling emerged as a valid alternative to the bootstrap. Our simulation study, in particular, reveals that 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). Subsampling does not display this behavior, thus allowing the easier recognition of the relevant variables (see, in particular, Figures 3 and 4). The superiority of subsampling in this regard is confirmed by the analysis of the AUC, which summarizes a method's ability to separate relevant and noise variables using all possible thresholds.

In the future, we would like to investigate the reasons for the high inclusion frequencies for noise variables in the bootstrap methods. As one possibility, this behavior may be 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., 2015, and references therein). However, other characteristics of the pseudo-samples generated via bootstrap may also play important roles: 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. (2015) may give further insights into this point.

A possible disadvantage of the use of subsample(m) is the selection of the weak 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 due to the decrease in the power of significance tests, as $m < n$. 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, Web Figure C.14 in Web Appendix C) and to underfitting issues.

Choice of the pseudo-sample size m. An important choice that may be related to this issue and that we did not consider in this article 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 weak effect variables (as a consequence of the increased power of the significance tests). If we increase m too much, however, we are no longer investigating instability, as the pseudo-samples become too similar. 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).

Variable selection procedures. In the article, we saw that our findings on the comparison of subsampling and bootstrapping do not depend on which classical variable selection procedure is implemented. From an inclusion frequency point of view, moreover, we saw that the possibility of reinclusion of previously excluded variables (stepwise selection) does not modify substantially the results obtained by backward elimination, no matter which selection criterion is used. In contrast, forward selection produced different inclusion frequencies for correlated variables. The differences concern both relevant and noise variables, and generally show decreased inclusion frequencies for forward selection. This result is clearest in the simulation study (variables x_{10} and x_{11}, see Figures 3 and 4), but it is also apparent in the Glioma (see, e.g., variables *surgd1* and *surgd2*) and the Ozone (especially between the variables *hochozon* and *fnoh24*) examples. A noticeably exception is the inclusion frequency for *gradd2* (Glioma data), which is smaller for backward elimination (and stepwise selection) than for forward selection.

Selection criteria. The choice of the selection criterion (four different significance levels, AIC and BIC) does not impact our conclusions concerning the comparison of the resampling strategies, the main aim of the article. With regard to variable selection criteria, we note that AIC and $\alpha = 0.157$ provide very similar results, as earlier noted by Sauerbrei (1999).

6. Supplementary Materials

Web Appendices A, B, C, referenced in Sections 2, 4, 5, and the R code for our analyses are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

RDB was financed by grant BO3139/4-1 to ALB, SJ 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 thank Rory Wilson for help with linguistic improvements.

References

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

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.

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.

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.

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

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester: 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*, 169–173. New York: Springer.

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. (2015). Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. *Biometrical Journal* , to appear.

Janitza, S., Strobl, C., and Boulesteix, A.-L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* **14**, 119.

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

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

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*. Boca Raton: CRC Press.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

Politis, D., Romano, J., and Wolf, M. (1999). *Subsampling*. New York: 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.

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*. Chichester: 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*, 108–113. MMV Munich: Medizin.

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

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., and Binder, H. (2015). On stability issues in deriving multivariable regression models. *Biometrical Journal* **57**, 531–555.

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*. New York: Cambridge University Press.

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.

Received October 2014. *Revised June* 2015. *Accepted July* 2015.