Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications

Silke Janitza*[,1], **Harald Binder**[2], and **Anne-Laure Boulesteix**[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

[2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Johannes Gutenberg University Mainz, Obere Zahlbacher Str. 69, 55131 Mainz, Germany

Received 14 November 2014; revised 22 April 2015; accepted 13 June 2015

The bootstrap method has become a widely used tool applied in diverse areas where results based on asymptotic theory are scarce. It can be applied, for example, for assessing the variance of a statistic, a quantile of interest or for significance testing by resampling from the null hypothesis. Recently, some approaches have been proposed in the biometrical field where hypothesis testing or model selection is performed on a bootstrap sample as if it were the original sample. *P*-values computed from bootstrap samples have been used, for example, in the statistics and bioinformatics literature for ranking genes with respect to their differential expression, for estimating the variability of *p*-values and for model stability investigations. Procedures which make use of bootstrapped information criteria are often applied in model stability investigations and model averaging approaches as well as when estimating the error of model selection procedures which involve tuning parameters. From the literature, however, there is evidence that *p*-values and model selection criteria evaluated on bootstrap data sets do not represent what would be obtained on the original data or new data drawn from the overall population. We explain the reasons for this and, through the use of a real data set and simulations, we assess the practical impact on procedures relevant to biometrical applications in cases where it has not yet been studied. Moreover, we investigate the behavior of subsampling (i.e., drawing from a data set without replacement) as a potential alternative solution to the bootstrap for these procedures.

Keywords: Bootstrap; Bootstrapped information criteria; Bootstrapped *p*-values; Bootstrapped test statistic; Tests on bootstrap samples.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

The bootstrap, introduced by Efron (1979), consists of generating a huge number of pseudo-samples from the original data set of interest. In the case of the nonparametric bootstrap (considered in this paper), a pseudo-sample is generated by randomly drawing observations with replacement from the original data. One then typically performs statistical analyses on each bootstrap sample, for instance the computation of an estimator of interest, yielding so-called bootstrapped estimates. Such procedures are becoming more and more widely used, as indicated by the now large number of reference textbooks on the subject (Davison, 1997; Good, 2005; Manly, 2006; Chernick, 2011). Bootstrapped estimates can be used to derive, for example, the variance of an estimator, a quantile of interest or a confidence interval (Davison, 1997).

*Corresponding author: e-mail: janitza@ibe.med.uni-muenchen.de, Phone: +49 89 440077755

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

In this paper, we are interested in the case where a *p*-value of a standard statistical test (such as, e.g., the *Z*-test or the likelihood ratio [LR] test) takes the role of the estimator which is being bootstrapped. More precisely, we mean *p*-values that result from statistical tests performed using a bootstrap sample as the data set as if it were the original data set, ignoring that it has actually been drawn with replacement from another sample. For example, a popular bootstrap-based method often applied in biometrical applications investigates the stability of stepwise model selection procedures. This procedure makes use of the bootstrap to generate pseudo-samples, and model selection is performed on each bootstrap sample, where *p*-values of the LR test are used to decide on the inclusion of variables in the model (Chen and George, 1985; Altman and Andersen, 1989; Sauerbrei and Schumacher, 1992).

More concisely, the problem we are addressing in this paper is that *p*-values computed from bootstrap samples are not valid and cannot be interpreted as *p*-values: when performing tests on bootstrap samples as if they were realizations from the true unknown distribution, the type I error is increased. This problem has already been reported in the literature for the special case of the LRtest (Bollen and Stine, 1992) and the χ^2-test (Strobl et al., 2007) and its consequences for the above-mentioned model stability investigations have also been very recently investigated (Rospleszcz et al., 2014; De Bin et al., 2015).

It is important to note that although we are interested in a problem also related to bootstrapping and testing, this problem is fundamentally different from obtaining *p*-values by the so-called bootstrap tests (Efron and Tibshirani, 1993). Bootstrap tests are an alternative to inference based on parametric assumptions when these assumptions are questionable or when such a method simply does not exist. A bootstrap test works roughly as follows: the estimator of interest is computed from a large number of bootstrap samples and the resulting empirical distribution is in some way compared to the null hypothesis. For example, if the null hypothesis states that the parameter of interest equals a certain value, one might look whether this value is within the confidence interval derived from the bootstrap estimates. Bootstrap tests as well as their pitfalls and some potential solutions have been extensively discussed in the literature in recent decades; see Efron and Tibshirani (1993) for an overview. It is important to note that in this paper we are never referring to *p*-values obtained by such bootstrap tests when we speak of bootstrapped *p*-values. Instead we are referring to the *p*-values that are obtained from performing any statistical test (such as, e.g., the *Z*-test or the LR test) using a bootstrap sample as the data set as if it were the original, as outlined in the previous paragraph and which is a completely different approach.

Bootstrapped *p*-values have been far less investigated than the bootstrap tests outlined in the previous paragraph. However, procedures based on bootstrapped *p*-values are not uncommon in the literature, especially in biometrical applications. Besides the example of stability investigations for stepwise model selection procedures mentioned above, *p*-values computed from bootstrap samples have been used in the statistics and bioinformatics literature for ranking genes with respect to their differential expression (Mukherjee et al., 2003), for estimating the variability of *p*-values which one would observe when repeating an experiment multiple times (Boos and Stefanski, 2011) or, in a completely different context, for deciding which variable should be selected for splitting in the recursive partitioning algorithm "random forest," which consists of building decision trees from bootstrap samples (Strobl et al., 2007).

Information criteria like the Akaike information criterion (AIC) or the Bayesian information criterion are also sometimes computed on bootstrap samples and used in a variety of contexts: examples include model stability investigations—as mentioned in the previous paragraph—and model averaging approaches, in which model weights are derived based on bootstrapped AICs (Buckland et al., 1997; Burnham and Anderson, 2002). Moreover, bootstrapped information criteria become relevant when estimating the error of model selection procedures which involve tuning parameters. When computed on bootstrap samples, it has been shown that bootstrapped information criteria deviate from information criteria that are computed based on the original sample. In the context of graphical models (Steck and Jaakkola, 2003) and model averaging procedures (Wagenmakers et al., 2004) this deviation has been shown to lead to a preference for models which are too complex. This issue may also be relevant when using the bootstrap, as an alternative to, say, cross-validation, for estimating the error of a prediction modeling strategy. For a large number of bootstrap samples drawn from the original data set, a prediction model is fit to the bootstrap sample using the considered strategy and is then used to make

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

predictions for the observations which were not included in this bootstrap sample and are thus considered test data. This yields an estimate for the prediction error of the model and the estimates from all bootstrap samples are averaged. Binder and Schumacher (2008) showed that the resulting error estimate is biased in the case where the prediction modeling strategy involves a parameter tuning step based on internal cross-validation. However, as we will show in this paper problems may also occur if the prediction modeling strategy involves a parameter tuning step based on an information criterion like the AIC.

In all these applications, it is essential that quantities such as p-values or model selection criteria evaluated on bootstrap samples represent what would be obtained on the original data or new data drawn from the overall population. Several articles, as previously outlined, suggest that this might often not be the case. However, the papers investigating the problems arising from bootstrapping p-values or information criteria are spread over a wide range of very heterogeneous journals which are not often read by biometricians and are to our knowledge not discussed in textbooks. Moreover, they handle very specific cases and a simple general theory to explain the problem is lacking. Further, the practical consequences for biomedical applications are to date largely unknown. The present paper addresses these problems. It gives new theoretical insights and investigates the practical consequences of three specific applications proposed in the literature which are based on either bootstrapped p-values (Applications 1 and 2) or bootstrapped information criteria (Application 3). Application 1 uses bootstrapped p-values for variable ranking, as proposed, for example, by Mukherjee et al. (2003). For each variable, bootstrapped p-values are computed for testing the null hypothesis that the variable is not associated with the response. A ranking is obtained by sorting the variables by their mean or median bootstrapped p-value (in our studies we use the median). Such resampling-based strategies might be helpful in minimizing the influence of some extreme observations which may greatly affect a variable ranking. Mukherjee et al. (2003) proposed their variable ranking approach in the context of ranking genes in small samples, where influential points are likely to have a huge impact. In our studies we use data of the National Health and Nutrition Examination Survey (NHANES) to rank variables based on bootstrapped p-values and we compare this ranking to that obtained by sorting variables based on the p-values of the original NHANES data. In Application 2, the variability of bootstrapped p-values is used to approximate the true p-value variability (Boos and Stefanski, 2011). It is motivated by the question of how much the p-values obtained from a study of an original data set would differ if one were to replicate this study. Boos and Stefanski (2011) propose computing the variability of bootstrapped p-values to report conjointly with the original p-value in real data applications to gain a better understanding of the variability of p-values. If the standard deviation is a large fraction of the p-value, there is high variability, which may explain the fact that identical experiments may lead to rather distinct p-values. For two classical tests, we conduct simulation studies to inspect whether the variability of bootstrapped p-values is a good approximation of the variability of p-values computed based on data drawn from the true underlying distribution. Application 3 concerns the aforementioned model selection procedures involving tuning parameters. Here, we consider gradient boosting models with the number of boosting steps as a tuning parameter. In our studies, we focus on the complexity of the resulting boosting models (in terms of the number of included variables) and compare the complexity of models derived based on bootstrap samples to that of models derived based on the original NHANES data; we furthermore compare the models' prediction accuracies. In all of our studies, we also investigate the behavior of subsampling (i.e., drawing from a data set without replacement) as an alternative to the bootstrap. Source code to reproduce the results of our studies is available as Supporting Information on the journal web page.

The structure of the paper is as follows. Section 2 starts with a description of the NHANES data which is used in our studies. We then specify the terms bootstrapped p-values and bootstrapped information criteria and briefly introduce the subsampling method. In Section 3, we present our studies on bootstrapped p-values for two classical widely used statistical tests, the Z-test and the LR test. We first show that tests performed on bootstrap samples have increased type I error, and that bootstrapped p-values often do not approximate p-values computed on the original data very well; we subsequently show the consequences for Applications 1 and 2. In Section 4, we present our studies on bootstrapped information criteria. We show that the aforementioned systematic deviation of bootstrapped

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

information criteria impacts model selection when basing the decision on the AIC and investigate the consequences for Application 3. Section 5 gives a discussion of our results and an outlook.

2 Data and methods

Section 2.1 gives a brief description of the NHANES data used in our studies. In Section 2.2, we describe the computation of p-values based on a single bootstrap sample when ignoring that the sample was drawn from the empirical distribution and not from the true distribution. We use the Z-test (Section 2.2.1) and the LR test (Section 2.2.2) as examples. Subsequently in Section 2.3 we describe the computation of information criteria based on bootstrap samples using the AIC as an example. In Section 2.4, we briefly describe the subsampling method investigated in this paper as a possible alternative to the bootstrap.

2.1 NHANES data

We consider data from the 2007 and 2008 cycle of the NHANES (National Center for Health Statistics, 2012) which is maintained by the Centers for Disease Control and Prevention. NHANES is designed as a series of cross-sectional surveys conducted in the US population. The data are freely available from the institution's homepage or from the Interuniversity Consortium for Political and Social Research. The considered data set comprises a total of $n = 1914$ subjects. For our investigations, we use the level of high-sensitive C-reactive protein (CRP) as the response. The CRP is a plasma protein involved in the acute phase response during inflammatory states (Black et al., 2004). We included 28 variables in our studies potentially related to the CRP level. These variables include information on the medical facility to which the subject most often goes, the subject's sex, age, body mass index, waist circumference, race, country of birth, education, marital status, income, smoking history, information on alcohol consumption, as well as laboratory values (white blood cell counts, systolic and diastolic blood pressure, cholesterol level) and health-related conditions including asthma, diabetes, history of stroke, history of heart failure, history of chronic bronchitis, history of any acute illnesses, history of alcohol abuse, self-rated general health, tooth condition, depressive mood, sleeping abnormalities (consisting of items on falling asleep and waking during the night). Many of the variables were obtained from interviews with the study persons. The corresponding interview questions, the abbreviations for the variables used in this paper and the measurement units of the variables are given in Table A1. Note that the data include metric predictors and categorical (nominal and ordinal) predictors. In our analyses, we treated the ordinal variables as nominal.

We also wanted to investigate settings where none of the predictors were associated with the response. To obtain a data set without any associations between the predictors and the response, we randomly permuted the response variable of the NHANES data to break any potential association between the 28 covariates and the response. This was repeated 1000 times to obtain a total of 1000 data sets in which no associations are present. The resulting data sets are called "permuted NHANES data" in this paper to distinguish them from the original NHANES data with unpermuted response.

2.2 Bootstrapping p-values

2.2.1 Z-test

Let $\boldsymbol{x}^\top = (x_1, \ldots, x_n)$ be realizations drawn from $N(\mu, \sigma^2)$ and let $\hat{F}$ denote the corresponding empirical distribution with known σ^2. The test statistic for testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$ is given by $Z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$, with $\bar{x}$ denoting the sample mean. Then Z follows a normal distribution with $\mathrm{E}(Z) = \sqrt{n}(\mu - \mu_0)/\sigma$ and $\mathrm{Var}(Z) = 1$.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Now let $\boldsymbol{x}^{*\top} = (x_1^*, \ldots, x_n^*)$ denote the realizations of a bootstrap sample that was drawn from $\hat{F}$ with replacement. The bootstrapped test statistic from a Z-test with hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ is defined as

$$Z^* = \sqrt{n}\frac{\bar{x}^* - \mu_0}{\sigma}, \tag{1}$$

with $\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*$ and known σ^2. If incorrectly assuming that Z^* follows a standard normal distribution, the corresponding bootstrapped p-value for an observed test statistic Z^* is computed as

$$p^* = 2 \cdot (1 - \Phi(|Z^*|)), \tag{2}$$

with Φ denoting the cumulative distribution function of the standard normal distribution. As will be shown in this paper, decisions based on bootstrapped p-values lead to increased type I error.

2.2.2 LR test

The LR test is used, for example, when comparing the fit of two nested models, where one model contains restrictions that are not imposed in the other. The likelihood of the restricted model, called the submodel in the following, is termed L_0, while L_1 corresponds to the likelihood of the unrestricted model. The test statistic for the LR test is defined as twice the difference in log-likelihoods:

$$T = -2(\log(L_0) - \log(L_1)). \tag{3}$$

The test statistic T asymptotically follows a noncentral χ^2-distribution with df degrees of freedom, which is calculated as the difference in degrees of freedom of the two models, and with noncentrality parameter κ. The asymptotic expectation of the test statistic is given by $\mathrm{AE}(T) = df + \kappa$ and the asymptotic variance is $\mathrm{AVar}(T) = 2df + 4\kappa$. Under the null hypothesis which states that the submodel is true, the noncentrality parameter is zero and thus T asymptotically follows a central $\chi^2(df)$-distribution and has asymptotic expectation $\mathrm{AE}(T) = df$ and asymptotic variance $\mathrm{AVar}(T) = 2df$.

The corresponding bootstrapped test statistic for the LR test is

$$T^* = -2(\log(L_0^*) - \log(L_1^*)), \tag{4}$$

with L_0^* and L_1^* denoting the likelihoods for the submodel and the unrestricted model, respectively, both evaluated on a bootstrap sample. The bootstrapped p-value for an observed T^* is defined as

$$p^* = Pr(\Lambda \geq T^* | H_0), \tag{5}$$

with $\Lambda \sim \chi^2(df)$.

2.3 Bootstrapping information criteria for model building

Information criteria are often used for the comparison of nonnested models. These measures compare models based on their goodness of fit to the data while penalizing the complexity of the model (see also Burnham and Anderson, 2002). AIC is a widely used measure for model selection. It is defined as

$$\mathrm{AIC} = -2\log(L) + 2p, \tag{6}$$

where L denotes the likelihood and p denotes the number of parameters included in the model. It has been shown that minimizing the AIC is approximately equivalent to minimizing the expected Kullback–Leibler distance between the true and the estimated density (Akaike, 1973).

The bootstrapped AIC is given by

$$\mathrm{AIC}^* = -2\log(L^*) + 2p, \tag{7}$$

with L^* denoting the likelihood computed for a model that was fit on a bootstrap sample.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The AICs for two nested models can be compared by using the LR test statistic (3). If AIC_1 denotes the AIC of the unrestricted model that includes p parameters and AIC_0 denotes the AIC of the submodel that includes $p-1$ parameters, then the LR test statistic on 1 degree of freedom can be expressed in terms of AIC_0 and AIC_1 as follows (cf. chapter 6.9.3 in Burnham and Anderson, 2002):

$$T = AIC_0 - AIC_1 + 2. \tag{8}$$

From Eq. (8) we see that if both models fit the data equally well according to the AIC (i.e., $AIC_0 = AIC_1$), we have $T = 2$. Further, the unrestricted model is chosen over the submodel if its AIC is smaller, corresponding to $AIC_0 - AIC_1 > 0$ and, according to Eq. (8), $T > 2$. In contrast, the submodel is chosen if $AIC_0 - AIC_1 < 0$, corresponding to $T < 2$. These considerations show that in the case of two nested models one can also use the value of the LR test statistic to decide which of the models is better in terms of the AIC; if the two models differ only in the inclusion of one parameter, values for the LR test statistic below 2 indicate the superiority of the submodel, values above 2 indicate that the unrestricted model is better, and both models are considered equally good if the test statistic T takes the value 2.

2.4 Subsampling as an alternative to the bootstrap

The subsampling procedure, also known as delete-d jackknife (Wu, 1986), is closely related to the bootstrap, but in contrast to the bootstrap a subsample is created by drawing m observations, with $m < n$, without replacement from the original sample. The optimal choice of the parameter m is delicate and is not treated here (see, e.g., Davison et al., 2003; Bickel and Sakov, 2008). In our studies, we chose m as the value $0.632n$, which corresponds to the expected number of unique observations in a bootstrap sample, in order to have on average the same number of unique observations in subsamples and bootstrap samples. The subsampling technique has been investigated in the literature and also contrasted to the bootstrap (Hartigan, 1969; Shao and Wu, 1989; Politis and Romano, 1994; Politis et al., 1999). It shows asymptotic consistency in cases where the bootstrap fails (Davison et al., 2003; Chernick, 2011). In particular, the type I error is not increased for test statistics computed on subsamples. For this reason it has been recommended to use subsampling instead of bootstrapping in the random forest algorithm (Strobl et al., 2007).

3 Studies on bootstrapping *p*-values

In Section 3.1, we show through theoretical and empirical results that the null hypothesis is too frequently rejected for both the Z-test and the LR test when using bootstrapped p-values, that is, there is increased type I error (Section 3.1.1). We also explore the conditional (i.e., conditional on $\boldsymbol{x}$) distribution of bootstrapped p-values under the null hypothesis (Section 3.1.2). In Sections 3.2 and 3.3, we investigate the consequences for two practices, namely, bootstrapping p-values for variable ranking and bootstrapping p-values for assessing the variability of p-values.

3.1 General results

3.1.1 Type I error

We use the following theorem to show that Z-tests for the test statistic Z^* (Eq. (1)) have increased type I error, or equivalently, that decisions made on bootstrapped p-values (Eq. (2)) lead to systematically too many false positive findings.

Theorem 1. *Let the bootstrapped test statistic for a Z-test with $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ be defined as in Eq.* (1)*. The unconditional expectation of this bootstrapped Z-test statistic Z^* is $E(Z^*) = E(Z)$, while the unconditional variance of Z^* is $Var(Z^*) = 2$.*

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1 Type I error when performing two-sided and one-sided upper Z-tests with predefined significance thresholds for test statistic $Z = \sqrt{n}(\frac{1}{n}\sum x_i - \mu_0)/\sigma$ with $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and for a bootstrapped test statistic $Z^* = \sqrt{n}(\frac{1}{n}\sum x_i^* - \mu_0)/\sigma$.

Hypotheses	Significance threshold	Type I error	
		for Z	for Z^*
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	$z_{0.95} = 1.64$	0.10	0.24
(two-sided test)	$z_{0.975} = 1.96$	0.05	0.17
	$z_{0.995} = 2.58$	0.01	0.07
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$z_{0.90} = 1.28$	0.10	0.18
(one-sided test)	$z_{0.95} = 1.64$	0.05	0.12
	$z_{0.99} = 2.33$	0.01	0.05

The proof of Theorem 1 is given in Section A.1. According to Theorem 1, the unconditional variance of the bootstrapped statistic Z^* is twice the variance of Z. Thus, under the null hypothesis that $H_0 : \mu = \mu_0$ (or $H_0 : \mu \leq \mu_0$; $H_0 : \mu \geq \mu_0$ for one-sided tests), the marginal distribution of the bootstrapped statistic Z^* is not the standard normal distribution (see the Supplementary Material for empirical results). Using the significance threshold $z_{1-\frac{\alpha}{2}}$, the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, the type I error is $2 \cdot (1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\frac{\alpha}{2}}))$, where Φ is the standard normal distribution function. For a one-sided lower (upper) test with null hypothesis $H_0 : \mu \geq \mu_0$ ($H_0 : \mu \leq \mu_0$), the significance threshold z_α ($z_{1-\alpha}$) is used and the type I error is $\Phi(\frac{1}{\sqrt{2}} z_\alpha)$ (and $1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\alpha})$, respectively). Table 1 shows examples for the type I error when performing Z-tests for test statistics Z and Z^*. It can be seen that the type I error is substantially increased when performing Z-tests on bootstrap samples as if they were the original samples.

Similar considerations apply when performing LR tests based on test statistic T^* from Eq. (4), or equivalently, when using the bootstrapped p-values from Eq. (5) for testing the null hypothesis (Bollen and Stine, 1992). It can be shown that the asymptotic distribution of T^* is not the same as that for T. Bollen and Stine (1992) gave an approximation for the unconditional asymptotic expectation of the test statistic T^*. They report it as being twice as large as the asymptotic expectation of T in the original sample. They also report the unconditional asymptotic variance of T^* to be larger than the asymptotic variance of T. However, we think that their derivations lack in theoretical foundations, since it is not clear that the asymptotic conditional variance of T^* equals $2df + 4T$. For this reason we performed simulation studies (shown in the Supplementary Material) to assess the validity of their results. Our empirical results are in line with the theoretical results of Bollen and Stine (1992). The probability mass in the tail of the distribution of T^* is larger than that of T, which leads to increased type I error for LR tests performed on bootstrap samples.

Since the marginal distribution of T^* is unknown, there is no straightforward derivation of the type I error for the LR test. To assess the increase in type I error for the LR test in a practical application, we performed empirical studies based on the original and permuted NHANES data, where we univariately tested the association between the CRP level and each of the 28 covariates by means of an LR test. The LR test was performed to test if the full model containing the intercept and covariate X_j, $j \in \{1, 2, \ldots, 28\}$ gives a better fit than the submodel containing only the intercept. An association was considered significant if the p-value was equal to or less than 0.05. Figure 1 shows the relative frequencies of significant associations in the B bootstrap samples for the unpermuted (left) and the permuted (right) NHANES data. For the unpermuted NHANES data on average (taken over $B = 10{,}000$ bootstrap samples), there were 18.4 significant associations, while in the original data 17

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

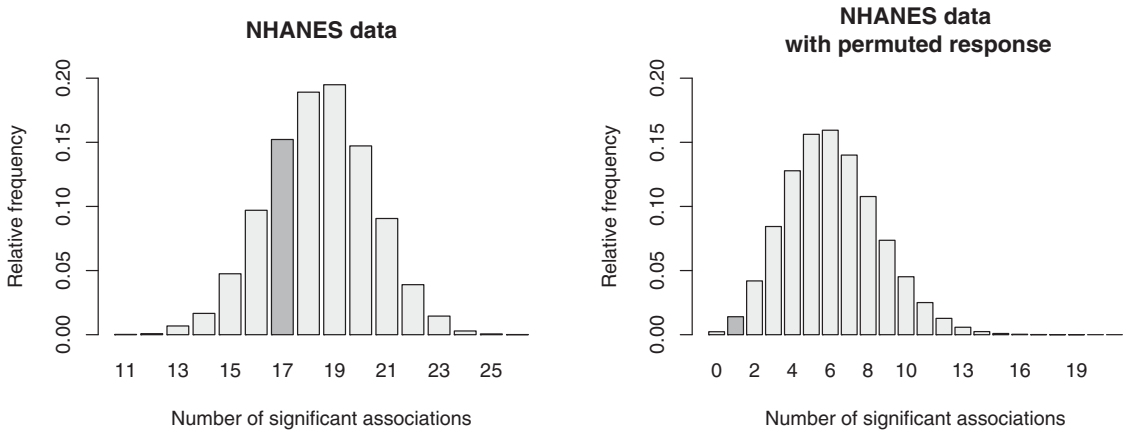

Figure 1 Relative frequency of bootstrap samples with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The total number of bootstrap samples was 10,000 for the unpermuted NHANES data, and 10,000 × 1000 for the permuted NHANES data. The dark gray bar indicates the number of significant associations in the unpermuted NHANES data (left) and the average number of significant associations in the 1000 permuted NHANES data sets (right).

of the 28 associations were significant. For the permuted NHANES data there were on average 1.36 significant associations (over 1000 original samples) and the average number (taken over all 1000 × B bootstrap samples) of significant associations according to bootstrapped p-values was 6.12.

We performed the same computations using subsamples instead of bootstrap samples, with results shown in Fig. 2. From theory it is clear that p-values obtained from subsamples systematically deviate from p-values obtained for the original sample due to the smaller sample size and the decreased power

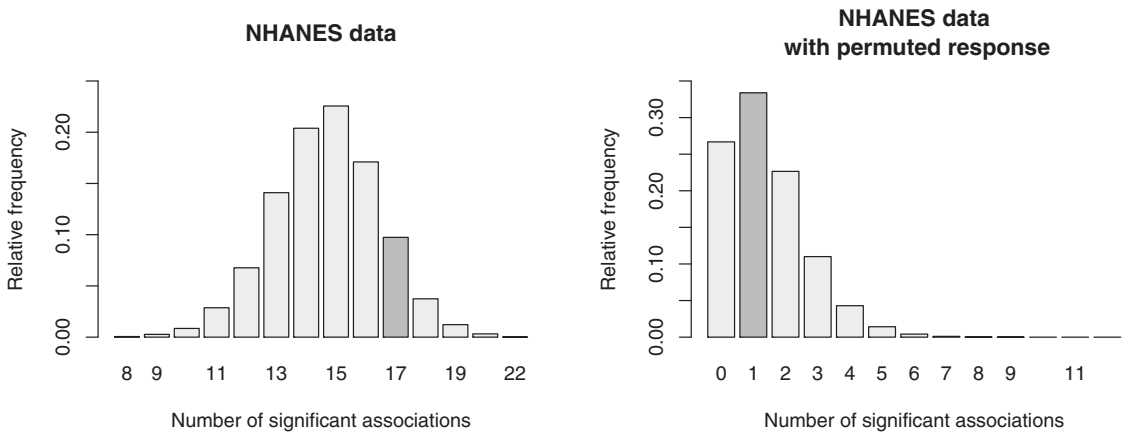

Figure 2 Relative frequency of subsamples with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The total number of bootstrap samples was 10,000 for the unpermuted NHANES data, and 10,000 × 1000 for the permuted NHANES data. The dark gray bars indicate the number of significant associations in the unpermuted NHANES data (left) and the average number of significant associations in the 1000 permuted NHANES data sets (right).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to detect associations in subsamples: this is clearly seen in Fig. 2. On average 14.7 of the 28 covariates were significantly associated with the CRP level in subsamples compared to 17 significant associations in the original sample.

In the case where no associations exist—the NHANES data with permuted response—a comparable number of significant findings can be observed in subsamples and in the 1000 original samples: there were on average 1.40 significant associations in subsamples compared to 1.36 significant findings in the 1000 original samples. This is in line with the fact that tests performed on subsamples—in contrast to tests performed on bootstrap samples—do not have increased type I error as shown, for example, by Sauerbrei et al. (2011). Accordingly, p-values derived on subsamples may be used for testing a specific hypothesis.

3.1.2 Distribution of bootstrapped p-values

Note that in Section 3.1.1 we considered the marginal distribution of the bootstrapped test statistic to prove that the type I error is increased when using bootstrapped p-values. However, a marginal representation does not give any information on the distribution of bootstrapped p-values for a specific sample $\boldsymbol{x}$. Moreover, it does not provide any information on whether the bootstrapped p-value can be expected to be similar to the p-value of an observed sample $\boldsymbol{x}$. In this subsection, we aim to address these issues. For this purpose, let us consider the setting of normally distributed variables and the null hypothesis which states that the population mean equals μ_0. Now let Z be the test statistic computed based on the observed sample $\boldsymbol{x}$, and let Z^* be the bootstrapped test statistic which follows a $N(Z, 1)$ distribution conditional on $\boldsymbol{x}$ (cf. Section 2.2.1). Figure 3 shows distributions for $Z^*|\boldsymbol{x}$, that is, the distributions are conditional on the sample $\boldsymbol{x}$. Conditional distributions are shown for three realizations of $\boldsymbol{x}$ with corresponding absolute Z values, namely (A) a large absolute Z value, (B) a small absolute Z value and (C) an intermediate absolute Z value. From this illustration, it can be seen that the distribution of bootstrapped p-values—and with that, the discrepancy between bootstrapped p-values and the original p-value—depends on the realized sample and the respective test statistic Z:

(A) If the observed $|Z|$ is large (upper panel of Fig. 3), there is approximately a 50% chance of having a bootstrapped p-value, p^*, which is larger than p, the observed p-value based on $\boldsymbol{x}$ (indicated by the dark gray area), and a 50% chance of having a smaller p^* (light gray area). In this scenario, we would consider p^* to be a good approximation of p.

(B) If the observed value for $|Z|$ is close to 0 (middle panel), Z^* follows approximately a standard normal distribution, and the bootstrapped p-values are uniformly distributed on [0, 1]. We thus expect a p^* of 0.5 (i.e., the expectation or median of a variable $U \sim U[0, 1]$). However, p is 1. In cases where Z is close to 0, the bootstrapped p-value is obviously not a good approximation of the p-value of the original sample.

(C) If $|Z|$ takes an intermediate value, say, 1, we have a similar situation to (A). However, in contrast to (A), there is a moderate probability for negative Z^* values smaller than $-|Z|$, or, in mathematical terms, $Pr(Z^* < -|Z||\boldsymbol{x})$ is much larger than 0 and cannot be ignored. Therefore, the probability of obtaining $p^* < p$ is greater than obtaining $p^* > p$. This shows that bootstrapped p-values are not a good approximation of the p-value of the original sample if $|Z|$ takes an intermediate value: in over 50% of the bootstrap samples we would expect p^* smaller than p.

To summarize, the smaller the $|Z|$ (or, the larger the p), the greater the difference between the median bootstrapped p-value and p. As $|Z|$ tends to infinity (or, p tending to 0), the difference becomes smaller. Empirical studies support these findings (not shown). Note, however, that these considerations are for the more commonly used two-sided test, but do not generalize to the one-sided Z-test as is shown in the Supplementary Material.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

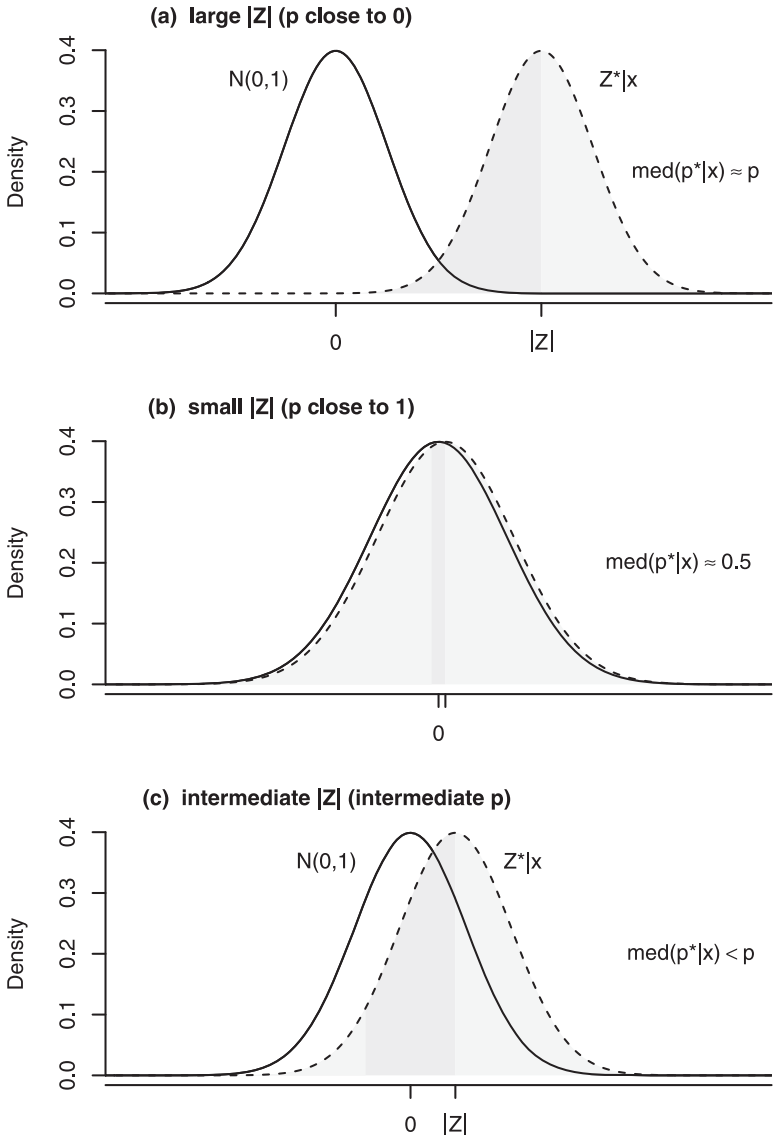

Figure 3 Conditional distribution of Z^* for a fixed sample $\boldsymbol{x}$ with observed test statistic Z. Three different scenarios are considered: (A) $|Z|$ is large, (B) $|Z|$ is small, and (C) $|Z|$ is intermediate. The standard normal distribution is indicated by the solid black line. The light (dark) gray area represents the bootstrapped test statistics Z^* with corresponding bootstrapped p-value smaller (larger) than the p-value derived for the observed test statistic Z.

One might argue that it was already shown in Section 3.1.1 that bootstrapped p-values do not approximate the original p-values very well. It is important to note that the increased type I error does not imply that bootstrapped p-values are a poor approximation of original p-values. For the one-sided Z-test, for example, tests performed using bootstrapped p-values have increased type I error, but bootstrapped p-values are a good approximation of the originals (see the Supplementary Material).

The considerations made for the two-sided Z-test apply to the LR test in a similar way. Let us assume for the moment that the null hypothesis (that the submodel is true) holds and that the LR test statistic

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

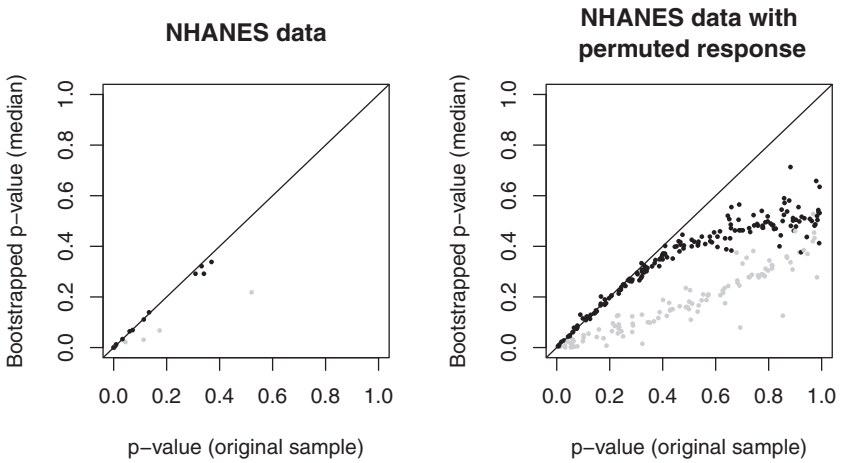

Figure 4 Median p-values obtained for testing the association between CRP level and each of the 28 covariates in 10,000 bootstrap samples, plotted against the p-values of the original sample, for the NHANES data. Black points represent the p-values for LR tests with 1 degree of freedom, and gray points correspond to LR tests with 3 or more degrees of freedom. Points lying on the diagonal line would indicate agreement between p-values derived on the original NHANES data and bootstrapped p-values. Left: Results obtained for the unpermuted NHANES data. Right: Results obtained for 10 permuted NHANES data sets in which there are no true associations between covariates and the CRP level (via permuting values for CRP level).

T equals zero for a given sample $\boldsymbol{x}$, which means that in the original data the derived likelihood of the submodel is exactly equal to the likelihood of the unrestricted model. Then the bootstrap samples are drawn from a distribution in which H_0 is true. Accordingly, the bootstrapped test statistic T^* follows a central χ^2-distribution and the bootstrapped p-value is uniformly distributed on [0, 1]. As with the Z-test, the median and expectation of the bootstrapped p-value is 0.5, while the p-value for the original data is 1. As with the two-sided Z-test, bootstrapped p-values for the LR test thus cannot be expected to be close to p-values computed on the original data.

Since the distribution of $T^*|\boldsymbol{x}$ is unknown, it is difficult to explore the distribution of bootstrapped p-values dependent on different values of T by theoretical arguments, as we have done for the Z-test. Therefore, we further investigated the discrepancy between bootstrapped p-values and original p-values using empirical studies of the NHANES data. We performed LR tests for each of the 28 covariates to test the null hypothesis (the submodel containing only the intercept is true) against the alternative hypothesis (the model containing the intercept plus the respective covariate is true). This was done for the original data as well as for $B = 10{,}000$ bootstrap samples and $B = 10{,}000$ subsamples drawn with and without replacement, respectively, from each original data set. As original data sets we used both the unpermuted NHANES data and 1000 permuted NHANES data sets.

Figure 4 (left) shows the median bootstrapped p-values for each of the 28 covariates plotted against the p-values obtained for the original sample. Black points represent the p-values for LR tests with 1 degree of freedom (performed for metric and binary covariates), while the gray points correspond to tests with 3 or more degrees of freedom (for categorical covariates with 4, 5, 6 or 12 levels). For the sake of clarity we show the results only for the first 10 permuted NHANES data sets in the right panel in Fig. 4. Note that since we have 10 data sets, 10×28 points are plotted. For LR tests with 1 degree of freedom we observe a similar situation to that for the Z-test: when the p-value is small, it is approximated well by the median bootstrapped p-value; however, for large p-values the approximation

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

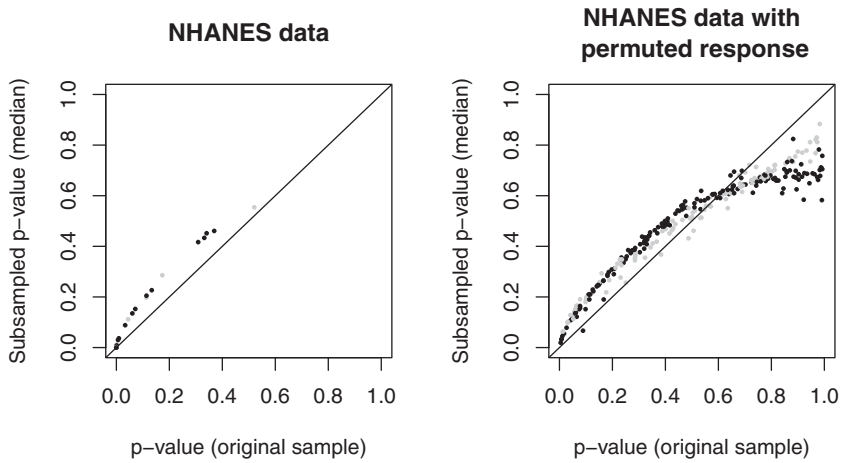

Figure 5 Median *p*-values obtained for testing the association between CRP level and each of the 28 covariates in 10,000 subsamples, plotted against the *p*-value of the original sample, for the NHANES data. Black points represent the *p*-values for LR tests with 1 degree of freedom, and gray points correspond to LR tests with 3 or more degrees of freedom. Points lying on the diagonal line would indicate agreement between *p*-values derived on the original NHANES data and *p*-values derived on subsamples. Left: Results obtained for the unpermuted NHANES data. Right: Results obtained for 10 permuted NHANES data sets in which there are no true associations between covariates and the CRP level (via permuting values for CRP level).

is not good. For LR tests with 3 or more degrees of freedom it seems bootstrapped *p*-values are never a good approximation, independent of whether the original *p*-values are small or large.

Figure 5 shows the median subsampled *p*-values plotted against the *p*-values obtained for the original sample. It can be observed that subsampled *p*-values were larger than *p*-values for the original sample if the latter were in the range $[0, 0.6]$. If *p*-values obtained for the original sample were above 0.6, subsampled *p*-values were smaller. It is clear that due to a different sample size subsampled *p*-values are not a good approximation of *p*-values for original samples.

3.2 Application 1: Bootstrapped *p*-values for variable ranking

In the following, we investigate the consequences when using bootstrapped *p*-values for ranking variables. Such an approach has been proposed by Mukherjee et al. (2003) for ranking genes with respect to their differential expression. The approach consists of computing the *p*-values for a large number of bootstrap samples, obtaining the median *p*-value for all considered genes, and sorting the genes by the median *p*-value. In the following, we apply this approach to the NHANES data to obtain a ranking of the 28 considered variables.

Figure 6 shows the variable rankings for the unpermuted NHANES data. The upper left panel corresponds to the rankings according to the *p*-values from the original NHANES sample and the upper right panel corresponds to rankings by the median bootstrapped *p*-values (i.e., the median of $B = 10{,}000$ bootstrapped *p*-values). In addition, results are shown when using the median *p*-value obtained from $B = 10{,}000$ subsamples (lower panel).

On the whole, the rankings are similar for the unpermuted NHANES data, especially among those variables with strong evidence for association. However, close inspection reveals some differences between the rankings based on the original sample and those based on bootstrap samples. More precisely, we observed the trend—in line with the results in Rospleszcz et al. (2014)—that categorical

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

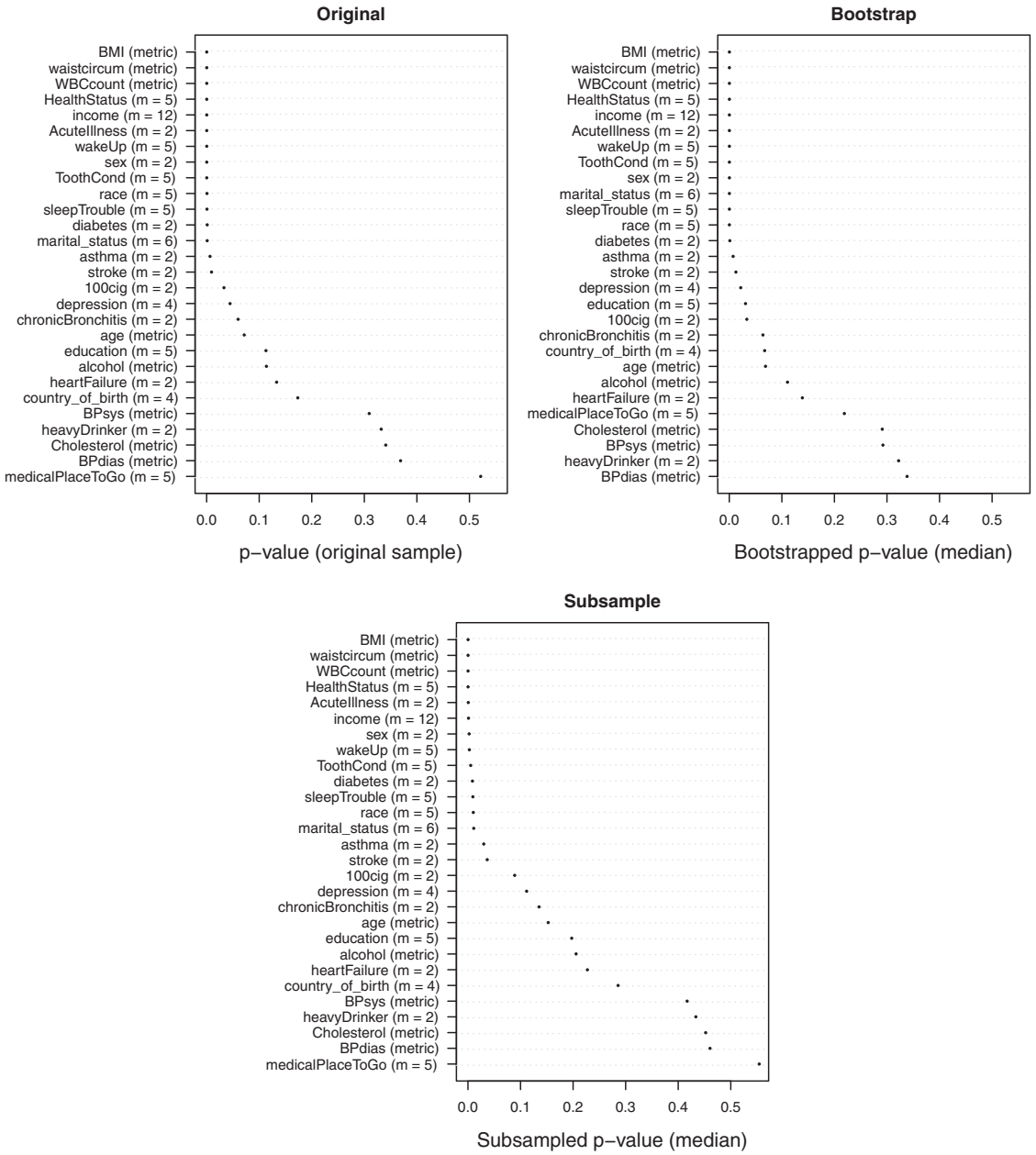

Figure 6 Variable ranking by p-values obtained for the original NHANES sample (upper left) and by the p-value obtained from the median over $B = 10{,}000$ bootstrapped p-values (upper right) and the median p-value from subsamples (lower). The parameter m denotes the number of levels of a categorical predictor variable.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Variable ranking for the unpermuted NHANES data.

Scale	Variable	Original	Bootstrap		Subsample	
		rank	rank (difference)		rank (difference)	
Metric or $m = 2$	BMI	1	1	(0)	1	(0)
	waistcircum	2	2	(0)	2	(0)
	WBCcount	3	3	(0)	3	(0)
	AcuteIllness	6	6	(0)	5	(+1)
	sex	8	9	(−1)	7	(+1)
	diabetes	12	13	(−1)	10	(+2)
	asthma	14	14	(0)	14	(0)
	stroke	15	15	(0)	15	(0)
	100cig	16	18	(−2)	16	(0)
	chronicBronchitis	18	19	(−1)	18	(0)
	age	19	21	(−2)	19	(0)
	alcohol	21	22	(−1)	21	(0)
	heartFailure	22	23	(−1)	22	(0)
	BPsys	24	26	(−2)	24	(0)
	heavyDrinker	25	27	(−2)	25	(0)
	Cholesterol	26	25	(+1)	26	(0)
	BPdias	27	28	(−1)	27	(0)
$m = 4$	depression	17	16	(+1)	17	(0)
	country_of_birth	23	20	(+3)	23	(0)
$m = 5$	HealthStatus	4	4	(0)	4	(0)
	wakeUp	7	7	(0)	8	(−1)
	ToothCond	9	8	(+1)	9	(0)
	race	10	12	(−2)	12	(−2)
	sleepTrouble	11	11	(0)	11	(0)
	education	20	17	(+3)	20	(0)
$m = 6$	medicalPlaceToGo	28	24	(+4)	28	(0)
	marital_status	13	10	(+3)	13	(0)
$m = 12$	income	5	5	(0)	6	(−1)

Notes: Variable rankings are obtained from p-values obtained for the original NHANES sample ("Original rank"), from the median bootstrapped p-value ("Bootstrap rank"), and from the median p-value from subsamples ("Subsample rank"). The difference to the "Original rank" is given in brackets for each variable. The parameter m denotes the number of levels of a categorical predictor variable.

predictors with many categories obtained systematically smaller bootstrapped p-values than metric predictors or categorical predictors with fewer categories. In our variable ranking, this can be seen from the fact that variables with many categories gain ranking positions closer to the top when ranked by the median bootstrapped p-value. Table 2 shows the ranking positions for each variable separately for variables of different scale. There are numerousness cases in which categorical variables with four or more categories gain ranking positions closer to the top when ranked by bootstrapped p-values. Conversely, the binary and metric variables are located at positions at the bottom of the ranking when the ranking is according to bootstrapped p-values.

In contrast, when using subsamples there are only minor differences in the ranking, with seemingly no effect of a variable's scale on its ranking position. This is also in line with the results presented in

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Rospleszcz et al. (2014) who investigated the use of subsampling as an alternative to bootstrap for a model building procedure.

The observed mechanisms are even more extreme for the permuted NHANES data sets where the response variable had been permuted (results shown in the Supplementary Material). For the permuted data sets there are very large differences in the variable ranking—with variables with many categories ranked at top positions and binary or metric variables at much lower positions when p-values are derived from bootstrap samples.

To conclude, our studies show that, though resampling procedures might be promising methods for obtaining stable variable ranking lists, bootstrapped p-values should not be compared with significance thresholds for making decisions on the significance of variables. In particular, care needs to be taken when the interest lies in ranking variables of different scales, which often occurs in epidemiological studies. An example of further relevance is gene ranking when single nucleotide polymorphisms are considered, which for some genes are represented by a categorical variable with three categories but for others only two categories. Moreover, associations between genes and a phenotype are usually weak or nonexistent, which is expected to be especially problematic as suggested by the results of the permuted NHANES data. Thus, in settings including categorical predictors bootstrapped p-values should not be applied for obtaining ranking lists.

Subsampling may be a reasonable alternative to the bootstrap for variable ranking: in our studies there were only minor differences between the ranking lists that were obtained by sorting variables by p-values obtained from the original sample and from subsamples. This might indicate that in the considered NHANES data set there are not many influential points that have a large impact on the results, but more research is needed on this topic. We conclude from these results that subsampling should be preferred over bootstrapping for obtaining variable rankings if variables are of different scales. We have to note, however, that in settings with very small sample sizes—for which the ranking approach was originally proposed (Mukherjee et al., 2003)—subsampling from a data set that consists of only a few observations may not be advisable.

3.3 Application 2: Bootstrapped p-values for assessing the variability of p-values

Recently, it has been proposed to compute the variance of p-values, or preferably the variance of $-\log_{10}(p\text{-value})$ (Boos and Stefanski, 2011). The question arises of whether the variance of bootstrapped p-values can be used to approximate the variability of p-values that would be observed if we repeatedly performed the same experiment. To investigate this issue, we performed simulation studies, which allow us to draw multiple times from the true distribution F.

In the first part of our simulation studies, we independently drew $n = 1000$ observations from $N(0, 1)$ and tested $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. A total of $B = 10{,}000$ bootstrap samples were generated by drawing from this data with replacement. A Z-test was performed for the original sample and for each bootstrap sample. Subsequently, we computed the standard deviation of the $B = 10{,}000$ bootstrapped p-values and the standard deviation of the negative logarithm of the bootstrapped p-values. This process was repeated 10,000 times and the standard deviation of the p-values and that of the negative logarithm of the p-values for the 10,000 original data sets were computed. The same analysis was done for the case where the n observations came from $N(0.08, 1)$.

In the second part of our simulation studies, 10 metric predictor variables $x_{i1}, \ldots, x_{i10}$ were independently drawn for $i = 1, \ldots, 1000$ from a multivariate normal distribution with expected value $\boldsymbol{\mu} = (0, \ldots, 0)^\top \in \mathbb{R}^{10}$ and variance $\boldsymbol{I}_{10}$, corresponding to the identity matrix of dimension 10. The response variable Y_i was generated according to the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

with $\epsilon_i \sim N(0, 1)$. The global null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$ states that none of the predictors is associated with the response, and the alternative hypothesis is that at least one of the

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

predictors is associated, that is, $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, 10\}$. The corresponding LR test compares the likelihood of the submodel L_0 which contains only the intercept, to the likelihood L_1 of the model which contains all predictor variables. If the null hypothesis is true, the LR test statistic (3) follows a central χ^2-distribution with 10 degrees of freedom. In our simulations, all β-coefficients were set to the value zero (i.e., the null hypothesis is true). As before, p-values and their standard deviations were derived. An additional analysis was performed in which the alternative hypothesis is true. For this simulation, all coefficients were set to 0.02.

Figure 7 shows the distributions of the standard deviations of the bootstrapped p-values and the standard deviations of the negative logarithm of the bootstrapped p-values for the first (Z-test) and the second simulation studies (LR test). The dotted line represents the standard deviation of the p-values or negative logarithm of the p-values computed from the 10,000 original samples. Results are also shown for subsampling. For the Z-test under H_0, we observe a systematic but probably negligible difference between the standard deviations of p-values for the original data and those of bootstrapped p-values (or $-\log_{10}(p\text{-value})$). Under H_1, in contrast, the standard deviation of bootstrapped p-values (or $-\log_{10}(p\text{-value})$) seems to be a good approximation of the true p-value variability. For the LR test with 10 degrees of freedom, the standard deviations of the p-value (or $-\log_{10}(p\text{-value})$) computed on bootstrap samples do not reflect the true p-value variability in our studies, neither under the null hypothesis nor under the alternative hypothesis. This result was expected in the light of the empirical results presented in Section 3.1.2 (Fig. 4), which showed that for LR tests with 3 or more degrees of freedom the distribution of bootstrapped p-values is always different from the true p-value distribution. Subsampling is not a reasonable alternative here since tests performed on subsamples do not reflect the p-value variability of the original samples either (Fig. 7).

4 Studies on bootstrapping information criteria

4.1 General results

Due to the correspondence between the LR test statistic and the AIC in the specific setting of nested models (cf. Section 2.3), it follows from Section 3.1.1 that bootstrapped information criteria like the AIC are thus not valid either. These considerations were also made by Wagenmakers et al. (2004).

The AIC is often used in model selection when choosing an appropriate model. Though it was shown that the AIC is biased (Steck and Jaakkola, 2003; Wagenmakers et al., 2004), it is unknown if this bias impacts the decision for a model. Thus, one might argue that the bias is not of any practical interest if it does not affect the model choice. To investigate if model choice is affected, we performed some experiments using the NHANES data. With 28 covariates in the NHANES data, there are $2^{28} = 268{,}435{,}456$ candidate models and due to computational effort it is not practicable to consider all. Though one usually considers models that include more than one covariate, for ease of illustration we narrowed it down to the 28 models each arising from the inclusion of exactly one of the covariates and investigated which of the models provides the best fit according to the AIC and bootstrapped AIC. Bootstrapped AIC values were computed for $B = 10{,}000$ bootstrap samples and an average AIC value was computed.

Figure 8 shows the difference between the AIC values computed on the original NHANES sample and the average bootstrapped AIC value. The difference seems to be bigger for models that include more parameters. Though all models have a systematically smaller bootstrapped AIC value, those models incorporating larger numbers of parameters have an exceedingly small AIC value. There are three exceptions: the model featuring *WBCcount*, that for *BMI* and that for *waistcircum*. Note that these models are the models with the best model fit according to the AIC. We expect that the observed phenomenon that models incorporating larger numbers of parameters have an exceedingly small AIC value will lead to a preferential selection of more complex models. We check our assumptions by

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

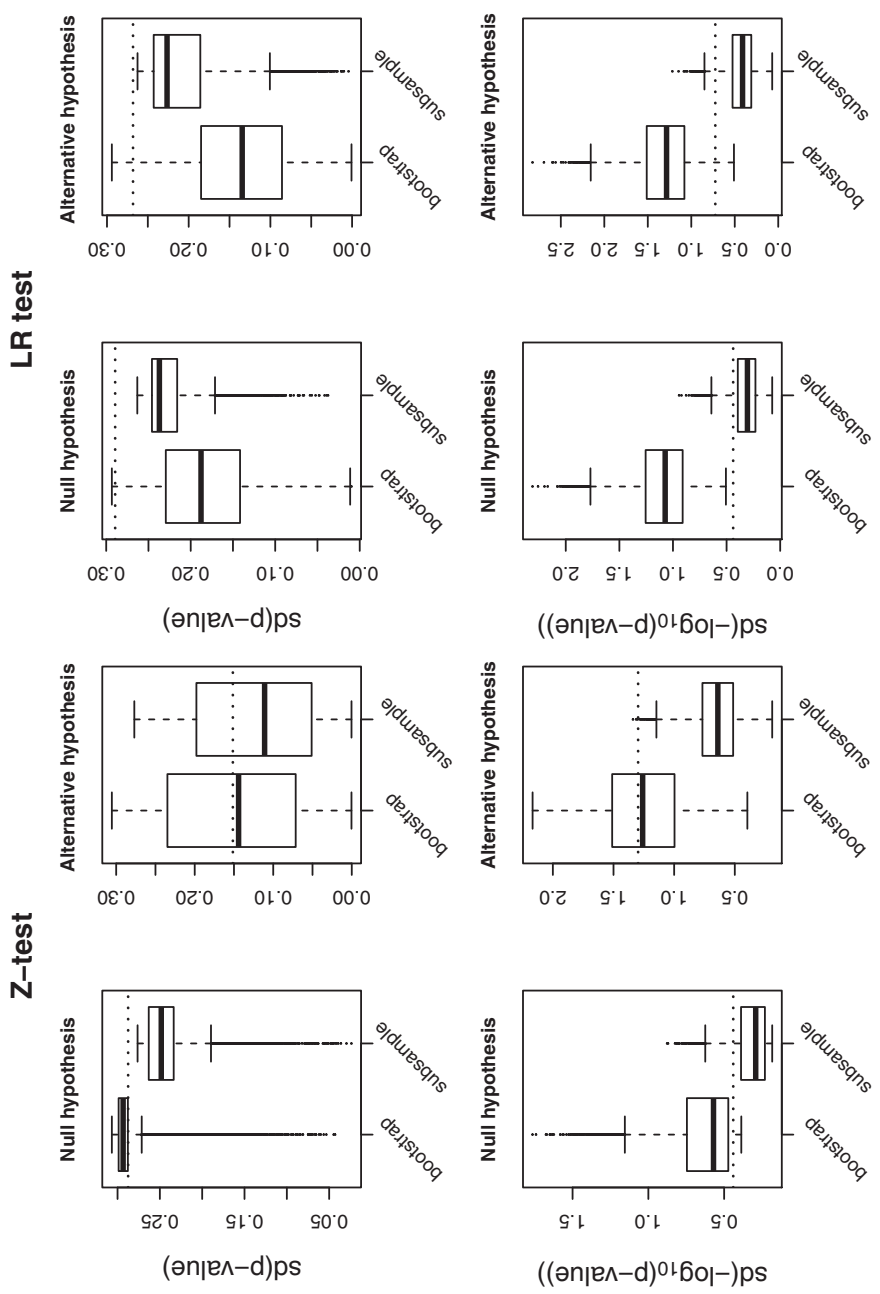

Figure 7 Standard deviations (sd) of bootstrapped and subsampled p-values or $-\log_{10}(p\text{-value})$ for the Z-test (left two columns) and the LR test with 10 degrees of freedom (right two columns). The dotted line represents the standard deviation of the p-values or negative logarithm of the p-values, computed from the 10,000 original samples. Each original sample gave rise to 10,000 bootstrap samples and 10,000 subsamples.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

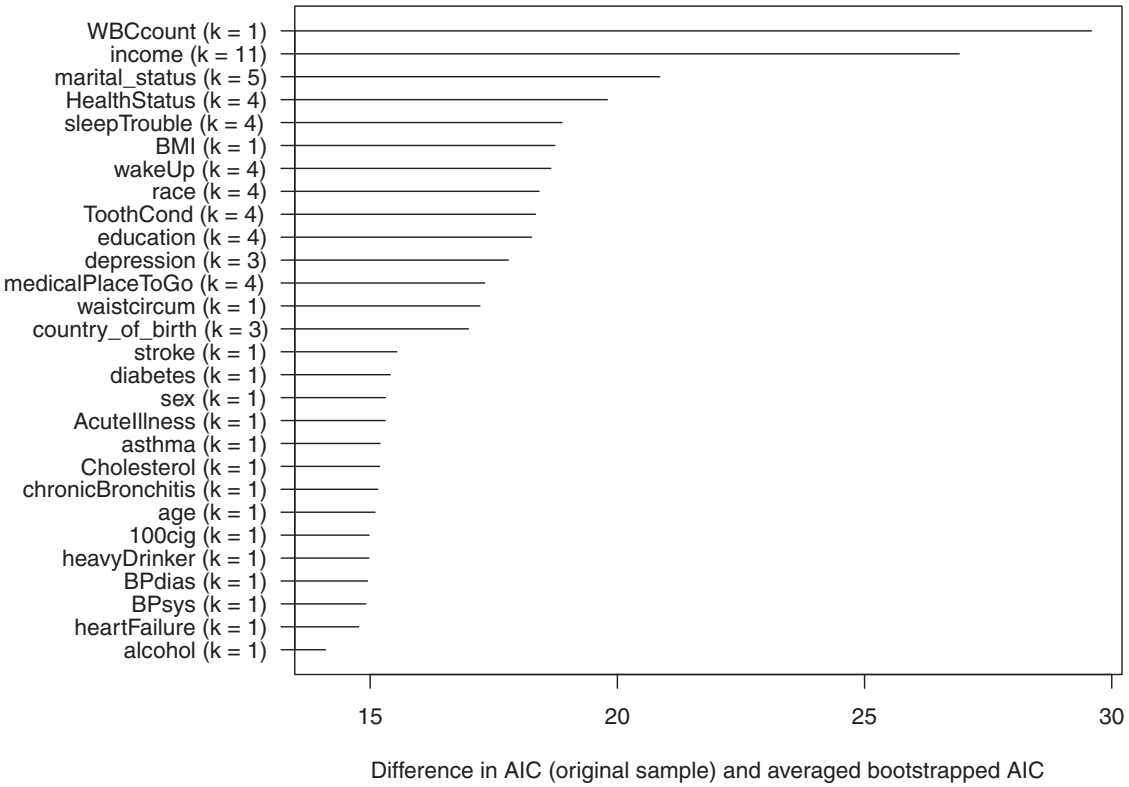

Figure 8 Difference between the AIC value computed on the original NHANES sample and the AIC value obtained from averaging over $B = 10{,}000$ bootstrapped AIC values for 28 univariate models. The parameter k denotes the number of parameters estimated for the respective variable in the univariate linear model.

ranking the models according to their average bootstrapped AIC and according to the AIC value obtained for the original sample.

The left panel of Fig. 9 shows the ranking of models by AIC value obtained for the original sample. The right panel of Fig. 9 shows the ranking by this average bootstrapped AIC. While the top and the bottom of the ranking lists are nearly identical, a number of differences can be observed in the middle: the model which includes $k = 5$ parameters coding marital status is ranked at the 12th position based on the original NHANES sample, while based on bootstrap samples it is ranked 9th. Conversely, the model which includes the variable *sex* ($k = 1$) is ranked 9th based on the original sample but only 12th when AICs were derived from bootstrap samples. Considerable differences in the ranking position can also be observed for the model which includes educational background ($k = 4$). For the original sample this model is ranked only 22nd, while for bootstrap samples it is ranked 17th. Overall, when looking at both rankings, one can see that models which include more parameters seem to obtain higher rankings when ranked by bootstrapped AICs. This applies for the models based on the covariates *wakeUp*, *sleepTrouble*, *marital_status*, *depression*, *education*, or *country_of_birth*. Models which include only one parameter (in addition to the intercept) have lower rankings for bootstrapped AICs (for covariates: *sex*, *acuteIllness*, *100cig*, *chronicBronchitis*, *age*, *alcohol*, *heartFailure*, *BPsys*, *heavyDrinker*). There are only two exceptions where it is reverse (*Cholesterol* and *race*). These results strongly suggest that there is a preferential selection of more complex models—that is, those that include more parameters—when using bootstrapped AICs.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

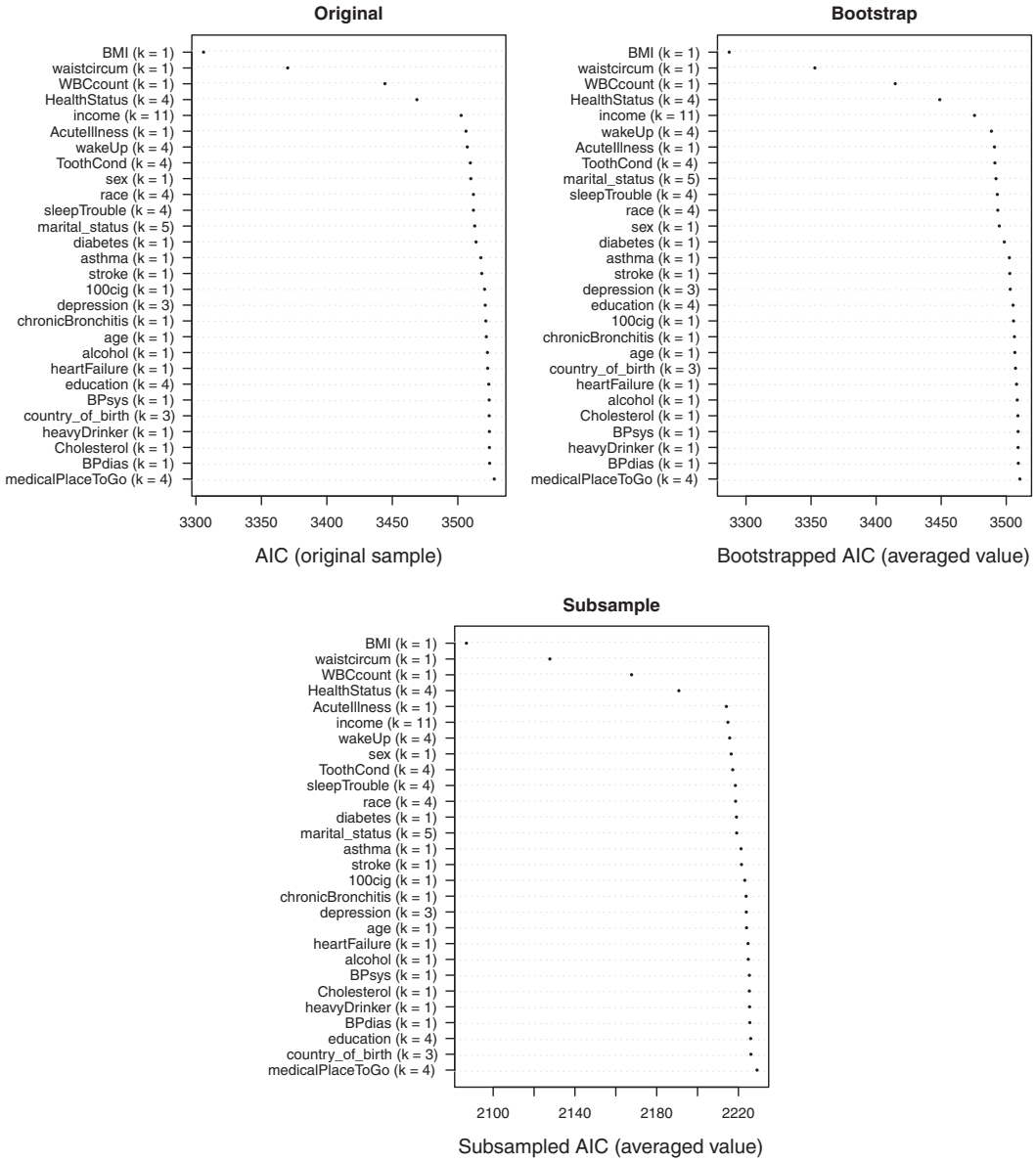

Figure 9 AIC values (in ascending order from top to bottom) obtained for the 28 models (each including exactly one covariate). The parameter k denotes the number of parameters included in the model for the respective variable. Upper left: AIC values derived on the original NHANES sample. Upper right: AIC values obtained from averaging over $B = 10{,}000$ bootstrapped AIC values. Lower: AIC values obtained from averaging over $B = 10{,}000$ AIC values computed based on subsamples.

Results were also obtained when using subsamples instead of bootstrap samples. Since subsamples contain fewer observations, AIC values obtained for models on subsamples are not comparable to those obtained for the original sample. However, it is interesting to explore if the decision for or against a model is different when the AIC is computed on subsamples instead of the original sample. This can again be seen when sorting the models according to their AIC values (Fig. 9, lower panel).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Indeed there are some characteristic changes in the ordering of the models according to the average AIC obtained from subsamples. But in contrast to the bootstrap, it seems as if more complex models (in terms of included parameters) are rather disfavored (see also results in the Supplementary Material). This can be explained as follows: From the definition of the AIC in Eq. (6), we can see that the AIC is dominated by the penalty term $2p$ (which penalizes the complexity of the model) if the first term $-2\log(L)$ is small, or equivalently, if the likelihood is large. Conversely, the AIC is dominated by the first term, $-2\log(L)$ (which is a measure of the model fit to the data), if the likelihood is small. The likelihood, as a product of n probabilities, becomes automatically smaller with increasing n. As a consequence the likelihood derived from a subsample is larger than the likelihood of the original sample. From these considerations, it is clear that for subsamples the AIC is more driven by the penalty term than for the original sample, which leads to the observed phenomenon that more complex models are more disfavored in subsamples than in the original sample.

To conclude, AICs obtained from subsamples and original samples do not lead to the same conclusion regarding the choice of optimal models as well.

4.2 Application 3: Bootstrapped information criteria for model selection

In this section, we investigate whether there is a preference for more complex models (in terms of included parameters) when constructing models based on bootstrap samples in the special context of gradient boosting (Friedman, 2001; Hothorn et al., 2010). Gradient boosting has become a popular method in biometrical applications to find sparse models by only making use of relevant predictor variables, which greatly facilitates model interpretation. Briefly the idea of gradient boosting algorithms is to combine weak learners in an iterative fashion to obtain a strong learner with high prediction accuracy. The prediction accuracy depends highly on the number of iterations, also called the number of boosting steps. With too many boosting steps, many weak learners are constructed and the resulting strong learner might be overfit to the data and thus have poor prediction accuracy on new data. If the number of boosting steps is too small, the number of weak learners might be too small to appropriately model the relationship between the covariates and the response. Thus, the number of boosting steps has to be carefully chosen, for example, through application of information criteria or internal cross-validation. For more details on gradient boosting we refer the reader to the literature.

In the following analysis, we applied the gradient boosting method firstly to the original NHANES sample and then to bootstrap samples. Note that in contrast to the earlier analysis we here modeled the association between CRP level and the covariates in a multivariate fashion. We used the AIC for choosing the number of boosting steps.

For the original NHANES sample, the number of boosting steps for the model with the smallest AIC was 309, the result being a model of 42 parameters (not including the intercept term). When performing tuning parameter selection on bootstrap samples we obtained systematically larger values for the number of boosting steps: in almost all (978 of $B = 1000$) bootstrap samples the chosen number of boosting steps was greater than 309 (see left boxplot in Fig. 10). The mean number of boosting steps in bootstrap samples was 468. The resulting models included a larger number of parameters on average: the average number was 44.3, two parameters more than the model which was obtained for the original NHANES sample. The left panel in Fig. 11 shows the relative frequency of models with a specific number of parameters. In 68.3% of the bootstrap samples, the model included more than 42 parameters, in 24.7% the number of parameters was lower and in 7% the models included exactly 42 parameters.

We performed the same calculations using subsamples instead of bootstrap samples. As one would expect, sparser models were selected (on average 34.7 parameters) than for the original sample or bootstrap samples (right panel in Fig. 11). Or equivalently, for subsamples a smaller number of boosting steps (254 on average) was chosen, seen in Fig. 10 (right boxplot).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

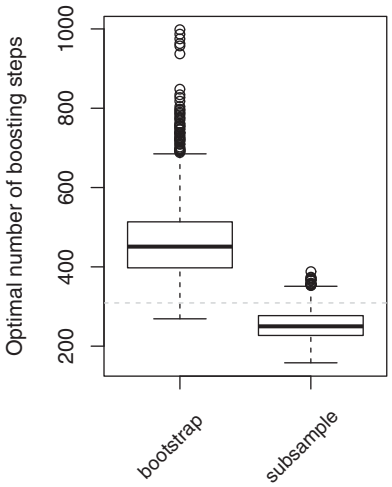

Figure 10 Optimal number of boosting steps selected via AIC in $B = 1000$ bootstrap samples and subsamples of the NHANES data. The dashed horizontal line indicates the chosen number of boosting steps in the original NHANES data.

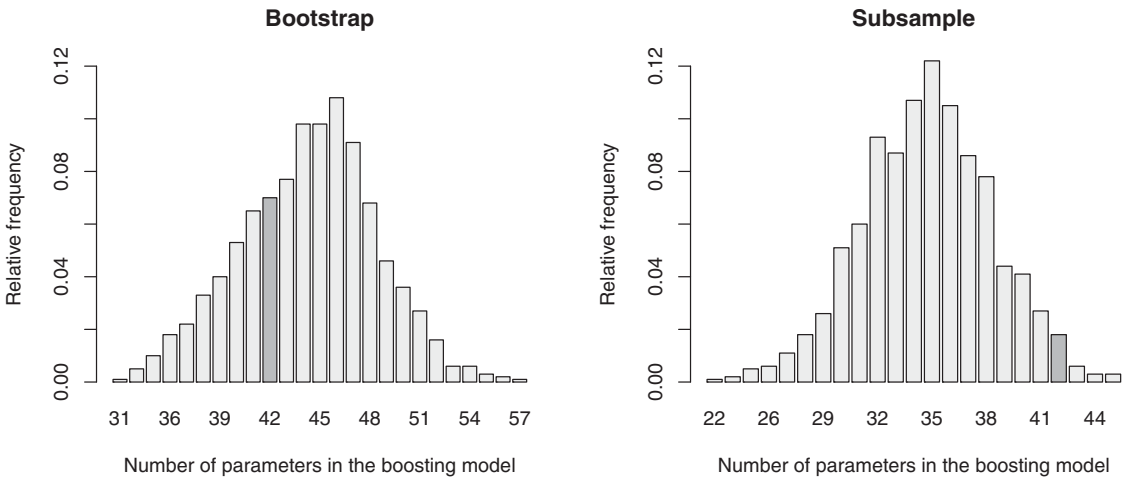

Figure 11 Relative frequency of boosting models (out of $B = 1000$) fitted on bootstrap samples (left) and on subsamples (right) with specified number of parameters (not including the intercept term). The dark gray bars indicate the number of parameters in the model that was fit on the original NHANES sample.

We also evaluated the models with respect to their predictive accuracy, using the observations that were not drawn into the bootstrap sample and subsample, respectively. Though models constructed on subsamples included fewer parameters, their predictive accuracy was comparable to the accuracy of models constructed on bootstrap samples: on average, even a marginally smaller mean squared error was obtained for models fit on subsamples (0.00075 compared to 0.00085 when using bootstrap samples), which suggests that the additional parameters in the models from bootstrap samples do not have any additional predictive value.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5 Discussion and outlook

In this paper, we applied selected bootstrap-based approaches (making either use of bootstrapped *p*-values or bootstrapped information criteria) on a large real data set from a population-based study to investigate whether results are affected by the systematic deviation in bootstrapped *p*-values and information criteria. When univariately testing the association between the level of high-sensitive CRP and various factors, we observed that bootstrapped *p*-values are often considerably smaller than *p*-values that are obtained for the original data. Also seen in our studies was that making decisions based on bootstrapped *p*-values results in increased number of false positive results. Further, the variability of bootstrapped *p*-values was shown not to reflect the variability of *p*-values for the LR test when repeating the same experiment several times, thus making the reliability of the approach suggested by Boos and Stefanski (2011) questionable.

We also observed a bias in bootstrapped information criteria when these were compared to information criteria that were derived from the original sample. In our studies, this led to a preferential selection of models which included more parameters, since these models systematically had smaller bootstrapped AIC values. Further, bootstrapped AIC values are sometimes used in the context of gradient boosting models. Here the tuning parameter selection (via AIC) and model fitting is performed based on a bootstrap sample while the remaining observations that were not drawn into the bootstrap sample are used for evaluating the model. In our application on real data, we observed higher values for the tuning parameter for bootstrap samples. This led to more complex boosting models (i.e., more parameters) than the model fit on the original sample. These results are in line with those reported by Steck and Jaakkola (2003) in the context of graphical models who show that more complex models (in terms of included parameters), have actually too high a likelihood, or equivalently, too small an AIC value, when fit on bootstrap samples. Thus, when using the AIC to select a model that was built from a bootstrap sample, one gives preference to more complex models that would possibly not have been selected had the original sample been used.

We also investigated the use of subsampling as a promising alternative strategy to circumvent problems induced by the bootstrap. The properties of subsampling have been theoretically investigated in the literature; it has been shown that subsampling has desirable properties even in situations where the bootstrap fails (see, e.g., Horowitz, 2001, and references therein for issues related to the inconsistency of the bootstrap). A recent approach to stability selection based on subsampling was introduced by Meinshausen and Bühlmann (2010). Their studies impressively show that subsampling is a powerful tool in investigating the stability of models, such as penalized likelihood models and graphical models. Further Strobl et al. (2007) proposed the use of subsampling instead of bootstrapping in the context of random forests to circumvent the problem of preferential selection of certain types of predictors for a split. However, our results show that subsampling should not be regarded as an universally applicable alternative to the bootstrap. For selecting the optimal number of boosting steps via information criteria, for example, with the subsampling procedure a smaller number of boosting steps is selected, which might lead to too sparse models. If the aim is to investigate the distribution of model complexity parameters, the subsampling procedure is thus not recommended; prediction performance might similarly be affected. For investigating the variability of *p*-values, subsampling is again not appropriate, if more than type I error control is wanted. Our investigations make it clear that subsampling is not a reliable alternative to the bootstrap for all types of applications, even if it has shown important advantages in some situations (Strobl et al., 2007; De Bin et al., 2015).

Applied researchers should be careful when using approaches to problems in which hypothesis tests or information criteria are computed based on a bootstrap sample. If no investigations exist that indicate the reliability of a bootstrap approach, simulation studies are a helpful investigative tool. It is important to keep in mind that one cannot directly apply any procedure to bootstrap samples as if they were the original sample. It is advisable for methodologists to check the validity of their proposed bootstrap approaches by using simulation studies, and then comparing these results to those obtained when using original samples from the true underlying distribution. In this way, unexpected results can be easily discovered and adjustments may be made.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acknowledgments S.J. was financed by grant BO3139/2-2 to A.L.B. (DFG Einzelförderung). The authors thank Rory Wilson for helpful comments.

Conflict of interest
The authors have declared no conflict of interest.

A Appendix

A.1 Proof of Theorem 1

We derive

$$\mathrm{E}(Z^*) = \mathrm{E}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{E}(Z).$$

The variance of Z^* can be split into two parts,

$$\mathrm{Var}(Z^*) = \mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) + \mathrm{E}(\mathrm{Var}(Z^*|\hat{F})). \tag{A1}$$

Table A.1 Variables and corresponding interview question or description for the considered NHANES data.

Abbreviation	Interview question/description	Categories/units
race	Recode of reported race and ethnicity information	Mexican American Other Hispanic Non-Hispanic White Non-Hispanic Black Other Race - Including Multi-Racial
country of birth	In what country (were you/was sample person) born?	50 US States or Washington, DC Mexico Other Spanish Speaking Country Other Non-Spanish Speaking Country
education	What is the highest grade or level of school (you have/sample person has) completed or the highest degree (you have/she/he has) received?	Less than 9th Up to 11th High school Some college Graduate
marital status	Marital status	Married Widowed Divorced Separated Never married Living with partner

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table A.1 Continued

Abbreviation	Interview question/description	Categories/units
HealthStatus	Would you say (your/sample person's) health in general is . . .	Excellent Very good Good Fair Poor
depression	Over the last two weeks, how often have you been bothered by the following problems: little interest or pleasure in doing things? Would you say . . .	Not at all Several days Over half the days Nearly every day
ToothCond	Now I have some questions about the condition of your teeth and gums. How would you describe the condition of (your/sample person's) teeth? Would you say . . .	Excellent Very good Good Fair Poor
sleepTrouble	In the past month, how often did (you/sample person) have trouble falling asleep?	Never Rarely Sometimes Often Almost always
wakeUp	In the past month, how often did (you/sample person) wake up during the night and had trouble getting back to sleep?	Never Rarely Sometimes Often Almost always
medicalPlaceToGo	What kind of place (do you/does sample person) go to most often: is it a clinic, doctor's office, emergency room, or some other place?	Clinic Doctor's office Hospital emergency Hospital outpatient Other
income	Total household income (reported as a range value in dollars)	Under \$5k \$5k to under \$10k \$10k to under \$15k \$15k to under \$20k \$20k to under \$25k \$25k to under \$35k \$35k to under \$45k \$45k to under \$55k \$55k to under \$65k \$65k to under \$75k \$75k to under \$100k Over \$100k

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table A.1 Continued

Abbreviation	Interview question/description	Categories/units
AcuteIllness	Did (you/sample person) have a head cold or chest cold that started during the last 30 days? *or* Did (you/sample person) have flu, pneumonia, or ear infections that started during those 30 days? *or* Did (you/sample person) have a stomach or intestinal illness with vomiting or diarrhea that started during those 30 days?	No Yes
100cig	(Have you/Has sample person) smoked at least 100 cigarettes in (your/his/her) entire life?	Yes No
diabetes	(Have you/Has sample person) ever been told by a doctor or health professional that (you have/(he/she/sample person) has) diabetes or sugar diabetes?	Yes No
asthma	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) have/has asthma?	Yes No
heartFailure	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had congestive heart failure?	Yes No
stroke	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had a stroke?	Yes No
chronicBronchitis	Has a doctor or other health professional ever told (you/sample person) that (you/she/he) had chronic bronchitis?	Yes No
heavyDrinker	Was there ever a time or times in (your/sample person's) life when (you/he/she) drank 5 or more drinks of any kind of alcoholic beverage almost every day?	Yes No
waistcircum	Circumvence of waist	cm
Cholesterol	Cholesterol level	mg/dL
WBCcount	White blood cell count	1k cells/μL
BPsys	Systolic blood pressure	mmHg
BPdias	Diastolic blood pressure	mmHg
age	Age	Years
BMI	Body mass index	kg/m^2
alcohol	Alcohol consume	Units

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The first term reduces to

$$\mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{Var}(Z|\hat{F}) = 1. \tag{A2}$$

As far as the second term in (A1) is concerned, the basic assumption underlying bootstrap estimation of the variance, which can be easily shown in the present simple case (Davison, 1997), is that $\mathrm{Var}(Z^*|\hat{F})$ approximates $\mathrm{Var}(Z)$. Using this result one obtains for the second term

$$\mathrm{E}(\mathrm{Var}(Z^*|\hat{F})) = \mathrm{E}(\mathrm{Var}(Z)) = 1. \tag{A3}$$

Summing (A2) and (A3), Eq. (A1) results in $\mathrm{Var}(Z^*) = 2$. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. and Caski, F. (Eds.), *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest, HU, 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 Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* **18**, 967–985.

Binder, H. and Schumacher, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology* **7**, Article 12.

Black, S., Kushner, I. and Samols, D. (2004). C-reactive protein. *Journal of Biological Chemistry* **279**, 48487–48490.

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research* **21**, 205–229.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician* **65**, 213–221.

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY.

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* (2 edn.). John Wiley & Sons, New York, NY.

Davison, A. C. (1997). *Bootstrap Methods and Their Application*. Vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, UK.

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

De Bin, R., Janitza, S., Sauerbrei, W. and Boulesteix, A.-L. (2015). Subsampling versus bootstrap in resampling-based model selection for multivariable regression. Biometrics (in press).

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

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.

Good, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses* (3 edn). Springer, New York, NY.

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

Horowitz, J. L. (2001). The bootstrap. *Handbook of Econometrics* **5**, 3159–3228.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research* **11**, 2109–2113.

Manly, B. F. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3 edn.). CRC Press, FL.

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

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Mukherjee, S., Roberts, S. J., Sykacek, P. and Gurr, S. J. (2003). Gene ranking using bootstrapped p-values. *ACM SIGKDD Explorations Newsletter* **5**, 16–22.

National Center for Health Statistics (2012). NHANES 2007 to 2008 public data general release file documentation. http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/generaldoc_e.htm.

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.

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

Rospleszcz, S., Janitza, S. and Boulesteix, A.-L. (2014). Categorical variables with many categories are preferentially selected in model selection procedures for multivariable regression models on bootstrap samples. *Technical Report 164*. Department of Statistics, University of Munich, Munich, DE.

Sauerbrei, W., Boulesteix, A.-L. and Binder, H. (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data. *Journal of Biopharmaceutical Statistics* **21**, 1206–1231.

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.

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

Steck, H. and Jaakkola, T. S. (2003). Bias-corrected bootstrap and model uncertainty. In: Saul, L. K. Thrun, S. and Scholkopf, B. (Eds.), *Advances in Neural Information Processing Systems*, 16. Cambridge, MA: MIT Press, 521–528

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25.

Wagenmakers, E.-J., Farrell, S. and Ratcliff, R. (2004). Naïve nonparametric bootstrap model weights are biased. *Biometrics* **60**, 281–283.

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

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

15214036, 2016, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201400246 by <Shibboleth>-member@wcupa.edu, Wiley Online Library on [04/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License