STATISTICS IN MEDICINE
Statist. Med. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3104

Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination

Peter C. Austin[1, 2, 3, *, †]

[1]*Institute for Clinical Evaluative Sciences, Toronto, Ont., Canada*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, Ont., Canada*
[3]*Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ont., Canada*

SUMMARY

Applied researchers frequently use automated model selection methods, such as backwards variable elimination, to develop parsimonious regression models. Statisticians have criticized the use of these methods for several reasons, amongst them are the facts that the estimated regression coefficients are biased and that the derived confidence intervals do not have the advertised coverage rates. We developed a method to improve estimation of regression coefficients and confidence intervals which employs backwards variable elimination in multiple bootstrap samples. In a given bootstrap sample, predictor variables that are not selected for inclusion in the final regression model have their regression coefficient set to zero. Regression coefficients are averaged across the bootstrap samples, and non-parametric percentile bootstrap confidence intervals are then constructed for each regression coefficient. We conducted a series of Monte Carlo simulations to examine the performance of this method for estimating regression coefficients and constructing confidence intervals for variables selected using backwards variable elimination. We demonstrated that this method results in confidence intervals with superior coverage compared with those developed from conventional backwards variable elimination. We illustrate the utility of our method by applying it to a large sample of subjects hospitalized with a heart attack. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: regression models; backwards variable elimination; bootstrap; confidence intervals; Monte Carlo simulations; automated variable selection

*Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ont., Canada M4N 3M5.
†E-mail: peter.austin@ices.on.ca

Contract/grant sponsor: Ontario Ministry of Health and Long Term Care
Contract/grant sponsor: Natural Sciences and Engineering Research Council (NSERC)
Contract/grant sponsor: CIHR (Institute of Health Services and Policy Research)

Received 25 January 2007
Accepted 12 September 2007

1. INTRODUCTION

Applied researchers frequently use automated variable selection methods, such as backwards elimination, for the purpose of developing parsimonious regression models [1–3]. While automated variable selection methods are popular with applied researchers, statisticians have several concerns with the use of these methods. Studies have shown that the use of automatic variable selection methods in ordinary least squares regression results in spurious noise variables being mistakenly identified as independent predictors of the outcome [4–6]. Additionally, global measures of goodness of fit have been shown to be overly optimistic [5, 7]. The use of automated variable selection methods with logistic regression has been shown to result in the selection of non-reproducible sets of independent variables [8]. *P*-values resulting from models obtained using automated variable selection methods are biased downwards [9], while estimated regression coefficients are biased high in absolute value [9]. Finally, its use results in confidence intervals that have coverage rates lower than the advertised nominal level [10]. Despite the statistical concerns with automated variable selection methods, these methods are implemented in many statistical software programs and are popular with applied researchers who wish to develop parsimonious regression models.

The objective of the current study is to examine whether using backwards variable elimination in repeated bootstrap samples can result in improved estimation of regression coefficients and improved confidence intervals for regression coefficients selected using backwards variable elimination. This paper is divided into four sections. First, we describe the data upon which both the Monte Carlo simulations and the subsequent case study will be based. Second, we describe the Monte Carlo simulations that were conducted to examine whether using backwards variable elimination in repeated bootstrap samples can lead to improved estimation and confidence intervals. Third, we illustrate the application of our methods by estimating regression models and constructing associated 95 per cent confidence intervals using a large sample of patients hospitalized with a heart attack. Finally, we summarize our findings.

2. DATA SOURCES

Data used for generating parameters for the Monte Carlo simulations and for the subsequent case study consisted of 9484 patients discharged with a diagnosis of acute myocardial infarction (AMI or heart attack) from 102 Ontario hospitals between April 1, 1999 and March 31, 2001. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, an ongoing initiative intended to improve the quality of care for patients with cardiovascular disease in Ontario [11]. All variables used in the current study were either continuous or dichotomous. Dichotomous variables denoted the presence or absence of a specific condition or risk factor. The following 34 variables were available for the current study as potential predictor variables: demographic factors: age and gender; presenting signs and symptoms: cardiogenic shock and acute pulmonary edema; vital signs on admission: systolic blood pressure, diastolic blood pressure, heart rate, and respiratory rate; cardiac risk factors: diabetes, smoking history, history of stroke or transient ischemic attack, hyperlipidemia, hypertension, and family history of coronary artery disease; laboratory tests: glucose, white blood count, hemoglobin, sodium, potassium, urea, and creatinine; and comorbid conditions and vascular history: angina, cancer, dementia, previous AMI, asthma, depression, hyperthyroidism, peptic ulcer disease, peripheral vascular disease, previous

DOI: 10.1002/sim

angioplasty, chronic congestive heart failure, previous coronary artery bypass graft surgery, and aortic stenosis. We used death within 30 days of admission to hospital as a dichotomous outcome variable. These data are described elsewhere [8, 12–16].

3. MONTE CARLO SIMULATIONS FOR EXAMINING INFERENCE FOR BOOTSTRAP MODEL SELECTION METHODS

In this section, we first describe our approach of using backwards variable elimination in repeated bootstrap samples to improve the estimation of regression coefficients and the associated confidence intervals for predictor variables selected using conventional backwards variable elimination. We then describe and report on the Monte Carlo simulations that were conducted to examine the performance of our proposed method.

3.1. Backwards variable elimination in repeated bootstrap samples

In the proposed method, B bootstrap samples are drawn from the original sample. In each of the B bootstrap samples, conventional backwards variable elimination using a significance-based stopping rule is used to develop a reduced or parsimonious regression model. The initial regression model consists of the full regression model that includes all candidate predictor variables. Variables are sequentially eliminated from the initial regression model until all retained variables are significant at a pre-specified significance level α. The regression coefficients for the selected variables are noted. Those variables that are not selected for inclusion have their regression coefficients set to zero. This process is repeated in each of the B bootstrap samples. For each candidate variable, an estimated regression coefficient is obtained as the mean of the regression coefficients for that variable across the B bootstrap samples. Similarly, for each of the candidate predictor variables, a 95 per cent confidence interval is constructed using the 2.5th and 97.5th percentiles of the regression coefficients as endpoints for those variables across the B bootstrap samples (non-parametric bootstrap percentile confidence intervals) [17]. We refer to this method as the zero-corrected bootstrap model selection method. We also considered a variation in the above approach. In this variation, those variables that are not selected for inclusion in a given bootstrap sample have their regression coefficients set to missing rather than to zero. We refer to this variant as the conditional bootstrap model selection method, since we consider the mean of the bootstrap coefficients, conditional on the variable having been selected for inclusion in the final model in a given bootstrap sample.

For comparative purposes, we also examined an alternative approach. Backwards variable elimination is used in the original sample to develop a parsimonious regression model using the same significance-based stopping rule as above (with significance level α). The coefficients from this parsimonious model are then re-estimated in each of the B bootstrap samples. Mean regression coefficients are estimated across the B bootstrap samples, and percentile bootstrap confidence intervals are obtained as above. We refer to this method as the naïve bootstrap method. This method does not attempt to account for uncertainty in which variables were selected for inclusion in the regression model.

We thus considered three different approaches: zero-corrected bootstrap model selection, conditional bootstrap model selection, and the naïve bootstrap method.

DOI: 10.1002/sim

3.2. Data-generating process for the Monte Carlo simulations

We used a random sample of 5000 subjects from the EFFECT data described in Section 2 (a random sample of size 5000 was used to decrease computational demands of the simulations. Using a sample size of 5000, the simulations used approximately 24 days of CPU time). A logistic regression model was fit using the 34 variables listed in Section 2, with 30-day mortality as the dichotomous outcome variable. We assumed that continuous variables (e.g. age) were linearly related to the log odds for 30-day mortality. The 35 estimated regression coefficients (intercept and 34 variables) were used as the basis of our data-generating process for the Monte Carlo simulations. For each subject in the random sample of 5000 subjects from the EFFECT data, the linear predictor was computed: $X_i\beta$, where X_i denotes the vector of variables for a given subject, and β denotes the vector of estimated regression coefficients. This set of 5000 linear predictors was held fixed and used in the subsequent data-generating process. Given a subject-specific linear predictor, we used the inverse logit transformation to determine that subject's predicted probability of death within 30 days of admission. We then randomly generated an outcome for that subject from a Bernoulli distribution with probability of success equal to the subject-specific probability of death within 30 days of admission. This process allowed us to randomly generate outcomes that, on average, resembled the EFFECT sample. Furthermore, the distribution of the covariates and the correlations between the covariates were always equal to that of the EFFECT sample and thus reflected real-world covariate correlations.

3.3. Monte Carlo simulations—methods

We randomly generated outcomes for 5000 subjects using the design matrix and vector of regression coefficients described in Section 3.2. We used the methods described in Section 3.1, with $B=500$ bootstrap samples. The above process was then repeated 1000 times. For each variable, the proportion of constructed 95 per cent confidence intervals that contained the true regression parameter was determined. We examined four different significance-based stopping rules: using significance levels (α) of $0.01, 0.05, 0.157$ (equivalent to the AIC criterion), and 0.50. The choice of these significance levels was based on the previous work by Steyerberg *et al.* [18] examining the use of automated variable selection methods for developing predictive models.

For comparative purposes, we examined the performance of conventional backwards variable elimination in the randomly generated data sets. In each of the 1000 randomly generated data sets, conventional backwards variable elimination was used to obtain a parsimonious regression model, using the same significance-based stopping rules as above. For each model obtained using backwards variable elimination, 95 per cent confidence intervals were constructed for each selected variable using model-based standard errors. For each variable, the mean regression coefficient was determined across the 1000 randomly generated data sets. Similarly, for each variable, the proportion of constructed 95 per cent confidence intervals that contained the true regression parameter was obtained in Section 3.1.

The Monte Carlo simulations were conducted using the R statistical programming language, version 2.2.0 [19]. Backwards variable elimination was done using the 'fastbw' function in the 'Design' package for R. Harrell states that 'this method uses the fitted complete model and computes approximate Wald statistics by computing conditional (restricted) maximum likelihood estimates assuming multivariate normality of estimates' [9, p. 114].

DOI: 10.1002/sim

Table I. Frequency of selecting variables in 1000 randomly generated samples.

Significance level for selection (α)	Minimum frequency of variable selection	Maximum frequency of variable selection	Number of variables selected in greater than 90 per cent of the data sets	Number of variables selected in less than 10 per cent of the data sets
0.01	0.010	1.000	6	15
0.05	0.017	1.000	6	11
0.157	0.034	1.000	6	7
0.50	0.096	1.000	6	1

3.4. Monte Carlo simulations—results

3.4.1. Frequency of selecting different candidate variables. The range in the proportions of the 1000 simulated data sets in which individual predictors were selected for inclusion using conventional backwards variable elimination is described in Table I for each of the significance levels for variable retention. In all five settings, smoking status was the variable selected the least frequently, being selected for inclusion in fewer than 10 per cent of the simulated data sets, regardless of the significance level for variable retention. Depending on the significance level, either three or four variables were selected for inclusion in 100 per cent of the simulated data sets. Age, systolic blood pressure, and white blood count were selected for inclusion in 100 per cent of the simulated data sets regardless of the significance level for variable retention. Cardiogenic shock was also selected in 100 per cent of the samples when the significance level for variable retention was 0.157 and 0.50. Six variables were selected for inclusion in greater than 90 per cent of the simulated samples regardless of the significance level employed: age, cardiogenic shock, systolic blood pressure, white blood count, glucose level, and creatinine. The number of variables selected for inclusion in fewer than 10 per cent of the simulated samples ranged from 1 ($\alpha = 0.50$) to a high of 15 ($\alpha = 0.01$). These results highlight the instability of regression models selected using backwards variable elimination, confirming prior empirical studies [8].

3.4.2. Estimating regression coefficients. The mean estimated regression coefficients for the 34 candidate predictor variables are depicted in Figures 1–4 for significance levels $0.01, 0.05, 0.157$, and 0.50, respectively. Each figure consists of four panels. The mean estimated regression coefficients for zero-corrected bootstrap model selection, conditional bootstrap model selection, the naïve bootstrap method, and backwards variable elimination are plotted against the true regression parameters used in the data-generating process. Figures 1–4 are qualitatively similar in form.

3.4.3. Coverage of 95 per cent confidence intervals. Coverage of 95 per cent confidence intervals for the 34 candidate predictor variables obtained using the different methods is displayed in Figures 5–8 for significance levels (α) 0.01, 0.05, 0.157, and 0.50, respectively. As above, the four figures are qualitatively similar in form. In the upper left panel of each figure, the coverage rates for confidence intervals obtained using the zero-corrected bootstrap model selection method are compared with those using conventional backwards elimination. The large majority of the 34 points lie above the line with unit slope, indicating that coverage rates for the 95 per cent confidence intervals were higher for zero-corrected bootstrap model selection compared with those

DOI: 10.1002/sim

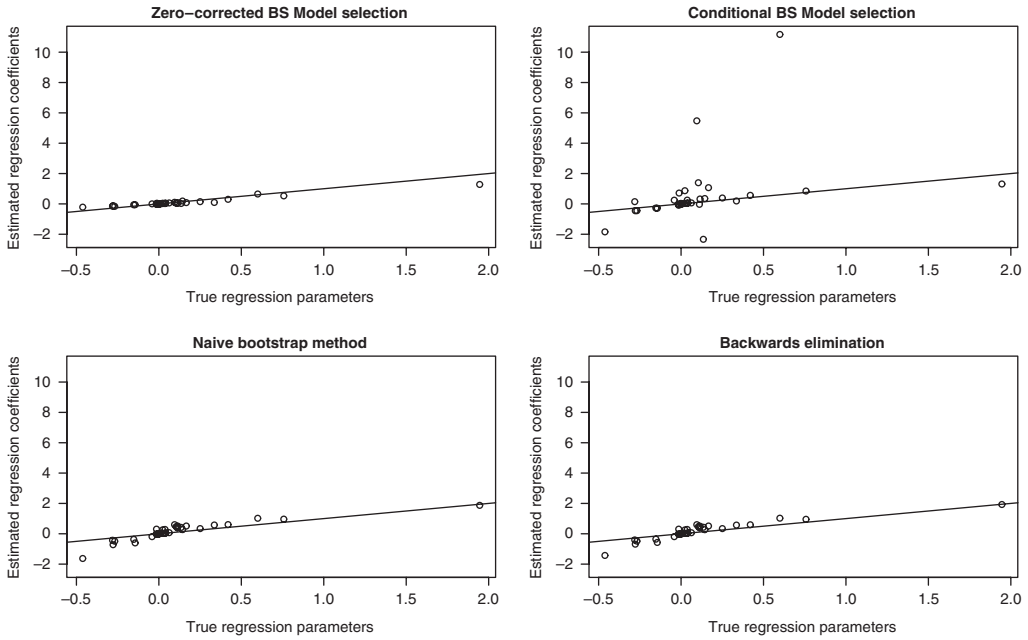

Figure 1. Estimation of regression coefficients ($\alpha = 0.01$).

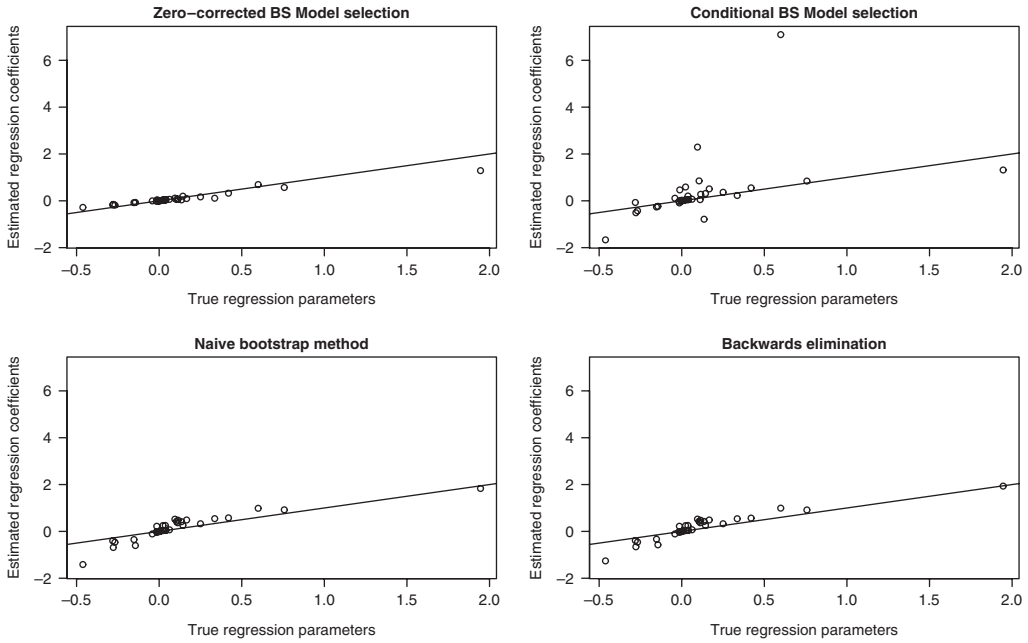

Figure 2. Estimation of regression coefficients ($\alpha = 0.05$).

DOI: 10.1002/sim

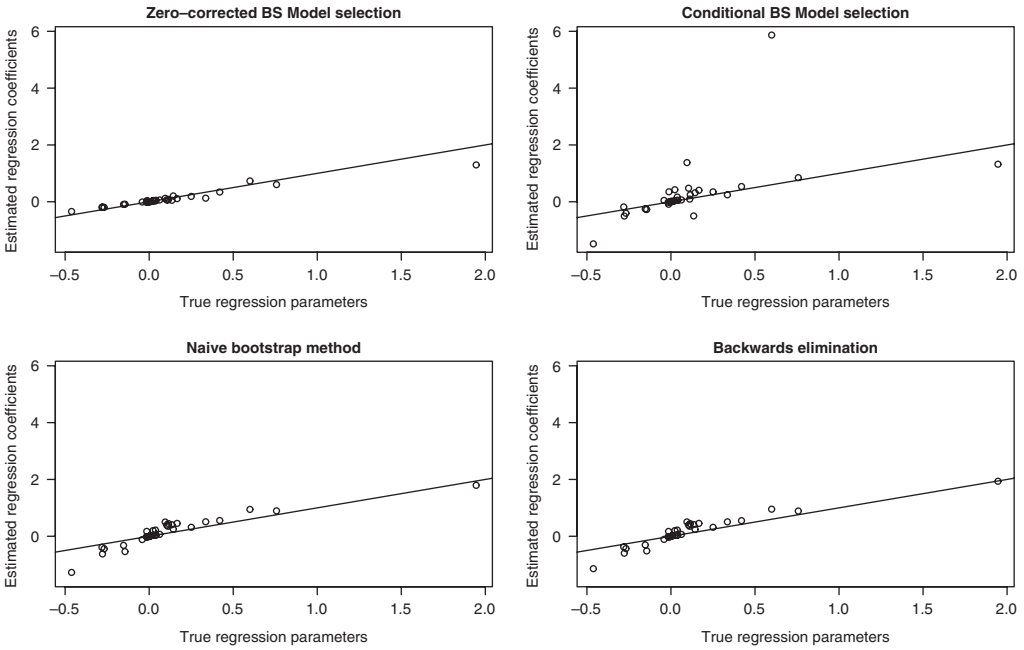

Figure 3. Estimation of regression coefficients ($\alpha = 0.157$).

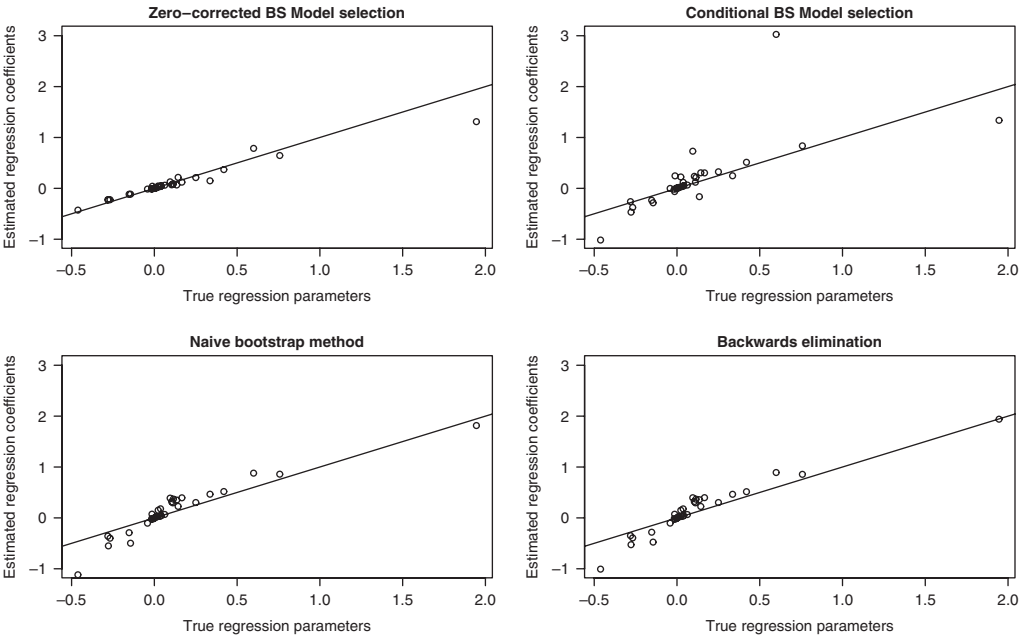

Figure 4. Estimation of regression coefficients ($\alpha = 0.50$).

DOI: 10.1002/sim

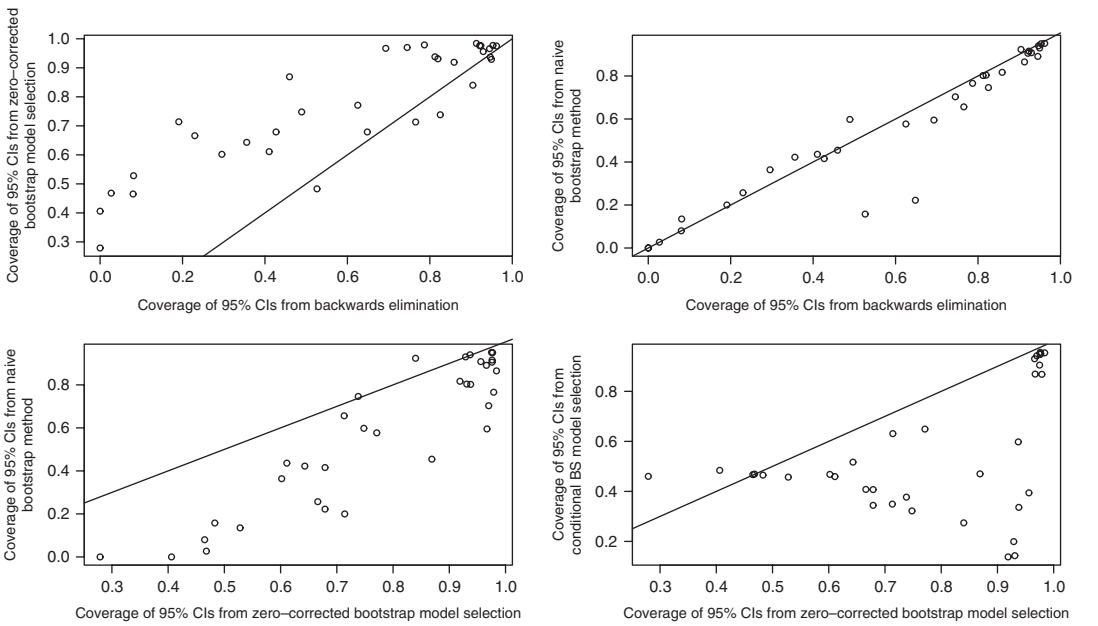

Figure 5. Coverage of 95 per cent confidence intervals ($\alpha = 0.01$).

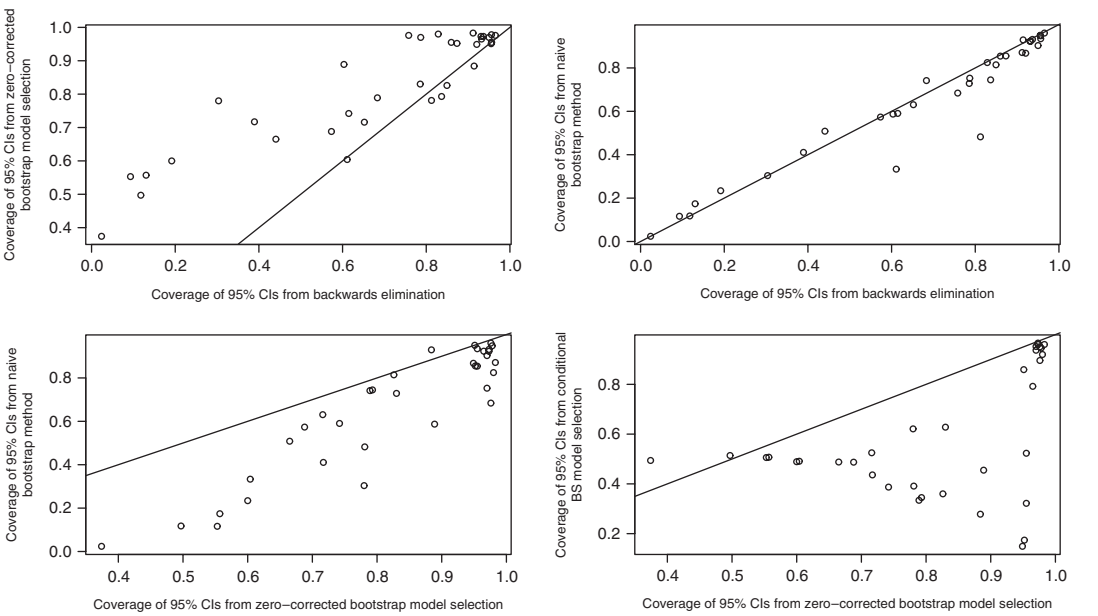

Figure 6. Coverage of 95 per cent confidence intervals ($\alpha = 0.05$).

DOI: 10.1002/sim

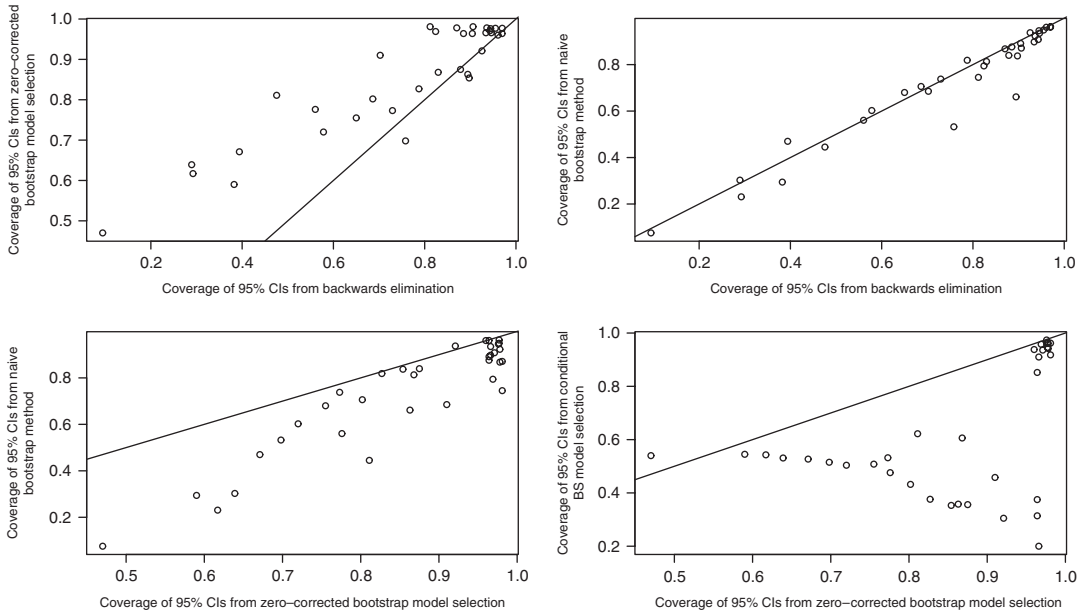

Figure 7. Coverage of 95 per cent confidence intervals ($\alpha = 0.157$).

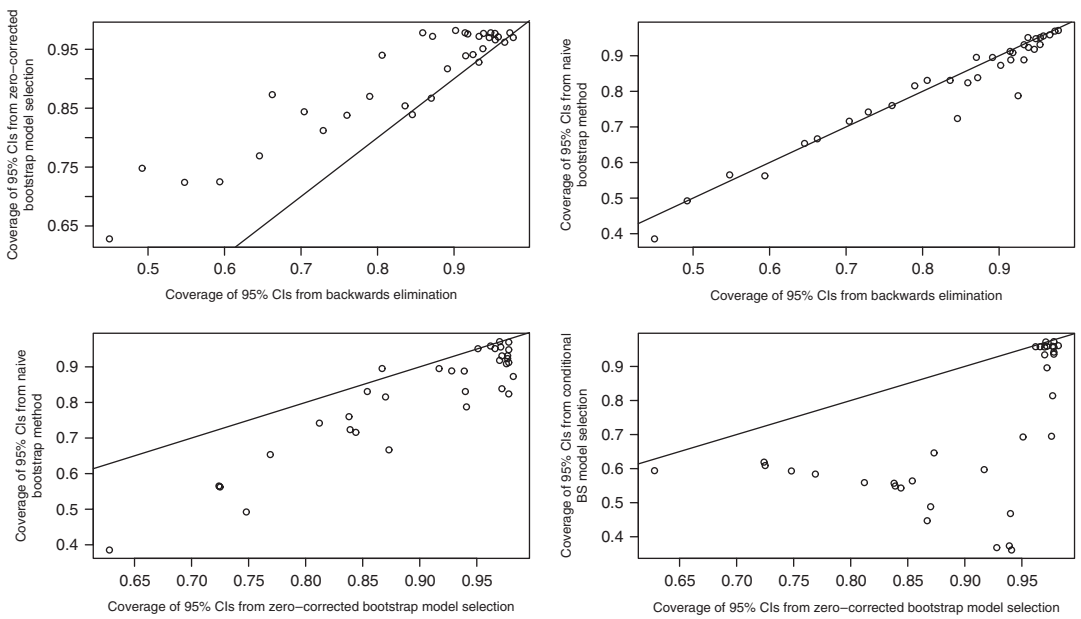

Figure 8. Coverage of 95 per cent confidence intervals ($\alpha = 0.50$).

DOI: 10.1002/sim

for backwards elimination. In the lower left panel of each figure, the large majority of the points lie below the line with unit slope, indicating that zero-corrected bootstrap model selection produced confidence intervals with greater coverage rates than did the naïve bootstrap method. In the upper right panel of each figure, the majority of points lie along the line with unit slope, indicating that backwards variable elimination and the naïve bootstrap method tended to result in confidence intervals with similar coverage rates. In the lower right panel of each figure, the large majority of points lie below the line with unit slope, indicating that zero-corrected bootstrap model selection results in 95 per cent confidence intervals with better coverage rates than did conditional bootstrap model selection.

Due to our use of 1000 iterations in our Monte Carlo simulations, confidence intervals with coverage rates of at least 93.9 per cent would not be statistically different from 95 per cent using a one-tailed significance test with a 0.05 significance level. Using zero-corrected bootstrap model selection, of the 34 predictor variables, the number of confidence intervals whose coverage rates were at least 93.9 per cent was $10, 15, 15$, and 19, when the significance level (α) was 0.01, 0.05, 0.157, and 0.50, respectively. In comparison, the use of conventional backwards variable elimination resulted in only $5, 5, 7$, and 8 confidence intervals whose coverage rates were at least 93.9 per cent when the significance level (α) was $0.01, 0.05, 0.157$, and 0.50, respectively.

4. CASE STUDY

The case study used all 9484 subjects from the EFFECT data set described in Section 2. Three separate logistic regression models were fit to the data. First, we fit a logistic regression model consisting of all 34 main effects. Second, a logistic regression model was obtained using backwards variable elimination from the full model. Using this approach, a significance-based stopping rule was employed such that only variables with a significance level of less than 0.50 were retained in the final model. We chose to use a significance level of 0.50 due to the findings in the Monte Carlo simulations that this significance level resulted in the lowest mean difference between estimated regression coefficients and the true regression coefficients. Third, we used zero-corrected bootstrap model selection, using a significance level of 0.50 for the significance-based stopping rule. One thousand bootstrap samples were drawn from the original data set for this approach. The mean regression coefficient was determined for each predictor variable across the 1000 bootstrap samples.

The results from the above three analyses are reported in Table II. For the full regression model, we report the estimated regression coefficient and associated 95 per cent confidence intervals for each of the 34 predictor variables. Fifteen of the 34 candidate predictor variables were selected as significant predictors of 30-day mortality using conventional backwards variable elimination with a significance level of 0.50 for variable retention. We report the estimated regression coefficients and associated confidence intervals for these 15 predictor variables. Finally, we report the estimated coefficients and associated confidence intervals for these 15 predictor variables as estimated using the zero-corrected bootstrap model selection.

In comparing the estimated regression coefficients, one notes that for three of the 15 variables, zero-corrected bootstrap model selection and backwards variable selection resulted in identical coefficient estimates. For five of the 15 variables, zero-corrected bootstrap model selection resulted in estimated regression coefficients that were closer to those of the full regression model, while

DOI: 10.1002/sim

Table II. Results of fitting three models in EFFECT sample.

Variable	Full model (34 main effects)	Backwards elimination	Bootstrap model selection
Age	0.066 (0.057, 0.074)	0.068 (0.061, 0.075)	0.066 (0.052, 0.076)
Female	0.081 (−0.081, 0.243)		
Cardiogenic shock	1.903 (1.49, 2.316)	1.915 (1.505, 2.325)	1.871 (1.304, 2.366)
Acute pulmonary edema	0.102 (−0.161, 0.365)		
Diabetes	0.006 (−0.178, 0.189)		
High BP	0.05 (−0.105, 0.206)		
Current smoker	0.029 (−0.163, 0.222)		
CVA/TIA	0.119 (−0.086, 0.323)		
Dyslipidemia	−0.237 (−0.43, −0.045)	−0.231 (−0.417, −0.045)	−0.206 (−0.438, 0)
Family history of CAD	−0.173 (−0.386, 0.039)		
Angina	0.056 (−0.106, 0.219)		
Cancer	−0.023 (−0.384, 0.338)		
Dementia	0.381 (0.115, 0.648)	0.395 (0.131, 0.66)	0.353 (0, 0.684)
Peptic ulcer disease	−0.267 (−0.591, 0.056)		
Previous AMI	−0.137 (−0.315, 0.041)		
Asthma	−0.044 (−0.361, 0.273)		
Depression	0.266 (0.011, 0.521)	0.296 (0.043, 0.548)	0.23 (0, 0.541)
Peripheral arterial disease	0.141 (−0.097, 0.38)		
Previous PCI	−0.368 (−0.901, 0.165)		
Congestive heart failure (chronic)	0.216 (−0.036, 0.469)	0.218 (−0.029, 0.465)	0.16 (0, 0.482)
Hyperthyroidism	0.351 (−0.179, 0.882)		
Previous CABG	0.148 (−0.155, 0.452)		
Aortic stenosis	0.4 (−0.025, 0.824)	0.443 (0.023, 0.864)	0.321 (0, 0.848)
Systolic BP	−0.018 (−0.022, −0.015)	−0.017 (−0.02, −0.015)	−0.017 (−0.022, −0.012)
Diastolic BP	0.002 (−0.004, 0.008)		
Heart rate	0.004 (0.002, 0.007)	0.005 (0.002, 0.008)	0.004 (0, 0.008)
Respiratory rate	0.021 (0.009, 0.032)	0.022 (0.011, 0.033)	0.021 (0, 0.035)
Hemoglobin	−0.001 (−0.006, 0.003)		
White blood count	0.04 (0.029, 0.051)	0.04 (0.029, 0.051)	0.04 (0.02, 0.055)
Sodium	0.001 (−0.016, 0.019)		
Potassium	0.112 (−0.01, 0.234)	0.115 (−0.005, 0.235)	0.084 (0, 0.239)
Glucose	0.042 (0.029, 0.055)	0.043 (0.032, 0.055)	0.042 (0.022, 0.057)
Urea	0.031 (0.017, 0.046)	0.033 (0.019, 0.047)	0.032 (0, 0.049)
Creatinine	0.002 (0.001, 0.003)	0.002 (0.001, 0.003)	0.002 (0, 0.003)

Note: Each cell contains the estimated regression coefficient and the associated 95 per cent confidence interval.

for the remaining seven variables, backwards model selection resulted in estimated regression coefficients that were closer to those of the full regression model.

5. DISCUSSION

In the current study, we examined the use of backwards variable elimination in multiple bootstrap samples to improve estimation of regression coefficients and confidence intervals for regression coefficients from models obtained using backward variable selection. Using Monte Carlo simula-

DOI: 10.1002/sim

tions, we demonstrated that zero-corrected bootstrap model selection results in better estimation of both regression coefficients and confidence intervals compared with estimates derived directly from a model obtained using conventional backwards elimination. We then illustrated the utility of our method in a large sample of patients hospitalized with a heart attack.

We examined two different methods based on bootstrap resampling. In the first approach, which we termed zero-corrected bootstrap model selection, regression coefficients for variables that were not selected for inclusion in a given bootstrap sample were set to zero. In an alternative approach, which we termed conditional bootstrap model selection, regression coefficients for variables that were not selected in a given bootstrap sample were set to missing. Regression coefficients were then averaged across the bootstrap samples. We found that conditional bootstrap model selection had substantially worse performance compared with zero-corrected bootstrap model selection. This result can likely be explained by the known bias in regression coefficients selected using automated variable selection methods. The conditional bootstrap method only averaged over coefficients for variables that had been selected for inclusion. Thus, for a given variable, the estimated regression coefficient was the mean of a set of biased coefficients. By contrast, zero-corrected bootstrap model selection sets regression coefficients to zero when that variable was not selected for inclusion in a given bootstrap sample. Thus, for a given variable, the estimated regression coefficient will be a weighted average of zero (the bootstrap samples in which the variable was not selected) and the mean of a set of biased coefficient estimates (the bootstrap samples in which the variable was not selected). Thus, the bias in the zero-corrected bootstrap model selection method is less than that of the conditional bootstrap model selection method. The low bias of the zero-corrected bootstrap model selection method is apparent in the upper left panels of Figures 1–4, in which it was shown that zero-corrected bootstrap model selection resulted in estimates that were, on average, very similar to the true regression parameters used in the data-generating process. A similar argument explains the poorer coverage of the confidence intervals generated using the conditional bootstrap method compared with the zero-corrected bootstrap model selection method.

In our Monte Carlo simulations, we found that zero-corrected bootstrap model selection resulted in confidence intervals with improved coverage compared with conventional backwards variable elimination. However, it still resulted in some variables having confidence intervals with coverage rates lower than the nominal level. We hypothesize that this is related to the low frequency with which some variables were selected for inclusion. For instance, when the significance level for variable retention was set to 0.05, then 11 of the candidate predictor variables were selected for inclusion in fewer than 10 per cent of the bootstrap samples (see Table I). Thus, for each of these 11 variables, at least 90 per cent of the regression coefficients would be set to 0. Computing percentile bootstrap confidence intervals may have had sub-optimal performance due to the low frequency of inclusion of these variables.

Strong arguments can be made for fitting the full model consisting of all predictor variables. While this can be a valid approach in many settings, there may be scenarios in which some form of variable reduction is desirable. For instance, the regression model may be intended for use prospectively. If variables must be collected prospectively, then investigators may wish to avoid collecting variables that are expensive or difficult to obtain if these variables are minimally important. Thus, in this setting, some form of variable reduction may be desirable.

A limitation to our Monte Carlo simulations is that they were limited to large data sets ($N = 5000$). There are computational issues that can make the implementation of our Monte Carlo simulations difficult in small data sets. For instance, some of the risk factors have a low prevalence

DOI: 10.1002/sim

(e.g. cardiogenic shock at admission). In small samples containing only a few subjects with a given risk factor, bootstrap samples may be selected in which no subjects have that risk factor. This will lead to a singular design matrix. Modifications of our methods need to be developed for application in small data sets. Similarly, with a low event rate, there may be a low number of deaths in small samples. This limits both the number of variables that may be included in the regression model [20] and the need for variable reduction.

Our proposed method shares some similarities with Bayesian model averaging. We used backwards variable selection in repeated bootstrap, averaging estimated regression coefficients over the multiple bootstrap samples. We then report the estimated regression coefficients and associated confidence intervals using this method for those variables that had been selected using backwards variable selection in the initial sample. Bayesian model averaging, in which models are weighted according to their posterior probability, allows for a more formal method of incorporating model uncertainty into model selection. Owing to the computationally intensive nature of the Monte Carlo simulations of Bayesian model averaging, the comparison of Bayesian model averaging with bootstrap model selection methods is beyond the scope of the current study. However, the relative performance of these competing methods should be examined in future research.

Our study adds to the literature examining the use of bootstrap methods to improve variable and model selection. Both Sauerbrei and Schumacher [21] and Austin and Tu [22] have proposed model selection methods based on using automated variable selection methods in multiple bootstrap samples. In the variation of this method proposed by Austin and Tu, in each bootstrap sample, backwards variable elimination or forward variable selection was used to identify the independent predictors of the outcome [22]. Candidate predictor variables are then ranked according to the proportion of the bootstrap samples in which they were selected as independent predictors of the outcome. In their empirical examination of this method, this method was found to result in a model with predictive accuracy that exceeded that of models developed using Akaike information criterion, Bayesian information criterion, or cross-validation [22]. Several studies in the cardiovascular surgical literature have employed this approach to develop parsimonious regression models [23–29]. In these cardiovascular applications, those variables that were selected for inclusion in at least 50 per cent of the bootstrap samples were retained for inclusion in the final regression model.

Automated variable selection methods are currently implemented in many statistical software packages and are popular with applied researchers. Many statisticians have concerns about the use of automated variable selection methods. The concern that we have addressed in this article is that the use of automated variable selection results in biased estimation of regression coefficients and confidence intervals with coverage rates that are lower than the advertised levels [10]. While our zero-corrected bootstrap model selection should not be seen as a panacea to the many potential problems with automated variable selection methods, its use can result in better estimation of regression coefficients and confidence intervals compared with that of model-based confidence intervals naively derived from the model obtained using conventional backwards variable elimination. Applied researchers are attracted to backwards variable elimination, in part, due to its ability to create a more parsimonious regression model. Our proposed method can result in regression coefficients and confidence intervals for the coefficients included in this reduced model that have superior performance compared with model-based confidence intervals derived from the selected model. However, coverage of 95 per cent confidence intervals can still be lower than the advertised nominal level.

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results, and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC). Dr Austin is supported in part by a New Investigator Award from the CIHR (Institute of Health Services and Policy Research). The data used were obtained from the EFFECT study, which was funded by a grant from the CIHR to the Canadian Cardiovascular Outcomes Research Team (CCORT).

REFERENCES

1. Miller AJ. Selection of subsets of regression variables. *Journal of the Royal Statistical Society*, *Series A* 1984; **147**:389–425.
2. Miller A. *Subset Selection in Regression* (2nd edn). Chapman & Hall/CRC: Boca Raton, FL, 2002.
3. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics* 1976; **32**:1–49.
4. Derkson S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 1992; **45**:265–282.
5. Flack VF, Chang PC. Frequency of selecting noise variables in subset regression analysis: a simulation study. *The American Statistician* 1987; **14**:84–86.
6. Murtaugh PA. Methods of variable selection in regression modeling. *Communications in Statistics—Simulation and Computation* 1998; **27**:711–734.
7. Copas JB, Long T. Estimating the residual variance in orthogonal regression with variable selection. *The Statistician* 1991; **40**:51–59.
8. Austin PC, Tu JV. Automated variable selection methods for logistic regression result in unstable models for predicting AMI mortality. *Journal of Clinical Epidemiology* 2004; **57**:1138–1146.
9. Harrell Jr FE. *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: New York, NY, 2001.
10. Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. *The American Statistician* 1990; **44**:214–217.
11. Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM. *Quality of Cardiac Care in Ontario*. Institute for Clinical Evaluative Sciences: Toronto, Ont., 2004.
12. Austin PC. A comparison of classification and regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine* 2007; **26**:2937–2957.
13. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**:2084–2106.
14. Austin PC, Tu JV. Comparing clinical data with administrative data for producing AMI report cards. *Journal of the Royal Statistical Society*, *Series A* (*Statistics in Society*) 2006; **169**:115–126.
15. Austin PC, Mamdani MM, Juurlink DN, Alter DA, Tu JV. Missed opportunities in the secondary prevention of myocardial infarction: an assessment of the effects of statin underprescribing on mortality. *American Heart Journal* 2006; **151**:969–975.
16. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine* 2005; **24**:1563–1578.
17. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
18. Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small datasets. *Statistics in Medicine* 2000; **19**:1059–1079.
19. R Core Development Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2005.
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
21. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**:2093–2109.

DOI: 10.1002/sim

22. Austin PC, Tu JV. Bootstrap methods for developing predictive models in cardiovascular research. *The American Statistician* 2004; **58**:131–137.
23. Sabik JF, Gillinov AM, Blackstone EH *et al.* Does off-pump coronary surgery reduce morbidity and mortality? *Journal of Thoracic and Cardiovascular Surgery* 2002; **124**:698–707.
24. Koch CG, Khandwala F, Nussmeier N, Blackstone EH. Gender and outcomes after coronary artery bypass grafting: a propensity matched comparison. *Journal of Thoracic and Cardiovascular Surgery* 2003; **126**:2032–2043.
25. Rice TW, Khuntia D, Rybicki LA, Adelstein DJ, Vogelbaum MA, Mason DP, Murthy SC, Blackstone EH. Brain metastases from esophageal cancer: a phenomenon of adjuvant therapy? *Annals of Thoracic Surgery* 2006; **82**(6):2042–2049, 2049, e1–e2.
26. DeCamp MM, Blackstone EH, Naunheim KS, Krasna MJ, Wood DE, Meli YM, McKenna Jr RJ, NETT Research Group. Patient and surgical factors influencing air leak after lung volume reduction surgery: lessons learned from the National Emphysema Treatment Trial. *Annals of Thoracic Surgery* 2006; **82**(1):197–206.
27. Svensson LG, Blackstone EH, Rajeswaran J, Sabik 3rd JF, Lytle BW, Gonzalez-Stawinski G, Varvitsiotis P, Banbury MK, McCarthy PM, Pettersson GB, Cosgrove DM. Does the arterial cannulation site for circulatory arrest influence stroke risk? *Annals of Thoracic Surgery* 2004; **78**(4):1274–1284.
28. Sabik JF, Nemeh H, Lytle BW, Blackstone EH, Gillinov AM, Rajeswaran J, Cosgrove DM. Cannulation of the axillary artery with a side graft reduces morbidity. *Annals of Thoracic Surgery* 2004; **77**(4):1315–1320.
29. Shishehbor MH, Seshadri N, Aktas M, Acharya N, Gillinov AM, Blackstone EH, Houghtaling PL, Migrino RQ, Ghaffari S. Comparison of outcomes in patients undergoing coronary bypass of patent versus restenosed bare metal stented coronary arteries. *American Journal of Cardiology* 2005; **96**:1416–1419.

DOI: 10.1002/sim