Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

A simulation based method for assessing the statistical significance of logistic regression models after common variable selection procedures

Tristan R. Grogan & David A. Elashoff

To cite this article: Tristan R. Grogan & David A. Elashoff (2017) A simulation based method for assessing the statistical significance of logistic regression models after common variable selection procedures, Communications in Statistics - Simulation and Computation, 46:9, 7180-7193, DOI: 10.1080/03610918.2016.1230216

To link to this article: https://doi.org/10.1080/03610918.2016.1230216

© 2017 The Author(s). Published with license by Taylor & Francis Group, LLC© Tristan R. Grogan and David A. Elashoff

Published online: 23 May 2017.

Submit your article to this journal

Article views: 4272

View related articles

View Crossmark data

Citing articles: 3 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lssp20

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 9, 7180–7193
https://doi.org/10.1080/03610918.2016.1230216

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A simulation based method for assessing the statistical significance of logistic regression models after common variable selection procedures

Tristan R. Grogan and David A. Elashoff

Department of Medicine Statistics Core, University of California, Los Angeles, California, USA

ABSTRACT

Classification models can demonstrate apparent prediction accuracy even when there is no underlying relationship between the predictors and the response. Variable selection procedures can lead to false positive variable selections and overestimation of true model performance. A simulation study was conducted using logistic regression with forward stepwise, best subsets, and LASSO variable selection methods with varying total sample sizes (20, 50, 100, 200) and numbers of random noise predictor variables (3, 5, 10, 15, 20, 50). Using our critical values can help reduce needless follow-up on variables having no true association with the outcome.

ARTICLE HISTORY
Received 26 April 2016
Accepted 22 August 2016

KEYWORDS
AUC; Logistic regression; Simulation study; Validation methods; Variable selection

1. Introduction

Recently, the medical field has been under fire for the frequency with which published research does not seem to be reproducible despite the billions of dollars invested in it. The journalist, Michael Hilzik (2013) published an article in the Los Angeles Times about this topic. The article was based on an influential *Nature* publication written by Glenn Begley, former head of Hematology and Oncology Research at Amgen. In his article (2012), Dr. Begley described how his team tried to reproduce results of influential papers from the past decade in the fields of cancer research and blood biology. He claimed that "Of the 53 landmark papers, only 6 could be proved valid." Similar findings were reported by a group at Bayer HealthCare in Germany where only 25% of the published papers which they were basing their research and development projects on could be categorized as reproducible (Prinz et al., 2011). Furthermore, the majority of *Nature* readers responding to a recent survey (66%) expressed a high level of concern that current levels of reproducibility are a major problem (Reality check on Reproducibility, 2016).

There are many reasons for lack of reproducibility in these fields, including low power (too small a sample size), differing lab protocols, pre-processing variation between sites, different patient populations, and use of non-random samplesto name a few. In addition, research results may not be reproducible due to the misinterpretation or misuse of statistical test results. As a consequence, investigators may be overoptimistic about their models, which could suffer from overfitting or may include spurious variables not truly related to the response. Nate

CONTACT Tristan Grogan Tgrogan@mednet.ucla.edu Department of Medicine Statistics Core, University of California, Los Angeles, CA 90024, USA.

Published with license by Taylor & Francis Group, LLC © 2017 Tristan R. Grogan and David A. Elashoff
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Silver, creator of the popular website www.FiveThirtyEight.com, stated in his book (2012) that forecasters often develop an elaborate model which is an overly specific solution to a general problem where the model has little or no value in making predictions. He calls this "overfitting," which is "the most important scientific problem you've never heard of."

Medical researchers are often interested in selecting a panel of predictor variables for diagnostic or prognostic models. A standard statistical approach is the use of logistic regression to identify markers of patient status such as cancer or control. This scenario is especially common in biomarker validation studies which can include large numbers of predictor variables relative to the sample size. See for example, the metabolomic comparison studies by Shen et al. (2013) and Zhang et al. (2014). Situations where the number of metabolites is larger than the number of subjects are very common. Marozzi (2014, 2015b) discusses that in these situations, traditional methods like the Hotelling test cannot be used, and proposes more appropriate methods. A frequently used method for evaluating the performance of these multivariable logistic regression models is to assess the magnitude and overall statistical significance of the area under the receiver operating characteristic (ROC) curve after an automated variable selection procedure has been performed.

This manuscript proposes to improve statistical reproducibility and accountability in the field by providing useful guidance and better methodology for a more relevant test of statistical significance from logistic regression models after variable selection has been performed. While some approaches have been developed to lower the magnitude of variable selection bias, they often require possession of the dataset and considerable statistical programming experience to implement. Our tables for adjusting the critical value can be used even when the dataset is unavailable to a researcher. The results of two cancer biomarker datasets which we have acquired from collaborations at our institution are included to help illustrate our method and assess how our thresholds compare to the permutation test. The goal of this manuscript is not to find the best variable selection technique, but to more accurately evaluate the significance of commonly used variable selection procedures.

2. Background

For a recent discussion on the misuse of statistics in practice, see Marozzi (2015a) who emphasizes that a context where the debate is particularly fierce is medicine. Driven by the wide availability of powerful and relatively cheap computers as well as statistical software, medical researchers tend to use more complex methods than before with a correspondent increased risk of misusing such methods. In the age of bioinformatics, genetics and high-throughput genomic/proteomic/metabolomic studies, this problem is exacerbated by the complexity of analyzing large datasets. However, it should be underlined that very standard methods like Student t, chi-squared and Fisher's exact test are still very commonly used, and despite being well known methods, they are very often applied incorrectly (Ferraris and Ferraris 2003; Fong et al. 2008; Gandhi et al. 2011; Ludbrook and Dudley 1998; Lucena et al. 2011; McKinney et al. 1989 and Podoll et al. 2012).

Many scholars have emphasized that a large share of published medical research contains statistical errors. Strasak et al. (2007) and Fernandes–Taylor et al. (2011) underline that also top class journals like Nature Medicine and The New England Journal of Medicine publish a considerable proportion of papers containing statistical errors (more or less severe). They conclude with caution that the journal impact factor is not a very meaningful predictor for the statistical quality of published research. Medical researchers can also be under tremendous pressure from institutions to publish manuscripts for career advancement, personal prestige, grant funding opportunities, and/or financial incentives. This pressure can cause a researcher to be more desperate to publish quickly before getting the proper statistical guidance.

A less commonly reported misuse of statistics includes variable selection bias, which can be defined as finding a spurious relationship between a predictor variable and the response; that is, concluding that there is a statistically significant relationship when in reality the result was just a product of multiple hypothesis testing. This type of mistake is called a Type 1 error and can be controlled by setting the alpha (α) level before an analysis is run. When hypothesis tests are conducted for more than one variable it is often desirable to control the family-wise error rate (FWER) which is defined as the chance of selecting any noise predictors. The FWER after implementing a variable selection algorithm is much harder to control and may increase dramatically in unpredictable ways.

Variable selection bias can lead to developing models that fit well in the original set of samples but do not generalize well to external samples due to lack of true association between the model predictors and outcomes. An example would be a study trying to discriminate between 10 cancer and 10 control samples using 50 biomarkers. There would certainly be some combination of markers which would separate the two groups perfectly whether or not any real biological signal was captured by the markers. However, if a model were chosen and then applied to another set of 10 cancer and 10 control samples from a similar cohort of patients, it would be unlikely to validate. In such cases of overfitting, logistic regression models sometimes suffer from quasi or complete separation of the groups in the dataset at hand and the parameter estimates may fail to converge. Even if the model does converge, scenarios with small sample sizes and large numbers of predictor variables are unlikely to give very useful results.

In research, we often set the Type 1 error rate (α) for any particular testing procedure to be 5% for one hypothesis test and set the FWER to be 5% for multiple hypothesis tests. This means that we are willing to tolerate the mistake of declaring a result to be statistically significant about 5% of the time. If a variable selection algorithm was controlling the FWER appropriately, the percentage of final models selecting at least one predictor should be around 5% when all the predictors are noise.

2.1. Common approaches

One way researchers have attempted to control the FWER in classification problems is not to include too many predictor variables relative to the number of cases and controls. A common rule of thumb is to compute the events per variable (EPV), defined as the ratio of events (E: the smaller of either response category) to the number of predictor variables. Harrell (1985) found that the minimum EPV to give reliable results was around 10 by using simulation studies, later confirmed by Peduzzi (1996). For example, with 50 cases and 50 controls, one would not want to include more than five predictor variables because $50/5 = 10$. Although this rule of thumb may help to prevent convergence problems, it does not address variable selection bias.

One way to evaluate the FWER is to carry out a permutation test. This type of test is useful in assessing significance of a model produced from automated variable selection techniques because it preserves the correlation structure between the predictor variables and the p-value resulting from this test will be adjusted for multiple testing. One can construct a null distribution empirically for the statistic of interest by randomly permuting only the outcome variable and rerunning the procedure many times. After the null distribution of that statistic is constructed, the observed test statistic from the unscrambled dataset can be plotted and an adjusted p-value can be computed (Pesarin and Salmaso, 2010). Permutation based tests have been shown to have many desirable properties above and beyond the more common parametric based tests in biomedical research (Ludbrook and Dudley, 1998). Although this method seems to give better results, it requires the original dataset and considerable statistical programming experience to implement.

2.2. Variable selection techniques

A common strategy to reduce the number of predictors is to test each one individually with the response and then fit a final model using only the variables which were significant on univariate analysis. This is known as "data snooping" or "pre-screening" and has been shown to give poor results (Harrell et al., 1985). One problem with using this approach is that the reported p-values are not adjusted for multiple hypothesis testing; thus greatly increasing the chances of an inflated FWER. Another approach for variable selection is to use statistical software to select variables by maximizing or minimizing a pre-specified fit criterion. The Akaike's Information Criterion (AIC) has been shown to have useful properties for selecting variables (Beal, 2005). The formula for the AIC is given by $2(m+1) - 2\ln(L)$ where m is the number of predictor variables in the model and L is the likelihood function for the model (Akaike, 1974). The penalty $2(m+1)$ discourages overfitting because extra predictor variables which aren't very helpful will lead to a suboptimal model. The panel of predictor variables with the lowest AIC is deemed to be the best. The AIC criterion was used for our forward stepwise and best subsets scenarios. However, the software does not make any adjustment for multiple hypothesis tests which can inflate the FWER.

2.3. Assessing predictive performance

The area under the ROC curve (AUC) is a measure of discrimination between two groups and is standardly reported for binary classification models. The curve is constructed from several different cutpoints of a marker for which sensitivity and specificity values can be computed. Numerical integration is used to calculate the total area under the curve which can range from 0.5 (useless model) to 1.0 (perfect discrimination). The AUC can be calculated for a panel of markers using logistic regression models by calculating predicted probabilities and has been shown to be useful in hypothesis testing (Chen et al., 2013). The statistical significance of the AUC can be assessed using the Wilcoxon test statistic, which although it has been shown to be the appropriate test of significance for AUC on a pre-specified set of variables, does not adjust for multiple hypothesis tests (Hanley and McNeail, 1982).

The traditional test of significance for a logistic regression model is the Likelihood Ratio Test (LRT) (Agresti, 2007). One problem with using this test after variable selection is that the p-value is not very informative. It is difficult to construct an adjusted p-value for this test statistic (e.g., Bonferroni) because it is unknown a priori how many tests (or steps) will be run using these algorithms.

A method of valid post-selection inference or "PoSI" has been proposed in the literature (Berk et al., 2013). This method adjusts the standard errors for variables chosen by variable reduction techniques. However, the dataset is required to implement their method and it is limited to scenarios with less than 20 predictor variables (computation time restraint). Although the method may be helpful, it can be difficult to implement.

2.4. Forward stepwise

Perhaps the most common variable selection strategy for logistic regression is to use a statistically automated process which tests all (or several) combinations of potential predictor variables and chooses the best model for you. For instance, in epidemiology journals, stepwise selection methods were reported as the predominant method for variable selection (Walter and Tiemeier, 2009). In the forward stepwise method, the process involves a series of steps. In step one, all single variable models are run in order to see which variable is the best predictor on its own. In step two, all two variable models are tried after retaining the variable selected from step one. The process is repeated until adding extra variables does not improve the AIC or other fit criteria (Efroymson, 1960). Although many statisticians are aware of the limitations of implementing this method, this unadjusted procedure is routinely (if not prominently) implemented in the research community. One reason may be that it is widely taught in introductory regression courses without mentioning its limitations. It is also available and easy to implement in all standard statistical software packages. Other reasons for its popularity may include the ability for the researcher to feed in all variables without much thought, the speed at which the final model is produced, and because of the seemingly high accuracy and statistically significant p-values it typically produces.

2.5. Cross-validation

Cross-validation for forward stepwise models can provide a better estimate of prediction to external samples because the data for which the model was constructed are not included in the assessment of model performance. However, if the available dataset is small, there may be no sizable cross-validation set. The concept of leave-one-out cross-validation (LOO-CV) was introduced by Mosteller and Tukey (1968).

The approach has been extended to k-fold cross-validation, in which instead of one value being removed at a time, the dataset is split up into k partitions. If we set k equal to 10, the algorithm splits the dataset into 10 equal subsets and uses 9 of the subsets to build a model and predict values in the 10th subset. The process is then repeated 9 times, omitting a different subset each time. Classification ability for the "hold out" 1/10 part of the samples is then computed and averaged. This technique often gives some level of comfort to researchers because they assume this solves the variable selection bias problem, thinking that their results can better generalize to the population of interest. We assessed our stepwise models with both leave-one-out and 10-fold cross-validation. Although cross-validation techniques can be extended to other selection strategies, they are often not reported.

2.6. Best subsets

Best subsets is a technique which is similar to forward stepwise, except instead of selecting variables in a series of steps, it models all possible combinations of variables (Hosmer et al., 1989). The process includes fitting every one variable model, two variable model, three variable model, and so on. The total number of models assessed will be $2m - 1$ where m is the number of predictor variables. With only five variables, there are 31 models to test and with ten variables, there are 1023 models. An advantage of this approach is that all possible combinations of variables are given a chance to appear together. An obvious disadvantage is that the computation time required when m is greater than 10 can be quite large. The final model is chosen by observing the minimum AIC or other fit criteria from all models. This method has high potential for variable selection bias.

2.7. LASSO

A solution proposed by Tibshirani to the variable selection problem is the least absolute shrinkage and selection operator (LASSO) selection technique (1996). LASSO is a regularization procedure which places a bound on the sum of the absolute values of the regression coefficients, shrinking them towards zero with some being exactly zero. The shrinkage is controlled by the regularization parameter, λ, which is commonly chosen with cross-validation. LASSO has been shown to perform quite well in small datasets and likely better than stepwise (Steyerberg et al., 2000). Implementing the LASSO procedure generally does not generate test statistics/p-values for the coefficient estimates although recent methods have been proposed (Lockhart et al., 2014). The predicted probabilities from the model coefficients produced by LASSO can be extracted and those can be used to generate ROC curves and significance can be assessed as described earlier.

2.8. Method

The goal of our simulation was to assess the frequency of convergence problems, the rate at which any noise predictors are selected, and to compute the FWER. We evaluated forward stepwise, best subsets, and LASSO variable selection techniques. From the simulation results, AUC thresholds which control the FWER at our chosen alpha level of 5% were tabulated and plotted.

The first input for the simulation is the sample size, which we define as the number of events and equal number of non-events (total sample size = 2E). The other inputs include the number of noise predictor variables and selection technique (Table 1). One thousand iterations were performed for each combination of the simulation inputs setting $\alpha = 0.05$ for each model. Independent predictor variables were randomly generated from the standard normal distribution for each of the 1,000 iterations of the simulation.

First, a dataset was simulated where Y is a $2E \times 1$ outcome vector and X is a $2E \times$ m0 design matrix where E is the number of events and non-events and m0 is the number of noise predictors. Forward stepwise, best subsets, and LASSO variable section methods were run for each of the 1,000 datasets and the AUC was extracted from each of the final fitted models. The 95th percentiles calculated from the simulations serve as thresholds to control the FWER at the pre-specified α level of 5%. Figures were constructed with number of events on the square root scale for ease of interpolation.

Table 1. Methods table.

Parameters	Number	Description
Events (E)	4	Events and non-events ($E = 10, 25, 50, 100$) Total Sample size $= 2 \times E$
Noise Predictors (m_0)	6	Number of potential noise predictor variables ($m_0 = 3,5,10,15,20,50$)
Variable selection technique	3	Forward stepwise, Best-subsets, LASSO
Iterations	1000	Simulations per each combination listed earlier

Table 2. Simulation results and proposed critical values.

Inputs		Stepwise Results				LASSO Results	
Events[a]	Predictor variables	Failed to converge	Noise predictors chosen[b]	Critical AUC	Critical AUC with CV[c]	Noise predictors chosen[b]	Critical AUC
10	3	1%	0(0–1)	0.85	0.75	0(0–1)	0.86
10	5	2%	1(0–2)	0.89	0.75	0(0–1)	0.90
10	10	32%	2(1–3)	0.92	0.75	0(0–2)	0.98
10	15	65%	2(1–3)	0.91	0.75	0(0–2)	1.00
10	20	86%	2(1–3)	0.93	0.85	0(0–2)	1.00
10	50	99%	4(4–4)	1.00	1.00	0(0–2)	1.00
25	3	0%	0(0–1)	0.72	0.64	0(0–1)	0.72
25	5	0%	1(0–1)	0.75	0.66	0(0–1)	0.75
25	10	0%	2(1–3)	0.82	0.70	0(0–1)	0.80
25	15	0%	3(2–4)	0.89	0.74	0(0–2)	0.85
25	20	4%	4(2–5)	0.93	0.74	0(0–2)	0.87
25	50	95%	6(4–8)	0.94	0.78	0(0–3)	0.95
50	3	0%	0(0–1)	0.65	0.60	0(0–1)	0.65
50	5	0%	1(0–1)	0.68	0.61	0(0–1)	0.68
50	10	0%	1(1–2)	0.71	0.64	0(0–1)	0.72
50	15	0%	2(1–4)	0.76	0.66	0(0–1)	0.74
50	20	0%	4(2–5)	0.79	0.68	0(0–1)	0.75
50	50	9%	13(8–16)	0.93	0.75	0(0–2)	0.82
100	3	0%	0(0–1)	0.61	0.57	0(0–1)	0.61
100	5	0%	1(0–1)	0.62	0.58	0(0–1)	0.63
100	10	0%	2(1–2)	0.66	0.60	0(0–1)	0.65
100	15	0%	2(1–3)	0.68	0.61	0(0–1)	0.67
100	20	0%	3(2–4)	0.69	0.62	0(0–1)	0.68
100	50	0%	9(7–13)	0.79	0.67	0(0–2)	0.73

[a] Total sample size = 2 × Events.
[b] Values reported as Median (IQR).
[c] 10-fold cross-validation.

2.9. Software

All analyses were performed in R Version 3.1.2 (http://www.R-project.org) utilizing the following functions: *bestGLM* for bestsubsets selection, *stepAIC* for forward stepwise selection, *cvbinary* for calculating cross-validated AUC values, *Lroc* and *pROC*for calculating AUC values and confidence intervals, *cv.glmnet* for the LASSO technique, and *mvrnorm*for generating the noise predictors. The source code is available upon request.

3. Results

We investigated three issues for each of the automatic variable selection techniques. The first was quantifying how often the models failed to converge. The second was observing the actual FWER, that is, how often any noise predictors were selected. The third was assessing the number of models reported as statistically significant from the Wilcoxon test on the predicted probabilities from the final chosen model when no included variables had been generated to be predictive. More realistic thresholds for distinguishing real statistical significance compared to software reported statistical significance were calculated for each of the different modeling and design scenarios (Table 2).

3.1. Forward stepwise

The forward stepwise models often had convergence problems in situations where the number of predictor variables was similar to the sample size. For example, in forward stepwise

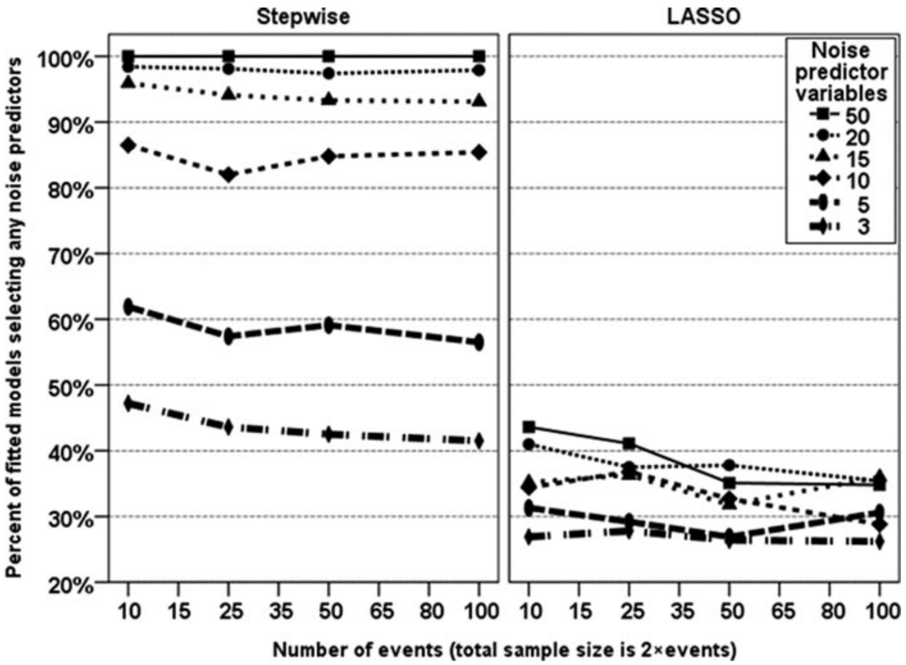

Figure 1. The percent of models selecting one or more noise predictor variables for Stepwise and LASSO. The 95% confidence interval for these points gives a margin of error around +/− 3%.

simulations with $E = 10$, 32% of models with 10 predictors failed to converge and the proportion which failed to converge jumped up to 65%, 86%, and 99% for 15, 20, and 50 predictors, respectively (Table 2). The other scenarios converged most of the time.

The number of noise predictors selected by forward stepwise was exceedingly high, indicating inadequate control of the FWER. With only three predictor variables, 40%–50% of the stepwise models selected at least one of the predictors across all sample sizes we modeled (Fig. 1). With five predictors, the percent of models with any noise predictors was around 60%. When we modeled 10, 15, 20, or 50 predictors, over 80% of the models included some noise predictors. One hundred percent of the models with 50 predictors across all sample sizes included at least one noise predictor. When testing even a small number of predictor variables, the percent of models selecting at least one of them was high. Whenever possible, the potential pool of predictor variables should be carefully considered and reduced in number before the stepwise process is implemented.

Another way to define the FWER is to be the percentage of fitted models reported to have a statistically significant AUC from the Wilcoxon Test. With only 3 noise predictors, the proportion of fitted models which were reported as statistically significant was around 20% for the 4 different sample sizes (Fig. 2). We observed a steady 10–20% increase in the proportion reported as significant as the number of predictors rose to 5, 10, 15, and 20. With 50 predictors, 100% of the scenarios incorrectly reported statistically significant final models. Whether we define the FWER to be any noise predictors selected or a statistically significant AUC of the fitted model, the FWER was well above 5%.

3.2. Cross-validation

The cross-validated stepwise AUC estimates were better than the unadjusted estimates, but still too high. The leave-one-out and 10-fold cross-validated AUC values were nearly identical,

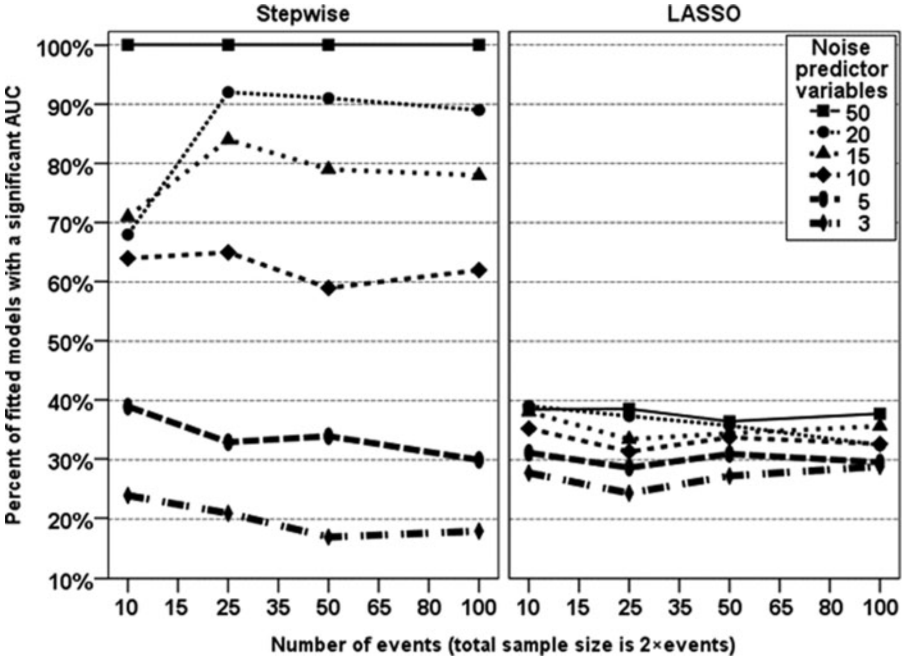

Figure 2. The percent of models reporting a statistically significant AUC at an unadjusted ?level of 0.05 for both Stepwise and LASSO. The 95% confidence interval for these points results in a margin of error around +/− 3%.

so we only presented the critical values for 10-fold CV (Table 2). The difference between the 95th percentiles of the unadjusted and CV adjusted AUC values was about 0.10 across all simulation scenarios. Using cross-validation did not fix the variable selection bias problem, but it did reduce the effect of variable selection bias compared to the unadjusted AUC.

3.3. Best subsets

The best subsets method yielded nearly identical results to forward stepwise in terms of convergence failures, variable selection, and critical value estimates for the AUC (results not shown). The average difference between the critical AUCs was about 0.01 between the two methods. The number of models run in the best subsets algorithm is much larger and requires a lot more computation time than forward stepwise, yet produces similarly poor results.

3.4. LASSO

Using the forward stepwise or best subsets method in these scenarios led to serious convergence problems, the selection of many noise variables, and far too many final models incorrectly reported as statistically significant. The cross-validation did not help correct the over optimism much either. However, LASSO does not have the same convergence problems as forward stepwise due to its regularization properties.

LASSO did better in terms of selecting fewer noise predictors than forward stepwise. The median number of noise predictors chosen by Stepwise ranged from 0 to 13 across all scenarios, whereas for LASSO the median number of predictors chosen was always 0 (Table 2). The proportion of models selecting any noise predictors was roughly half of the forward stepwise models across all scenarios (Fig. 1). In the most extreme cases with 50 predictor variables, all (100%) stepwise models selected at least one noise predictor. However, LASSO only selected

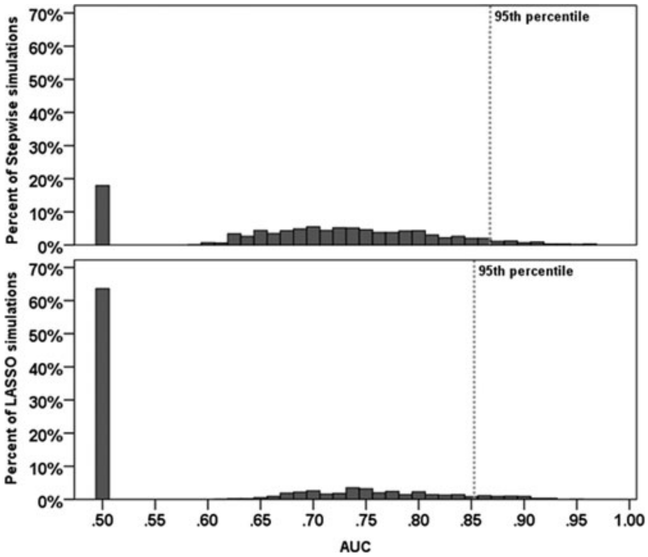

Figure 3. For the scenario with 20 events and 10 predictor variables the distribution of the AUC is plotted for both Stepwise and LASSO for the same 1,000 datasets. The critical value (95th percentile) for each method is shown by the dashed line.

a noise variable in 35%–40% of those specific simulations. Furthermore, the 95th percentiles of the AUC estimates for LASSO were similar to those for forward stepwise (Table 2). The distribution of the AUC between the two techniques for a specific scenario is shown in Fig. 3. Although the median and overall distribution of the AUC between the two methods is very different, the 95th percentiles are similar (Stepwise: 0.87, LASSO: 0.85). Therefore, none of the three methods performs well in terms of controlling the FWER.

3.5. Application

In order to compare the critical AUC estimate from our simulation with a critical AUC estimate from a permutation test, we analyzed two cancer biomarker examples. Our first example has 10 biomarkers measured in 20 cancer and 20 control samples ($E = 20$). After running the forward stepwise procedure, two predictors were selected for the model resulting in an observed AUC of 0.78 and a reported p-value of 0.002 (10-fold CV AUC of 0.70). We performed a specific simulation for this case by specifying m0 = 10 and $E = 20$, resulting in an estimated 5% critical AUC of just under 0.87. More generally, we can use interpolation with Fig. 4 and estimate the 5% critical AUC for this scenario to be around 0.85. The permutation test performed for this example resulted in a 95th percentile for the AUC of 0.83. For this study we would be skeptical of the significance of the observed model since the AUC of 0.78 fell below the simulated 5% critical value, suggesting that the set of markers may not be informative for discrimination in cancer status. This study was the scenario used in Fig. 3, which illustrates the higher frequency of no variables selected (AUC = 0.50) for LASSO, while the 95th percentile remains about the same.

The second example includes 15 biomarkers with 30 cases and 30 controls (m0 = 15 and $E = 30$). The forward stepwise procedure was run resulting in a final model with 1 variable selected and an AUC of 0.71 with a reported p-value of 0.006 (10-fold CV AUC of 0.63). The

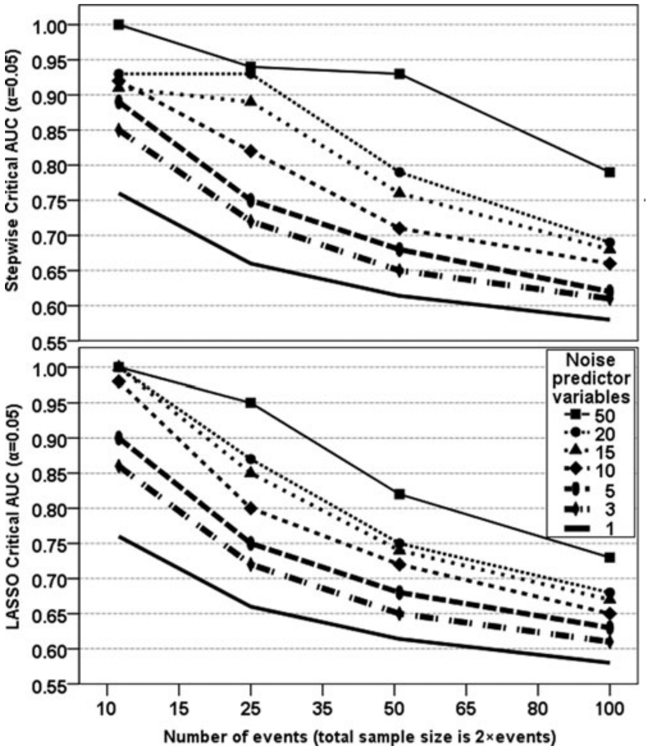

Figure 4. Our adjusted critical AUC values which would control the FWER at ? = 0.05 for Stepwise and LASSO. The 1 noise predictor scenario was added to this figure to show the unadjusted or naive significance level assuming no variable selection bias.

exact case specification yielded a simulated critical AUC of 0.85 with Fig. 4, also giving a value near 0.85 using interpolation. The permutation test resulted in a critical AUC of 0.83. Again our conclusion would be to view this final model skeptically because 0.71 failed to surpass the threshold.

The permutation test yielded approximately the same critical AUC values as our tables. For both examples, the simulated critical values between the exact scenarios and closest ones provided in Table 2 and Fig. 4 were essentially the same. The distribution of the AUC for the stepwise simulations looked very similar to the distribution from the permutation test. Our tables provide similar information to the permutation test without needing to obtain the data or running the specific test ourselves.

The critical values in Table 2 and Fig. 4 can be used to give a better assessment of true significance without relying on obtaining the specific dataset. Scenarios not directly simulated may be interpolated from Fig. 4 or determined from the code accessible in the uploaded digital content portion of the journal. This result gives researchers a tool to quickly assess the significance and believability of studies using variable selection algorithms.

3.6. Limitations

Although our method controls the FWER, there are limitations. Our method was only tested with an equal number of events and non-events. The predictor variables were all generated to be independent and normally distributed which is often not true in biomarker research. However, for two application problems which included correlated and skewed markers, our method gave similar critical values to the permutation test, which does account for the correlation structure of the predictor variables because only the outcome variable is permuted. The AIC was used as the information criterion, and others such as BIC and AICC were not explored. We presented several common selection techniques, but many more exist such as least angle regression (LARS), classification and regression trees (CART models), neural networks, Dantzig selector, elastic net, gradient boosting, and support vector machines (SVM), to name a few. However, we wanted to present critical values for relatively simple techniques standardly used by clinical researchers and not focus on finding the "best" and most current method.

4. Conclusion

We have found that using automated variable selection techniques often leads to artificially high AUC values when the generated predictor variables have no underlying predictive ability. Cross-validation methods and more recent selection strategies (LASSO) helped slightly, but did not overcome this problem. We have demonstrated for many common scenarios that the traditional statistical significance level reported for the AUC in a logistic regression model after variable selection is inaccurate and should not be relied upon. Many statisticians recommend that researchers should not to use automated methods but these methods are still frequently implemented. Therefore, the tables and figures presented in this article are relevant because they better control the FWER at a pre-specified alpha level of 0.05.

Financial support

This research was supported by NIH/National Center for Advancing Translational Science (NCATS) UCLA CTSI Grant Number UL1TR000124.

Competing interests

The authors declare that they have no competing interests. The authors alone are responsible for the content and writing of the article.

Authors' contributions

DAE conceived the study and participated in its design and interpretation of the results. TRG carried out all statistical analyses, constructed all tables/figures, and wrote the manuscript. Both authors read and approved the final version.

Acknowledgments

This research was supported by NIH/National Center for Advancing Translational Science (NCATS) UCLA CTSI Grant Number UL1TR000124.

ORCID

Tristan R. Grogan http://orcid.org/0000-0001-9471-2938

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* 2nd. ed. New Jersey: John Wiley and Sons, Inc.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

Beal, D. (2005). Selecting the best multiple linear regression model for multivariate data using information criteria. *SESUG SAS Institute* paper SA105_05.

Begley, G. C., Ellis, M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature* 483:531–533.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics* 41:802–837.

Chen, W., Samuelson, F. W., Gallas, B. D., Kang, L., Sahiner, B., Petrick, N. (2013). The assessment of the added value of new predictive biomarkers. *BMC Medical Research Methodology* 13:98.

Efroymson, M. A. (1960). *Multiple regression analysis: Mathematical Methods for Digital Computers*. New York: John Wiley.

Fernandesâ€"Taylor, S., Hyun, J. K., Reeder, R. N., Harris, A. H. S. (2011). Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Research Notes* 4:304–308.

Ferraris, V. A., Ferraris, S. P. (2003). Assessing the Medical Literature: Let the Buyer Beware. *The Annals of Thoracic Surgery* 76:4–11.

Fong, D. Y. T., Lee, C. F., Lau, S. P. (2008). Contingency table analysis in obstetrics and gynaecology. *Hong Kong Journal Gynaecology, Obstetrics and Midwifery* 8:42–50.

Gandhi, R., Smith, H. N., Mahomed, N. N, Rizek, R., Bhandari, M. (2011). Incorrect use of the student *t* test in randomized trials of bilateral hip and knee arthroplasty patients. *Journal of Arthroplasty* 26:811–816.

Hanley, J. A., McNeail, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36.

Harrell, F. E.Jr., Lee, K. L., Matchar, D. B., Reichert, T. A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports* 69:1071–1077.

Hiltzik, M. (2013). Science has lost its way, at a big cost to humanity. *The LA Times*. Retrieved from http://www.latimes.com

Hosmer, D. W., JovanovicB., Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics* 45:1265–1270.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* 42:413–468.

Lucena, C., Lopez, J. M., Abalos, C., Robles, V., Pulgar, R. (2011). Statistical errors in microleakage studies in operative dentistry.A survey of the literature 2001–2009. *European Journal of Oral Sciences* 119:504–510.

Ludbrook, J., Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician* 52:127–132.

Marozzi, M. (2014). Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Statistical Methods in Medical Research* 0:1–18. doi: 10.1177/0962280214529104.

Marozzi, M. (2015a). Does bad inference drive out good? *Clinical and Experimental Pharmacology and Physiology* 42:727–733.

Marozzi, M. (2015b). Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Statistics in Medicine* 34:1511–1526.

McKinney, P. W., Young, M. J., Hartz, A., Lee, M.B. (1989). The inexact use of fisher's exact test in six major medical journals. *Journal of the American Medical Association* 261:3430–3433.

Mosteller, F., Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of Social Psychology* 2: 1–26.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49:1373–1379.

Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications, and Software*. New Jersey: John Wiley and Sons.

Podoll, A. S., Bell, C. S., Molony, D. A. (2012). Evidence-based practice in nephrology: critical appraisal of nephrology clinical research: Were the correct statistical tests used? *Advances in Chronic Kidney Disease* 19:27–33.

Prinz, F., Schlange, T., Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10(712):712–713.

Reality Check on Reproducibility. (2016, May 16). Editorial. *Nature* 533:437.

Shen, H., Xu, W., Zhang, J., Chen, M., Martin, F.L., Xia, Y., hellip; Zhu, Y. G. (2013). Urinary metabolic biomarkers link oxidative stress indicators associated with general arsenic exposure to male infertility in a Han Chinese population. *Environmental Science and Technology* 47:8843–8851.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. England: The Penguin Press.

Steyerberg, W., Eijkemans, J. C., Harrell, F., Habbema, J. (2000). Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine* 19:1059–1079.

Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., Ulmer, H. (2007). The use of statistics in medical research: A comparison of *the new England journal of medicine* and *nature medicine*. *The American Statistician* 61:47–55.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* 58:267–288.

Walter, S., Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology* 12:733–736.

Zhang, J., Huang, Z., Chen, M., Xia, Y., Martin, F. L., Hang, W., Shen, H. (2014). Urinary metabolome identifies signatures of oligozoospermic infertile men. *Fertility and Sterility* 102:44–53.