On the Discovery and Use of Disease Risk Factors with Logistic Regression: New Prostate Cancer Risk Factors

David E. Booth[1], Venugopal Gopalakrishna-Remani[2], Matthew Cooper[3], Fiona R. Green[4] Margaret P. Rayman[5]

[1]M&IS Dept., Kent State University, Kent OH 44242, [2]Dept. of Management, University of Texas-Tyler, Tyler, TX 75799, [3]Dept. of Internal Medicine, Washington University School of Medicine, St. Louis MO, 63110, [4]University of Manchester, UK, [5]Dept. of Nutritional Sciences, University of Surrey, Guildford GU27XH UK

Abstract

We begin by arguing that the often used algorithm for the discovery and use of disease risk factors, stepwise logistic regression, is unstable. We then argue that there are other algorithms available that are much more stable and reliable (e.g. the lasso and gradient boosting). We then propose a protocol for the discovery and use of risk factors using lasso or boosting variable selection with logistic regression. We then illustrate the use of the protocol with a set of prostate cancer data and show that it recovers known risk factors. Finally we use the protocol to identify new risk factors for prostate cancer.

1. Introduction

As Austin and Tu (2004) remark, researchers as well as physicians are often interested in determining the independent predictors of a disease state. These predictors, often called risk factors, are important in disease diagnosis, prognosis and general patient management as the attending physician tries to optimize patient care. In addition, knowledge of these risk factors help researchers evaluate new treatment modalities and therapies as well as help make comparisons across different hospitals (Austin and Tu, 2004). Because risk factors are so important in patient care it behooves us to do the best job possible in the discovery and use of disease risk factors. Because new statistical methods (Ayers and Cordell (2010), Yuan and Liu (2006), Steyerberg et al (2000), Wiegand (2009), Breiman (1995), Tibshirani (1996), Dahlberg (2010), Efron and Hastie (2016)) have been and are being developed, (Dahlgren (2010)) it is important for risk factor researchers to be aware of these new methods and to adjust their discovery and use of risk factor protocols as is necessary. In this paper, we argue that now is such a time. For a number of years in risk factor research a method of automatic variable selection called stepwise regression and its variants forward selection and backward elimination (Chatterjee and Price 1977 (chapter 9)) have been used even as new methods have become available (see Neter, Wasserman and Kutner, 1983, Chapter 12, Kutner et al 5th ed. 2005 P 364ff, Labidi et al 2009, Queiroz et al 2010, Qui et al 2013, Guo et al 2016, Khan et al 2016 and many others). The last three cited are risk factor studies. We do not argue for a change of protocols in risk factor discovery and use simply because newer methods are available. As literature shows(Austin and Tu, 2004) the older methods are often unreliable and the newer methods are much less so. We point out that the purpose of this paper is the following:

1. To summarize some of the studies that show that stepwise regression and its variants, as now used more often than they should be in risk factor studies ,are unreliable and in fact may cause some of the irreproducibility of life sciences research as discussed by Arnaud (2014) as we shall discuss later.

2. To argue on the basis of current research that there are methods available that are considerably more reliable.
3. To propose a modern statistical protocol for the discovery and use of risk factors when using logistic regression as is commonly done.
4. To illustrate the use of the protocol developed in 3 using a set of prostate cancer data (Cooper et al 2008).
5. To report the finding of new prostate cancer risk factors using the modern procedures.

We further note that nothing in the way of statistical methods is new in this paper. What is new is the introduction of a clear protocol to identify and use disease risk factors that involve much less problematic methods than stepwise regression. We then use the proposed methodology to identify a known prostate cancer risk factor and then discover new prostate cancer risk factors.

2. What then should replace these automatic variable selection methods?

From the references in Section 1, we see that the shrinkage methods have done well when compared to the current stepwise and all subsets methods and thus we follow the suggestion of Steyerburg et al and look at shrinkage methods. The question then becomes what shrinkage method might we choose as the next variable selection method? We are impressed by the work of Ayers and Cordell (2010) in this regard. First we note that shrinkage estimators are also called penalized estimators. In particular the lasso (Tibshirani 1996) as defined by Zou (2006) can be considered. We note that the factor lambda is said to be the penalty.

Now Ayers and Cordell (2010) studied "the performance of penalizations in selecting SNPs as predictors in genetic association studies." Their conclusion is: "Results show that penalized methods outperform single marker analysis, with the main difference being that penalized methods allow the simultaneous inclusion of a number of markers, and generally do not allow correlated variables to enter the model in which most of the identified explanatory markers are accounted for." At this point, penalty estimators (i.e. shrinkage) look very attractive in risk factor type studies. (Efron and Hastie (2016), Chapter 16.)

Another paper (Zou 2006) helps us make our final decision. Zou (2006) considers a procedure called adaptive lasso in which different values of the parameter λ are allowed for each of the regression coefficients. Furthermore, Zou shows that an adaptive lasso procedure is an oracle procedure such that $\beta(\delta)$ (asymptotically) has the following properties

a) It identifies the right subset model and
b) It has the optimal estimated rate.

Zou then extends these results to the adaptive lasso for logistic regression. Wang and Leng (2007) developed an approximate adaptive lasso (i.e. a different λ for each β is allowed) by least squares approximation for many types of regression. Boos (2014) shows how easy it is to implement this software in the statistical language R for logistic regression. Thus, we choose to use the least squares approximation to their adaptive lasso logistic regression in the

next section. We note here that a special variant of lasso, group lasso (Meier et al (2008)) is needed for categorical predictor variables.

In the next section, we propose and discuss a protocol for the discovery and use of risk factors in logistic regression models. In the following section we illustrate the use of the protocol using the data of Cooper et al (2008) to look at some risk factors for prostate cancer. We will show that currently known risk factors can be identified as well as new risk factors discovered using these methods.

In addition a new method of variable selection called gradient boosting has been developed. (Ridgeway (2015), Kendziorski (2016) James et al (2013), Chapter 8, Maloney et al (2012), Efron and Hastie (2016), Chapter 17.) This method has some of the same advantages as lasso and we add it to the protocol and test it as well.

3. A suggested protocol for using logistic regression to discover and use disease risk factors.

Our suggested protocol is shown below. We discuss the protocol in this section and illustrate its use with prostate cancer risk factors in the following section. This protocol uses the R statistical language.

The Logistic Regression

Protocol for use with Risk Factors

1. Ready data for analysis.
2. Input to R.
3. Regress a suitable dependent variable ((say) 0- Control, 1 – Has disease) on X (a potential risk factor) as described by Harrell(2001 Chapter 10) for logistic regression.
4. Select a set of potential risk factors. If an X variable is continuous, we suggest use of the Bianco-Yohai robust (outlier resistant, see Hauser and Booth (2011)) estimator and further suggest putting outliers aside for further analysis as they may give rise to extra information.
5. Now build a full risk factor prediction model.
6. Use potential risk factors (Xs) to form a full model with the appropriate dependent variable (as in 3).
7. If any variables are continuous repeat 4 using the entire potential full model.
8. With any outliers set aside for further study, regress the dependent variable on the logistic regression full model using the adaptive lasso method, least squares approximation, as described by Boos (2014) which is easiest in R.
9. Using a Bayesian Information Criterion (BIC) select variables without zero lasso regression coefficients to be predictors in a risk factor based reduced model. If categorical risk factors are present use group lasso regression (Meier et al (2008)). Use graphs like Fig. 1 in Meier et al (2008) to identify the zero lasso regression coefficients that may exist for the categorical variables.
10. Repeat Step 8 for gradient boosting as described by Kendziorski (2016).
11. Validate the reduced model, with the similar validation of the full model of step 6, if there is any doubt about variables discarded from the full model using bootstrap cross validation (Harrell, 2001) and then check the usual model diagnostics (Pregibon, 1981) for either lasso or boosting or both.

12. Predict with the reduced model containing the appropriate risk factors as described in Harrell (2001), Chapter 11 and Ryan (2009), Chapter 9.

Notes to the protocol.

A. We note that for the genome wide case of predictors one should refer to Li et al (2011) and Wu et al (2009).
B. All logistic regression assumptions should be checked and satisfied as in Pregibon (1981).

5.The prostate cancer example including new risk factors

This example is taken from Cooper et al (2008) where the data and biological system are described. The data set used in this paper is a subset of the Cooper et al data set with all observations containing missing values removed. We note that all potential predictor variables are categorical so no imputation was performed. The coding assignments and the variable definitions are given in the Appendix. The simple and multiple logistic regressions are carried out as described in Harrell (2001). Robust logistic regressions, when needed, are carried out as described in Hauser and Booth (2011). Variable selection is carried out using the adaptive lasso (Zou, 2006) with the least squares approximation of Wang and Leng (2007) for continuous independent variables and by group lasso (Meier et al (2008)) for categorical independent variables. Gradient boosting is carried out using R Package gbm Ridgeway (2015) as described by Kendziorski (2016), Ho (2012), Maloney (2012). All computations are carried out using the R statistical language. The R functions for variable selection (adaptive lasso and group lasso) along with the papers are available from Boos (2014), and used as described there. The use of the group lasso R function is covered in R help for packages grplasso and grpreg. The data sets and R programs are available from the authors (DEB). The variables studied as potential risk factors are listed in the X column of Table 1. The dependent variable is current status.

We now follow the protocol and explain each step in detail. We begin by considering the one predictor logistic regressions in Table 1. First note that all potential risk factors in this data set are categorical (factors) so we do not have to consider the Bianco-Yohai (Bianco and Martinez (2009)) estimator of protocol Step 4 for this data. Cooper et al (2008) hypothesize a SNP-SNP interaction as a risk factor for prostate cancer. We now test this hypothesis and attempt to answer the question is there such an interaction? In order to answer this question, we first note that the answer is not completely contained in Table 1. Second, we recall that we have a gene-gene interaction of two genes if both affect the final phenotype of the individual together. To be specific, we now consider the two genes representing the relevant alleles of the SEPP1 and SOD2 genes. If there is a gene-gene interaction, we must see the following statistically. The relevant alleles of the SEPP1 and SOD2 genes must be selected to be in a reasonable prediction equation for the disease state by the appropriate lasso or boosting algorithm (see Figures 1,2,3, and 4). The appropriate lasso algorithm here is the group lasso for logistic regression because the predictor variables are categorical. We now note that in our data set we have four candidate predictor variables from which to search for our gene-gene interaction MnSOD_DOM_Final, SeP_Ad_Final, MnSOD_AD_Final and SeP_DOM_Final. Either observation of the Variable Values or a simple trial shows that we cannot include all four variables in the model at once because they are pairwise collinear. Hence we have to separate the variables into the two cases, the models of Figure 1 and Figure 2. We also note that lasso generally does not allow correlated variables to enter the model (Ayers and Cordell(2010)).

We now begin our search using lasso with the model of Figure 1. This gives us a candidate for an interaction. We then perform the group lasso analysis of Figure 1. Here we must determine if the relevant alleles are included in the group lasso selected prediction equation. Roughly this is the case if the lasso regression coefficients are not zero at the end of the algorithm's execution as shown on the coefficient path plot of Figure 1. By looking at equation (2.2) of Meier et al (2008) we see that $0 \leq \lambda < \infty$ hence as $\lambda \to \infty$, $s_\lambda(\beta) \to 0$ and thus $\beta_i \to 0$ but not uniformly. Hence the question is what value of λ do we choose to determine if the coefficients are close enough to zero to discard that term from the model as a zero coefficient. Based on Table 2 where we compute the

Table 1 Simple Logistic Regression Results
Dependent variable CURRENTSTATUS Intercepts are not listed

X		Coeff.	SE	P
X_STRATUM		-.055132	.005646	$<2x10^{-16}$
MnSOD_AD_Final	0	-0.4334	.1241	0.000477
	1	-0.2478	.1157	0.032196
	2	-0.3140	.1233	0.010879
SeP_Ad_Final	0	0.21219	0.10309	0.039557
	1	0.12890	0.10754	0.230675
	2	0.23484	0.15797	0.137117
MnSOD_DOM_Final	0	0.4334	0.1241	0.000477
	1	0.2704	0.1126	0.016369
SeP_DOM_Final	0	0.21219	0.10309	0.039557
	1	0.14445	0.10568	0.171679
Smoke_ever	0	-.00339	.08161	0.967
	1	-.03791	.07016	0.589
Alco_ever	0	-0.428943	0.142425	0.0026
	1	0.002951	0.062317	0.9622
FAMHIST		0.84619	0.09497	$<2x10^{-16}$

Table 2

Optimal λs Computed from R Packages

grplasso and grpreg for Indicated Models

Predictors in Model	λmin	λmax	λopt			
MnSOD_AD_Final SeP_DOM_Final	.009	70.55	.635			
MnSOD_DOM_Final SeP_Ad_Final	.017	83.99	1.428			

Note: λmin computed by package grpreg using a Bayesian Information Criterion λmax was computed by package grplasso.

Figure 1 – The Group Lasso Coefficient plot for the logistic regression –

Containing MnSOD_DOM_FINAL and SeP_Ad_Final

We note that for lambda=λopt none of the paths shrink to zero suggesting that a SNP-SNP interaction, as reported in Cooper et al (2008) exists.

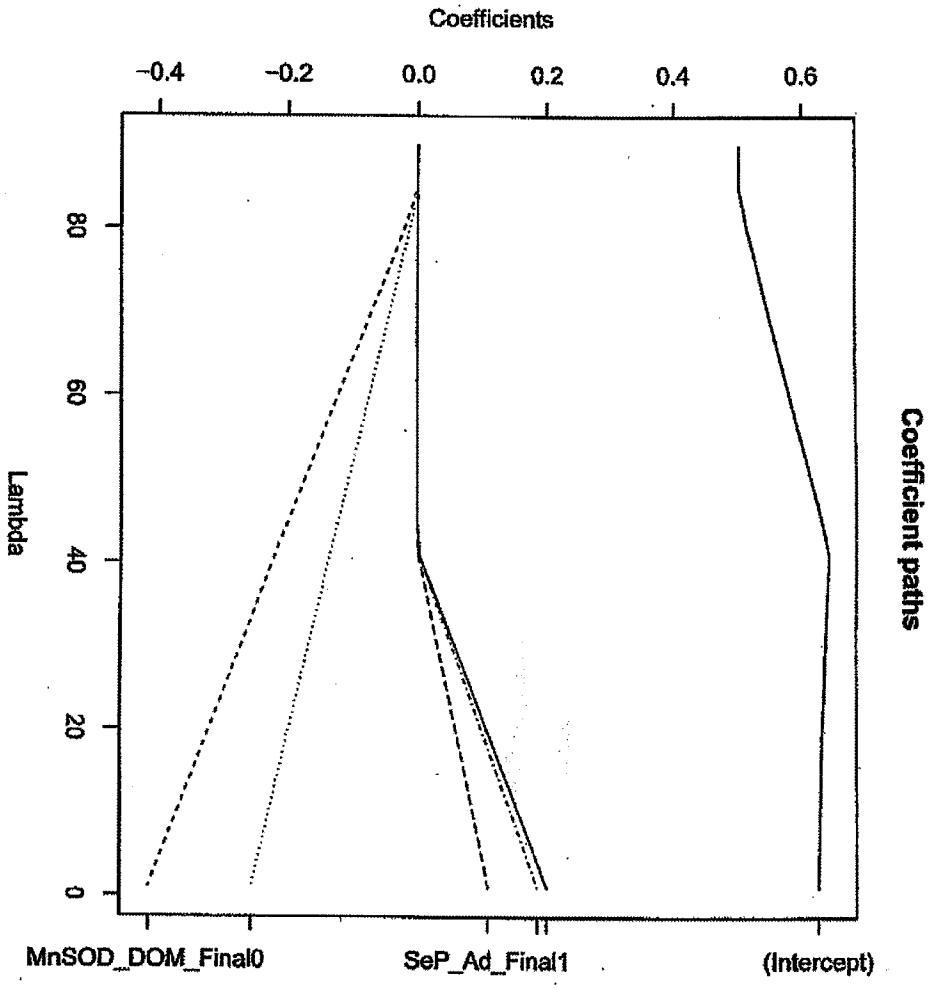
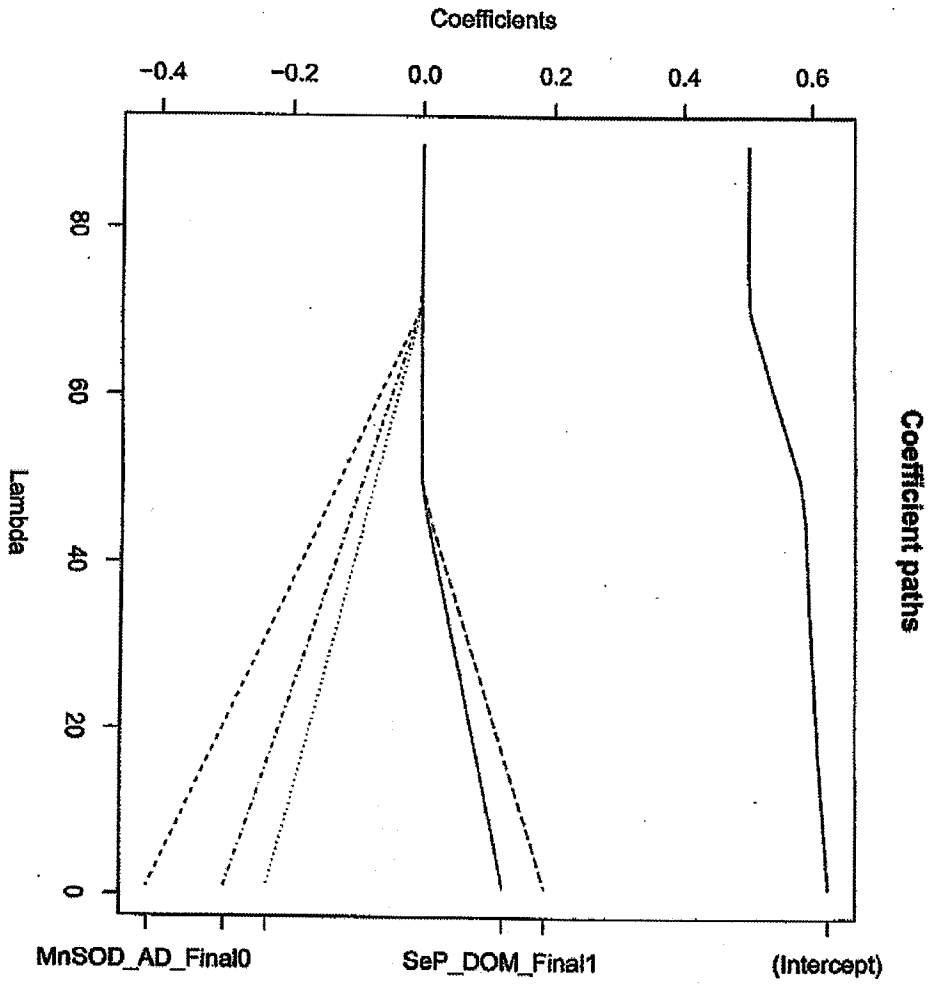

Coefficients
−0.4
−0.2
0.0
0.2
0.4
0.6
Lambda
0
20
40
60
80
MnSOD_DOM_Final0
SeP_Ad_Final1
(Intercept)
Coefficient paths

Figure 2 – Grouplasso Coefficient Plot for Model Containing MnSOD_AD_Final and SeP_SOM_Final


Coefficients
−0.4
−0.2
0.0
0.2
0.4
0.6
Lambda
0
20
40
60
80
MnSOD_AD_Final0
SeP_DOM_Final1
(Intercept)
Coefficient paths

optimal λ to use we choose λ=1.428 to be the cutoff point. Hence we can now apply the condition of the previous paragraph. We now check Figure 1 to see which if any of these candidate alleles are selected for the group lasso prediction equation which was our criterion. We now examine the Figure 1 plot at λopt=1.428. We note that at this λ none of the candidate alleles have coefficients of zero. Hence using our criterion we can summarize as follows:

1. We need Figure 1 selection to show interaction. SeP_Ad_Final0 was Ala/Ala so this is one allele that qualifies. Similarly for SeP_Ad_Final1 and 2 which are Ala/Thr and Thr/Thr respectively.
2. Both MnSOD_DOM_Final0 and MnSOD_DOM_Final1 (i.e. Ala/Ala and +/Ala) satisfy so this shows that for MnSOD the result is +/Ala. Hence the identified interaction alleles are

SEPP1	SOD2
Ala/Ala	+/Ala

which agrees with the Cooper et al (2008) finding on a gene-gene interaction risk factor. Similarly we have from SeP_Ad_Final 1 and 2

Ala/Thr	+/Ala
Thr/Thr	+/Ala

which are also risk factors.

We now repeat this analysis for the model which contains the other possible candidate alleles. By our criterion for gene-gene interaction we need $\beta_i \neq 0$ for λopt=0.635, from observing Table 2. Now by observing Figure 2 we see that for MnSOD_AD_ Final the 0, 1 and 2 values meet the criteria while for SeP_DOM_Final only the 0 and 1 alleles do. By consulting the Appendix we see that

SeP_DOMFinal1 is ala/Thr and Thr/Thr

SeP_DOM_Final0 is Ala/Ala

MnSOD_AD_Final0 is Val/Val

1 is Val/Ala

2 is Ala/Ala

Hence we conclude that we have additional gene-gene interactions that are risk factors. Since one combination was identified using the first model. We now have

SEPP1	SOD2
Ala/Ala	Val/Val
Ala/Ala	Val/Ala
+/Thr	Val/Val
+/Thr	Val/Ala

as risk factors. None of these have been reported in the prior literature as far as we can determine

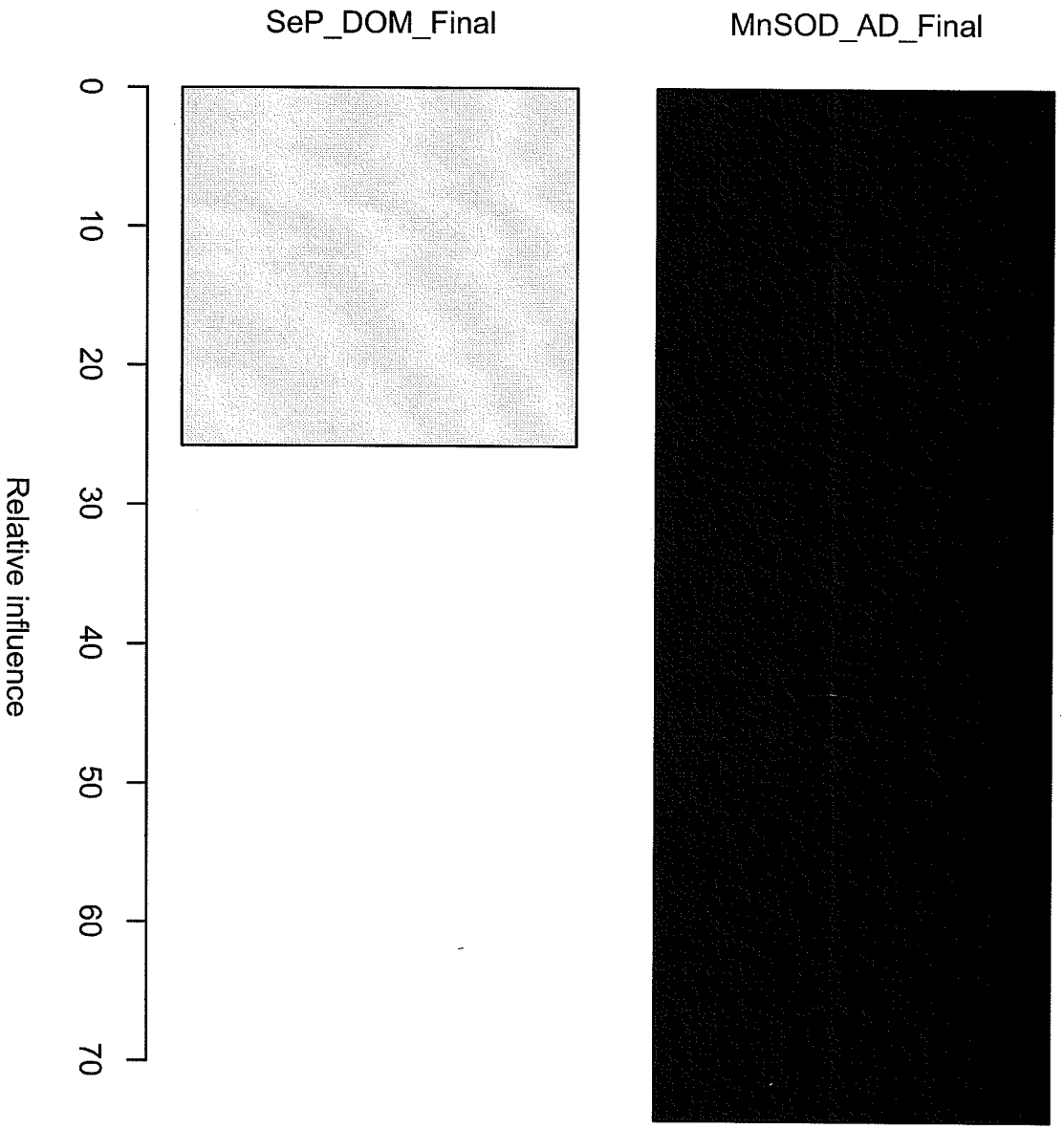
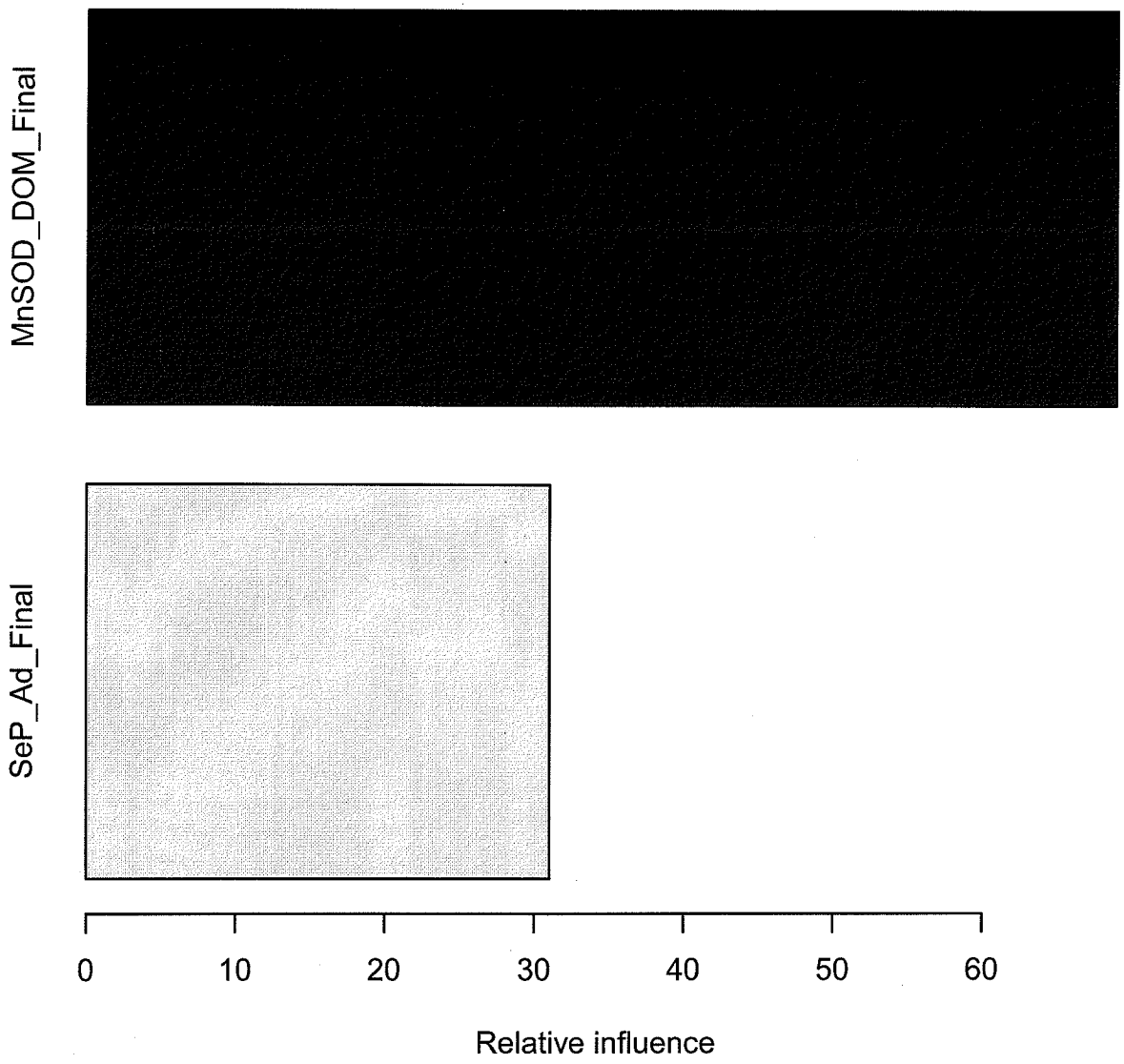


Fig. 3 Boosting Results Pkg gbm, AdaBoost Corresponds to Fig. 2

Figure 4 Corresponds to Fig.1, Same Conditions as in Fig. 3



We can now make prediction equations using our now known risk factors which will give our predicted diagnosis of whether or not a patient is at risk for prostate cancer based on our variable values assuming that we use a new observation not one which is included in our current data set.. We recommend the use of bootstrap cross validation to validate this equation and full details are included in (Harrell, 2001). As a final reminder, all of the other assumptions of logistic regression need to be checked each and every time. The reader is referred to Pregibon (1981) for further details. These new risk factor results are particularly important since the SEPP1 gene product is in the same metabolic path as a tumor suppressor for prostate cancer (Ansong et al 2015).

We now repeat the analysis using gradient boosting. The results are shown in Figures 3 and 4. The results are identical to the lasso results.

6.Limitations of the proposed Protocol and Future Research

As much as we would like this to be the last word on the discovery and use of disease risk factors with logistic regression, it is not. We will mention a few possible limitations and our hope for some future work perhaps by us or others that we would like to see.

First, Ayers and Cordell (2010) mention a limitation of this suggestion, the fact that there is no known way to get confidence intervals and p-values for lasso estimates. Fortunately this is changing. Currently, there is a paper by Lockhart et al (2012) entitled "A significance test for the lasso". While this is a complicated paper that doesn't solve all problems a strong beachhead has been established.

Next, we discussed the advantages of adaptive lasso earlier (esp. the oracle property) but no algorithm currently exists to solve the adaptive group lasso problem in the case of logistic regression. We conjecture based on the results of the linear regression case extended to the logistic case that if we could extend adaptive lasso to the group lasso for logistic regression cases that the same desirable properties of adaptive lasso would hold, especially the oracle property.

Finally the usual problems of outliers, etc., as always, raise their head. The Bianco-Yohai algorithm (Bianco and Martinez (2011)) is a start but this hasn't been extended to any penalized shrinkage regression method. We conclude that there is much work to be done and fully expect to see other papers like this one in the future and hopefully statistical practice can continue to evolve and even better solutions can be applied to these interesting and important problems.

7. Conclusion

We have attempted in this paper to bring up to date statistical thinking to the problem of the identification and use of disease risk factors, where stepwise regression is still too often used. Much remains to be done, but we hope that the ideas presented here will improve statistical practice in this very important area. In the process of bringing this thinking up to date, we have shown that we recover a currently known risk factor and identify new risk factors which suggest the value of our approach. These new risk factor results are particularly important since the SEPP1 gene product has recently been shown to be in the same metabolic pathway as a tumor suppressor for prostate cancer (Ansong et al 2015)

Appendix Data Set Variable

INCLUSIONSTATUS	Cancer status at inclusion	0 = Control 1 = Cancer
X_INCLUSIONAGE_YRS	Age	Age
CURRENTSTATUS	Updated cancer status	0 = Control 1 = Cancer
T	T- Stage	Staging 1 to 4 -1 = Control 9 = No data
N	N - Stage	0 = N- 1 = N+ -1 = Control 99 = No data
M	M - Stage	0 = M- 1 = M+ -1 = Control 99 = No data
DIFF	Tumour Differentiation	Staging 1 to 3 -1 = Control 99 = No data
GLEASON	Gleason Score	Staging 1 to 10 -1 = Control 99 = No data
PSA	Prostate specific antigen	µg/ml -1 = Data not available -2 = Control
ADV	Advanced stage cancer in at least one of the above markers (TNM, Diff, Gleason, PSA) See below for how the cancers were classified	0 = Not aggressive 1 = Aggressive -1 = Control 99 = No data
X_STRATUM	Stratification of data based on age and geographical location	
FAMHIST	Family history	0 = No 1 = Yes
smoke_ever	Smoking	0 = Never 1 = Ever 99 = Data missing
alco_ever	Alcohol consumption	0 = Never 1 = Ever 99 = Data missing
X_BMI	Body Mass Index	-1 = No Data 1<, = BMI
MnSOD_AD_Final	SOD2 Genotype	0 = Val/Val 1 = Val/Ala 2 = Ala/Ala
MnSOD_DOM_Final	SOD2 Dominant Model	0 = Val/Val 1= Val/Ala and Ala/Ala
SeP_Ad_Final	SePP1 Genotype	0 = Ala/Ala 1 = Ala/Thr 2 = Thr/Thr
SeP_DOM_Final	SePP1 Dominant Model	0 = Ala/Ala 1 = Ala/Thr and Thr/Thr
inclusion_age_banded	Age banded within 10 years	
Ad_control_100_final	Aggressive and Control. All other cases excluded	0 = Control 1 = Aggressive
Loc_control_100	Non-aggressive and Control. All other cases excluded	0 = Control 1 = Non Aggressive

Cases were classified as either non-aggressive at diagnosis (tumor stage 1 and 2, Gleason score < 8, Differentiation G1-G2, NP/NX, MO/MX, PSA < 100 μg/L; NPC) or aggressive at diagnosis (tumor stage 3-4, Gleason score ≥ 8, Differentiation G3-G4, N+, M+, PSA ≥ 100 μg/L;APC).

References

Ansong, E., Ying, Q., Ekoue, D. N., Deaton, R., Hall, A. R., Kajdacsy-Galla, A., Yang, W., Gann, P. H., Diamond, A. M. (2015) Evidence that Selenium Binding Protein 1 is a Tumor Suppressor in Prostate Cancer. PLoS ONE 10(5); e0127295. doi:10.1371/jouenal.pone.0127295

Arnaud, D. H. (2014), Confronting Irreproducibility, Chemical and Engineering News 92 (50), 28-30..

Austin, P. and Tu, J (2004), Automated Variable Selection Methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality, J. Clinical Epidemiology 57, 1138-1146.

Ayers, K and Cordell, H (2010), SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression, Genetic Epidemiology 34: 879-891.

Bianco, A and Martinez, E. (2009), Robust testing in the logistic regression model, Computational Statistics and Data Analysis 53, 4095-4105.

Boos, D. (2014) au., Adaptive Lasso in R, 2/9/2014, http://www.stat.ncsu.edu/~boos/var.select/lasso.adaptive.html

Breiman, L (1995), Better Subset Regression Using the Nonnegative garrote, Technometrics 37 (4), 373-384

Chatterjee, S and Price B, (1977), Regression Analysis by Example, New York: John Wiley and Sons.

Cooper, M., Adami, H., Gronberg, H., Wiklund, F., Green, F., Rayman, M. (2008), Interaction between Single Nucleotide Polymorphisms in Selenoprotein P and Mitochondrial Superoxide Dismutase Determines Prostate Cancer Risk, Cancer Res 2008: 68: (24), 10171-10177

Dahlgren, J (2010), Alternative Regression Methods are not considered in Murtaugh (2009) or by ecologists in general, Ecology Letters (2010) 13: E7-E9.

Efron, B. and Hastie, T. (2016), Computer Age Statistical Inference, Cambridge, Cambridge University Press.

Guo, L., Guo, X., Chang, Y., Yang, T., Zhang, L., Li, T., and Sun, Y. Prevalence and Risk Factors of Heart Failure with the Preserved Injection Fraction, Int. J. Environ., Res. Public Health 2016, 13(8), 770.

Harrell, Jr., F., (2001) Regression Modeling Strategies; New York: Springer.

Hauser, R. and Booth, D (2011), Predicting Bankruptcy with robust logistic regression, J. Data Sci 9(4), 585-605.

Ho, R. (2012), Big Data Machine Learning, DZoneRefCard z #158, Carey NC:DZone Inc..

James, G., Wilten, D., Hastie, T., and Tibshwani, R. (2013), An Introduction to Statistical Learning, N.Y.: Springer.

Kendziorski, C. (2016), https://www.biostat.wisc.edu/~Kendzior/stat877/illustration.pdf accessed 9/1/2016.

Khan, MS, Pervaiz, MK, Javed, I, Biostatistical Study of Clinical Risk Factors in Myocardial Infarction, PAFMJ 2016; 66(3): 354-360.

Kutner, M, Nachtsheim, Neter, j, Li, W, (2005) Applied Linear Statistical Models, 5th ed., New York; McGraw-Hill Irwin.

Labidi, M, Baillot, R, Dionne, B, LaCasse, Y, Maltais, F and Boulet, L (2009), Pleural Effusions following Cardiac Surgery, Chest 2009: I 36 : 1604-1611

Li, H, Das, K, Fu, G, Li, R, Wu, R (2011), The Bayesian lasso for genome-wide association studies, Bioinformatics (2011) 27 (4), 516-523

Lockhart, R, Taylor, J, Tibshirani, R. J, Tibshirani, R, A significance test for the lasso (2012), Department of Statistics, paper 131, http://repository.cmu.edu/statistics/131

Maloney, K., Schmid, M., Weller, D., Applying Additive Modeling and Gradient Boosting to Assess the Effects of Watershed and Reach Characteristics on Riverine Assemblages, Methods in Ecology and Evolution, 2012, 3, 116-128.

Meier, L, Van der Geer, S, Buhlmann, P (2008), The group lasso for logistic regression J.R. Statist, Soc B, 70, part 1, 53-71

Neter J, Wasserman, W and Kutner, M (1983) Applied Linear Regression Models, Homewood: Richard D. Irwin
Pregibon, D (1981), Logistic Regression Diagnostics, Annals of Statistics 9: 705-721
Qiu, l, Cheng, X, Wu, J, Liu, J, Xu, T, Ding, H, Liu, Y, Ge, Z, Wang, Y, Han, H, Liu, J, Zhu, G, 2013, Prevalence of hyperuricemia and its related risk factors in healthy adults from Northern and Northeastern Chinese Provinces, BMC Public Health, 2013, 13:664
Queiroz, N, Sampaio, D, Santos, E, Bezerra, A. 2012, Logistic model for determining factors, associated with HIV infection among blood donor candidates at the Fundacao HEMOPE
Rev Bras Hematologia Hemoterapia, 2012; 34(3): 217-21
Ridgeway, G. (2015), Package 'gbm', http://cran.r-project.org 9/17/2016.
Ryan, T (2009), Modern Regression Methods 2nd Ed, Hoboken, NJ: Wiley
Steyerberg, E, Eijkemans, M, Harrell, Jr, F, Habbema, J. (2000), Prognostic Modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets, Statist. Med. 2000: 19: 1059-1079
Tibshirani, R (1996), Regression Shrinkage and Selection via the lasso, Journal of the Royal Statistical Society: series B58 (1), 267-288
Wang, H and Leng, C. (2008), A note on adaptive group lasso, Computational Statistics and Data Analysis 52 (2008), 5277-5286
Wiegand, R (2009), Performance of Using Multiple Stepwise algorithms for variable selection Statist. Med. 2010, 29, 1647-1659
Wu, T, Chen, Y F, Hastie T, Sobel, E, Lange, K, 2009 Genome wide association analysis by lasso penalized logistic regression, Bioinformatics 25: 714-721
Yuan, M. and Lin, Y (2006), Model Selection and Estimation in Regression with Grouped Variables, J. Royal Statistical Society: Series B 68(1), 49-67
Zou, H. (2006) The Adaptive lasso and its Oracle Properties, Journal of the American Statistical Association 101:476, 1418-1429