Evaluating Somatic Mutation–Based Scores for Predicting Prostate Cancer Aggressiveness

Juhnar C Esmeralda and Hui-Yi Lin, PhD
Biostatistics & Data Science Program, School of Public Health, LSUHSC

LSU Health NEW ORLEANS School of Public Health

Louisiana State University Health Sciences Center School of Public Health

publichealth.lsuhsc.edu

Background

Prostate cancer (PCa) is the most common cancer and the 2nd leading cause of cancer death among American men. The existing PCa risk stratification of PCa aggressiveness is not sufficient, so there is a need to develop tools to increase prediction accuracy. Somatic mutations, particularly in genes like *TP53*, are powerful predictors of aggressive PCa. To improve risk prediction accuracy and stability, developing risk scores by integrating mutations in multiple genes is needed. Somatic data have unique features (small sample size and data sparsity), so variable selection using a logistic model is not suitable. To overcome these limitations, this study intends to develop somatic mutation scores of PCa aggressiveness by using advanced machine learning approaches.

Methods

- Our study focused on 413 Caucasian PCa patients in The Cancer Genome Atlas – Prostate Adenocarcinoma data (TCGA-PRAD).
- We compared the performance of regularization regressions (LASSO and elastic net) with the traditional approach using the univariate logistic model. In addition, we are also interested in evaluating modified regularization regressions, which fine-tune lambda based on p-values, and the two-stage approach using the Support Vector Machine (SVM) as the 2nd stage.
- The performance was evaluated using AUC, accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Youden's statistic was used as the Receiver Operating Curve cutoff to obtain optimal performance measure.

Table 1. Summary of 9 somatic mutation scoring methods used in this Study

Approach	Details of gene Selection & Tuning	Scoring or Classification[1]
Univar. logistic model	Univariate logistic model	Direct
LASSO	LASSO with deviance-based tuning	Direct
LASSO+ SVM	LASSO with deviance-based tuning	SVM
LASSO-p	LASSO with p-value-based tuning	Direct
LASSO-p + SVM	LASSO with p-value-based tuning	SVM
EN	Elastic net with deviance-based tuning	Direct
EN + SVM	Elastic net with deviance-based tuning	SVM
EN-p	Elastic net with p-value-based tuning	Direct
EN-p+ SVM	Elastic net with p-value-based tuning	SVM

[1]Direct method uses weights based on model coefficients

Results

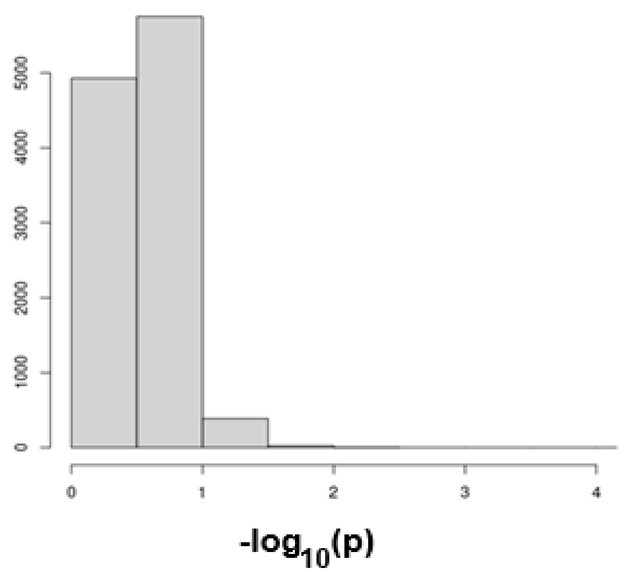

Fig 1. Plot of univariate p-values of genes from TCGA-PRAD

We used a Bonferroni cutoff to determine the genes to be retained for the univariate weighting approach. We have examined the distribution of $-\log_{10}(p)$ and determined that a cutoff of 10^{-3} was appropriate.

Table 2. Results of the Classification Evaluation Measures on 6 Combinations of Gene Selection Algorithms and Classification Procedures

Score	Score p-val.	No. of Genes	AUC	Sens.*	Spec.*	PPV*	NPV*	Accuracy*
Univariate logistic model	2.44 x 10^{-5}	1	0.57	0.18	0.97	0.83	0.57	0.57
LASSO	0.183	216	0.54	0.52	0.58	0.52	0.57	0.56
LASSO+SVM	NA	216	1	1	1	1	1	1
LASSO-p	4.89 x 10^{-15}	84	0.93	0.73	0.91	0.95	0.8	0.85
LASSO-p + SVM	NA	84	0.95	0.73	1	1	0.81	0.87
EN	0.129	304	0.54	0.52	0.58	0.52	0.58	0.55
EN + SVM	NA	304	1	1	1	1	1	1
EN-p	NA	339	1	1	1	1	1	1
EN-p + SVM	NA	339	1	1	1	1	1	1

*The metrics are calibrated from the training data, TCGA-PRAD.

The traditional univariate weighting approach yielded the lowest sensitivity among the 9 methods. Using the univariate weighting approach as our risk classification tool for PCa aggressiveness generates the lowest detection rate among patients who had aggressive prostate cancer. Although specificity and PPV were at least 0.83, AUC, NPV and accuracy were at 0.57.

Results (cont.)

- Under the direct method, LASSO-p retained fewer genes (84 genes) and higher AUC of 0.93 when compared against LASSO. LASSO-p generated a significant score ($p=4.89 \times 10^{-15}$) while LASSO did not ($p=0.18$). The former's score p-value was also smaller than the univariate weighting approach (2.44×10^{-5}).
- LASSO-p's two-stage counterpart showed a relatively better specificity than its direct version (from 0.91 to 1.00). LASSO, on the other hand, selected more genes than LASSO-p (216 vs. 84 genes).
- EN-p under the direct method selects more genes than the EN method (339 vs. 304 genes).
- Three (LASSO + SVM, EN + SVM, and EN-p + SVM) of the 9 algorithms yielded a value of 1.0 for AUC, sensitivity, specificity, PPV, NPV and accuracy while selecting 216 and 339.

Conclusion

- Among the 9 methods considered in this study, LASSO-p (direct method) outperformed the other 8. LASSO-p obtained the smallest p-value among direct methods. Although coming in 2nd to LASSO-p + SVM in terms of AUC, specificity, NPV and accuracy, it produced the smallest p-value.
- LASSO-p under the two-stage method (LASSO-p + SVM) obtained the most optimal set of metrics.
- SVM with linear kernel have only boosted LASSO-p when examining across the 5 gene selection procedures.
- *TP53* and *MAP1A*, genes that are commonly associated to cancer occurrences, were among the genes that had the highest effect sizes in the LASSO-p model. LASSO-p extracted *MAP1A* despite the univariate logistic model failing to select the gene on account of quasi-complete separation.