Evaluating Somatic Mutation–Based Scores for Predicting Prostate Cancer Aggressiveness

Authors and Affiliation

Juhnar Esmeralda; Hui-Yi Lin, PhD

Biostatistics and Data Science Program, School of Public Health, LSU Health Sciences Center

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.

Results

The traditional univariate weighting approach yielded the lowest sensitivity among the 9 methods. Under the direct method, LASSO-p retained fewer genes (84 genes) and higher AUC of 0.93 when compared against LASSO. LASSO-p's two-stage counterpart showed a relatively better specificity than its direct version.

Conclusions

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. *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.