Abstract

Background: The highly variable clinical behavior of localized prostate cancer poses challenges to individualized risk assessment of patients.
Objective: The objective is to provide methodological groundwork for gene-based survival analysis in prostate cancer patients in a small n large p data set. Based on these methods, I develop an RNA expression-based score for the prediction of prostate cancer prognosis.
Data and Methods: I used the RNA sequencing data of 332 patients as training data and microarray gene expression data of 164 patients as a validation set. I included ~20,000 genes in my analyses. Death of disease and biochemical recurrence were the main outcome measures.
I used Cox proportional hazard regression, random forest permutation importance, elastic net regularized regression and Weibull regression for feature selection. The robustness of these methods was assessed by adding permuted genes, without association with prognosis. I implemented a method for data augmentation to account for technical variances in the sequencing process. I applied a variety of machine learning methods inter alia Cox proportional hazard models and support vector machines to the real and augmented cohorts. Additionally, I incorporated biological information based on the hallmarks of cancer in a multivariate Cox proportional hazard model.
Results and Limitations: The robustness of the feature selection methods against genes without prognostic information was highly dependent on the cohort. My analyses resulted in a risk score that consists of 229 genes and predicts the risk highly accurately with a survival support vector machine [Concordance-index: 0.81; 95% CI: 0.60-0.93]. Data augmentation could improve prediction accuracy. The incorporation of biological information did not lead to meaningful improvements. Overall, a version of the support vector machine [Range concordance index: 0.75-0.86; n = 19] and a score based on Cox proportional hazard models [Range concordance index: 0.72-0.81; n = 19] yielded high prediction accuracy independent of the feature selection method and cohort. The low number of events in the validation data set resulted in large confidence intervals.
Conclusion: I developed methods to accurately assess the risk of prostate cancer prognosis, provided a methodological framework for survival analysis with a small n large p data set and showed the benefits and limitations of my newly developed methods.