Cancer Detection Using Biomarkers: SAP

Biomarkers for Predicting Prostate Cancer Recurrence

Denis Ostroushko

Introduction

Prostate cancer (PCa) stands as the most prevalent cancer among the male population in the United States. Typically, PCa tumors are identified through prostate-specific antigen (PSA) blood test screenings. However, an elevated PSA level can stem from various factors, necessitating a biopsy for confirmation. Alarmingly, research indicates that a significant portion of tumors, ranging from 50% to 62%, would remain unnoticed without screening, categorized as 'indolent.' These latent tumors might only manifest symptoms 7 to 14 years post-detection. The majority of cancers detected through PSA screening are localized and low-risk, characterized by a Gleason score of 6 or lower.

Upon detecting high PSA levels, a biopsy is often recommended. For localized, low-risk cancers, active surveillance, or "watchful waiting," is advocated, where treatment is administered only upon disease progression. However, for non-localized or high-risk cancers, prompt treatment via surgery or radiation is typically advised.

Despite recommendations for active surveillance in low-risk cases, many opt for definitive therapy, such as surgery, often leading to adverse side effects. This inclination is fueled by discomfort among both patients and physicians in delaying treatment, despite the fact that some individuals with low-grade prostate cancer succumb to the disease.

Furthermore, ambiguity surrounds the optimal course of action for men with moderate-grade disease (Gleason score = 7). Consequently, there is a pressing need to identify biomarkers that can predict PCa mortality and recurrence, particularly from initial biopsy results. With over 40 candidate biomarkers identified by various authors, the question arises: can these biomarkers, combined with clinical covariates, culminate in a predictive model for PCa recurrence within 5 years of prostatectomy? Moreover, does integrating biomarkers with clinical covariates enhance predictive accuracy compared to relying solely on clinical parameters? These questions underscore the significance of exploring novel approaches to improve PCa prognostication and treatment decision-making.

Data Set

The dataset comprises tumor samples obtained from 400 men who underwent radical prostatectomy for prostate cancer (PCa) at the University of Minnesota Medical Center between 1999 and 2008. Demographic and clinical data were extracted from medical records, encompassing variables such as age, preoperative PSA levels, Gleason score, and an indicator for non-localized tumors.

Tumor samples were subjected to immunohistochemical (IHC) staining for various biomarkers. Our primary outcome of interest is PCa recurrence, defined as the time from prostatectomy to biochemical recurrence, with biochemical recurrence marked by a prostate-specific antigen (PSA) value of 0.2 ng/mL or higher. Time-to-recurrence was censored at the last contact date for participants who did not experience recurrence during the follow-up period of at least 5 years. This allows for the creation of a binary outcome variable, eliminating the need to address censoring.

A set of 40 candidate biomarkers, standardized to have a standard deviation of 1, will be considered in our analysis. Additionally, clinical covariates such as age, preoperative PSA levels, Gleason score, and tumor localization status were extracted, all of which are known to be associated with PCa recurrence. This

Table 1: Baseline demograohic characteristics of the study sample

	No Progression	Progressed	SMD
N	243	157	
Mean PreOp. PSA (SD)	7.66 (5.77)	6.69 (4.95)	0.181
Mean Age (SD)	60.67 (6.70)	62.78 (6.52)	0.320
N localized tumor (%)	127 (52.3)	106 (67.5)	0.315
Mean Gleason Score (SD)	6.78 (0.67)	6.83 (0.74)	0.065

comprehensive dataset provides a robust foundation for investigating the predictive utility of biomarkers and clinical covariates in forecasting PCa recurrence following prostatectomy.

Methods

The primary aim of this analysis is to assess the predictive capability of statistical models containing solely clinical covariates against models integrating biomarker information. Our key focus lies in discerning which biomarkers significantly contribute to predictive power and distinguishing them from those that do not. Given the complexity of our dataset, we will explore both traditional regression techniques for binary outcomes and more adaptable tree-based methodologies such as Random Forest.

To facilitate model development, we will construct and compare models encompassing varying numbers of biomarkers. Initially, baseline models will solely incorporate clinical predictors. Subsequently, we will develop and contrast these with Random Forest and logistic regression models. We aim to construct a parsimonious model with biomarkers, employing a LASSO logistic regression approach with shrinkage penalties solely applied to biomarkers. As an alternative, for the random forest we will conduct variable selection utilizing the mean decrease in Gini index for biomarkers.

Our final suite of models will include logistic regression models incorporating clinical covariates alongside a ridge penalty applied exclusively to biomarkers. Additionally, a Random Forest model incorporating all variables will be included for comparison.

Hyperparameters for each model type will be meticulously selected via cross-validation. For Random Forest, hyperparameter tuning will solely consider the number of trees and the number of variables utilized at each split.

The outcomes of cross-validation on out-of-sample data will be detailed in the appendix and presented as supplementary material, providing rationale for our model selection procedures.

Evaluation of all models will be based on two primary metrics: AUC (Area Under the Curve) and PPV (Positive Predictive Value). Given the balanced class distribution in our dataset (60% did not experience recurrence within 5 years), we anticipate specificity and sensitivity to demonstrate symmetrical behavior, with minimal impact on AUC of the ROC curve. PPV will be scrutinized to ensure predictive models allocate the smallest number of patients into the 'at-risk' category, thus minimizing unnecessary intervention. The cutoff for probability dichotomization will be determined based on Youden's index.

AUC and PPV collected based on predictions on the testing data for all six models set will be compared using pairwise differences. To account for multiple comparison we will use Benjamini Hochberg correction controlling false discovery rates. Standard errors for point estimates will be obtained using bootstrap resampling of predicted values on the training data set.

For training purposes, 75% of the available data, equivalent to 300 observations, will be utilized, with the remaining 100 observations reserved for testing. Data partitioning will be conducted using class stratification to preserve class balance. Results for cross validation, in sample, and out of sample predictions will be delivered in tables like Table 2.

All analyses will be conducted using R version 4.3.1.

Table 2: Resust Table

	Regression Models		Random Forest	
	AUC	PPV	AUC	PPV
Baseline covariates				
Baseline covariates + some biomarkers				
Baseline covatiates + all biomarkers				

[a] Regression Model with some Biomarkers uses LASSO regularization to select biomarkers

[a] Final staisitics will have standard errors for estimates in parentheses