Predictive model for prostate cancer recurrence within five years of prostatectomy

Introduction

Prostate cancer is the most common cancer among men in the United States, with a significant number of new cases and deaths reported annually. Although the overall survival rate is high, the variability in the disease's progression necessitates improved predictive models for recurrence post-prostatectomy. This study aims to develop and validate a predictive model for prostate cancer recurrence within five years of prostatectomy using a combination of clinical covariates and biomarkers. The research question is whether these biomarkers can be combined with clinical covariates to develop a predictive model for prostate cancer recurrence within five years of prostatectomy and whether the model, including biomarkers, improves prediction over clinical covariates alone.

Dataset

The data was collected from a group of 400 men who underwent radical prostatectomy at the University of Minnesota Medical Center between 1999 and 2008. The data contains demographic information, clinical variables, and measurements on 40 biomarkers. Age, Preoperative PSA, Gleason score, and Non-localized tumor indicators were extracted from the patients' medical files, which are known to be associated with PCa recurrence. Variables were standardized, and a binary outcome for prostate cancer recurrence was created based on biochemical failure. The Gleason score was given as character. It was split and converted to a numeric format, and the sum of the numbers was done to get the gleason score. The follow-up time was converted into years, and to create an outcome variable with 5-year follow-up recording will be done. For all those with a follow-up time under 5 years and those who had recurrence will be coded as 1, and the rest with recurrence over 5 years will be coded as 0.

Exploratory Data Analysis

Exploratory Data Analysis will involve summary tables and graphical summaries. Table. 1 shows the summary statistics of the continuous variables: age, pre-operative PSA levels, and the top 5 important biomarkers, namely HMMR, SIAH2_nuc, HAS2, adfp, and IGF1. The average age of participants is 61.5 years, with a standard deviation of 6.7, suggesting a moderate spread around the mean. The median age is 61, with half of the participants aged between 56.5 and 66.5 years, as indicated by the interquartile range 10. Pre-operative PSA levels, crucial in prostate cancer detection and monitoring, average 7.28 ng/mL, with a standard deviation of 5.48, indicating significant patient variability. The median PSA level is 5.55 ng/mL, with an interquartile range of 4.83, further highlighting the diverse disease presentations in the study cohort.

Regarding biomarkers, HMMR levels have a mean and median that are very close to each other (7.65 and 7.64, respectively). SIAH2_nuc has a mean level of 1.44 and a median of 1.12, with its values dispersed as indicated by a standard deviation of 0.97 and an interquartile range of 1.15. HAS2, adfp, and IGF1 biomarkers also exhibit varied levels across the cohort, reflecting the molecular heterogeneity typical of prostate cancer.

The categorical data in the table, including Gleason score, tumor localization, biochemical recurrence, and overall outcome, underscore the clinical facets of the study population. The Gleason score, an important prognostic indicator in prostate cancer, shows that the majority of patients (59%) have a score of 7, followed by 33% with a score of 6, while scores 8 and 9 are less common, representing 5% and 4% of the patients respectively. Tumor localization data reveal that 58% of patients have non-localized tumors, indicative of more advanced disease. Biochemical recurrence, a marker of disease recurrence or progression, is observed in 48% of the patients, illustrating the substantial burden of aggressive disease in this group. Finally, the outcome variable indicates that 39.25% of patients experienced recurrence in 5-year follow-up. Figure 1 shows the bar plot depicting the feature importance derived from a Random Forest model. It is arranged in descending order of importance, and the top 5 biomarkers are used for the full model.

Methods

We will employ a logistic regression model to estimate the risk of PCa recurrence. The modeling will proceed in two steps. The first is the clinical model, which will include only clinical covariates like age, preoperative PSA, Gleason score, and tumor stage to establish a baseline predictive performance. The full model expanded upon the base model by incorporating significant biomarkers (HMMR, SIAH2_nuc, HAS2, adfp, IGF1) identified through a Random Forest analysis. These clinical biomarkers are the top 5 biomarkers among the 40 given in the data. A Random Forest model was specifically employed to rank the biomarkers based on their importance, utilizing the Mean Decrease Gini criterion and the top 5 for the full model. The outcome variable is a binary indicator of prostate cancer recurrence within five years post-prostatectomy. The predictor variables are the clinical covariates (age, preoperative PSA, Gleason score, tumor stage) for the baseline and full models and the top 5 standardized biomarkers for the full model are included.

Inclusion of variables based on clinical relevance and statistical significance, with consideration for multicollinearity. Bootstrapping for internal validation to assess model stability and reliability. A 5-fold cross-validation for both logistic regression and random forest models. This method splits the data into five subsets, using four for training and one for validation in each iteration, providing a robust estimate of the model's performance. The models are also evaluated using bootstrapping (with 1000 replications), which measures accuracy and stability by resampling the data with replacement and fitting the model multiple times. Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics are calculated for each model to clearly compare model performance in terms of distinguishing between the two outcome classes. Performance metrics such as accuracy, sensitivity, specificity, positive predictive value (precision), negative predictive value, and F1 score are calculated. Feature importance is visualized using a bar plot, clearly representing which features are most influential in the random forest model. ROC curves are plotted for the logistic regression and random forest models to compare their performance visually.

Results

Table 1. shows the values for the models. The cross-validated logistic regression model shows moderate discrimination capability (AUC: 0.58) with high sensitivity (0.86) but poor specificity (0.29), indicating it's better at identifying true positives than true negatives. However, its precision (0.65) and F1 score (0.74) suggest decent performance in correctly identifying positive cases. The bootstrapped logistic regression model has low sensitivity (0.31) but high specificity (0.84), implying it's better at identifying true negatives but struggles with true positives. The cross-validated Random Forest outperforms logistic regression significantly with a higher AUC (0.88), indicating better overall performance. It demonstrates excellent sensitivity (0.98) and specificity (0.78), resulting in high accuracy (0.90) and a strong F1 score (0.92), indicating robust performance across both positive and negative classes. Similar to the cross-validated random forest, the Bootstrapped Random Forest model performs well across all metrics, demonstrating high sensitivity (0.81), specificity (0.92), accuracy (0.87), and a strong F1 score (0.83), indicating its effectiveness in both sensitivity and specificity.

Figure 2. shows the ROC curve for the best and base models. The logistic regression model's performance is less than that of the Random Forest model, as indicated by its curve being closer to the diagonal line of no discrimination. The graph for the RF model is above the blue curve, indicating that it can better distinguish between the classes (patients with and without recurrence of prostate cancer). The AUC for the Random Forest is 0.86, which is generally considered good. The points marked on the blue line represent cut-off values used to classify the results as positive or negative. For example, one of the points has a sensitivity of approximately 0.88 and a 1-specificity of about 0.30. This point represents a threshold that catches a high percentage of actual positive cases (high sensitivity), with a moderate rate of false positives. An AUC close to 1 indicates a great model, while an AUC close to 0.5 indicates a model with no discriminative ability. The closer the ROC curve is to the top left corner, the higher the test's overall accuracy. In your case, the Random Forest model's AUC of 0.86 suggests that it is a strong model.

Conclusion and Discussion

From the performance of the random forest model, integrating the top 5 candidate biomarkers with clinical covariates for predicting prostate cancer recurrence could enhance the model's predictive power. Random forests are known for their ability to handle high-dimensional datasets effectively, making them suitable for integrating numerous biomarkers into predictive models. Due to the relatively low discriminative capability of the logistic regression model and the superior performance of the random forest model, it is reasonable to hypothesize that incorporating the relevant biomarkers from the 40 candidate biomarkers with clinical covariates could improve predictive accuracy compared to using clinical covariates alone. The study's limitations include potential overfitting, as the models were tested on the same dataset used for training. Another issue is that Random Forests can be less interpretable, which may limit the clinical application to understand the decision-making process.

References:

1. Rizzardi, A.E., Rosener, N.K., **Koopmeiners, J.S.**, Vogel, R.I., Metzger, G.J., Forster, C.L., Marston, L.O., Tiffany, J.R., McCarthy, J.B., Turley, E.A., Warlick, C.A., Henriksen, J.C., Schmechel, S.C. "Evaluation of Protein Biomarkers of Prostate Cancer Aggressiveness." BMC Cancer, **14**, Article 244, 2014.

Exploratory data Analysis

Table 1: Table 1. Model Performance Comparison

Model	AUC	Sensitivity	Specificity	PPV	NPV	Accuracy	Kappa	F1_Score
Cross-Validated Logistic Regression	0.66	0.86	0.29	0.65	0.58	0.64	0.17	0.74
Cross-Validated Random Forest	0.97	0.98	0.78	0.87	0.95	0.90	0.78	0.92
Bootstrapped Logistic Regression	0.66	0.86	0.29	0.65	0.58	0.64	0.17	0.74
Bootstrapped Random Forest	0.97	0.98	0.78	0.87	0.95	0.90	0.78	0.92

Figure 1. Feature Importances from Random Fore

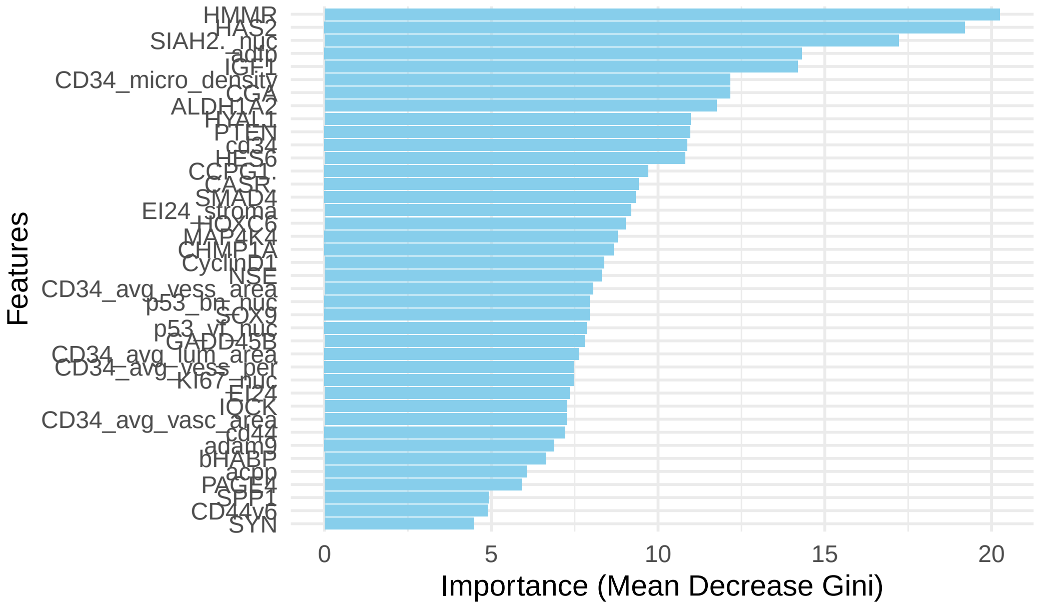


Figure 2. ROC Curves

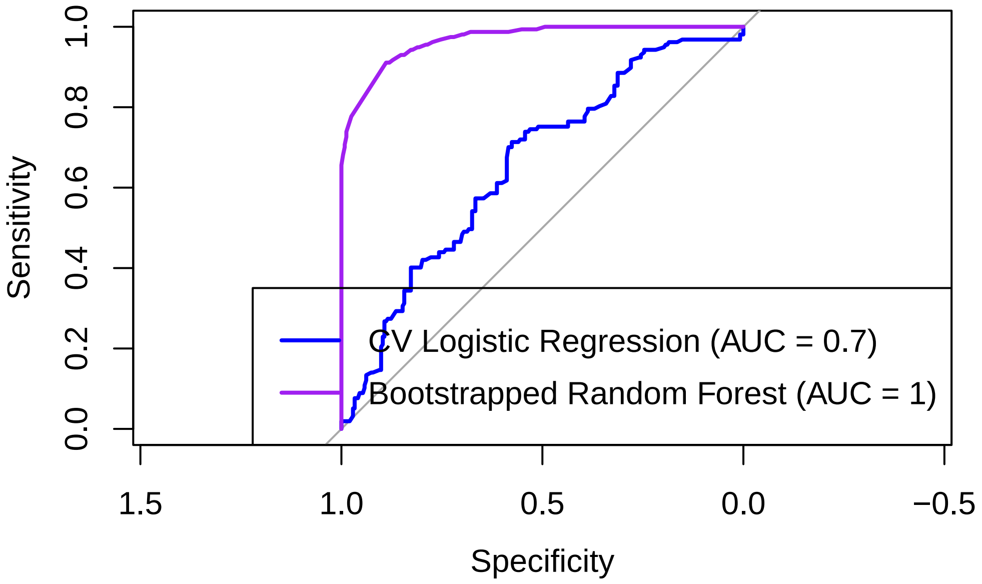
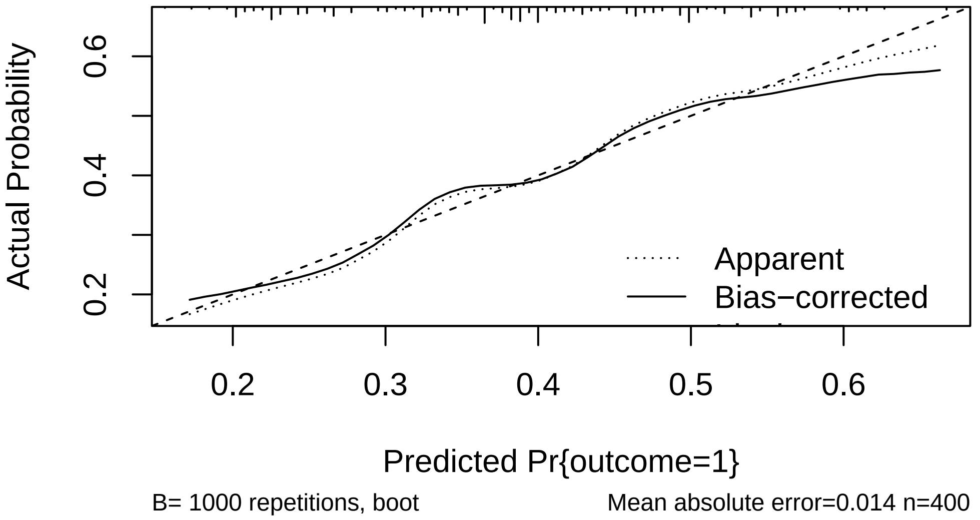
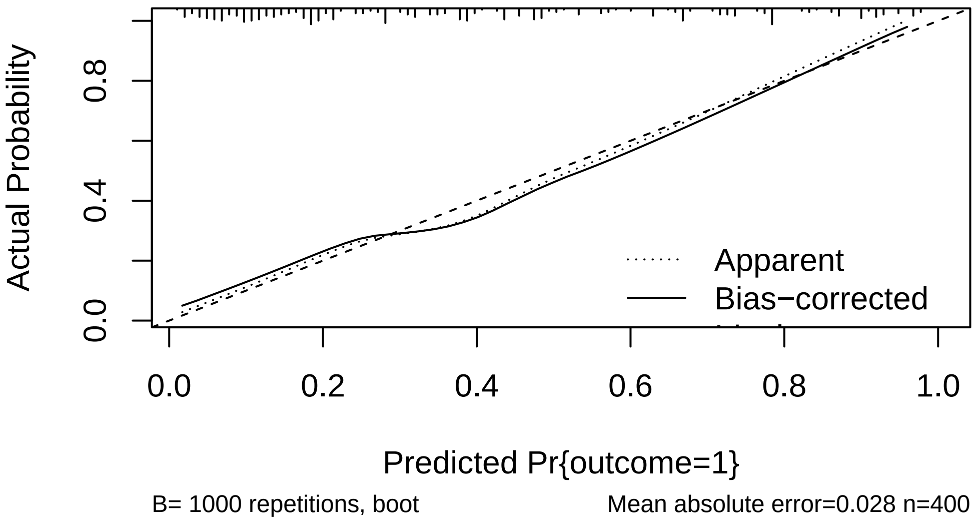


Figure 3. Calibration Plot for Base Model



```
n=400   Mean absolute error=0.014   Mean squared error=0.00034
0.9 Quantile of absolute error=0.027
```

Figure 4. Calibration Plot for Full Model



```
n=400   Mean absolute error=0.028   Mean squared error=0.00096
0.9 Quantile of absolute error=0.047
```