Prostate Cancer Prognosis: Predicting Capsular Penetration Using Baseline Clinical Variables

Group 9

Miles Keitz, Edy Reynolds, Samantha Ward, and Raymond Yang

1. **Introduction/Background**

Prostate cancer is the most common cancer in men, excluding skin cancer, and early detection is crucial for effective treatment. This study aims to determine whether certain baseline clinical variables can predict whether a prostate tumor has penetrated the prostatic capsule. This means that the cancerous cells have grown beyond the outer boundary of the prostate gland, indicating a more advanced stage of prostate cancer and a worse prognosis.

This dataset consists of 376 patients, of which 151 (40.16%) were diagnosed with tumor penetration. The primary objective is to use logistic regression to analyze the relationship between various predictors and the likelihood of capsular penetration.

The response variable in this study is CAPSULE, which is a dichotomous variable indicating whether the tumor has penetrated the prostate capsule (1 = penetration, 0 = no penetration). In the context of logistic regression, “success” is defined as a tumor penetrating the prostatic capsule (CAPSULE = 1). Identifying significant predictors of capsular penetration can help improve early detection and inform clinical decision-making.

Several explanatory variables are included in the dataset:

**AGE** (in years) – Patient’s age at the time of examination.

**RACE** (categorical: 1 = White, 2 = Black) – Patient’s racial background.

**DPROS** (categorical) – Digital rectal exam results, classified as:

* 1 = No nodule detected
* 2 = Unilobar nodule (left)
* 3 = Unilobar nodule (right)
* 4 = Bilobar nodule

**DCAPS** (categorical: 1 = No, 2 = Yes) – Indicated whether capsular involvement was detected during a rectal exam.

**PSA** (mg/ml) – Prostate-specific antigen level, a key biomarker used for prostate cancer screening.

**VOL** () – Tumor volume obtained from ultrasound measurements.

**GLEASON** (score 1-10) – Total Gleason Score, which measures tumor aggressiveness based on biopsy samples.

This dataset provides an opportunity to evaluate the relationship between these clinical predictors and tumor penetration using statistical methods. By applying logistic regression, we aim to determine which variables are significantly associated with capsular penetration and assess their predictive power.

1. A graph of a number of years

   AI-generated content may be incorrect.A graph of a number of psa levels

   AI-generated content may be incorrect.**Graphs and Summary Statistics**

A graph of a graph

AI-generated content may be incorrect.A graph of a number of tumor cells

AI-generated content may be incorrect.

**Table 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Mean** | **Median** | **Standard Deviation** |
| AGE | 66.00 | 67.00 | 6.43 |
| PSA | 15.28 | 8.75 | 19.89 |
| VOL | 15.88 | 14.25 | 18.41 |
| GLEASON SCORE | 6.383 | 6.000 | 1.092 |

The patients' ages ranged from 47 to 79 years, with a mean of 66 years. The distribution of prostatic-specific antigen (PSA) is right-skewed (mean > median), indicating the presence of extreme PSA values in some patients. Similarly, the Tumor Volume distribution shows that many patients had no tumor volume detected, resulting in a large concentration of zeros in the corresponding graph. The observed Gleason Scores ranged from zero to nine, with the majority falling between 6 and 7 (mean = 6.383). Clinically, a Gleason score of 6 is considered low-grade, indicating slow-growing cancer, while a score of 7 is considered intermediate-grade with a moderate growth rate.

A graph of a purple rectangular object

AI-generated content may be incorrect.A graph showing a rectangle and a rectangle

AI-generated content may be incorrect.A graph of a rectangle

AI-generated content may be incorrect.A graph with a bar and a bar

AI-generated content may be incorrect.

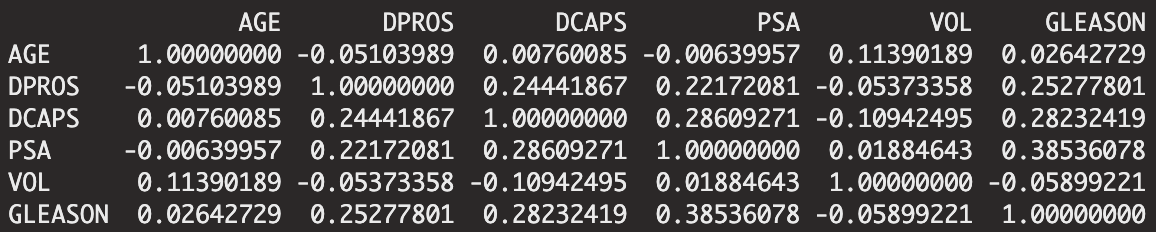
**Table 2**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Category** | **Percentage (%)** |
| RACE | White (1) | 90.43% |
|  | Black (2) | 9.57% |
| DPROS | No Nodule (1) | 26.06% |
|  | Unilobar Left (2) | 34.84% |
|  | Unilobar Right (3) | 25.27% |
|  | Bilobar (4) | 13.83% |
| DCAPS | No (1) | 89.36% |
|  | Yes (2) | 10.64% |
| CAPSULE | No Penetration (0) | 59.84% |
| (response) | Penetration (1) | 40.16% |

*\*Values found using pivot tables in Excel*

The dataset is predominantly composed of White (90.43%) patients, with Black patients making up only 9.57% of the study population. Among the digital rectal exam (DPROS) results, the most common finding was a left-side unilobar nodule, observed in approximately 35% of patients. Capsular involvement (DCAPS) was detected in only 10.64% of patients during the rectal exam. However, about 40% of patients exhibited tumor penetration (CAPSULE = 1), suggesting that DCAPS alone may not be a strong predictor of capsular penetration.

1. **Logistic Regression Model** 
   1. **Correlated Quantitative Variables (Multicollinearity)**



Based on the correlation matrix between the quantitative variables from the dataset, no coefficients seem to be strongly correlated with one another. The strongest correlation we retrieved was between the GLEASON and PSA variables with a coefficient of 0.3854. This suggests a somewhat moderate positive correlation, indicating that elevated PSA levels might be associated with more severe cases of prostate cancer. However, the association between these two variables is not strong enough to indicate multicollinearity, as the general threshold is a correlation greater than 0.7, so they were retained in the model. Aside from that, the other coefficients are generally small, suggesting no sign of multicollinearity among the other variables.

* 1. **Interaction Terms**

For this prostate cancer study, an interaction term should involve the combination of two variables (since we are not including three-way interactions) that might have a meaningful effect on the likelihood of the tumor penetration (CAPSULE response variable). In this case, it is important to account for a few interactions before conducting the analysis.

PSA x VOL (Prostatic Specific Antigen Value x Tumor Volume)

PSA is a key indicator for prostate cancer, and tumor volume represents the physical growth of the cancer. This can mean that a high PSA value may have a greater effect on tumor penetration when the volume of the tumor is larger.

DPROS x VOL (Digital Rectal Exam x Tumor Volume)

The Results of the Digital Rectal Exam can help indicate the presence and extent of nodules (none, left unilobar nodule, right unilobar nodule, bilobar nodule), while volume quantifies the size of the cancer tumor. This interaction can be relevant because if a tumor is smaller and lower in volume, it might not be detectable via a digital rectal exam, making DPROS less informative and vice versa. By including this interaction term, the model can account for cases where tumor volume amplifies the impact of detecting nodules on the risk of tumor penetration.

PSA x GLEASON (Prostatic Specific Antigen Value x Gleason Score)

PSA measures the activity of the cancer while the Gleason Score assesses tumor aggressiveness. The effect of PSA on the penetration of tumors may depend on the aggressiveness of the tumor. This can suggest that a high PSA value combined with a high Gleason Score might indicate a stronger, more aggressive cancer, potentially causing a higher likelihood of tumor penetration.

These interactions highlight the importance of considering variables together, as they may provide a more comprehensive understanding of the tumor’s potential for prostatic capsule penetration. However, it is important to note that these are just potential interaction ideas prior to conducting any analysis, and the significance of each variable will be evaluated to determine if any should be excluded from the model.

* 1. **Final Model**

To decide if there were any predictor variables that we could remove from the model, we conducted a Likelihood-Ratio Test using the Anova() function in R with the following null and alternative hypotheses: H0: βi = 0, Ha: βi ≠ 0 with β representing the variable coefficients and i = 1, 2, ..., 7. Against a significance level of 0.05, AGE (p-value: 0.32), RACE (p-value: 0.06), DCAPS (p-value: 0.16), and VOL (p-value: 0.14) all failed to reject the null hypothesis, meaning that each variable is an insignificant predictor of tumor penetration of the prostatic capsule. To confirm these results, we conducted an analysis of deviance between our full model with all 7 variables versus a reduced model that excludes AGE, RACE, DCAPS, and VOL with the following null and alternative hypotheses: H0: The reduced model is sufficient relative to the full model, Ha: The reduced model is insufficient relative to the full model. The full model has a residual deviance of 374.54, and our reduced model has a residual deviance of 381.12. By taking the difference of the residual deviances between the reduced model and the full model, we get an LRT statistic of 6.58. With 4 degrees of freedom, we get a p-value of 0.16, failing to reject the null hypothesis at a significance level of 0.05. Now, with the remaining variables, we want to see if there are any two-way interaction terms to make a more accurate model. Our saturated model to make comparisons includes all possible two-way interactions with the four remaining variables. The step() function reveals that a two-way interaction is not necessary for the model. However, we decided to include an interaction term between PSA and GLEASON. The AIC with this interaction term does not differ heavily from the model without the interaction term. To confirm, we conducted an analysis of deviance with our reduced and saturated models. The test has the following hypotheses: H0: The reduced model is sufficient relative to the saturated model; Ha: The reduced model is insufficient relative to the saturated model. The saturated model has a residual deviance of 381.04, and our reduced model has a residual deviance of 377.61. By taking the difference of the residual deviances between the reduced model and the full model, we get an LRT statistic of 3.43. With 2 degrees of freedom, we get a p-value of 0.18, rejecting the null hypothesis at a significance level of 0.05. With this, we have

Final model: -8.437 + 0.767(DROPS2) + 0.1549(DROPS3) + 1.420(DROPS4) + 0.052(PSA) + 1.040(GLEASON) - 0.004(PSA\*GLEASON)

Reference level: DPROS1 (No nodule)

1. **Interpretation and Prediction**

**A. Interpretation**

One explanatory variable included in our final model is Gleason Score. Gleason Score is a scoring system used to grade the severity of prostate cancer in patients. It is the summation of two numbers, each scoring the degree of observed tumor cells. In our model, the regression coefficient for Gleason Score is 1.040. This represents the estimated multiplicative effect of Gleason Score on whether the tumor has penetrated the prostatic capsule or not. In other words, for every point increase in Gleason Score, the odds ratio of success, or the tumor penetrating the capsule, increases by 4%.

**B. Prediction**

Mean values for numeric variables included in the final model:

Gleason: 6.382979

PSA: 15.27947

Categorical variable, value chosen at random:

Dpros: level 3

-8.437 + 0.767(DROPS2) + 0.1549(DROPS3) + 1.420(DROPS4) + 0.052(PSA) + 1.040(GLEASON) - 0.004(PSA\*GLEASON)

=>

-8.437 + 0.767(0) + 0.1549(1) + 1.420(0) + 0.052(15.280) + 1.040(6.383) - 0.004(15.280\*6.383)

= -1.2393

Probability: Probability:

For a patient with DPROS rectal exam score of 3, a Gleason score of 6.383, and a prostate-specific antigen level of 15.280, the estimated probability of the tumor penetrating the prostatic capsule is 22.45%.

1. **Summary**

This study examined whether baseline clinical variables could predict tumor penetration in prostate cancer patients. Using logistic regression, we analyzed a dataset of 376 patients, with 151 (40.16%) exhibiting capsular penetration. Our goal was to identify significant predictors of penetration to aid in early detection. One of the most surprising findings was that capsular detection in a rectal exam (DCAPS) was not a strong predictor of actual tumor penetration. Although only 10.64% of patients had capsular involvement detected during a rectal exam, 40.16% of patients had conformed penetration, indicating a potential drawback of physical exams in determining tumor presence. Additionally, PSA and Gleason Score showed a moderate correlation (0.385), suggesting that higher PSA levels might be associated with an increase in Gleason Score. A challenge faced during the analysis was determining relevant interaction terms. Initially, variables AGE, RACE, DCAPS, and VOL were included in the model but were removed due to their insignificant contribution to the model based on the likelihood-ratio test. After testing multiple interactions, PSA:GLEASON was the last interaction term retained by the stepwise selection model, suggesting that the impact of PSA on tumor penetration risk may be dependent on Gleason Score.

If this study were to be improved, additional clinical and lifestyle factors could improve the model’s accuracy. One valuable addition would be smoking status (e.g., smoker vs. non-smoker), as smoking has been linked to increased cancer aggressiveness. Another useful variable would be body mass index (BMI), given its association with cancer progression and metabolic health. Incorporating these factors could improve the model and provide a better understanding of prostate cancer prognosis.

Overall, this study showed the importance of considering multiple clinical variables and their interactions when attempting to predict tumor progression. Future research could benefit from adding genetic markers and lifestyle factors to further improve prostate cancer risk assessment.

1. **R Code**

# Load Data

library(readr)

library(dplyr)

library(ggplot2)

pros <- read\_csv("PROS.csv")

# Summary Statistics (Quantitative)

## AGE

pros |>

summarise(mean(AGE), sd(AGE), median(AGE))

ggplot(pros, aes(x = AGE)) +

geom\_histogram(binwidth = 5, fill = "darkblue", color = "black") +

labs(title = "Distribution of Age", x = "Age (years)", y = "Frequency") + theme\_minimal()

## PSA

pros |>

summarise(mean(PSA), sd(PSA), median(PSA))

ggplot(pros, aes(x = PSA)) +

geom\_histogram(binwidth = 2, fill = "darkred", color = "black") +

labs(title = "Distribution of PSA Levels", x = "PSA (mg/ml)", y = "Frequency") +

theme\_minimal()

## Tumor Volume (VOL)

pros |>

summarise(mean(VOL), sd(VOL), median(VOL))

ggplot(pros, aes(x = VOL)) +

geom\_histogram(binwidth = 5, fill = "darkgreen", color = "black") +

labs(title = "Distribution of Tumor Volume", x = "Tumor Volume (cm³)", y = "Frequency") + theme\_minimal()

## Gleason Score

pros |>

summarise(mean(GLEASON), sd(GLEASON), median(GLEASON))

ggplot(pros, aes(x = GLEASON)) +

geom\_histogram(binwidth = 1, fill = "purple4", color = "black") + labs(title = "Distribution of Gleason Scores", x = "Gleason Score", y = "Frequency") + theme\_minimal()

# Summary Statistics for Quantitative Variables

quantitative\_vars <- pros |>

select(AGE, PSA, VOL, GLEASON)

summary(quantitative\_vars)

sd(pros$AGE, na.rm = TRUE)

sd(pros$PSA, na.rm = TRUE)

sd(pros$VOL, na.rm = TRUE)

sd(pros$GLEASON, na.rm = TRUE)

# Summary Statistics (Categorical)

## RACE

pros |>

group\_by(RACE) |>

summarise(n())

ggplot(pros, aes(x = as.factor(RACE))) + geom\_bar(fill = "darkblue", color = "black") + labs(title = "Distribution of Race", x = "Race (1 = White, 2 = Black)", y = "Count") + theme\_minimal()

## DPROS

ggplot(pros, aes(x = as.factor(DPROS))) + geom\_bar(fill = "darkred", color = "black") + labs(title = "Results of Digital Rectal Exam", x = "DPROS (1-4)", y = "Count") + theme\_minimal()

## DCAPS

ggplot(pros, aes(x = as.factor(DCAPS))) + geom\_bar(fill = "darkgreen", color = "black") + labs(title = "Capsular Detection in Rectal Exam", x = "DCAPS (1 = No, 2 = Yes)", y = "Count") + theme\_minimal()

## CAPSULE

ggplot(pros, aes(x = as.factor(CAPSULE))) + geom\_bar(fill = "purple4", color = "black") + labs(title = "Capsular Penetration Status", x = "CAPSULE (0 = No, 1 = Yes)", y = "Count") + theme\_minimal()

# Correlation Matrix

cor(pros[,3:9])

# Full model

full.model <- glm(CAPSULE ~ AGE + factor(RACE) + factor(DPROS) + factor(DCAPS) + PSA + VOL + GLEASON, family = binomial, data = pros)

summary(full.model)

# LRT for Significant Predictors

Anova(full.model)

# Reduced model without Age, Race, DCAPS, and VOL

reduced.model\_3 <- glm(CAPSULE ~ factor(DPROS) + PSA + GLEASON, family = binomial, data = pros)

summary(reduced.model\_3)

## Analysis of Deviance

# 381.12 - 374.54 = 6.58

# Delta = 7 - 3 = 3, df = 4

1 – pchisq(6.58, 4)

# Model Effects with Interactions

model.interact.effects <- glm(CAPSULE ~ factor(DPROS) + PSA + GLEASON +

factor(DPROS):PSA + factor(DPROS):GLEASON +

PSA:GLEASON,

family = binomial, data = pros)

summary(model.interact.effects)

# Significant Two-Way Interactions with Step Function

model.main.effects <- glm(CAPSULE ~ factor(DPROS) + PSA + GLEASON, family = binomial, data = pros)

step(model.main.effects, scope = ~.^2)

## Analysis of Deviance for Model with Interaction

# 381.04 - 377.61 = 3.43

# Delta = 6 - 4 = 2, df = 2

1 – pchisq(3.43, 2)

#Final Model

final.model <- glm(CAPSULE ~ factor(DPROS) + PSA + GLEASON + PSA:GLEASON, family = binomial, data = pros)

summary(final.model)

#Mean Values

mean(pros$GLEASON)

mean(pros$PSA)

1. **Sources**

“Key Statistics for Prostate Cancer: Prostate Cancer Facts.” *Prostate Cancer Facts |*

*American Cancer Society*, 16 Jan. 2025,

[www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html](http://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html).

“Understanding Your Pathology Report: Prostate Cancer.” *Understanding Your Pathology*

*Report: Prostate Cancer | American Cancer Society*, 7 July 2023,

[www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/prostate-pathology/prostate-cancer-pathology.html](http://www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/prostate-pathology/prostate-cancer-pathology.html).

“Genetics of Prostate Cancer (PDQ®)”

[Genetics of Prostate Cancer (PDQ®) - NCINational Cancer Institute (.gov)https://www.cancer.gov › types › prostate-genetics-pdq](https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.cancer.gov/types/prostate/hp/prostate-genetics-pdq%23:~:text%3DProstate%2520cancer%2520is%2520highly%2520heritable,risk%2520and%2520high%252Drisk%2520genes.&ved=2ahUKEwjt2d704YKMAxVNFzQIHQPFLLkQFnoECBcQAw&usg=AOvVaw3qCWPQfF-vO_48i72RMuPC)