Machine Learning Mini-Project

MEIN40330 - AI for Personalised Medicine - 2024/25 Summer

Yu Du

Table of contents

1 Background

According to the report by the **National Cancer Registry Ireland (NCRI)**, the most common incident cancers from 2020 to 2022 were prostate cancer in males and breast cancer in females. These two types of cancer had significantly higher incidence rates than others, each accounting for approximately 32% of cases (*Cancer in Ireland 1994-2022: Annual Statistical Report of the National Cancer Registry (2024)*, 2024). Furthermore, since breast cancer has been studied extensively in previous coursework, I chose to focus on exploring **prostate cancer** in this mini-project.

Prostate cancer refers to the uncontrolled growth of cells in the prostate, a gland in the male reproductive system located below the bladder (Wikipedia Contributors, 2019). From a genetic perspective, structural variants involving the **TMPRSS2** and **ERG** genes are among the most common alterations associated with prostate cancer, with variant frequencies of 23% and 24%, respectively (*CBioPortal for Cancer Genomics*, 2025). These variants are strongly linked to cancer aggressiveness and prognosis. Therefore, this project aims to explore whether the **TMPRSS2-ERG fusion** is associated with worse clinical outcomes, such as reduced survival rate. If not, I will explore what other impacts the TMPRSS2-ERG fusion may have compared to non-fusion cases.

To answer this research question, it is essential to obtain patient sample data, particularly **omics data**[1], for further analysis. Applying statistical and machine learning techniques to omics data enables us to identify differences in specific genes, transcripts, proteins, and epigenetic modifications. At the same time, it can help reveal key gene signaling pathways and cancer-driving factors, which may support the development of targeted drugs for precision medicine.

2 Data Preparation

In this mini-project, I primarily used the following three data files:

- `data_sv.txt`: to extract TMPRSS2-ERG fusion information.
- `data_clinical_patient.txt`: to obtain clinical information, including survival status, survival time, and disease recurrence.
- `data_clinical_sample.txt`: to map sample IDs to patient IDs.

```
# Read structural variation data
sv_data_raw <- read.delim("./prostate_msk_2024/data_sv.txt", stringsAsFactors = FALSE)

# Read patient clinical information
clinical_data_raw <-
  read.delim("./prostate_msk_2024/data_clinical_patient.txt", stringsAsFactors = FALSE)

# Clean the clinical data (remove the first 4 rows of meta information and reset column names)
clinical_data <- clinical_data_raw[-c(1:4), ]
colnames(clinical_data) <- clinical_data_raw[4, ]

# Read sample-patient mapping
sample_data_raw <-
  read.delim("./prostate_msk_2024/data_clinical_sample.txt", stringsAsFactors = FALSE)

# Clean the sample data
sample_data <- sample_data_raw[-c(1:4), ]
colnames(sample_data) <- sample_data_raw[4, ]
```

Next, I identified the TMPRSS2-ERG fusion samples and mapped them to the corresponding patient IDs.

```
# Filter out TMPRSS2-ERG fusion samples
fusion_cases <- sv_data_raw[
  (sv_data_raw$Site1_Hugo_Symbol == "TMPRSS2" & sv_data_raw$Site2_Hugo_Symbol == "ERG") |
  (sv_data_raw$Site1_Hugo_Symbol == "ERG" & sv_data_raw$Site2_Hugo_Symbol == "TMPRSS2"),
]

# Map sample ID on patient ID
fusion_patient_ids <- unique(
  sample_data[sample_data$SAMPLE_ID %in% fusion_cases$Sample_Id, "PATIENT_ID"]
)
```

[1]For this mini-project, I used prostate cancer data downloaded from the cBioPortal for Cancer Genomics database for analysis.

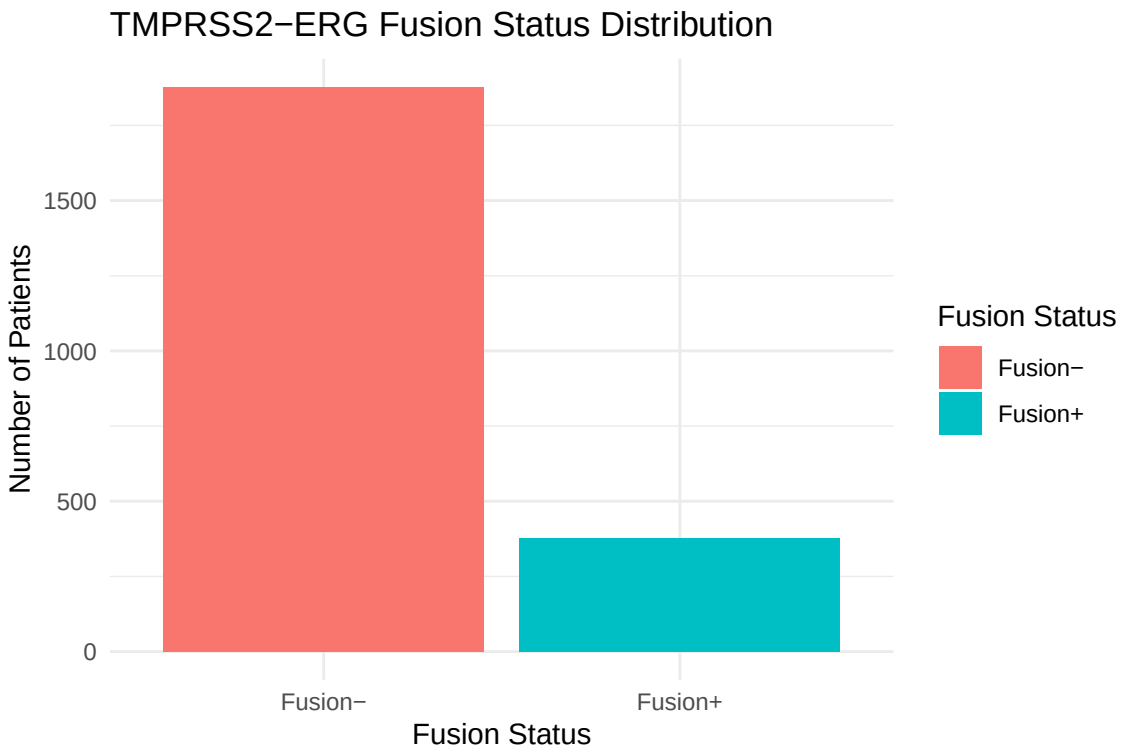
Finally, I created a new fusion status variable by assigning two values: `Fusion+` (indicating the presence of the TMPRSS2-ERG fusion) and `Fusion-` (indicating its absence). The corresponding data was then cleaned and integrated to facilitate subsequent analyses.

```
# Add fusion status column
clinical_data$fusion_status <- ifelse(
  clinical_data$PATIENT_ID %in% fusion_patient_ids, "Fusion+", "Fusion-"
)

# Count fusion status
fusion_freq <- as.data.frame(table(clinical_data$fusion_status))
colnames(fusion_freq) <- c("Fusion_Status", "Count")

# Extract life time and status
clinical_data$OS_MONTHS <- as.numeric(clinical_data$OS_MONTHS)
clinical_data$OS_STATUS <- clinical_data$OS_STATUS
clinical_data$event <- ifelse(grepl("DECEASED", clinical_data$OS_STATUS), 1, 0)

# Convert age and Gleason to numerical values
clinical_data$CURRENT_AGE <- as.numeric(clinical_data$CURRENT_AGE)
clinical_data$GLEASON_HIGHEST_REPORTED <- as.numeric(clinical_data$GLEASON_HIGHEST_REPORTED)

# Convert TMB to a numeric value
sample_data$TMB_NONSYNONYMOUS <- as.numeric(sample_data$TMB_NONSYNONYMOUS)

# Merge fusion status into sample_data (imported from clinical_data)
sample_data$fusion_status <-
  clinical_data$fusion_status[match(sample_data$PATIENT_ID, clinical_data$PATIENT_ID)]
```

3 Methods

3.1 Fusion vs Non-fusion Frequency Analysis

```
# Plot distribution bar chart
ggplot(fusion_freq, aes(x = Fusion_Status, y = Count, fill = Fusion_Status)) +
  geom_bar(stat = "identity") +
  labs(title = "TMPRSS2-ERG Fusion Status Distribution", x = "Fusion Status",
       y = "Number of Patients") +
  scale_fill_discrete(name = "Fusion Status") + theme_minimal()
```

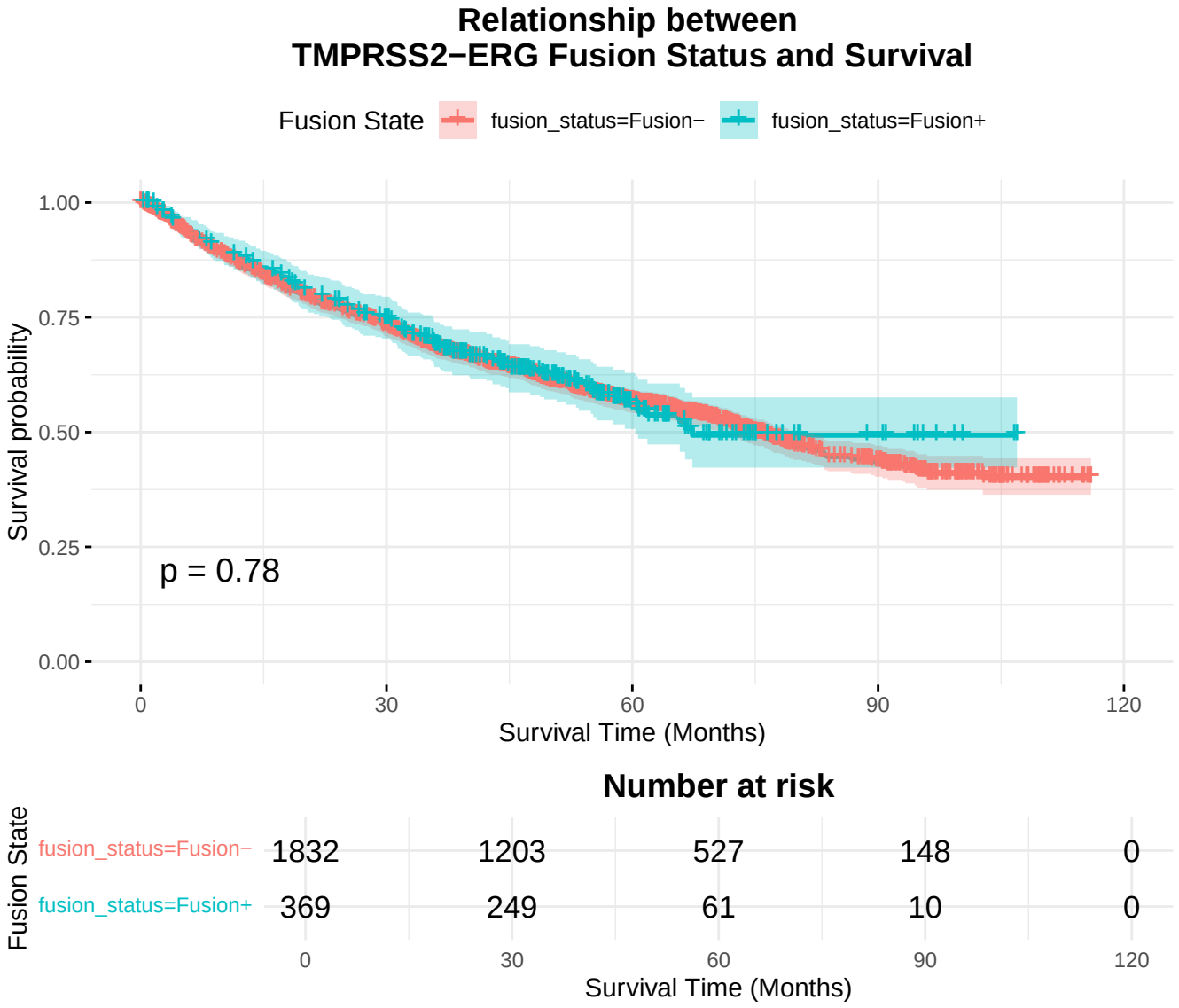


I first examined the distribution of patients in the two groups and found an imbalance: the number of `Fusion-` patients was approximately four times greater than that of `Fusion+`. However, previous studies have shown that TMPRSS2–ERG fusion is the most common gene fusion event in prostate cancer, with a prevalence ranging from 40% to 70% of cases (Rubin et al., 2011). This discrepancy suggests that the conclusion derived from this dataset may not fully reflect the general population.

3.2 Kaplan-Meier Survival Analysis

```
# Construct survival object
surv_obj <- Surv(clinical_data$OS_MONTHS, clinical_data$event)

# Fit survival models
fit <- survfit(surv_obj ~ fusion_status, data = clinical_data)

# Plot survival curves
ggsurvplot(
  fit,
  data = clinical_data,
  pval = TRUE,
  conf.int = TRUE,
  risk.table = TRUE,
  title = "Relationship between \nTMPRSS2-ERG Fusion Status and Survival",
  xlab = "Survival Time (Months)",
  legend.title = "Fusion State",
  ggtheme = theme_minimal() +
```

```
            theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
)
```



The results show that the two patient groups have almost identical survival curves, especially during the first 75 months. However, this conclusion is based on a limited number of TMPRSS2-ERG fusion samples, and the survival curve for the fusion group exhibits greater uncertainty, as indicated by a wider confidence interval—an observation further supported by the summary table below the survival curves. The p-value also supports this conclusion: there is no statistically significant difference between the two groups ($p = 0.78 \gg 0.05$). Therefore, based on these visual and statistical results alone, TMPRSS2-ERG fusion may not serve as an independent prognostic factor for survival time. Further analysis involving additional variables is required.

3.3 Cox multivariate survival analysis

```r
# Fit the Cox model
cox_fit <-
  coxph(Surv(OS_MONTHS, event) ~ fusion_status + CURRENT_AGE + GLEASON_HIGHEST_REPORTED,
        data = clinical_data)

# Output the fitted model
summary(cox_fit)
```

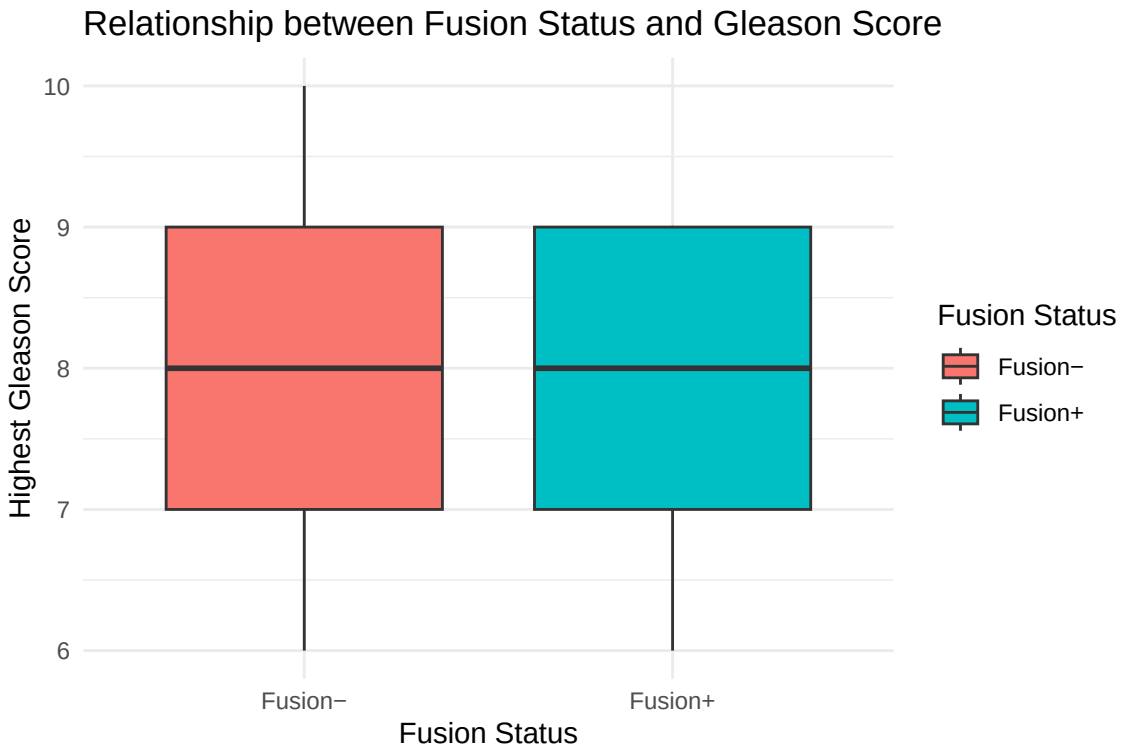
```
Call:
coxph(formula = Surv(OS_MONTHS, event) ~ fusion_status + CURRENT_AGE +
    GLEASON_HIGHEST_REPORTED, data = clinical_data)

  n= 1995, number of events= 783
   (262 observations deleted due to missingness)

                              coef exp(coef)  se(coef)      z Pr(>|z|)
fusion_statusFusion+     -0.021373  0.978854  0.101220 -0.211    0.833
CURRENT_AGE              -0.002290  0.997713  0.004183 -0.547    0.584
GLEASON_HIGHEST_REPORTED  0.322871  1.381088  0.042062  7.676 1.64e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
fusion_statusFusion+        0.9789     1.0216    0.8027     1.194
CURRENT_AGE                 0.9977     1.0023    0.9896     1.006
GLEASON_HIGHEST_REPORTED    1.3811     0.7241    1.2718     1.500

Concordance= 0.586  (se = 0.011 )
Likelihood ratio test= 63.31  on 3 df,   p=1e-13
Wald test            = 60  on 3 df,   p=6e-13
Score (logrank) test = 61.34  on 3 df,   p=3e-13
```

I considered two other potentially relevant factors: **current age** and **Gleason score**, and included them in a multivariable Cox proportional hazards model. The results showed that TMPRSS2-ERG fusion still had no statistically significant association with overall survival (HR = 0.98, p = 0.83). In contrast, Gleason score emerged as a clearly significant factor—each one-point increase in Gleason score was associated with approximately a 38% increase in the risk of death—highlighting its importance in prostate cancer prognosis.

3.4 Correlation Analysis between Fusion Status and Clinical Index Gleason

```r
# Draw the box-plot
ggplot(
  clinical_data, aes(x = fusion_status, y = GLEASON_HIGHEST_REPORTED, fill = fusion_status)) +
  geom_boxplot() +
  labs(title = "Relationship between Fusion Status and Gleason Score",
```

```
       x = "Fusion Status", y = "Highest Gleason Score") +
  scale_fill_discrete(name = "Fusion Status") + theme_minimal()
```



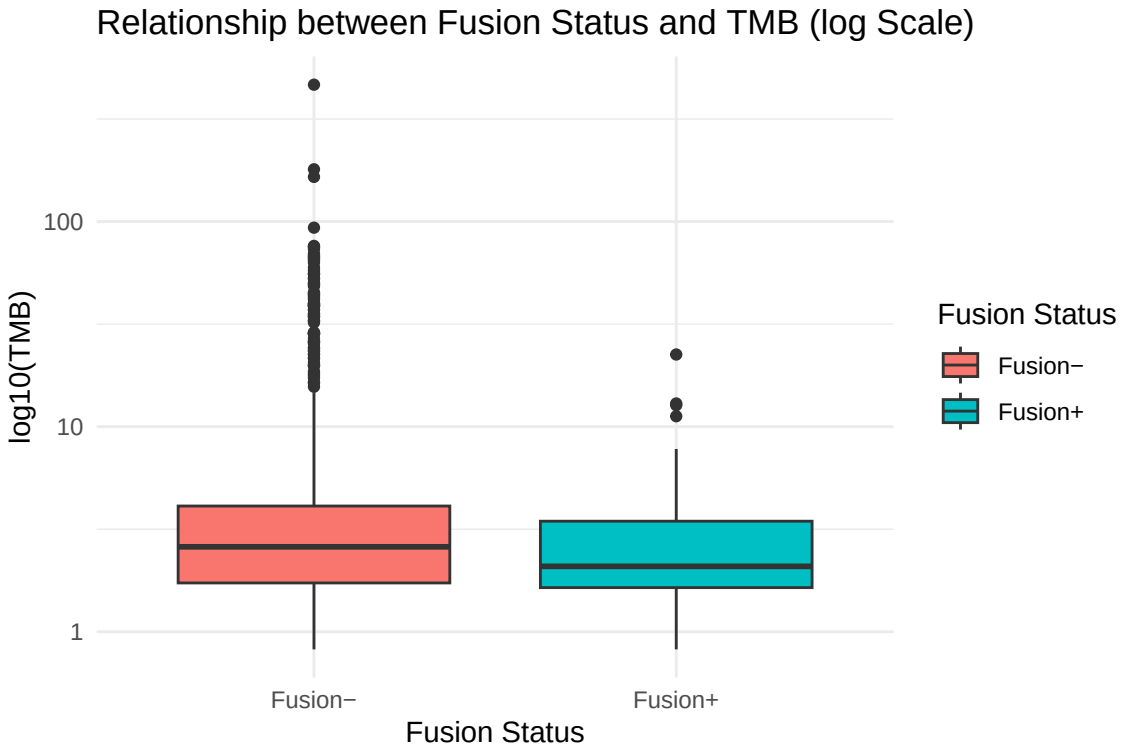

```
# T-test
gleason_ttest <- t.test(GLEASON_HIGHEST_REPORTED ~ fusion_status, data = clinical_data)
print(gleason_ttest)
```

```
    Welch Two Sample t-test

data:  GLEASON_HIGHEST_REPORTED by fusion_status
t = 1.998, df = 487.35, p-value = 0.04627
alternative hypothesis: true difference in means between group Fusion- and group Fusion+ is not equal
95 percent confidence interval:
 0.001804787 0.215810061
sample estimates:
mean in group Fusion- mean in group Fusion+
             8.158515              8.049708
```

Although the box plots of Gleason scores for the two fusion status groups appear similar, the **t-test** revealed a statistically significant difference ($p = 0.046$), suggesting that the `Fusion+` group may tend to have a lower level of tumor differentiation.

3.5 Correlation Analysis between Fusion Status and Clinical Index Tumor Mutational Burden (TMB)

```
# Draw the box-plot
ggplot(sample_data, aes(x = fusion_status, y = TMB_NONSYNONYMOUS, fill = fusion_status)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(title = "Relationship between Fusion Status and TMB (log Scale)",
       x = "Fusion Status",
       y = "log10(TMB)") +
  scale_fill_discrete(name = "Fusion Status") + theme_minimal()
```




```
# T-test
tmb_ttest <- t.test(TMB_NONSYNONYMOUS ~ fusion_status, data = sample_data)
print(tmb_ttest)
```

```
    Welch Two Sample t-test

data:  TMB_NONSYNONYMOUS by fusion_status
t = 6.7922, df = 2153.2, p-value = 1.424e-11
alternative hypothesis: true difference in means between group Fusion- and group Fusion+ is not equal
95 percent confidence interval:
 1.703851 3.087103
sample estimates:
```

```
mean in group Fusion- mean in group Fusion+
             4.663252              2.267775
```

Based on a comparison involving approximately 2,200 patients, the Welch's t-test showed that the average tumor mutation burden (TMB) for the `Fusion-` group was 4.66, whereas for the `Fusion+` group it was only 2.27. This difference was highly statistically significant ($p < 1e\text{-}10$). The box plot also illustrates this result clearly: the `Fusion-` group displays a wider spread and includes more extreme high-TMB samples. Together, these findings suggest that the TMPRSS2-ERG fusion may be associated with lower gene mutation burden.

4 Results

In this study, I investigated TMPRSS2–ERG gene fusion is associated with worse clinical outcomes in prostate cancer. Cox multivariate survival analysis revealed no significant difference in overall survival between fusion-positive and fusion-negative patients (HR = 0.98, p = 0.83), indicating that TMPRSS2–ERG fusion is not an independent prognostic factor for survival. Although a marginal difference in Gleason scores was observed (Fusion− mean = 8.16 vs. Fusion+ = 8.05, p = 0.046), the magnitude of this difference was clinically negligible. I still identified a strong association between TMPRSS2–ERG fusion and tumor mutational burden (TMB): fusion-negative tumors had significantly higher TMB values ($p < 1e\text{-}10$), suggesting greater genomic instability and potential responsiveness to immunotherapy.

Taken together, while TMPRSS2–ERG fusion does not appear to affect survival directly, it marks a biologically distinct subtype of prostate cancer characterized by lower mutation load and potentially lower immunogenicity.

5 Reference

[1] *Cancer in Ireland 1994-2022: Annual statistical report of the National Cancer Registry (2024).* (2024). National Cancer Registry Ireland. https://www.ncri.ie/en/reports-publications/reports/cancer-in-ireland-1994-2022-annual-statistical-report-of-the-national.

[2] Wikipedia Contributors. (2019, April 12). *Prostate cancer.* Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Prostate_cancer.

[3] *cBioPortal for Cancer Genomics.* (2025). Cbioportal.org. https://www.cbioportal.org/study/summary?id=prostate_msk_2024.

[4] Rubin, M. A., Maher, C. A., & Chinnaiyan, A. M. (2011). Common Gene Rearrangements in Prostate Cancer. *Journal of Clinical Oncology, 29*(27), 3659–3668. https://doi.org/10.1200/jco.2011.35.1916.