

****Abstract****

This project aims to analyze which treatment option for prostate cancer leads to the highest recovery rate among patients. Prostate cancer is one of the most common cancers among men and is ranked as the "second most frequently diagnosed cancer"(Rawla, 2019). It also remains the "fifth leading cause of death worldwide" (Rawla, 2019), making the identification of the most effective treatment essential. As research in cancer therapy continues to expand, it is imperative to compare the overall effectiveness of different treatment methods, including hormonal therapy, radiation therapy, surgery, and combination therapies, to determine whether a single or combined approach yields better recovery outcomes. This study will evaluate tumor size, initial PSA levels, and biopsy Gleason scores over a one-year period to compare patient recovery and cancer recurrence. Statistical analyses will identify patterns of effectiveness among various treatments and determine which method best improves survival rates while reducing recurrence. It is hypothesized that combination treatments will decrease survival and decrease recurrence, whereas single therapies may yield higher recurrence rates but increased survival rates. By analyzing these relationships, this project seeks to highlight how treatment choice influences recovery rates, recurrence, and long-term survival outcomes for prostate cancer patients.

1) Introduction:

The primary purpose of this analysis is to compare and quantify the long-term differential effectiveness of major treatment modalities for prostate cancer—specifically surgery, radiotherapy, hormonal therapy, and a combination of hormonal therapy with radiotherapy. The goal is to apply meticulous statistical methods and rigorous data analysis to determine which specific treatment provides the best outcomes for patient survival and recovery, as well as whether combining therapies is more effective than single-treatment approaches. This study directly addresses the central research question: “What is the differential effectiveness of primary treatment modalities on clinical survival rates and biochemical recurrence outcomes for prostate cancer patients after controlling for prognostic factors?” The analysis is designed to benefit patients by identifying treatments with the most durable effectiveness, while also assisting families and clinicians in making informed decisions based on reliable statistical evidence. By providing outcome-based comparisons, this research aims to guide more personalized treatment strategies tailored to each patient’s risk profile and cancer stage. Patient-specific factors, including initial cancer stage (severity and aggressiveness), overall health status, tumor size, initial PSA levels, and biopsy Gleason scores, will be considered. A multivariable regression model will be employed to isolate the effectiveness of treatment type while controlling for these prognostic factors to test the hypothesis that significant differences exist in long-term outcomes among first-line treatments..

2) Data:

The data used for this analysis were retrieved from the Zenodo repository, a well-known open-access repository mainly used by science communities. A dataset we’re using, titled “Comprehensive Clinical, Pathological and Follow-up Dataset of Prostate Cancer Patients,” was collected and published by Mert Başaranoğlu at Mersin Üniversitesi Hastanesi (Mersin University Hospital) in southern Turkey. It contains detailed information on factors related to Prostate Cancer, including 30 variables spanning four broader categories: Clinical Parameters, Treatment Information, Pathological Outcomes, and Follow-up Data.

2-1) Dataset Variables:

```
##      Variable_Name
## 1      Hasta_ID
## 2      Yas
```

```

## 3      Tani_Tarihi
## 4      PSA_Tani
## 5      Klinik_Evre
## 6 Biyopsi_Gleason
## 7      Risk_Grubu
##
## 1
## 2
## 3
## 4
## 5 Clinical cT-Stage determined by pre-treatment examinations (cT1c < cT2a < cT2b < cT2c < cT3a < cT3b)
## 6      Biopsy Gleason Score (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, high grade)
## 7      Risk Group Classification
##      Type_of_Data
## 1      Categorical
## 2      Discrete
## 3      Date
## 4      Continuous
## 5 Ordinal/Categorical
## 6 Ordinal/Categorical
## 7 Ordinal/Categorical

##      Variable
## 1      Albumin
## 2      Lenfosit
## 3      CRP
## 4      NLR
## 5      CALLY_Index
## 6 Komorbidite_Skor
##
## 1      ASerum albumin level (g/dL). Indicator of malnutrition
## 2      Lymphocyte count (x10^9/L). Indicator of lymphopenia
## 3      C-Reactive Protein (mg/dL). Indicator of inflammation
## 4      Neutrophil-to-Lymphocyte Ratio. A prognostic indicator for systemic inflammation
## 5      CALLY Index. A composite index, likely related to the C-Reactive Protein
## 6 Comorbidity Score indicating the severity of other co-existing chronic diseases ( 0 (No comorbidity) to 5 (Severe comorbidity))
##      Type_of_Data
## 1      Continuous
## 2      Discrete
## 3      Continuous
## 4      Continuous
## 5      Continuous
## 6 Ordinal/Categorical

```

Treatment Information

```

##      Variable
## 1      Tedavi_Tipi
## 2 Tedavi_Tarihi
## 3      RT_Dozu
## 4      ADT_Tipi
## 5      ADT_Suresi
##

```

```

## 1 Main Treatment Type received (1 for Radical Prostatectomy, 2 for\nRadiotherapy/RT, 3 for Hormone T
## 2
## 3
## 4
## 5
##   Type_of_Data
## 1   Categorical
## 2       Date
## 3   Continuous
## 4   Categorical
## 5   Continuous

##           Variable
## 1 Patolojik_Evre
## 2   Cerrahi_Sinir
## 3   Final_Gleason
##
## 1                               Final Tumor Pathological Stage determined after surgery o
## 2 Surgical Margin Status indicating if cancer cells were present at the edge of the removed tissue. (
## 3                               Final Gleason Score confirmed from the final c
##           Type_of_Data
## 1 Ordinal/Categorical
## 2   Binary/Categorical
## 3 Ordinal/Categorical

##           Variable
## 1           PSA_Nadir
## 2 PSA_Takip_3ay / 6ay / 12ay
## 3           BCR_Durum
## 4           BCR_Tarihi
## 5           Metastaz_Durum
## 6           Metastaz_Tarihi
## 7           Son_Durum
## 8           Son_Takip_Tarihi
##
## 1                               The lowest PSA level reached after treatment (ng/mL).
## 2
## 3 Biochemical Recurrence (BCR) Status whether the PSA level rise above a recurrence threshold? (True
## 4
## 5                               Metastasis Status whether distan
## 6
## 7                               Patient's Survival :
## 8
##           Type_of_Data
## 1           Continuous
## 2           Continuous
## 3 Binary/Categorical
## 4           Date
## 5 Binary/Categorical
## 6           Date
## 7 Binary/Categorical
## 8           Date

```

Data Cleaning

To make data sets more consistent, we then decided to delete unnecessary information by removing the corresponding columns. Afterward, converting all categorical variables into factor type was required to ensure that our program, R(Studio), correctly interprets them as discrete categories rather than as continuous values or even strings. Moreover, this foreign data set has named its critical variables in another language that we need to rename in English. Lastly, addressing missing variables and visualizing were involved in each of them in various forms.

Visualization:

The preliminary visualizations and numerical summaries are crucial for preparing and interpreting the three multivariate logistic regression models, as this step is essential for diagnosing data quality, identifying confounding factors, and assessing the signal strength before the formal statistical analysis begins. We've established a few bar graphs on the frequency of treatment type, frequency of biochemical recurrence status, survival status, and risk factors. The box plots on PSA level at diagnosis and after treatment will be used to directly confound our numerical data among different treatment groups, as well as interpreting information from risk factors through our numerical summary. We will not use any scatter plot in this model, but we are preparing and interpreting the three multivariate logistic regression modes.

Analysis aims:

The main purpose of our analysis is to compare and quantify the long-term discriminatory effects of major first-line treatment methods for patients with prostate cancer. Some of the analyses we aim to complete will include linear regression models, box plot visualizations, R plots, and numerical summaries. We will apply statistical tests such as ANOVA and t-tests to compare treatment effectiveness across different clinical stages, Gleason scores, and recovery outcomes. If prostate cancer is less aggressive (lower Gleason score and early clinical stage), we expect surgery to yield the best outcomes due to its practicality and reduced long-term side effects compared to radiation or hormonal therapy. To test this, we will use box plots to visually compare recovery rates and PSA level reductions among treatment groups, followed by a one-way ANOVA to test for significant differences in mean recovery rates. We hypothesize that patients with localized, low-stage tumors (Stages 1–2) receiving surgery will have significantly higher recovery rates and lower recurrence probabilities. If the cancer is more aggressive or has metastasized, we hypothesize that hormonal therapy will be more effective because hormones circulate through the bloodstream and can target systemic tumor growth. A multivariate logistic regression model will be used to assess the relationship between treatment type and recurrence or metastasis, controlling for PSA level, tumor size, and patient age. For early-stage cancers (Stages 1–2), radiation therapy may also be effective as a minimally invasive treatment. We will visualize PSA levels and Gleason scores across radiation and surgery groups using box plots, and conduct a t-test to evaluate whether differences in mean PSA reduction between these groups are statistically significant. In advanced stages (Stages 3–4), a combination of hormonal and radiation therapy is expected to provide the best results. We will test this using multivariate logistic regression models for survival and recurrence outcomes, hypothesizing that combination therapy will be associated with the highest survival rates but also with greater treatment intensity and side effects. Overall, these analyses will allow us to statistically evaluate how treatment modality interacts with cancer aggressiveness, Gleason score, and clinical stage to influence recovery, recurrence, and long-term survival outcomes.

Work Cited (Used in Abstract)

Rawla P. (2019). Epidemiology of Prostate Cancer. World journal of oncology, 10(2), 63–89. <https://doi.org/10.14740/wjon1191>