Title: Treatment Effectiveness in Patients with Prostate Cancer

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.

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..

Data:

The dataset we are using for this analysis is retrieved from the Zendo repository (https://zenodo.org/records/15007105), titled "Comprehensive Clinical, Pathological and Follow-up Dataset of Prostate Cancer Patients." It was collected and published by Mert Başaranoğlu at Mersin Üniversitesi Hastanesi (Mersin University Hospital), in southern Turkey. There are a lot of variables and numbers that we can collect through this data set since it includes 600 observations of individual prostate cancer patients' records. Overall, the dataset contains 30 variables related to clinical parameters, including patient demographics, diagnostic and treatment information, pathological outcomes, and follow-up data. There are a few steps that we've used to clean the data. These are elaborated in this order: identify which variables we're going to use in this project, delete unnecessary information (columns) afterward, convert all categorical variables into factor form, changing unnoticeable name to English (since it's a data from other country), addressing missing variable and visualizing each of them in various form. Our data will not be generalized through a randomized simulation.

Dataset Variables:

Initial Diagnosis

Hasta_ID (Categorical): Patient ID

Yas (Discrete): Age

Tani_Tarihi (Date): Diagnosis Date

PSA_Tani (Continuous): Serum Prostate-Specific Antigen (PSA) level at diagnosis (ng/mL)

Klinik_Evre (Ordinal/Categorical): Clinical cT-Stage determined by pre-treatment examinations (cT1c < cT2a < cT2b < cT2c < cT3a < cT3b for increasing extent of tumor invasion)

Biyopsi_Gleason (Ordinal/Categorical): Biopsy Gleason Score (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, higher score indicates higher aggressiveness)

Risk_Grubu (Ordinal/Categorical): Risk Group Classification (1 for Low, 2 for Intermediate, 3 forHigh)

Risk Factors

Albumin (Continuous): ASerum albumin level (g/dL). Indicator of nutritional status and systemic health

Lenfosit (Discrete): Lymphocyte (Immune system component) Count

CRP (Continuous): C-Reactive Protein (mg/L). Indicator of inflammation

NLR (Continuous): Neutrophil-to-Lymphocyte Ratio. A prognostic indicator for systemic inflammation and cancer aggressiveness.

CALLY_Index (Continuous): CALLY Index. A composite index, likely related to inflammation or blood components.

Komorbidite_Skor (Ordinal/Categorical): Comorbidity Score indicating the severity of other co-existing chronic diseases (0 (No comorbidities) < ... < 5 (Severe comorbidities))

Treatment Information

Tedavi_Tipi (Categorical): Main Treatment Type received (1 for Radical Prostatectomy, 2 for Radiotherapy/RT, 3 for Hormone Therapy, 4 for Combination of Radiotherapy and Hormone Therapy)

Tedavi_Tarihi (Date): Treatment Date

RT_Dozu (Continuous): Total Radiation Dose (in Gy), if radiotherapy was performed

ADT_Tipi (Categorical): Androgen Deprivation Therapy (ADT, hormone therapy) Type used

ADT_Suresi (Continuous): ADT(hormone therapy) Duration

Pathological Markers

Patolojik_Evre (Ordinal/Categorical): Final Tumor Pathological Stage determined after surgery on the removed tissue (pT2a < pT2b < pT2c < pT3a < pT3b < pT4, NaN indicates patient did not undergo surgery)

Cerrahi_Sinir (Binary/Categorical): Surgical Margin Status indicating if cancer cells were present at the edge of the removed tissue. Crucial for recurrence prediction (0: Negative, 1: Positive, NaN indicates patient did not undergo surgery).

Final_Gleason (Ordinal/Categorical): Final Gleason Score confirmed from the final excised tissue (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, higher score indicates higher aggressiveness)

Follow-up & Outcomes

PSA_Nadir (Continuous): The lowest PSA level reached after treatment (ng/mL). A lower nadir generally indicates better treatment success

PSA_Takip_3ay / 6ay / 12ay (Continuous): Follow-up PSA levels (ng/mL) measured at at 3/6/12 Months

BCR_Durum (Binary/Categorical): Biochemical Recurrence (BCR) Status whether the PSA level rise above a recurrence threshold? (True for Recurrence occurred, False for no Recurrence occurred)

BCR_Tarihi (Date): Date when biochemical recurrence was confirmed

Metastaz_Durum (Binary/Categorical): Metastasis Status whether distant metastasis occur during follow-up? (0 for No, 1 for Yes)

Metastaz_Tarihi (Date): Date when metastasis was confirmed

Son_Durum (Binary/Categorical): Patient's Survival Status at the last follow-up (0 for Alive, 1 for Deceased)

Son_Takip_Tarihi (Date): Date of the last recorded patient information.

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