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.

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, especially surgery, radiation therapy, hormone therapy, and a combination of radiation and hormone methods. Our main goal is to use careful statistical methods to determine which specific treatment provides the best results, and whether combining the treatments is more effective than a single method. This directly addresses our central research question, “What is the differential effect of first-line treatment methods on clinical (survival) and biochemical outcome (recurrence) after controlling prognostic factors in patients with prostate cancer?” This analysis will help both patients and their families by helping inform and personalize treatment strategies based on initial risk profiles. It will also provide an analysis that will help clinicians identify PSA regulation and prediction of survival among various risk groups. We will consider the patient’s initial cancer severity and health status, and test several points to verify the hypothesis that significant differences exist in the long-term outcomes between the first-line treatments. More specifically, it is likely that multivariate regression models will be used to isolate the effect of treatment type from other strong prognostic factors such as initial PSA, tumor size, Gleason score/risk group, and patient age. Therefore, to test these assumptions, we will construct three separate multivariate logistic regression models targeting binary outcomes of recurrence, metastasis, and overall survival.

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