Treatment Effectiveness in Patients with Prostate Cancer

Jennifer Cho, Narya Elhawary, Minsol Seo

Due 2025-12-05

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:

Table 1: Initial Diagnosis Variables

Variable_Name	English_Translation	Type_of_Data
Hasta_ID	Patient ID	Categorical
Yas	Age	Discrete
Tani_Tarihi	Diagnosis Date	Date
PSA_Tani	Serum Prostate-Specific Antigen (PSA) level at diagnosis (ng/mL)	Continuous
Klinik_Evre	Clinical cT-Stage determined by pre-treatment examinations (cT1c < cT2a < cT2b < cT2c < cT3a < cT3b for increasing extent of tumor invasion)	Ordinal/Categorical
Biyopsi_Gleason	Biopsy Gleason Score (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, higher score indicates higher aggressiveness)	Ordinal/Categorical
Risk_Grubu	Risk Group Classification (1 for Low, 2 for Intermediate, 3 forHigh)	Ordinal/Categorical

Table 2: Risk Factors

Variable	In_English	Type_of_Data
Albumin	ASerum albumin level (g/dL). Indicator of nutritional status and systemic health	Continuous
Lenfosit	Lymphocyte (Immune system component) Count	Discrete
CRP	C-Reactive Protein (mg/L). Indicator of inflammation	Continuous
NLR	Neutrophil-to-Lymphocyte Ratio. A prognostic indicator for systemic inflammation and cancer aggressiveness.	Continuous
CALLY_Index	CALLY Index. A composite index, likely related to inflammation or blood components.	Continuous
Komorbidite_Skor	Comorbidity Score indicating the severity of other co-existing chronic diseases (0 (No comorbidities) < ... < 5 (Severe comorbidities))	Ordinal/Categorical

Treatment Information

Table 3: Treatment Information

Variable	In_English	Type_of_Data
Tedavi_Tipi	Main Treatment Type received (1 for Radical Prostatectomy, 2 for Radiotherapy/RT, 3 for Hormone Therapy, 4 for Combination of Radiotherapy and Hormone Therapy)	|Categorical
Tedavi_Tarihi	Treatment Date	Date
RT_Dozu	Total Radiation Dose (in Gy), if radiotherapy was performed	Continuous
ADT_Tipi	Androgen Deprivation Therapy (ADT, hormone therapy) Type used	Categorical
ADT_Suresi	ADT(hormone therapy) Duration	Continuous

Table 4: Pathological Markers

Variable	In_English	Type_of_Data
Patolojik_Evre	Final Tumor Pathological Stage determined after surgery on the removed tissue (pT2a < pT2b < pT2c < pT3a < pT3b < pT4, NaN indicates patient did not undergo surgery)	Ordinal/Categorical
Cerrahi_Sinir	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).	Binary/Categorical
Final_Gleason	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)	Ordinal/Categorical

Table 5: Follow-up & Outcomes

Variable	In_English	Type_of_Data
PSA_Nadir	The lowest PSA level reached after treatment (ng/mL). A lower nadir generally indicates better treatment success	Continuous
PSA_Takip_3ay / 6ay / 12ay	Follow-up PSA levels (ng/mL) measured at at 3/6/12 Months	Continuous
BCR_Durum	Biochemical Recurrence (BCR) Status whether the PSA level rise above a recurrence threshold? (True for Recurrence occurred, False for no Recurrence occurred)	Binary/Categorical

BCR_Tarihi	Date when biochemical recurrence was confirmed	Date
Metastaz_Durum	Metastasis Status whether distant metastasis occur during follow-up? (0 for No, 1 for Yes)	Binary/Categorical
Metastaz_Tarihi	Date when metastasis was confirmed	Date
Son_Durum	Patient's Survival Status at the last follow-up (0 for Alive, 1 for Deceased) |	inary/Categorical |
Son_Takip_Tarihi	Date of the last recorded patient information.	Date

2-2) 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. These process allow us to observe that missing values were only found for the Hormone Type (ADT_Tipi) variable in patients who did not receive Hormone Therapy (as it is not applicable). Also, information is only present for the specific treatment a patient received (e.g., Radiation Dose is missing for surgery patients and vise versa).

3) Visualization:

We use preliminary visualizations and numerical summaries in order to prepare and interpreting the two multivariate logistic regression models and one linear regression, 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.

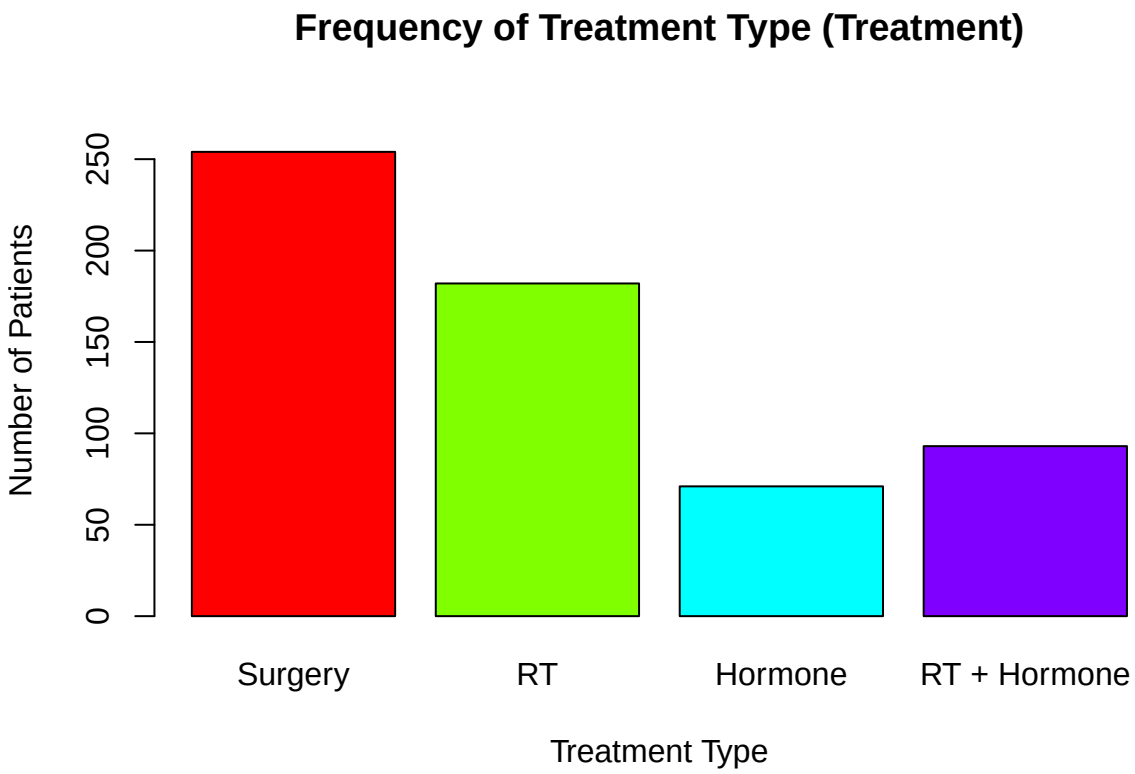


Figure 1: Illustration of distribution of the four main treatment groups across the patient cohort.

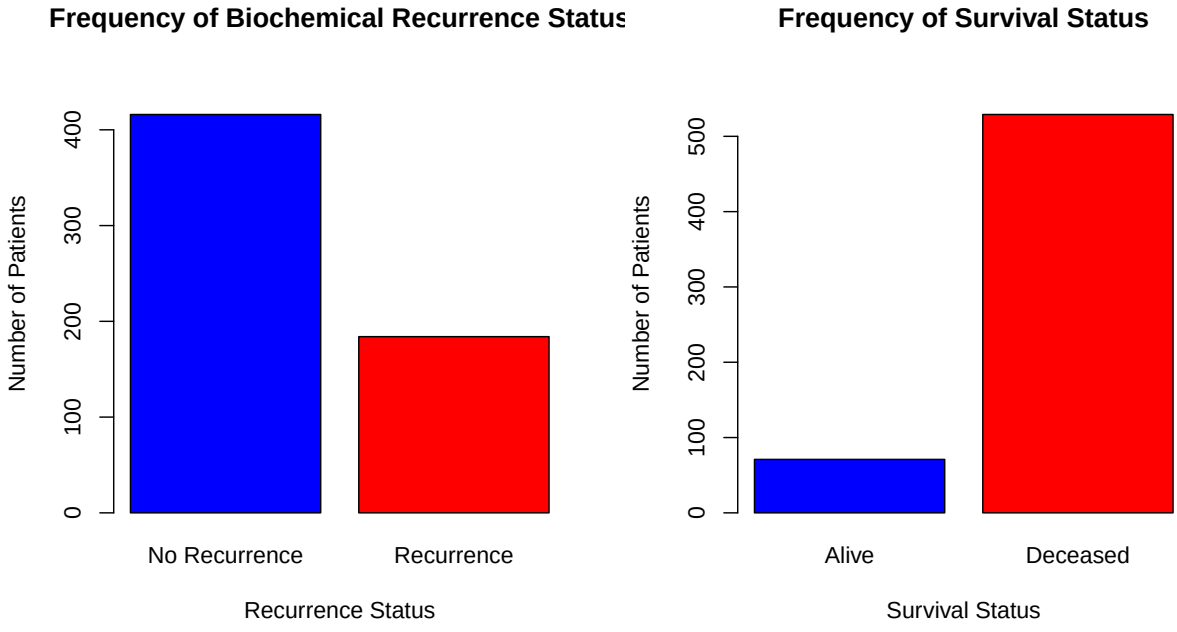


Figure 2: Side-by-side bar plots for Biochemical Recurrence (BCR) and Survival Status to display the main outcomes.

As shown in Figure 2, the data exhibit a low overall recurrence rate of Biochemical Recurrence (BCR), with the majority of patients (69.3%) successfully avoiding recurrence, which is represented as 'False'. However, the surprising and ironic cohort is Survival Status, where 0 = Alive and 1 = Deceased, showing that the survival rate is significantly low, with only 11.8% alive.

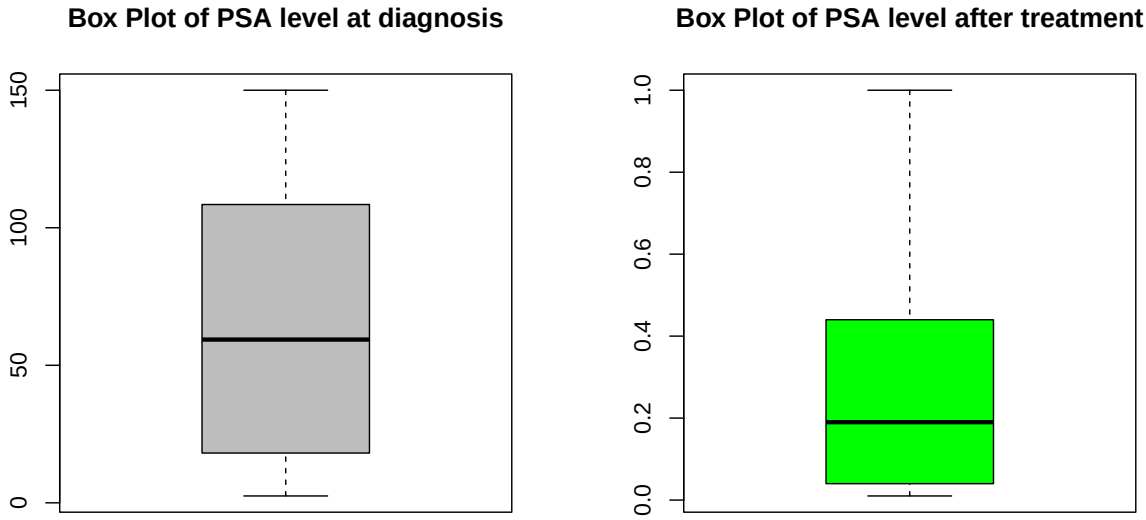


Figure 3: Comparison of PSA level before and after treatment on Diagnosis

Figure 3 displays the distributions and central tendencies of Prostate-Specific Antigen (PSA) levels at the time of diagnosis and after primary treatment. The box plots and numerical summaries illustrate a significant and positive effect of the treatment modalities on reducing PSA levels in the cohort. Before treatment (PSA_before), the median PSA level was 59.35 ng/mL, with values ranging up to 150.00 ng/mL. Following treatment (PSA_after), the median level dropped dramatically to just 0.1900 ng/mL, with most values lying significantly below the pre-treatment range. This substantial decrease in both the median and quartiles (as evidenced by the significant drop from the 1st Quartile of 18.10 to 0.0400 indicates that the treatments were immediately successful in reducing the primary tumor burden and systemic PSA activity in the majority of patients. This observation aligns with clinical expectations, in which successful treatment should drive PSA levels to near-undetectable levels (nadir).

4) Analysis

PART A) TEST

4A-1) Chi-squared Test on Treatment vs. BCR

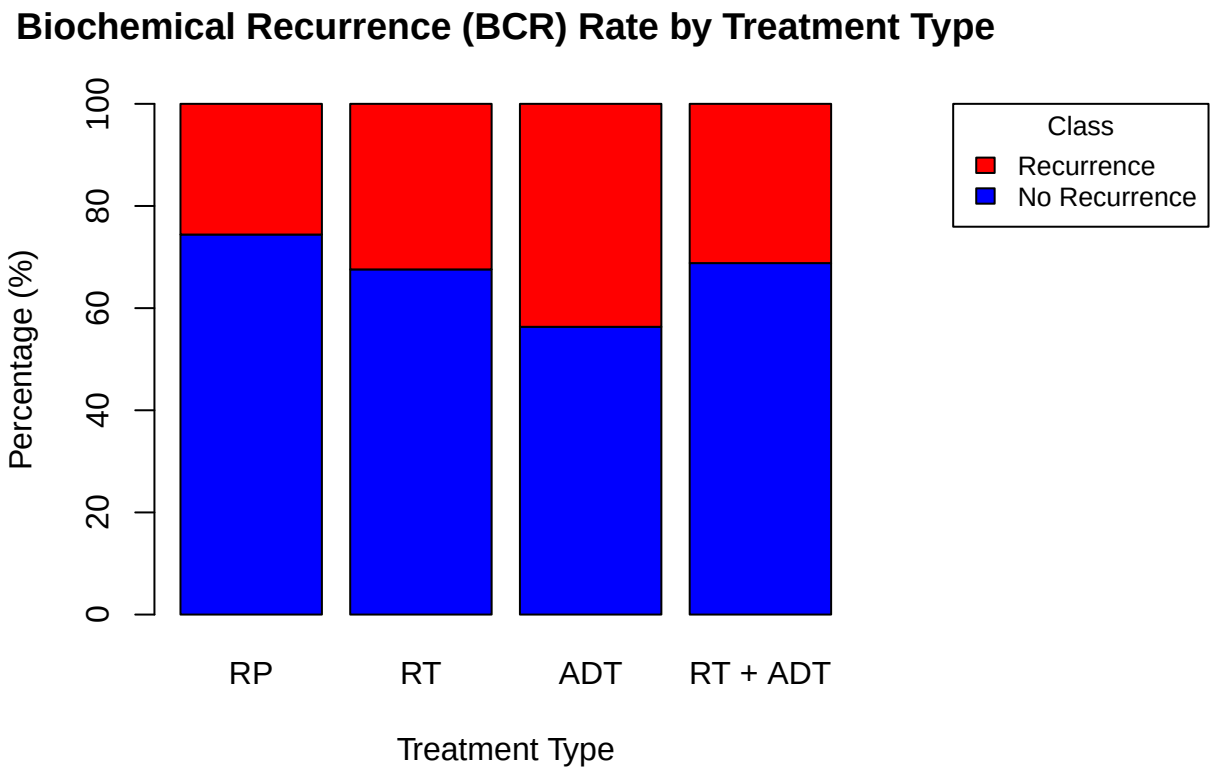
A Chi-squared test was performed to examine the association between Biochemical Recurrence (BCR) and treatment type.

- Null Hypothesis (H0): Treatment Type and Biochemical Recurrence (BCR) Status are independent (The recurrence rate is the same across all treatment groups.)
- Alternative Hypothesis (Ha): Treatment Type and Biochemical Recurrence (BCR) Status are not independent (The recurrence rate is significantly different for at least one treatment group.)

	Treatment Type	Recurrence Count (False)	Recurrence Count (True)	No Recurrence (%) (False)	Recurrence (%) (True)
Radical Prostatectomy	Radical Prostatectomy	189	65	74.4	25.6
Radiotherapy (RT)	Radiotherapy (RT)	123	59	67.6	32.4
Hormone Monotherapy	Hormone Monotherapy	40	31	56.3	43.7
Combination (RT + Hormone)	Combination (RT + Hormone)	64	29	68.8	31.2

```
## Chi-squared = 8.9915, df = 3, p-value = 0.0294
```

The Chi-squared test yielded a test statistic of chi^2 = 8.9915 with a corresponding p-value of 0.0294. Since the p-value is less than the significance level (alpha (a) = 0.05), we reject the null hypothesis. There is a statistically significant association between the type of treatment a patient receives and the likelihood of experiencing Biochemical Recurrence (BCR).



Hormone Monotherapy had the highest recurrence rate at 43.7%, while Radical Prostatectomy had the lowest at 25.6%. The bar plot visually confirms this, with the red recurrence portion for Hormone Monotherapy

(ADT) notably larger and the recurrence portion for Radical Prostatectomy (RP) noticeably smaller compared to other treatment methods. This result showcases a visual pattern that aligns with the statistical analysis of the chi square test. Thus a multivariate logistic regression analysis can further quantify treatment differences amongst probability of reoccurence of cancer post surgery.

4A-2) Chi-squared Test on Treatment vs. Survival

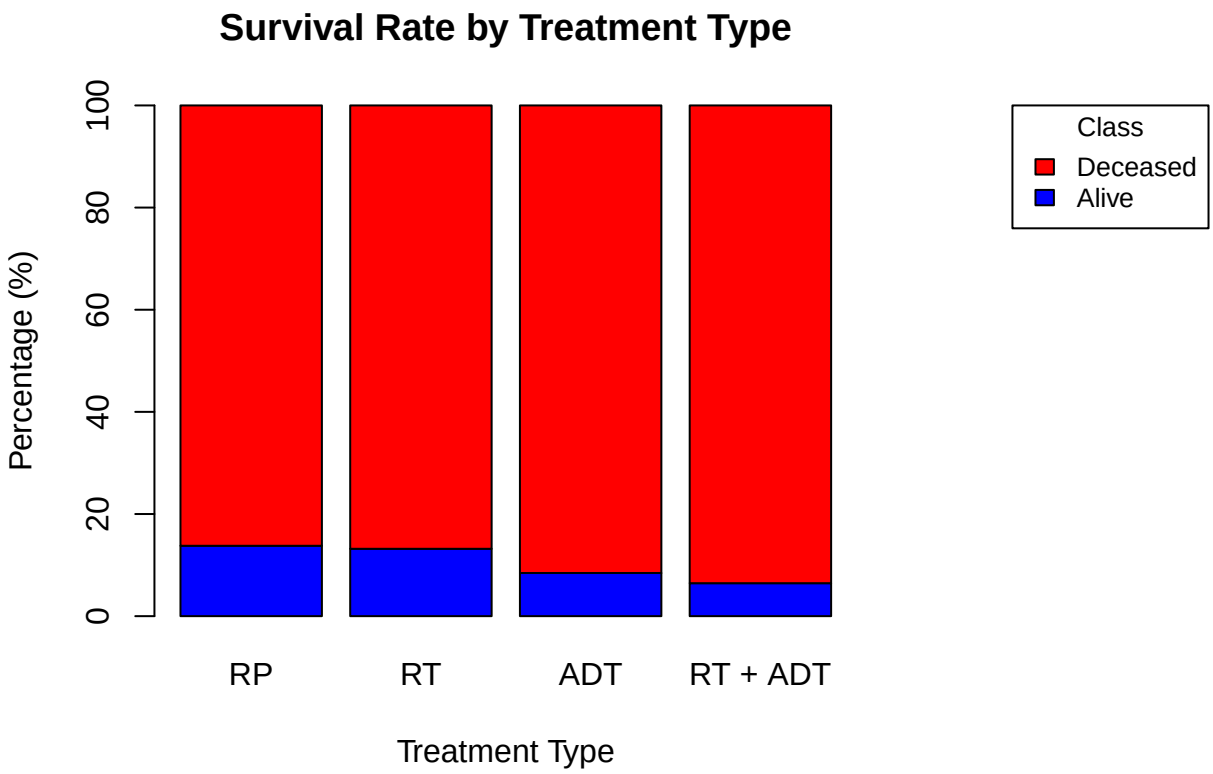
A Chi-squared test was performed to examine the association between Biochemical Recurrence (BCR) and treatment type.

- Null Hypothesis (H0): Treatment Type and Biochemical Recurrence (BCR) Status are independent (The survival rate is the same across all treatment groups.)
- Alternative Hypothesis (Ha): Treatment Type and Biochemical Recurrence (BCR) Status are not independent (The survival rate is significantly different for at least one treatment group.)

	Treatment Type	Count (Alive)	Count (Deceased)	Survival (%)	Mortality (%)
Radical Prostatectomy	Radical Prostatectomy	35	219	13.8	86.2
Radiotherapy (RT)	Radiotherapy (RT)	24	158	13.2	86.8
Hormone Monotherapy	Hormone Monotherapy	6	65	8.5	91.5
Combination (RT + Hormone)	Combination (RT + Hormone)	6	87	6.5	93.5

```
## Chi-squared = 4.6021, df = 3, p-value = 0.2034
```

The Chi-squared test yielded a test statistic of chi^2 = 4.6021 with a corresponding p-value of 0.2034. Since the p-value is greater than the significance level (alpha (a) =0.05), we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is no statistically significant association between the type of treatment a patient receives and the likelihood of their Survival Status (Alive vs. Deceased). The bar plot also supports this conclusion, showing that the distribution of survival rates is not substantially different across the four treatment types. This is likley due to the fact that survival is influenced by multiple clincal factors as well as enviromental factors outside of pateint data. A multivariate logistic approach can be a better assessment to determine best treatment type.



The bar plot also supports this conclusion, showing that the survival rate distribution is not substantially different across the four treatment types.

4A-3) ANOVA Test, Treatment vs PSA Difference

An ANOVA was conducted to compare mean PSA at diagnosis (PSA_before) across the four Treatment Type groups.

- Null Hypothesis (H0): PSA at diagnosis is the same across all four Treatment Type groups
- Alternative Hypothesis (Ha): Alternative Hypothesis (H1) PSA at diagnosis is significantly different for at least one treatment group

Table 8: Table 4.5: Tukey's HSD Post-Hoc Test for Delta PSA

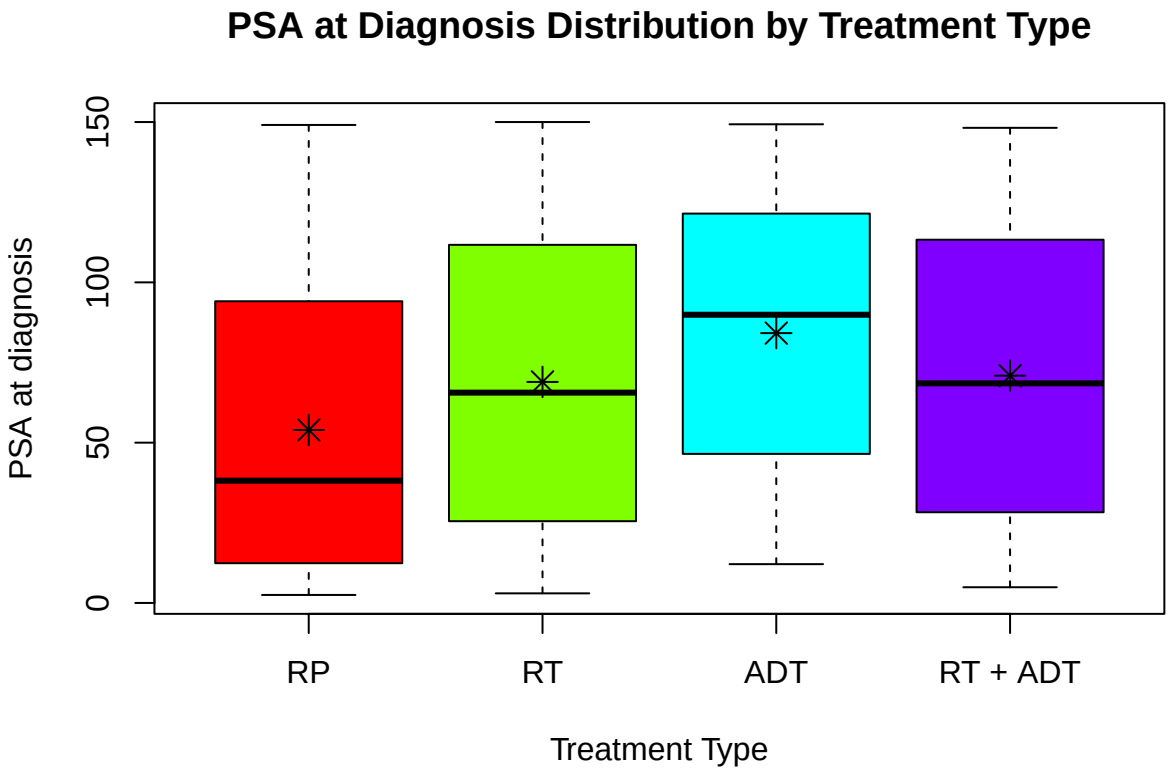
Group 1	Group 2	Mean Diff.	p-value	Sig. ($p < 0.05$)
Radical Prostatectomy	Radiotherapy (RT)	14.598	0.0056	Yes
Radical Prostatectomy	Hormone Monotherapy	29.754	0.0000	Yes
Radical Prostatectomy	Combination (RT + Hormone)	16.523	0.0151	Yes
Radiotherapy (RT)	Hormone Monotherapy	15.157	0.0821	No
Radiotherapy (RT)	Combination (RT + Hormone)	1.925	0.9874	No

Hormone Monotherapy	Combination (RT + Hormone)	-13.232	0.2536	No

The ANOVA test yielded a p-value of 1.6e-6 (0.0000016). Since the p-value is much less than the significance level (alpha (a) =0.05), we reject the null hypothesis (H0). Thus, there is a statistically significant difference in the mean PSA at diagnosis (PSA_before) among the different treatment groups.

Since the ANOVA test confirmed a significant difference, a Tukey HSD post hoc test was performed to identify which specific pairs of groups differ significantly.

The mean PSA at diagnosis for the Radical Prostatectomy (Treatment 1) group is significantly lower than that for all other treatment groups (Treatment 2, 3, and 4). There is no significant difference in the mean PSA at diagnosis among the three non-surgical groups (Treatment 2, 3, and 4) in their pairwise comparisons. The comparison between Hormone Therapy Monotherapy (Treatment 3) and Radical Prostatectomy (Treatment 1) showed the most considerable mean difference (diff = 30.21). This indicates that the Treatment 3 group had the highest average PSA at diagnosis, whereas the Treatment 1 group had the lowest.

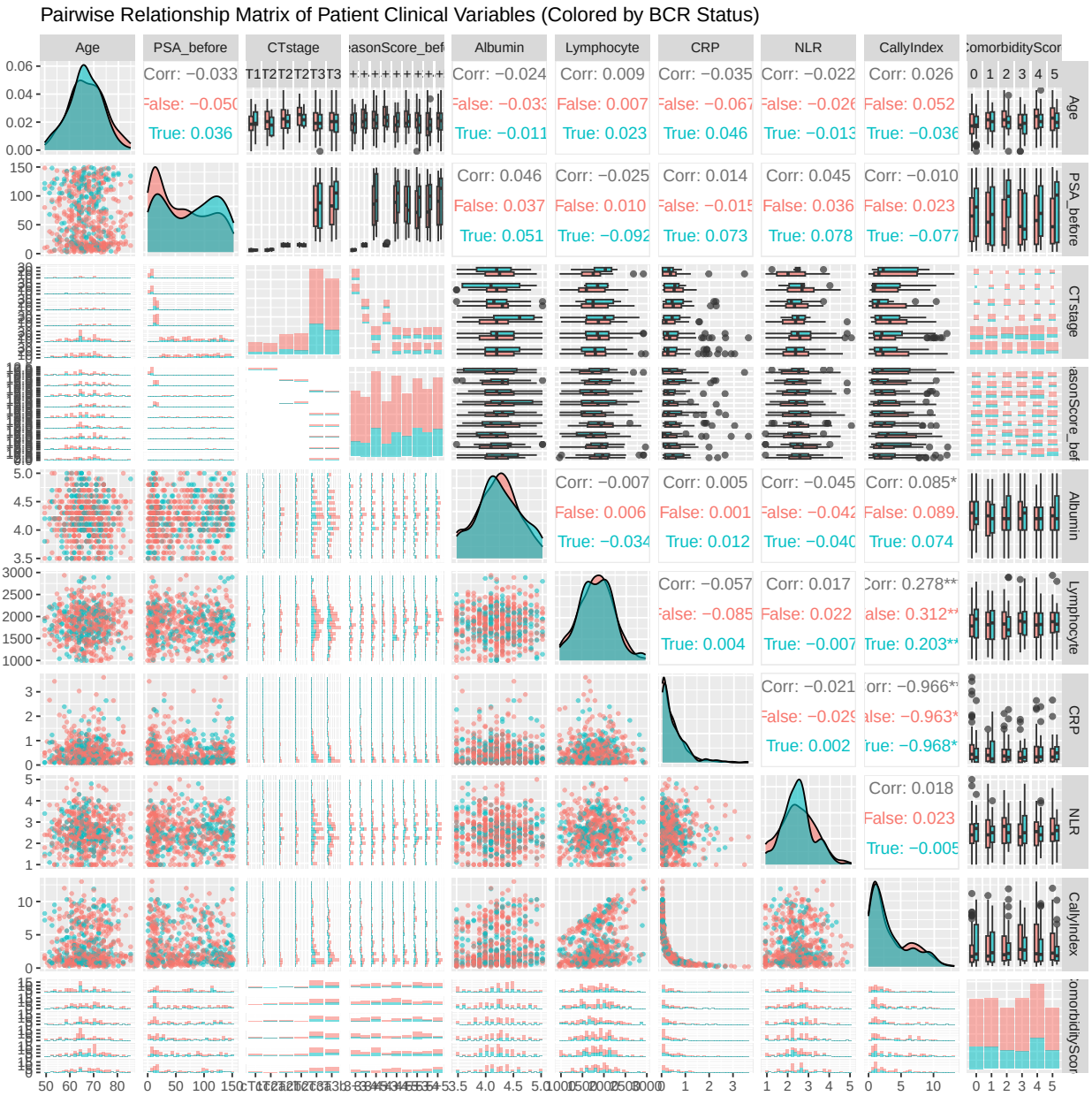


The box plot visually confirms these findings: the PSA distribution for the Radical Prostatectomy (RP) group is noticeably lower than that of all other groups, and its mean value (often marked by a symbol) is the lowest. Conversely, the distribution and mean value for the Hormone Monotherapy (ADT) group are distinctly higher than those for the RP group, visually confirming that patients selected for RP had the lowest initial PSA levels at diagnosis. This showcases how treatment choice is influenced by intial disease severity, and the best treatemnt can be determined through multivaraite models.

PART B) MODELING

Data Analysis 0: Correlation between Variables

A Scatterplot Matrix was generated to visually confirm the correlation between clincal and biological variables. This visualization shows the relationship between each pair of variables and the distribution of each variable. Which will detect potential multicollinearity between variables before fitting the regression models.

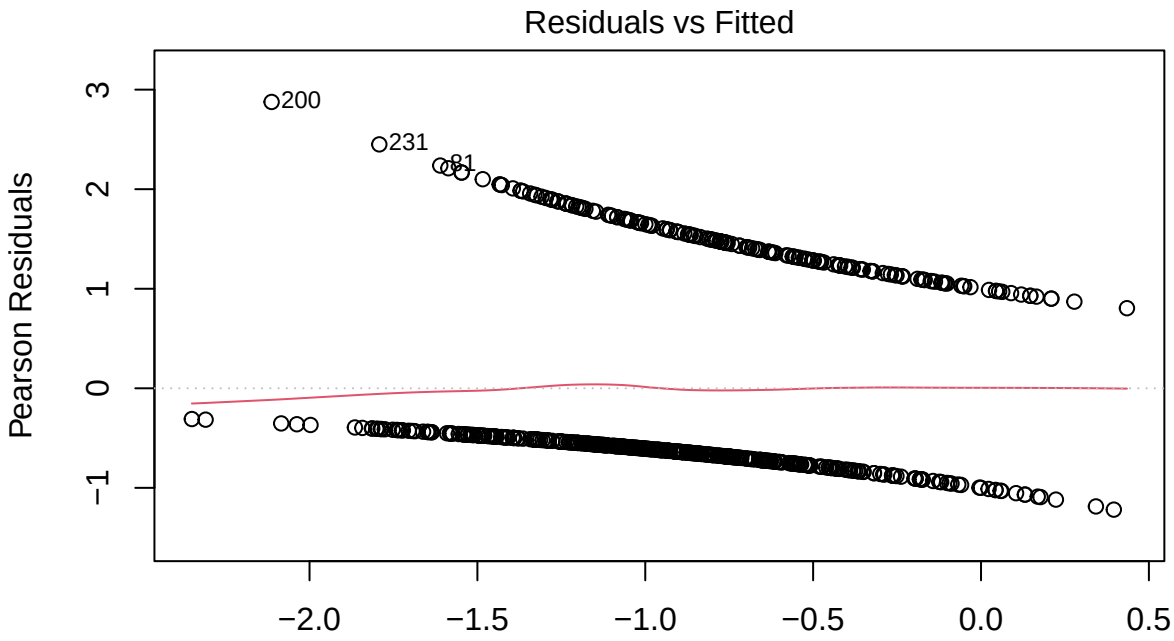


Data Analysis 1: Logistic Regression (BCR)

To evaluate which treatment type appears best for preventing biochemical recurrence, a logistic regression model was constructed with BCR as the dependent variable and clinical and biological variables as independent variables. BCR is a binary outcome (recurrence vs. no recurrence), logistic regression is an appropriate

method for comparing the odds of recurrence across treatment groups while controlling for baseline disease severity and patient health factors. The full multivariate model included treatment type, baseline PSA, clinical stage, biopsy Gleason score, age, inflammatory markers, and comorbidity score.

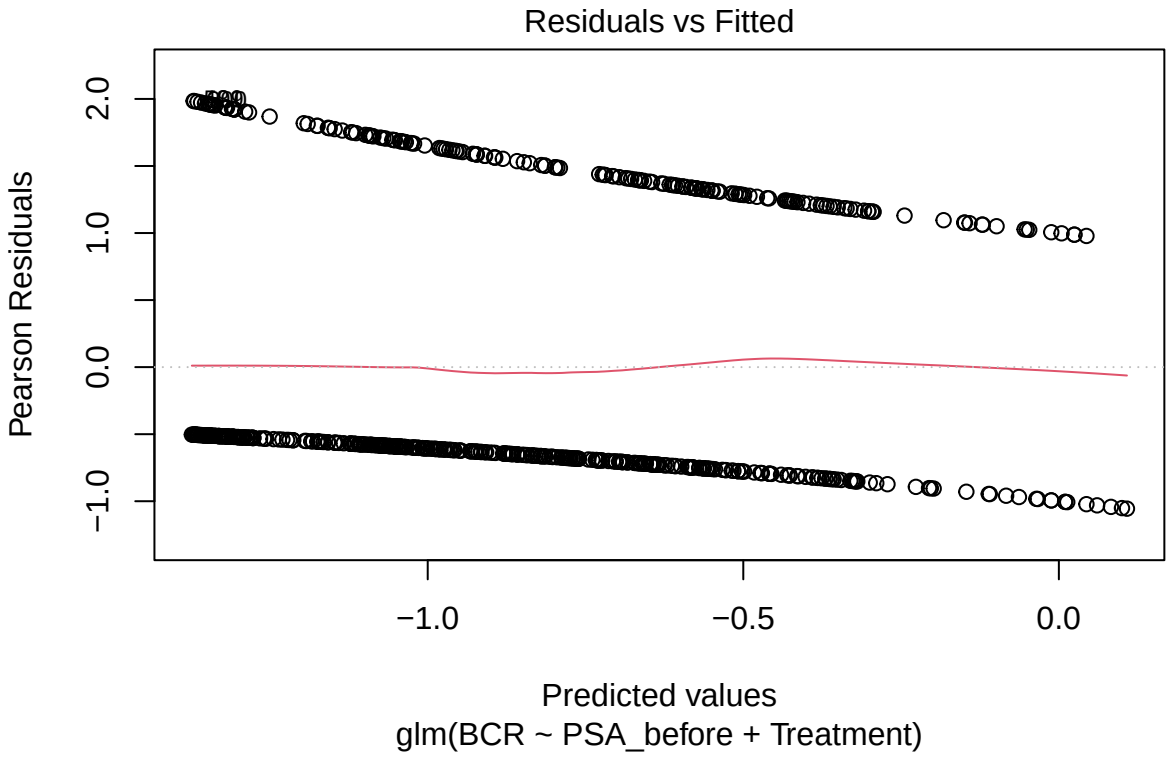
```
##
## Call:
## glm(formula = BCR ~ Age + PSA_before + CTstage + GleasonScore_before +
##     Albumin + Lymphocyte + CRP + NLR + CallyIndex + ComorbidityScore +
##     Treatment, family = binomial, data = data, na.action = na.omit)
##
## Coefficients: (2 not defined because of singularities)
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.081e+00  1.542e+00  -0.701   0.4833
## Age                    -2.019e-02  1.311e-02  -1.541   0.1234
## PSA_before              5.780e-03  2.794e-03   2.069   0.0386 *
## CTstagecT2a             2.026e-01  6.178e-01   0.328   0.7430
## CTstagecT2b             3.239e-01  6.244e-01   0.519   0.6040
## CTstagecT2c            -2.387e-01  5.881e-01  -0.406   0.6848
## CTstagecT3a             9.524e-02  5.697e-01   0.167   0.8672
## CTstagecT3b            -1.203e-01  5.755e-01  -0.209   0.8345
## GleasonScore_before3+4 -1.374e-01  4.881e-01  -0.282   0.7783
## GleasonScore_before3+5 -9.053e-02  3.498e-01  -0.259   0.7958
## GleasonScore_before4+3         NA         NA      NA       NA
## GleasonScore_before4+4 -2.344e-01  3.614e-01  -0.648   0.5167
## GleasonScore_before4+5 -6.841e-02  3.788e-01  -0.181   0.8567
## GleasonScore_before5+3  2.789e-01  3.626e-01   0.769   0.4417
## GleasonScore_before5+4  9.191e-02  3.656e-01   0.251   0.8015
## GleasonScore_before5+5         NA         NA      NA       NA
## Albumin                 2.257e-01  2.477e-01   0.911   0.3621
## Lymphocyte              1.387e-04  2.624e-04   0.529   0.5970
## CRP                    -1.539e-01  2.462e-01  -0.625   0.5319
## NLR                     8.682e-05  1.173e-01   0.001   0.9994
## CallyIndex             -3.165e-02  4.543e-02  -0.697   0.4860
## ComorbidityScore1       6.865e-02  3.187e-01   0.215   0.8294
## ComorbidityScore2      -5.122e-02  3.308e-01  -0.155   0.8770
## ComorbidityScore3      -3.583e-01  3.268e-01  -1.096   0.2729
## ComorbidityScore4       2.711e-01  2.973e-01   0.912   0.3619
## ComorbidityScore5      -2.400e-03  3.301e-01  -0.007   0.9942
## Treatment2              2.849e-01  2.276e-01   1.252   0.2107
## Treatment3              6.202e-01  2.975e-01   2.084   0.0371 *
## Treatment4              2.014e-01  2.818e-01   0.715   0.4747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 739.69  on 599  degrees of freedom
## Residual deviance: 711.93  on 573  degrees of freedom
## AIC: 765.93
##
## Number of Fisher Scoring iterations: 4
```



The multivariate model reveals that PSA_before is statistically significant, suggesting that higher PSA at diagnosis is associated with increased odds of biochemical recurrence. This aligns with earlier visual patterns showing PSA as a key indicator of disease burden and risk. The model also shows that Treatment 3 (hormone therapy) is statistically significant compared with Treatment 1 (surgery as the reference group). This indicates that, after controlling for other predictors, patients receiving hormone therapy demonstrate higher recurrence risk relative to surgery. The coefficients for GleasonScore_before 4+3 and 5+5 were undefined due to singularities.

```
##
## Call:
## glm(formula = BCR ~ PSA_before + Treatment, family = binomial,
##     data = data, na.action = na.omit)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.387813   0.186466  -7.443 9.86e-14 ***
## PSA_before   0.005629   0.001964   2.867  0.00415 **
## Treatment2   0.254127   0.216914   1.172  0.24137
## Treatment3   0.655230   0.285619   2.294  0.02179 *
## Treatment4   0.185449   0.269549   0.688  0.49145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 739.69  on 599  degrees of freedom
## Residual deviance: 722.69  on 595  degrees of freedom
```

```
## AIC: 732.69
##
## Number of Fisher Scoring iterations: 4
```



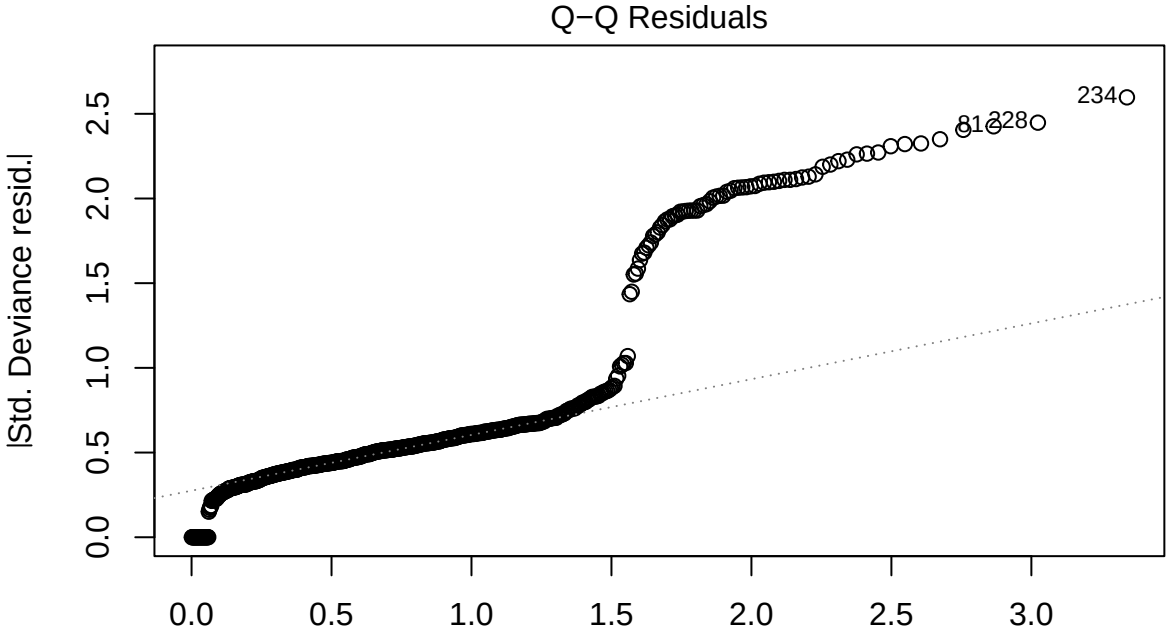
The Residuals vs Fitted plot does not show the ideal random scatter around zero and instead shows two distinct linear patterns. While this pattern can occur in logistic regression diagnostics, it suggests the full model may be over-parameterized for the amount of signal in the data. This supports simplifying the model for clearer inference and better efficiency.

Data Analysis 2: Logistic Regression (Survival)

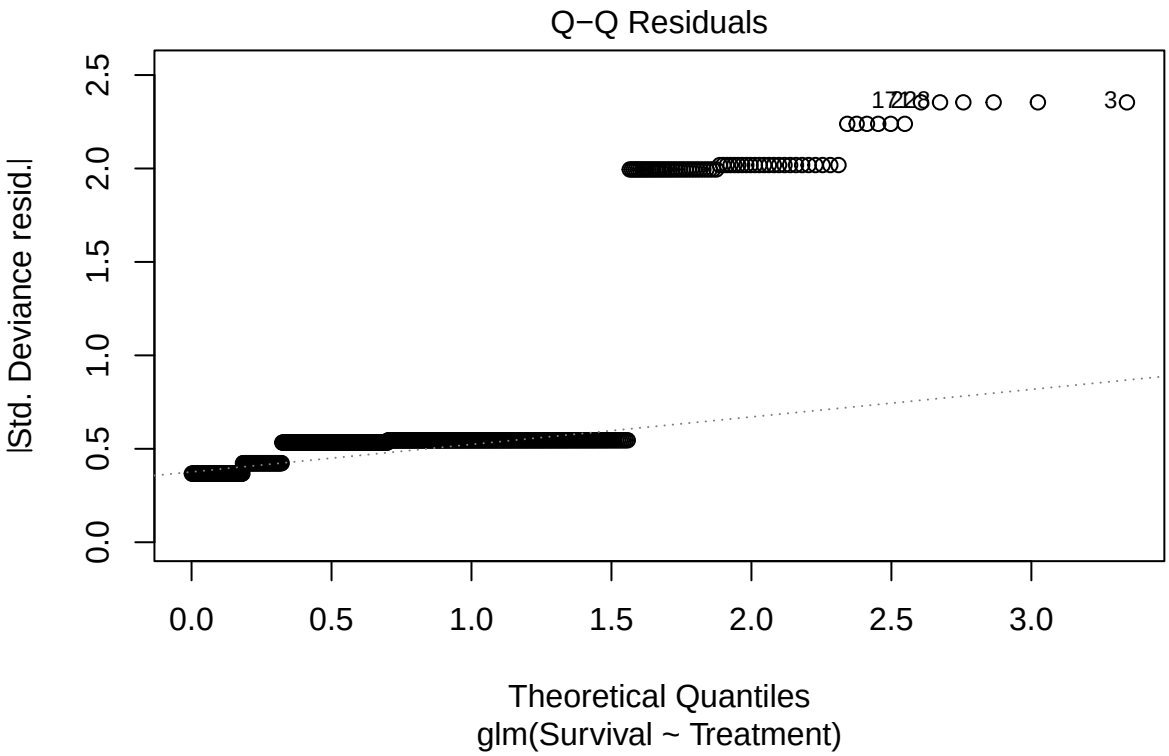
A logistic regression approach was used to evaluate survival outcomes across treatment types while accounting for clinical covariates.

```
##
## Call:
## glm(formula = Survival ~ Age + PSA_before + CTstage + GleasonScore_before +
##     Albumin + Lymphocyte + CRP + NLR + CallyIndex + ComorbidityScore +
##     Treatment, family = binomial, data = data, na.action = na.omit)
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.137e+00  2.305e+00   0.927   0.3539
## Age                    -1.432e-02  1.926e-02  -0.744   0.4571
## PSA_before             -5.326e-03  4.009e-03  -1.328   0.1840
```

```
## CTstagecT2a             1.501e+01  7.183e+02   0.021   0.9833
## CTstagecT2b            -5.715e-01  9.402e-01  -0.608   0.5433
## CTstagecT2c            -6.360e-01  8.983e-01  -0.708   0.4790
## CTstagecT3a            -5.481e-01  9.238e-01  -0.593   0.5529
## CTstagecT3b            -1.775e-01  9.300e-01  -0.191   0.8486
## GleasonScore_before3+4 -4.216e-01  6.089e-01  -0.692   0.4887
## GleasonScore_before3+5  1.017e-01  5.329e-01   0.191   0.8486
## GleasonScore_before4+3         NA         NA      NA       NA
## GleasonScore_before4+4 -3.853e-01  5.071e-01  -0.760   0.4474
## GleasonScore_before4+5  2.886e-01  6.112e-01   0.472   0.6368
## GleasonScore_before5+3 -2.008e-01  5.348e-01  -0.375   0.7074
## GleasonScore_before5+4 -5.629e-01  5.175e-01  -1.088   0.2767
## GleasonScore_before5+5         NA         NA      NA       NA
## Albumin                 1.316e-01  3.595e-01   0.366   0.7144
## Lymphocyte              6.403e-04  3.878e-04   1.651   0.0987 .
## CRP                     4.782e-01  4.251e-01   1.125   0.2606
## NLR                    -1.478e-01  1.679e-01  -0.880   0.3788
## CallyIndex              2.322e-02  6.964e-02   0.333   0.7389
## ComorbidityScore1      -2.827e-01  4.637e-01  -0.610   0.5421
## ComorbidityScore2      -4.300e-01  4.675e-01  -0.920   0.3577
## ComorbidityScore3       8.570e-02  4.905e-01   0.175   0.8613
## ComorbidityScore4      -1.129e-01  4.612e-01  -0.245   0.8066
## ComorbidityScore5      -4.427e-01  4.752e-01  -0.932   0.3516
## Treatment2              1.879e-01  3.009e-01   0.624   0.5324
## Treatment3              8.841e-01  4.861e-01   1.819   0.0690 .
## Treatment4              1.040e+00  4.756e-01   2.188   0.0287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 436.31  on 599  degrees of freedom
## Residual deviance: 403.52  on 573  degrees of freedom
## AIC: 457.52
##
## Number of Fisher Scoring iterations: 16
```




```
##
## Call:
## glm(formula = Survival ~ Treatment, family = binomial, data = data,
##     na.action = na.omit)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.83372    0.18204  10.073   <2e-16 ***
## Treatment2   0.05082    0.28484   0.178   0.8584
## Treatment3   0.54890    0.46388   1.183   0.2367
## Treatment4   0.84042    0.45967   1.828   0.0675 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 436.31  on 599  degrees of freedom
## Residual deviance: 431.23  on 596  degrees of freedom
## AIC: 439.23
##
## Number of Fisher Scoring iterations: 5
```



We performed a similar logistic regression analysis to predict patient survival. Although the initial model suggested potential significance for Lymphocyte count and Treatment 4, the diagnostic plots reveal substantial issues with the model fit. The Q-Q plot of the residuals shows a stepped pattern that deviates significantly from the theoretical diagonal line. This indicates that the residuals do not follow the expected distribution. Consequently, this specific model appears unreliable for drawing conclusions about survival factors in this dataset.

Model Interpretation 1: Logistic Regression (BCR ~ Treatment)

```
##
## Call:
## glm(formula = BCR ~ PSA_before + Treatment, family = binomial,
##     data = data, na.action = na.omit)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.732583   0.293763  -2.494  0.01264 *
## PSA_before   0.005629   0.001964   2.867  0.00415 **
## Treatment1  -0.655230   0.285619  -2.294  0.02179 *
## Treatment2  -0.401103   0.290011  -1.383  0.16665
## Treatment4  -0.469781   0.330558  -1.421  0.15526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 739.69  on 599  degrees of freedom
## Residual deviance: 722.69  on 595  degrees of freedom
## AIC: 732.69
##
## Number of Fisher Scoring iterations: 4
```

Given the significance of Treatment 3 in our previous BCR model, we re-leveled the Treatment factor to set Treatment 3 as the reference group. This allows for a direct comparison against the other treatment types. The analysis confirms a significant difference between Treatment 1 and Treatment 3. Calculating the odds ratio from the model coefficients reveals that patients receiving Treatment 3 have approximately 1.93 times higher odds of experiencing biochemical recurrence compared to those receiving Treatment 1, holding PSA levels constant

Model Interpretation 2: Logistic Regression (BCR ~ Combination)

```
##
## Call:
## glm(formula = BCR ~ PSA_before + Combination, family = binomial,
##     data = data, na.action = na.omit)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.250651   0.163902  -7.630 2.34e-14 ***
## PSA_before      0.006496   0.001914   3.394  0.00069 ***
## CombinationYes -0.017033   0.246312  -0.069  0.94487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 739.69  on 599  degrees of freedom
## Residual deviance: 728.04  on 597  degrees of freedom
## AIC: 734.04
##
## Number of Fisher Scoring iterations: 4
```

We also investigated whether using a combination therapy (Treatment 4) provides a distinct advantage regarding BCR risk. By creating a binary variable to compare the combination therapy against all other treatments, we found that the coefficient for the combination group is not statistically significant ($p = 0.94$). This result implies that, in this dataset, simply combining treatments does not inherently result in a better outcome regarding biochemical recurrence compared to single-mode therapies.

Model Evaluation:

Model Fitting and Predicting

Evaluating: AUC (Area Under the Curve)

```
## Setting levels: control = False, case = True
```

```
## Setting direction: controls < cases
```

```
## AUC (Area Under the Curve): 0.5685
```

Confusion Matrix (Threshold 0.5)

Table 9: Confusion Matrix (Threshold 0.5)

	Actual Status	
Predicted	False	True
False	72	26
True	53	29

Table 10: Performance Metrics (Threshold 0.5)

Metric	Value
Accuracy	0.5611
Sensitivity (Recall)	0.5273
Specificity	0.5760

Finally, we evaluated the predictive performance of the reduced BCR model. The data was split into a 70% training set and a 30% testing set, with oversampling applied to the training data to handle class imbalance. Upon testing the model on the unseen data, the Area Under the Curve (AUC) was calculated to be 0.5685, indicating that the model's predictive capability is only slightly better than random chance. The confusion matrix yields an overall accuracy of 56.11%. With a sensitivity of 52.73% and a specificity of 57.60%, the model struggles to reliably classify positive BCR cases, suggesting that while PSA and Treatment are significant risk factors, they are not sufficient on their own to form a high-precision predictive tool.

5) Conclusion:

This project compared four primary prostate cancer treatment, Surgery(1), Radiotherapy(2), Hormone Therapy(3), and Combination Therapy (Hormone +Radiation). The main focus was to determine which treatment type appears the best across post-treatment outcomes. Our analyses focused on biochemical recurrence, PSA-based response patterns, and survival-related modeling, supported by both visual and inferential methods. The most consistent finding across tests and modeling is that treatment type is statistically associated with biochemical recurrence. The significant Chi-squared test supports the visual differences in recurrence rates across treatments, and the logistic regression confirms that these differences persist even after accounting for clinical severity. In particular, higher PSA at diagnosis increases recurrence risk, and Treatment 3 is associated with higher recurrence odds relative to Treatment 1. The reduced BCR model provides a more efficient explanation of recurrence patterns than the full model, supported by the substantial AIC improvement (765.93 $\rightarrow$ 732.69) and cleaner residual behavior. In contrast, survival differences are less stable in this dataset. While exploratory results suggested that Treatment 4 and Lymphocyte count may relate to survival, diagnostic issues—particularly the stepped deviation in the Q-Q plot—indicate that the survival model is not reliable enough to support strong treatment conclusions. This outcome emphasizes the importance of diagnostic checks and suggests that survival may require either additional predictors, a larger follow-up window, or a larger sample to model effectively. The combination-therapy comparison for BCR suggests that Treatment 4 does not show an inherent recurrence advantage over other treatments in this dataset ($p = 0.94$). This finding may reflect treatment selection patterns and baseline differences across groups, supported by the ANOVA result showing that baseline PSA significantly differs by treatment type.Based on recurrence-focused evidence, surgery (Treatment 1) appears to be the most favorable option for minimizing BCR in this dataset, while hormone monotherapy (Treatment 3) appears less favorable for recurrence prevention. However, because baseline PSA differs across treatment groups, as well as hormone therapy is more commonly asisgned to pateints with worse risks. Overall, these conclusions are not definitive in regards to overall best treatment for patients but it gives insight that surgery is likely the most beneficial to reduce likelihood of cancer returning to a patient .

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