Survival Compass*

Statistical Insights into Lung Cancer Patients Journey Post Diagnosis

Lexi Knight

April 15, 2024

This study investigates the impact of pathogenic stage and treatment modalities on lung cancer survival post-diagnosis. Analysis of patient data reveals significant correlation between pathogenic stage, seeking treatment and survival outcomes. Notably, patients at advanced stages with metastases in distant sites beyond the lung, extensive lymph node involvement and tumors with extensive growth, invading nearby structures demonstrate lower survival rates. These findings underscore the critical importance of early detection, tailored treatment strategies and ongoing research efforts to enhance lung cancer survival rates globally.

Table of contents

*Code and data are available at: https://github.com/LexiKnight/Lung_Cancer/tree/main

1 Introduction

Clinging to life amidst the shadows of lung cancer, where every breath becomes a battleground. Survival becomes not just a statistic but an interplay between several individual characteristics. We explore the hidden keys to defying the odds and emerging victorious against one of the deadliest adversaries of our time. Lung cancer is the leading cause of cancer-related deaths in the world (Park, 2017). It is a disease that develops in the lining of the airways in lung tissues. Non-small cell lung cancer (NSCLC) is the most common type, accounting for 80-85% of all lung cancers according to the American Cancer society (Markman, 2023). Staging is important for prognosis and making treatment decisions. Common treatments include surgery, radiation therapy and chemotherapy (Kai, 2021). Pathogenic stage is determined by presence of nearby metastasis, lymph node involvement as well as tumor spread and size (Markman, 2023). This paper investigates the relationship between lung cancer patients' survival and pathogenic stage. The estimand is the median survival time in days post-diagnosis. We also look at whether patients decided to have treatment and if so, which method; radiation therapy or chemotherapy. Through analysis of a dataset made up of 981 patients in Sydney, Australia, we offer insight into the prognostic markers.

Tumor size is often the main determinant of stage and treatment. As tumor categories increase, the tumor expands, invading nearby structures (Zhang, 2015). A study involving 52,287 patients diagnosed between the years 1998 and 2003 found tumor size to be an independent

prognostic factor in estimating overall survival. The authors found that patients presenting with larger tumors predicted a worse prognosis and thus are associated with a decrease in survival. There is a similar relationship between extensive lymph node involvement and patient survival (Zhang, 2015). Initial spread of cancer cells are localized, then become regional, involving nearby lymph nodes and the most severe cases comprises expansion to other organs such as the brain, liver and bones (Markman, 2023). A study looked at five year survival rates based on the severity of spread. 62.8% of patients with localized spread, 34.8% of patients with regional and 8% of patients with distant, advanced spread were found to survive for 5 years post diagnosis. More than half of these lung cancer patients have advanced spread to other organs when diagnosed(Markman, 2023). Overall, it is found that patients with no regional lymph node metastases, and smaller tumors are easier to be treated and thus are associated with improved survival rates (Zhang, 2015).

Presence of metastatic LN is one of the most important determinants of prognosis of NSCLC cases (Kai, 2021). In the early stage, cancer has not spread to lymph nodes. As severity increases, lymph node metastasis sequentially spreads to more distant lymph nodes such as mediastinal and there is severe lymph node involvement (Park, 2017). Lymph node involvement, also termed lymph node ratio, is a crucial factor in guiding treatment options (Kai, 2021). A study made up of 97 patients with a mean age of 63 who have undergone surgery between the years 2009 and 2015 in Korea find that increased lymph node involvement is associated with a more advanced disease status and hence affiliated with prognosis (Park, 2017). Another study looked at 11,341 NSCLC patients between the years 2004 to 2015, from 18 geographically diverse populations, covering approximately 28% of the population of the United States. These patients were treatment naive and underwent surgical resection of the tumor. Although 5757 patients died, the rest showed great results, with a median survival of 22 months (Kai, 2021). The authors found that patients with low lymph node involvement lead to higher survival compared to patients with high lymph node ratios. A regression analysis revealed that lymph node ratio is an independent and significant predictor of patient survival. The authors also observed that disease burden and anatomical location of the lymph nodes involved may influence the patients survival (Kai, 2021).

After tumor size, LN involvement and presence of distant metastasis are categorized, the pathogenic stage of the cancer is then determined (Eldridge, 2022). The most valuable prognostic factor in non-small cell lung cancer is the pathogenic stage (Park, 2017) .Stage is determined by tumor size, number of tumors and where the cancer has spread. Stage 1 is localized spread, stage 2 and 3 is regional spread while stage 4 is distant spread of the tumor (Eldridge, 2022). Cancer stage was determined using the seventh American Joint Committee on Cancer staging system (AJCC) (Park, 2017). A study done in Australia including 2119 lung cancer patients illustrated those with stage IV disease, the most advanced stage, showed shorter survival than those at lower stages (Denton, 2016). The earlier the cancer is found, that is the lower the pathogenic stage, the greater the likelihood curative radiation therapy is an effective treatment (Eldridge, 2022). However, there is minimal literature looking at post-diagnosis survival rates based on pathogenic stage and method of treatment. The extent of this disease illustrates the importance of living a healthy lifestyle, undergoing regular screening and

development of improved treatment methods. Over the past decade, there has been great improvement of lymph node assessment in cancer patients (Kai, 2021). Experts hope survival rates continue to improve with new therapies and treatment approaches (Markman, 2023).

The remainder of this paper is structured as follows. In Section 2, we visualize the exploration of variables constituting the pathogenic stage and treatment types. Section 3, outlines the model employed to analyze the relationship between these variables and the duration of survival post-diagnosis. Moreover, Section 4 offers visual depictions of the study's outcomes. Finally, in Section 5, we summarize the primary findings, propose avenues for enhancement, and identify potential areas for future research.

2 Data

2.1 Software and R-packages

This project was created using statistical software, `R` (R Core Team 2023). For data cleaning and manipulation, we used `tidyverse` (**tidy?**) package. For data visualization such as creating the figures, we used `ggplot2` (**ggplot2?**). For converting into Parquet file, we used `arrow` (**arrow?**), managing file paths `here` (**here?**), statistical modeling `rstanarm` (Goodrich et al. 2022). Furthermore, `showtext` (**showtext?**) was used for aesthetic purposes.

2.2 Methodology

The data for this study were collected from a comprehensive database comprising 981 lung cancer patients diagnosed between 1991 to 2013 from `Center for Open Science` (**cfos?**), a dataset acquired in Sydney Australia. The dataset included information on patient demographics, clinical characteristics, treatment modalities, and survival outcomes.

2.2.1 Data Collection

We obtained data on lung cancer patients meeting the following criteria: histologically confirmed lung cancer diagnosis, availability of complete clinical data, treatment-naive patients, single malignancy and located in Australia. Patients with missing or incomplete information were excluded from the analysis.

2.2.2 Data Cleaning

After obtaining the dataset, there were a series of rigorous preprocessing steps undertaken to ensure data quality and consistency. This included selecting the columns of interest namely; days to death post diagnosis, presence of distant metastasis, lymph node involvement, pathogenic stage, tumor size and treatment type. Next, we cleaned the column names and excluded missing values. Additionally, we converted the days to death column to numeric.Tests where included to ensure accuracy, reliability and validity of the datset for subsequent analysis and interpretation.

2.2.3 Data Analysis

Descriptive and inferential statistical analyses were conducted to explore the dataset and derive meaningful insights. These included linear regression modeling.

2.3 Features

The dataset comprised several key features relevant to lung cancer prognosis, including pathogenic stage, presence of distant metastasis, lymph node involvement, tumor size, and treatment type.

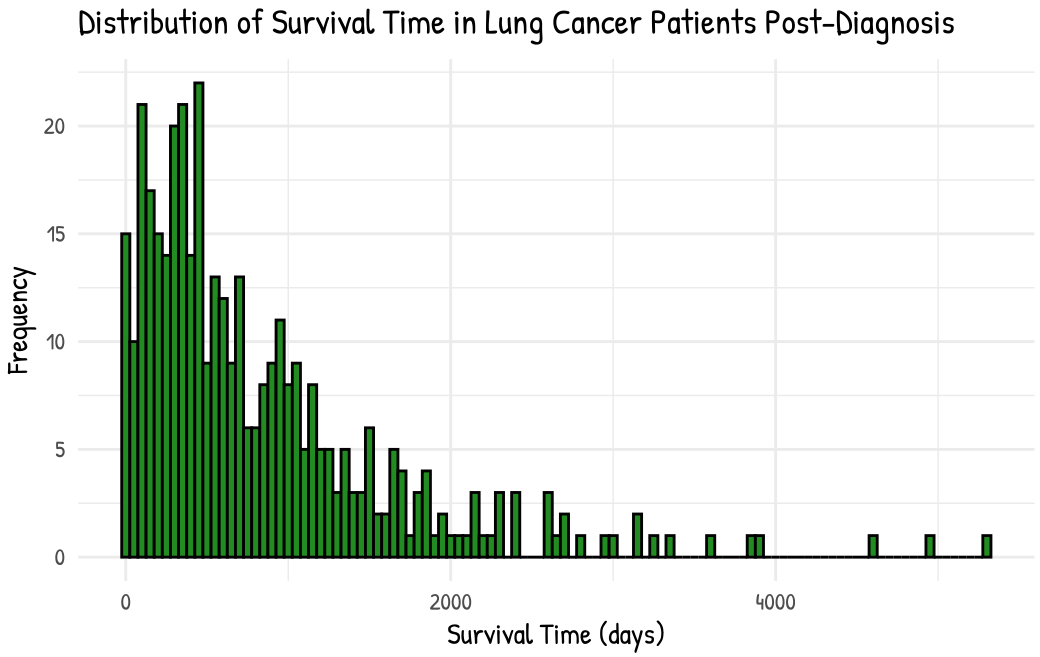


Figure 1: Distribution of survival time in patients

2.3.1 Days to Death

One of the critical features analyzed in this study is the duration between the date of lung cancer diagnosis and the date of death, referred to as "days to death." This metric serves as a key indicator of patient survival and provides valuable insights into the disease trajectory and prognosis. By examining the distribution of survival times among lung cancer patients post-diagnosis, we aim to characterize the temporal patterns of disease progression and assess the impact of various clinical and demographic factors on survival outcomes. Understanding the time course from diagnosis to death is crucial for guiding treatment decisions, predicting patient outcomes, and identifying opportunities for intervention to improve survival rates. Through comprehensive analysis of days to death data, we seek to elucidate the factors influencing patient survival in lung cancer and contribute to the refinement of prognostic models for clinical practice.

2.3.2 Pathogenic Stage

Pathogenic stage, a critical determinant of lung cancer prognosis, was classified according to the [mention staging system used, e.g., TNM classification]. As depicted in Figure 1, the distribution of patients varied across different stages, with [mention any notable trends, e.g., a higher proportion of patients diagnosed at advanced stages].

2.3.3 Presence of Distant Metastasis

The presence of distant metastasis, indicative of disease spread beyond the primary tumor site, significantly influences treatment decisions and patient outcomes. Figure 2 illustrates the percentage of patients with and without distant metastasis, highlighting the impact of metastatic disease on prognosis.

2.3.4 Lymph Node Involvement

Lymph node involvement is a key prognostic factor in lung cancer, reflecting the extent of disease spread to regional lymph nodes. Figure 3 presents the distribution of patients based on lymph node involvement, demonstrating its association with survival outcomes.

2.3.5 Tumor Size

Tumor size, often measured as the diameter of the primary tumor, is closely linked to disease progression and treatment response in lung cancer patients. Figure 4 showcases the distribution of patients across different tumor size categories, elucidating its significance as a prognostic factor.

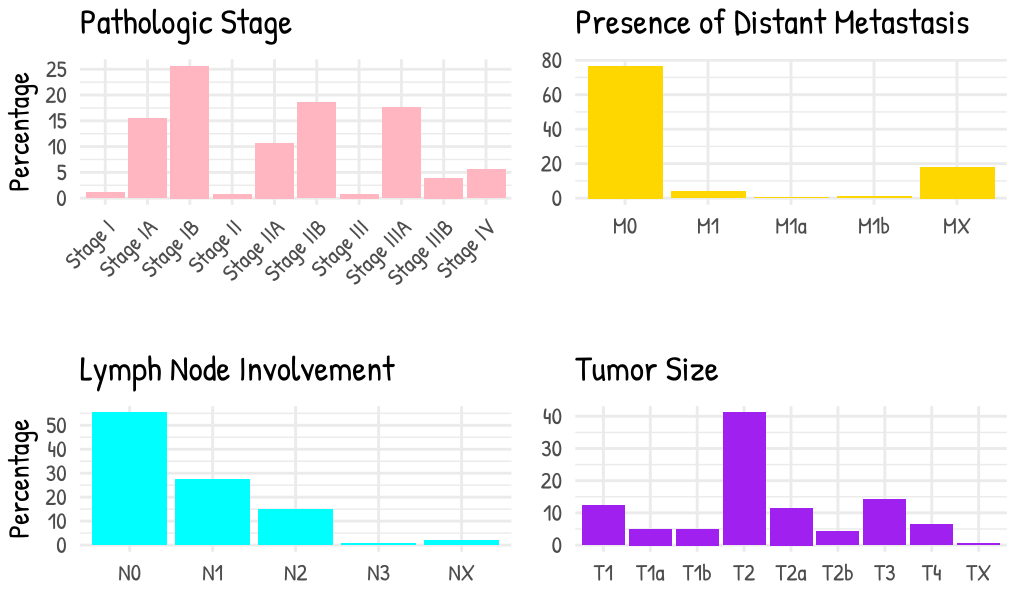


Figure 2: ADD FIGURE 2 CAPTION HERE

3 Model

3.1 Model set-up

In this section, we aim to predict the survival outcomes of lung cancer patients post-diagnosis with a linear regression model framework. We consider several predictors including pathogenic stage, lymph node involvement, presence of distant metastasis, tumor size, and treatment type.We specify the model and subsequently justify its appropriateness for our analysis.

3.1.1 Model Specifications

We employ a linear regression model to predict the number of days from diagnosis to death for each lung cancer patient. The model is defined as follows:

$$y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$

where:

- y_i represents the number of days from diagnosis to death for patient i.
- μ_i denotes the expected number of days to death for patient i.
- σ represents the standard deviation of the survival times.

The linear predictor μ_i is specified as:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{pathologic\_stage}} \times \text{pathologic\_stage}_i \\
&\quad + \beta_{\text{lymph\_node}} \times \text{lymph\_node\_involvement}_i \\
&\quad + \beta_{\text{metastasis}} \times \text{presence\_of\_distant\_metastasis}_i \\
&\quad + \beta_{\text{tumor\_size}} \times \text{tumor\_size}_i \\
&\quad + \beta_{\text{treatment\_type}} \times \text{treatment\_type}_i
\end{aligned}
$$

where:

- α represents the intercept term, capturing the baseline number of days to death.
- β {pathologic_stage}, β {lymph_node}, β {metastasis}, β {tumor_size}, β {treatment_type} are the coefficients associated with each predictor variable.

3.1.2 Model justification

Linear regression models are most appropritate in predicting continuous outcomes. As survival time is continuous, this model allows us to quantify the relationships between these predictors and survival outcomes, providing valuable insights into the factors influencing the prognosis of lung cancer patients.

3.1.2.1 Response Variable

Out variable of interest is survival time in lung cancer patient afte they have been diagnosed

We model the survival time (y_i) as a continuous variable, reflecting the duration from diagnosis to death for each patient. This continuous characterization is appropriate for capturing the temporal aspect of survival outcomes in medical contexts.

3.1.2.2 Input Variables

We consider several clinically relevant predictors including pathologic stage, lymph node involvement, presence of distant metastasis, tumor size, and treatment type. These variables are chosen based on their established associations with lung cancer prognosis, encompassing key aspects of disease severity and treatment strategies.

3.1.2.3 Model Structure

The linear regression model relates the expected survival time (μ_i) to the linear combination of predictor variables, allowing us to quantify the impact of each predictor on the expected duration of survival. This framework facilitates interpretation of the associations between clinical variables and survival outcomes, providing valuable insights for patient prognosis.

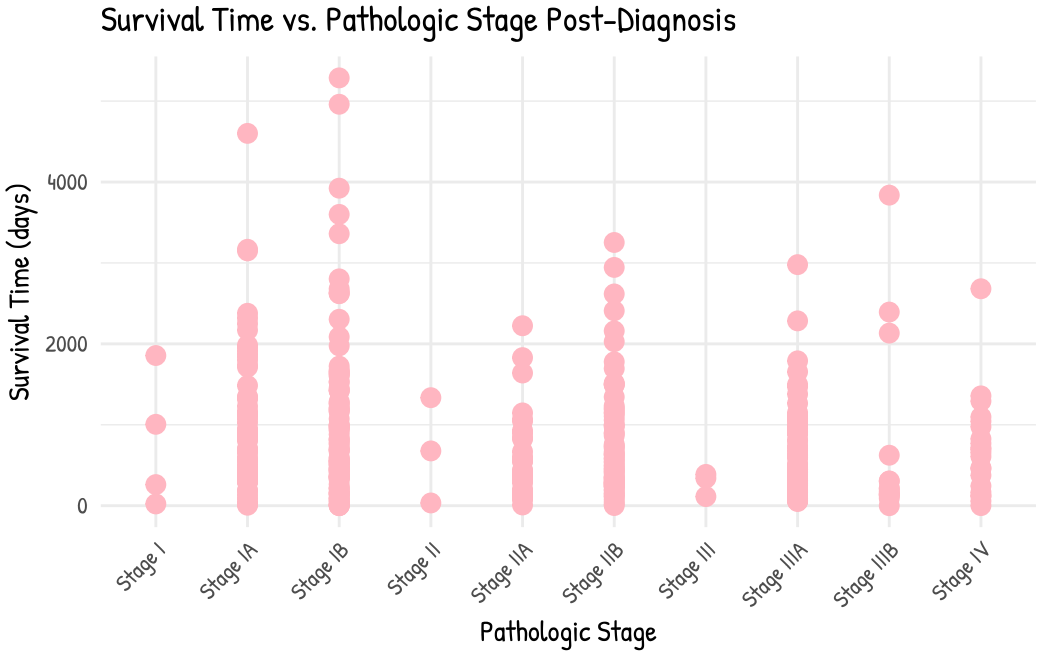
3.1.2.4 Parameter Estimation

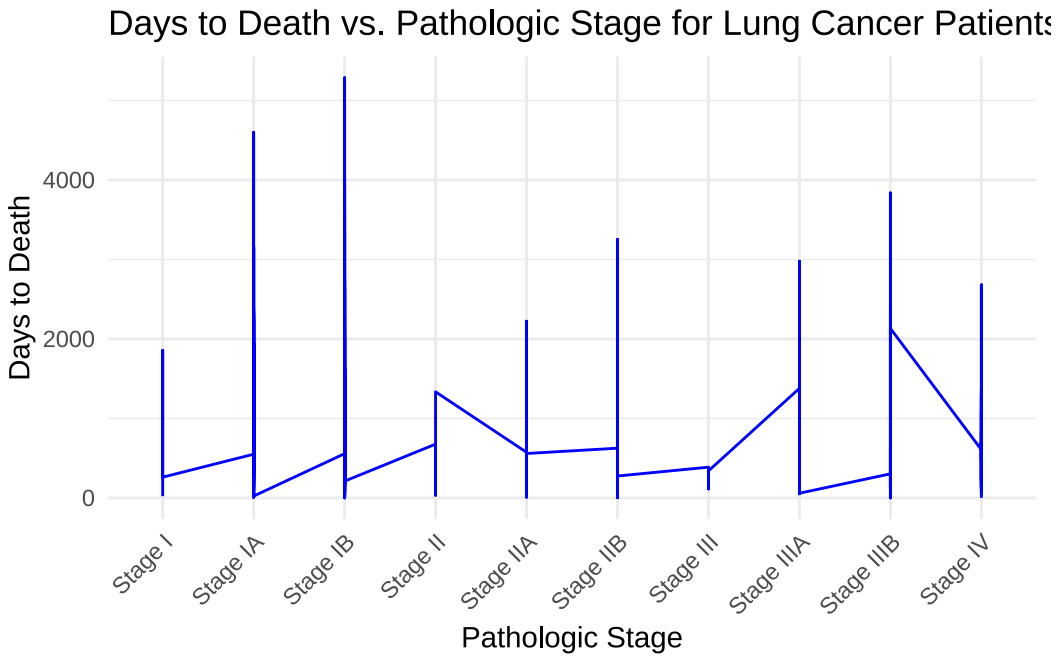
We anticipate that the survival time of lung cancer patients post-diagnosis will be influenced by various clinical factors such as pathologic stage, extent of lymph node involvement, presence of distant metastasis, tumor size, and treatment type. Specifically, we expect that advanced pathologic stages, increased lymph node involvement, presence of distant metastasis, larger tumor sizes, and certain treatment types will be associated with shorter survival times.

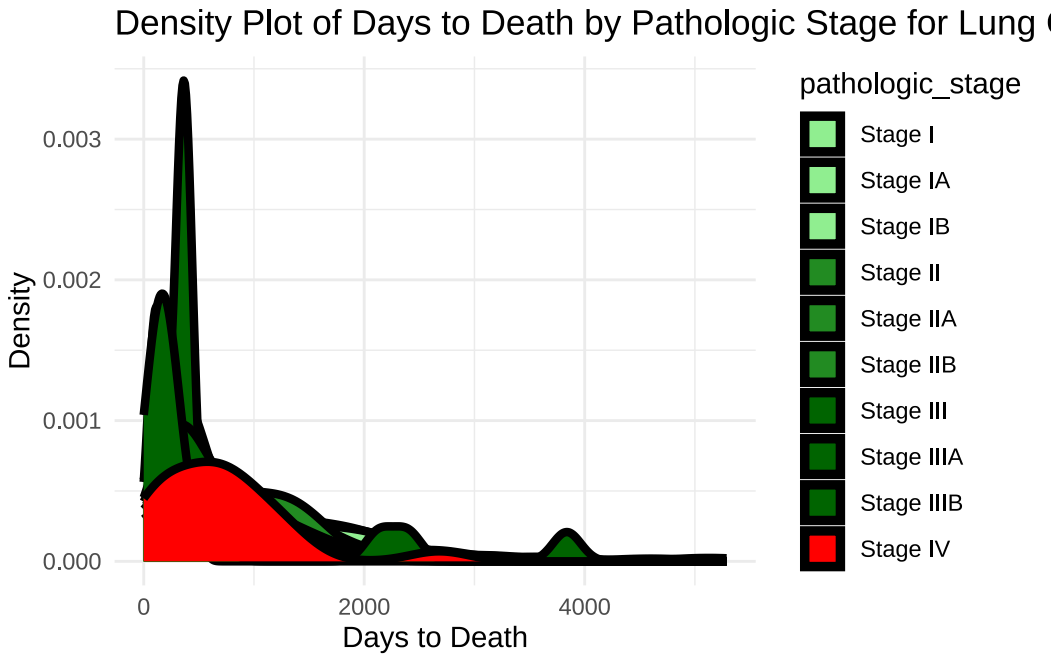
We run the model in R (R Core Team 2023) estimating the model coefficients (α and β) using Bayesian inference via the 'stan_glm()' function from the Goodrich et al. (2022) package. This approach leverages Markov Chain Monte Carlo (MCMC) algorithms to obtain posterior distributions for the model parameters, enabling robust estimation of parameter uncertainties and inference on the effects of predictor variables.

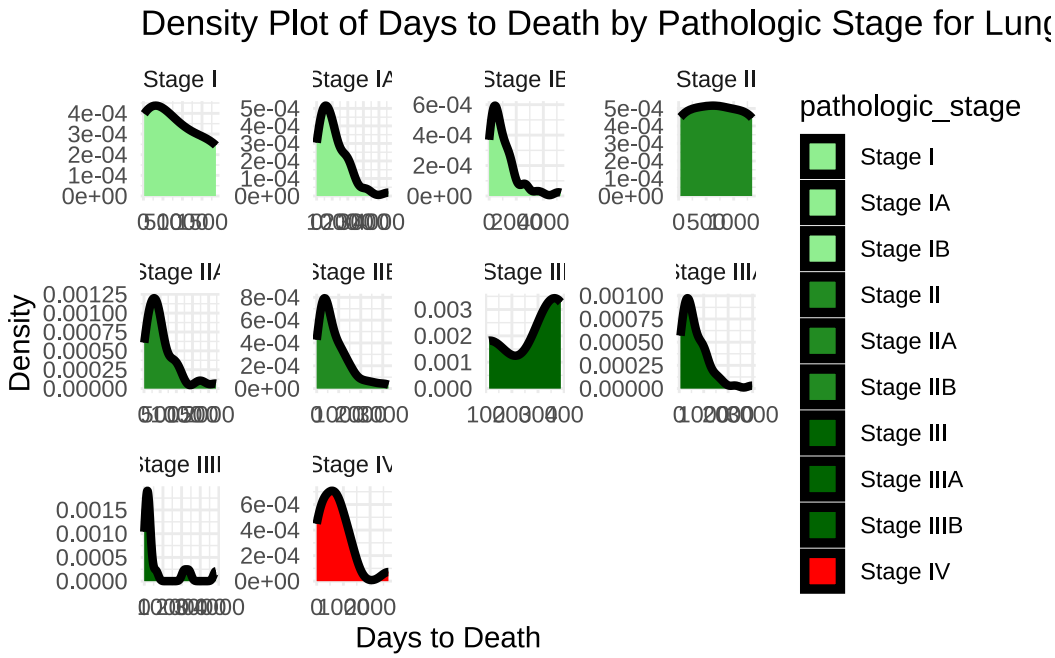
4 Results

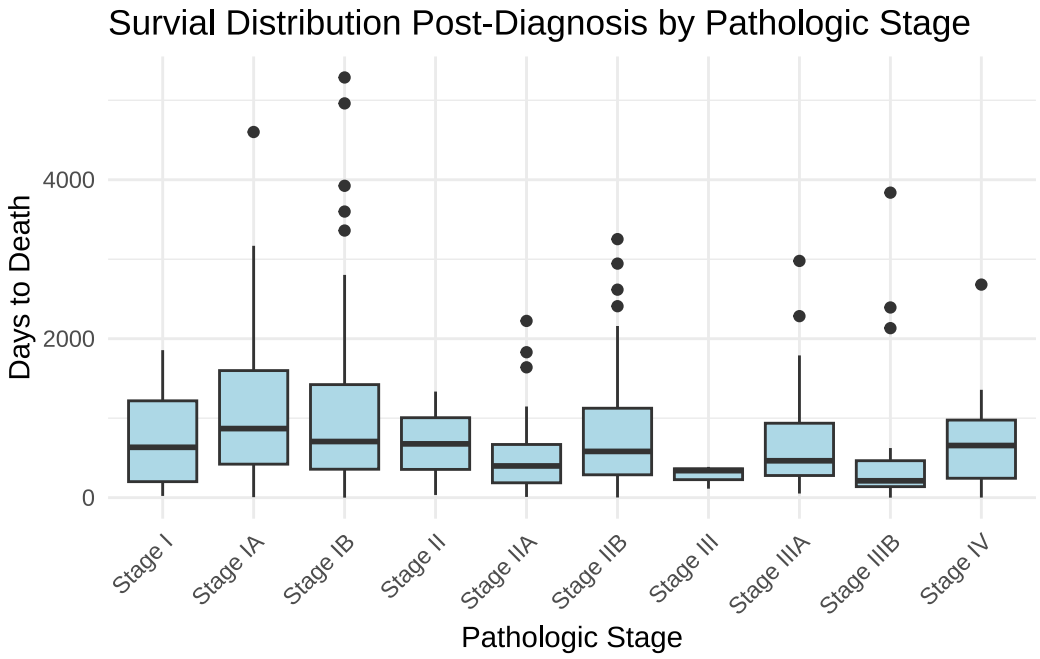
Our results are summarized in @.

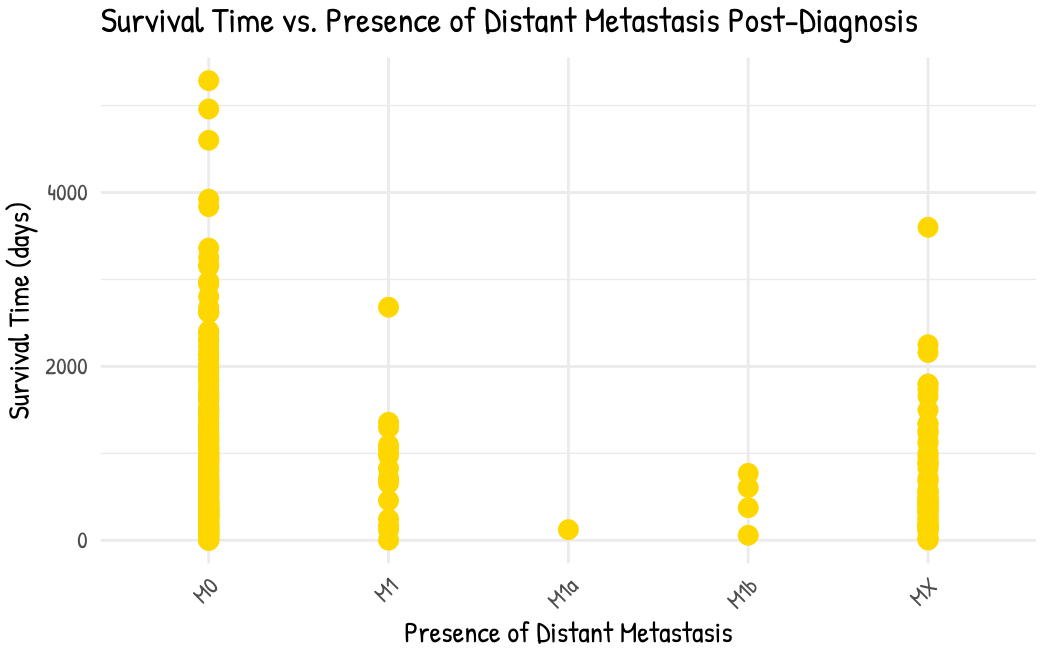


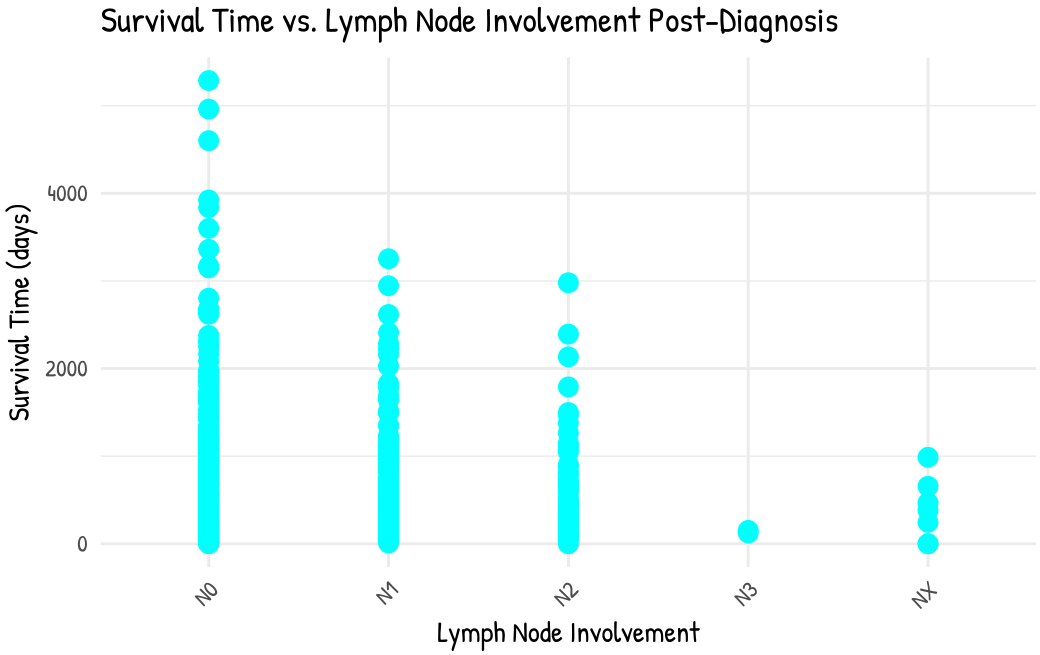

Days to Death vs. Pathologic Stage for Lung Cancer Patients
Days to Death
4000
2000
0
Stage I
Stage IA
Stage IB
Stage II
Stage IIA
Stage IIB
Stage III
Stage IIIA
Stage IIIB
Stage IV
Pathologic Stage

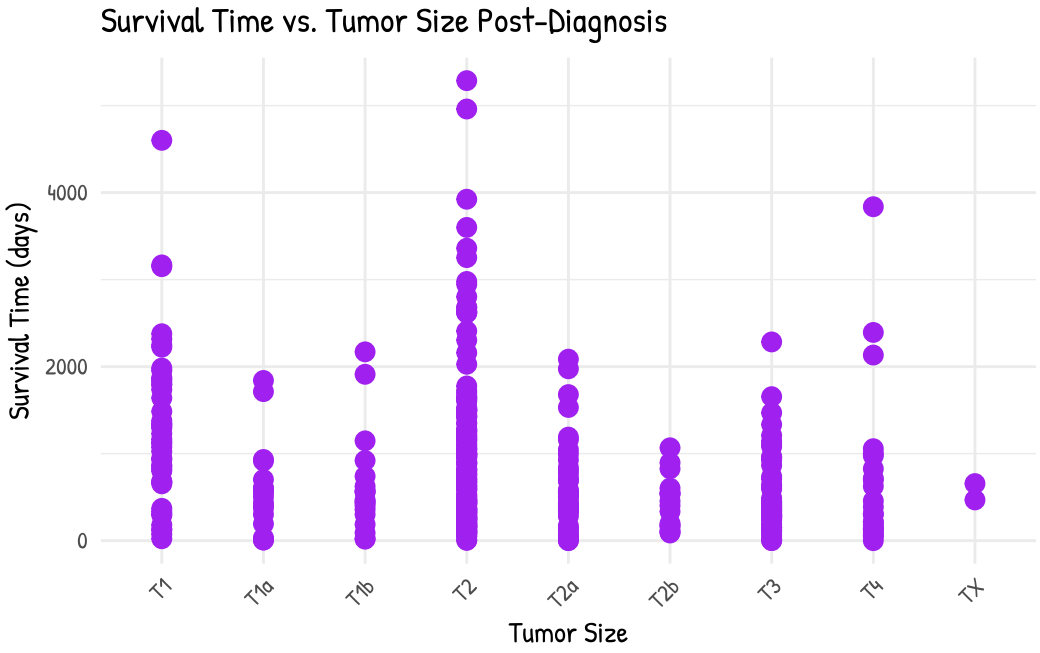

Density Plot of Days to Death by Pathologic Stage for Lung
Density
0.003
0.002
0.001
0.000
0
2000
4000
Days to Death
pathologic_stage
Stage I
Stage IA
Stage IB
Stage II
Stage IIA
Stage IIB
Stage III
Stage IIIA
Stage IIIB
Stage IV

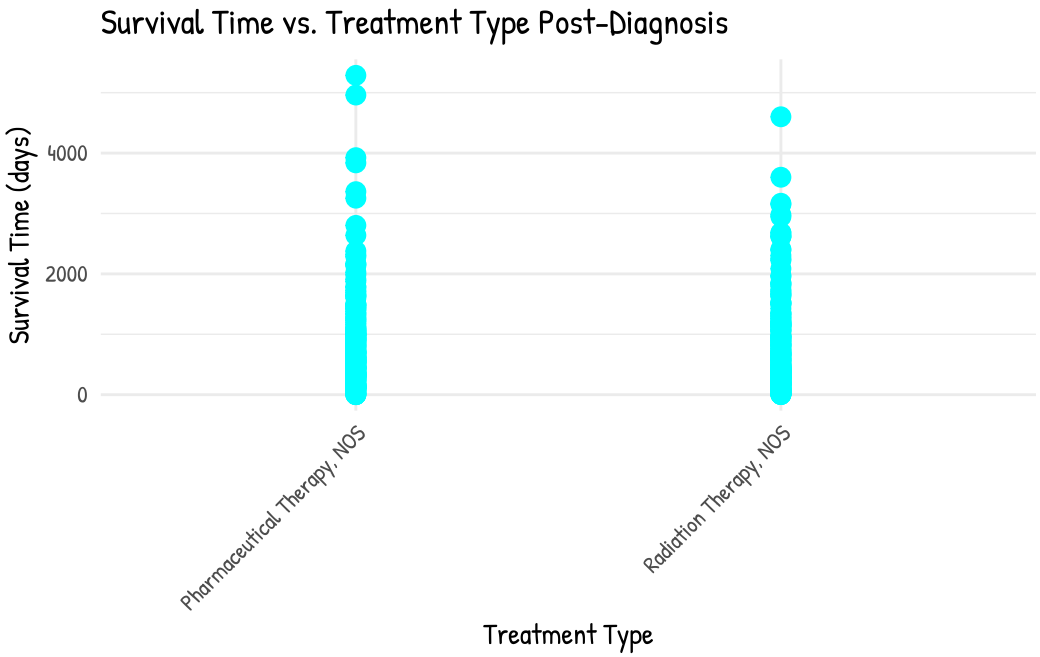

Density Plot of Days to Death by Pathologic Stage for Lung
Stage I
Stage IA
Stage IB
Stage II
Stage IIA
Stage IIB
Stage III
Stage IIIA
Stage IIIB
Stage IV
Density
Days to Death
pathologic_stage
Stage I
Stage IA
Stage IB
Stage II
Stage IIA
Stage IIB
Stage III
Stage IIIA
Stage IIIB
Stage IV

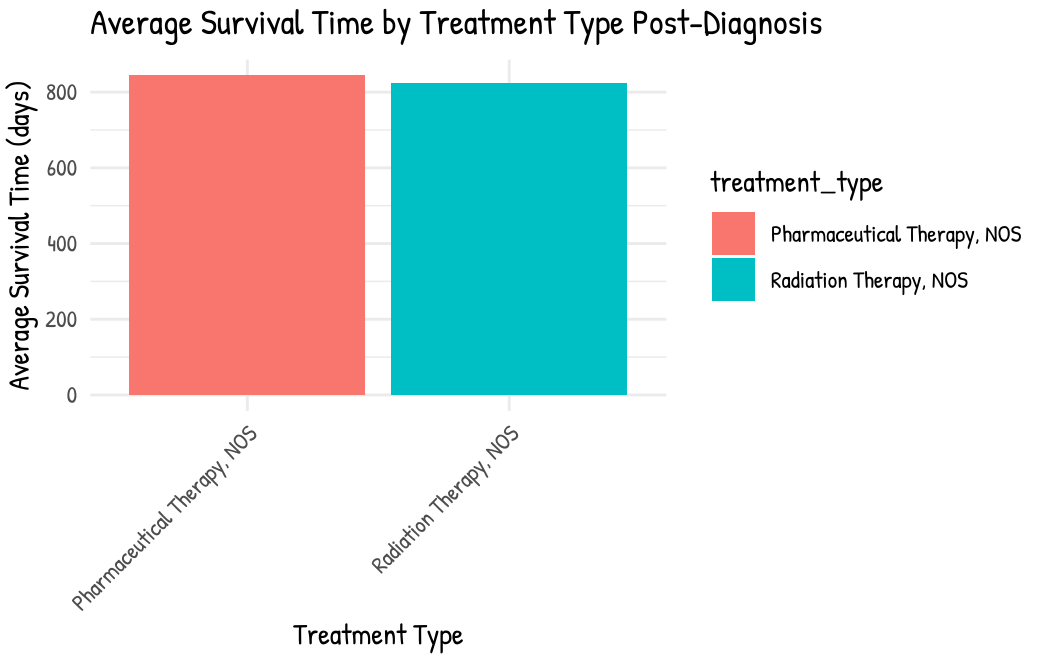

Survial Distribution Post-Diagnosis by Pathologic Stage
Days to Death
4000
2000
0
Stage I
Stage IA
Stage IB
Stage II
Stage IIA
Stage IIB
Stage III
Stage IIIA
Stage IIIB
Stage IV
Pathologic Stage

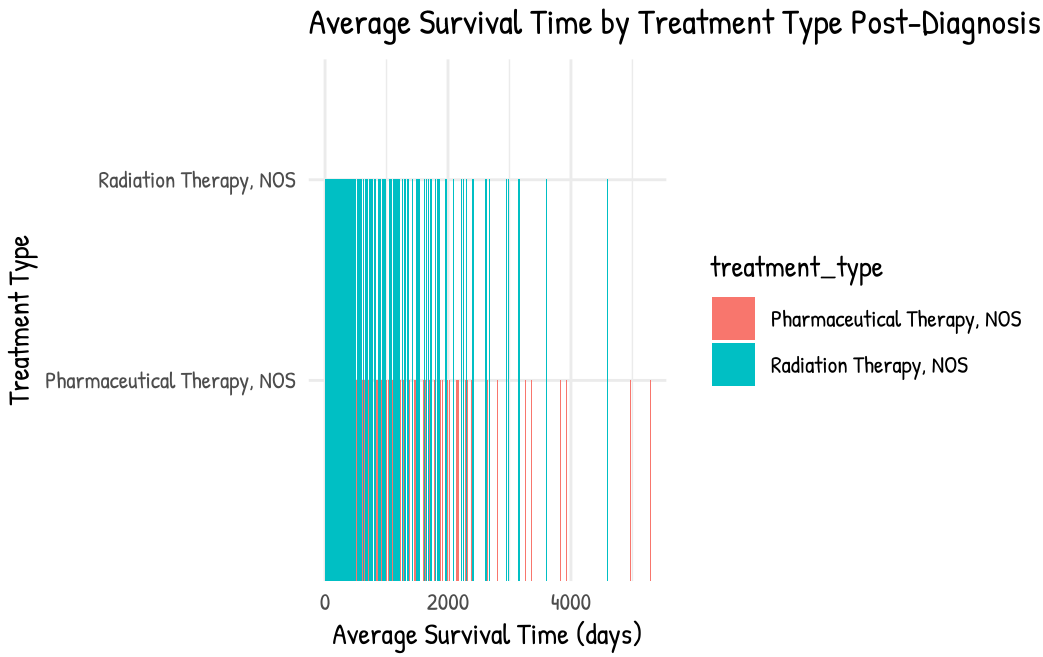

Survival Time vs. Presence of Distant Metastasis Post-Diagnosis
Survival Time (days)
4000
2000
0
M0
M1
M1a
M1b
MX
Presence of Distant Metastasis

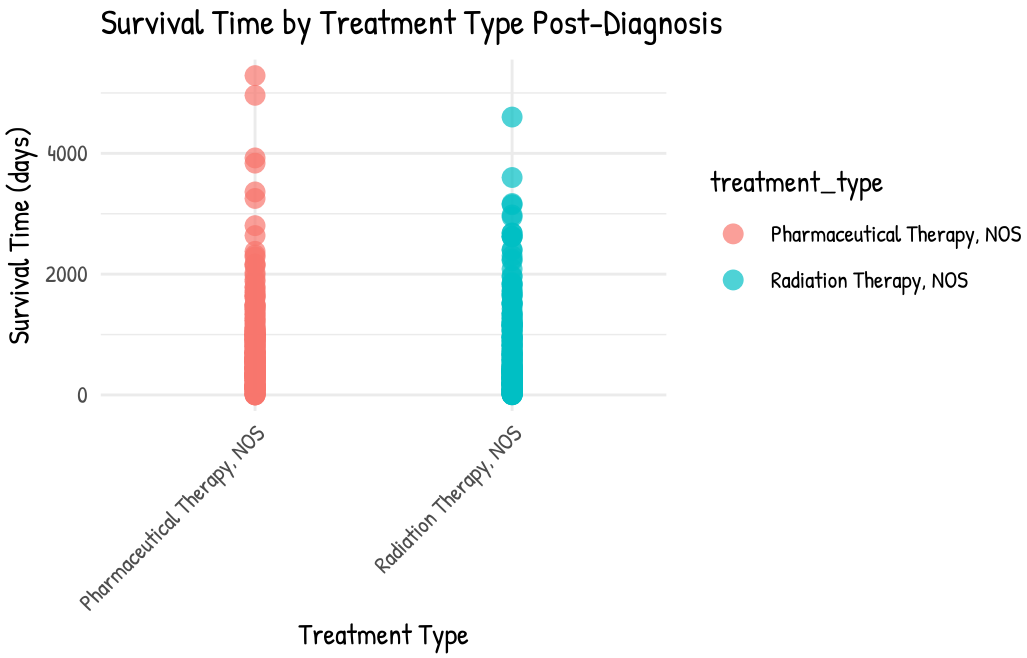

Survival Time vs. Lymph Node Involvement Post-Diagnosis
Survival Time (days)
4000
2000
0
N0
N1
N2
N3
NX
Lymph Node Involvement

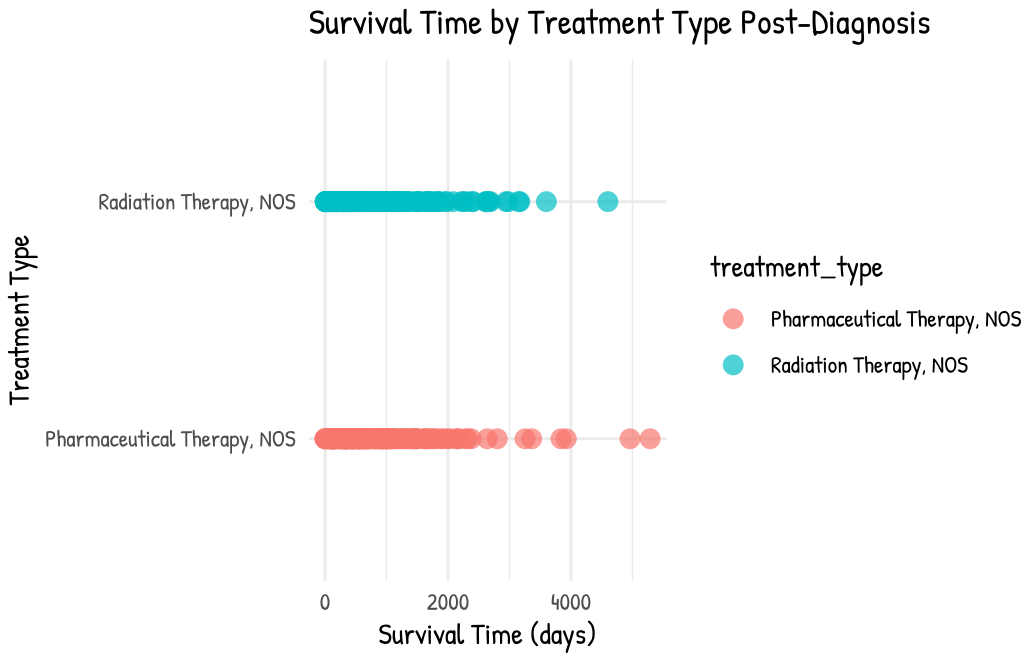

Survival Time vs. Tumor Size Post-Diagnosis
Survival Time (days)
4000
2000
0
T1
T1a
T1b
T2
T2a
T2b
T3
T4
TX
Tumor Size

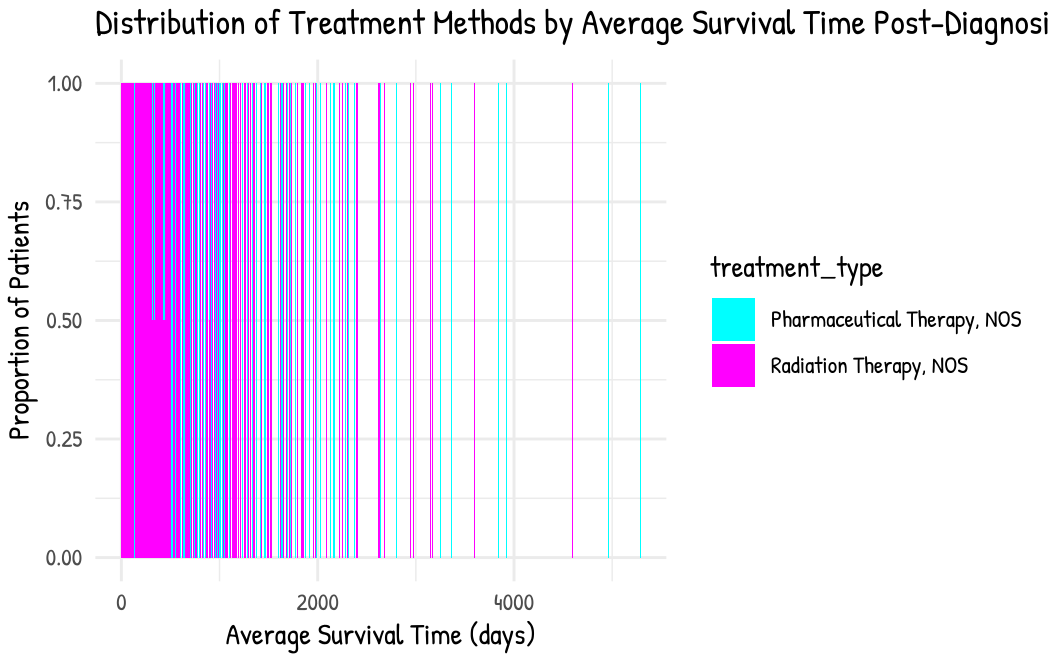

Survival Time vs. Treatment Type Post-Diagnosis
Survival Time (days)
4000
2000
0
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS
Treatment Type

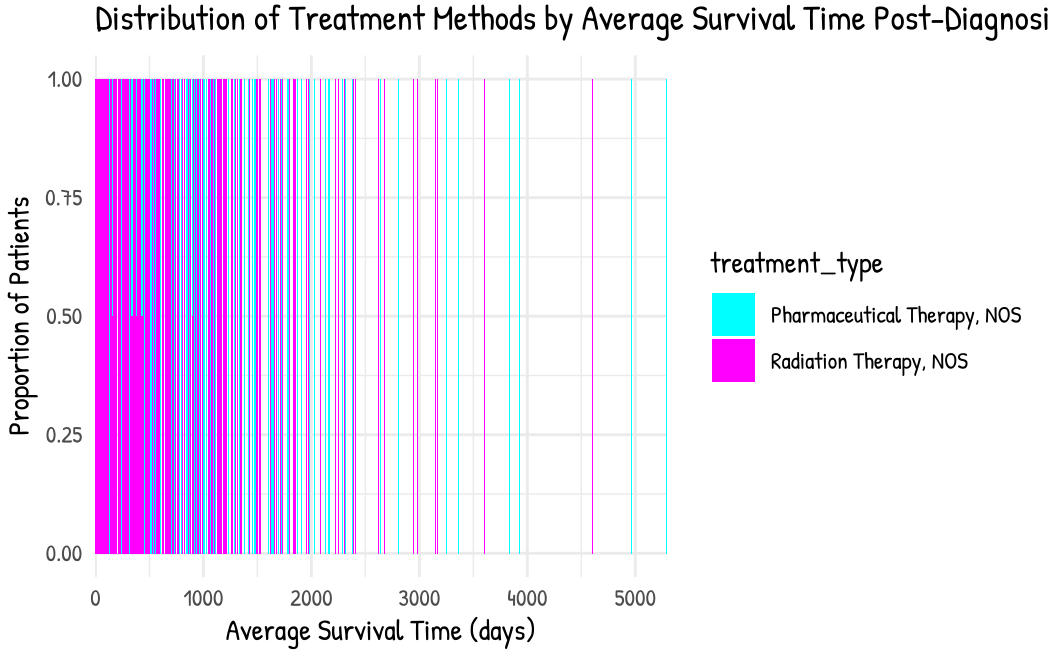

Average Survival Time by Treatment Type Post-Diagnosis
Average Survival Time (days)
800
600
400
200
0
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS
Treatment Type
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS

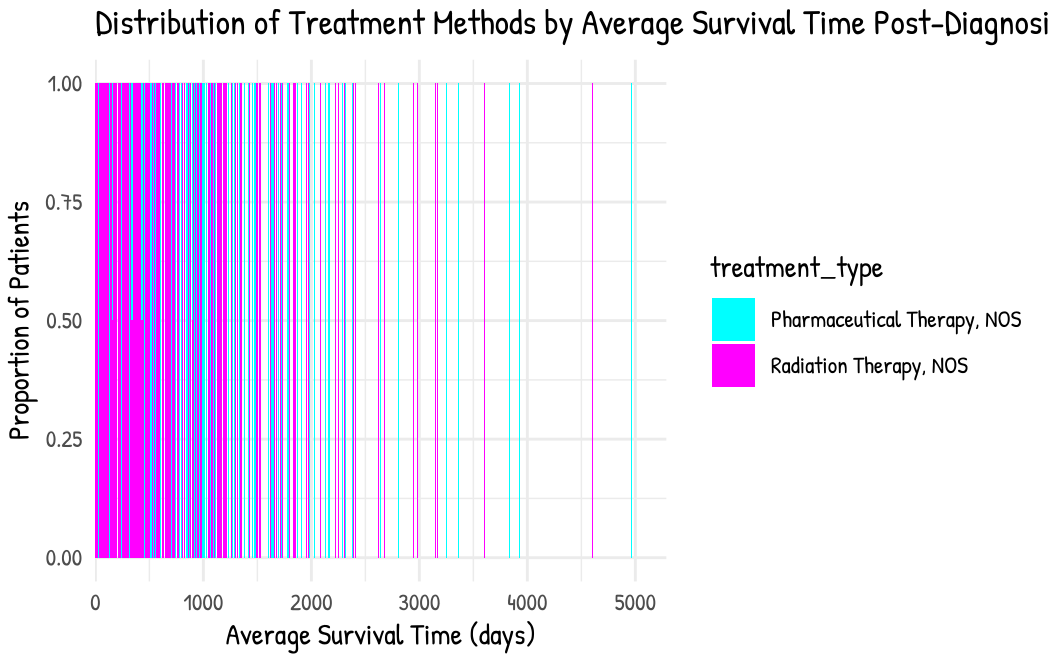

Average Survival Time by Treatment Type Post-Diagnosis
Treatment Type
Radiation Therapy, NOS
Pharmaceutical Therapy, NOS
0
2000
4000
Average Survival Time (days)
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS

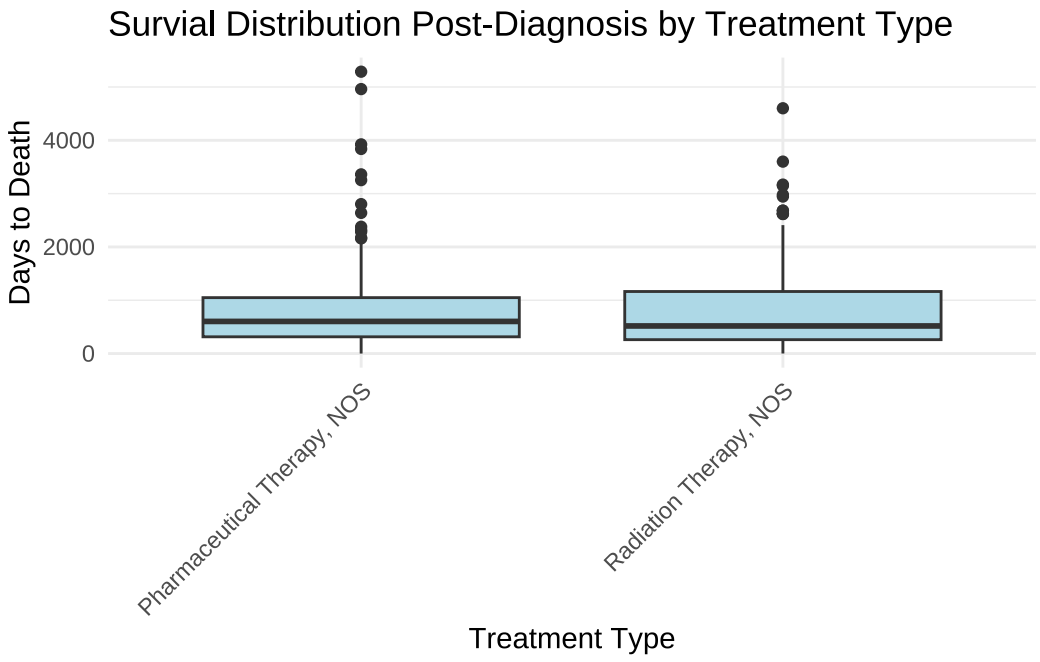

Survival Time by Treatment Type Post-Diagnosis
Survival Time (days)
4000
2000
0
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS
Treatment Type
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS


Survival Time by Treatment Type Post-Diagnosis
Radiation Therapy, NOS
Pharmaceutical Therapy, NOS
Treatment Type
0
2000
4000
Survival Time (days)
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS


Distribution of Treatment Methods by Average Survival Time Post-Diagnosi
Proportion of Patients
1.00
0.75
0.50
0.25
0.00
0
2000
4000
Average Survival Time (days)
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS


Distribution of Treatment Methods by Average Survival Time Post-Diagnosi
Proportion of Patients
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
Average Survival Time (days)
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS


Distribution of Treatment Methods by Average Survival Time Post-Diagnosi
Proportion of Patients
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
Average Survival Time (days)
treatment_type
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS


Survial Distribution Post-Diagnosis by Treatment Type
Days to Death
4000
2000
0
Pharmaceutical Therapy, NOS
Radiation Therapy, NOS
Treatment Type

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

A Appendix

B Additional data details

C Model details

we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected by, the data

Figure 3: **?(caption)**

C.1 Diagnostics

Is this needed?

Checking the convergence of the MCMC algorithm

Figure 4: **?(caption)**

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.