

Survival Compass*

Statistical Insights into Lung Cancer Patients Journey Post Diagnosis

Lexi Knight

April 17, 2024

This study investigates the impact of pathologic stage and treatment modalities on lung cancer survival post-diagnosis. Analysis of patient data reveals significant correlation between pathologic 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

1	Introduction	2
2	Data	4
2.1	Software and R-packages	4
2.2	Methodology	5
2.2.1	Data Collection	5
2.2.2	Data Cleaning	5
2.2.3	Data Analysis	5
2.3	Features	6
2.3.1	Days to Death	6
2.3.2	pathologic Stage	7
2.3.3	Presence of Distant Metastasis	7
2.3.4	Lymph Node Involvement	7
2.3.5	Tumor Size	8

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

3	Model	8
3.1	Model set-up	8
3.1.1	Model Specifications	9
3.1.2	Model justification	9
4	Results	10
5	Discussion	19
5.1	First discussion point	19
5.2	Second discussion point	19
5.3	Third discussion point	19
5.4	Weaknesses and next steps	19
A	Appendix	20
B	Additional data details	20
C	Model details	20
C.1	Diagnostics	20
	References	21

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). pathologic 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 pathologic 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 pathologic stage of the cancer is then determined (Eldridge, 2022). The most valuable prognostic factor in non-small cell lung cancer is the pathologic 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 pathologic 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 pathologic 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).

Radiation therapy is a local treatment, targeting the tumor directly, damaging the DNA within cancer cells with the aim of shrinking the tumor. The success rate of radiation therapy treatment is dependent on the location, stage of the cancer as well as individual factors. It is the primary treatment for early stage NSCLC and palliative treatment. Chemotherapy is a systemic therapy utilizing drugs to kill and inhibit cancer cell growth and is often the primary treatment for stage 3 and 4 cancers. NSCLC patients with signs of lymph node metastases have shown great benefit in survival when treated with chemotherapy (Kai, 2021). A study looking at patients with pathologic stage 3 and 4 NSCLC showed that those who received chemotherapy survived for an average of 10.5 months whereas those that received RT only survived for 3.7 months (Eldridge, 2022). Patients with early pathologic stages such as 1, 2 and 3 undergo curative treatment. Stage 4, the advanced stage however, is treated as palliative treatment in that the cancer is far too advanced and thus all that can be done is to minimize symptoms and try to improve quality of life. Stage 3b and stage 4 tumors inoperable (Eldridge, 2022). Overall, the choice between treatments is made based on assessment of the patient’s condition, pathologic stage with tradeoff between providing effective treatment and optimizing quality of life. (Eldridge, 2022)

The remainder of this paper is structured as follows. In Section 2, we visualize the exploration of variables constituting the pathologic 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 the `tidyverse` (Wickham, Averick, et al. 2024) package, which includes `dplyr` (Wickham, François, et al. 2024), `readr` (Wickham, Hester, et al. 2024). Specifically, `readr` was employed for efficient reading of rectangular text data, while `dplyr` facilitated data manipulation tasks such as filtering, summarizing, and joining datasets. For unit testing dataset, we utilized the `testthat` (Wickham et al. 2024) package. This allowed us to systematically test our functions and ensure that they behaved as expected across various scenarios. In our statistical modeling process, we utilized the `rstanarm` (Goodrich et al. 2022) package for Bayesian applied regression modeling. This package leverages the Stan probabilistic programming language for efficient computation of Bayesian models, allowing

us to perform complex regression analyses while incorporating uncertainty. To tidy up and summarize mixed effects models, we employed the `broom.mixed` (Bauer et al. 2024) package. This package provides functions to visualize the results of mixed effects models, making it easier to interpret the findings from such analyses. For arranging and combining plots in our visualizations, we utilized the `patchwork` (Pedersen et al. 2024) package. This allowed us to seamlessly arrange multiple plots into a single coherent visual representation, facilitating the communication of complex relationships and patterns in the data. For aesthetic purposes in our visualizations, we employed the `showtext` (Qiu 2024) package. This enabled us to use a wide range of fonts in our plots, enhancing the visual appeal and customization of our graphical outputs.

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** ((Magic) 2023), 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-naïve 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, we selected the columns of interest namely; days to death post diagnosis, presence of distant metastasis, lymph node involvement, pathologic stage, tumor size and treatment type. Next, we renamed the columns, giving them meaningful names and excluded the data containing missing values. This left us with a sample of 382 lung cancer patients. Additionally, we converted the days to death column to numeric. Tests were included to ensure accuracy, reliability and validity of the dataset 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, see Section 3.

2.3 Features

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

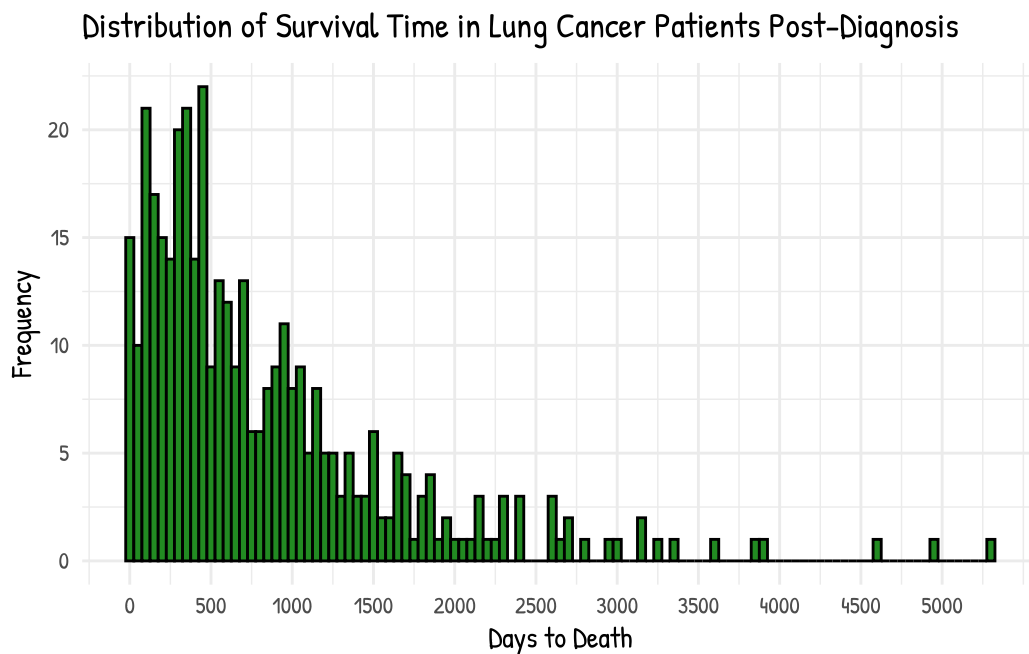


Figure 1: Distribution of survival time in patients

2.3.1 Days to Death

The main feature 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 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.

Figure 1 illustrates the survival curve for lung cancer patients post-diagnosis where frequency is on the y-axis and survival time in days is on the x-axis. There is a clear trend of a decrease in

frequency of survival as time increases. Mortality is most abundant 500 days after diagnosis, that is equivalent to about a year and four months. After about 1500 days, just over four years, the trend line plateaus.

Figure 1, the distribution of survival time post-diagnosis suggests that there is a high mortality rate shortly after diagnosis, thus most lung cancer patients die within one year of diagnosis (500 days). There is a steady decline in death and hence incline in survival up until four years (1500 days). The plateau after this indicates that there is a subgroup of patients who survive beyond four years. Section 4 explores which patients had longer survival.

2.3.2 pathologic Stage

Pathologic stage, a critical determinant of lung cancer prognosis, was classified according to the TNM (tumor, lymph node, metastasis) Cancer Staging. As depicted in ?@fig-two, illustrating the the distribution of pathologic stages with pathologic stage on the x-axis and percentage on the y-axis. The distribution of patients varied across different stages, with the highest proportion of patients diagnosed at StageIB. At this stage, the cancer is still found early however the tumor has regional spread to nearby lymph nodes. Interestingly, at diagnosis, patients had a fairly even spread across the different stages.

2.3.3 Presence of Distant Metastasis

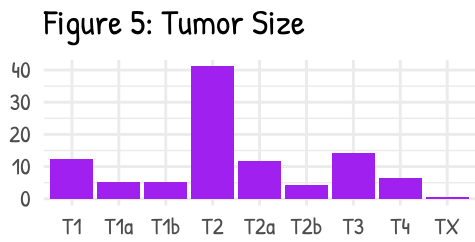
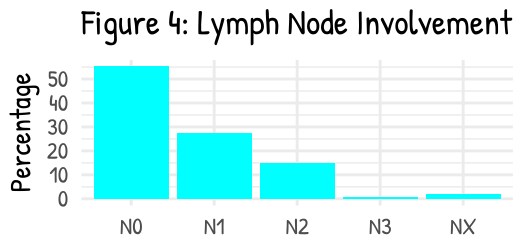
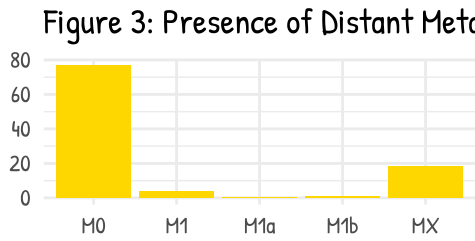
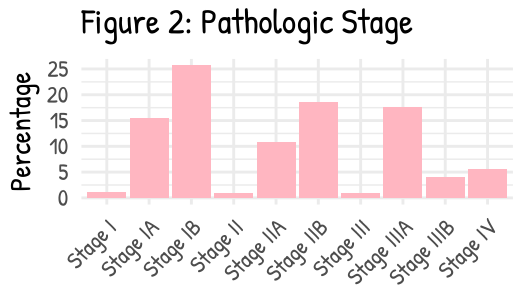
The presence of distant metastasis, indicative of cancer spread beyond the site of origin to nearby lymph nodes, significantly influences treatment decisions and patient outcomes. ?@fig-three illustrates the percentage of patients with distant metastasis where the x-axis is the classification and the y-axis is the percentage of patients. The figure indicates that when patients are diagnosed, most are found in category 'M0' having no distant metastasis and thus cancer has not spread beyond the site of origin. It is important to note that category 'MX', where tumor could not be evaluated has the next largest proportion of patients. Additionally, category 'M1a' appears to show no data however this is because only 1 out of the 382 patients were classified this way.

2.3.4 Lymph Node Involvement

Lymph node involvement is another key prognostic factor in lung cancer, reflecting the extent of disease spread to regional lymph nodes. ?@fig-four presents the distribution of patients based on lymph node involvement where lymph node involvement is on the x-axis and percentage is on the y-axis. Here, there is a clear trend that as lymph node involvement becomes more severe, the proportion of patients decline. This demonstrates that upon diagnosis, most patients fall under the category 'N0' where there is no regional lymph node metastasis hence cancer has not spread to the lymph nodes.

2.3.5 Tumor Size

Tumor size is measured based on the diameter of the primary tumor in lung cancer patients. **Figure 2** showcases the distribution of patients across different tumor size categories with tumor size on the x-axis and percentage of patients on the y-axis. The highest proportion of patients were categorized as 'T2' where the tumor is between 3 to 5 centimeters in diameter has grown into the inner lining of the lung, possibly leads to swelling and or collapse of the lung. This category was overwhelmingly more prominent than all others by twofold. All other categories were relatively evenly distributed.



two, fig-three, fig-four, fig-five} {#fig-

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 pathologic 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.
- $\beta_{\{\text{pathologic_stage}\}}, \beta_{\{\text{lymph_node}\}}, \beta_{\{\text{metastasis}\}}, \beta_{\{\text{tumor_size}\}}, \beta_{\{\text{treatment_type}\}}$ are the coefficients associated with each predictor variable.

3.1.2 Model justification

Linear regression models are most appropriate 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

Our variable of interest is survival time in lung cancer patient after 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

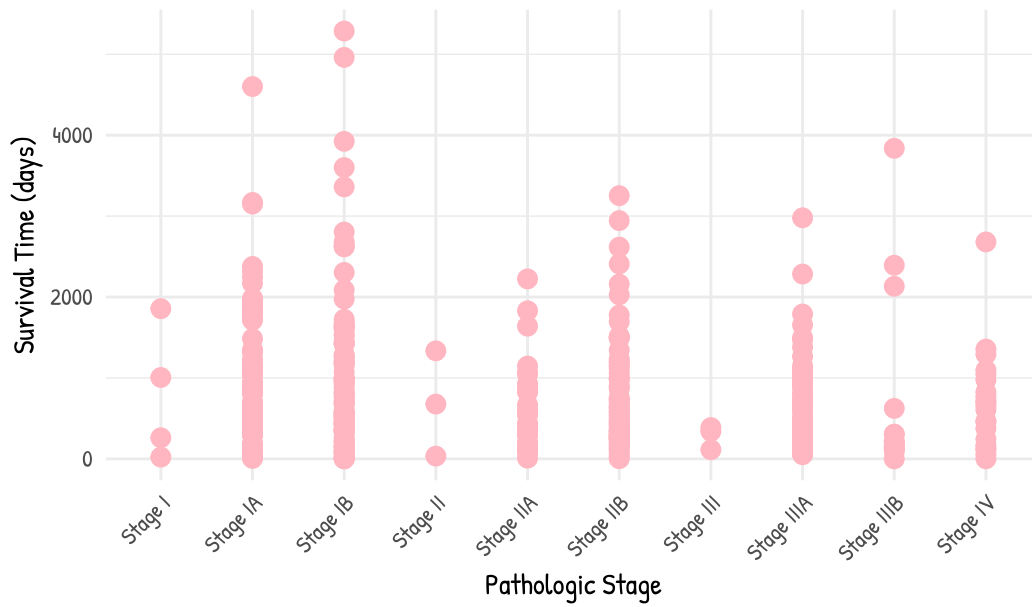
We anticipate that the survival time of lung cancer patients post-diagnosis will be influenced by various clinical factors such as pathologic stage, extent of lymph node involvement, presence of distant metastasis, tumor size, and treatment type. Specifically, we expect that advanced pathologic stages,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

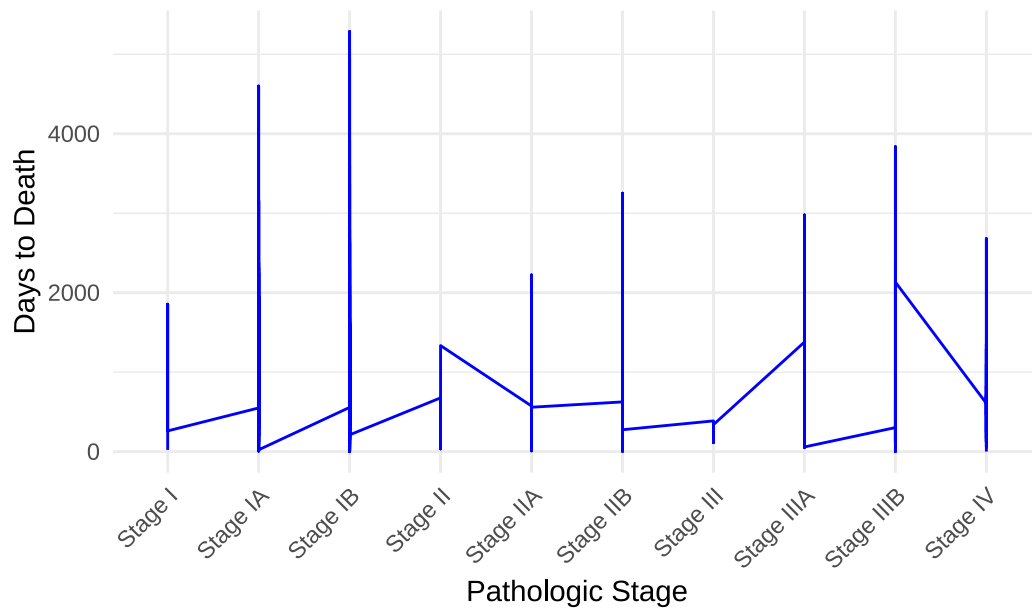
4 Results

Our results are summarized in @.

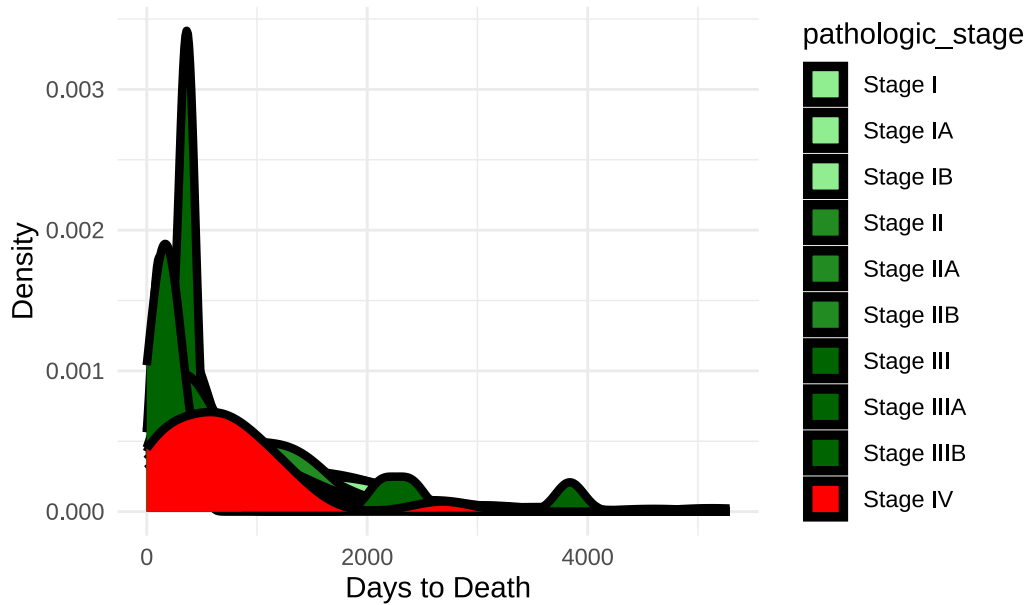
Survival Time vs. Pathologic Stage Post-Diagnosis



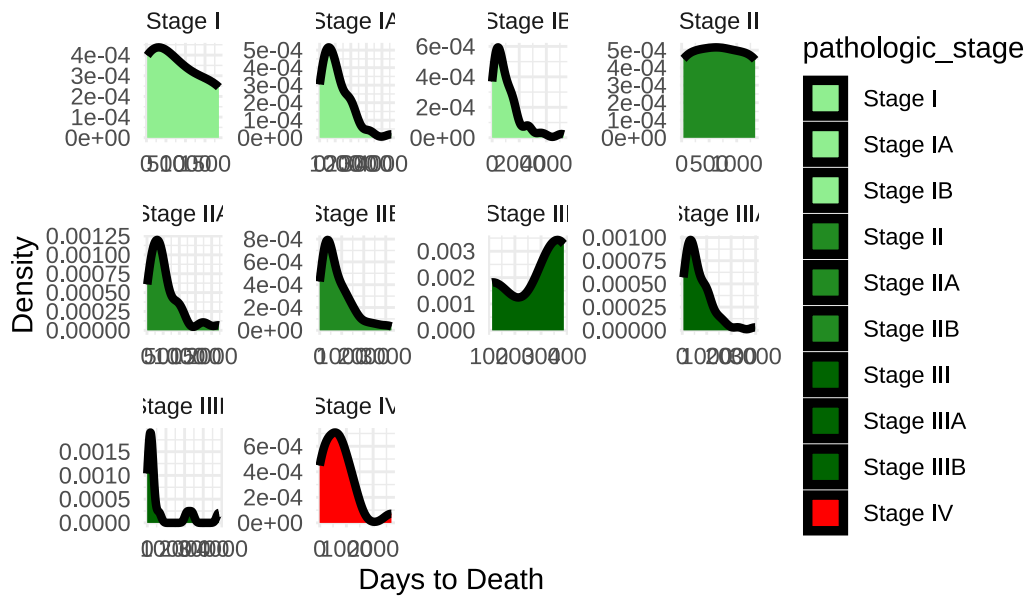
Days to Death vs. Pathologic Stage for Lung Cancer Patients

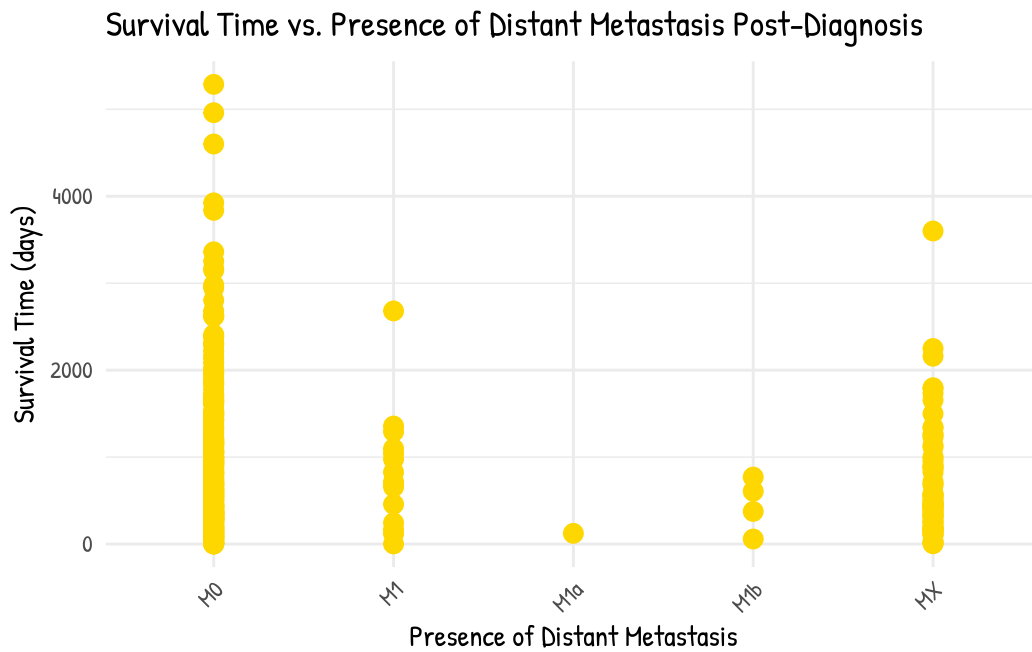
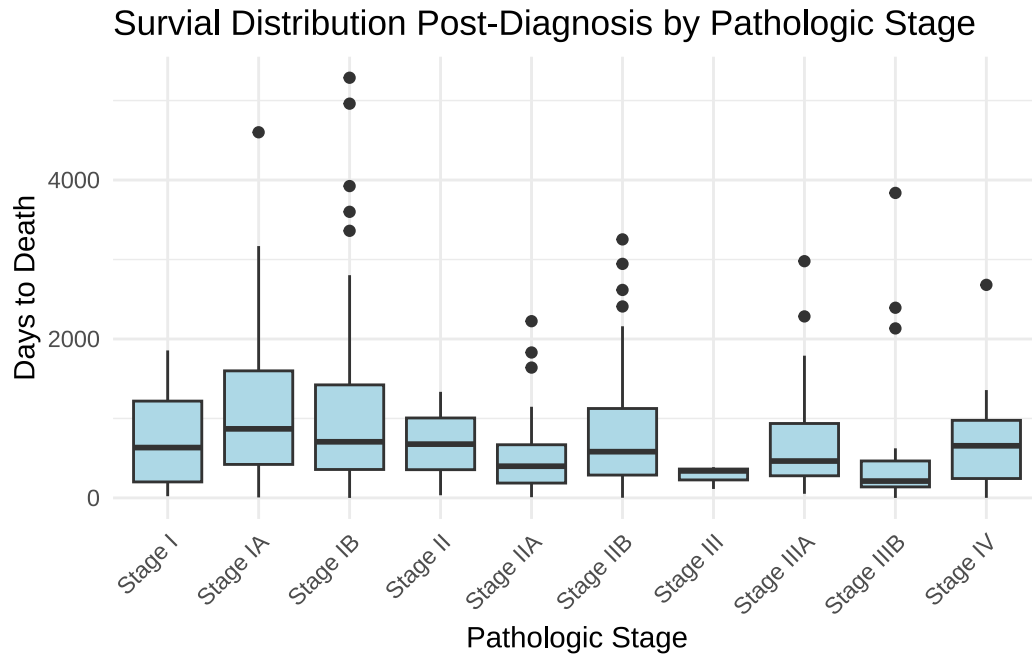


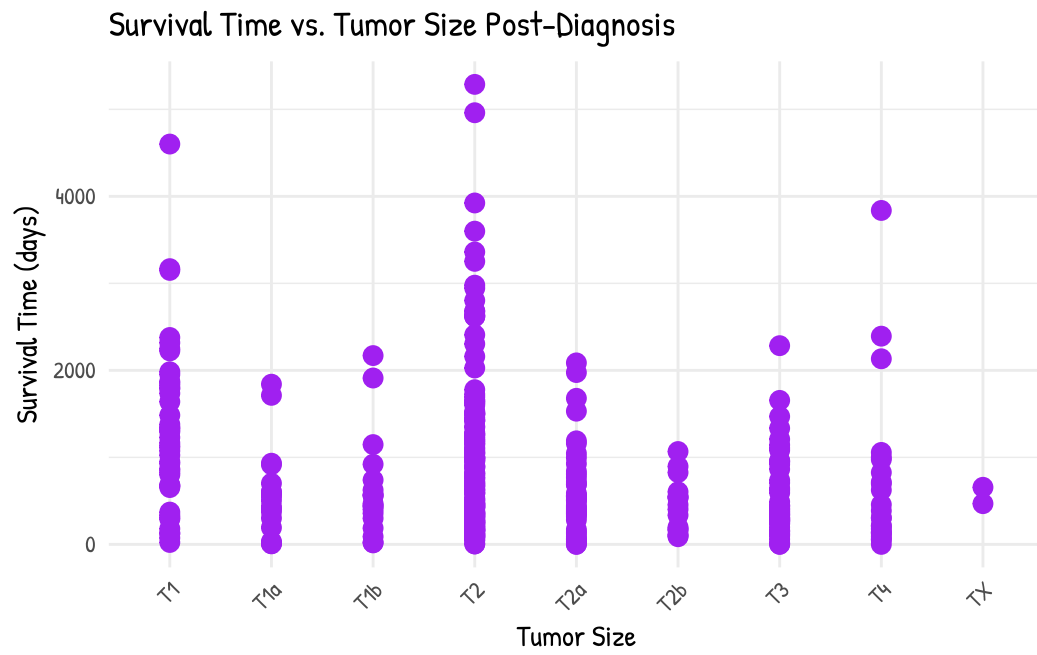
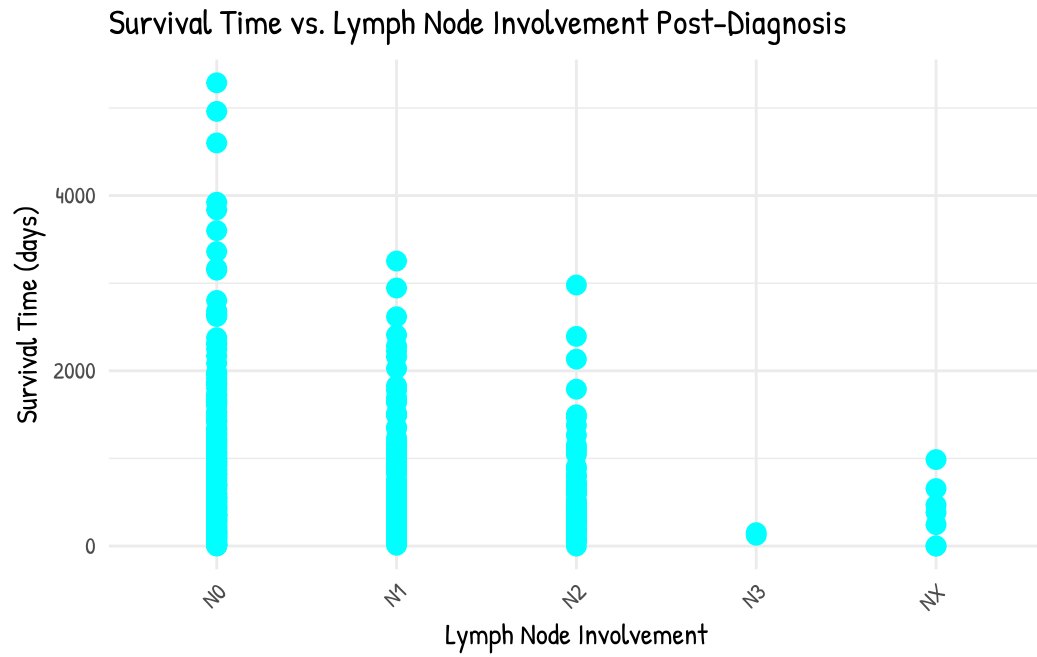
Density Plot of Days to Death by Pathologic Stage for Lung

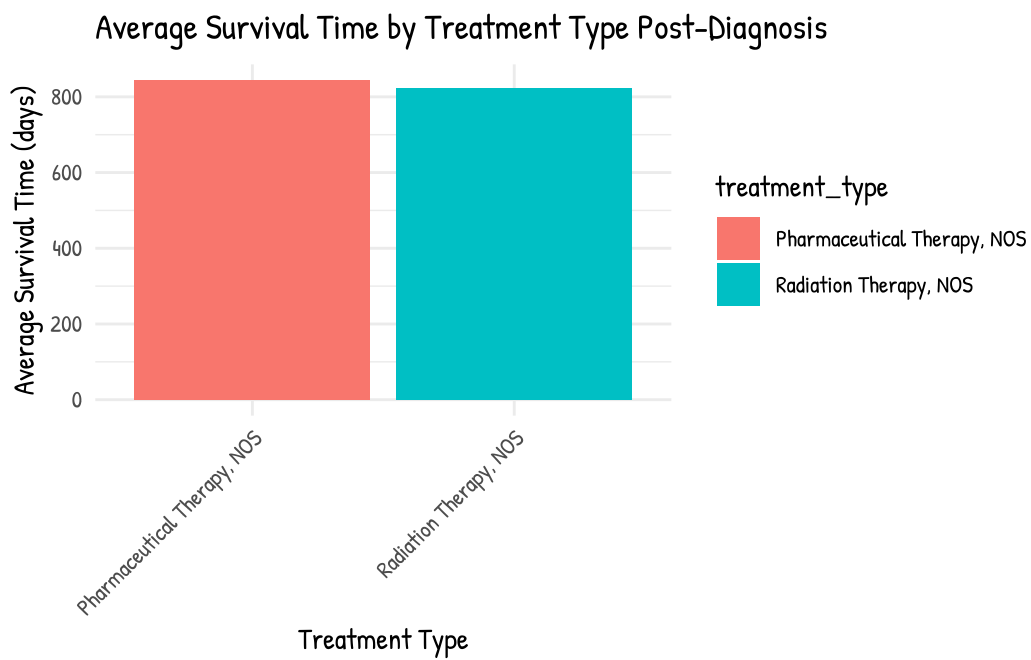
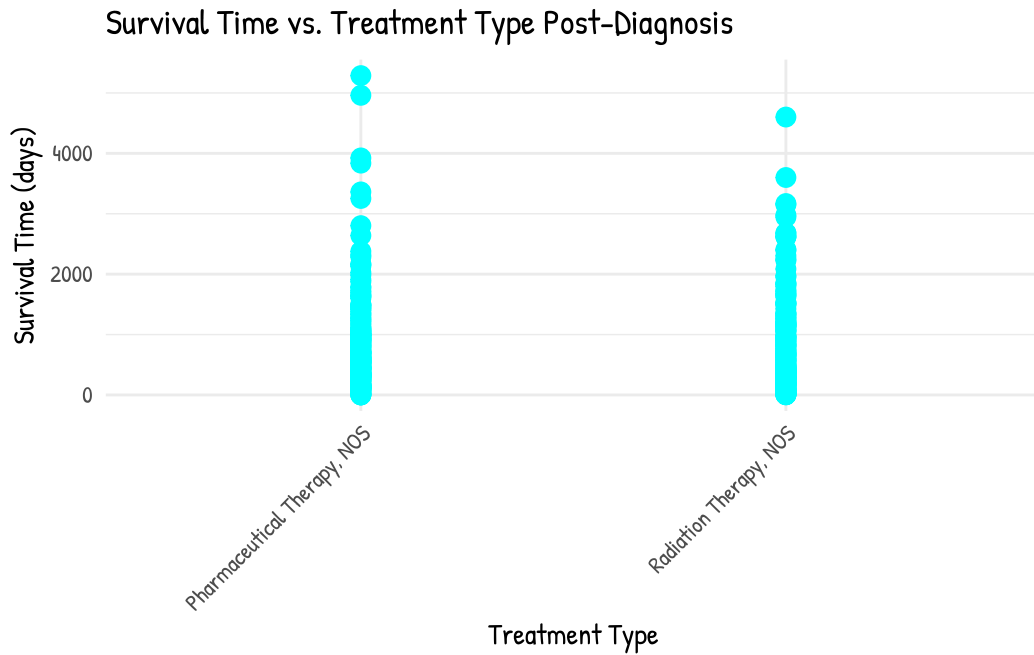


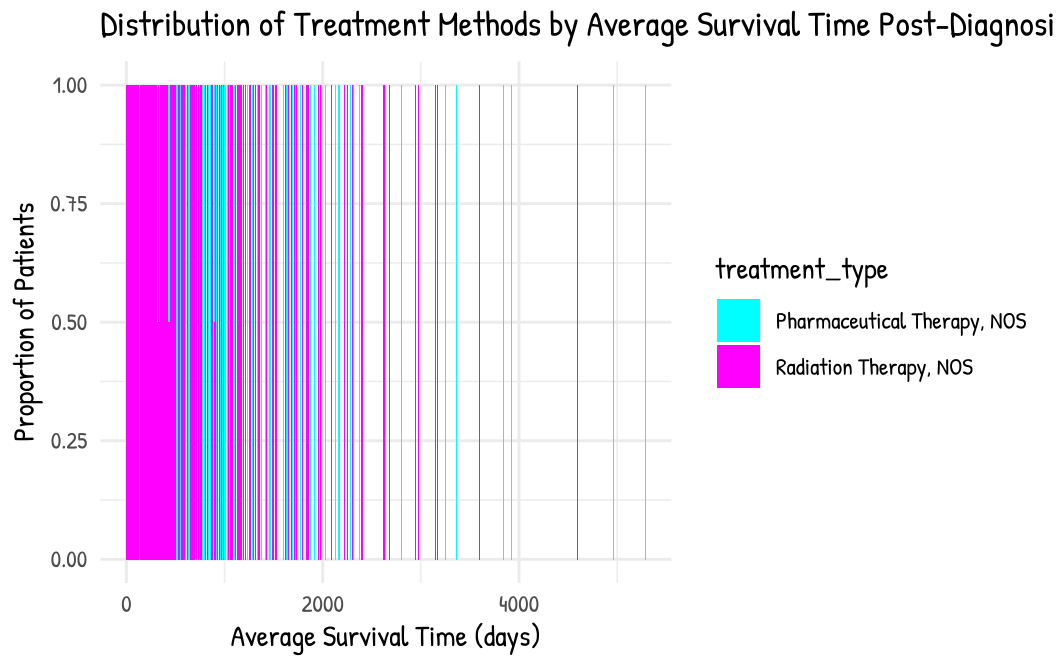
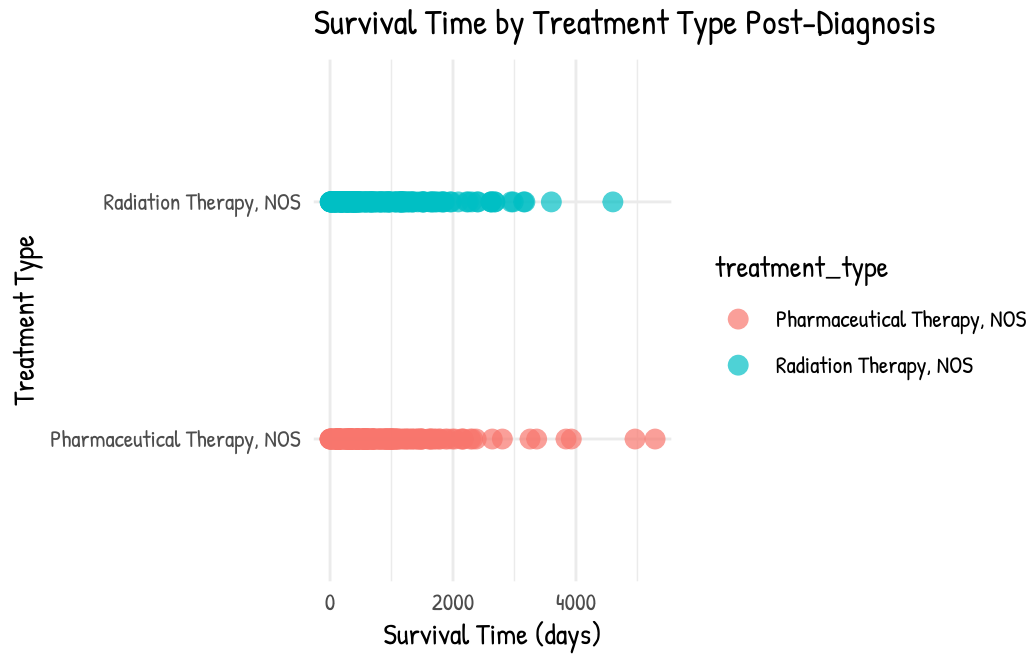
Density Plot of Days to Death by Pathologic Stage for Lung



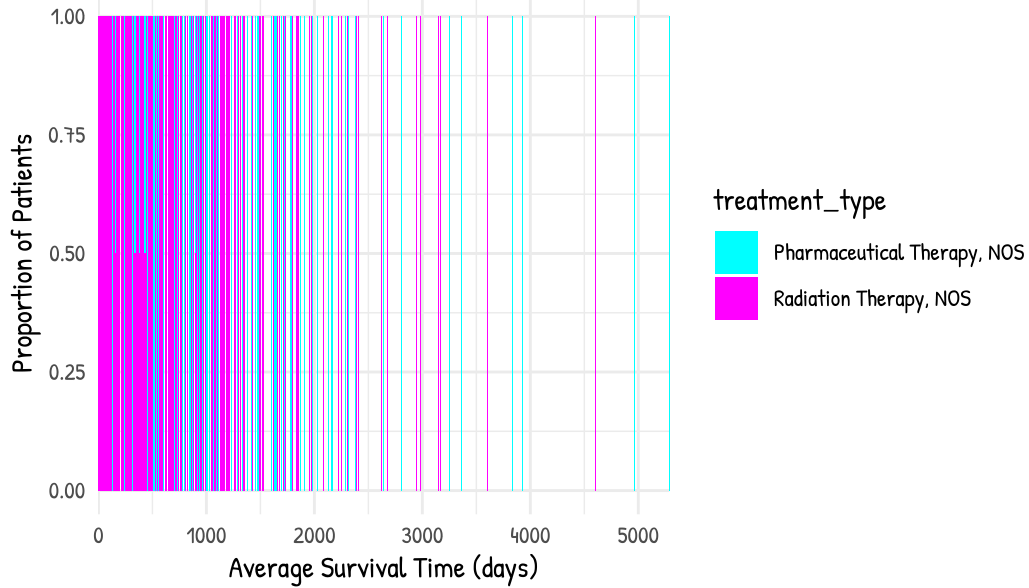




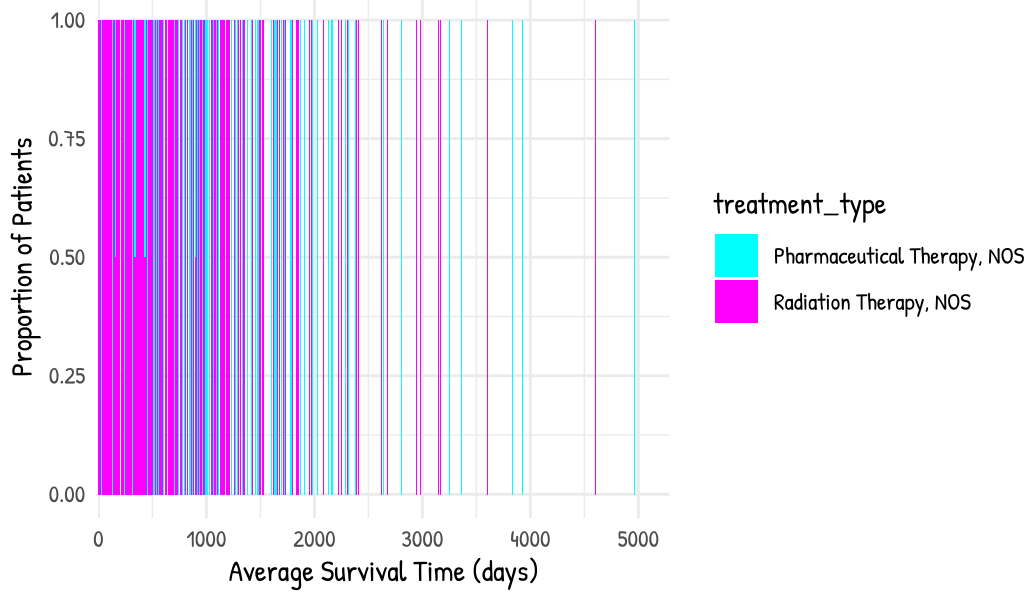


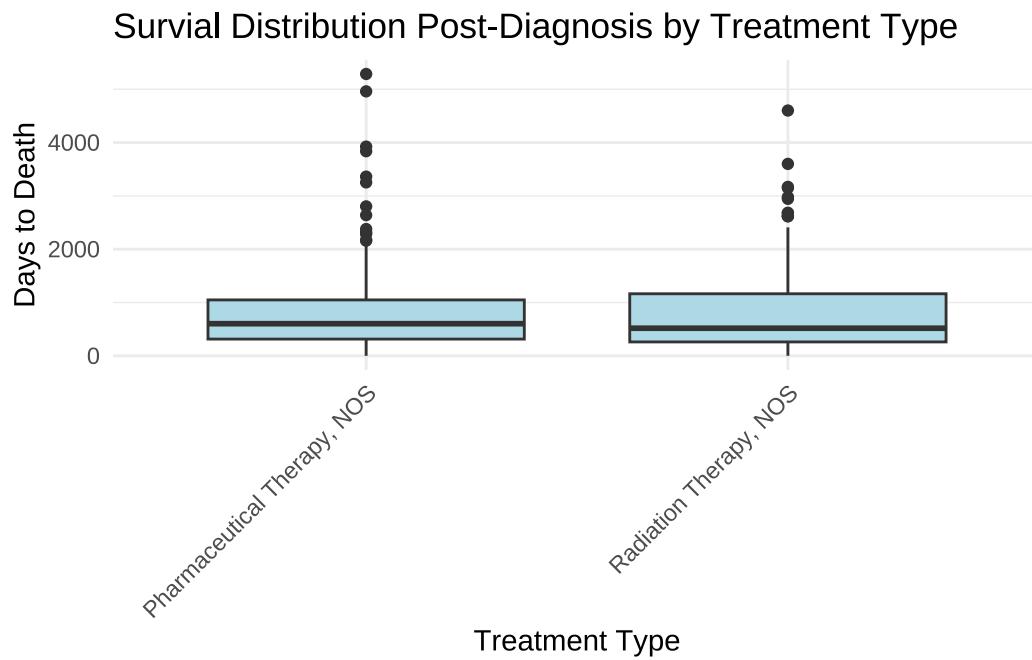


Distribution of Treatment Methods by Average Survival Time Post-Diagnosis



Distribution of Treatment Methods by Average Survival Time Post-Diagnosis





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 2: ?(caption)

C.1 Diagnostics

Is this needed?

Checking the convergence of the MCMC
algorithm

Figure 3: ?(caption)

References

- Bauer, Paul C. et al. 2024. *Broom.mixed: Tidy Summaries of Mixed Effects Models*. <https://CRAN.R-project.org/package=broom.mixed>.
- Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. “Rstanarm: Bayesian Applied Regression Modeling via Stan.” <https://mc-stan.org/rstanarm/>.
- (Magic), Sicong Chen. 2023. “Lung Cancer Prediction.” <https://osf.io/qk34b/>.
- Pedersen, Thomas Lin et al. 2024. *Patchwork: The Composer of Plots*. <https://CRAN.R-project.org/package=patchwork>.
- Qiu, Yixuan. 2024. *Showtext: Using Fonts More Easily in r Graphs*. <https://CRAN.R-project.org/package=showtext>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley et al. 2024. *Testthat: Unit Testing for r*. <https://CRAN.R-project.org/package=testthat>.
- Wickham, Hadley, Mara Averick, et al. 2024. *Tidyverse: Easily Install and Load the 'Tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, Romain François, et al. 2024. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, Jim Hester, et al. 2024. *Readr: Read Rectangular Text Data*. <https://CRAN.R-project.org/package=readr>.