Survival Compass*

Understanding the Impact of Pathologic Stage and Treatment Modalities on Lung Cancer Patients' Survival Post-Diagnosis.

Lexi Knight

April 21, 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

*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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 2015).

Presence of metastatic LN is one of the most important determinants of prognosis of NSCLC cases (Kai et al. 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 et al. 2017). Lymph node involvement, also termed lymph node ratio, is a crucial factor in guiding treatment options (Kai et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 2021). A study looking at patients with pathologic stage 3 and 4 NSCLC showed at 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. 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` (Chen 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-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, 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 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, 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.

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 plateus.

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

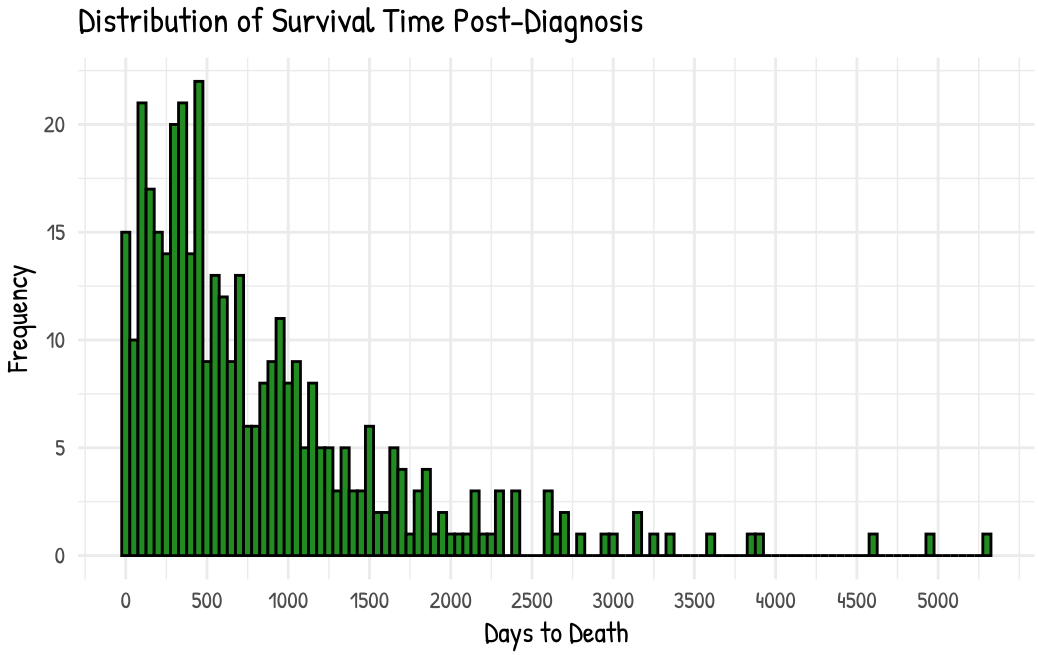

Figure 1: The histogram shows the distribution of survival time (days_to_death) in lung cancer patients post-diagnosis.

(500 days). There is a steady decline in death and hence incline in survial up until four years (1500 days). The plateu 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 Figure 2, 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. Figure 2 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. We decided to include the classification 'MX', where tumor could not be evaluated as it contained the next largest proportion of patients. If this data was omitted, there would be a large skew in the data. Additionally, category 'M1a' appears to show no data however this can be explained as 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. Figure 2 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.

2.3.6 Treatment Type

Treatment type analysis examines the distribution of lung cancer patients based on the type of treatment they received. Figure 3 illustrates this distribution, with treatment type on the x-axis and the percentage of patients on the y-axis. Pharmaceutical Therapy and Radiation Therapy were the two primary treatment categories observed. The split between usage of these two treatment types was very minimal with pharmaceutical therapy at 50.3% and radiation therapy at 49.7%.

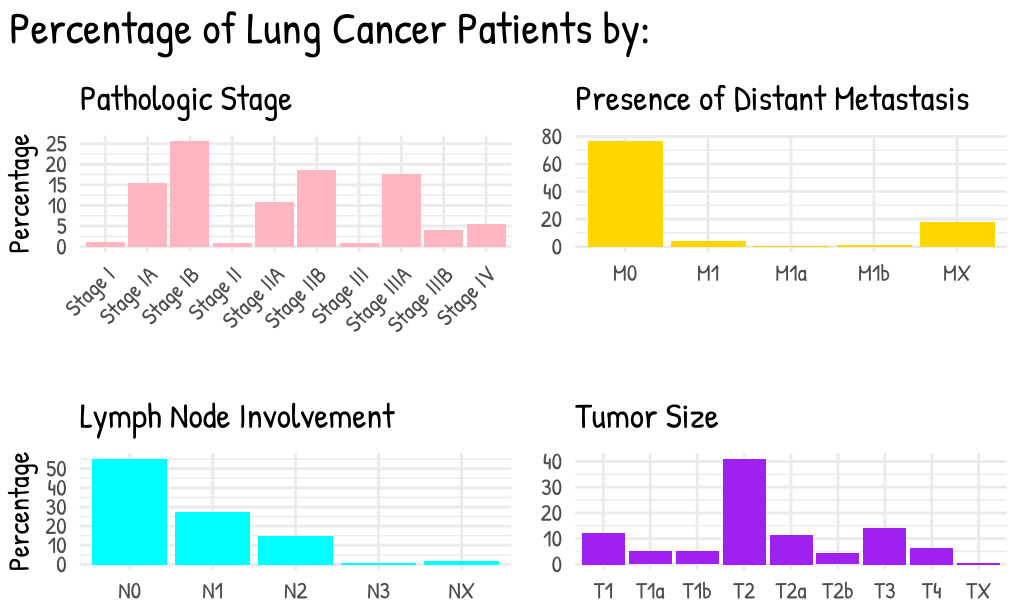

Figure 2: The histogram shows the percentage of lung cancer patients by pathologic stage, presence of distant metastasis, lymph node involvement, tumor size

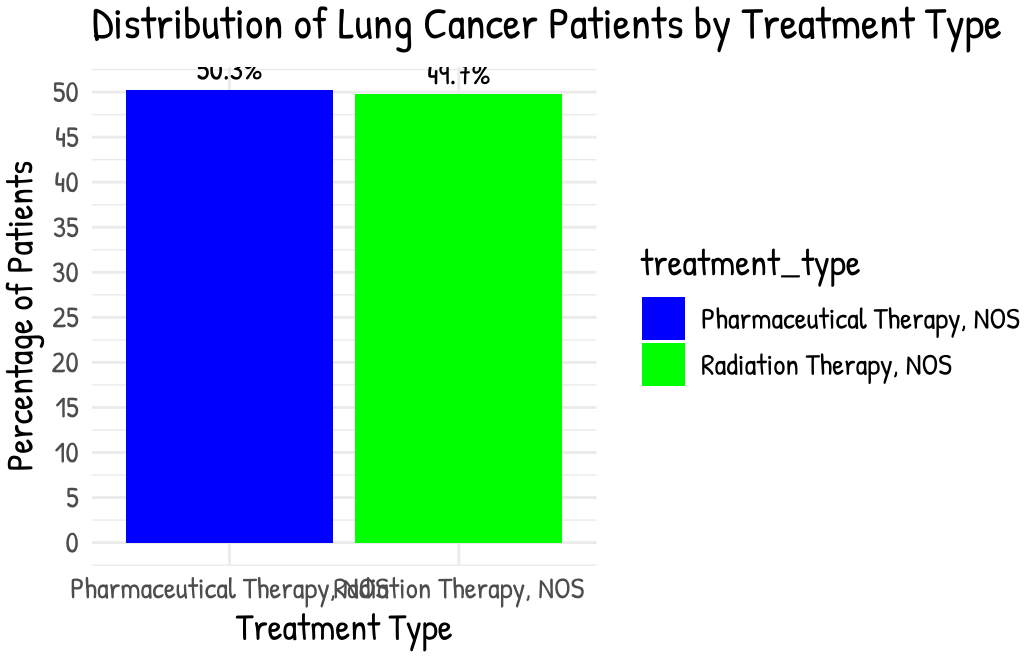

Figure 3: The histogram shows the percentage of lung cancer patients by treamtent type.

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.
- 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 Figure 4, Figure 5, Figure 6 and Figure 7. This analysis investigates lung cancer treatment outcomes given either radiation therapy or pharmaceutical therapy as well as pathologic stage, presence of distant metastasis, lymph node involvement and tumor size.

4.1 Treatment Type by Pathologic Stage

This scatter plot Figure 4 provides a view of how pathologic stage correlates with the time to death among lung cancer patients, differentiated by treatment type. Each dot represents an individual patient, color-coded to represent whether they received pharmaceutical therapy or radiation therapy. On the x-axis we have days to death after diagnosis and on the y-axis we have the pathologic stages. There are considerably fewer patients part of stages I, II, III, and IIB at diagnosis compared to the other ones. As the time from diagnosis to death increases along the x-axis, we observe a general trend downwards in pathologic stage, indicating disease progression over time. The trend lines fitted for each treatment type further highlight this progression, with radiation therapy showing a steeper decline compared to pharmaceutical therapy.

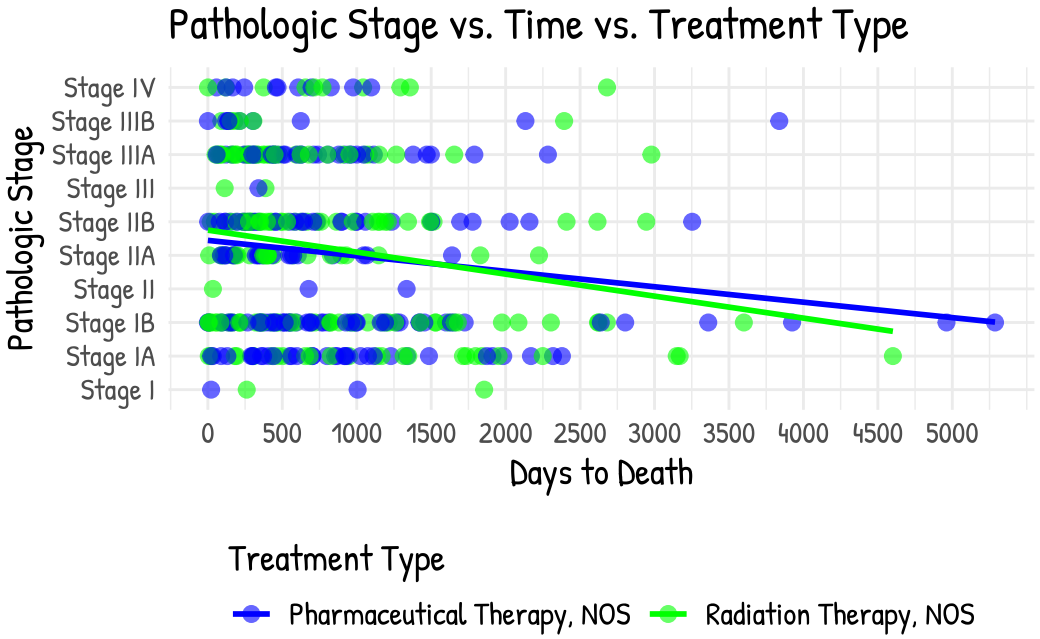

Figure 4: The scatterplot shows the relationship between pathologic stage, time to death, and treatment type in lung cancer patients post-diagnosis.

4.2 Treatment Type by Presence of Distant Metastasis

In this scatter plot, Figure 5 we examine the relationship between the presence of distant metastasis, time to death, and treatment type for lung cancer patients. Similarly to figure above, the dots on the plot represent individual patients, color-coded based on their treatment regimen. Days to death after diagnosis is on the x-axis and on the y-axis we have the degree

of the presence of distant metastasis. The majority of patients are diagnosed at classification 'MO' and hence are caught before cancer is spread from the primary site of origin. As patients survival time increases, excluding category 'MX', there is a decrease in the presence of distant metastasis, denoted by the downward trend in the scatter plot. Patients undergoing pharmaceutical therapy show a sharper decrease in distant metastasis compared to those receiving radiation therapy.

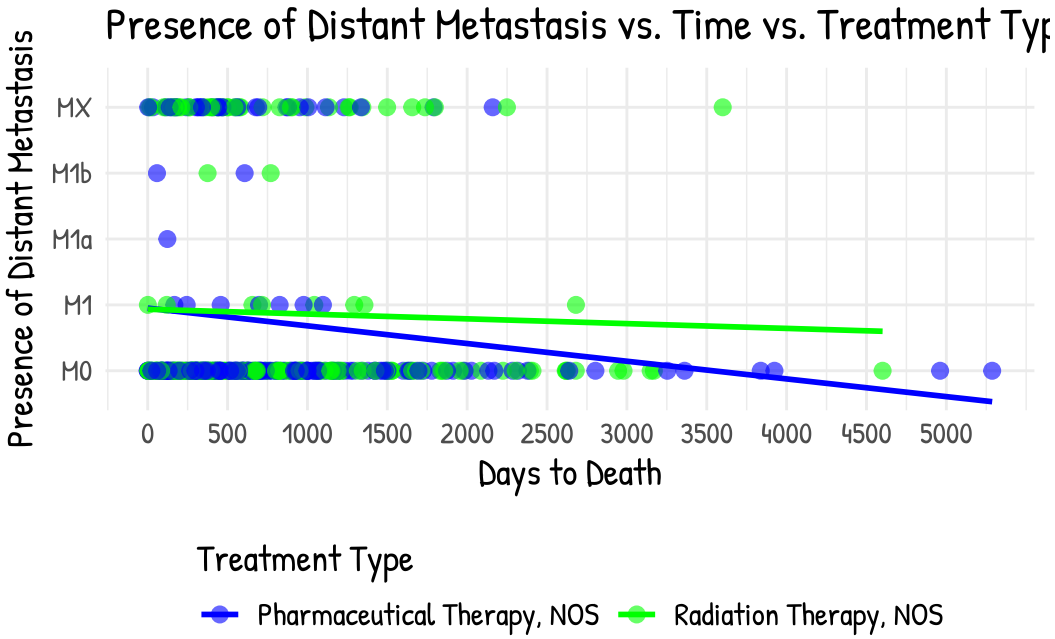

Figure 5: The scatterplot illustrates the relationship between distant metastasis, time to death, and treatment type in lung cancer patients post-diagnosis.

4.3 Treatment Type by Lymph Node Involvement

This scatter plot Figure 6 explores the relationship between lymph node involvement, time to death, and treatment type among cancer patients. Like above, each dot symbolizes an individual patient, with colors indicating their treatment type; pharmaceutical therapy or radiation therapy. Days to death is on the x-axis and lymph node involvement classifications are on the y-axis. At diagnosis, most patients are classified under either NO, N1 or N2. As survival time increases, there is an observable decline in lymph node involvement, reflected by the downward trend in the scatter plot. The trend lines further illuminate this pattern, with patients undergoing pharmaceutical therapy exhibiting a slightly steeper decline in lymph node involvement over time compared to those receiving radiation therapy, however this is very minimal and insignificant difference between the two treatment types.

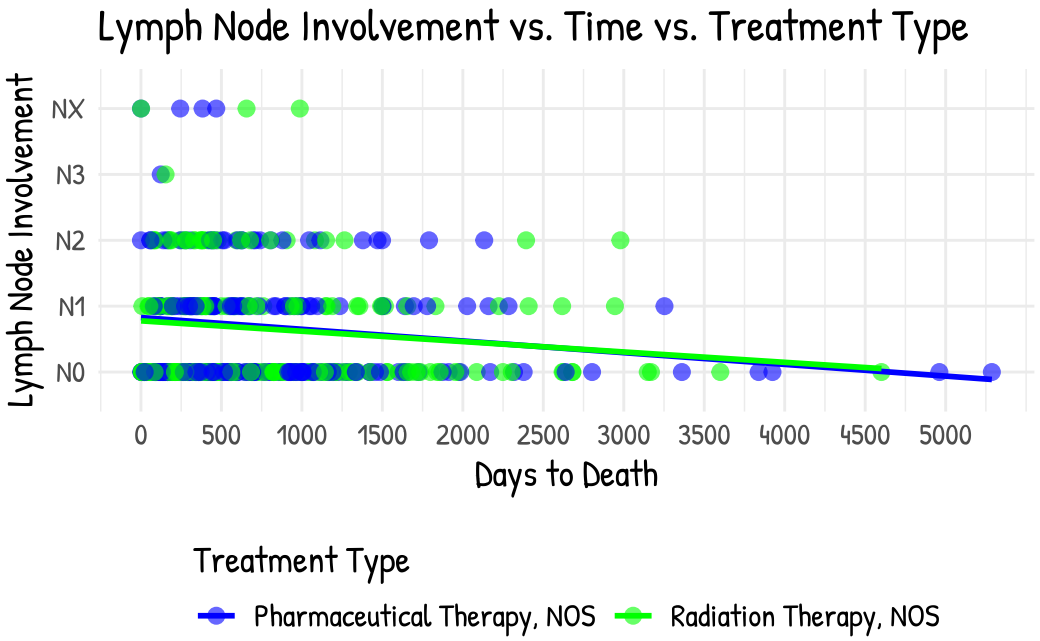

Figure 6: **?(caption)**

4.4 Treatment Type by Tumor Size

In this scatter plot, Figure 7 we investigate the relationship between tumor size, time to death, and treatment type for cancer patients. Each dot on the plot corresponds to an individual patient, categorized by their treatment regimen whether it is pharmaceutical therapy or radiation therapy. The x-axis represents the time from diagnosis until death and the y-axis is comprised of the tumor sizes. At diagnosis most patients are classified with a T2 tumour. As survival time increases, there is a noticeable reduction in tumor size, depicted by the downwards trajectory in the scatter plot. The trend lines further accentuate this trend, with patients undergoing radiation therapy demonstrating a more pronounced decrease in tumor size over time compared to those receiving pharmaceutical therapy.

5 Discussion

Lung cancer presents significant challenges in oncology, with survival rates influenced by various factors. This study aimed to elucidate the interplay between pathologic stage, treatment modalities, and survival outcomes in lung cancer patients post-diagnosis.

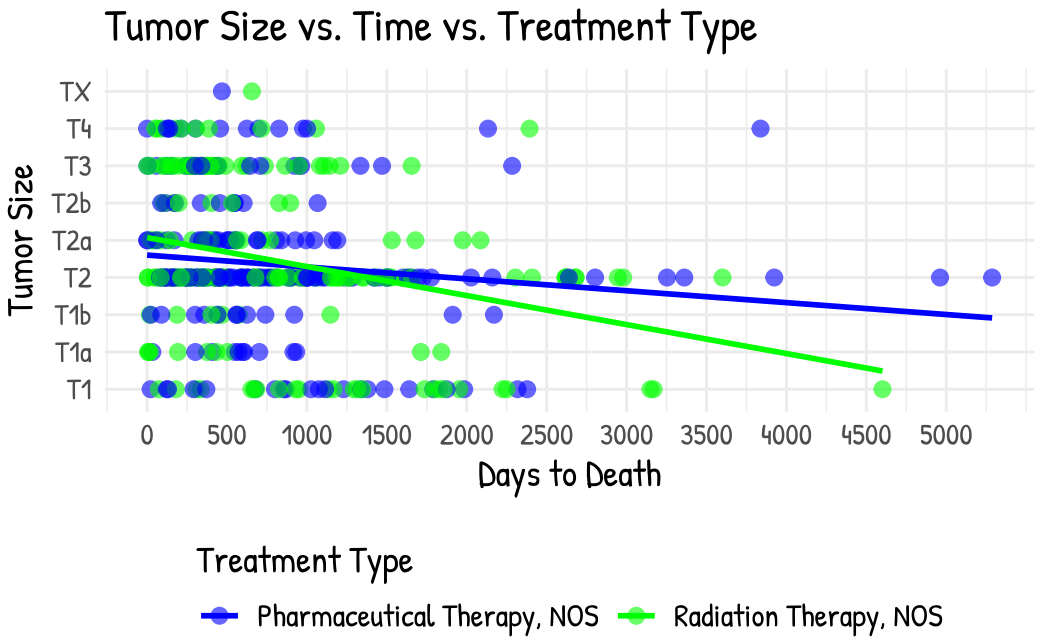

Figure 7: **?(caption)**

5.1 Survival Dynamics

Eldridge explains that the severity of lung cancer is staged by pathologic stage which is determined by all three presence of distant metastasis, lymph node involvement and tumor size (**Eldrige?**). Previous research found that patients with larger tumors and more extensive lymph node involvement are associated with a decrease in survival (Zhang et al. 2015). Additionally, greater presence of distant metastasis and thus advanced spread to other organs is shown to lower a patients survival (Markman 2023). Our analysis revealed compelling associations between various clinical factors and survival times in lung cancer patients. Upon diagnosis, lung cancer patients with more advanced pathologic stages (Figure 4), classification of presence of distant metastasis (Figure 5), higher greater involvement of lymph nodes (Figure 6) and larger tumors (Figure 7) were all shown to be associated with a shorter survival time.

5.2 Treatment Efficacy

An intriguing finding from our study was the differential impact of treatment modalities on survival outcomes. It was observed that Figure 4 which looked at pathologic stage, demonstrates that patients receiving pharmaceutical therapy tend to exhibit a slower progression in pathologic stage over time, thus linked to having greater survival time compared to those

undergoing radiation therapy. As pathologic stage is determined based on presence of distant metastasis, lymph node involvement and tumor size, this means that the main finding is that pharmaceutical therapy is the more effective treatment. In looking at these variables separately however, we see that not all the variables indicate that pharmaceutical therapy is superior.

Results looking at tumor size Figure 7, pharmaceutical therapy is associated with slower tumor growth rate and thus pharmaceutical therapy is more potent at treating lung cancer patients, aligning with the results of pathologic stage Figure 4. On the other hand, the plateaued trend line for radiation therapy compared to the sharper decline of pharmaceutical therapy in Figure 5 of presence of distant metastasis, this portrays radiation therapy was more effective in prolonging survival. Finally, with lymph node involvement Figure 6, pharmaceutical therapy and radiation therapy are both equally associated with increasing survival time based on lymph node involvement. These nuanced insights underscore the need for tailored treatment strategies based on individual patient characteristics and disease profiles.

5.3 Limitations

Despite the insights gained from our study, several limitations warrant acknowledgment. As mentioned earlier, the data set included 981 patients and after removing patients with incomplete data, we were left with 382 patients. This is a much smaller sample and thus there is a chance of bias and impact validity of our findings. Another area of weakness could be that values for 'MX' presence of distant metastasis and 'NX" lymph node involvement which indicate cases that could not be assed, these were left in the data set as to not decrease the sample size anymore. It likely that had these been taken out, the results would be different. We made the executive decision to include these values as reasons for inability to be assessed could be because the spread was too small to be detected. Additionally, our study did not account for certain variables such as performance status, smoking status, and genetic mutations, which may influence treatment responses and survival outcomes. Another limitation is that this study treatment types were pharmaceutical and radiation treatment individually. In previous studies, pharmaceutical treatment is often used in combination with radiation therapy to treat lung cancer as its shown better results in destroying cancer cells. In some patients, pharmaceutical therapy ensures the tumor remains small in size and the radiation destroys the cancer.

5.4 Future Directions

The dataset we used also offered information regarding demographics such as race, gender and year of birth. It would have been interesting to explore whether the treatment type and success would have been affected by these demographics. For example, radiation oncologists are looking into gender-dependent radiation therapies and are finding that female patients are more curative however they portray more side effects. Future research efforts should aim

to address these gaps and explore the role of novel treatment modalities, biomarkers, and personalized medicine approaches in improving lung cancer prognosis.

6 Conclusion

In conclusion, our study provides valuable insights into the complex factors influencing survival outcomes in lung cancer patients post-diagnosis. By elucidating the relationships between pathologic stage, treatment modalities, and survival times, we aim to inform clinical practice and guide future research endeavors. Although we established a general trend for those more likely to survive longer post-diagnosis, it's important to note that no two lung cancer patients are the same due to gene mutations present in the tumor and thus survival will vary person to person Eldridge (2022). Additionally, treatments are constantly improving and changing. Ultimately, our collective efforts must continue to focus on advancing early detection methods, refining treatment strategies, and enhancing the quality of care for lung cancer patients worldwide.

"words and phrases to avoid": enduring; landscape or evolving landscape; intricate; meticulous approach; nuanced; robust; pivotal; crafting; crucial; layered perspective. doing extensive/exhaustive statistical analysis/statistical endeavors on comprehensive/substantial data

A Appendix

A.0.1 Pathologic Stage

- stage is determined by tumor size, number of tumors found and where the cancer has spread Stage I: cancer is found early, is in one lung, has not spread to LN. Localized Stage II: found early but has spread inside or just ouside of one lung and may have spread to nearby LN on same side of the body. Regional spread. Stage III: cancer found in one lung and has spread to the chest and LN further further away from the lungs. Regional spread. Stage IV: advanced cancer and has spread to the lining or fluid around the heart or lungs, or has spread to distant parts of the body. Includes having tumors in both lungs, no matter the size of the tumors.

A.0.2 Tumor size

TX: size cannot be assessed - too small for testing T1: <3cm diameter, only in lungs T1a: <1cm, have not reached tissues surrounding lung/main branches of the airways T1b: 1-2cm, have not reached tissues surrounding lung/main branches of the airways T2: 3-5cm, grown into main bronchus/inner lining of lungs or caused lung collapse or swelling. T2a: 3-4cm, grown into main airways and/or tissue around the lungs T2b: 4-5cm, grown into main airways and/or tissue around the lungs T3: 5-7cm, growth beyond primary site, into chest wall/ nearby site or other tumors in the same lobe of the lung T4: >7cm, extensive tumor growth, invading nearby structures/organs. Growth into center of chest, to the diaphragm, heart or its major blood vessels, windpipe, esophagus, spine or other tumors in same lobe of the lung

A.0.3 Presence of distant metastasis

MX: not evaluated M0: no distant metastasis, cancer not spread beyond site of origin/ nearby LN. Localized. M1a: presence of metastasis to distant sites within the same lung/pleura. Regional. M1b: metastasis to distant sites beyong lung and pleura, ex. brain, liver, bones. Distant.

A.0.4 Lymph Node (LN) involvement

NX: LN cannot be assessed N0: no regional LN metastasis, cancer not spread to LN N1: involvement of regional LN N2: more extensive LN involvement

A.0.5 Treatment type

Pharmaceutical therapy Radiation Therapy (RT)

(AJCC, 2017) Stage I = T1 to T2a, N0, M0 Stage II = T1 to T3, N0 to N1, M0 Stage III = T1 to T4, N0 to N3, M0 Stage IV = Any T, Any N, M1s

References

Bauer, Paul C. et al. 2024. *Broom.mixed: Tidy Summaries of Mixed Effects Models*. https://CRAN.R-project.org/package=broom.mixed.

Chen, Sicong. 2023. "Lung Cancer Prediction." https://osf.io/qk34b/.

Denton, E J, D Hart, Z Wainer, G Wright, P A Russell, and M Conron. 2016. "Changing Trends in Diagnosis, Staging, Treatment and Survival in Lung Cancer: Comparison of Three Consecutive Cohorts in an Australian Lung Cancer Centre." *Internal Medicine Journal* 46 (8): 946–54. https://doi.org/10.1111/imj.13132.

Eldridge, Lynne. 2022. "Can Radiation Therapy Cure Lung Cancer?" https://www.verywellhealth.com/lung-cancer-radiation-success-rate-5209320.

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

Kai, Liu, Chen Zhoumiao, Xu Shaohua, Chen Zhao, Li Zhijun, He Zhengfu, and Cai Xiujun. 2021. "The Lymph Node Ratio Predicts Cancer-Specific Survival of Node-Positive Non-Small Cell Lung Cancer Patients: A Population-Based SEER Analysis." *Journal of Cardiothoracic Surgery* 16 (13). https://doi.org/10.1186/s13019-020-01390-x.

Markman, Maurie. 2023. "Lung Cancer Stages and Survival Rate." https://www.cancercenter.com/cancer-types/lung-cancer/stages.

Park, Jae Kil, Jae Jun Kim, Seok Whan Moon, and Kyo Young Lee. 2017. "Lymph Node Involvement According to Lung Adenocarcinoma Subtypes: Lymph Node Involvement Is Influenced by Lung Adenocarcinoma Subtypes." *Journal of Thoracic Disease* 9 (10): 3903–10. https://doi.org/10.21037/jtd.2017.08.132.

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.

Zhang, Jianjun, Kathryn A Gold, Heather Y Lin, Stephen G Swisher, Yan Xing, J Jack Lee, Edward S Kim, and William N William Jr. 2015. "Relationship Between Tumor Size and Survival in Non–Small-Cell Lung Cancer (NSCLC): An Analysis of the Surveillance, Epidemiology, and End Results (SEER) Registry." *Journal of Thoracic Oncology* 10 (4): 682–90. https://doi.org/10.1097/jto.0000000000000456.