

Survival Compass: Statistical Insights into Lung Cancer Patients' Journey Post-Diagnosis*

Lexi Knight

April 4, 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.

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.

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

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 ??, we visualize the exploration of variables constituting the pathogenic stage and treatment types. Section ??, outlines the model employed to analyze the relationship between these variables and the duration of survival post-diagnosis. Moreover, Section ?? offers visual depictions of the study's outcomes. Finally, in Section ??, we summarize the primary findings, propose avenues for enhancement, and identify potential areas for future research.

2 Data

Our data is

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

```
Warning: package 'showtext' was built under R version 4.2.3
```

```
Loading required package: sysfonts
```

```
Warning: package 'sysfonts' was built under R version 4.2.3
```

Loading required package: showtextdb

Warning: package 'showtextdb' was built under R version 4.2.3

Warning: package 'arrow' was built under R version 4.2.3

Attaching package: 'arrow'

The following object is masked from 'package:utils':

timestamp

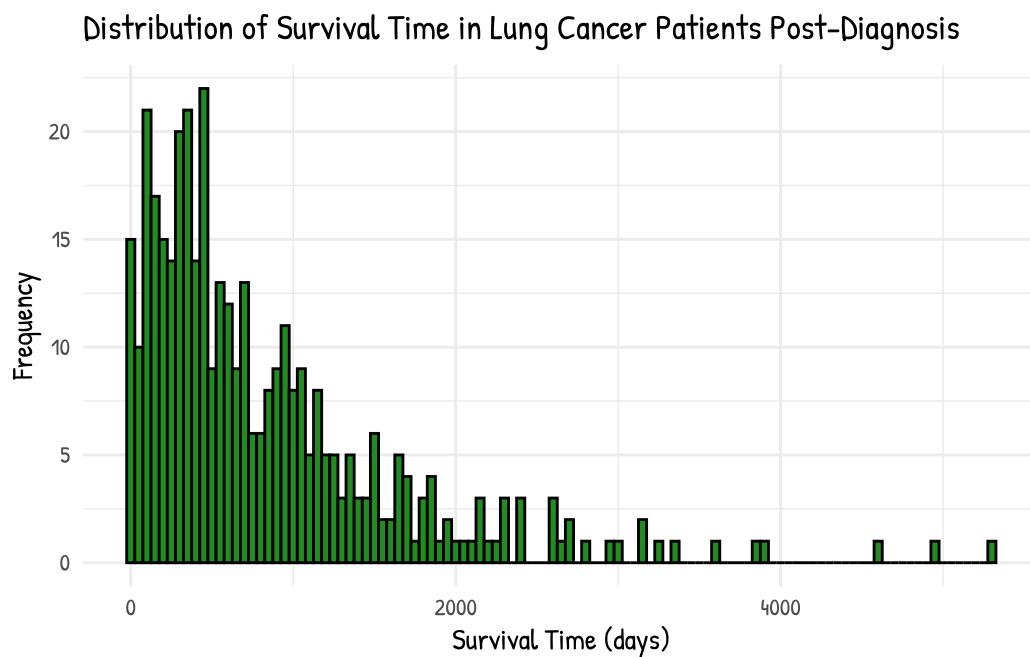


Figure 1: bb

Talk more about it.

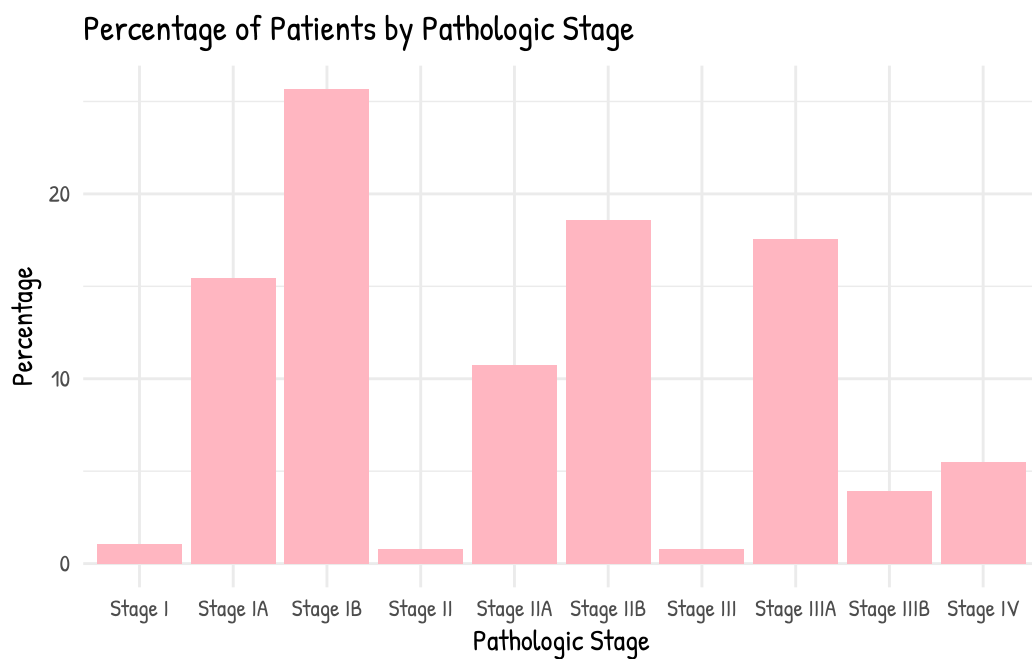


Figure 2: bb

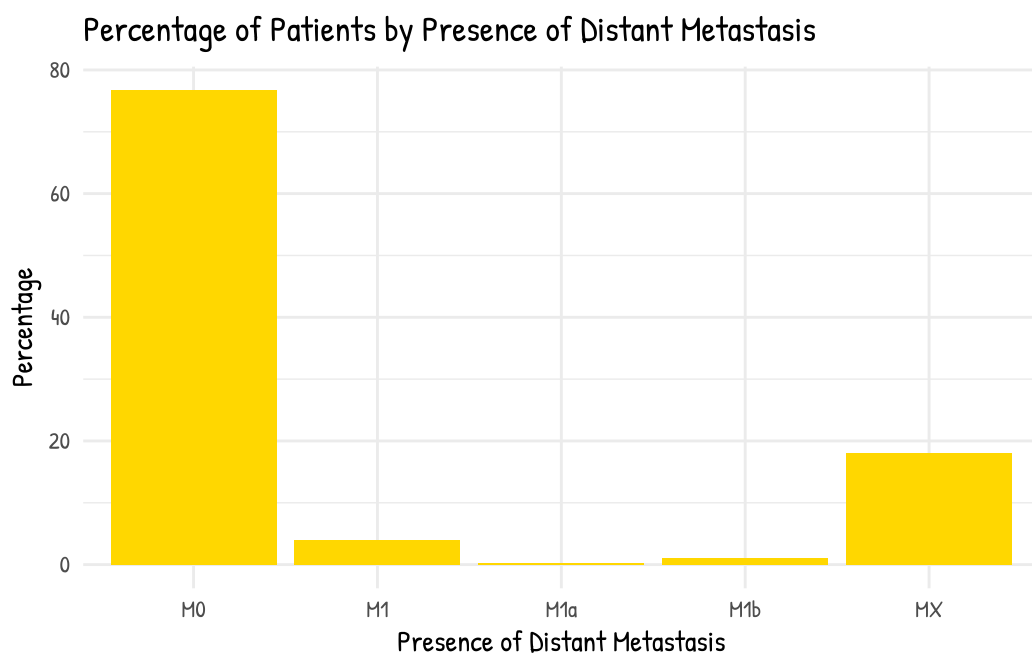


Figure 3: bb

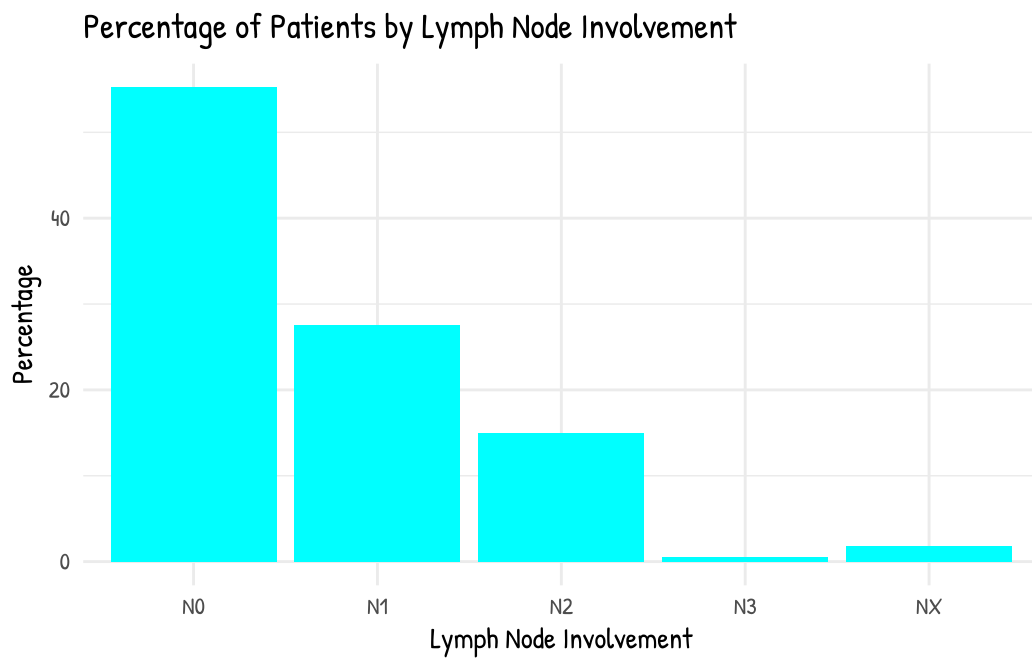


Figure 4: bb

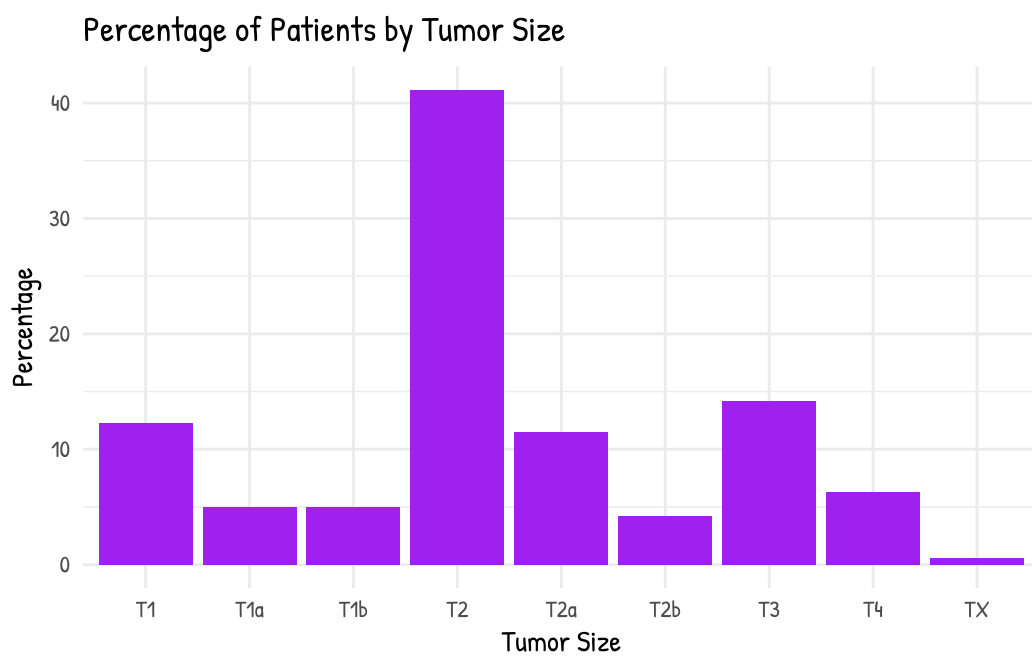


Figure 5: bb

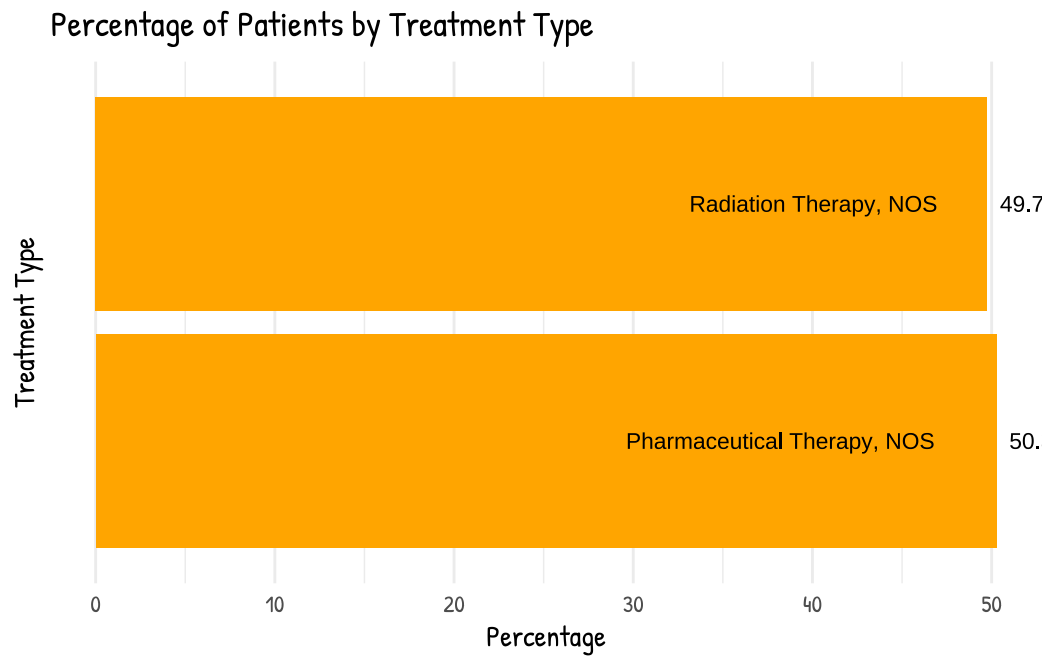


Figure 6: bb

3 Model

3.1 Linear Regression Model

- 4 interested in predicting days_to_death, a continuous outcome variable based on
- 5 multiple predictor variables (pathogenic_stage, lymph_node_involvement,
- 6 presence_of_distant_metastasis, tumor_size, treatment_decision, treatment_type)
- 7 Linear regression model are suitable for predicting continuous outcomes
- 8 quantifying relationship between predictor variables and outcome variable.
- 9 Fit a linear regression model to predict continuous outcome variable
- 10 'days_to_death'

10.1 Model set-up

In this study, we aim to investigate the survival outcomes of lung cancer patients post-diagnosis. Let y_i denote the number of days from diagnosis to death for patient i . We consider several predictors including pathogenic stage, lymph node involvement, presence of distant metastasis, tumor size, and treatment type.

We model the survival time y_i using a linear regression framework:

$$y_i | \mathbf{x}_i \sim \text{Normal}(\mu_i, \sigma^2)$$

where μ_i is the expected survival time for patient i , and σ represents the standard deviation of the survival times. The linear predictor μ_i is specified as: