Assignment 2: Lung Cancer Patients in Sweden

*Analysis of Survival Data/Survival Analysis (Joint Section) HT2025

Liang-Jen Huang
Department of Statistics, Uppsala University

Hadeel Elhassan
Department of Statistics, Uppsala University

I. INTRODUCTION

Lung cancer patients may receive different types of treatment depending on their medical condition and personal background. An important practical question is whether these treatments are associated with different chances of survival over time. In this study, not all patients die during the observation period, as some are still alive when follow-up ends. For this reason, simply comparing the number of deaths across treatments would not give a complete or accurate picture. Instead, the analysis considers both how long patients are followed and whether death occurs during that time.

The main aim of this analysis is to examine whether the risk of death over time differs between treatment options. Before making formal comparisons, we first explore the data to better understand the patient population. In particular, we examine whether key patient characteristics, such as age, stage of cancer, and smoking history, are similar across treatment groups, since these factors may influence both treatment decisions and survival outcomes. We also assess whether some patient characteristics convey similar information, so that the analysis remains clear and interpretable. These steps help ensure that any differences observed between treatments are meaningful and not driven by underlying differences in the patient groups.

II. DATA DESCRIPTION

The dataset used for this assignment is the *Lung Cancer Dataset*, obtained from Kaggle[1].

It is a large European dataset of patients diagnosed with lung cancer. The analysis focuses on 33,161 patients from Sweden, who were followed for different periods of time. At the end of the observation period, 7,165 patients were still alive, while 25,996 had died, which means that the outcome was not observed for approximately a fifth of patients during follow-up.

The dataset includes information on patient age, diagnosis and treatment timelines, cancer stage, smoking history, and treatment type, along with other relevant characteristics. These data allow for meaningful comparisons of survival outcomes across different treatment groups. A full description of the recorded information is provided in Table

A detailed description of all variables is provided in Table I.

TABLE I: Description of variables used in the analysis

Variable	Description
diagnosis_date	Date of lung cancer diagnosis.
end_treatment_date	Date of treatment end or death.
survived	Survival status (0 = death, 1 = alive).
cancer_stage	Stage of lung cancer (0 = Stage I, 1 = Stage II, 2 = Stage III, 3 = Stage IV).
gender	Gender of the patient (0 = Female, 1 = Male).
age	Age at diagnosis (years).
bmi	Body Mass Index (kg/m^2).
cholesterol_level	Cholesterol level (mg/dL).
smoking_status	Smoking behavior (0 = Never Smoked, 1 = Passive Smoker, 2 = Former Smoker, 3 = Current Smoker).
family_history	Family history of cancer (0 = No, 1 = Yes).
hypertension	High blood pressure (0 = No, 1 = Yes).
asthma	Asthma condition (0 = No, 1 = Yes).
cirrhosis	Liver cirrhosis (0 = No, 1 = Yes).
other_cancer	History of other cancers (0 = No, 1 = Yes).
treatment_type	Type of treatment received (0 = Radiation, 1 = Chemotherapy, 2 = Surgery, 3 = Combined).

III. DATA PREPROCESSING

Different patients have different diagnosis dates and end-of-treatment dates. To obtain a consistent measure of time, we calculated treatment days as follows:

$$\text{treatment\_days} = \text{end\_treatment\_date} - \text{diagnosis\_date}$$

This measure was used as the survival time in all subsequent analyses. The original dataset includes information on 33,161 patients. Due to computational limitations within the SAS environment, it was not feasible to work with the full dataset. To ensure that the analysis could be carried out efficiently, a random selection of 500 patients was used. The final processed dataset used in the analyses is available on GitHub.[2].

IV. DESCRIPTIVE ANALYSIS OF PATIENT CHARACTERISTICS

Before comparing survival outcomes across treatment types, we first explored the data using simple visual summaries to

[1]https://www.kaggle.com/datasets/khwaishsaxena/lung-cancer-dataset?resource=download

[2]https://github.com/edogawa-liang/LungCancer-Survival-Sweden/blob/main/data/Sweden_Lung_Cancer_500.csv

check whether patients receiving different treatments had similar background characteristics. This step is important because factors such as age, stage of cancer, and smoking history can influence both the choice of treatment and patient outcomes. If these factors differ across treatment groups, any observed differences in survival may reflect differences in the patients themselves rather than the treatments.

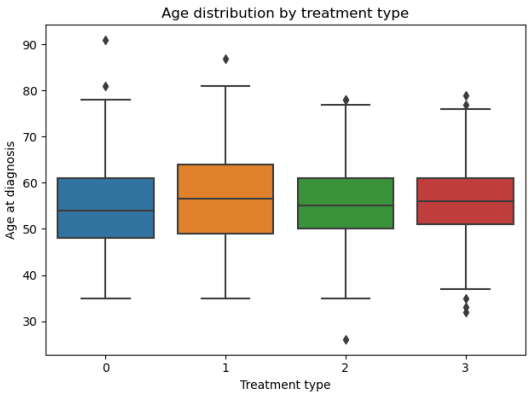


Fig. 1: Age distribution by treatment type

First, the relationship between age and type of treatment (Figure 1) shows that while the age ranges are broadly similar between treatment groups, there are noticeable differences in the typical ages and the overall spread. This suggests that some treatments are more commonly given to older or younger patients. Since age can influence both treatment decisions and patient outcomes, it should be considered when comparing survival across treatments.

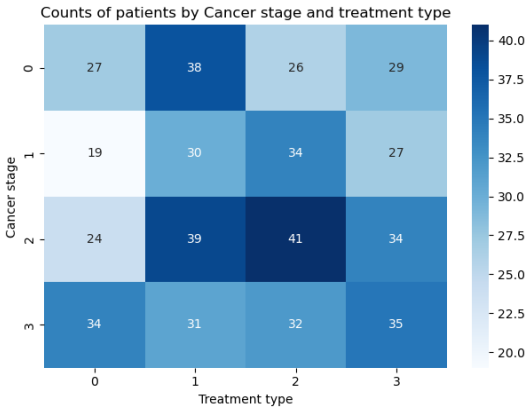


Fig. 2: Counts of patients by Cancer stage and treatment type

Next, the comparison between cancer stage and treatment type (Figure 2) shows that the mix of disease stages differs across treatment groups, although no single treatment is consistently linked to more advanced cases. This is clinically expected, as disease severity influences treatment choices. Because cancer stage is closely related to patient outcomes, it should be considered when comparing treatments fairly.

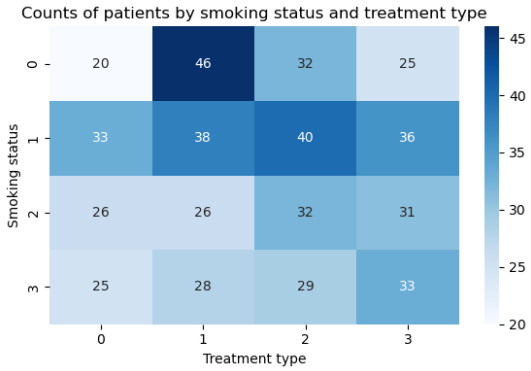


Fig. 3: Counts of patients by smoking status and treatment type

Moreover, the comparison between smoking history and treatment type (Figure 3) shows that smoking patterns differ across treatment groups. Smoking history is related to overall health and lung function, which can influence both treatment decisions and patient outcomes. For this reason, smoking history should be taken into account when comparing survival across treatments.

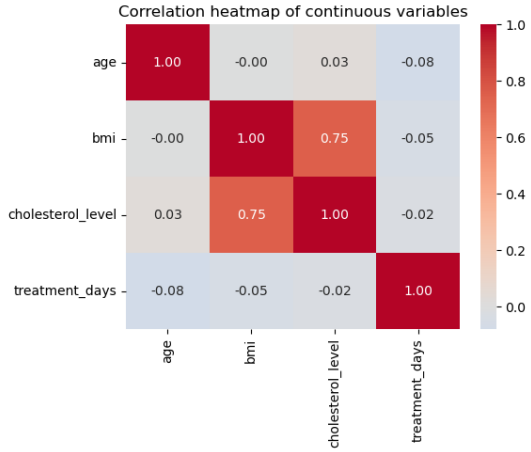


Fig. 4: Correlation heatmap of continuous variables

In addition to checking whether treatment groups were comparable, we examined whether some numerical patient characteristics were closely related to each other (Figure 4). Most of these measures showed little overlap, suggesting that they reflect different aspects of patient health. However, body mass index and cholesterol levels were strongly related, indicating that they provide similar information. This relationship should be considered to avoid unnecessary repetition in later analyses.

Finally, a simple comparison of survival across treatment types is shown in Figure 5 for background context. Although small differences in survival can be seen between treatments, these comparisons do not take into account differences in patient characteristics such as age, disease severity, or smoking

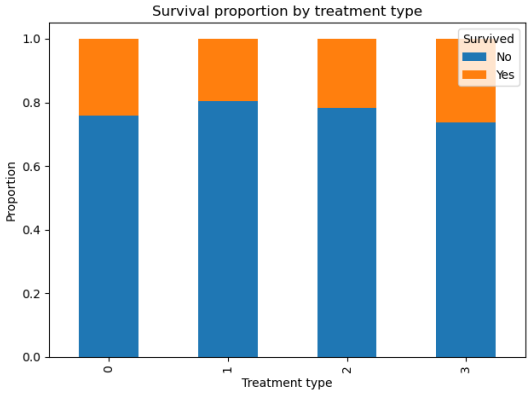


Fig. 5: Survival proportion by treatment type

history. As a result, these differences should be interpreted with caution and not viewed as direct effects of the treatments.

Some patient characteristics, such as gender and family history, are expected to play a more indirect role in treatment decisions and are therefore not shown here. However, they are still taken into account and discussed in the later stages of the analysis.

V. Analysis

A. Model Specification

Several statistical models were evaluated to understand which patient characteristics are associated with survival time. The final model was selected based on standard statistical criteria, balancing good model fit with simplicity and ensuring that key model assumptions were reasonably satisfied. Following the model selection process, gender, family history, smoking status, asthma, and other cancer were excluded due to a lack of evidence for a clear association with survival.

For completeness, we briefly present the mathematical formulation of the final model below. While the subsequent analysis and results can be understood without reference to the equations, the equations may help some readers better understand the model.

Let T_i denote the survival time (treatment_days) for individual i, and let $\mathbf{Z}_i$ denote the vector of covariates. We fitted a stratified Cox proportional hazards model with a time-dependent effect for hypertension.

$$\begin{aligned} h_i(t \mid \mathbf{Z}_i) = h_{0,\,\text{stage}(i)}(t) \exp\Big(&\beta_{\text{age}}\,\text{age}_i + \beta_{\text{cir}}\,\text{cirrhosis}_i \\ &+ \beta_{t1}\,\text{treatment}_{i1} + \beta_{t2}\,\text{treatment}_{i2} + \beta_{t3}\,\text{treatment}_{i3} \\ &+ \beta_{\text{ht}}\,\text{hypertension}_i + \beta_{\text{ht},t}\,\text{hypertension}_i \log t\Big), \end{aligned}$$

where

- $h_{0,\,\text{stage}(i)}(t)$ is the baseline hazard function specific to the cancer stage stratum of individual i;
- age_i and cirrhosis_i denote age and cirrhosis status;
- treatment_{i1}, treatment_{i2}, and treatment_{i3} are dummy variables representing treatment categories, with the reference category omitted;
- hypertension_i is an indicator of hypertension;
- $\beta_{\text{ht},t}$ captures a time-dependent effect of hypertension through the interaction with $\log t$.

Before presenting the model results, we first introduce an important quantity commonly used in survival analysis: the hazard ratio (HR). The hazard ratio describes the relative risk of experiencing the event at a given time. An HR greater than 1 indicates a higher risk, whereas an HR less than 1 indicates a lower risk.

B. Model Results and Interpretation

Overall, the model shows that age and hypertension are the most important factors associated with patient survival time, while differences between treatment types are relatively small after accounting for these patient characteristics. Importantly, the effect of hypertension is not constant over time, meaning that its impact on patient risk changes as time passes. In the analysis, patients are compared within the same cancer stage, allowing overall risk to differ between stages, while assuming that patient characteristics affect risk in a similar way across stages. In other words, factors such as age, hypertension, and treatment have the same type of impact on risk across different cancer stages.

Age. There is sufficient evidence to suggest that age is associated with survival. Each additional year of age increases the risk by approximately 1.2% (HR = 1.012), meaning that older patients tend to face a higher risk of the event compared with younger patients, assuming all other characteristics are the same.

Cirrhosis. Although patients with cirrhosis show a slightly higher risk (HR = 1.16), there is insufficient evidence to conclude that cirrhosis has a clear impact on survival after accounting for other factors. This suggests that the observed increase in risk may be due to random variation rather than a systematic effect.

Treatment type. After accounting for differences in patient characteristics and disease severity, there is no clear evidence of meaningful differences in survival between the treatment types. Overall, patients receiving different treatments appear to have similar survival outcomes.

Hypertension. There is sufficient evidence to suggest that hypertension is strongly associated with survival, and that its effect changes over time. Patients with hypertension have a substantially higher risk early in follow-up, but this excess risk decreases as time passes. For example, patients with hypertension have an estimated risk about 147 times higher after 7 days, 26 times higher after 30 days, and about 3 times higher after 180 days, compared with patients without hypertension. This suggests that hypertension is most critical early in follow-up, and its impact decreases over time. This indicates that hypertension is particularly critical in the early period, highlighting the importance of early monitoring and management.

Additional technical details are provided in the Appendix for reference.

C. Conclusion

Although this study was motivated by an interest in identifying differences in survival across treatment types, the analysis did not find strong evidence that treatment type substantially affects the risk of death over time once differences in patient characteristics and cancer stage are taken into account. In contrast, patient characteristics, particularly age and hypertension, play a more prominent role, with hypertension showing a clear time-dependent effect.

VI. Task 2

The extent to which explanatory variables can be changed to favour the hospital depends on the sensitivity of the issue and the potential consequences for patients, the public, and how the results are used. Selectively including variables only because they reduce differences in survival risks producing misleading conclusions and may negatively affect public health decisions.

According to the Royal Statistical Society's Code of Conduct, statisticians are required to act with integrity and objectivity and to indicate the risks and consequences if their conclusions are overruled. The hospital manager should therefore be informed of the ethical risks and moral responsibilities involved. Providing biased or selectively presented results would compromise professional integrity, for which statisticians are accountable.

Rather than altering the analysis to achieve a desired outcome, a more ethical approach would be to investigate the data more thoroughly, for example by examining data collection methods, assessing confounders, or considering factors such as patient comorbidities. Negative results remain valid and should be reported transparently.

Furthermore, the American Statistical Association's Ethical Guidelines emphasise that statisticians should avoid favoritism, resist attempts to influence analyses, and not condone statistical misconduct. Complying with the manager's request would therefore be inconsistent with professional ethical standards.

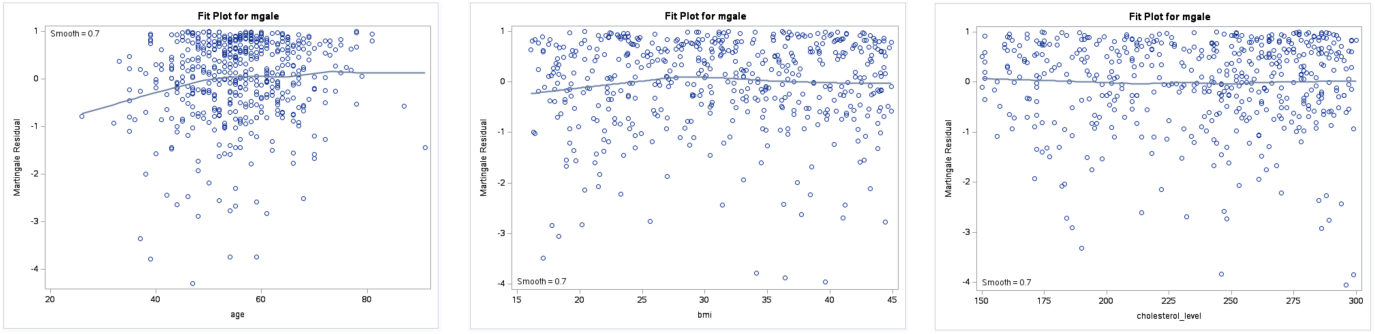
VII. Appendix

A. Analysis of Maximum Likelihood Estimates

TABLE II: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Std. Error	Chi-square	Pr > ChiSq	Hazard Ratio
age	1	0.01174	0.00542	4.7027	0.0301	1.012
cirrhosis	1	0.14538	0.12025	1.4617	0.2267	1.156
treatment_type1	1	0.02300	0.15047	0.0234	0.8785	1.023
treatment_type2	1	0.03152	0.15274	0.0426	0.8365	1.032
treatment_type3	1	-0.23874	0.15764	2.2938	0.1299	0.788
hypertension	1	7.32809	2.61368	7.8610	0.0051	1522.467
Int_hypertension	1	-1.20230	0.43021	7.8103	0.0052	0.301

B. Assessment of Functional Form for Continuous Covariates



(a) Plot of martingale residual versus age with LOWESS smooth

(b) Plot of martingale residual versus BMI with LOWESS smooth

(c) Plot of martingale residual versus cholesterol level with LOWESS smooth

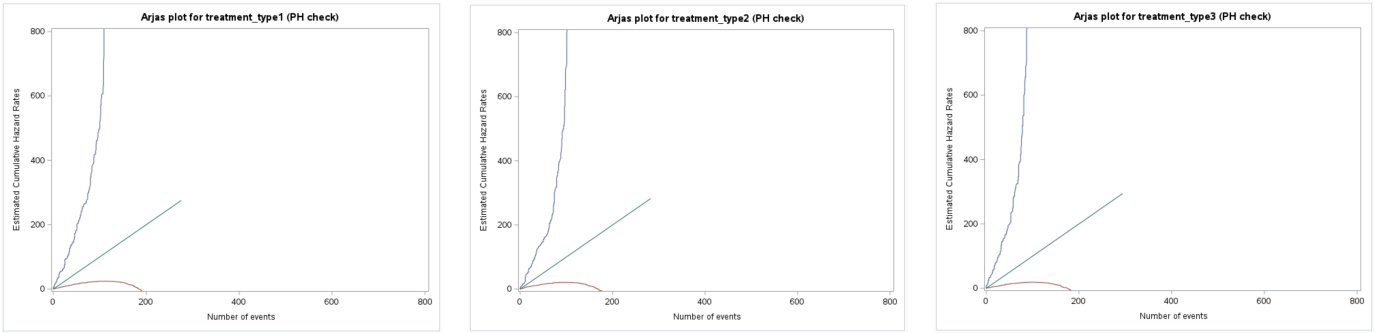
Fig. 6: Martingale residual plots versus continuous covariates with LOWESS smooth

To assess whether the continuous covariates were appropriately modeled using a linear functional form, Martingale residual plots were examined for each continuous variable. Specifically, we plotted Martingale residuals against BMI, age, and cholesterol level (Figure 6), together with a smoothed curve to highlight potential systematic patterns.

For all three covariates, the smoothed curves were approximately flat and showed no strong or systematic departures from linearity across the range of observed values. While minor local fluctuations were observed, these did not indicate a clear non-linear relationship that would justify the use of transformations or more flexible functional forms.

Based on these diagnostic plots, we concluded that modeling BMI, age, and cholesterol level as linear terms is adequate for the purposes of this analysis. Therefore, no transformations or spline terms were introduced in the final model.

C. Arjas Plot for checking PH assumption



(a) Treatment type 1

(b) Treatment type 2

(c) Treatment type 3

Fig. 7: Arjas plot for checking the proportional hazards assumption of treatment

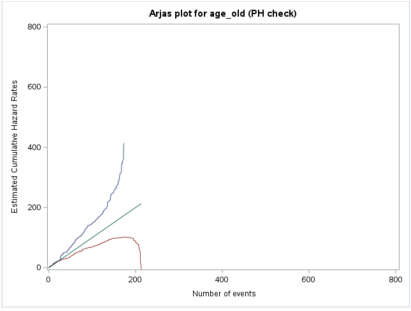


Fig. 8: Arjas plot for checking the proportional hazards assumption of Age

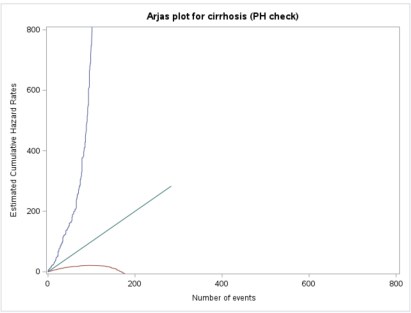


Fig. 9: Arjas plot for checking the proportional hazards assumption of Cirrhosis

The Arjas plots were constructed after stratifying by cancer stage and were fitted without including the time-dependent effect of hypertension. As shown in Figure 789, the plots display noticeable deviations from the reference line, which may suggest violations of the proportional hazards assumption. The corresponding covariates passed the initial time-dependent covariate tests, and the Cox–Snell residuals indicate that the overall model fit is reasonable. No further model modifications were made based on the Arjas plots alone.

D. Model diagnostics

To assess the overall fit of the final model, Cox–Snell residuals were examined. Since the model includes a time-dependent covariate (hypertension × log(time)), the Cox–Snell residuals were estimated excluding the time-dependent covariate, as recommended. Figure 10 shows the Nelson–Aalen estimate of the cumulative hazard plotted against the Cox–Snell residuals. The curve follows the 45-degree line reasonably well, indicating an adequate overall model fit.

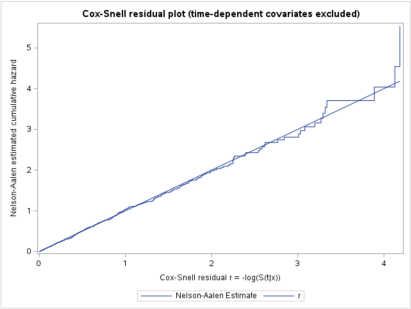


Fig. 10: Cox–Snell residuals to check the overall fit of the model

The generalized $R^2 = 0.0386$ suggests that the factors included in the model have a limited ability to fully explain why some patients live longer than others. In other words, although variables such as age and hypertension are related to survival, many other influences on patient outcomes are not captured in the available data.