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].

[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

IV. Descriptive analysis of patient characteristics

A. Survival Curves by Patient Characteristics

First, we look at the survival curves for our main variable of interest, treatment type, and also examine whether other patient characteristics are associated with differences in survival.

From Figure 1, we observe that at the beginning of the follow-up period, survival is quite similar across treatment groups. Over time, however, combined treatment appears to be associated with slightly better survival compared to the other treatments.

Next, Figure 2 presents the survival curves by smoking status. The curves are very close to each other throughout the follow-up period, suggesting no clear difference in survival between smoking groups.

In Figure 3, patients are divided into four age groups based on percentiles. Patients younger than 49 years tend to have better survival. However, among the older age groups, the survival curves overlap substantially and do not separate clearly, indicating limited differences in survival across these groups.

Figure 4 shows survival curves stratified by cancer stage. Here, we observe clear differences in survival, with patients diagnosed at earlier stages exhibiting substantially higher survival probabilities. This pattern is clinically intuitive.

Finally, Figure 5 displays the survival curves for hypertension status. The curves cross over time: patients with hypertension appear to have lower survival early on, but higher survival later in the follow-up period. This pattern suggests that the impact of hypertension on survival may change over time.

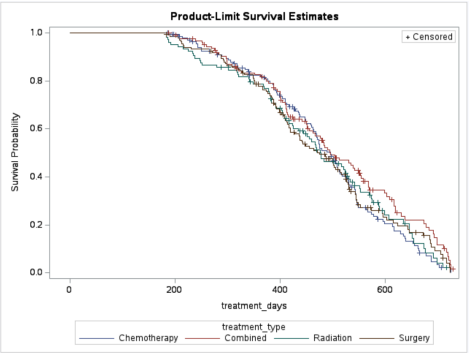


Fig. 1: Survival Curves by Treatment Type

B. Relationships Between Treatment and Patient Characteristics

Next, we examine how treatment type is related to other patient characteristics, to assess whether additional factors may influence treatment choice.

First, the relationship between age and type of treatment (Figure 6) 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

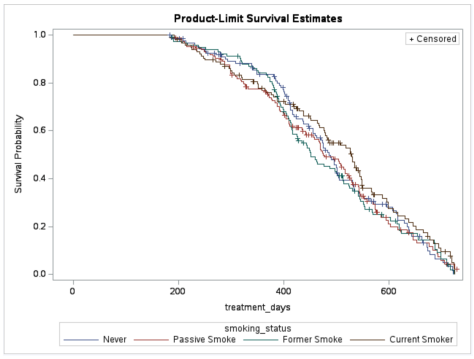


Fig. 2: Survival Curves by Smoking Status

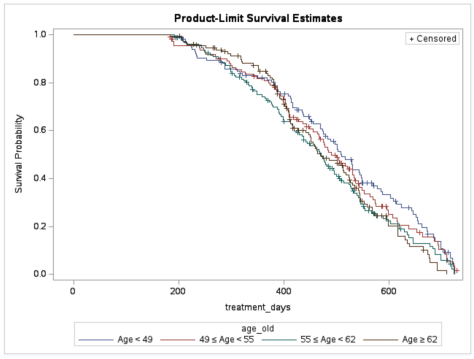


Fig. 3: Survival Curves by Age

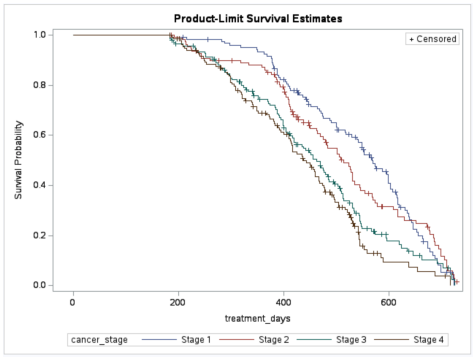


Fig. 4: Survival Curves by Cancer Type

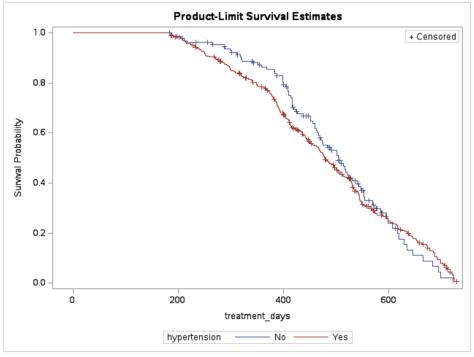


Fig. 5: Survival Curves by Hypertension

patients. Since age can influence both treatment decisions and patient outcomes, it should be considered when comparing survival across treatments.

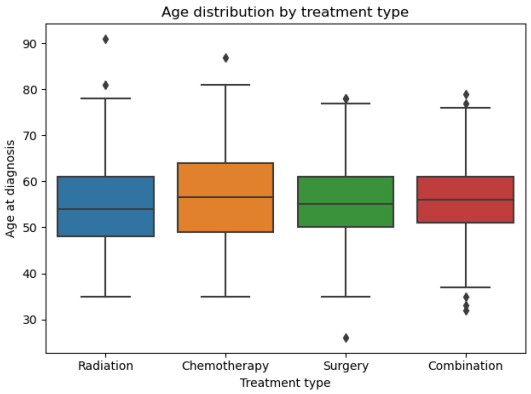


Fig. 6: Age distribution by treatment type

Next, the comparison between cancer stage and treatment type (Table II) 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.

TABLE II: Crosstable of cancer stage by treatment type

Cancer stage	Radiation	Chemotherapy	Surgery	Combination
0	27	38	26	29
1	19	30	34	27
2	24	39	41	34
3	34	31	32	35

Moreover, the comparison between smoking history and treatment type (Table III) 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.

TABLE III: Crosstable of smoking status by treatment type

Smoking status	Radiation	Chemotherapy	Surgery	Combination
Never	20	46	32	25
Passive	33	38	40	36
Former	26	26	32	31
Current	25	28	29	33

Overall, Figure 6 and Tables II and III show that treatment choice depends on patient characteristics. Different treatments are often given to patients at different cancer stages, and chemotherapy is more common among older patients. As a result, differences in survival between treatments may reflect differences in patient conditions rather than the treatment itself. For this reason, treatment will be examined together with other patient characteristics in the later analysis.

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. Estimated Risk Effects from the Analysis Results

Table IV shows how different patient characteristics are related to the risk of death. Each hazard ratio compares patients who differ in one characteristic, assuming all other characteristics are the same. Values greater than 1 indicate a higher risk of death compared to the reference group, while values less than 1 indicate a lower risk.

The associated p-values provide information on whether the observed effects are statistically meaningful, with smaller p-values indicating stronger evidence that the corresponding factor has a real impact on survival. In this analysis, a conventional threshold of 0.05 was used as a reference for deciding whether differences are considered meaningful.

TABLE IV: Summary of Risk Effects on Survival

Covariate	Hazard Ratio	p-value
Age	1.012	0.0301
Cirrhosis	1.156	0.2267
Chemotherapy v.s Radiation	1.023	0.8785
Surgery v.s Radiation	1.032	0.8365
Combined v.s Radiation	0.788	0.1299
Hypertension	Strong time-dependent effect	< 0.01

Overall, the analysis result 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.

B. Visual Comparison of Survival for a Typical Patient

In simple terms, we define a typical patient and keep this patient fixed. This patient is defined as 55 years old, without hypertension, without cirrhosis, and receiving radiation therapy. We then change one characteristic at a time while holding the others constant, so that the effect of each factor on survival can be viewed separately.

Age. When comparing different age groups (Figure 7), younger patients generally show higher survival probabilities over time, consistent with the finding that risk increases with age.

Cirrhosis. When comparing patients with and without cirrhosis (see Figure 8), those with cirrhosis generally show lower survival over time. Although the statistical analysis does not allow us to draw a firm conclusion, the survival curves still suggest that patients without cirrhosis tend to have slightly better survival than those with cirrhosis.

Hypertension. For hypertension, the survival curves show a more complex pattern (see Figure 9). Patients with hypertension tend to have lower survival early in follow-up, but later show higher survival compared to those without hypertension. This visual pattern indicates that the impact of hypertension on survival changes over time rather than remaining constant.

Treatment type. Finally, when comparing different treatment types while holding patient characteristics constant (see Figure 10), the survival curves are very close to each other. This suggests that, once other relevant patient factors are taken into account, the choice of treatment type makes only a small difference in survival.

VI. 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,

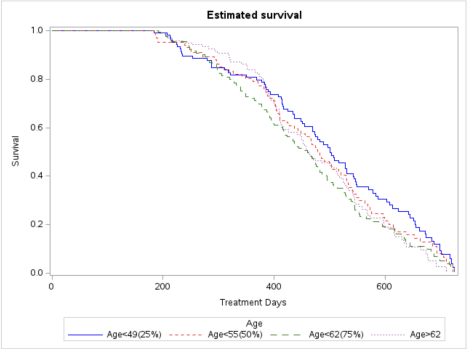


Fig. 7: Estimated Survival by Age

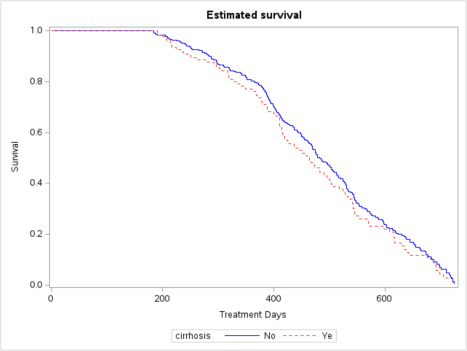


Fig. 8: Estimated Survival by Cirrhosis

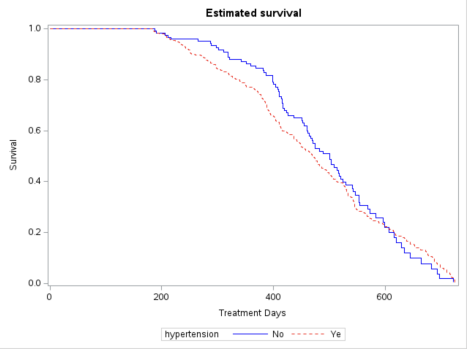


Fig. 9: Estimated Survival by Hypertension

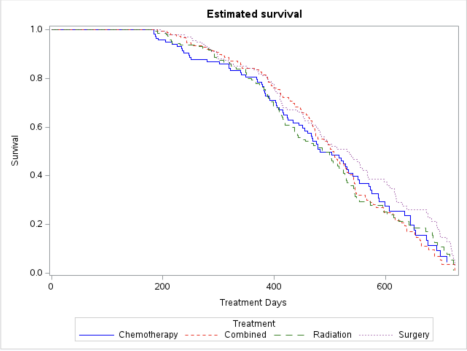


Fig. 10: Estimated Survival by Treatment

play a more prominent role, with hypertension showing a clear time-dependent effect.

VII. Task 2

The extent to which explanatory variables can be changed to favour a particular treatment rather than another depends on the sensitivity of the issue and the potential consequences for patients, the public, and how the results are used. The issue here is whether analytical choices are made to reflect the data objectively, or to make one treatment appear more effective than others. 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. In the context of this study, this means being transparent when comparing treatment effects, even if the results do not support a preferred interpretation. 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, such as making one treatment appear more favourable, 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.

VIII. Appendix

A. Investigate the correlation/multicollinearity among the possible covariates.

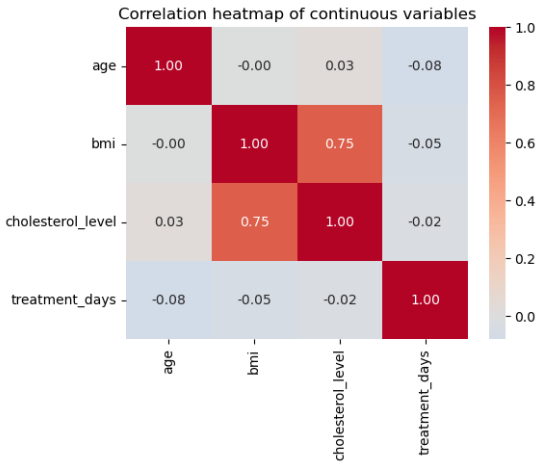


Fig. 11: Correlation heatmap of continuous variables

Before fitting the model, we examined whether some numerical patient characteristics were closely related to each other (Figure 11). BMI and cholesterol level showed a relatively strong correlation, suggesting that they provide similar information. However, since neither of these variables was retained in the final model after backward selection, this correlation does not affect the interpretation of the final model.

B. Analysis of Maximum Likelihood Estimates

Table V presents the maximum likelihood estimates from the stratified Cox proportional hazards model, including coefficient estimates, standard errors, test statistics, and the corresponding hazard ratios for each covariate.

TABLE V: Analysis of Maximum Likelihood Estimates

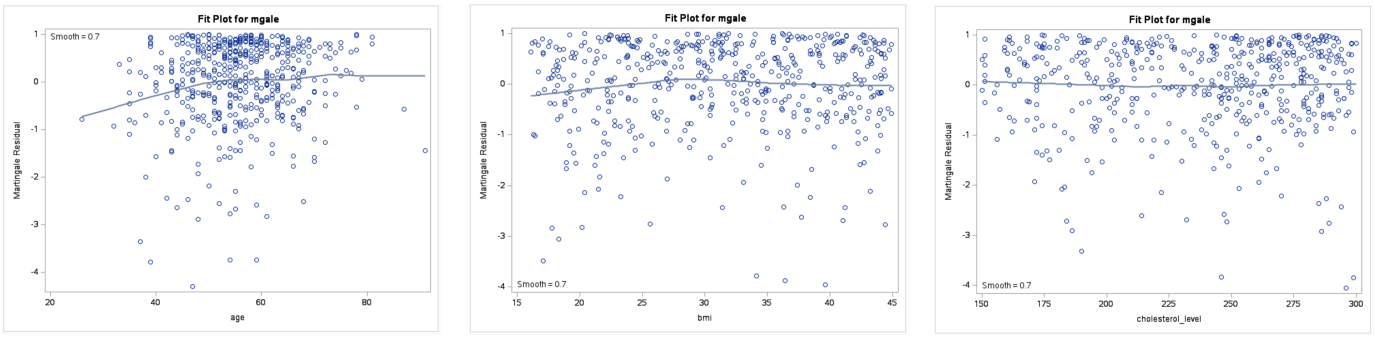
Parameter	DF	Estimate	Std. Error	Chi-square	Pr > ChiSq	Hazard Ratio
age	1	0.01174	0.00542	4.7027	0.0301	1.012
cirrhosis	1	0.14538	0.12025	1.4617	0.2267	1.156
treatment_type1	1	0.02300	0.15047	0.0234	0.8785	1.023
treatment_type2	1	0.03152	0.15274	0.0426	0.8365	1.032
treatment_type3	1	-0.23874	0.15764	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

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 $\log t$.



(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. 12: Martingale residual plots versus continuous covariates with LOWESS smooth

C. Assessment of Functional Form for Continuous Covariates

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 12), 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.

D. Arjas Plot for checking PH assumption

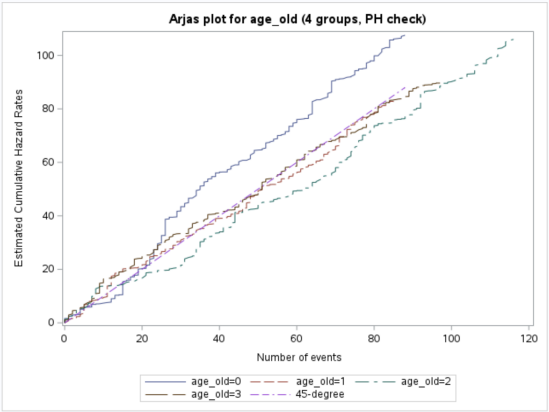


Fig. 13: Arjas plot for checking the proportional hazards assumption of Age

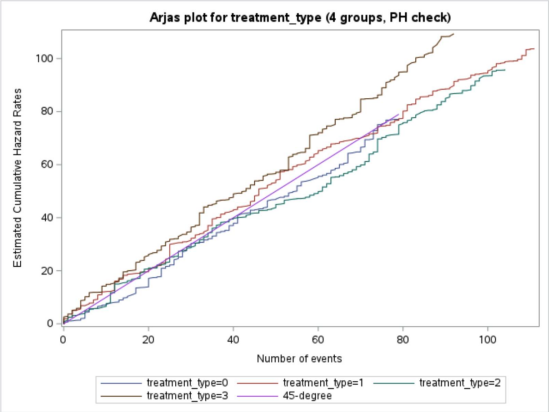


Fig. 14: Arjas plot for checking the proportional hazards assumption of Treatment

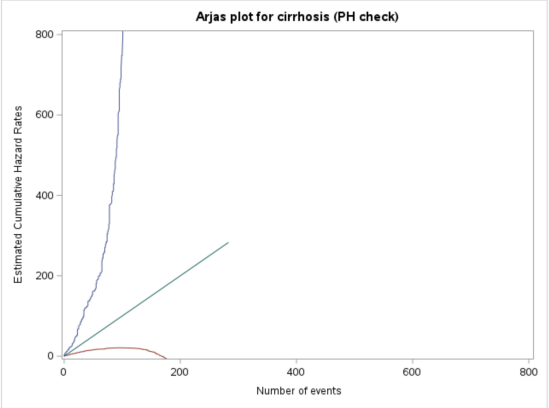


Fig. 15: Arjas plot for checking the proportional hazards assumption of Cirrhosis

The Arjas plots were constructed after stratifying by cancer stage and fitting the Cox model without including the time-dependent effect of hypertension. As shown in Figures 13–15, the plots were used to assess the proportional hazards assumption for age, treatment type, and cirrhosis.

For age and treatment type (Figures 13 and 14), the Arjas plots do not show clear non-linear patterns, indicating that the proportional hazards assumption is reasonably satisfied for these covariates.

For age (Figures 13), some age categories (e.g., $49 < \text{age} \leq 55$ and $55 < \text{age} \leq 62$) lie very close to the 45-degree reference line, indicating limited separation when age is categorized. In contrast, the youngest age group shows a more noticeable deviation, consistent with better survival among younger patients. As age is included as a continuous variable in the analysis, and no violation of the proportional hazards assumption is indicated by the Arjas plots, age is retained in the model as specified.

For treatment type (Figure 14), the Arjas plot shows that the curves for different treatment groups, especially treatment types 0 and 1, stay very close to the 45-degree reference line. This suggests that survival patterns across treatment groups are quite similar, and treatment type alone does not create large differences in risk over time. However, since the curves do not show clear non-linear patterns, there is no indication that treatment type violates the proportional hazards assumption. As treatment is the main variable of interest in this analysis, treatment type is therefore retained in the analysis.

For cirrhosis (Figures 15), the Arjas plot suggests that the proportional hazards assumption may not hold perfectly. In addition, the statistical evidence is not strong, so this result should be interpreted with caution.

E. 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 16 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.

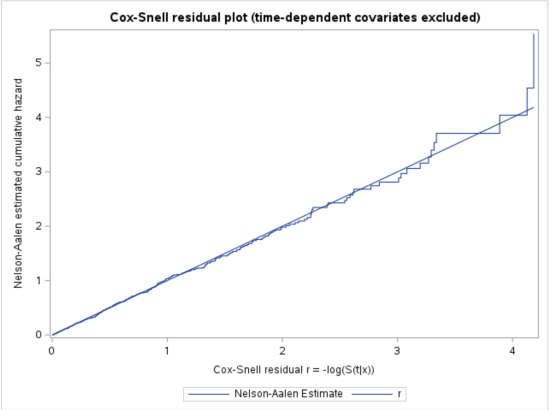


Fig. 16: Cox–Snell residuals to check the overall fit of the model

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.

Given that the Cox–Snell residuals suggest an adequate overall fit, we also considered a model without a time-dependent effect for hypertension. The model without time-dependent covariates has a generalized R^2 of 0.0217, whereas the model including time-dependent hypertension yields a slightly higher R^2. This suggests a small improvement in explanatory power when allowing the effect of hypertension to vary over time.