Assignment 1: 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. Data Description

The dataset used for this assignment is the *Lung Cancer Dataset*, obtained from Kaggle[1]. It contains information related to lung cancer mortality among the European population and provides a comprehensive collection of patient data, focusing on individuals who have been diagnosed with lung cancer.

After extracting the observations corresponding to Sweden, a total of 33161 observations were selected for analysis. Among these patients, 7165 were still alive at the end of treatment, while 25996 people had died. Hence, the degree of censoring is approximately 21.6%, representing the proportion of patients whose event (death) had not yet occurred by the end of observation.

The dataset consists of 14 explanatory variables and one outcome variable, `survived`. A detailed description of all variables is provided in Table I.

The variables currently included in the analysis are `diagnosis_date`, `end_treatment_date`, `survived`, `smoking_status`, and `treatment_type`. The diagnosis date and end treatment date are used to define the time-to-event variable, while smoking status and treatment type are the main variables of interest and stratification variables.

II. Data Preprocessing

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

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

This variable was used as the survival time in all subsequent analyses.

The original dataset contains 33,161 observations. Because the full dataset was too large for our SAS environment and caused computational issues when fitting survival models, we randomly selected 500 patients to create a manageable subset for the analysis. Among these patients, 386 experienced the event of interest (death), while 114 were still alive at the end of follow-up. 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

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. Task1

Our data set includes several variables with 2 to 4 groups. For task one, the variable treatment type was chosen, consisting of four groups (0 = Radiation, 1 = Chemotherapy, 2 = Surgery, 3 = Combined). Additionally, the variable survival has two groups (0 = death and 1 = survived). For analysis purposes, 0 = death was defined as the event, and therefore 1 = survived was treated as a censored observation.

The survival probabilities for time to death were estimated using the Kaplan–Meier (product-limit) estimator, and the resulting survival curves for the four treatment types are shown in Figure 1. The curves indicate that survival is broadly similar across treatment groups, with no clear separation between them.

Although both the Kaplan–Meier and Nelson–Aalen estimators can be used to estimate survival, the Kaplan–Meier estimator was chosen because the censoring rate in the data is moderate (approximately 23%) and the sample size is

relatively large (500 patients). Under these conditions, the bias of the Kaplan–Meier estimator under heavy censoring is not a major concern, and the Nelson–Aalen estimator does not offer clear advantages, as discussed by Bohoris (1994) [1]. In addition, the Kaplan–Meier estimator directly estimates survival probabilities, which are easier to interpret for descriptive analysis, a property commonly emphasized in applied survival studies (Ramadurai & Ponnuraja, 2011) [2].

To describe the uncertainty of the estimated survival curves over the entire follow-up period, Hall–Wellner confidence bands with a log–log transformation were used. The Hall–Wellner bands provide simultaneous confidence coverage for the entire Kaplan–Meier curve and allow the width of the bands to vary over time, reflecting changing uncertainty across the follow-up period.

The log–log transformation was chosen because it has been shown to provide good coverage properties and reasonably narrow confidence bands for the Kaplan–Meier estimator. In addition, the log–log transformation ensures that the confidence bands remain within the valid range of survival probabilities [0,1]. Given the relatively large sample size and moderate censoring in this study, this transformation provides reliable inference and is a commonly recommended choice for Kaplan–Meier confidence bands (Talsma, 2023) [3].

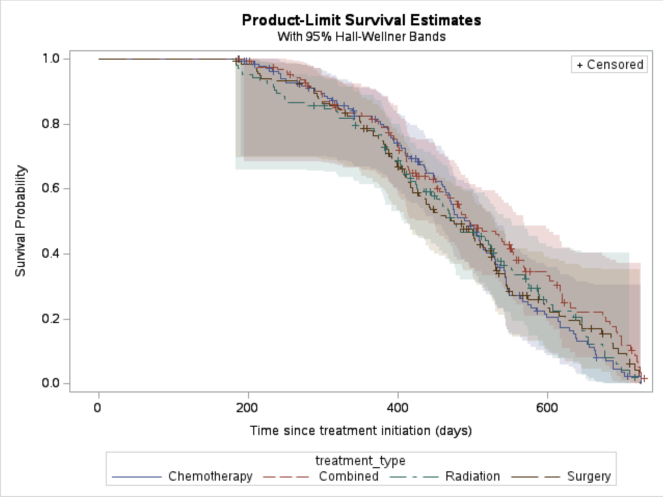

Fig. 1. Survival probabilities for different treatment types

To apply the Kaplan–Meier estimator, several assumptions need to be considered, including random sampling, independent observations, non-informative censoring, and right-censored data. Our data consist of right-censored survival times, as some patients were still alive at the end of follow-up. We randomly selected 500 patients from the original cohort of 33161 Swedish patients, which suggests that the random sampling assumption is likely reasonable, although this would ideally be confirmed with the data collectors. Similarly, the assumption of independent observations cannot be directly verified from the data and would also require confirmation from the data collectors or experts. The assumption of non-informative censoring cannot be formally tested. However, by plotting censored and uncensored observations, we can check for obvious censoring patterns. As shown in Figure 2, no clear censoring patterns are observed, suggesting that the non-informative censoring assumption is reasonable in this dataset.

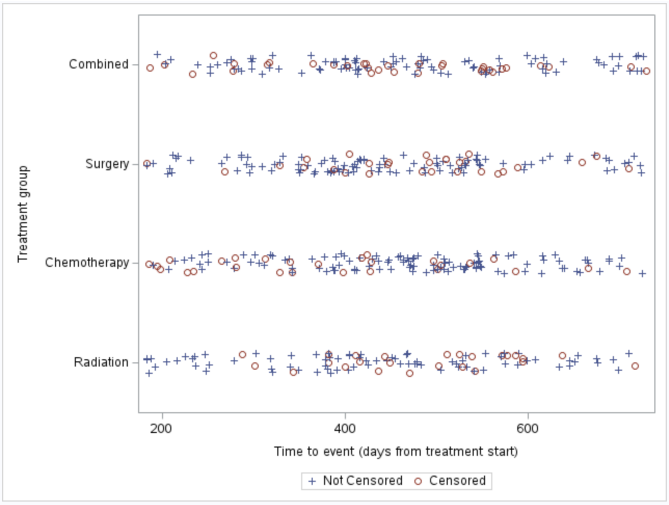

Fig. 2. Censoring Patterns by Treatment Type

To test whether the risk of experiencing the event differs between the treatment groups, the following hypotheses were formulated:

H_0 : The hazard rate is the same for all treatment groups over time,

H_1 : At least one treatment group has a different hazard rate at some time point.

In practice, the choice of the significance level should depend on the consequences of incorrectly rejecting the null hypothesis and is often discussed with experts. In this analysis, incorrectly rejecting the null hypothesis would mean claiming a survival difference between treatment groups when none exists. Since the practical impact of such an error is unclear, we use the commonly adopted significance level of 5%.

To compare survival between treatment groups, we apply the log-rank test. The choice of the log-rank test is motivated by the study focus and the observed pattern of the survival curves. As shown in Figure 1, the survival curves are similar across treatments, with only small differences appearing at later times. Because we are not specifically interested in early differences in risk, the log-rank test, which weights all event times equally, is an appropriate choice here.

From Table II, The log-rank p-value is 0.2402, which is greater than 0.05. Therefore, there is no evidence of a difference in survival between the treatment groups.

The median time to death, defined as the time point at which 50% of the patients in a group have experienced the event, was estimated for each treatment group using the Kaplan–Meier method. The corresponding 95% confidence intervals were constructed using a log–log-transformed approach, which pro-

TABLE II
TEST OF EQUALITY AMONG TREATMENT TYPES

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.2044	3	0.2402
Wilcoxon	2.1623	3	0.5394
-2Log(LR)	0.7185	3	0.8688

vides better coverage properties in the presence of censoring and ensures valid confidence limits.

The results, summarized in Table III, show no large differences in median time to death between the treatment groups. For example, for patients receiving Radiation therapy, the estimated median time to death is 476 days, with a 95% confidence interval of [425, 531] days. Similar interpretations apply to the other treatment groups.

Although treatment types 1 and 3 have slightly longer median times to death (496 days) compared to treatment types 0 and 2 (476 and 475 days, respectively), the confidence intervals overlap substantially, indicating that these differences are small and may be due to random variation rather than meaningful differences between treatments.

The confidence intervals are based on the same assumptions as the Kaplan–Meier estimator, including independent observations and non-informative right censoring. In addition, the construction of confidence intervals relies on large-sample theory, assuming asymptotic normality of the Kaplan–Meier estimator after appropriate transformation. Given the relatively large sample size in this study (500 patients), this assumption is considered reasonable and is therefore fulfilled.

TABLE III
QUARTILE ESTIMATES BY TREATMENT TYPE

		95% Confidence Interval	
Treatment type	Point Estimate	Lower	Upper
Radiation	476.000	425.000	531.000
Chemotherapy	496.000	464.000	514.000
Surgery	475.000	420.000	517.000
Combined	496.000	455.000	553.000

IV. TASK2

To ensure consistency with Task 1, the cumulative hazard curves were derived from the Kaplan–Meier estimator. Specifically, the survival function $\hat{S}(t)$ was estimated using the Kaplan–Meier (product-limit) estimator and transformed into a cumulative hazard function via $\hat{H}(t) = -\log\{\hat{S}(t)\}$. This approach ensures that the survival and cumulative hazard curves are based on the same estimator and differ only by a simple transformation. Given the relatively large sample size ($n = 500$) and a moderate censoring rate (114 out of 500 observations, approximately 23%), the Kaplan–Meier estimator provides stable and appropriate nonparametric estimates in this setting.

The cumulative hazard curves for the four treatment types were plotted to examine how the risk of experiencing death accumulates over time. The results presented in Figure 3 show that the cumulative hazard curves for the treatment groups remain close throughout the study period. This indicates that the risk accumulates in a comparable way between treatment groups, with no group showing a consistently faster or slower increase over time. Overall, the cumulative hazard patterns suggest that the underlying risk processes are broadly similar across treatments, which is consistent with the findings of the log-rank test.

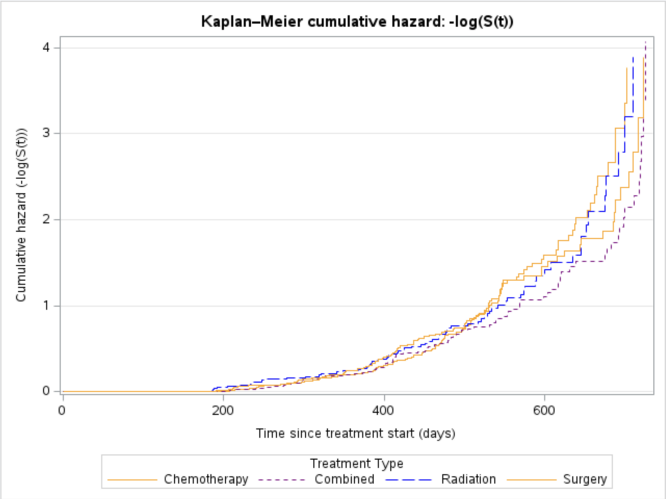

Fig. 3. The cumulative hazard for all treatment type groups

V. TASK3

In Task 3, a log-transformation of the time-to-event variable was applied. Based on this transformed time scale, we examined whether there are differences in time to event between the treatment groups. The results are presented in Table IV.

The log-rank test yields a p-value of 0.2402, which is not statistically significant at the 5% significance level. Since we do not have clear information about the consequences of treatment differences, a commonly used significance level of 0.05 was applied. In addition, Gehan's test, which places more weight on early differences, produces an even larger p-value. This indicates that even when early differences are emphasized, there is little evidence of differences between the treatment groups, which can also be observed in Figure 1.

These findings are consistent with the results obtained in Task 1. While log-transformation is often used in other modeling contexts to reduce the influence of large observation values, for rank-based survival tests the transformation mainly changes the time scale but not the ordering of event times. As a result, the test statistics and p-values remain unchanged, and no significant differences in survival time between treatment groups are observed.

VI. TASK4

Survival differences between treatment groups were assessed using a *stratified log-rank test*, with smoking status (Never Smoked, Passive Smoker, Former Smoker, Current Smoker) used as the stratification variable. The four compared groups correspond to the four treatment types.

TABLE IV
TEST OF EQUALITY AMONG TREATMENT TYPES OVER LOG(TIME)

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.2044	3	0.2402
Wilcoxon	2.1623	3	0.5394
-2Log(LR)	0.4704	3	0.9253

The hypotheses for the stratified tests (stratified by smoking status) are:

$$H_0: \ h_{1s}(t) = h_{2s}(t) = h_{3s}(t) = h_{4s}(t), \quad s = 1, \ldots, M, \ 0 < t < \tau,$$

H_a : At least one of the treatment–specific hazards

$$h_{ks}(t) \text{ differs for some } t < \tau.$$

Here, s indexes the strata defined by smoking status.

From Table V, we can see that the stratified test in Task 4 resulted in a p-value of 0.2371. Since this value is above the chosen significance level of 0.05, we do not reject the null hypothesis. This means that, after adjusting for smoking status, there is still no statistically significant difference in time to event between the treatment types.

TABLE V
TREATMENT GROUP COMPARISON (STRATIFIED BY SMOKING STATUS)

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.2358	3	0.2371
Wilcoxon	1.6294	3	0.6528

Compared with Task 1, the p-value in Task 4 is slightly smaller, indicating that stratifying on smoking status provides a minor improvement in the ability to detect group differences. However, the improvement is very small and the result remains non-significant. Therefore, stratification does not materially change the conclusion, and in this case it is not essential to include the stratifying variable.

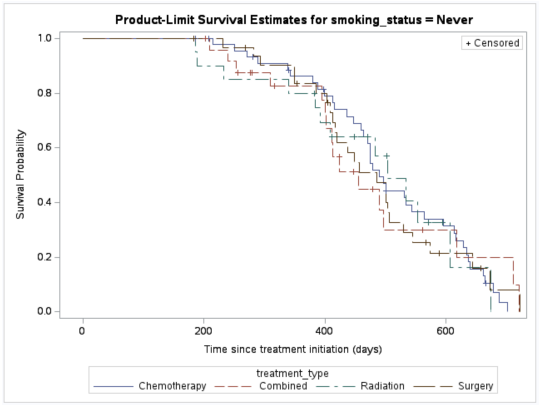

Fig. 4. Treatment Types (Never Smoked)

From Figures 4–7, we observe that across all four smoking-status strata, the survival curves for the different treatment types show substantial overlap. Although the exact shapes of the curves vary slightly between strata, no treatment group consistently demonstrates higher or lower survival across the smoking categories. In particular, the curves within each

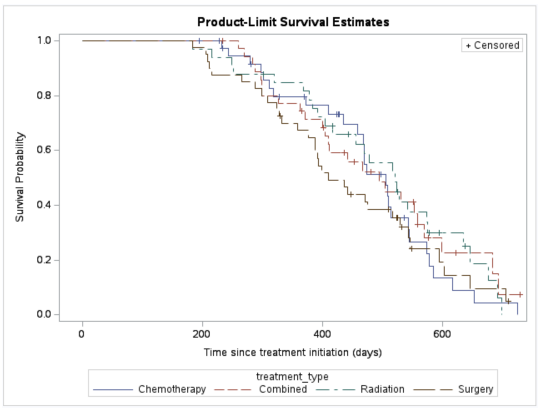

Fig. 5. Treatment Types (Passive Smoker)

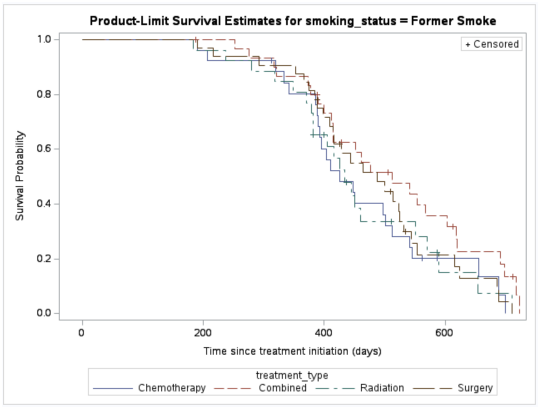

Fig. 6. Treatment Types (Former Smoker)

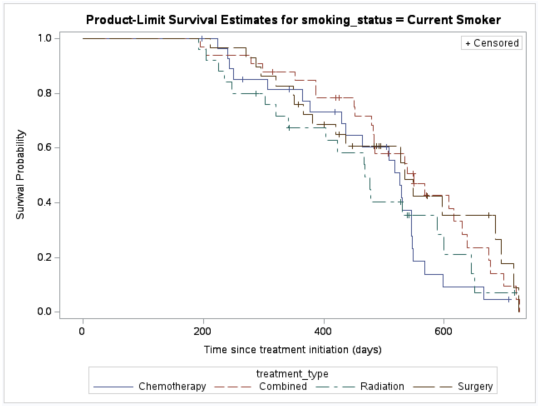

Fig. 7. Treatment Types (Current Smoker)

stratum remain very close to one another, and there is no clear indication that the treatment effects differ in direction across smoking status. Therefore, maintaining a stratified analysis by smoking status appears appropriate in this setting.

REFERENCES

[1] G. A. Bohoris, "Comparison of the cumulative-hazard and kaplan–meier estimators of the survivor function," *IEEE Transactions on Reliability*, vol. 43, no. 2, pp. 230–232, 1994.

[2] R. Ramadurai and C. Ponnuraja, "Non-parametric estimation of the survival probability of children affected by tb meningitis," *International Refereed Research Journal*, vol. 2, no. 2, pp. 216–228, 2011.

[3] J. Talsma, "Estimation of median survival time and its 95% confidence interval using SAS PROC LIFETEST," *Journal of Biopharmaceutical Statistics*, vol. 34, no. 3, pp. 366–378, 2023.