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`. All variables are currently included in the analysis. The time-related variables are `diagnosis_date`, `end_treatment_date`, and `age`. The main grouping variables include `cancer_stage`, `smoking_status`, and `treatment_type`. Other variables are treated as potential covariates; however, in further analyses, the roles of grouping variables and covariates may be interchanged depending on the analytical results. A detailed description of all variables is provided in Table I.

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 this sample size was too large for our SAS environment and caused computational issues when running survival models, we randomly selected 500 patients to create a manageable subset for the analysis. The final processed dataset used in the analyses is available on GitHub.[2].

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 the time to death were estimated using the Kaplan-Meier (product-limit) estimator, and the resulting plots are shown in Figure 1 comparing the four treament types. To use the Kaplan-Meier estimator, several assumptions must be considered.

The assumptions of random samples, independent samples, and right censored data must be discussed with the data collectors and the experts. Noninformative censoring cannot be tested directly and also requires consultation with the researchers. However, plotting the censored observations 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

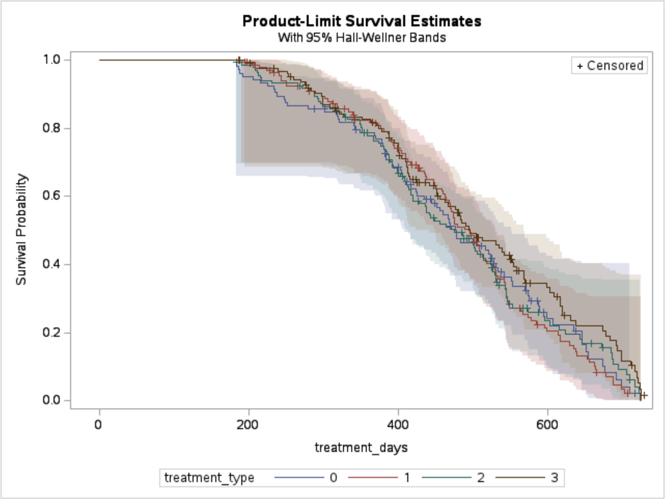


Fig. 1. survival probabilities for different treatment types

inspect for obvious censoring patterns can help support this discussion. Therefore, a censoring plot was created in Figure 2,and no clear censoring patterns appear in the data.

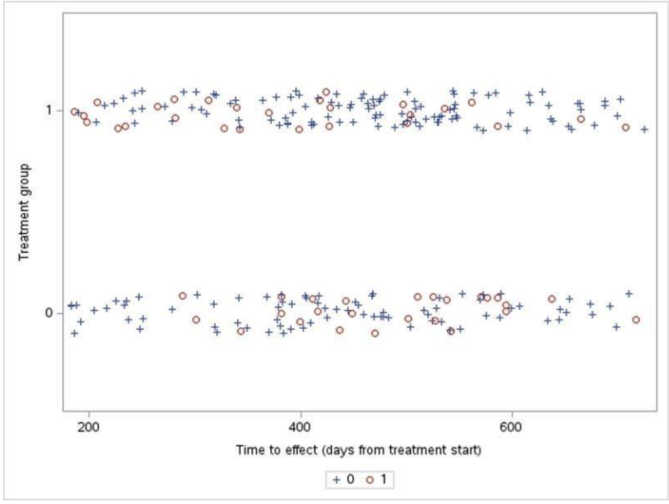


Fig. 2. Censoring plot

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

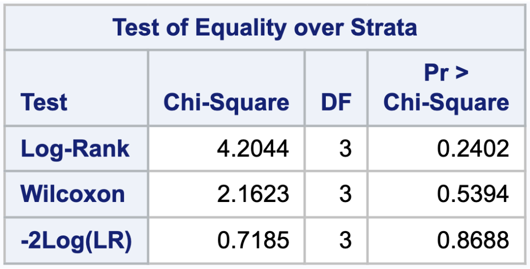
H_0 : The hazard rate is the same for all treatment groups over time,

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

The standard level is 5

The choice of significance level depends on the consequences of rejecting the null hypothesis when it is true. The standard level is 5% and this level is used here. To test the hypotheses,log-rank test was applied.

From Figure 3, The log-rank p-value is 0.2402, which is greater than 0.05, indicating that there is no evidence that the hazard rate differs between treatment groups. Thus, we cannot reject the null hypothesis.

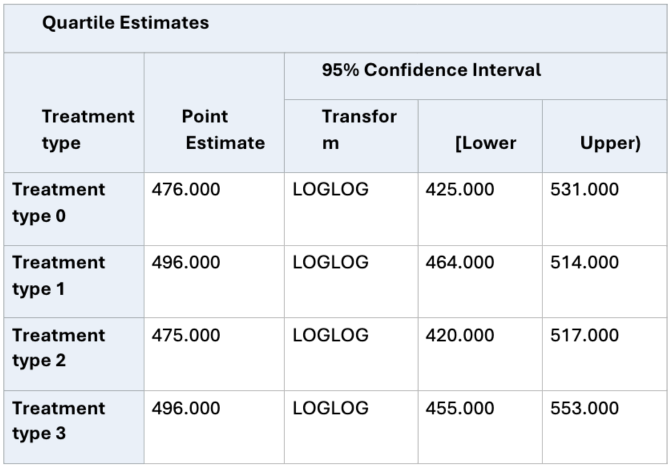


Test of Equality over Strata

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

Fig. 3. Test of equality

The median time to death estimates and the corresponding 95% confidence interval were extracted from the Kaplan-Meier results for each group and summarized in Figure 4 to make the comparison easier. There are no large differences between the median time to death between the groups. However, treatment types 1 and 3 had a slightly longer median times to death (496 days for both groups) compared to treatment type 0 and 1 (476 and 475 days respectively) indicating slightly better survival in treatment types 1 and 3.



Quartile Estimates

Treatment type	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
Treatment type 0	476.000	LOGLOG	425.000	531.000
Treatment type 1	496.000	LOGLOG	464.000	514.000
Treatment type 2	475.000	LOGLOG	420.000	517.000
Treatment type 3	496.000	LOGLOG	455.000	553.000

Fig. 4. The median time to experience death

IV. TASK2

The cumulative hazard curves for all four treatment types were also plotted to examine how the risk of experiencing death accumulates over time. The results presented in figure 5 5 show that the cumulative hazard curves for the treatment groups are close together throughout the study period. The risk of experiencing the event over time appears similar between the groups. This finding is consistent with the log-rank test, which showed no statistically significant differences between the treatment groups.

V. TASK3

To examine whether the log-transformed time-to-event differs across the treatment groups, we performed a one-way

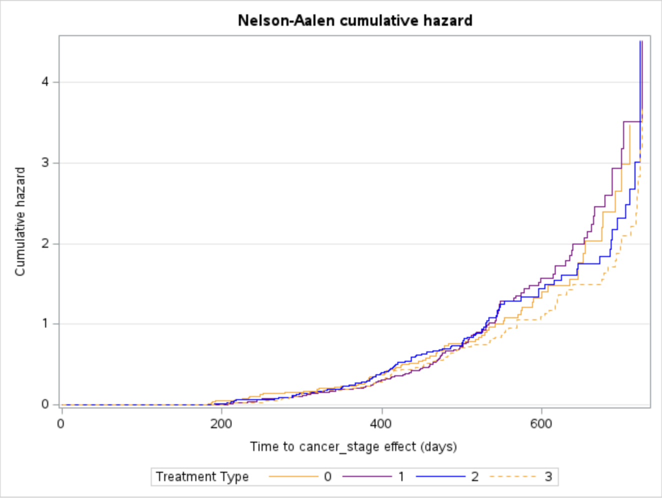
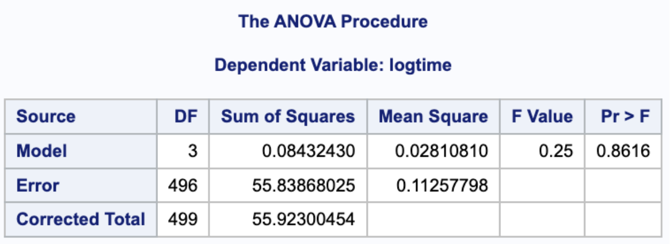


Fig. 5. the cumulative hazard for all treatment type groups

ANOVA with `treatment_type` as the grouping variable and `logtime` as the response.

The hypotheses for the ANOVA are:

$$H_0: \ \mu_0 = \mu_1 = \mu_2 = \mu_3,$$
$$H_1: \ \text{At least one group mean differs.}$$



The ANOVA Procedure

Dependent Variable: logtime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.08432430	0.02810810	0.25	0.8616
Error	496	55.83868025	0.11257798		
Corrected Total	499	55.92300454			

Fig. 6. ANOVA table

The ANOVA results were showed in Figure 6. In Task 1, the log-rank test gave a p-value of 0.2402. This is not statistically significant at the usual 0.05 level, but it is still much smaller than the ANOVA p-value in Task 3, which was 0.8616. This means that the survival analysis approach is ”closer” to detecting a difference between groups compared to the ANOVA.

The main reason for this is that survival analysis is a time-to-event method, and it properly handles censored observations. It keeps the information from individuals who did not experience the event during follow-up. ANOVA, on the other hand, treats all observed times as complete and ignores censoring entirely. Because of this loss of information, ANOVA has much less power to detect group differences in this kind of data.

VI. Task4

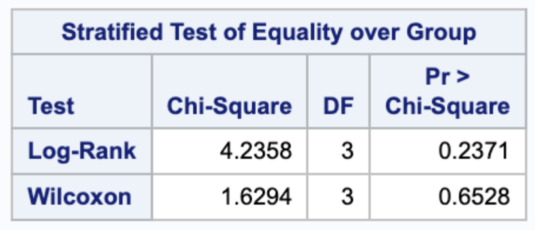
The hypotheses for the Stratified tests are:

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

$$H_a: \ \text{At least one of the treatment–specific hazards } h_{ks}(t) \text{ differs for some } t < \tau.$$

From Figure 7, we can see that the stratified test in Task 4 resulted in a p-value of 0.2371. Since this value is above the 0.05 significance level, 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.

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.



Stratified Test of Equality over Group

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

Fig. 7. Stratified Test of Equality over Group

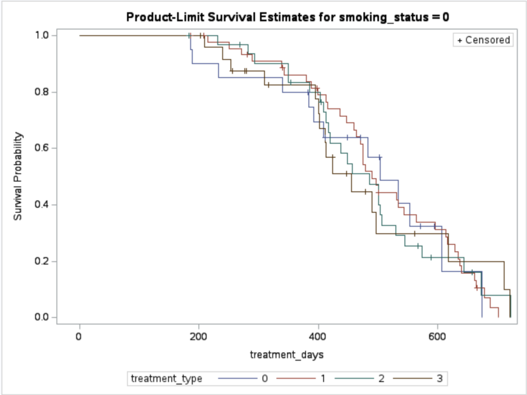


Fig. 8. Treatment Types (Never Smoked)

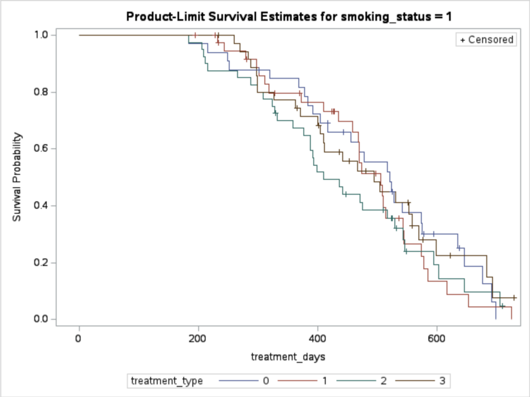


Fig. 9. Treatment Types (Passive Smoker)

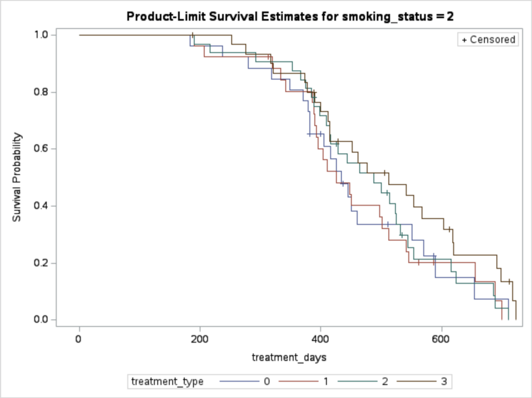


Fig. 10. Treatment Types (Former Smoker)

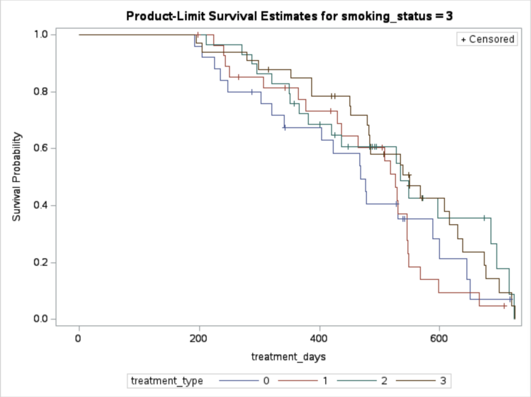


Fig. 11. Treatment Types (Current Smoker)

From Figures 891011, we observe that across all four smoking-status strata (0, 1, 2, and 3), 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.

Appendix A. SAS Code

```
proc import
    dbms=csv
    datafile=’/home/u64389050/Lung cancer dataset/lungdata.csv’
    out=lungdata
    replace;
run;

/* Group: treatment_type */

proc lifetest data=lungdata plots=survival (cb); time treatment_days * survived (1); * 0= die
    1=survived (censored)*/; strata treatment_type; run;


/* Censoring plot*/
ods graphics / reset attrpriority=none; /* Need to override the built-in attributes in proc
    sgplot */
proc sgplot data=lungdata; styleattrs / datasymbols=(plus circle) datacolors=(black red);
scatter x=treatment_days y=treatment_type /group=survived jitter;
xaxis label= "Time to effect (days from treatment start)";
yaxis integer label="Treatment group" min=-0.2 max=1.2;
format survived cens.;
label survived="00"x; /* Remove the label    censored    from x axis legend */
run;
```

Listing 1. SAS code for Task 1

```
proc lifetest data=lungdata method=breslow nelson; /*    nelson    option */
    time treatment_days * survived(1);  /* 0=die, 1=survived(censored) */
    strata treatment_type;
    ods output BreslowEstimates=lungdata_nelson;
run;

proc sgplot data=lungdata_nelson;
styleattrs datacontrastcolors=(orange purple blue)
datalinepatterns=(solid shortdash mediumdash);
step x=treatment_days y=cumhaz / group=treatment_type;
title "Nelson-Aalen cumulative hazard";
xaxis label="Time to cancer_stage effect (days)";
yaxis label="Cumulative hazard" grid;
label treatment_type="Treatment Type";
run;
```

Listing 2. SAS code for Task 2

```
data lungdata_logtime;
    set lungdata;
    logtime = log(treatment_days);
run;
proc anova data=lungdata_logtime;
    class treatment_type;
    model logtime = treatment_type;
run;
```

Listing 3. SAS code for Task 3

```
proc lifetest data=lungdata;
    time treatment_days * survived(1); /* 0=die, 1=survived(censored) */
    strata smoking_status / group=treatment_type;
run;
```

Listing 4. SAS code for Task 4