Survival Analysis Project

Karl Do Santos Zounon, Kateryna Draganova, Hanna Hellgren, Laura Lacombe

2025-03-16

Introduction

Cancer is one of the leading causes of mortality worldwide and understanding the factors that influence patient survival is crucial for improving treatment strategies. Survival analysis provides a powerful statistical framework for studying the time until a critical event, such as death or disease recurrence.

In this analysis, we explore the survival patterns of lung cancer patients using the survival package in R. Our goal is to understand how different factors, such as age, sex, and performance scores influence survival outcomes.

First, we load the necessary library and dataset. The *cancer* dataset contains information about cancer patients, including their survival time and various clinical factors.

```r
library(survival)
data(cancer)
head(cancer)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

```r
df <- cancer
```

The dataset contains the following columns:

- inst: Institution code
- time: Survival time in days
- status: Censoring status (1=censored: this means that the event (death) was not observed during the study period. The patient may still be alive at the end of the study, the patient may have been lost to follow-up, the patient may have died after the study period ended, 2=dead)
- age: Age in years
- sex: Male=1, Female=2
- ph.ecog: ECOG performance score as rated by the physician (0=asymptomatic, 1=symptomatic but completely ambulatory, 2=in bed <50% of the day, 3=in bed > 50% of the day but not bedbound, 4=bedbound)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient

- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in the last six months (pounds)

Exploratory Data Analysis

We begin by examining the dataset for missing values.

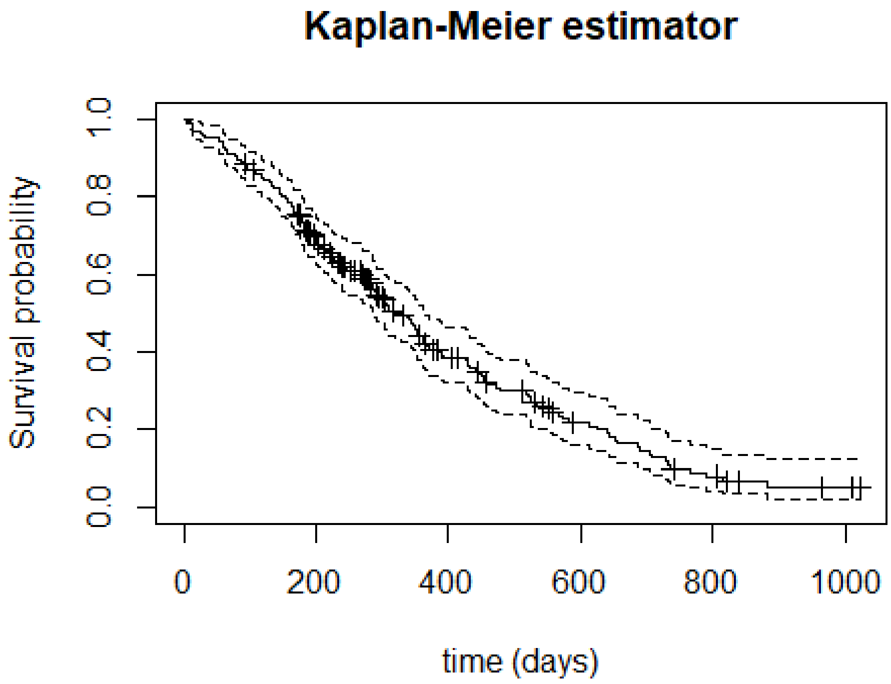
```
colSums(is.na(df))

##      inst      time    status       age       sex   ph.ecog  ph.karno pat.karno
##         1         0         0         0         0         1         1         3
##  meal.cal   wt.loss
##        47        14
```

To handle the missing values, we remove the rows with missing discrete data and input the average of meal.cal and wt.loss.

```
library(dplyr)

##
## 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

clean_data <- df %>%
  filter(!is.na(ph.ecog) & !is.na(ph.karno) & !is.na(pat.karno)) %>%
  mutate(
    meal.cal = ifelse(is.na(meal.cal), mean(meal.cal, na.rm = TRUE), meal.cal),
    wt.loss = ifelse(is.na(wt.loss), mean(wt.loss, na.rm = TRUE), wt.loss)
  )

colSums(is.na(clean_data))

##      inst      time    status       age       sex   ph.ecog  ph.karno pat.karno
##         1         0         0         0         0         0         0         0
##  meal.cal   wt.loss
##         0         0
```

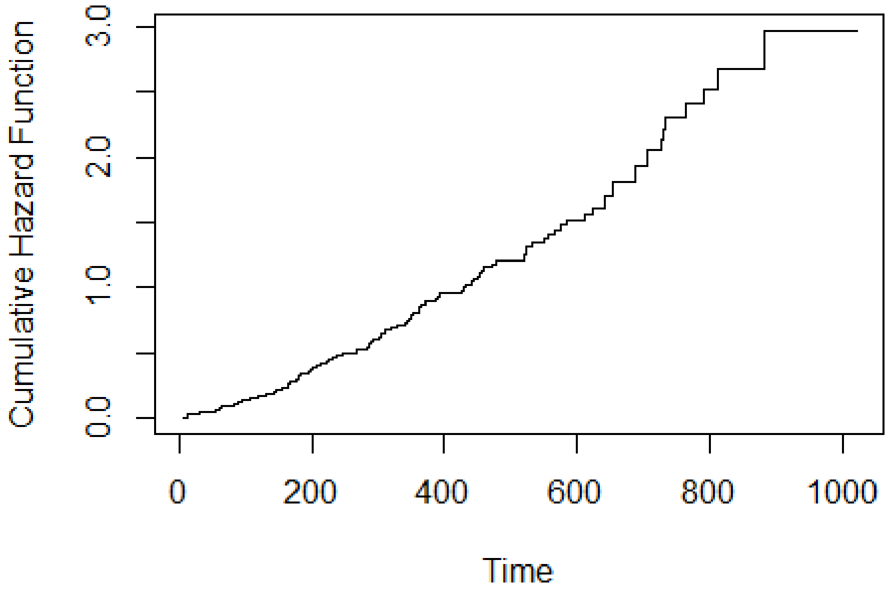
Now, that we have clean data we can make our first model. We can plot the survival function with the Kaplan-Meier estimator.

```
survival <- survfit(formula = Surv(time, status) ~ 1, data = clean_data)
survival

## Call: survfit(formula = Surv(time, status) ~ 1, data = clean_data)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 223    160    329     286     364

plot(survival, mark.time = TRUE,
     main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (days)")
```



We can plot the cumulative hazard function, which is the accumulated risk over time, with the Nelson-Aalen estimator.

```
cumul_hazard <- -log(survival$surv)

plot(survival$time, cumul_hazard, type = "s", xlab = "Time", ylab = "Cumulative Hazard
Function")
```



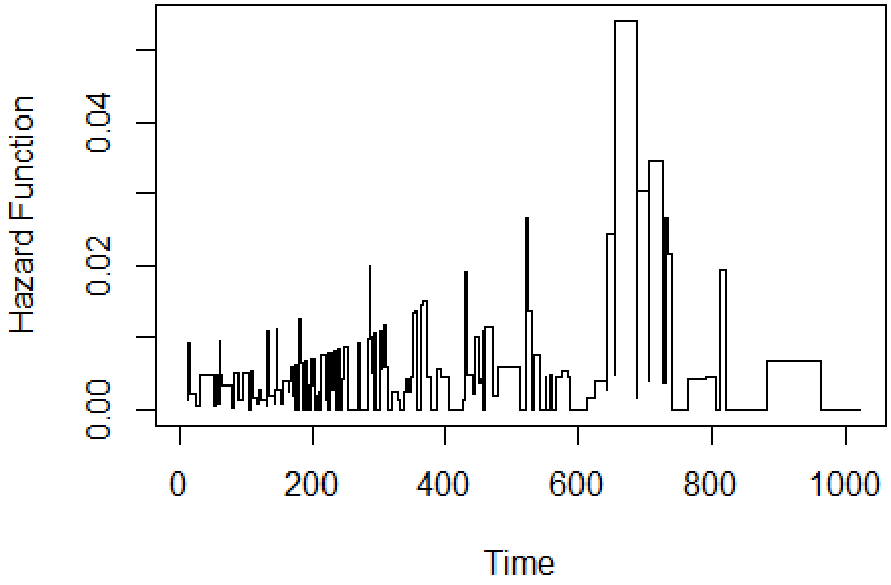
```
H1 <- survival$time[which.min(abs(cumul_hazard - 1))]

H1
```

```
## [1] 429
```

H(t)=1 is the time at which an individual is expected to have an event. So, the event is expected on the 429th day.

The hazard function h(t) is the rate of change of the cumulative hazard function H(t). It represents the instantaneous rate of failure at time t (or the occurrence of an event).

```
hazard <- diff(cumul_hazard) / diff(survival$time)

hazard <- c(NA, hazard)

plot(survival$time, hazard, type = "s", col = "black", xlab = "Time", ylab = "Hazard
Function")
```



Cox Proportional Hazard Model

```
cox_model <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data = clean_data)
summary(cox_model)

## Call:
## coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = clean_data)
##
##   n= 223, number of events= 160
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## age      0.009801  1.009849  0.009335  1.050  0.29377
## sex     -0.544009  0.580417  0.169549 -3.209  0.00133 **
## ph.ecog  0.460528  1.584910  0.115369  3.992 6.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        1.0098     0.9902    0.9915    1.0285
## sex        0.5804     1.7229    0.4163    0.8092
## ph.ecog    1.5849     0.6310    1.2642    1.9870
##
```

```
## Concordance= 0.635  (se = 0.025 )
## Likelihood ratio test= 28.66  on 3 df,   p=3e-06
## Wald test            = 28.24  on 3 df,   p=3e-06
## Score (logrank) test = 28.75  on 3 df,   p=3e-06
```

HR = 0.5804, meaning males have 42% lower hazard than females (because HR < 1). Significant (p = 0.00133) → Sex significantly affects survival.

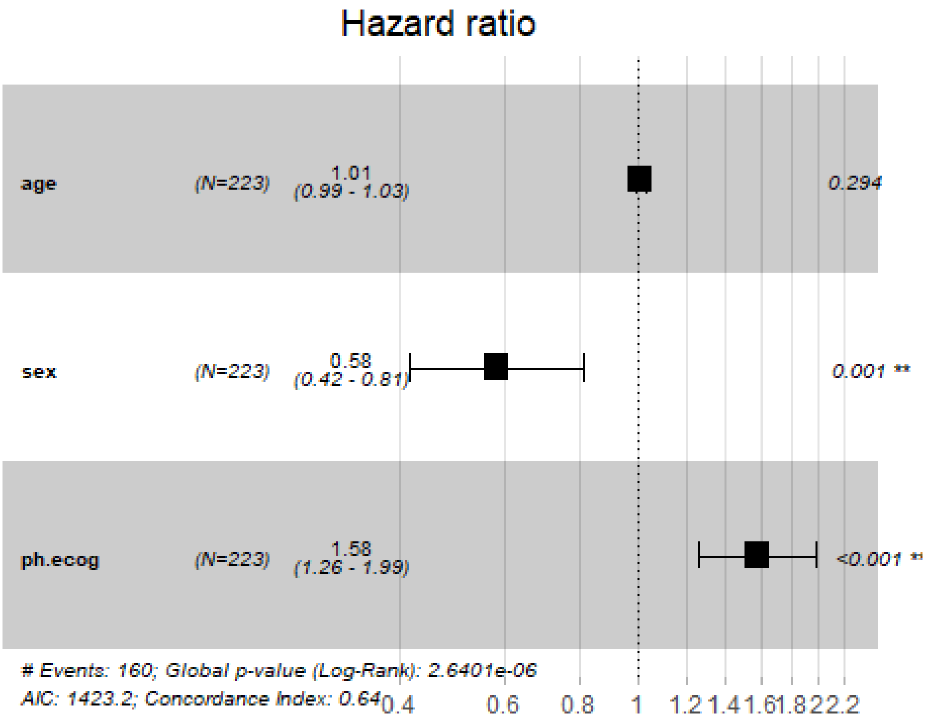
HR = 1.5849, meaning a higher ph.ecog score increases the hazard of death by 58.5%. Highly significant (p < 0.001) → ECOG score strongly impacts survival.

The 95% CI gives the range in which the true hazard ratio likely falls. If the CI excludes 1, the variable is statistically significant. For sex (0.4163 to 0.8092) and ph.ecog (1.2642 to 1.9870), the confidence intervals do not include 1, confirming their statistical significance.

Concordance = 0.635 → Indicates the model's predictive accuracy (values closer to 1 are better).

All three tests confirm that at least one covariate significantly affects survival.

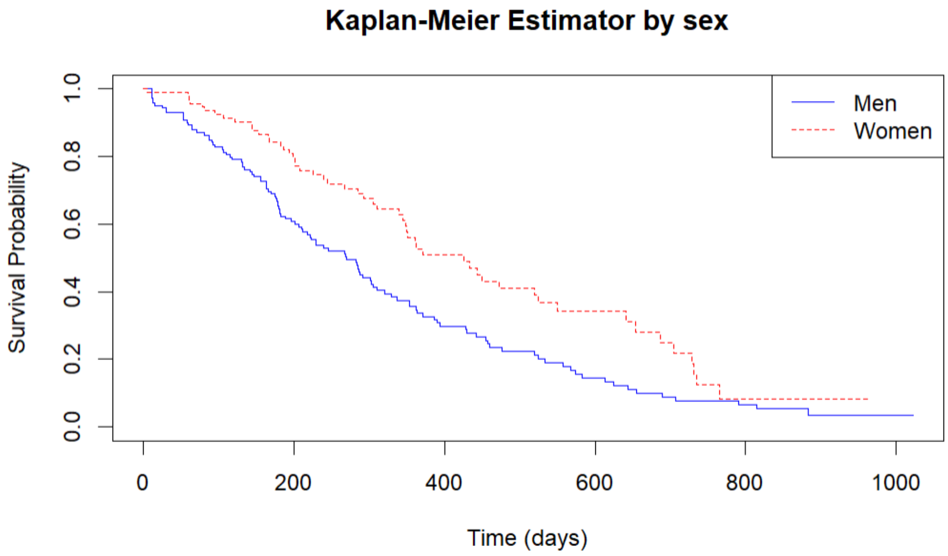
```
library(survminer)

## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

## The following object is masked _by_ '.GlobalEnv':
##
##     myeloma

## The following object is masked from 'package:survival':
##
##     myeloma

# visualization of CI
ggforest(cox_model, data = clean_data)
```



```
print(paste(sum(clean_data$sex == 1), "males"))

## [1] "134 males"

print(paste(sum(clean_data$sex == 2), "females"))

## [1] "89 females"
```

We have more observations of males than females, which we will have to consider for the results.

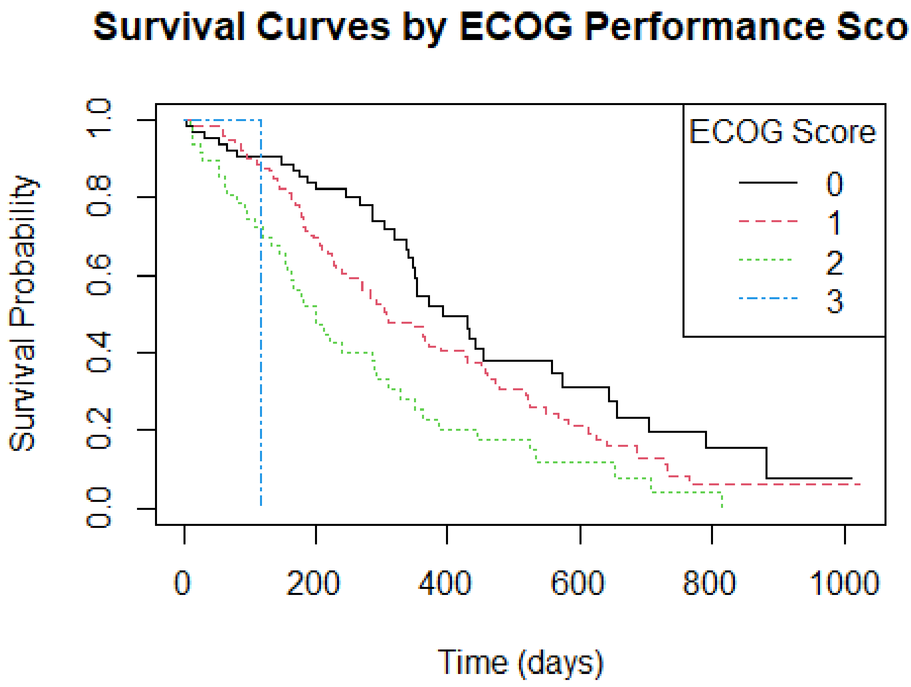
```
survival_male <- survfit(Surv(time, status) ~ 1, data = clean_data[clean_data$sex == 1,
])
survival_female <- survfit(Surv(time, status) ~ 1, data = clean_data[clean_data$sex == 2,
])



plot(survival_male, conf.int = FALSE,
     main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (days)",
     col = "blue")

lines(survival_female, conf.int = FALSE,
     main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (days)",
     col = "red")
```



The data suggests that females have a higher survival probability over time compared to males. This could imply that females respond better to treatment, have a slower disease progression, or other factors that contribute to better survival outcomes. The more significant difference in survival probabilities in the early stages suggests that initial treatment or disease characteristics may differ between males and females, leading to different survival outcomes.

As we explained previously, ph.ecog is a strongly significant value, so we can compare different groups.

```r
# Fit the survival model by ECOG performance score
fit_ecog <- survfit(Surv(time, status == 2) ~ ph.ecog, data = clean_data)

# Plot the survival curves
plot(fit_ecog, col = 1:4, lty = 1:4, xlab = "Time (days)", ylab = "Survival Probability",
main = "Survival Curves by ECOG Performance Score")
legend("topright", legend = levels(factor(cancer$ph.ecog)), col = 1:4, lty = 1:4, title =
"ECOG Score")
```



```
as.data.frame(table(clean_data$ph.ecog))
```

```
##   Var1 Freq
## 1    0   62
## 2    1  113
## 3    2   47
## 4    3    1
```

Patients with an ECOG score of 0 have the highest survival probability over time, followed by those with scores of 1, 2. The score 3 has only one observation which is why there is the vertical line after the event. This indicates that a lower ECOG score is associated with better survival outcomes.

The separation between the curves suggests that the ECOG score is a significant prognostic factor in predicting patient survival.

Conclusion

Our analysis revealed that sex and ECOG performance score are significant predictors of lung cancer patient survival. Females tend to have better survival outcomes than males, although we had bias because of a higher percentage of male observations. Patients with higher ECOG scores, indicating poorer physical function, have a significantly higher risk of death.

In conclusion, this study highlighted the importance of sex and ECOG performance score as predictors of lung cancer patient survival. These results underscore the need for a personalized approach to cancer treatment, considering patients’ individual characteristics. While some variables did not show a significant impact, they may play a role in specific contexts or in interaction with other factors. Further research is needed to deepen our understanding of the complex factors that influence lung cancer survival. Ultimately, this information can contribute to improving treatment strategies and enhancing outcomes for lung cancer patients. Longitudinal studies could also be conducted to follow patients over a longer period and assess the impact of changes in variables over time. It would be interesting to include data on molecular and genetic biomarkers, which could provide additional insights into the underlying mechanisms of lung cancer survival. The analysis of subgroups of patients according to the type of lung cancer would also be a path to consider.