Survival Analysis

Zachary Fuller

2024-05-03

###The goal of this document is to demonstrate different methods for survival analysis. ####A concise presentation of results will be in a separate document.

#####Load data and libraries

```r
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
```

```r
library(readr)
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: ’survminer’

## The following object is masked from ’package:survival’:
##
##     myeloma
```

```r
library(ggplot2)

#note that I set the working directory with setwd().
data <- read_csv("cirrhosis.csv")
```

```
## Rows: 418 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): Status, Drug, Sex, Ascites, Hepatomegaly, Spiders, Edema
## dbl (13): ID, N_Days, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

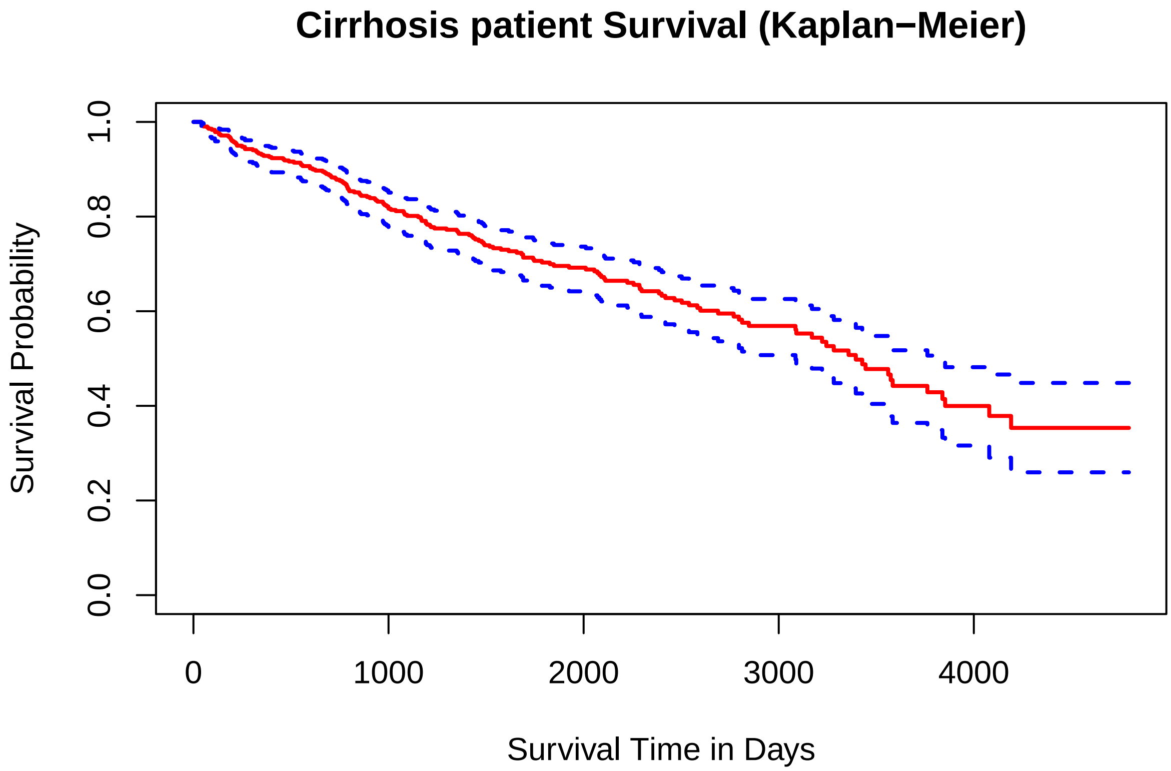
###Status: C=Censored, CL=Censored due to liver transplant, D=Death

#####In order to make the code easily reproducible, use this block to set parameters. You will need to change labels on graphs for different data.

```r
#time till event
time <- data$N_Days
#name of status column in dataset
status <- data$Status
#event of interest (in this example, death)
s <- "D"

#gender
female <- "F"
male <- "M"
```

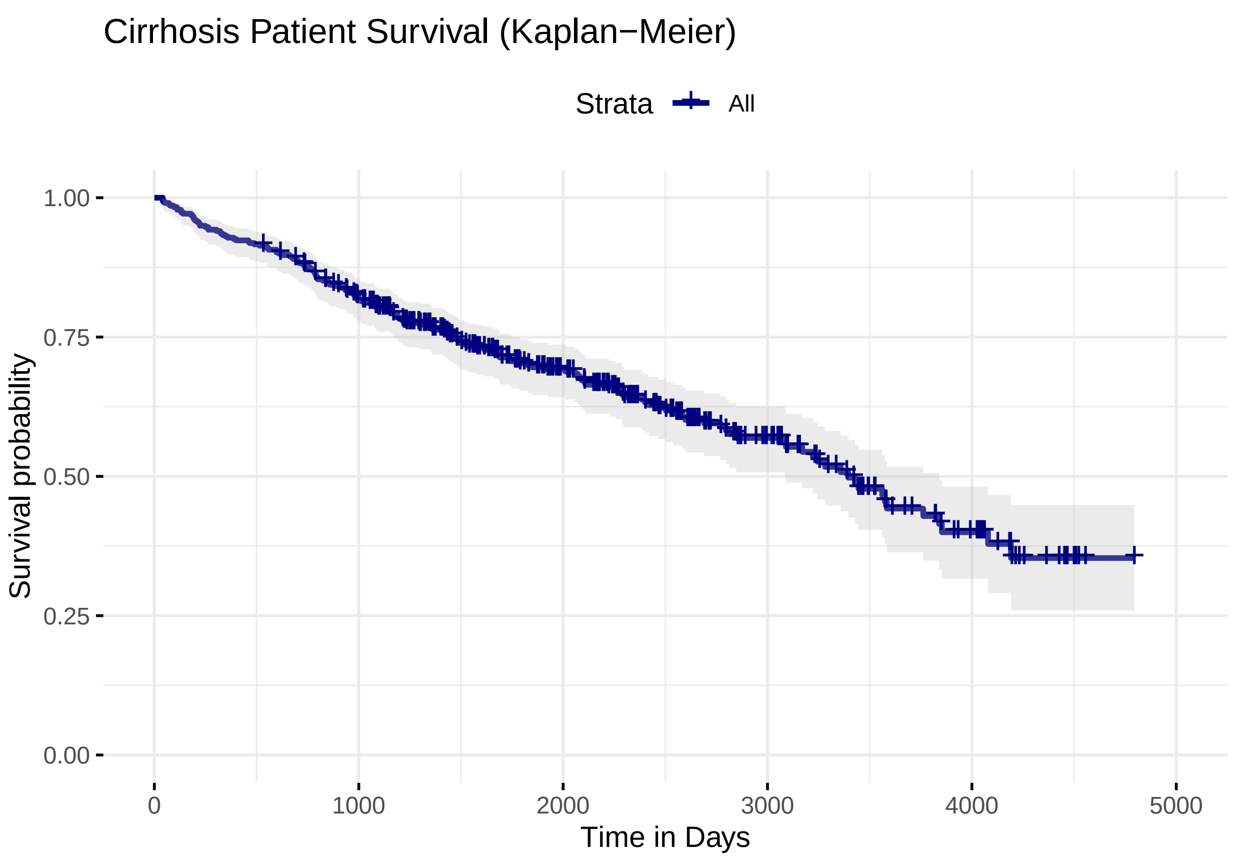
#####Kaplan-Meier Approach, without covariates

```r
#Parameters to set
#time till event
time <- data$N_Days
#name of status column in dataset
status <- data$Status
#event of interest (in this example, death)
s <- "D"

#Calculations
data$SurvObj <- with(data, Surv(time, status == s))
km_surv <- survfit(SurvObj ~ 1, data = data, conf.type = "log-log")

#Standard Plot
plot(km_surv, col=c("red", "blue", "blue"), lwd=2,
     main="Cirrhosis patient Survival (Kaplan-Meier)",
     xlab="Survival Time in Days",
     ylab = "Survival Probability")
```



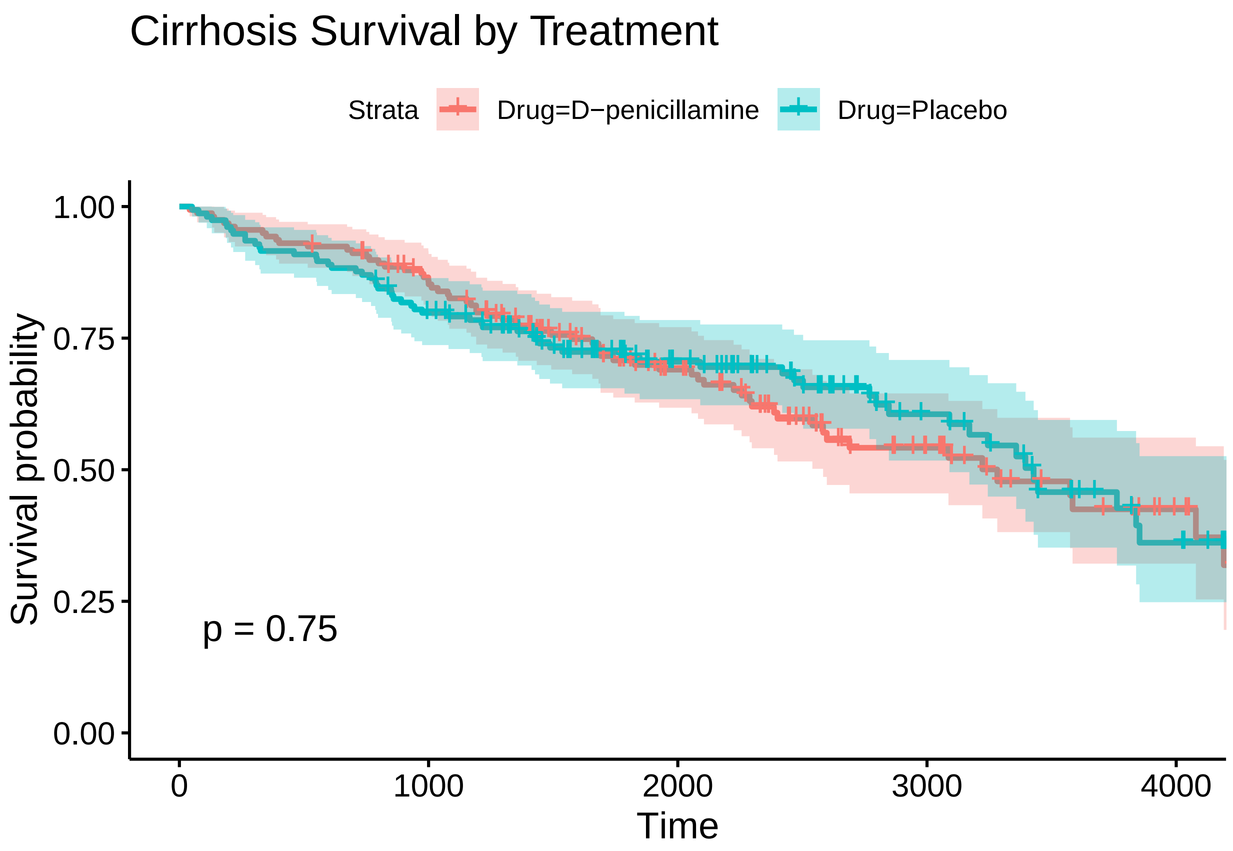
####Kaplan-Meier using ggsurvplot

```
ggsurvplot(
  km_surv,
  data=data,
  size=1,
  palette="navy",
  conf.int=TRUE,
  ggtheme= theme_minimal(),
  xlab="Time in Days")+
  labs(title = "Cirrhosis Patient Survival (Kaplan-Meier)")
```



####Survival Curves by Treatment #####Using ggsurvplot only

```r
bytreat <- survfit(SurvObj~Drug, data=data)
ggsurvplot(bytreat, data=data, conf.int=TRUE, pval=TRUE)+labs(title="Cirrhosis Survival by Treatment")
```



###From the plot, we see that gender is not a significant predictor of survival probability. At least not without including covariates. ###The p-value shown is calculated from the log-rank test and is used to compare the two curves.The issue here is that the curves cross, so log-rank may not perform well in this case.

####Use non-parametric test for the difference between the curves.

```
survdiff(SurvObj~Drug, data=data, rho=0)
```

```
## Call:
## survdiff(formula = SurvObj ~ Drug, data = data, rho = 0)
##
## n=312, 106 observations deleted due to missingness.
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## Drug=D-penicillamine 158       65     63.2    0.0502     0.102
## Drug=Placebo         154       60     61.8    0.0513     0.102
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```

###The p-value from the non-parametric test, still shows that there is not a significant difference between the curves.

####Find a model with significant covariates

```r
#reduce data to complete cases.
data <- data[complete.cases(data),]
mdl <- survreg(SurvObj~.-N_Days-Status-ID, data=data)
set.seed(1234)
mdl_reduced <- capture.output(step(mdl)) #capture.output is just blocking the iterations of the step fu
mdl_reduced[187:215]
```

####Now, compare a model with all significant covariates to one with all plus drug.

```r
#
drug_model <- survreg(SurvObj~Age+Edema+Bilirubin+Albumin+Copper+SGOT+Prothrombin+Stage+Drug, data=data

nodrug_model <- survreg(SurvObj~Age+Edema+Bilirubin+Albumin+Copper+SGOT+Prothrombin+Stage, data=data)

anova(nodrug_model, drug_model)
```

```
##                                                                              Terms
## 1        Age + Edema + Bilirubin + Albumin + Copper + SGOT + Prothrombin + Stage
## 2 Age + Edema + Bilirubin + Albumin + Copper + SGOT + Prothrombin + Stage + Drug
##   Resid. Df    -2*LL Test Df  Deviance  Pr(>Chi)
## 1       265 1937.051      NA        NA        NA
## 2       264 1936.313    =  1 0.7376036 0.3904296
```

###What we see here is that even when significant covariates and included, the drug does not have a significant effect on survival outcome.

Data Downloaded from fedesoriano. (August 2021). Cirrhosis Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset.