

EDA

Chen Liang

2024-11-16

Load data

```
cirrhosis <- read_csv("data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365),
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes"))

## 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.
# Check for missing values
missing_data <- colSums(is.na(cirrhosis))
missing_data
```

##	id	n_days	status	drug	age
##	0	0	0	106	0
##	sex	ascites	hepatomegaly	spiders	edema
##	0	106	106	106	0
##	bilirubin	cholesterol	albumin	copper	alk_phos
##	0	134	0	108	106
##	sgot	tryglicerides	platelets	prothrombin	stage
##	106	136	11	2	6

Histogram Plots for continuouse variables

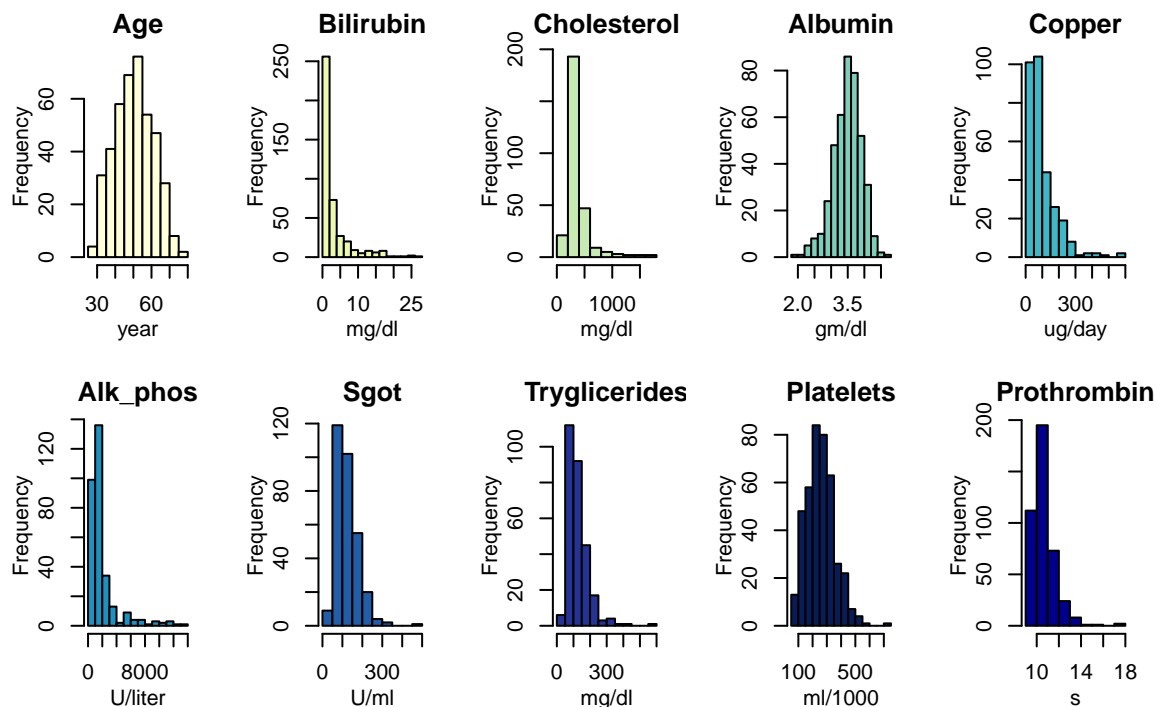
```
conti_vars = cirrhosis |>
  select(age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, tryglicerides, platelets,prothrombin)

par(mfrow = c(2, 5), # 2 rows, 5 columns
    oma = c(2, 2, 3, 1), # Outer margins
    mar = c(4, 4, 2, 1), # Inner margins for individual plots
    mgp = c(2, 1, 0)) # Margins for axis labels and titles
```

```
colors <- c(brewer.pal(9, "YlGnBu"), "darkblue")
```

```
# Plot each histogram using a color from the Set3 palette
```

```
hist(conti_vars$age, main = "Age", xlab = "year", ylab = "Frequency", col = colors[1])
hist(conti_vars$bilirubin, main = "Bilirubin", xlab = "mg/dl", ylab = "Frequency", col = colors[2])
hist(conti_vars$cholesterol, main = "Cholesterol", xlab = "mg/dl", ylab = "Frequency", col = colors[3])
hist(conti_vars$albumin, main = "Albumin", xlab = "gm/dl", ylab = "Frequency", col = colors[4])
hist(conti_vars$copper, main = "Copper", xlab = "ug/day", ylab = "Frequency", col = colors[5])
hist(conti_vars$alk_phos, main = "Alk_phos", xlab = "U/liter", ylab = "Frequency", col = colors[6])
hist(conti_vars$sgot, main = "Sgot", xlab = "U/ml", ylab = "Frequency", col = colors[7])
hist(conti_vars$tryglicerides, main = "Tryglicerides", xlab = "mg/dl", ylab = "Frequency", col = colors[8])
hist(conti_vars$platelets, main = "Platelets", xlab = "ml/1000", ylab = "Frequency", col = colors[9])
hist(conti_vars$prothrombin, main = "Prothrombin", xlab = "s", ylab = "Frequency", col = colors[10])
```



Boxplot for continuous variables

```
# Boxplot for all continuous variables
```

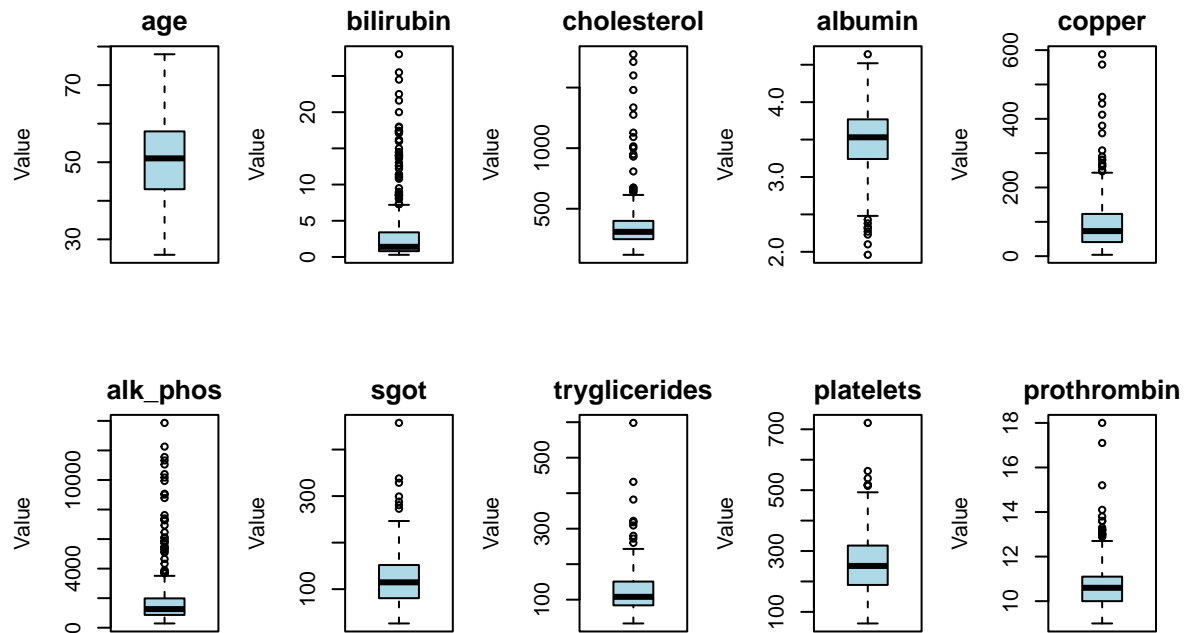
```
par(mfrow = c(2, 5), oma = c(2, 2, 3, 1), mar = c(4, 4, 2, 1))
conti_names <- names(conti_vars)
```

```
for (i in seq_along(conti_names)) {
  boxplot(conti_vars[[conti_names[i]]],
    main = conti_names[i],
    ylab = "Value",
    col = "lightblue",
    outline = TRUE) # Show outliers
}
```

```
# Add an overall title
```

```
mtext("Boxplots for Continuous Variables", outer = TRUE, cex = 1.5, line = 1)
```

Boxplots for Continuous Variables

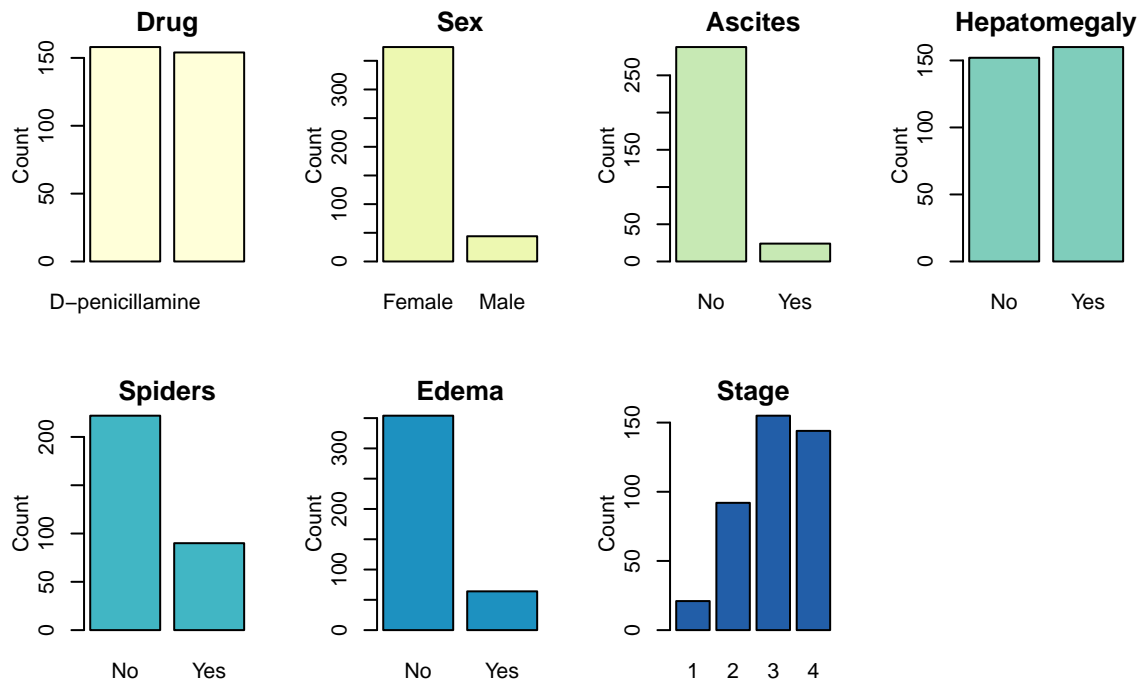


Bar Plots for categorical variables

```
cate_vars = cirrhosis |>
  select(drug, sex, ascites, hepatomegaly, spiders, edema, stage)

par(mfrow = c(2, 4), # 2 rows, 5 columns
    oma = c(2, 2, 3, 1), # Outer margins
    mar = c(4, 4, 2, 1), # Inner margins for individual plots
    mgp = c(2, 1, 0)) # Margins for axis labels and titles

barplot(table(cate_vars$drug), main = "Drug", ylab = "Count", , col = colors[1])
barplot(table(cate_vars$sex), main = "Sex", ylab = "Count", , col = colors[2])
barplot(table(cate_vars$ascites), main = "Ascites", ylab = "Count", col = colors[3])
barplot(table(cate_vars$hepatomegaly), main = "Hepatomegaly", ylab = "Count", col = colors[4])
barplot(table(cate_vars$spiders), main = "Spiders", ylab = "Count", col = colors[5])
barplot(table(cate_vars$edema), main = "Edema", ylab = "Count", col = colors[6])
barplot(table(cate_vars$stage), main = "Stage", ylab = "Count", col = colors[7])
```



Correlation Plot

```
numeric_cirr <- cirrhosis |>
  select_if(is.numeric)

cor_matrix <- cor(numeric_cirr, use = "complete.obs")

corrplot(cor_matrix, method = "circle", type = "lower", order = "hclust")
```

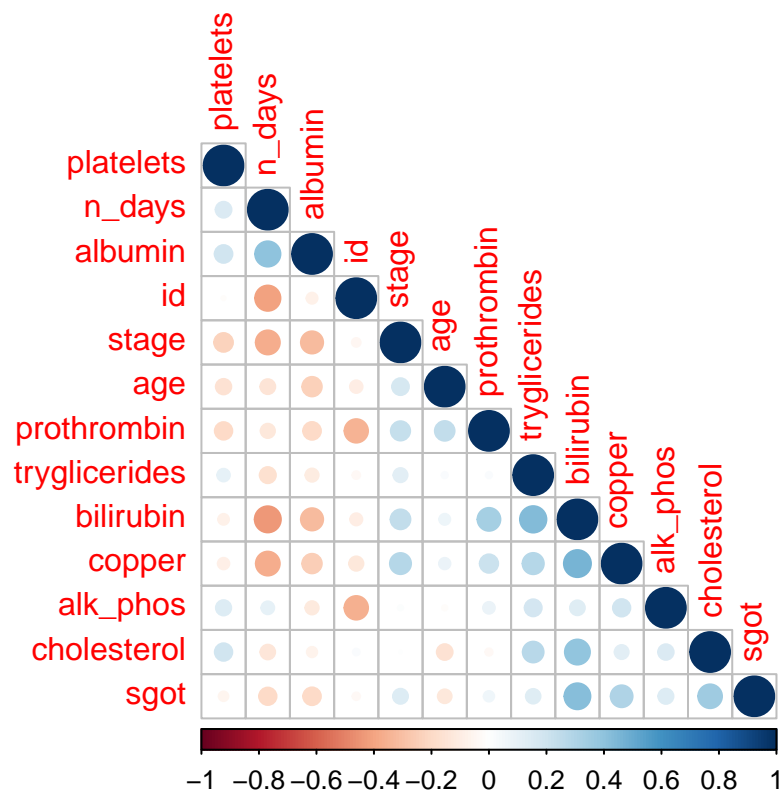


Table 1: Baseline Characteristics

```
theme_gtsummary_journal(journal = "nejm")

## Setting theme "New England Journal of Medicine"

cirrhosis_df <- cirrhosis |>
  mutate(
    status = case_when(
      status == "C" ~ "Censored",
      status == "CL" ~ "Censored due to liver tx",
      status == "D" ~ "Death",
      TRUE ~ status))

table_1 <- cirrhosis_df |>
  select(-id) |>
  tbl_summary(
    by = status,
    statistic = list(
      all_continuous() ~ "{mean} / {median} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    digits = all_continuous() ~ 1,
    missing = "no",
    label = list(
      n_days ~ "N_days",
      drug ~ "Drug",
      age ~ "Age",
```

```

sex ~ "Sex",
ascites ~ "Ascites",
hepatomegaly ~ "Hepatomegaly",
spiders ~ "Spiders",
edema ~ "Edema",
bilirubin ~ "Bilirubin",
cholesterol ~ "Cholesterol",
albumin ~ "Albumin",
copper ~ "Copper",
alk_phos ~ "Alk_phos",
sgot ~ "SGOT",
tryglicerides ~ "Tryglicerides",
platelets ~ "Platelets",
prothrombin ~ "Prothrombin",
stage ~ "Stage"
)) |>
modify_caption("Baseline Characteristics") |>
as_flex_table() |>
line_spacing(space = 0, part = "body")

```

table_1

Warning: fonts used in `flextable` are ignored because the `pdflatex` engine is
used and not `xelatex` or `lualatex`. You can avoid this warning by using the
`set_flextable_defaults(fonts_ignore=TRUE)` command or use a compatible engine
by defining `latex_engine: xelatex` in the YAML header of the R Markdown
document.

Table 1: Baseline Characteristics

Characteristic	Censored N = 232 ¹	Censored due to liver tx N = 25 ¹	Death N = 161 ¹
N_days	2,333.2 / 2,186.5 (994.7)	1,546.2 / 1,435.0 (753.1)	1,376.9 / 1,083.0 (1,000.0)
Drug			
D-penicillamine	83 (49%)	10 (53%)	65 (52%)
Placebo	85 (51%)	9 (47%)	60 (48%)
Age	49.6 / 50.0 (10.4)	41.6 / 41.0 (6.3)	54.0 / 54.0 (9.8)
Sex			
Female	215 (93%)	22 (88%)	137 (85%)
Male	17 (7.3%)	3 (12%)	24 (15%)
Ascites	1 (0.6%)	0 (0%)	23 (18%)
Hepatomegaly	60 (36%)	12 (63%)	88 (70%)
Spiders	33 (20%)	5 (26%)	52 (42%)
Edema	16 (6.9%)	3 (12%)	45 (28%)
Bilirubin	1.6 / 0.9 (1.9)	3.6 / 3.1 (3.6)	5.5 / 3.2 (5.8)
Cholesterol	326.5 / 292.0 (165.8)	439.5 / 343.5 (335.5)	415.8 / 339.0 (275.0)

Table 1: Baseline Characteristics

Characteristic	Censored N = 232 ¹	Censored due to liver tx N = 25 ¹	Death N = 161 ¹
Albumin	3.6 / 3.6 (0.4)	3.5 / 3.5 (0.5)	3.4 / 3.4 (0.5)
Copper	66.6 / 52.0 (57.1)	124.0 / 102.0 (100.1)	135.4 / 111.0 (98.5)
Alk_phos	1,578.1 / 1,107.5 (1,633.1)	1,535.2 / 1,345.0 (837.7)	2,594.4 / 1,664.0 (2,670.0)
SGOT	107.3 / 94.6 (52.8)	130.1 / 127.0 (36.9)	141.9 / 134.9 (58.4)
Tryglicerides	111.8 / 104.0 (48.3)	133.9 / 124.0 (70.5)	140.5 / 122.0 (79.3)
Platelets	261.2 / 256.0 (88.6)	309.6 / 304.0 (102.7)	242.5 / 224.0 (107.0)
Prothrombin	10.5 / 10.4 (0.9)	10.4 / 10.3 (0.5)	11.2 / 11.0 (1.0)
Stage			
1	19 (8.3%)	0 (0%)	2 (1.3%)
2	64 (28%)	5 (20%)	23 (15%)
3	97 (42%)	10 (40%)	48 (31%)
4	50 (22%)	10 (40%)	84 (54%)

¹Mean / Median (SD); n (%)

Stratification

Note: To select variables for stratification, we used categorical variables that are more clinically relevant (recoded age, drug, stage) and variable that have similar sample size between each group (Hepatomegaly). We then perform logrank test, Gehan Wilcoxon test, and KM just for visualization/verification purpose.

Library (add this to the top)

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:gtsummary':
##
##     select

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(survminer)
```

```
## Loading required package: ggpubr

##
## Attaching package: 'ggpubr'

## The following objects are masked from 'package:flextable':
##
##     border, font, rotate
```

```
##
## Attaching package: 'survminer'
## The following object is masked from 'package:survival':
##
##      myeloma
library(glmnet)

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##      expand, pack, unpack
## Loaded glmnet 4.1-8
library(PHInfiniteEstimates)

## Loading required package: lpSolve
## Loading required package: coxphf
## Loading required package: nph
library(kableExtra)

##
## Attaching package: 'kableExtra'
## The following objects are masked from 'package:flextable':
##
##      as_image, footnote
## The following object is masked from 'package:dplyr':
##
##      group_rows
cirrhosis = cirrhosis |>
  mutate(
    status = case_when(
      status == "D" ~ 1, # Event of interest (death)
      status == "C" | status == "CL" ~ 0, # Censored data
      TRUE ~ as.numeric(status)))

## Warning: There was 1 warning in `mutate()`.
## i In argument: `status = case_when(...)`.
```

Caused by warning:

! NAs introduced by coercion

Recoded age

```
# Age divide to 4 groups - seperated by quantile
age_quantile = cirrhosis |>
  pull(age) |>
  quantile(probs = c(0.25, 0.5, 0.75))
cirrhosis_age = cirrhosis |>
  mutate(age_ord = case_when(
```



```

age <= age_quantile[1] ~ paste0("<=", age_quantile[1]),
age <= age_quantile[2] ~ paste0(age_quantile[1], "-", age_quantile[2]-1),
age <= age_quantile[3] ~ paste0(age_quantile[2], "-", age_quantile[3]-1),
.default = paste0(">", age_quantile[3])) |>
  factor( levels = c(paste0("<=", age_quantile[1]),
    paste0(age_quantile[1], "-", age_quantile[2]-1),
    paste0(age_quantile[2], "-", age_quantile[3]-1),
    paste0(">", age_quantile[3])),
    ordered = T))
)

# Logrank
logrank_age_recoded_4 = survdiff(Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
logrank_age_recoded_4

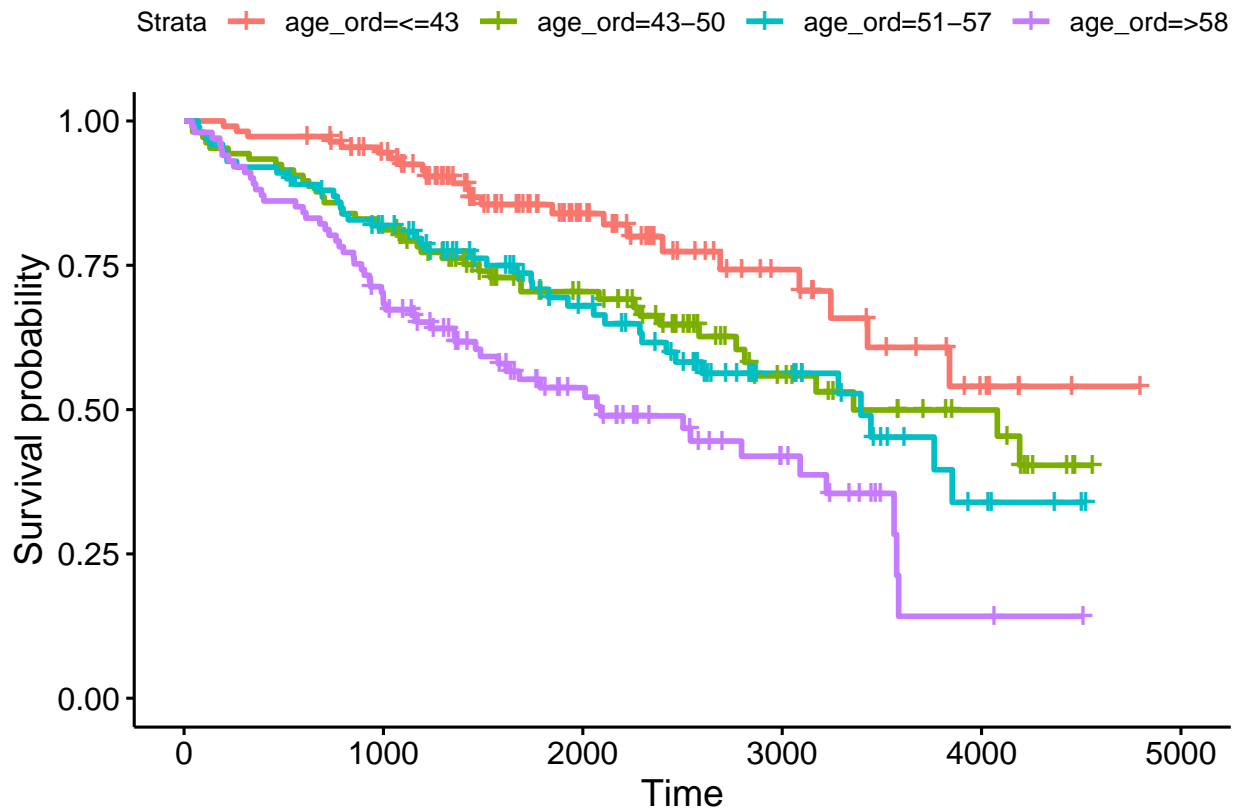
## Call:
## survdiff(formula = Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## age_ord=<=43  111         23    44.2   10.1811   14.0679
## age_ord=43-50 106         42    45.0    0.1964    0.2756
## age_ord=51-57 100         41    39.3    0.0726    0.0962
## age_ord=>58   101         55    32.5   15.5766   19.6845
##
##  Chisq= 26.1  on 3 degrees of freedom, p= 9e-06

# Gehan Wilcoxon test
wilcoxon_age_recoded_4 = gehan.wilcoxon.test(Surv(n_days, status) ~ age_ord, data =
                                             cirrhosis_age)
wilcoxon_age_recoded_4

##
##  Gehan-Wilcoxon
##
## data:
## = 25.62, p-value = 1.146e-05
## alternative hypothesis: two-sided

# KM Curve
survfit(Surv(n_days, status) ~ age_ord, data = cirrhosis_age) |>
  ggsurvplot()

```



```
# divide quantile to 3 then
```

```
survdif(Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
```

```
## Call:
```

```
## survdiff(formula = Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
```

```
##
```

```
##           N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## age_ord=<=43 111      23    44.2   10.1811   14.0679
```

```
## age_ord=43-50 106      42    45.0    0.1964    0.2756
```

```
## age_ord=51-57 100      41    39.3    0.0726    0.0962
```

```
## age_ord=>58 101      55    32.5   15.5766   19.6845
```

```
##
```

```
## Chisq= 26.1 on 3 degrees of freedom, p= 9e-06
```

```
# Age divide to 3 groups - seperated by quantile
```

```
age_quantile = cirrhosis |>
```

```
  pull(age) |>
```

```
  quantile(probs = c(1/3, 2/3))
```

```
cirrhosis_age = cirrhosis |>
```

```
  mutate(age_ord = case_when(
```

```
    age <= age_quantile[1] ~ paste0("<=", age_quantile[1]),
```

```
    age <= age_quantile[2] ~ paste0(age_quantile[1], "-", age_quantile[2]-1),
```

```
    .default = paste0(">", age_quantile[2])) |>
```

```
    factor( levels = c(paste0("<=", age_quantile[1]),
```

```
                  paste0(age_quantile[1], "-", age_quantile[2]-1),
```

```
                  paste0(">", age_quantile[2])),
```

```
    ordered = T))
```

```

)

# Logrank
logrank_age_recoded_3 = survdiff(Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
logrank_age_recoded_3

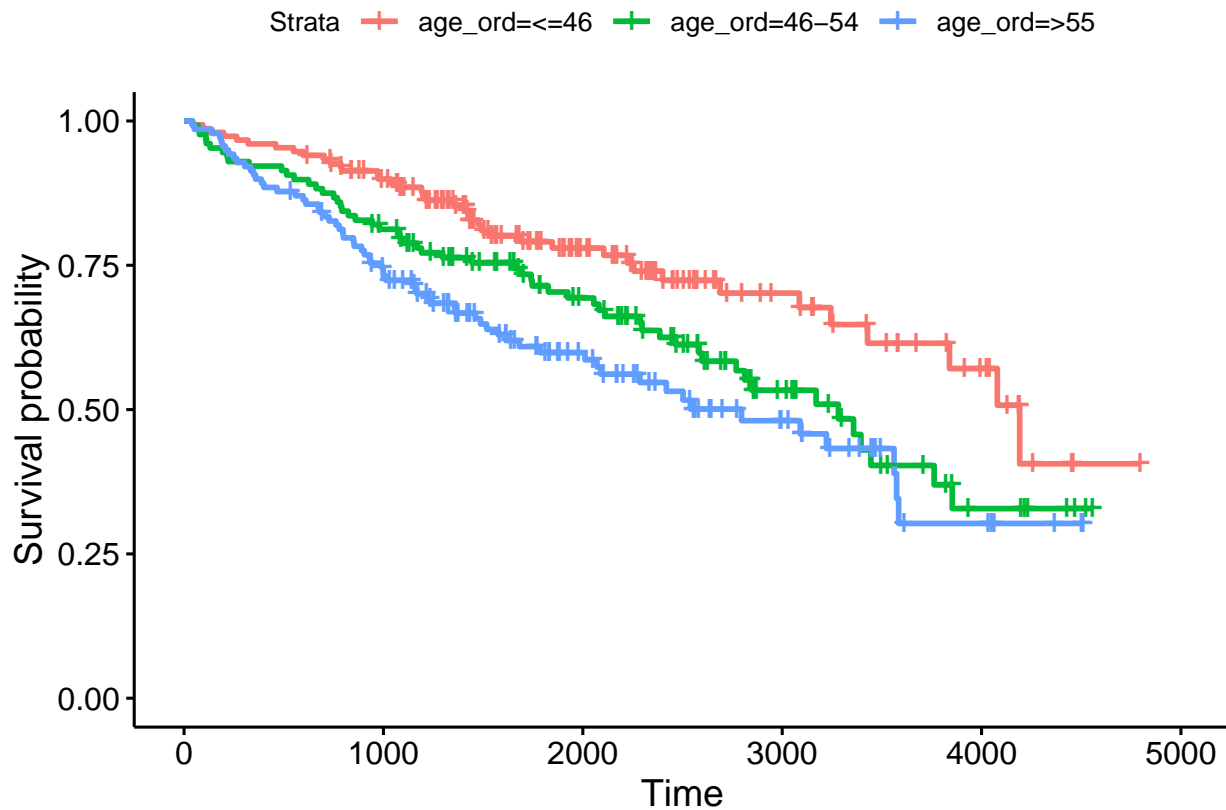
## Call:
## survdiff(formula = Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## age_ord<=46 151         40     60.5      6.959     11.177
## age_ord=46-54 128         56     52.7      0.202      0.301
## age_ord=>55  139         65     47.7      6.242      8.914
##
##  Chisq= 13.4  on 2 degrees of freedom, p= 0.001

# Gehan Wilcoxon test
wilcoxon_age_recoded_3 = gehan.wilcoxon.test(Surv(n_days, status) ~ age_ord, data = cirrhosis_age)
wilcoxon_age_recoded_3

##
##  Gehan-Wilcoxon
##
## data:
## = 14.127, p-value = 0.0008559
## alternative hypothesis: two-sided

# KM Curve
survfit(Surv(n_days, status) ~ age_ord, data = cirrhosis_age) |>
  ggsurvplot()

```



Hepatomegaly

Hepatomegaly - enlarged liver Q: Does having hepatomegaly, a symptom of cirrhosis, affect survival probability?

```
# Logrank
logrank_hepatomegaly = survdiff(Surv(n_days, status) ~ hepatomegaly, data = cirrhosis)
logrank_hepatomegaly

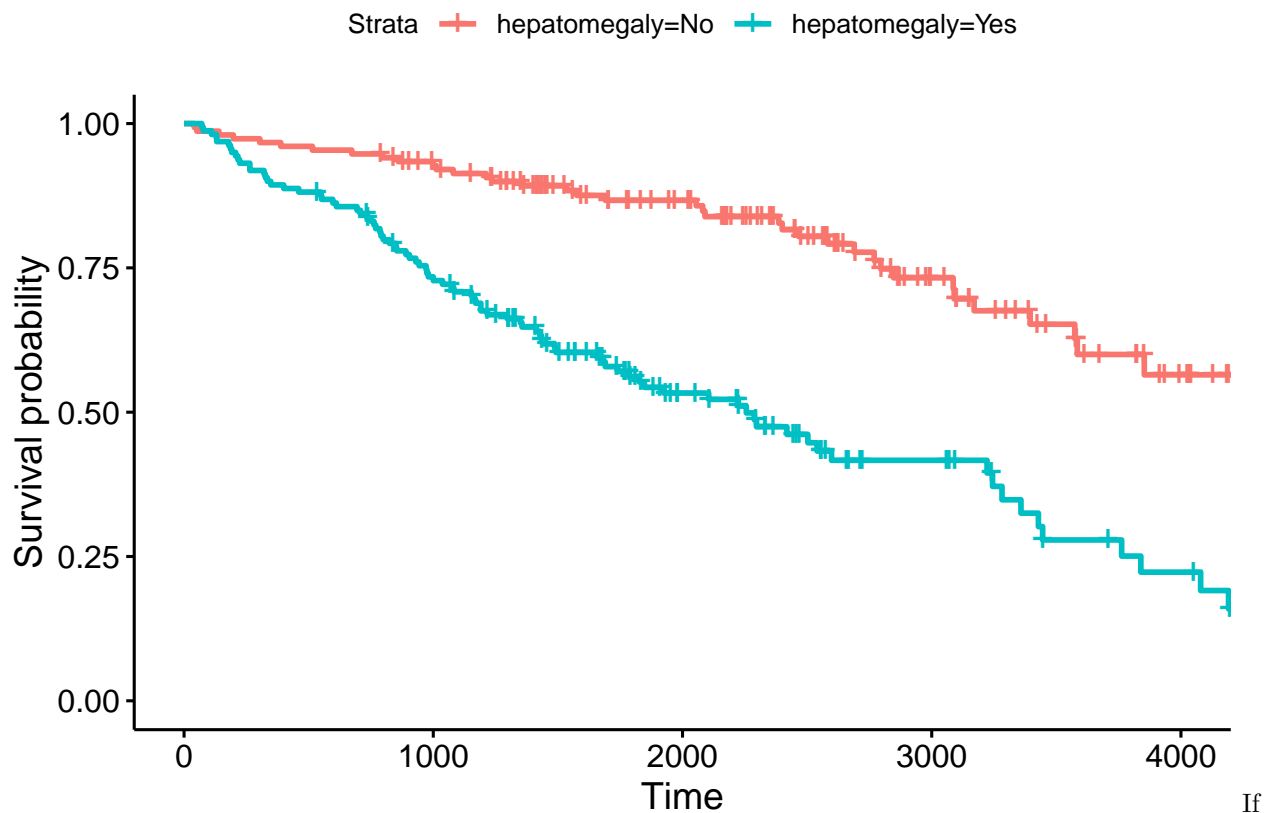
## Call:
## survdiff(formula = Surv(n_days, status) ~ hepatomegaly, data = cirrhosis)
##
## n=312, 106 observations deleted due to missingness.
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## hepatomegaly=No 152      37    71.7     16.8     40.2
## hepatomegaly=Yes 160      88    53.3     22.6     40.2
##
## Chisq= 40.2  on 1 degrees of freedom, p= 2e-10

# Gehan Wilcoxon test
wilcoxon_hepatomegaly = gehan.wilcoxon.test(Surv(n_days, status) ~ hepatomegaly, data = cirrhosis)
wilcoxon_hepatomegaly

##
## Gehan-Wilcoxon
##
## data:
## = 36.159, p-value = 1.818e-09
## alternative hypothesis: two-sided
```

```
# KM Curve
```

```
survfit(Surv(n_days, status) ~ hepatomegaly, data = cirrhosis) |>  
  ggsurvplot()
```



If the both logrank and Wilcoxon test are significant, having hepatomegaly significantly decreases survival probabilities.

Stage

Q: Does different stage of cirrhosis affect mortality?

```
# Log rank test
```

```
logrank_stage = survdiff(Surv(n_days, status) ~ stage, data = cirrhosis)  
logrank_stage
```

```
## Call:
```

```
## survdiff(formula = Surv(n_days, status) ~ stage, data = cirrhosis)
```

```
##
```

```
## n=412, 6 observations deleted due to missingness.
```

```
##
```

```
##           N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## stage=1  21         2    11.4      7.78      8.46
```

```
## stage=2  92        23    44.1     10.12     14.25
```

```
## stage=3 155        48    61.3      2.87      4.73
```

```
## stage=4 144        84    40.2     47.81     65.29
```

```
##
```

```
##  Chisq= 70.1  on 3 degrees of freedom, p= 4e-15
```

```
# Gehan Wilcoxon test
```

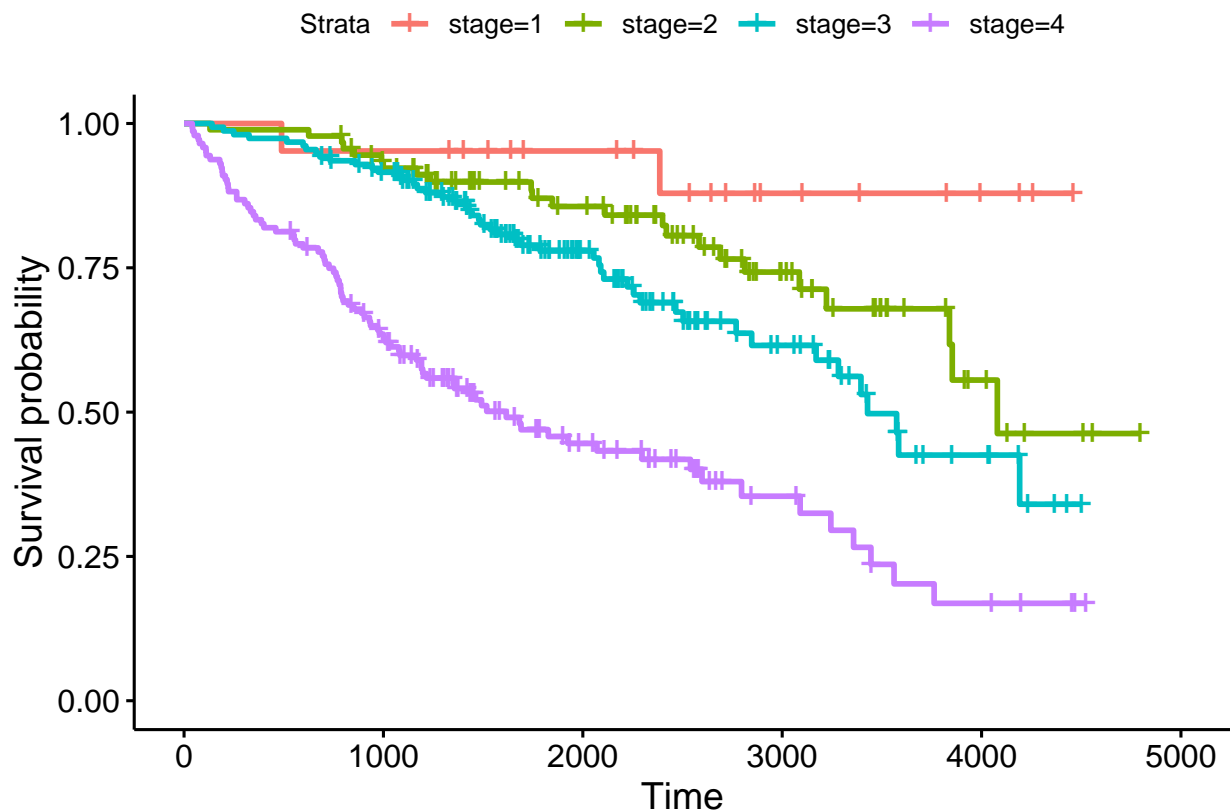
```
wilcoxon_stage = gehan.wilcoxon.test(Surv(n_days, status) ~ stage, data = cirrhosis)
```

```
wilcoxon_stage
```

```
##  
## Gehan-Wilcoxon  
##  
## data:  
## = 76.26, p-value = 2.22e-16  
## alternative hypothesis: two-sided
```

```
# KM Curve
```

```
survfit(Surv(n_days, status) ~ stage, data = cirrhosis) |>  
ggsurvplot()
```



If the both logrank and Wilcoxon test are significant, having later stage of cirrhosis significantly lowers survival probabilities compared to earlier stage of cirrhosis.

Drug

Q: Does D-penicillamine improve risk of mortality?

```
# Log rank test
```

```
logrank_drug = survdiff(Surv(n_days, status) ~ drug, data = cirrhosis)  
logrank_drug
```

```
## Call:
```

```
## survdiff(formula = Surv(n_days, status) ~ drug, data = cirrhosis)
```

```
##
```

```
## 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

# Gehan Wilcoxon test
wilcoxon_drug = gehan.wilcoxon.test(Surv(n_days, status) ~ drug, data = cirrhosis)
wilcoxon_drug

##
## Gehan-Wilcoxon
##
## data:
## = 0.0017763, p-value = 0.9664
## alternative hypothesis: two-sided

# KM Curve
custom_palette = c("#edf8fb", "#b2e2e2", "#66c2a4", "#238b45", # Similar shades for drug 1
                    "#ffffd4", "#fed98e", "#fe9929", "#cc4c02") # Similar shades for drug 2
survdifff(Surv(n_days, status) ~ drug, data = cirrhosis)

## Call:
## survdifff(formula = Surv(n_days, status) ~ drug, data = cirrhosis)
##
## 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

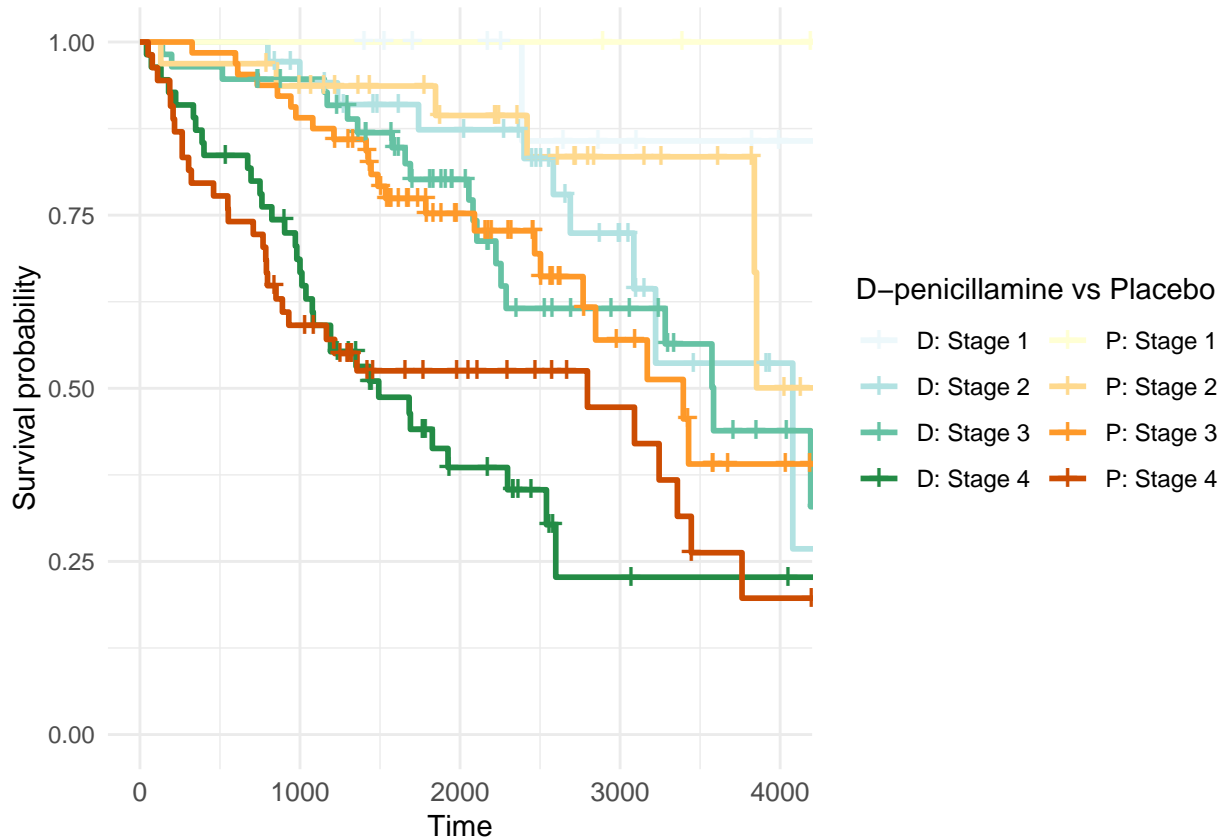
legend_labels <- c(
  "D: Stage 1", "D: Stage 2",
  "D: Stage 3", "D: Stage 4",
  "P: Stage 1", "P: Stage 2",
  "P: Stage 3", "P: Stage 4"
)

km_plot = ggsurvplot(
  survfit(Surv(n_days, status) ~ drug + strata(stage), data = cirrhosis),
  palette = custom_palette, # Apply the custom palette
  legend.title = "D-penicillamine vs Placebo",
  legend.labs = legend_labels
)

km_plot$plot = km_plot$plot +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 10),
        legend.position = "right",
        plot.title = element_text(face="bold", size = 15),
        legend.labs = legend_labels) +
  guides(
    color = guide_legend(ncol = 2) # Make the legend two columns
  )
```

```
print(km_plot)
```

```
## Warning in plot_theme(plot): The `legend.labs` theme element is not defined in the element hierarchy
## The `legend.labs` theme element is not defined in the element hierarchy.
```



If the both logrank and Wilcoxon test are not significant, D-penicillamine does not affect survival probabilities. Supporting argument: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8846335/>

```
test_comparison = data.frame(
  Variable = c(
    "Age Recoded 4",
    "Age Recoded 3",
    "Hepatomegaly",
    "Stage",
    "Drug"
  ),
  Logrank = c(
    logrank_age_recoded_4$pvalue,
    logrank_age_recoded_3$pvalue,
    logrank_hepatomegaly$pvalue,
    logrank_stage$pvalue,
    logrank_drug$pvalue
  ),
  Wilcoxon = c(
    wilcoxon_age_recoded_4$p.value,
    wilcoxon_age_recoded_3$p.value,
    wilcoxon_hepatomegaly$p.value,
    wilcoxon_stage$p.value,
```



```

    wilcoxon_drug$p.value
  )
)
test_comparison |>
  knitr::kable(digits = 4,
               caption = "P-Values from Log-Rank and Wilcoxon Tests")

```

Table 2: P-Values from Log-Rank and Wilcoxon Tests

Variable	Logrank	Wilcoxon
Age Recoded 4	0.0000	0.0000
Age Recoded 3	0.0012	0.0009
Hepatomegaly	0.0000	0.0000
Stage	0.0000	0.0000
Drug	0.7498	0.9664

Feature Selection

Note: Also tried ChatGPT's R implementation of **Collett's Model Selection Approach** (involves p-value of log likelihood test. Gives the same thing as backward/stepwise selection. Since the implementation could be wrong and it's too lengthy. I won't put it there.

Performed forward, backward, and stepwise selection and LASSO to select relevant features.

Forward Selection

```

forward_model = stepAIC(
  coxph(Surv(n_days, status) ~ ., data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  direction = "forward",
  trace = F
)

# Forward already has drug

forward_model_interact = stepAIC(
  coxph(Surv(n_days, status) ~ 1, data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  scope = list(lower = ~ 1, upper = as.formula(paste(
    "~ . +",
    paste0(
      "drug:",
      cirrhosis |> dplyr::select(-id, -n_days, -status, -drug) |> colnames(),
      collapse = " + "
    )
  ))),
  trace = F,
  direction = "forward"
)

## Start:  AIC=1100.38
## Surv(n_days, status) ~ 1

```

```

##
##              Df      AIC
## + drug:bilirubin      2 1028.4
## + drug:copper         2 1052.9
## + drug:albumin        2 1056.4
## + drug:ascites        3 1059.5
## + drug:stage          2 1060.5
## + drug:edema          3 1070.5
## + drug:prothrombin     2 1078.2
## + drug:hepatomegaly    3 1078.6
## + drug:age            2 1082.2
## + drug:spiders        3 1082.6
## + drug:sgot           2 1087.1
## + drug:tryglicerides  2 1091.4
## + drug:cholesterol     2 1094.3
## + drug:platelets       2 1098.9
## + drug:alk_phos        2 1099.4
## + drug:sex             3 1100.0
## <none>                1100.4
##
## Step:  AIC=1028.38
## Surv(n_days, status) ~ drug:bilirubin
##
##              Df      AIC
## + drug:stage          2  999.77
## + drug:albumin        2 1003.36
## + drug:copper         2 1007.57
## + drug:ascites        2 1010.16
## + drug:age            2 1014.06
## + drug:prothrombin     2 1019.03
## + drug:edema          2 1021.27
## + drug:hepatomegaly    2 1022.68
## + drug:spiders        2 1023.94
## + drug:sex            2 1025.63
## + drug:platelets       2 1026.21
## <none>                1028.38
## + drug:sgot           2 1029.22
## + drug:alk_phos        2 1030.17
## + drug:tryglicerides  2 1031.79
## + drug:cholesterol     2 1032.31
##
## Step:  AIC=999.77
## Surv(n_days, status) ~ drug:bilirubin + drug:stage
##
##              Df      AIC
## + drug:albumin        2  985.02
## + drug:copper         2  986.22
## + drug:ascites        2  992.54
## + drug:age            2  993.10
## + drug:prothrombin     2  993.26
## + drug:sgot           2  995.06
## + drug:edema          2  996.52
## + drug:sex            2  997.04
## + drug:platelets       2  999.60

```

```

## <none>          999.77
## + drug:alk_phos      2 1000.30
## + drug:tryglicerides 2 1000.60
## + drug:spiders       2 1001.78
## + drug:hepatomegaly  2 1003.08
## + drug:cholesterol   2 1003.72
##
## Step:  AIC=985.02
## Surv(n_days, status) ~ drug:bilirubin + drug:stage + drug:albumin
##
##           Df      AIC
## + drug:copper      2 974.52
## + drug:sex         2 979.11
## + drug:prothrombin  2 980.15
## + drug:age         2 980.35
## + drug:sgot        2 982.98
## + drug:tryglicerides 2 983.63
## <none>             985.02
## + drug:ascites     2 985.43
## + drug:edema       2 986.47
## + drug:platelets   2 986.97
## + drug:spiders     2 987.25
## + drug:alk_phos    2 988.02
## + drug:hepatomegaly 2 988.31
## + drug:cholesterol 2 988.76
##
## Step:  AIC=974.52
## Surv(n_days, status) ~ drug:bilirubin + drug:stage + drug:albumin +
##      drug:copper
##
##           Df      AIC
## + drug:prothrombin  2 970.36
## + drug:age         2 971.75
## + drug:tryglicerides 2 973.20
## <none>             974.52
## + drug:sex         2 974.86
## + drug:sgot        2 975.26
## + drug:edema       2 975.52
## + drug:platelets   2 976.76
## + drug:ascites     2 976.89
## + drug:alk_phos    2 977.63
## + drug:spiders     2 977.71
## + drug:hepatomegaly 2 977.81
## + drug:cholesterol 2 978.29
##
## Step:  AIC=970.36
## Surv(n_days, status) ~ drug:bilirubin + drug:stage + drug:albumin +
##      drug:copper + drug:prothrombin
##
##           Df      AIC
## + drug:age         2 968.83
## + drug:sgot        2 970.14
## + drug:sex         2 970.31
## <none>             970.36

```

```

## + drug:tryglicerides 2 970.83
## + drug:hepatomegaly 2 972.58
## + drug:platelets 2 973.03
## + drug:edema 2 973.36
## + drug:spiders 2 973.66
## + drug:alk_phos 2 973.99
## + drug:ascites 2 974.20
## + drug:cholesterol 2 974.23
##
## Step: AIC=968.83
## Surv(n_days, status) ~ drug:bilirubin + drug:stage + drug:albumin +
## drug:copper + drug:prothrombin + drug:age
##
##           Df      AIC
## + drug:sgot 2 966.29
## <none> 968.83
## + drug:tryglicerides 2 970.15
## + drug:sex 2 970.55
## + drug:hepatomegaly 2 971.24
## + drug:edema 2 971.33
## + drug:platelets 2 971.50
## + drug:spiders 2 971.61
## + drug:cholesterol 2 971.97
## + drug:ascites 2 972.64
## + drug:alk_phos 2 972.66
##
## Step: AIC=966.29
## Surv(n_days, status) ~ drug:bilirubin + drug:stage + drug:albumin +
## drug:copper + drug:prothrombin + drug:age + drug:sgot
##
##           Df      AIC
## <none> 966.29
## + drug:edema 2 967.81
## + drug:tryglicerides 2 967.97
## + drug:hepatomegaly 2 968.80
## + drug:spiders 2 968.96
## + drug:sex 2 969.38
## + drug:platelets 2 969.64
## + drug:cholesterol 2 969.77
## + drug:ascites 2 969.79
## + drug:alk_phos 2 970.19

forward_model_interact = coxph(as.formula(
  paste(
    "Surv(n_days, status) ~",
    forward_model_interact$formula[3],
    "+ drug"
  )
), data = cirrhosis |>
  dplyr::select(-id) |> na.omit())

forward_model_interact |>
  tbl_regression(exponentiate = TRUE) |>

```

Table 3: Cox Model From Forward Selection

Characteristic	HR [†]	95% CI [†]	p-value
drug			
D-penicillamine	—	—	
Placebo	0.61	0.00 to 564	0.89
drug * bilirubin			
D-penicillamine * bilirubin	1.07	1.00 to 1.14	0.049
Placebo * bilirubin	1.12	1.07 to 1.17	<0.001
drug * stage			
D-penicillamine * stage	1.46	0.97 to 2.20	0.067
Placebo * stage	1.80	1.19 to 2.73	0.005
drug * albumin			
D-penicillamine * albumin	0.38	0.18 to 0.78	0.008
Placebo * albumin	0.46	0.22 to 0.96	0.039
drug * copper			
D-penicillamine * copper	1.00	1.00 to 1.01	0.073
Placebo * copper	1.00	1.00 to 1.01	0.029
drug * prothrombin			
D-penicillamine * prothrombin	1.42	1.02 to 1.98	0.036
Placebo * prothrombin	1.28	1.01 to 1.62	0.041
drug * age			
D-penicillamine * age	1.03	1.00 to 1.05	0.044
Placebo * age	1.03	1.00 to 1.07	0.052
drug * sgot			
D-penicillamine * sgot	1.01	1.00 to 1.01	0.026
Placebo * sgot	1.00	1.00 to 1.01	0.18

[†]HR = Hazard Ratio, CI = Confidence Interval

```
modify_caption("Cox Model From Forward Selection")
```

Backward Selection

```
backward_model = stepAIC(
  coxph(Surv(n_days, status) ~ ., data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  direction = "backward",
  trace = F
)

backward_model = coxph(as.formula(
  paste("Surv(n_days, status) ~", backward_model$formula[3], "+ drug")
), data = cirrhosis |>
  dplyr::select(-id) |> na.omit())
```

Table 4: Cox Model From Backward Selection

Characteristic	HR ¹	95% CI ¹	p-value
drug			
D-penicillamine	—	—	
Placebo	2.92	0.56 to 15.4	0.21
age	1.04	1.01 to 1.06	0.001
edema			
No	—	—	
Yes	2.29	1.26 to 4.16	0.007
bilirubin	1.07	1.02 to 1.13	0.004
cholesterol	1.00	1.00 to 1.00	0.062
albumin	0.48	0.28 to 0.84	0.009
copper	1.00	1.00 to 1.01	0.003
sgot	1.01	1.00 to 1.01	0.003
tryglicerides	1.00	0.99 to 1.00	0.046
platelets	1.00	1.00 to 1.00	0.27
prothrombin	1.20	0.98 to 1.47	0.078
stage	1.73	1.28 to 2.35	<0.001
drug * sgot			
Placebo * sgot	0.99	0.99 to 1.00	0.034
drug * tryglicerides			
Placebo * tryglicerides	1.01	1.00 to 1.01	0.007
drug * platelets			
Placebo * platelets	1.00	0.99 to 1.00	0.072

¹HR = Hazard Ratio, CI = Confidence Interval

```
backward_model_interact = stepAIC(
  coxph(as.formula(
    paste(
      "Surv(n_days, status) ~ . +",
      paste0(
        "drug:",
        cirrhosis |> dplyr::select(-id, -n_days, -status, -drug) |> colnames(),
        collapse = " + "
      )
    )
  ), data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  direction = "backward",
  trace = F
)

backward_model_interact |>
  tbl_regression(exponentiate = TRUE) |>
  modify_caption("Cox Model From Backward Selection")
```

Stepwise Selection

```
stepwise_model = stepAIC(
  coxph(Surv(n_days, status) ~ ., data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  direction = "both",
  trace = F
)
stepwise_model = coxph(as.formula(
  paste("Surv(n_days, status) ~", stepwise_model$formula[3], "+ drug")
), data = cirrhosis |>
  dplyr::select(-id) |> na.omit())

stepwise_model_interact = stepAIC(
  coxph(as.formula(
    paste(
      "Surv(n_days, status) ~ . +",
      paste0(
        "drug:",
        cirrhosis |> dplyr::select(-id, -n_days, -status, -drug) |> colnames(),
        collapse = " + "
      )
    )
  ), data = cirrhosis |>
    dplyr::select(-id) |> na.omit()),
  direction = "both",
  trace = F
)

stepwise_model_interact |>
  tbl_regression(exponentiate = TRUE) |>
  modify_caption("Cox Model From Stepwise Selection")
```

LASSO

```
x = model.matrix(Surv(n_days, status) ~ . - id, cirrhosis |> na.omit())[, -1]
y = Surv(cirrhosis |>
  na.omit() |>
  pull(n_days),
  cirrhosis |>
  na.omit() |>
  pull(status))
cv_model = cv.glmnet(x, y, family = "cox", alpha = 1)
best_lambda = cv_model$lambda.min
selected_coefficients = coef(cv_model, s = best_lambda)
selected_var_name = rownames(selected_coefficients)[which(selected_coefficients != 0)]
selected_var_name = map(selected_var_name, str_remove_all, "Yes") |> unlist()
lasso_model = coxph(as.formula(paste0("Surv(n_days, status) ~ ", paste(selected_var_name, collapse = "+
  dplyr::select(-id) |> na.omit())

lasso_model = coxph(as.formula(
  paste(
    "Surv(n_days, status) ~",
```

Table 5: Cox Model From Stepwise Selection

Characteristic	HR [†]	95% CI [†]	p-value
drug			
D-penicillamine	—	—	
Placebo	2.92	0.56 to 15.4	0.21
age	1.04	1.01 to 1.06	0.001
edema			
No	—	—	
Yes	2.29	1.26 to 4.16	0.007
bilirubin	1.07	1.02 to 1.13	0.004
cholesterol	1.00	1.00 to 1.00	0.062
albumin	0.48	0.28 to 0.84	0.009
copper	1.00	1.00 to 1.01	0.003
sgot	1.01	1.00 to 1.01	0.003
tryglicerides	1.00	0.99 to 1.00	0.046
platelets	1.00	1.00 to 1.00	0.27
prothrombin	1.20	0.98 to 1.47	0.078
stage	1.73	1.28 to 2.35	<0.001
drug * sgot			
Placebo * sgot	0.99	0.99 to 1.00	0.034
drug * tryglicerides			
Placebo * tryglicerides	1.01	1.00 to 1.01	0.007
drug * platelets			
Placebo * platelets	1.00	0.99 to 1.00	0.072

[†]HR = Hazard Ratio, CI = Confidence Interval

```

lasso_model$formula[3],
  "+ drug"
)
), data = cirrhosis |>
  dplyr::select(-id) |> na.omit())

```

```

lasso_model |>
  tbl_regression(exponentiate = TRUE) |>
  modify_caption("Cox Model From LASSO")

```

```
str_replace_all(paste(forward_model$formula[[3]][2]), " \\+", ",")
```

```
## [1] "drug, age, sex, ascites, hepatomegaly, spiders, edema, bilirubin, cholesterol, albumin, copper,
```

Model Comparison

```

model_comparison = data.frame(
  Model = c(
    "Forward",

```


Table 6: Cox Model From LASSO

Characteristic	HR [†]	95% CI [†]	p-value
age	1.03	1.01 to 1.05	0.004
ascites			
No	—	—	
Yes	0.99	0.49 to 2.00	0.97
spiders			
No	—	—	
Yes	1.14	0.72 to 1.79	0.58
edema			
No	—	—	
Yes	1.48	0.88 to 2.47	0.14
bilirubin	1.09	1.05 to 1.13	<0.001
albumin	0.47	0.27 to 0.83	0.009
copper	1.00	1.00 to 1.00	0.005
sgot	1.00	1.00 to 1.01	0.012
prothrombin	1.28	1.05 to 1.56	0.014
stage	1.52	1.13 to 2.05	0.006
drug			
D-penicillamine	—	—	
Placebo	0.94	0.63 to 1.42	0.78

[†]HR = Hazard Ratio, CI = Confidence Interval

```

    "Forward with Interaction",
    "Backward",
    "Backward with Interaction",
    "Stepwise",
    "Stepwise with Interaction",
    "LASSO"
),
Log_Lik = c(
  logLik(forward_model),
  logLik(forward_model_interact),
  logLik(backward_model),
  logLik(backward_model_interact),
  logLik(stepwise_model),
  logLik(stepwise_model_interact),
  logLik(lasso_model)
),
AIC = c(
  AIC(forward_model),
  AIC(forward_model_interact),
  AIC(backward_model),
  AIC(backward_model_interact),
  AIC(stepwise_model),
  AIC(stepwise_model_interact),

```

```

    AIC(lasso_model)
  ),
  Kept_Variable = c(
    str_replace_all(paste(forward_model$formula[[3]][2]), " \\+", ","),
    str_replace_all(paste(forward_model_interact$formula[3]), " \\+", ","),
    str_replace_all(paste(formula(backward_model)[3]), " \\+", ","),
    str_replace_all(paste(formula(backward_model_interact)[3]), " \\+", ","),
    str_replace_all(paste(formula(stepwise_model)[3]), " \\+", ","),
    str_replace_all(paste(formula(stepwise_model_interact)[3]), " \\+", ","),
    str_replace_all(paste(formula(lasso_model)[3]), " \\+", ",")
  )
)
model_comparison |>
  knitr::kable() |>
  kable_styling(full_width = TRUE) |>
  column_spec(1, width = "2cm") |>
  column_spec(2, width = "2cm") |>
  column_spec(3, width = "2cm") |>
  column_spec(4, width = "7cm")

```

Model	Log_Lik	AIC	Kept_Variable
Forward	-467.8089	969.6179	drug, age, sex, ascites, hepatomegaly, spiders, edema, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, tryglicerides, platelets, prothrombin
Forward with Interaction	-469.1367	968.2734	drug:bilirubin, drug:stage, drug:albumin, drug:copper, drug:prothrombin, drug:age, drug:sgot, drug
Backward	-469.5553	957.1105	age, edema, bilirubin, albumin, copper, sgot, prothrombin, stage, drug
Backward with Interaction	-463.5469	957.0937	drug, age, edema, bilirubin, cholesterol, albumin, copper, sgot, tryglicerides, platelets, prothrombin, stage, drug:sgot, drug:tryglicerides, drug:platelets
Stepwise	-469.5553	957.1105	age, edema, bilirubin, albumin, copper, sgot, prothrombin, stage, drug
Stepwise with Interaction	-463.5469	957.0937	drug, age, edema, bilirubin, cholesterol, albumin, copper, sgot, tryglicerides, platelets, prothrombin, stage, drug:sgot, drug:tryglicerides, drug:platelets
LASSO	-469.4033	960.8066	age, ascites, spiders, edema, bilirubin, albumin, copper, sgot, prothrombin, stage, drug

Since stepwise model have the lowest AIC and best log likelihood statistics, “age + edema + bilirubin + albumin + copper + sgot + prothrombin + stage” will be used for the following models.

Precaution: ChatGPT said we need to choose a variable of interest before variable selection (forcing the variable in). From the context of the data, it seems like they are testing the effect of drug. However, by test statistics and unadjusted association between drug and survival probability. There is no apparent link. Maybe this could be the central topic of the following parts?

Table 8: Multivariate Cox Proportional Hazards Analysis

Characteristic	HR ¹	95% CI ¹	p-value
age	1.02	1.00 to 1.04	0.019
sex			
Female	—	—	
Male	1.30	0.75 to 2.26	0.35
bilirubin	1.12	1.08 to 1.16	<0.001
albumin	0.35	0.22 to 0.56	<0.001
copper	1.00	1.00 to 1.01	0.002
prothrombin	1.32	1.12 to 1.57	0.001
stage	1.46	1.13 to 1.88	0.003

¹HR = Hazard Ratio, CI = Confidence Interval

Multivariate analysis

```

cirrhosis <- cirrhosis |>
  mutate(
    status = case_when(
      status == "D" ~ 1, # Event of interest (death)
      status == "C" | status == "CL" ~ 0, # Censored data
      TRUE ~ as.numeric(status))

# Fit the Cox proportional hazards model
cox_model <- coxph(Surv(n_days, status) ~ age + sex + bilirubin + albumin + copper + prothrombin + stage,
  data = cirrhosis)

# Summarize the results
cox_summary <- tbl_regression(cox_model, exponentiate = TRUE) %>%
  modify_caption("Multivariate Cox Proportional Hazards Analysis")

cox_summary

```