Final Project: Cirrhosis Patients Survival Analysis

Ruisi Liu, Yuqi Niu

December 6, 2023

1 Introduction

1.1 Data Source

Our dataset, sourced from the UC Irvine Machine Learning Repository, encompasses data on cirrhosis resulting from prolonged liver damage, often attributed to conditions such as hepatitis or chronic alcoholism. This dataset is derived from a Mayo Clinic study on primary biliary cirrhosis (PBC), conducted from 1974 to 1984. It includes records of 424 patients who were diagnosed with PBC and met the criteria for a randomized, placebo-controlled trial of D-penicillamine conducted during this ten-year period. The dataset's initial 312 entries represent participants in this trial and feature comprehensive data. An additional 112 patients, not part of the clinical trial, agreed to basic measurement recording and survival tracking. However, six of these cases were lost to follow-up soon after diagnosis, so the dataset includes detailed information on 106 non-trial patients, in addition to the 312 trial participants.

The dataset comprises 20 columns: ID, N_Days, Status, Drug (D-penicillamine/placebo), Age, Sex, Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Triglycerides, Platelets, and Prothrombin. In our survival analysis, 'N_Days' (the time from registration to death, transplantation, or study analysis) and 'Status' (censored, censored due to liver transplantation, or death) are used as response variables. The remaining 17 columns serve as predictors, with 14 of these providing detailed medical condition data.We drop the rows that have at least one NA when we fit our models. Finally, the dataset includes 147 censored cases, 18 cases censored due to liver transplantation, and 111 deaths.

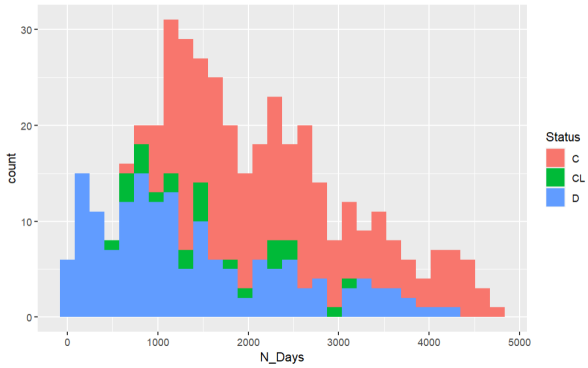

Figure 1: Distribution of Data(D: Death, C: Censored,CL:Censored because of Transplantation)

2 Methodology

2.1 Questions of Interest

Given the medical conditions of patients, hospitals may want to prioritize those who have higher risk of dying in the next few days/years. To help understand the survival time of cirrhosis patients, we propose the following question:

- What are the factors that are related to the hazard probability (and also survival probability) of cirrhosis patients over time?
- Will the factors change if we assume censoring to be informative?

Since there are many patients censored from the study either for unknown reason or for transplantation, we want to build models to assess the probability for all these events. The hospitals can also monitor those who are more likely to quit the study to take some actions to maintain them.

- What are the probability of censoring (with or without liver transplantation) over time?

2.2 Models

1. Non-parametric model: Kaplan-Meier

   We first utilize Kaplan-Meier estimator as a fundamental tool for visualizing the survival experience. It is a non-parametric statistic used to estimate the survival function from lifetime data, so it makes no assumptions about the survival distribution. The Kaplan-Meier estimator is formulated as

   $$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

   where $S(t)$ denotes the survival probability beyond time t, t_i are the distinct observed event times, d_i represents the number of events at time t_i, and n_i is the number of subjects at risk just prior to time t_i.

2. Parametric models: Weibull and Loglog
   Weibull Proportional Hazards Model

   $$\theta = e^{\mathbf{X}\beta}$$
   $$h(t) = \lambda p(\lambda t)^{p-1}\theta$$
   $$S(t) = e^{-(\lambda\theta^{1/p}t)^p}$$

   when $p > 1$ the hazard is increasing and when $p < 1$ the hazard is decreasing over time. i represents each patient.

   Loglog-based Discrete-Time Model

   $$log(\lambda) = \mathbf{X_i}\beta$$
   $$h(t) = \frac{\frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}}{1 + \left(\frac{t}{\lambda}\right)^{\alpha}}$$
   $$S(t) = (1 + \frac{t}{\lambda}^{\alpha})^{-1}$$

   where λ is the scale parameter and α is the shape parameter. When $\alpha > 1$, the hazard function initially increases and then decreases, showing a unimodal shape. When $\alpha < 1$, the hazard function is monotonically decreasing over time.

   Note that we did not choose exponential distribution, which is a special case of Weibull, to model because it assumes independence of time, which is unrealistic in our context.

3. Semi-parametric model: Cox Proportional Hazard

   $$h(t, \mathbf{X}) = h_0(t)e^{\mathbf{X}\beta}$$
   $$S(t) = exp(-\int_0^t h(u)du)$$

4. Competing Risk Model
   There are three events in our data: Death, Censored w/o transplantation, Censored w/ transplantation. In order to track the change for all the events, we use a competing risk model to assess the probability of events occuring. The estimated cumulative incidences are as described in Putter, Fiocco & Geskus (2007).

2.3 Inverse Probability Censoring Weighting

All of our models assume censoring to be non-informative. Though it is impossible to test this assumption explicitly, we want to measure if there is an imbalance in the covariates distribution by the absolute standardized mean difference between death and censored data. If so, our survival model may be potentially biased and people who are censored will be under-represented.

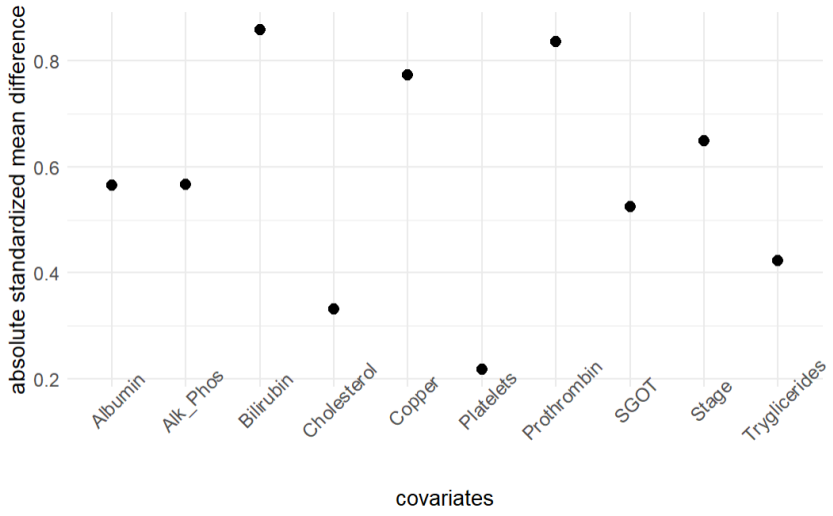

Figure 2: Absolute Standardized Mean Difference of Covariates between Patients who died and who were censored

From the plot we see that there is a significant divergence between the covariates in the death and censored data. Thus, we need to use weights to rebalance our model estimations.
We first use compute the propensity score, the probability of death, using the logit model.

$$log\frac{p}{1-p} = \beta_0 + \beta_i X_i + ...$$
$$p = P(\text{death})$$
$$i \in \{Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Trygliceride s, Platelets, Prothrombin, Edema_N, Edema_S\}$$

Then we calculate the inverse probability of weights for each individual.

$$Weights = \begin{cases} \frac{1}{\text{propensity score}} & \text{for dead patients} \\ \frac{1}{\text{(1-propensity score)}} & \text{for censored patients} \end{cases} \quad (1)$$

Ultimately, to ensure covariate balance between the two groups, we evaluate the standardized differences both before and after applying the weights.

2.4 Assess Causal Relationship

Since patients' data was collected at the END of their visits to the clinic, their medical conditions might be influenced by the drug(even though before the study they were randomly assigned to treatment). If this is the case, then we should only include drug as our predictor and exclude all other covariates.

Therefore, we decided to firstly use a logit model to calculate the propensity score to see if there is a difference between the distribution in the covariates of the treatment group and the control group.

$$log\frac{p}{1-p} = \beta_0 + \beta_i X_i + ...$$

$$p = P(\text{assigned to the treatment group})$$

$$i \in \{Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Trygliceride s, Platelets, Prothrombin, Edema_N, Edema_S\}$$

Then we use deviance test to compare this model with the null model(no covariates, only an offset term) to see if the drug group is different from the control group in terms of the patients' medical records.

$$2(l(\theta_{null}) - l(\theta_{full})) \sim \chi^2_{n-p-1}$$

where $l(\theta_{null})$ is the log-likelihood of the null model, $l(\theta_{full})$ is the log-likelihood of the full model.

2.5 Multiple Testing Correction

Since there are many coefficients in each model, and we test for whether $\beta_i = 0$ multiple times, we use Bonferroni correction and Benjamin-Hochberg methods to adjust the p-values. We then select the significant coefficients based on the updated p-values.

3 Results

3.1 Assumption Check

3.1.1 Non-infomative Censoring

We see that the standardized absolute difference between the death and censored data is quite different before and after weighting and is reduced significantly, so we could treat censored data as non-informative now.

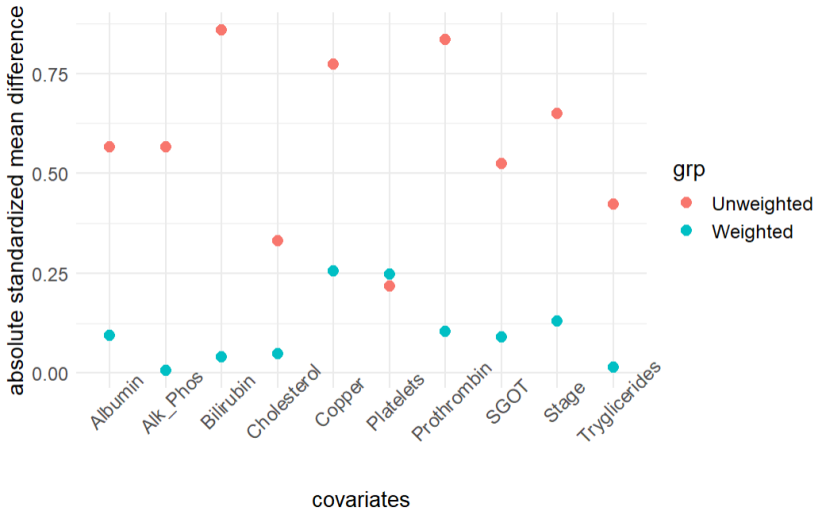

Figure 3: Absolute Standardized Mean Difference of Covariates between Patients who taken drugs and those did not (blue points are after weighting)

3.1.2 Weibull Model Assumption

The most distinctive feature of a Weibull cumulative hazard plot is that when plotted on a log-log scale, it should approximately form a straight line. So we also plot the relationship between log cumulative hazard and log time using Kaplan-Meier estimates.

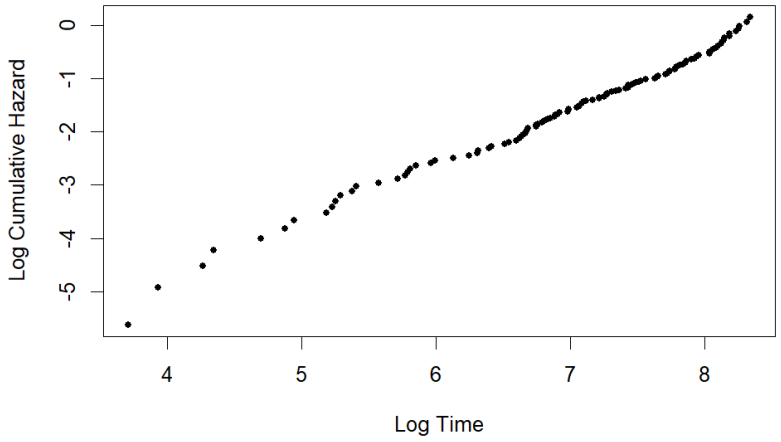

Figure 4: Cumulative Hazard Estimate

Since it is approximately a straight line, suggesting that the Weibull model may be suitable.

3.1.3 Proportional Hazard Function

In Cox proportional hazards models, we assume the ratio of the hazard rates for any two individuals is constant over time. In other words, the effect of the covariates on the hazard is constant over time. We use 'cox.zph()', which operates by examining the Schoenfeld residuals against the time to check for proportionality assumption. We find the global p-value is 0.051 and $1.4 * 10^{-9}$ before and after weighting respectively. Therefore, it suggests that the proportional hazards assumption is likely violated after weighting.

3.2 Inference

3.2.1 Kaplan-Meier Estimates

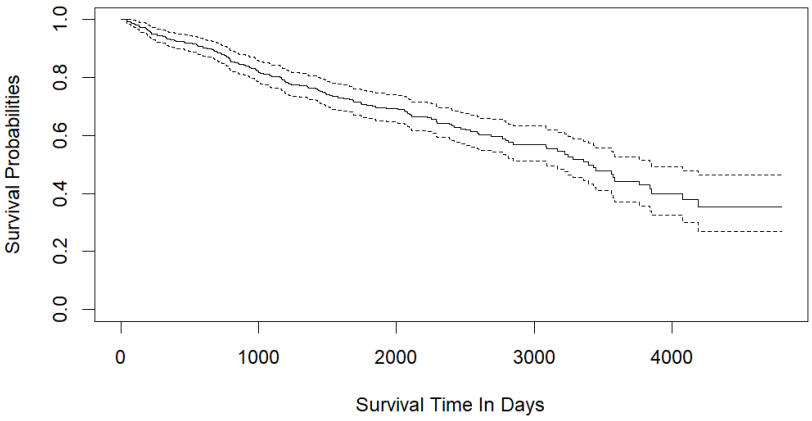

Figure 5: Kaplan-Meier Estimates for All Patients

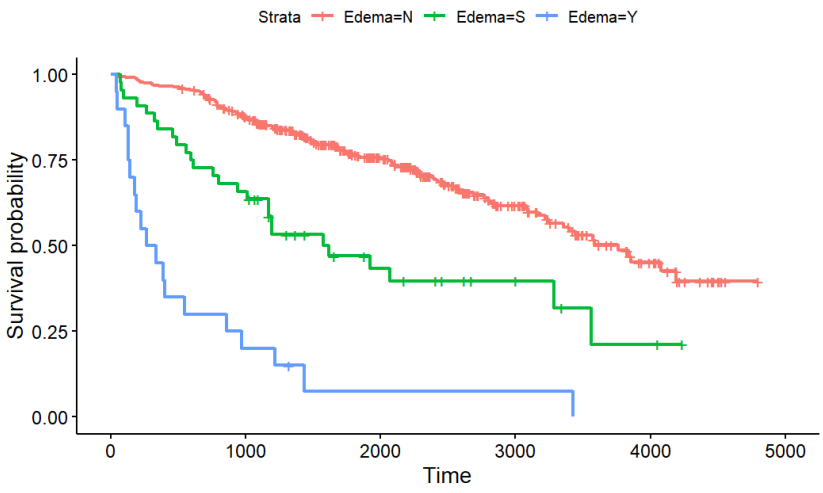

Figure 6: Kaplan-Meier Estimates for Patients with Different Endema Conditions

The first plot is the Kaplan-Meier Estimates for all patients. We see that the survival probability is decreasing at a nearly constant rate over time. However, we have to remind ourselves that censored data has an impact on the curve.

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

In the equation above, n_i represents the number of patients at risk prior to time t_i, which exclude patients who quit the study. In this case, the survival probability might be over-estimated if the censored patients died soon after leaving the study, or under-estimated if they indeed lived for a long time.
We also plot the three survival curves for patients with different Edema conditions, which seem to be divergent. Formally, we used a log-rank test to verify their difference(p value $\leq 2e-16$). Thus, Endema itself is already a good predictor for the survival paths. We want to see if we have more control variables, what are the other good predictors.

3.2.2 Weibull Estimates

Table 1: Significant p-value results for Weibull (Unweighted)

	Value	Std..Error
(Intercept)	10.00629	1.19371
Age	-0.02328	0.00722
Edema_N	0.52829	0.33768
Edema_S	0.34510	0.33634
Bilirubin	-0.03923	0.01703
Prothrombin	-0.14231	0.06707
Stage	-0.21302	0.09946
Log(scale)	-0.79352	0.07829

BIC of the whole model: 2045.74

Table 2: Significant p-value results for Weibull (Weighted)

	Value	Std..Error
(Intercept)	6.95341	0.72168
DrugPlacebo	0.19219	0.07279
Age	-0.00003	0.00001
Edema_N	0.73800	0.19226
Edema_S	0.54602	0.21209
Bilirubin	-0.05288	0.01135
Alk_Phos	0.00004	0.00002
Platelets	-0.00154	0.00043
Stage	-0.23033	0.05436
Log(scale)	-0.63557	0.05168

BIC of the whole model: 5109.354

The coefficients in the tables are the estimated coefficients and their corresponding standard errors that are significant after using Benjamin-Hochberg correction. Those are not shaded are the ones also significant under the Bonferroni correction, which is more conservative.

The coefficient of the model suggests that Edema and Bilirubin have significant impacts on the survival probability. This coincides with our domain knowledge of the cirrhosis. The weighted model has a higher BIC, which may indicate a lack of fit. It added two more significant coefficients, which indicates that platelets and/or the stage of the disease tend to influence the probability of survival.

The exponential and the Weibull are the only log-linear models that are simultaneously proportional hazards models. Other parametric distributions can be used for survival regression either as a proportional hazards model or as an accelerated failure time model.

Note that bonferroni correction is more conservative in terms of rejecting the nulls compared to Benjamin-Hochberg. Therefore, there are fewer parameters selected used Bonferroni (shaded coefficients are discarded).

3.2.3 Loglog Estimates

Table 3: Significant p-value results for Loglog(Unweighted)

	Value	Std..Error
(Intercept)	10.00629	1.19371
Age	-0.02328	0.00722
Copper	-0.00211	0.00077
Log(scale)	-0.79352	0.07829

BIC of the whole model: 2035.031

Table 4: Significant p-value results for Loglog(Weighted)

	Value	Std..Error
(Intercept)	7.79360	0.81777
DrugPlacebo	0.23269	0.07977
Age	-0.00004	0.00001
Edema_N	1.05025	0.21952
Edema_S	0.57652	0.24840
Bilirubin	-0.03856	0.01506
Alk_Phos	0.00005	0.00002
Platelets	-0.00206	0.00052
Stage	-0.25305	0.05740
Log(scale)	-0.81491	0.05228

BIC of the whole model: 5123.310

Interestingly, we see that when use the complementary log-log model to estimate the survival probability, Billirubin is no longer selected. Age becomes significant instead.

3.2.4 Cox-PH Estimates

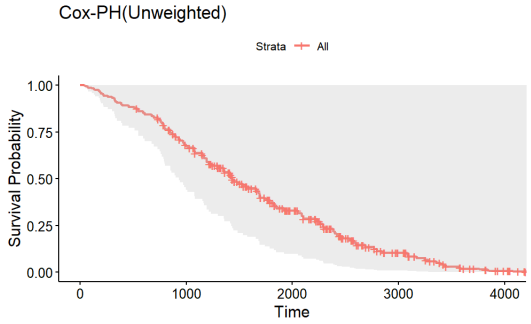

Figure 7: Cox-PH estimates(Unweighted) Estimates for all patients

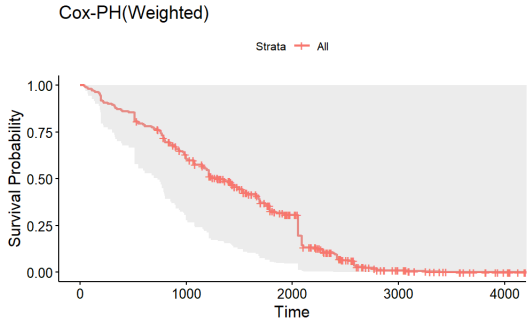

Figure 8: Cox-PH estimates(weighted) Estimates for all patients

Table 5: Significant p-value results for Cox-PH (Unweighted)

	coef	exp.coef.	se.coef.
Bilirubin	0.08	1.084	0.026

BIC of the whole model: 1016.441

Table 6: Significant p-value results for Cox-PH (Weighted)

	coef	exp.coef.	se.coef.	robust.se
Bilirubin	0.10914	1.11532	0.02125	0.03296

BIC of the whole model: 3038.525

The findings from the Cox regression analysis, both before and after weighting, indicate similar significance levels, with Bilirubin being the only significant predictor affecting survival probability. $\exp(\beta_{Bilirubin}) > 1$ suggests that each one-unit increase in Bilirubin is linked to a 10% rise in the hazard rate, pointing to a heightened risk. Similarly, the BIC for the unweighted model is considerably lower.

3.2.5 Competing Risk Model

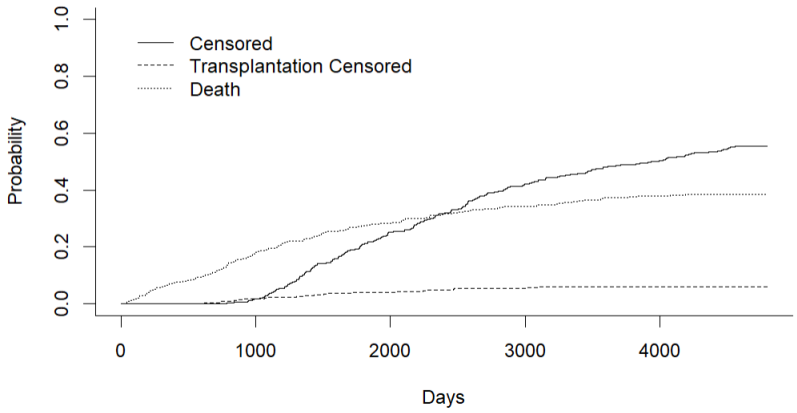

Figure 9: Probability of the Cumulative Incidents

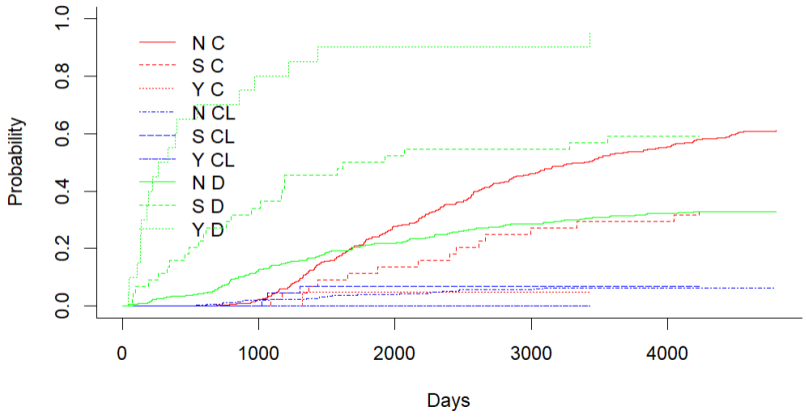

Figure 10: Probability of the Cumulative Incidents for patients with different Endema conditions (N(no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy))
C(Censored),CL(Censored with Transplantation),D(Death)

From the plots above we see that, patients without endema (marked by N) have a cumulative probability of being censored(the red solid line). This may indicate that the healthier they are, the more likely they quit the study. Thus, in order to get a more representative estimates of the survival probability, the hospital may need to take some actions to maintain records for this kind of patients.

4 Discussion

We used non-parametric, parametric, and semi-parametric survival models to assess the influence of factors on the survival probability. We also used inverse probability censoring weighting to account for the possibility of informative censoring. Our selected covariates are diverse across models, and the bast result(in terms of BIC) is given by the unweighted Cox PH model, in which Bilirubin is the only significant predictor.

In every comparison between unweighted and weighted versions of our models, we consistently observe that the weighted models exhibit significantly higher BIC values. One possible reason is that BIC penalizes models for the number of parameters they use. The weighted Cox regression might use more parameters or a more complex weighting scheme, leading to a higher BIC compared to the

simpler unweighted model. Moreover, we should be more careful when making inference from the Cox-PH weighted model because the proportional assumption was tested to be violated. Further study of the interaction between time and covariates may be needed.

The competing risk model shows that while censoring due to transplantation is quite rare, the probability of being censored due to other reasons starts to grow after 1000 days from the beginning of the study. Also, the group curves reveal that patients with different conditions of Endema have divergent paths of experiencing the three events(C,CL,D). Our findings may give some insights for future studies.

5 Limitation

1. In Cox regression, we observe the proportionality assumption is maintained in the unweighted model but becomes violated upon applying weighting. This suggests that the introduction of weighting methods can alter the fundamental relationship between covariates and the hazard rate.

2. We are working with a relatively small sample size, comprising 276 data points, yet it involves as many as 19 covariates. With limited data, these parameter estimations may not be very accurate, leading to unreliable results. Also, there's a higher chance that the dataset may not adequately represent the larger population, leading to issues with generalizing our findings.

5.1 Reference

Dickson,E., Grambsch,P., Fleming,T., Fisher,L., and Langworthy,A.. (2023). Cirrhosis Patient Survival Prediction. UCI Machine Learning Repository. https://doi.org/10.24432/C5R02G.

Fleming, & Harrington. (2005). Counting Processes and Survival Analysis. John Wiley & Sons, Inc.

Putter H, Fiocco M, Geskus RB (2007). Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine 26, 2389–2430.

5.2 Codes

```
library(survival)
library(condSURV)
library(JM)
library(dplyr)
library(survminer)
library(ggplot2)
library(flexsurv)
library(fastDummies)
library(nnet)
library(generalhoslem)
library(mlogit)
library(VGAM)
library(ResourceSelection)
library(modi)
library(cmprsk)

##########Codes for Data Cleaning
cir <- read.csv('cirrhosis.csv')
prop.table(table(cir$Status))
ggplot(data=cir,aes(x=N_Days,fill=Status))+geom_histogram()

####### Test whether the distributions of medical conditions of patients in treatment, and control ar
```

```r
#there are 106 patients that did not participate in the study, we call the group "Other"
#cir£Drug[is.na(cir£Drug)] <- "Other"

cir$Sex <- ifelse(cir$Sex=="F",1,0)
cir$Ascites <- ifelse(cir$Ascites=="Y",1,0)
cir$Hepatomegaly <- ifelse(cir$Hepatomegaly=="Y",1,0)
cir$Spiders <- ifelse(cir$Spiders=="Y",1,0)
cir <- fastDummies::dummy_cols(cir, select_columns = "Edema")

# Assuming 'treatment' is the categorical treatment indicator
# and 'covariates' is a data frame containing relevant covariates
cir$Drug <-as.factor(cir$Drug)

cirr <- data.frame(cir)
cirr$Drug <- ifelse(cirr$Drug=="D-penicillamine",1,0)
cirr <- na.omit(cirr)

#Since Age and Sex are already randomized before the study, we exclude them here
model <- glm(Drug~Ascites+Hepatomegaly+Spiders+Edema+Bilirubin+Cholesterol+Albumin+Copper+Alk_Phos+SG

# Deviance test
null_model <- glm(cirr$Drug ~ 1, family = binomial)
null_deviance <- null_model$deviance
fitted_deviance <- deviance(model)
df_diff <- df.residual(null_model) - df.residual(model)

chi2_stat <- null_deviance - fitted_deviance
p_value <- pchisq(chi2_stat, df_diff, lower.tail = FALSE)
p_value


### Non-parametric


cir$Death <- ifelse(cir$Status == "D",1,0)
km <- survfit(Surv(N_Days, Death) ~ 1, data = cir)
print(km, print.rmean = TRUE)


plot(km, xlab = "Survival Time In Days", ylab = "Survival Probabilities")

### Determine whether there are significant differences in the fitted survival distributions using a
survdiff(Surv(N_Days, Death) ~ Edema, data = cir,rho=0)

### Parametric weibull

cir2 <- data.frame(cir)
cir2$Drug <- ifelse(cir2$Drug=="D-penicillamine",1,0)
model1 <- survreg(Surv(N_Days, Death) ~ Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bil
model1_p <- summary(model1)$table[,4]
#model1_p <- na.omit(model1_p)


model1null <- survreg(Surv(N_Days, Death) ~ ., data = cir, dist = "weibull")
#model1_p_bonf[model1_p_bonf<=alpha]
#model1_p_BH[model1_p_BH<=alpha]
```

```r
#since we have many parameters to estimate and test
#we use bonferroni correction to account for the multiple testing issues
model1_p_bonf <- p.adjust(model1_p, method="bonferroni")
model1_p_BH <- p.adjust(model1_p, method="BH")

alpha=0.05

coef_df <- data.frame(summary(model1)$table[,c(1,2)])
coef_df <- round(coef_df,3)
coef_df[which(model1_p_BH<=alpha),]
coef_df[which(model1_p_bonf<=alpha),]

library(xtable)
xtable(coef_df[which(model1_p_BH<=alpha),],digits=c(0,5,5))

xtable(data.frame(BIC(model1)))

### Parametric loglogistic

model2 <- survreg(Surv(N_Days, Death) ~ Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bil
model2_p <- summary(model2)$table[,4]
model2_p <- na.omit(model2_p)

model2_p_bonf <- p.adjust(model2_p, method="bonferroni")
model2_p_BH <- p.adjust(model2_p, method="BH")

model2_p_bonf[model2_p_bonf<=alpha]
model2_p_BH[model2_p_BH<=alpha]

coef_df <- data.frame(summary(model2)$table[,c(1,2)])
coef_df <- round(coef_df,5)
coef_df[which(model2_p_BH<=alpha),]
coef_df[which(model2_p_bonf<=alpha),]

BIC(model2)
library(xtable)
xtable(coef_df[which(model2_p_BH<=alpha),],digits=c(0,5,5))

#### With weights

cirr1 <- na.omit(cir)
model <- glm(Death~Ascites+Hepatomegaly+Spiders+Edema+Bilirubin+Cholesterol+Albumin+Copper+Alk_Phos+S

cirr1$score <- predict(model, type = "response")

weight1<-numeric(length(cirr1$score))
for (i in 1:length(cirr1$score)){
  if(cirr1$Status[i]=="D"){
    weight1[i]<-1/(cirr1$score[i])
  }else{
    weight1[i]<-1/(1-cirr1$score[i])
  }
}

cirr1$Status <- ifelse(cirr1$Status=="D","Death","Censored")
```

```r
cirr1 <- cirr1[,c("Bilirubin","Cholesterol","Albumin","Copper","Alk_Phos","SGOT","Tryglicerides","Pla

df2 <- c()

# Calculate standardized mean difference for variable A
for(i in 1:(ncol(cirr1)-1)){
  mean_death <- weighted.mean(cirr1[cirr1$Status == "Death",i],weight1[which(cirr1$Status == "Death")
  mean_censor <- weighted.mean(cirr1[cirr1$Status == "Censored",i],weight1[which(cirr1$Status == "Cen
  sd_death <- sqrt(weighted.var(cirr1[cirr1$Status == "Death",i],weight1[which(cirr1$Status == "Death
  sd_censor <- sqrt(weighted.var(cirr1[cirr1$Status == "Censored",i],weight1[which(cirr1$Status == "C

  abs_smd <- abs((mean_death - mean_censor) / sqrt((sd_death^2 + sd_censor^2) / 2))
  df2 <- append(df2,abs_smd)
}

dff2 <- data.frame(covariates=c("Bilirubin","Cholesterol","Albumin","Copper","Alk_Phos","SGOT","Trygl

dff$grp <- "Unweighted"
dff2$grp <- "Weighted"
df_all <- rbind(dff,dff2)

ggplot(df_all, aes(covariates,abs_smd,col=grp))+
  geom_point(size=3)+
   theme_minimal()+
    theme(
    axis.title.x = element_text(size = 14),  # Adjust x-axis title size
    axis.title.y = element_text(size = 14),  # Adjust y-axis title size
    axis.text.x = element_text(angle = 45,size = 12),   # Adjust x-axis labels size
    axis.text.y = element_text(size = 12),    # Adjust y-axis labels size
    legend.title = element_text(size = 14),  # Adjust legend title size
    legend.text = element_text(size = 12)
  )+
  labs(y="absolute standardized mean difference")

### Parametric weibull (Weighted)
m1 <- survreg(Surv(N_Days, y) ~ Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bilirubin+C
s1<-summary(m1)

print(paste("AIC:", AIC(m1), "BIC:", BIC(m1)))

p_values<-s1$table[,4]
adjusted_p_values <- p.adjust(p_values, method = "bonferroni")
print(adjusted_p_values)
adjusted_p_values <- p.adjust(p_values, method = "BH")
print(adjusted_p_values)

#### Assumption checking
# Compute the cumulative hazard
s0<-summary(m0)
#km_cumhaz <- log(m0$surv)
plot(log(s0$time), log(s0$cumhaz), xlab = "Log Time", ylab = "Log Cumulative Hazard", cex=0.7,pch=21,

### Parametric Log logistic (Weighted)
m2<-survreg(Surv(N_Days, y) ~ Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bilirubin+Cho
s2<-summary(m2)
print(paste("AIC:", AIC(m2), "BCI:", BIC(m2)))
```

```r
p_values<-s2$table[,4]
adjusted_p_values <- p.adjust(p_values, method = "BH")
coef_df <- data.frame(summary(m2)$table[,c(1,2)])
coef_df[which(adjusted_p_values<=alpha),]

adjusted_p_values <- p.adjust(p_values, method = "bonferroni")
coef_df <- data.frame(summary(m2)$table[,c(1,2)])
coef_df <- round(coef_df,5)
coef_df[which(adjusted_p_values<=alpha),]

### Semi-parametric (Cox)

#### Unweighted
model3 <- coxph(Surv(N_Days, Death)~Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bilirub

model3_p <- summary(model3)$coefficient[,5]
model3_p <- na.omit(model3_p)

model3_p_bonf <- p.adjust(model3_p, method="bonferroni")
model3_p_BH <- p.adjust(model3_p, method="BH")

model3_p_bonf[model3_p_bonf<=alpha]
model3_p_BH[model3_p_BH<=alpha]

coef_df <- data.frame(summary(model3)$table[,c(1,2)])
coef_df <- round(coef_df,5)
coef_df[which(model3_p_BH<=alpha),]
coef_df[which(model3_p_bonf<=alpha),]

library(xtable)
xtable(coef_df[which(model3_p_BH<=alpha),],digits=c(0,5,5,5,5))

BIC(model3)

#### Weighted
m3 <- coxph(Surv(N_Days, y) ~ Drug+Age+Sex+Ascites+Hepatomegaly+Spiders+Edema_N+Edema_S+Bilirubin+Cho

s3<-summary(m3)
print(paste("AIC:", AIC(m3), "BIC:", BIC(m3)))

p_values<-summary(m3)$coefficients[,6]
adjusted_p_values <- p.adjust(p_values, method = "BH")
coef_df <- data.frame(summary(m3)$coefficients[,c(1,2,3,4)])
coef_df[which(adjusted_p_values<=alpha),]

adjusted_p_values <- p.adjust(p_values, method = "bonferroni")
coef_df <- data.frame(summary(m3)$coefficients[,c(1,2,3,4,6)])
coef_df <- round(coef_df,3)
coef_df[which(adjusted_p_values<=alpha),]

### test proportional hazard function
test_results1 <- cox.zph(m3)
test_results2 <- cox.zph(l3)

### Cox survival function
```

```r
surv_curve<-survfit(l3)
ggsurvplot(surv_curve, data = cir,  conf.int = TRUE,
           title = "Cox-PH(Unweighted)",
           xlab = "Time", ylab = "Survival Probability")

surv_curve<-survfit(m3)
ggsurvplot(surv_curve, data = cir,  conf.int = TRUE,
           title = "Cox-PH(Weighted)",
           xlab = "Time", ylab = "Survival Probability")
```