STAT 5140: Final Project

Andy Chuang, Kevin Nguyen, & Dominic Matriccino

May 5, 2023

1 Introduction to Data

1.1 Data Overview

The Mayo Clinic Primary Biliary Cirrhosis Data is a comprehensive dataset containing information on 424 patients with primary biliary cirrhosis (PBC), who were referred to the clinic over a ten-year period. Among these patients, 312 participated in a randomized placebo-controlled trial of the drug D-penicillamine and provided largely complete data. The remaining 112 patients did not participate in the trial but consented to have basic measurements recorded and to be followed for survival. This dataset includes data on both the randomized participants and the additional 106 cases, as well as six cases that were lost to follow-up shortly after diagnosis. The dataset provides a valuable resource for researchers studying PBC and related conditions.

1.2 Variable Description

- ID: Case number
- Time: Number of days between registration and the earlier of death, transplantation, or study analysis time in July, 1986

- Status: 0 = Censored, 1 = Censored due to liver transplant, 2 = Death
- Treatment: 1 = D-penicillamine, 2 = Placebo
- Age: Age in days
- Sex: 0 = Male, 1 = Female
- Ascites: Presence of ascites (0 = No, 1 = Yes)
- Hepato: Presence of hepatomegaly (0 = No, 1 = Yes)
- Spiders: Presence of spiders (0 = No, 1 = Yes)
- Edema: Presence of edema (0 = No edema and no diuretic therapy for edema, 0.5 = Edema present without diuretics or edema resolved by diuretics, 1 = Edema despite diuretic therapy)
- Bili: Serum bilirubin in mg/dl
- Chol: Serum cholesterol in mg/dl
- Albumin: Albumin in gm/dl
- Copper: Urine copper in ug/day
- Alk.phos: Alkaline phosphatase in U/liter
- Ast: SGOT in U/ml
- Trig: Triglicerides in mg/dl
- Platelet: Platelets per cubic ml/1000
- Protime: Prothrombin time in seconds
- Stage: Histologic stage of disease

2 Questions to Address

For this project, we have one primary question and two secondary questions that we aim to address:

1. **Primary Question**: What is the effect of risk factors on the survival of patients with PBC?

2. **Secondary Question 1**: Can we devise a method to test if type 1 censoring (status = 1, i.e., due to liver transplantation) is informative or non-informative?

3. **Secondary Question 2**: How should we deal with the type 1 censoring (status = 1, i.e., due to liver transplantation)?

3 Statistical Analysis

To improve the interpretability and ease of use in our statistical analysis, we conducted some preliminary data cleaning by converting the `age` and `time` variables from a per-day basis to a per-year basis. This conversion is more conventional and easier to interpret, which will alleviate any potential scaling issues that may arise during statistical analysis. Additionally, this conversion can alleviate potential scaling issues that may arise during statistical analysis. We also converted categorical variables from numerical to factor, which will allow us to run our statistical analysis more effectively. Finally, we simplified the `status` variable, which has three levels of 0 = censored, 1 = censored due to liver transplant, and 2 = death, by combining the censored levels to improve clarity. We combined the "0" and "1" levels as our new first level and used the "2" level as our new second level.

We utilized the Kaplan-Meier (KM) estimator to conduct preliminary elementary calculations on our categorical variables. This nonparametric method helps estimate the survival

function by observing changes in survival probabilities over time at event times. Our objective was to investigate the impact of variables of interest on survival time and assess their significance. Using the KM estimator, we can identify differences in survival probabilities among levels of a categorical variable, which will enable us to answer our primary question of interest.

We have plotted the KM estimates for the following variables in the left to right order, as shown in Figure 1: `sex`, `ascites`, `hepato`, `spiders`, `edema`, `stage`, and `treatment`.

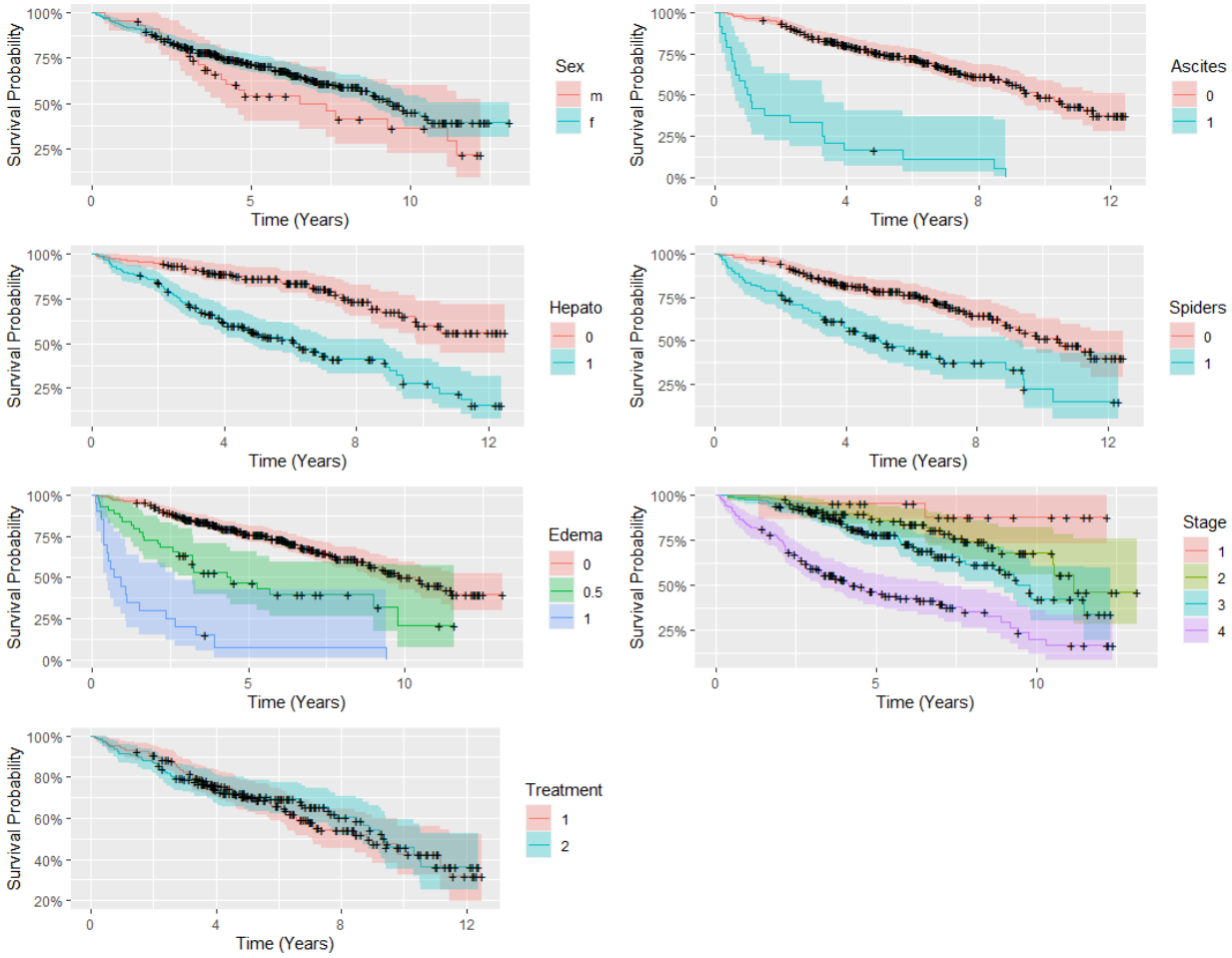

Figure 1: Kaplan-Meier Estimates

Although the KM estimator provides valuable insights into the significance of categorical

variables, the dataset under consideration also comprises of numerical variables. Hence, to study the impact of risk factors on the survival of patients with PBC, we leverage the Cox proportional hazards model. We use the `coxph()` function from the `survival` package in R to construct our model with every predictor, both numerical and categorical, after removing rows with missing data. However, we are interested not only in the effects of risk factors on patient survival but also in the interactions between those risk factors. Therefore, we incorporate two interaction terms into our model, one between the `treatment` and `age` variables and another between the `age` and `cholesterol` variables. The inclusion of interaction terms between `age` and `cholesterol`, and between `age` and `treatment`, is medically justified. It is well-established that age is an important factor to consider as it can affect the progression of the disease and the response to treatment. Cholesterol levels also tend to increase with age, particularly in men and women after the age of 50. Furthermore, older patients may have different treatment responses, adverse events, or medication tolerances compared to younger patients, highlighting the importance of assessing treatment effectiveness in different age groups. After running a full model, we discovered that several predictors were insignificant. To overcome this challenge and obtain the optimal model, we used the `stepAIC()` function from the `MASS` package to perform backward stepwise regression, which selects the model with the smallest Akaike Information Criterion (AIC). We present the output of the `summary()` function on this model in Figure 2.

```
coxph(formula = Surv(time, status) ~ age + edema + bili + albumin +
    copper + ast + protime + stage, data = pbc2)

  n= 276, number of events= 111

               coef  exp(coef)   se(coef)      z Pr(>|z|)
age        0.0313188  1.0318144  0.0102909  3.043  0.00234 **
edema0.5   0.1598036  1.1732804  0.3054890  0.523  0.60090
edema1     0.9121653  2.4897075  0.3545431  2.573  0.01009 *
bili       0.0869270  1.0908171  0.0198104  4.388 1.14e-05 ***
albumin   -0.7387129  0.4777284  0.2792471 -2.645  0.00816 **
copper     0.0027867  1.0027905  0.0009912  2.811  0.00493 **
ast        0.0039562  1.0039641  0.0018415  2.148  0.03168 *
protime    0.2642049  1.3023951  0.1122859  2.353  0.01862 *
stage2     1.3596258  3.8947355  1.0808743  1.258  0.20843
stage3     1.6823556  5.3782101  1.0478753  1.605  0.10839
stage4     2.0627073  7.8672396  1.0432133  1.977  0.04801 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
age          1.0318     0.9692    1.0112    1.0528
edema0.5     1.1733     0.8523    0.6447    2.1352
edema1       2.4897     0.4017    1.2427    4.9881
bili         1.0908     0.9167    1.0493    1.1340
albumin      0.4777     2.0932    0.2764    0.8258
copper       1.0028     0.9972    1.0008    1.0047
ast          1.0040     0.9961    1.0003    1.0076
protime      1.3024     0.7678    1.0451    1.6230
stage2       3.8947     0.2568    0.4682   32.3981
stage3       5.3782     0.1859    0.6897   41.9364
stage4       7.8672     0.1271    1.0182   60.7865

Concordance= 0.845  (se = 0.019 )
Likelihood ratio test= 165.7  on 11 df,   p=<2e-16
Wald test            = 176.9  on 11 df,   p=<2e-16
Score (logrank) test = 278.8  on 11 df,   p=<2e-16
```

Figure 2: Summary Output of Reduced Cox Proportional Hazards Model

After fitting our reduced Cox proportional hazards model, we performed model validation using residual analysis. We plotted the Martingale residuals against the numerical variables in our model, as shown in Figure 3.

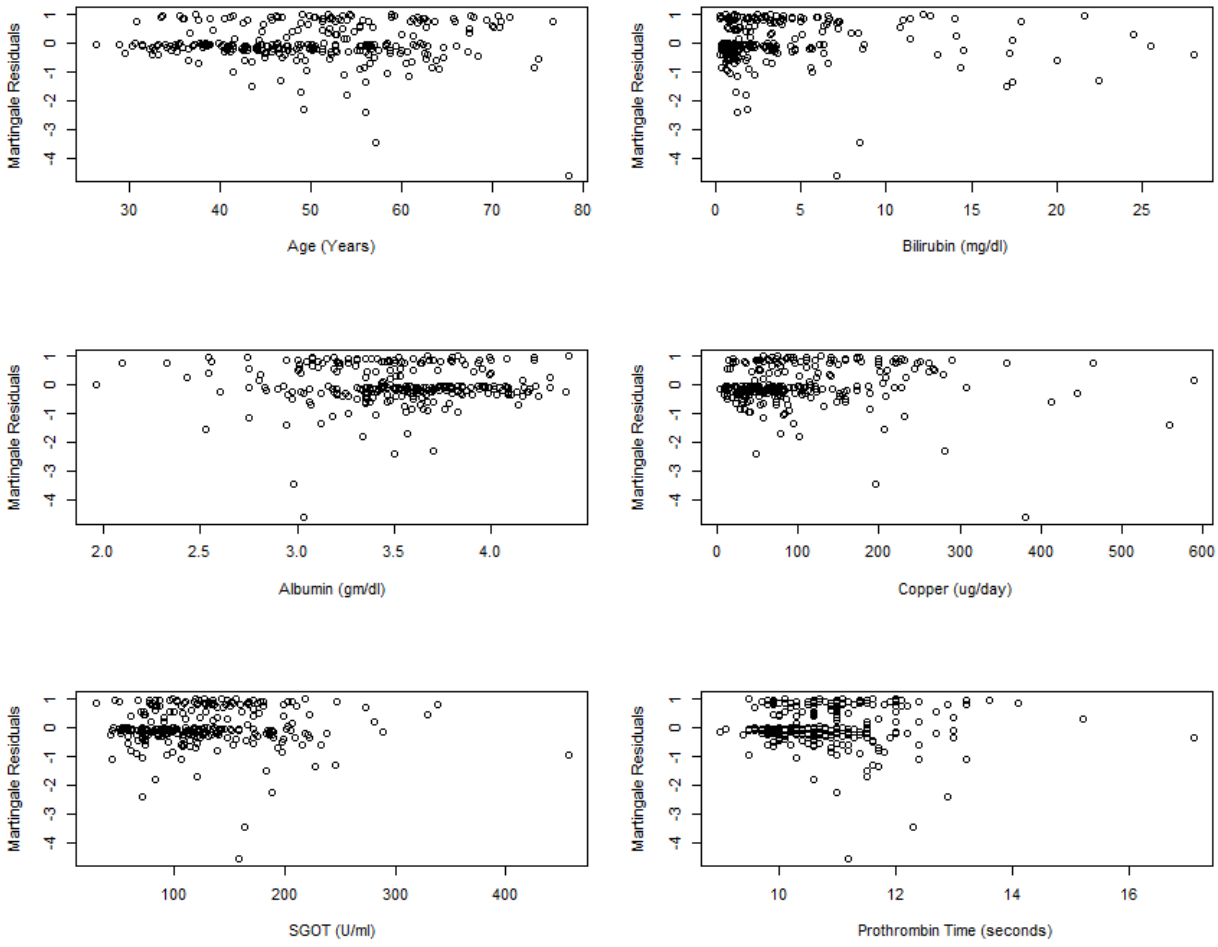

Figure 3: Martingale Residuals vs. Numerical Variables

Figure 3 shows that all variables displayed values close to 0 with no clear patterns or trends. Additionally, the shape of the residuals was roughly linear, indicating that the linearity assumption was satisfied. Based on these findings, we concluded that it is appropriate to retain these variables in our model.

We proceeded to fit a parametric regression model using the `survreg()` function from the

`survival` package. This model utilized all the variables in the dataset, including our two interaction terms, and was fitted to the dataset with the removed missing values. To ensure that the baseline hazard function followed a Weibull distribution, we specified `dist = "weibull"` in the model. However, some predictors still had high p-values, prompting us to select the optimal model using the lowest AIC through backward stepwise regression. We present the output of the `summary()` function on this model in Figure 4.

```
survreg(formula = Surv(time, status) ~ age + edema + bili + albumin +
    copper + ast + protime + stage, data = pbc2, dist = "weibull")
               Value Std. Error     z       p
(Intercept)  5.466526   1.363616  4.01 6.1e-05
age         -0.019130   0.006241 -3.07  0.0022
edema0.5    -0.115817   0.184296 -0.63  0.5297
edema1      -0.620899   0.207196 -3.00  0.0027
bili        -0.051030   0.011213 -4.55 5.3e-06
albumin      0.417757   0.164181  2.54  0.0109
copper      -0.001803   0.000593 -3.04  0.0024
ast         -0.002500   0.001108 -2.26  0.0241
protime     -0.171826   0.069318 -2.48  0.0132
stage2      -0.858934   0.664119 -1.29  0.1959
stage3      -1.030796   0.645908 -1.60  0.1105
stage4      -1.265734   0.643263 -1.97  0.0491
Log(scale)  -0.501244   0.075565 -6.63 3.3e-11

Scale= 0.606

Weibull distribution
Loglik(model)= -313   Loglik(intercept only)= -398.2
        Chisq= 170.45 on 11 degrees of freedom, p= 9.5e-31
Number of Newton-Raphson Iterations: 7
n= 276
```

Figure 4: Summary Output of Reduced Weibull Parametric Regression Model

The `survreg()` function estimates coefficients in the log-hazard scale, meaning that the coefficients represent changes in the logarithm of the hazard rate per unit change in the corresponding predictor variable. However, these coefficients are not easily interpretable in this form. To obtain more meaningful and commonly used results in survival analysis, we must take the exponential of the coefficients to interpret them in terms of hazard ratios.

The hazard ratio represents the multiplicative change in the hazard rate associated with a one-unit increase in the predictor variable, holding all other variables constant. Therefore, taking the exponent of the coefficients is necessary for proper interpretation of the results and for making meaningful conclusions about the effects of the predictor variables on the survival outcome. As a result, the exponentiated coefficients are displayed in Figure 5.

(Intercept)	age	edema0.5	edema1	bili	albumin	copper	ast	protime	stage2	stage3	stage4
236.6367671	0.9810516	0.8906380	0.5374612	0.9502503	1.5185519	0.9981985	0.9975033	0.8421254	0.4236134	0.3567229	0.2820322

Figure 5: Exponentiated Coefficients of Weibull Parametric Regression Model

Subsequently, we validate our model once more by conducting residual analysis. This involves computing the Cox-Snell residuals using our reduced parametric regression model, and generating a plot of the Nelson-Aalen cumulative hazard against the residuals, as shown in Figure 6.

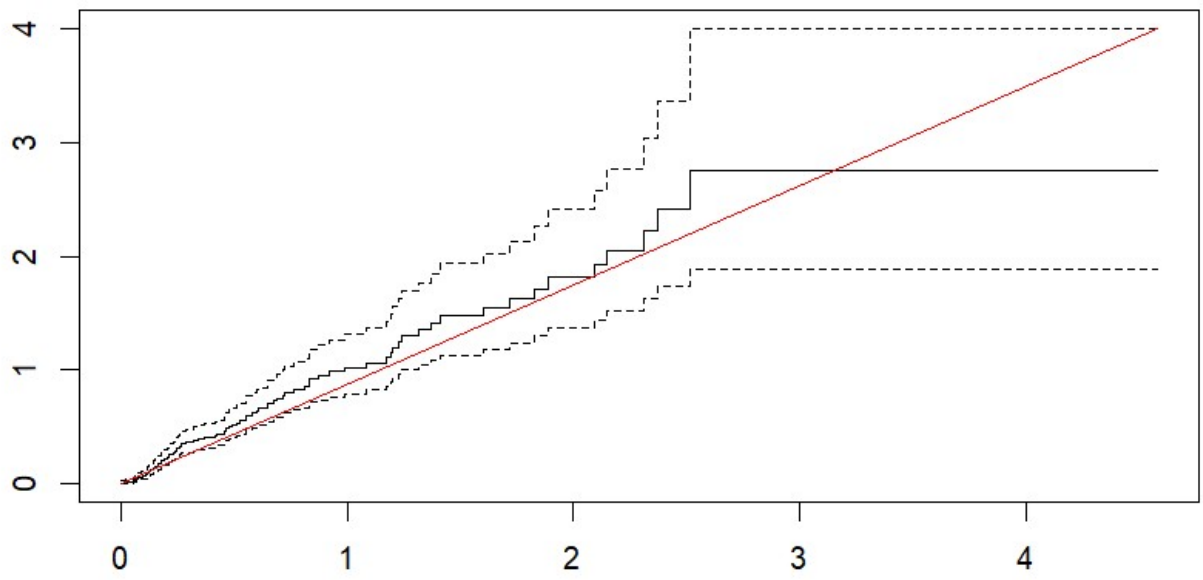

Figure 6: Nelson-Aalen Cumulative Hazard vs. Cox-Snell Residuals

Based on the plot in Figure 6, it can be observed that the Cox-Snell residuals display a reasonably linear pattern, indicating that our general diagnostic requirements have been met.

4 Results

4.1 Interpretation of Analysis

Figure 1 reveals several important findings. The plots for `sex` and `treatment` demonstrate that male and female patients, as well as those who received treatment and those who received a placebo, have similar results. This is evident as the lines representing both groups overlap throughout the study. The `ascites` plot reveals a stark contrast, with all patients with ascites dying before the end of the study and a sharp decline in their survival probability compared to those without ascites, whose probability of survival decreased more gradually over time. The `hepato` and `spiders` plots indicate that the presence of hepatomegaly or spiders corresponds with a lower chance of survival. Finally, the `edema` and `stage` plots illustrate that an increase in the level of edema or PBC stage is associated with a decrease in one's probability of survival, and it is worth noting that none of the patients in the "1" level of the edema variable survived until the end of the study.

Figure 2 shows the exponentiated coefficients (`exp(coef)`) which represent the factor by which the hazard (risk of death) changes for every one-unit increase in the predictor variable while holding all other variables constant. The hazard of death increases by a factor of 1.03 (or 3%) for every one-unit increase in `age`, and by a factor of 2.49 (or 149%) for every one-unit increase in `edema1`. Similarly, for every one-unit increase in `bili`, the hazard of death increases by a factor of 1.09 (or 9%), while for every one-unit increase in `albumin`, it decreases by a factor of 0.48 (or 48%). The hazard of death does not change significantly with a one-unit increase in `copper` or `ast`, or with `stage2`, `stage3`, or `edema0.5`. However, for every one-unit increase in `protime`, the hazard of death increases by a factor of 1.30 (or 30%). Additionally, for every one-unit increase in `stage4`, the hazard of death increases by a factor of 7.87 (or 687%).

The concordance statistic, with a value of 0.845, shows that the model has a high ability to correctly differentiate between patients who live longer versus those who do not. This means that the model can accurately predict the likelihood of survival based on the predictor variables. The likelihood ratio test, Wald test, and score (log-rank) test all indicate that the model is highly significant and provides a good fit to the data. These tests suggest that the predictor variables in the model are strongly associated with survival time, and therefore the model is a reliable tool for predicting survival outcomes.

Looking closely at Figure 5, we can see how different factors impact survival. We found that only one variable, `albumin`, had a positive effect on survival. For every one-unit increase in `albumin`, there was a 51% increase in survival probability, holding all other variables constant. The p-value for `albumin` was 0.0109, which is significant at the level of 0.05. On the other hand, `age` had a negative effect on survival. For each one-year increase in `age`, the survival probability decreased by 2% while holding all other variables constant. The p-value for `age` was lower than the p-value for `albumin`, indicating stronger evidence in favor of the fact that age does provide a more pronounced change in survival time, although the magnitude of the change is less. The stages of cirrhosis were also related to worse survival times, with the coefficients being definitively negative and the magnitudes relatively large. Out of the three stage variables, only the stage4 variable was significant, which shortened survival time by approximately 72%. Our top three variables in terms of statistical significance were `bili`, `age`, and `copper`.

4.2 Addressing Questions of Interest

Primary Question: Figure 1 shows that patients with ascites have a much lower chance of survival compared to those without ascites. The presence of hepatomegaly or spiders also corresponds with a lower chance of survival. An increase in the level of edema or PBC stage

is associated with a decrease in one's probability of survival. Figure 2 shows that as patients age or have higher levels of bilirubin, copper, or protime, their risk of death increases. On the other hand, higher levels of albumin decrease the risk of death. The stages of cirrhosis were also related to worse survival times, with higher stages indicating a shorter survival time. Figure 5 shows that albumin is the only variable that had a positive effect on survival. Age had a negative effect on survival, and the stages of cirrhosis were related to worse survival times. Overall, the information suggests that different risk factors can have a significant impact on the survival of patients with PBC.

Secondary Question 1: Censoring in survival analysis can be informative or non-informative, depending on whether it's related to survival time and underlying risk factors or not. Type 1 censoring due to liver transplantation may be informative if patients who are censored for this reason have a different survival experience than those censored for other reasons. To determine the informativeness of type 1 censoring, we can compare the survival curves of patients censored due to liver transplantation with those censored for other reasons. If the survival curves are similar, the censoring is likely non-informative, but if they differ significantly, the censoring is likely informative. The log-rank test or Wilcoxon test can be used to compare survival curves among different groups and assess whether the observed differences in survival are statistically significant. If the p-value of the test is less than a predetermined level of significance, such as 0.05, we can reject the null hypothesis that there is no difference in survival between the groups and conclude that censoring due to liver transplantation is likely to be informative. Conversely, if we fail to reject the null hypothesis, we conclude that censoring due to liver transplantation is likely to be non-informative.

Secondary Question 2: We decided to combine type 1 censoring with type 0 censoring. The main reason for combining type 1 censoring with type 0 censoring is to avoid loss of information and potential bias in the analysis. Dropping type 1 censoring would result in a

reduction of sample size and loss of valuable information, potentially leading to inaccurate estimates of the survival function and bias in the analysis. By combining type 1 censoring with type 0 censoring, we can utilize all the available information to estimate the survival function and evaluate the impact of potential risk factors on survival outcomes. Additionally, this approach allows us to consider the possibility that patients who were censored due to liver transplantation may have a different survival experience than those who were censored for other reasons.

5 Discussion

As previously mentioned, both of our models employed backward stepwise regression with AIC as the criterion for selecting the best model. However, AIC has limitations as a feature selection method. It can choose an overfitted model and prefers more complex models when the improvement in fit is small relative to the increase in parameters. AIC also assumes unbiased and normally distributed maximum likelihood estimates and overlooks estimation uncertainty, which can impact parameter inference or prediction based on the model. Moving on to our Cox proportional hazards model, it is flexible, interpretable, handles censored data, and does not require normality assumptions. However, it is sensitive to outliers, assumes constant hazard ratios over time, does not estimate baseline hazard, and can be challenging to assess model fit. Conversely, our parametric regression model is flexible, efficient, and interpretable, but depends on assumptions about the distribution of the time-to-event outcome. Additionally, a significant limitation of our data is the absence of information on race/ethnicity, making it challenging to generalize our findings. Cirrhosis, a chronic liver disease with multiple causes such as alcohol consumption, viral infections, and metabolic disorders, can have different progressions and outcomes depending on factors like genetics, lifestyle, and comorbidities. By not considering race, the study may fail to detect significant differences in the prevalence, incidence, and outcomes of cirrhosis across various racial and

ethnic groups. Certain racial groups may be more susceptible to certain types of cirrhosis due to genetic predispositions or cultural practices, like alcohol consumption.

Appendix

```
library(survival)
library(tables)
library(ggplot2)
library(ggfortify)
library(gridExtra)
library(MASS)

data(pbc)
table(pbc$status)
pbc$status <- ifelse(pbc$status == 2, 1, 0)
pbc$time <- pbc$time/365.25
colSums(is.na(pbc))

# Data cleaning
pbc$edema <- as.factor(pbc$edema)
pbc$trt <- as.factor(pbc$trt)
pbc$sex  <- as.factor(pbc$sex)
pbc$ascites <- as.factor(pbc$ascites)
pbc$hepato <- as.factor(pbc$hepato)
pbc$spiders <- as.factor(pbc$spiders)
pbc$stage <- as.factor(pbc$stage)
pbc2 <- na.omit(pbc)

km_fit2 <- survfit(Surv(time, status) ~ sex, data = pbc, type = "kaplan-meier")
g1 <- autoplot(km_fit2) + labs(fill = "Sex", color = "Sex", x = "Time (Years)",
                y = "Survival Probability")
km_fit3 <- survfit(Surv(time, status) ~ ascites, data = pbc, type = "kaplan-meier")
g2<- autoplot(km_fit3)  + labs(fill = "Ascites", color = "Ascites", x = "Time (Years)",
                y = "Survival Probability")
km_fit4 <- survfit(Surv(time, status) ~ hepato, data = pbc, type = "kaplan-meier")
g3<- autoplot(km_fit4) + labs(fill = "Hepato", color = "Hepato", x = "Time (Years)",
                y = "Survival Probability")
km_fit5 <- survfit(Surv(time, status) ~ spiders, data = pbc, type = "kaplan-meier")
g4 <- autoplot(km_fit5) + labs(fill = "Spiders", color = "Spiders", x = "Time (Years)",
                y = "Survival Probability")
km_fit6 <- survfit(Surv(time, status) ~ edema, data = pbc, type = "kaplan-meier")
g5 <- autoplot(km_fit6) + labs(fill = "Edema", color = "Edema", x = "Time (Years)",
                y = "Survival Probability")
km_fit16 <- survfit(Surv(time, status) ~ stage, data = pbc, type = "kaplan-meier")
g6<-autoplot(km_fit16) + labs(fill = "Stage", color = "Stage", x = "Time (Years)",
                y = "Survival Probability")
km_fit17 <- survfit(Surv(time, status) ~ trt, data = pbc, type = "kaplan-meier")
g7<-autoplot(km_fit17) + labs(fill = "Treatment", color = "Treatment", x = "Time (Years)",
                y = "Survival Probability")
```

```
grid.arrange(g1,g2,g3,g4,g5,g6,g7,nrow = 4, ncol = 2)

# Coxph
model <- coxph(Surv(time, status) ~ trt + age + sex + ascites + hepato +
               spiders + edema + bili + chol + albumin + copper + alk.phos + ast +
               trig + platelet + protime + stage + age:chol + trt:age, data = pbc2)
summary(model)
step <- stepAIC(model, direction = "backward")
summary(step)

mr <- residuals(step, type = "martingale")

par(mfrow = c(3,2))
plot(pbc2$age, mr, xlab = "Age (Years)", ylab = "Martingale Residuals")
plot(pbc2$bili, mr, xlab = "Bilirubin (mg/dl)", ylab = "Martingale Residuals")
plot(pbc2$albumin, mr, xlab = "Albumin (gm/dl)", ylab = "Martingale Residuals")
plot(pbc2$copper, mr, xlab = "Copper (ug/day)", ylab = "Martingale Residuals")
plot(pbc2$ast, mr, xlab = "SGOT (U/ml)", ylab = "Martingale Residuals")
plot(pbc2$protime, mr, xlab = "Prothrombin Time (seconds)", ylab = "Martingale Residuals")

ret <- survreg(Surv(time, status) ~ trt + age + sex + ascites + hepato +
               spiders + edema + bili + chol + albumin + copper + alk.phos + ast +
               trig + platelet + protime + stage + age:chol + trt:age, dist = "weibull", data = pbc2)
summary(ret)
step2 <- stepAIC(ret, direction = "backward")
summary(step2)

CS <- pbc2$status - residuals(step)
plot(survfit(Surv(CS, pbc2$status) ~ 1, type = "flem"), fun = "cumhaz")
lines(x = c(0, 4.58), y = c(0, 4), type = "l", lty = 1, col ="red")
```