Hailey Jensen
Joshua Lazaro
Nina De La Torre

Cirrhosis Final Report STA 4903

1. Background

According to the National Liver Foundation, in the United States alone more than 100 million people have some kind of liver disease, and in 2020 liver disease/cirrhosis was the 12th leading cause of death in America. Not only can cirrhosis be of immense damage to your liver but it also increases the risk of having a stroke. Cirrhosis is also the most common reason to get a liver transplant, which can be a risky procedure given that you can even find a donor in the first place. Liver disease is often caused by heavy alcohol use, tattoos/piercings, and obesity to name a few. In our project, we will be studying a dataset collected from patients who were diagnosed with Primary Biliary Cirrhosis (or PBC, but it is now called Primary Biliary *Cholangitis*). This type of liver disease is an autoimmune disease, meaning that the patient's immune system attacks healthy cells of the liver. PBC is a chronic disease where the liver's bile ducts become inflamed until they eventually collapse and are destroyed. Bile helps with a person's digestion, and having PBC can lead to death. Although liver cancer is twice more likely in men than women, Primary Biliary Cirrhosis is actually the other way around, as the ratio of women to men with PBC is 9 to 1. Conducting a survival analysis on cirrhosis patients is vital to understanding what covariates are most important when determining the severity of the disease. This analysis will also help us understand the expected lifetimes of patients, and whether or not the therapeutic drugs used were effective in treating the patients.

2. Data

The data was collected between 1974 and 1984 from a study conducted at the Mayo Clinic, a women's health center. The clinic was conducting a trial where patients with Primary Biliary Cirrhosis met eligibility requirements to participate, and were randomly given either a placebo or the drug D-penicillamine. There were initially 312 patients who participated in the actual trial, then for another 106 patients although they did not participate in the trial they consented to having basic measurements taken and were followed by survival. Out of the total 418 observations we only used the first 312 for our study, since there were a lot of missing values for the additional patients. Of the 312 observations used there are 276 females and 36 males. One of the variables indicated whether a patient lived, died, or had a liver transplant. There were only 19 patients who had a liver transplant and because having a liver transplant only happens for patients in dire situations we also decided to remove it from our study so that they wouldn't interfere with our censoring variable.

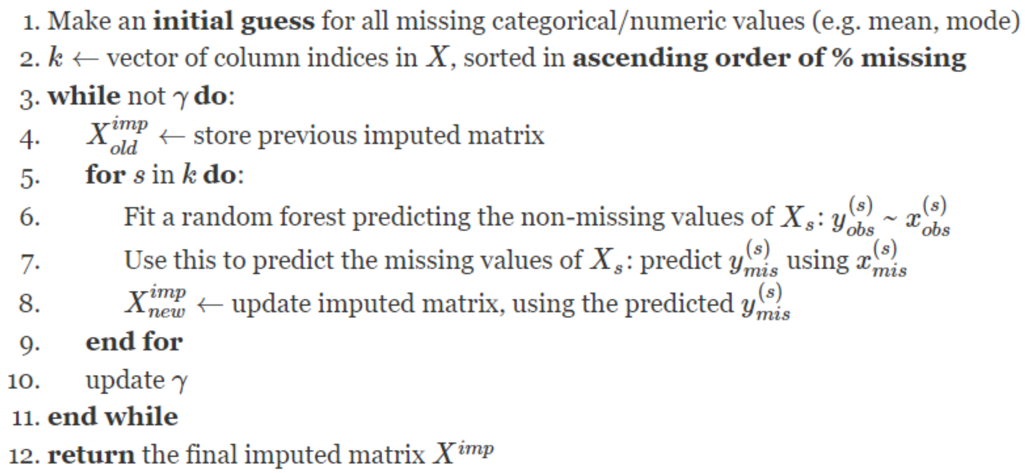
The data consisted of 22 covariates, 9 categorical and 13 numeric and are as follows: **Id:** identification **Time:** Number of days between registration and the earlier of death. **Event:** dead (0) or alive (1) ; we took out the observations of those who had a liver transplant. **Age:** age in days. **Bili:** therapy serum in mg/dL. Bilirubin is the yellowish pigment that's made during the breakdown of red blood cells that is found in bile. Assuming that this therapy serum is used to decrease high levels of bilirubin in the liver, higher values of "Bili" may indicate more liver damage. **Chol:** Cholesterol serum in mg/dL. Cholesterol serum is used to up the levels of cholesterol. Cirrhosis decreases cholesterol values, so high values of "Chol" may indicate increased progression of cirrhosis. **Alb:** Albumin in gm/dL. Albumin is a protein in blood plasma used to treat complications of cirrhosis with ascites and works to expand plasma and increase blood volume. Low values are bad. **Cu:** Copper in µg/day. Copper is an essential mineral for the body but too much is toxic so high values are worse. **Alkphos:** Alkaline phosphatase in Units/liter. Alkaline phosphatase is an enzyme used to break down proteins. High values are bad **Sgot:** SGOT in units/ml. SGOT is an enzyme found in the liver. High values are bad. **Trig:** Triglycerides in mg/dl. Triglycerides are a type of fat in the blood. High values are bad. **Plat:** platelets per cubic ml / 1000. Low values are bad. **Ptt:** Prothrombin time in seconds. This is how long it takes for a clot to form in a blood sample. **Stage:** histologic stage of disease (ranging from 1-4). **Rx D-penicillamine:** drug used, D-penicillamine is 1, Placebo is 0. **Sexmale:** Sex of the patient. Male is 1, female is 0. **Asciteseyes:** Presence of ascites. Ascites is a condition where too much fluid builds up in the abdomen. Yes is 1, No is 0. **Hepmegyes:** Presence of hepatomegaly which is abnormal enlargement of the liver. Yes is 1, No is 0. **Spidersyes:** Presence of spiders (spider veins; capillary branches radiate like "spider legs" carrying away free flowing blood). Yes is 1, No is 0. **EPWDOERBD:** Edema present without diuretics or edema resolved by diuretics. Yes is 1, No is 0. **EEANDTFE:** Edema and no diuretic therapy for edema. Yes is 1, No is 0. **EDDT:** Edema despite diuretic therapy. Yes is 1, No is 0.

3. Building a Model

Before building our model, 62 missing values had to be handled. Many methods were considered such as average, linear and random sampling imputations, however, a random forest was used instead as it:

- Can be applied to mixed data types (missings in numeric & categorical variables)
- No preprocessing required (no dummy-coding, standardization, data splitting, etc.)
- No assumptions required (aside from the normal assumption of being MAR/MCAR)
- Robust to noisy data, as random forests effectively have built-in feature selection. Methods like KNN imputation will have poor predictions in datasets with weak & non-informative predictors, whereas missForest() will make little to no use of these features
- Non-parametric: makes no assumptions about the relationship between the features, unlike MICE which assumes linearity
- Can leverage non-linear and interaction effects between features to improve imputation accuracy
- Gives an OOB error estimate for its predictions (Numeric: NRMSE/MSE, Categorical: PFC)

The algorithm is as follows:



```
Make an **initial guess** for all missing categorical/numeric values (e.g. mean, mode)
$k \leftarrow$ vector of column indices in $X$, sorted in **ascending order of % missing**
**while** not $\gamma$ **do**:
    $X^{imp}_{old} \leftarrow$ store previous imputed matrix
    **for** $s$ in $k$ **do**:
        Fit a random forest predicting the non-missing values of $X_s$: $y^{(s)}_{obs} \sim x^{(s)}_{obs}$
        Use this to predict the missing values of $X_s$: predict $y^{(s)}_{mis}$ using $x^{(s)}_{mis}$
        $X^{imp}_{new} \leftarrow$ update imputed matrix, using the predicted $y^{(s)}_{mis}$
    **end for**
    update $\gamma$
**end while**
**return** the final imputed matrix $X^{imp}$
```

By imputing our missing values with the R library “missForest”, we were able to perform exploratory data analysis. R packages such as “ggplot2” and “gridExtra” were utilized to better visualize and understand the data:

Probability density functions were created for appropriate variables as a useful way of representing the distribution of probabilities for a continuous random variable (Figure 1). By analyzing the shapes of these distributions we are able to determine characteristics such as central tendency, spread, and skewness. Most of our variables are in fact not normally distributed, which re-emphasizes the importance of the usage of random forest to impute for our missing values.

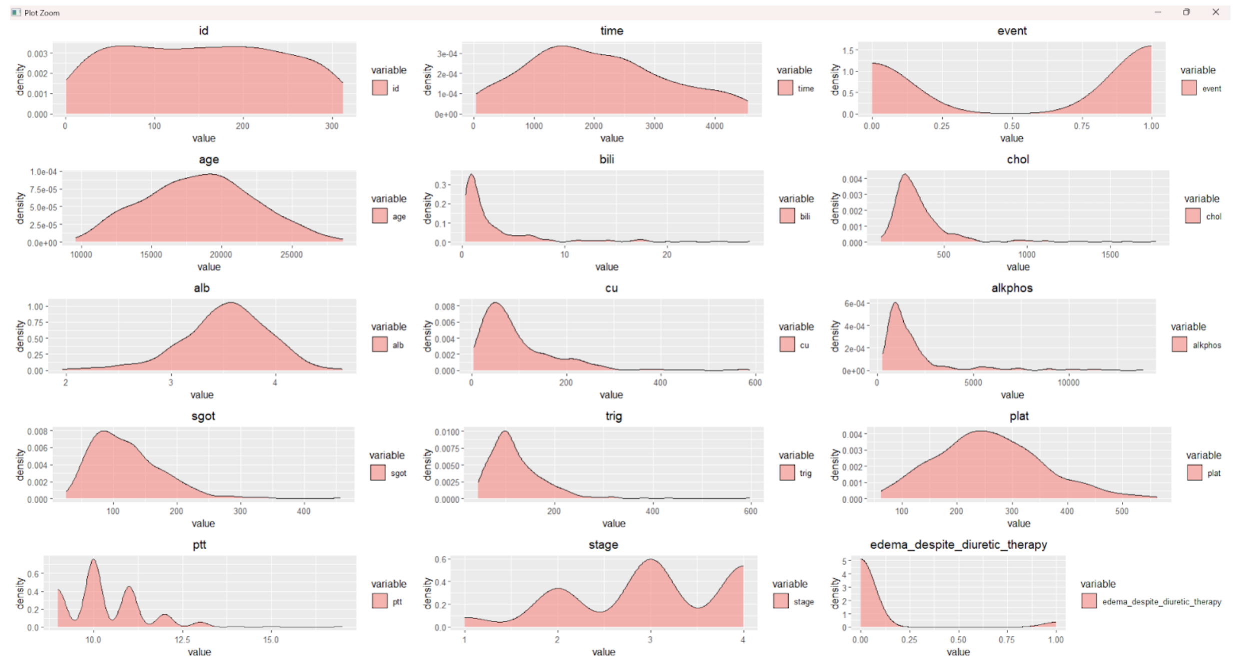


Figure 1: Probability density functions are iterated for each appropriate variable, thus demonstrating their distributions.

In Figure 2, we aimed to compare the distributions of the variables before and after applying random forest imputation. This comparison enabled us to assess whether the random forest model significantly altered our data. Upon examination, the distributions appear to be strikingly similar, with only minor variations observed between them.

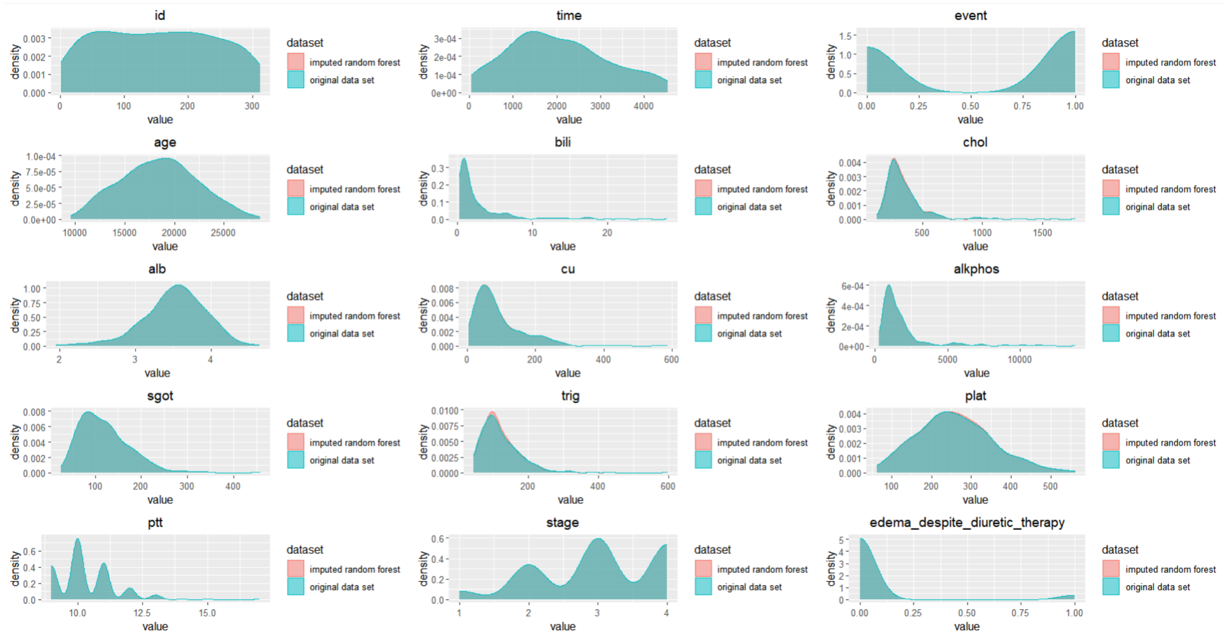


Figure 2: Demonstrations of superimposed graphs based on before and after imputations

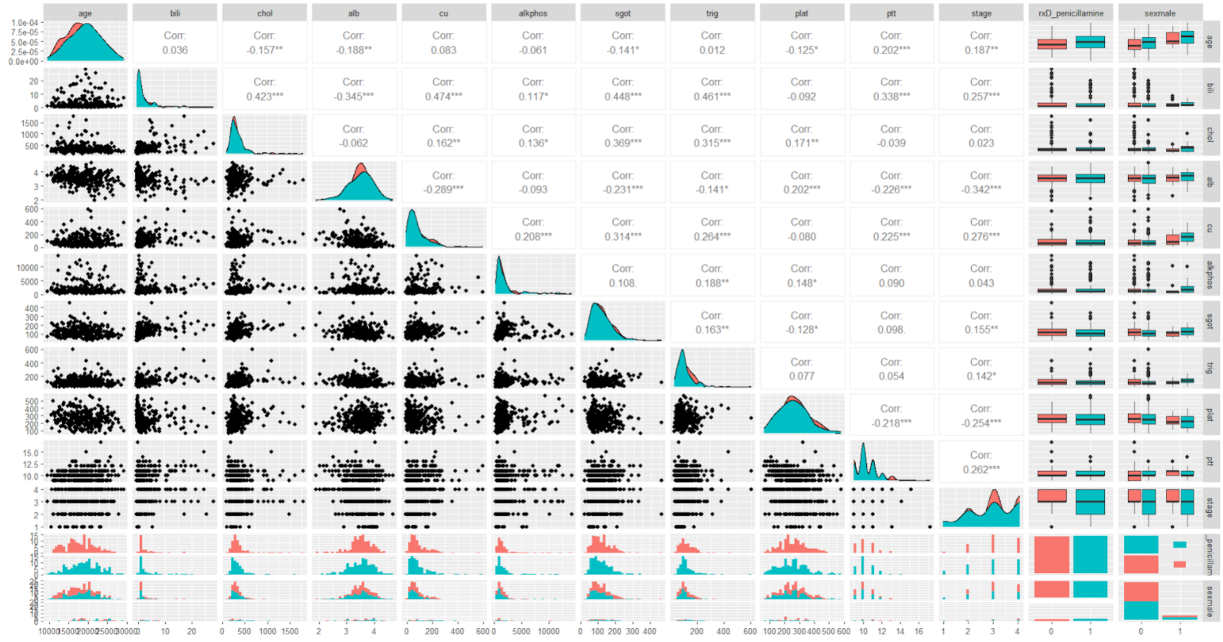


Figure 3: Distributions of patients with rxD Penicillamine and placebo treatments based on various covariates.

We also wanted to visualize the difference between patients who had received the rxD Penicillamine drug in comparison to patients who had received the placebo (Figure 3). However, there does not appear to be any significant difference between such groups, even when compared by gender and biological features such as bilirubin, cholesterol, and others.

In addition we also wanted to visualize the differences in stages when compared by gender (Figure 4). As previously stated, there is a huge difference between the males and the females due to the data that has been obtained from Mayo Clinics Women Healthcare Center.

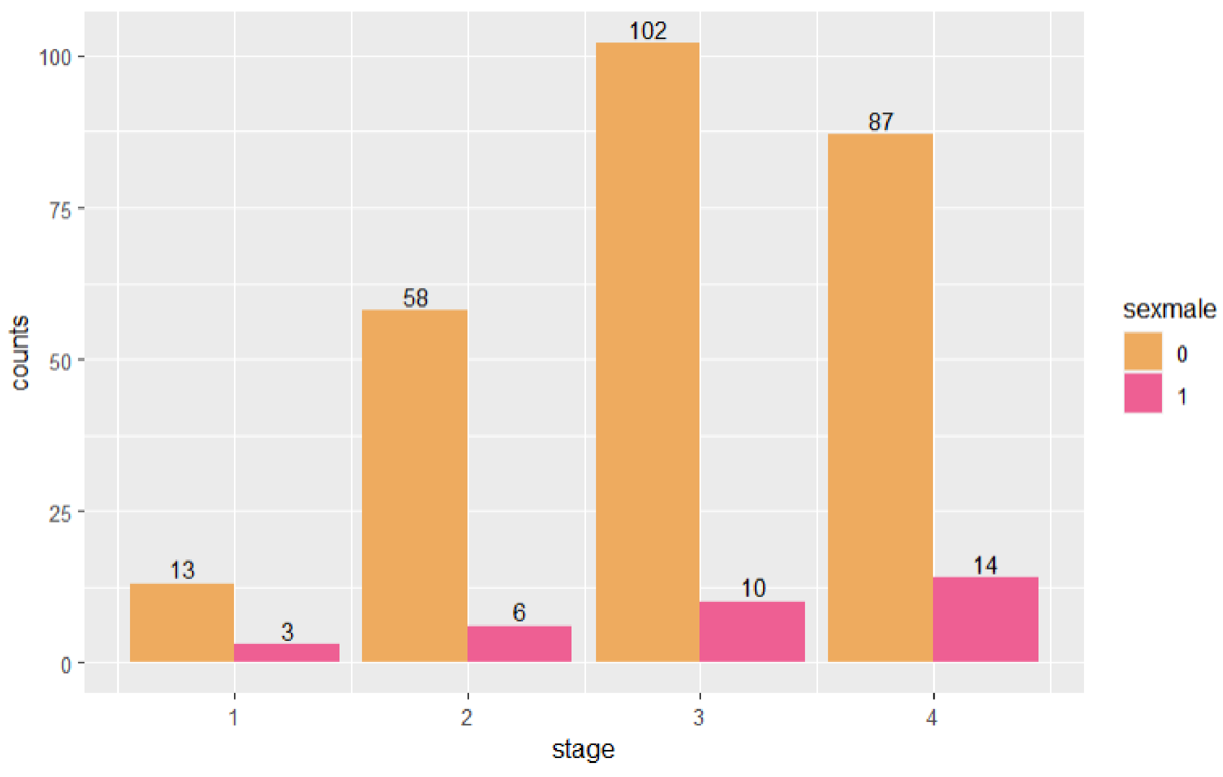
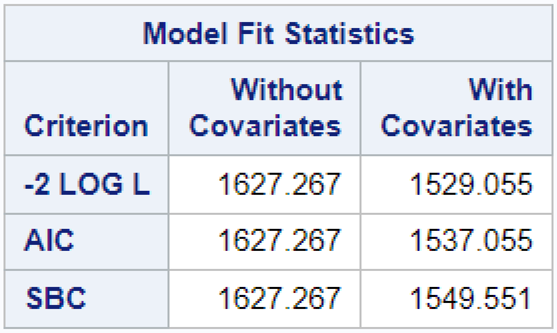


Figure 4: Differences of patients based on gender and patients. (0 represents female, and 1 represents male)

After conducting exploratory data analysis, our team proceeded the survival analysis by using the LIFEREG procedure. However, despite testing various distributions and transforming the age variable, we were unable to find a distribution that fit our data.

Instead, we opted out by using the semi-parametric procedure PHREG as it allowed us to conduct our survival analysis without requiring any assumptions about the underlying distribution of the survival times. This approach allowed us to use the Cox proportional hazards model, which is widely used in survival analysis for its flexibility and robustness. Additionally, we employed a stepwise procedure to select a subset of covariates for inclusion in the final model as this approach allows for the sequential removal and addition of covariates based on their significance level in the model. After running our model, we obtain the following results.

4. Interpreting the Results



Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1627.267	1529.055
AIC	1627.267	1537.055
SBC	1627.267	1549.551

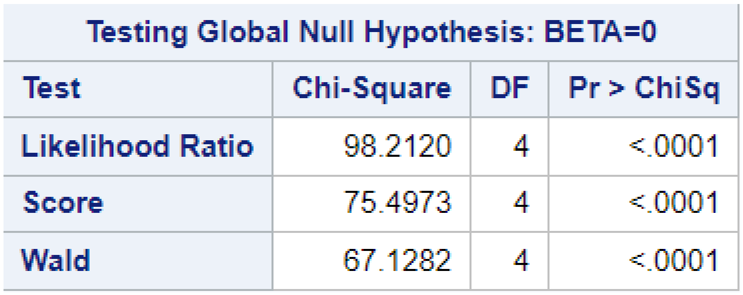
The given results show model fit statistics for two competing models: one without covariates and one with covariates. Covariates are variables that are controlled for in the analysis, potentially improving the model's fit. The table provides three common fit indices to evaluate and compare the models: -2 Log Likelihood (-2 LOG L), Akaike Information Criterion (AIC), and Schwarz Bayesian Criterion (SBC).

1) -2 Log Likelihood (-2 LOG L): This measures the goodness of fit of a model. Lower values indicate a better fit. In this case, the model with covariates has a lower -2 LOG L value (1529.055) compared to the model without covariates (1627.267), suggesting that the model with covariates fits the data better.

2) Akaike Information Criterion (AIC): AIC compares the goodness of fit of models while penalizing for model complexity. Lower AIC values indicate a better balance between fit and complexity. In this case, the model with covariates has a lower AIC value (1537.055) compared to the model without covariates (1627.267), suggesting that it is a better fit when considering model complexity.

3) Schwarz Bayesian Criterion (SBC): Also known as Bayesian Information Criterion (BIC), SBC is another measure that balances goodness of fit and model complexity. Like AIC, lower values are better. In this case, the model with covariates has a lower SBC value (1549.551) compared to the model without covariates (1627.267), indicating that it is a better fit when considering model complexity.

Overall, the inclusion of covariates in our model allows for a better fit than the model without covariates as indicated by all three fit indices.



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	98.2120	4	<.0001
Score	75.4973	4	<.0001
Wald	67.1282	4	<.0001

The given results show three tests for the Global Null Hypothesis, which states that all the coefficients in the model (Betas) are equal to zero, meaning that the predictors have no significant impact on the outcome. The three tests are the Likelihood Ratio Test, Score Test, and Wald Test. These tests help determine whether the predictors in the model have a significant effect on the outcome variable.

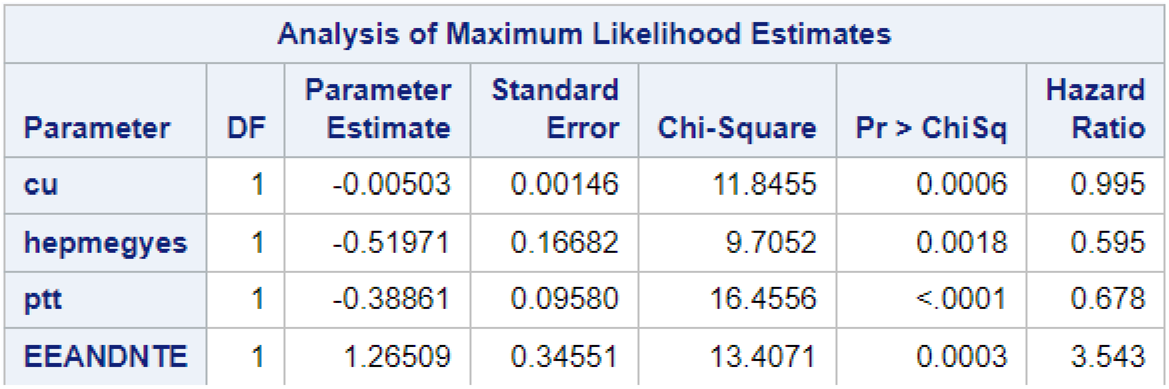
1) Likelihood Ratio Test: This test compares the difference in the -2 Log Likelihoods of the null model (with no predictors) and the alternative model (with predictors). A higher Chi-Square value indicates a better fit for the alternative model. In this case, the Chi-Square value is 98.2120 with 4 degrees of freedom (DF), and the p-value (Pr > ChiSq) is less than 0.0001. Since the p-value is less than the conventional significance level of 0.05, we reject the null hypothesis, concluding that at least one of the predictors in the model is significantly related to the outcome.

2) Score Test: This test is based on the score (or gradient) of the likelihood function. A higher Chi-Square value indicates that the predictors have a significant effect on the outcome. In this case, the Chi-Square value is 75.4973 with 4 degrees of freedom (DF),

and the p-value (Pr > ChiSq) is less than 0.0001. Since the p-value is less than 0.05, we reject the null hypothesis, concluding that at least one of the predictors in the model is significantly related to the outcome.

3) Wald Test: This test is based on the estimated coefficients (Betas) and their standard errors. A higher Chi-Square value indicates that the predictors have a significant effect on the outcome. In this case, the Chi-Square value is 67.1282 with 4 degrees of freedom (DF), and the p-value (Pr > ChiSq) is less than 0.0001. Since the p-value is less than 0.05, we reject the null hypothesis, concluding that at least one of the predictors in the model is significantly related to the outcome.

In conclusion, all three tests - Likelihood Ratio, Score, and Wald - provide strong evidence to reject the Global Null Hypothesis (BETA=0), suggesting that at least one of the predictors in the model has a significant impact on the outcome variable.



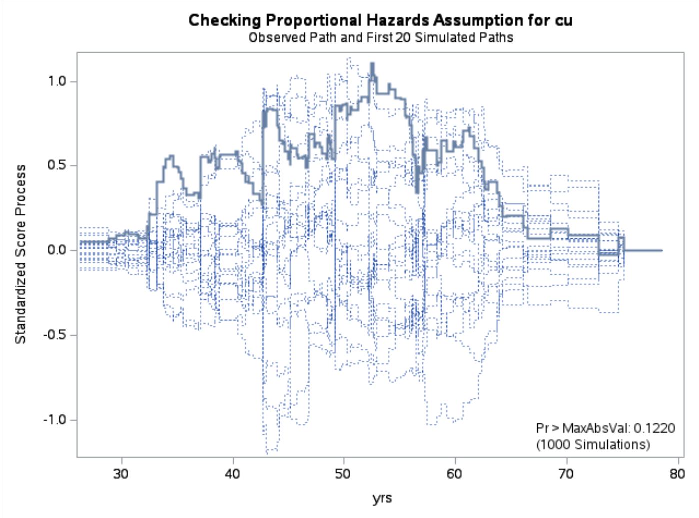
Analysis of Maximum Likelihood Estimates

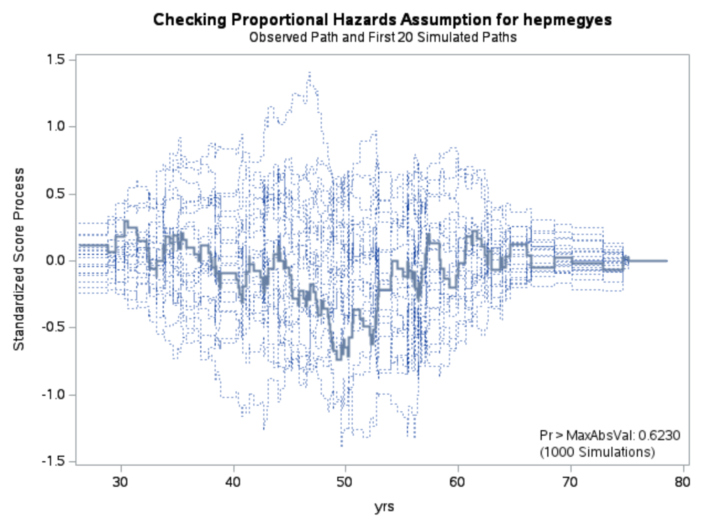
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
cu	1	-0.00503	0.00146	11.8455	0.0006	0.995
hepmegyes	1	-0.51971	0.16682	9.7052	0.0018	0.595
ptt	1	-0.38861	0.09580	16.4556	<.0001	0.678
EEANDNTE	1	1.26509	0.34551	13.4071	0.0003	3.543

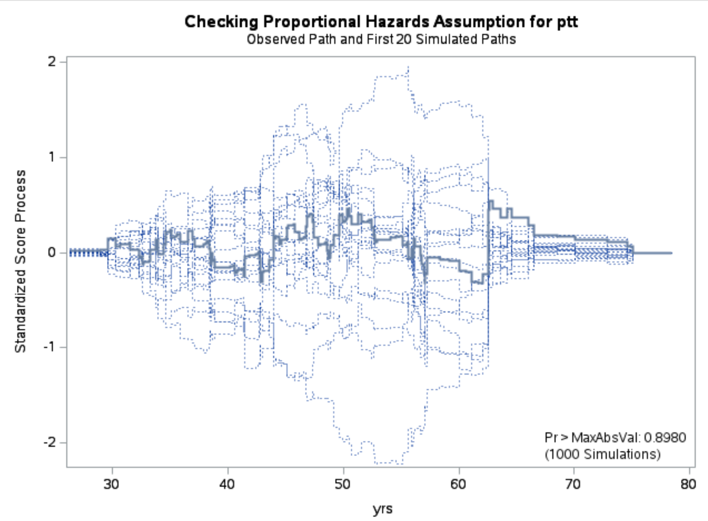
All the predictors have p-values less than the conventional significance level of 0.05, indicating that they are significantly associated with the outcome variable.

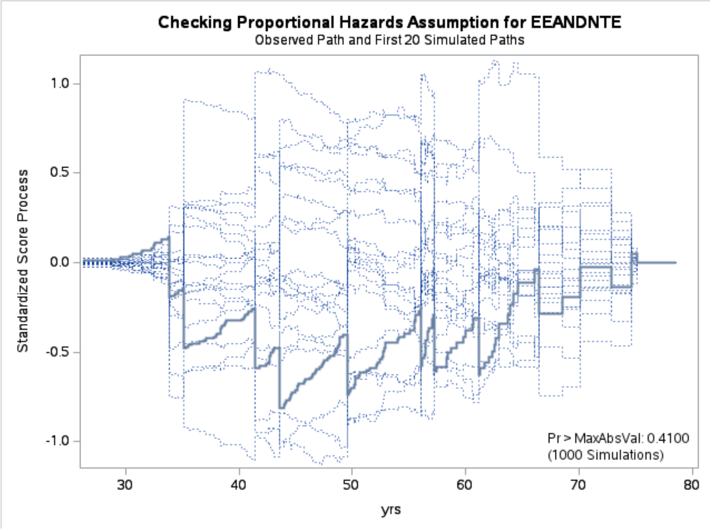
- For predictor cu, the hazard ratio is 0.995, suggesting that for each one-unit increase in cu, the hazard (risk) of the event occurring decreases by 0.5%.
- For predictor hepmegyes, the hazard ratio is 0.595, indicating that the hazard (risk) of the event occurring is 59.5% lower for those with hepmegyes=1 compared to those with hepmegyes=0.

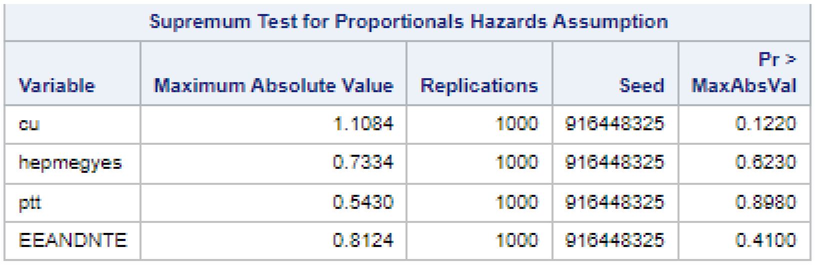
- For predictor ptt, the hazard ratio is 0.678, suggesting that for each one-unit increase in ptt, the hazard (risk) of the event occurring decreases by 32.2%.
- For predictor EEANDNTE, the hazard ratio is 3.543, implying that the hazard (risk) of the event occurring is 3.543 times higher for those with EEANDNTE=1 compared to those with EEANDNTE=0.









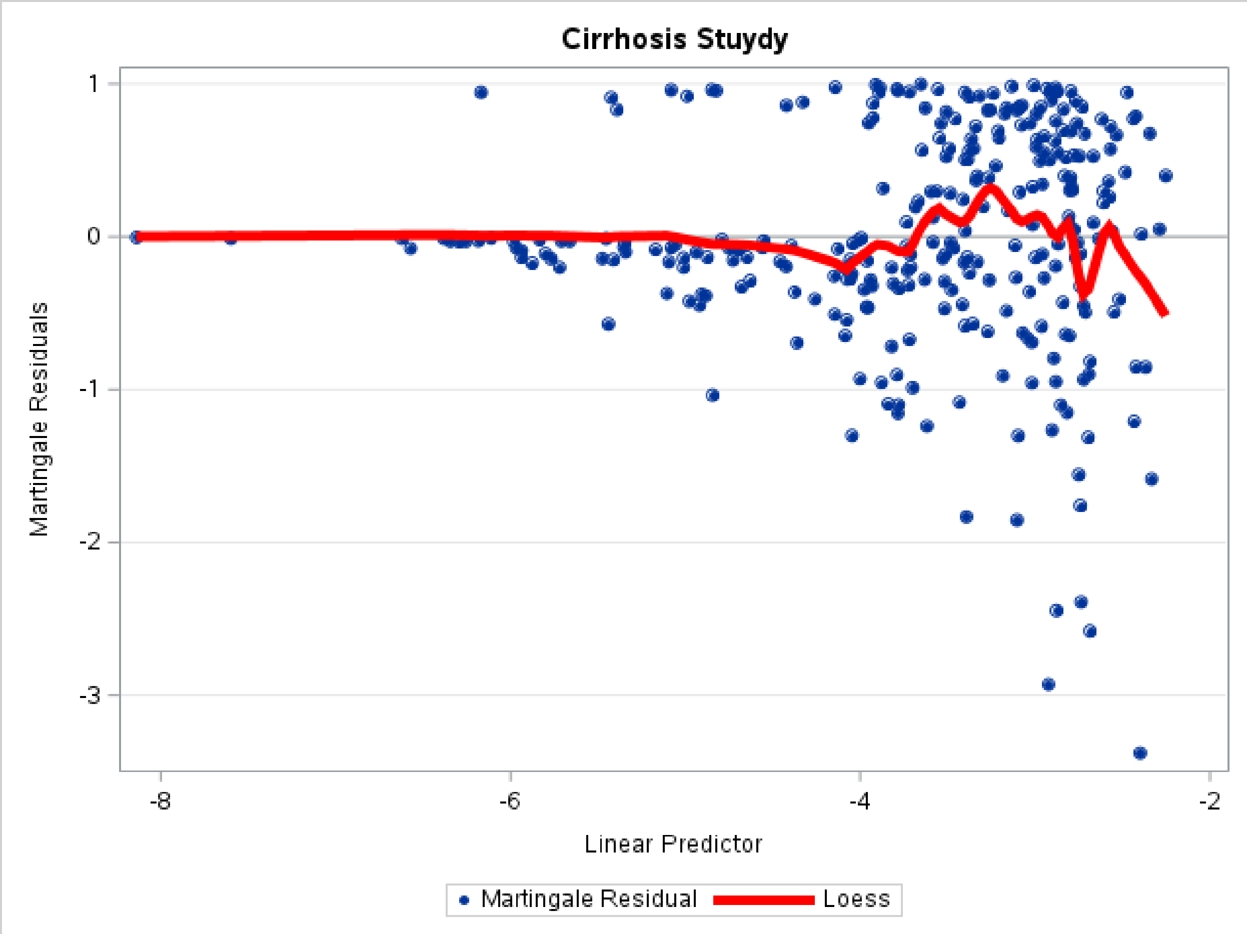


Supremum Test for Proportionals Hazards Assumption

Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
cu	1.1084	1000	916448325	0.1220
hepmegyes	0.7334	1000	916448325	0.6230
ptt	0.5430	1000	916448325	0.8980
EEANDNTE	0.8124	1000	916448325	0.4100

The given table presents the results of the Supremum Test for the Proportional Hazards Assumption. The Proportional Hazards Assumption is an essential assumption for the Cox Proportional Hazards model, which states that the hazard ratios for the predictors are constant over time. Violation of this assumption can lead to incorrect conclusions about the relationship between the predictors and the outcome variable.

The p-values for all predictors are greater than the conventional significance level of 0.05, indicating that there is no evidence to reject the null hypothesis that the Proportional Hazards Assumption holds for each predictor. In other words, the test results suggest that the hazard ratios for each predictor are constant over time, and the Cox Proportional Hazards model is appropriate for this data.

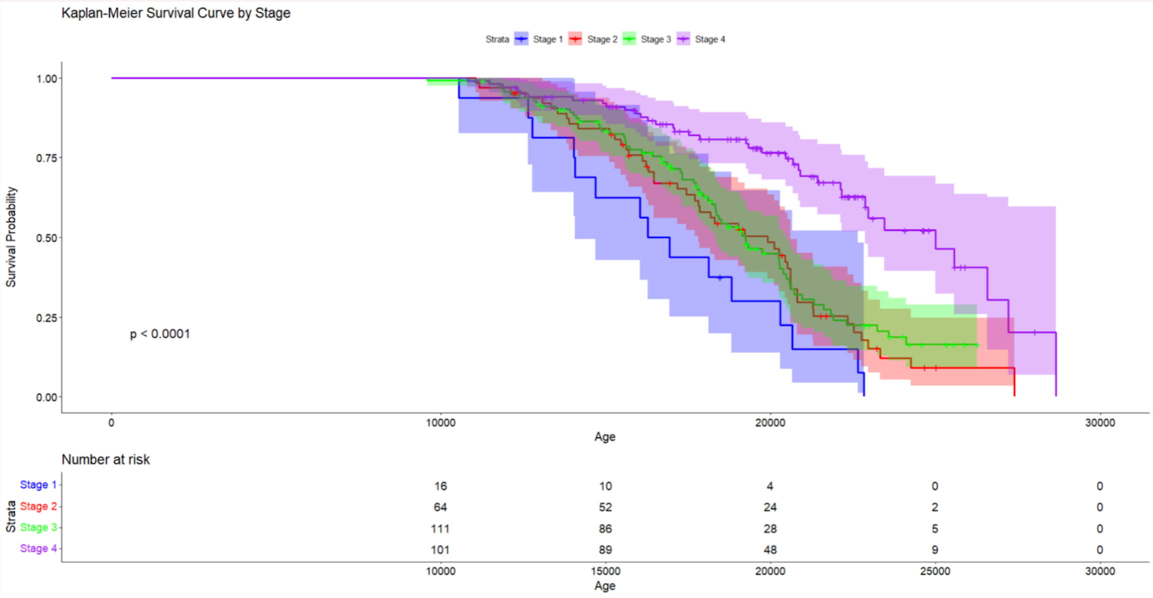


Lastly, discussing the residuals we decided to go with a martingale residual plot with a Loess fit line. The Loess line (represented in red) is a non-parametric technique that's used to fit a smooth curve to our residual data to determine if there is a trend or not. We can observe based on the Loess line that there is no trend in our residual data due to the fact that the trendline “wiggles” about zero or the x-axis. Looking at the scattered points alone we notice that there is no apparent pattern and it appears to be relatively symmetrical about the x-axis. There is an influential observation at the bottom right that exceeds the bounds of -3, however when cross-examining with the deviance residual plot we can conclude that the point is not determined to be a true outlier within the data.

5. Summary

As stated in the above results section, we decided to go with the PHREG model as opposed to the LIFEREG model due to its semi-parametric qualities. When running the LIFEREG procedure we noticed that our data does not truly follow a set distribution, so it was a safer option to go with the more flexible PHREG modeling approach. Going through the stepwise procedure, we were left with our four variables of interest; **cu, hepmegyes, ptt,** and **NEANDTFE.** We can conclude that all of these covariates are significant in predicting the outcome of the covariate **Age**. It appears that as time goes on (or **Age** increases), the survival probability of an individual with PBC (Primary Biliary Cirrhosis) decreases.

Some limitations we faced throughout the data was the fact that one of the things we decided to leave out of our model was the **Stage** of the patients liver disease. Plotting the survival graphs by stage showed interesting results, but seemed to be stacked backwards since the graph indicates that stage four patients were more likely to live longer than stage one patients. Another reason we decided to leave it out of our model was because of the large discrepancy in the number of stage patients. For instance, we only had 16 stage one patients while there were 112 stage 3 patients.



Next, this study was done at the Mayo Women's Clinic which meant that we noticed a higher concentration of female patients in the study vs. the male patients. This disease does, however, impact females more commonly than males so this inequality in the data may not be too problematic. Another issue we faced was the high volume of missing data observations. As mentioned in the previous sections, we imputed the data to try and fill those missing values with as accurate of a value as we could, but that being said we can't know for certain if our values are even within the same ballpark of the true ones. Lastly, we had to take out an entire subgroup of individuals who had liver transplants. We were given patients who were alive or dead after the trial as well as those who had received a liver transplant. Due to the unknown outcome of those with transplants we were unable to dictate whether or not they were alive or dead after the trials so we thought it was best to remove them as a whole. The removed patients still possessed valuable data and maybe in future work we could examine more closely on these patients and see how the act of getting a liver transplant could help to predict their age of survival.

Bibliography

- https://www.openml.org/search?type=data&sort=runs&id=524&status=active
- https://liverfoundation.org/about-your-liver/facts-about-liver-disease/how-many-people-have-liver-disease/
- https://www.mayoclinic.org/diseases-conditions/liver-problems/symptoms-causes/syc-20374502
- https://academic.oup.com/qjmed/article/100/8/534/1522064
- https://www.mayoclinic.org/tests-procedures/liver-transplant/about/pac-20384842#:~:text=The%20most%20common%20cause%20of,reason%20for%20a%20liver%20transplant.
- https://rpubs.com/lmorgan95/MissForest
- https://cran.r-project.org/web/packages/missForest/missForest.pdf