# Survival Analysis of Primary Biliary Cirrhosis Patients: Evaluating the Efficacy of D-penicillamine in a Mayo Clinic Clinical Trial

Cirrhosis is a severe and advanced form of liver scarring caused by a variety of liver diseases and conditions, including chronic intoxication and hepatitis. Primary Biliary Cirrhosis (PBC) is a type of cirrhosis that effects the bile ducts of the liver, resulting in impaired bile flow and progressive liver damage. Numerous medical advancements have been made to better understand and manage PBC over the years (Purohit & Cappell, 2015).

## About Dataset

Cirrhosis is a late stage of scarring (fibrosis) of the liver caused by many forms of liver diseases and conditions, such as hepatitis and chronic alcoholism. The following data contains the information collected from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A description of the clinical background for the trial and the covariates recorded here is in Chapter 0, especially Section 0.2 of Fleming and Harrington, Counting

Processes and Survival Analysis, Wiley, 1991. A more extended discussion can be found in Dickson, et al., Hepatology 10:1-7 (1989) and in Markus, et al., N Eng J of Med 320:1709-13 (1989).

A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo-controlled trial of the drug D-penicillamine. The first 312 cases in the dataset participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

Attribute Information

1) ID: unique identifier

2) N\_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986

3) Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)

4) Drug: type of drug D-penicillamine or placebo

5) Age: age in [days]

6) Sex: M (male) or F (female)

7) Ascites: presence of ascites N (No) or Y (Yes)

8) Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)

9) Spiders: presence of spiders N (No) or Y (Yes)

10) Edema: presence of edema 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)

11) Bilirubin: serum bilirubin in [mg/dl]

12) Cholesterol: serum cholesterol in [mg/dl]

13) Albumin: albumin in [gm/dl]

14) Copper: urine copper in [ug/day]

15) Alk\_Phos: alkaline phosphatase in [U/liter]

16) SGOT: SGOT in [U/ml]

17) Triglycerides: triglicerides in [mg/dl]

18) Platelets: platelets per cubic [ml/1000]

19) Prothrombin: prothrombin time in seconds [s]

20) Stage: histologic stage of disease (1, 2, 3, or 4)

## Step 1: Data Exploration and Cleaning

**1. Load the Dataset:**

The dataset was loaded from the provided CSV file, "cirrhosis.csv", into a SAS data set named "clinical\_trial."

**2. Inspect the Dataset Structure**:

I used the proc contents procedure to get an overview of the dataset's structure, including variable names, types, and formats.

**3. Check for Missing Values:**

To identify missing values in the dataset, I utilized the proc means procedure with the nmiss option.

A screenshot of a computer

Description automatically generated

**4. Handle Missing Values:**

In this step, I addressed missing values in the dataset. I employed imputation techniques to fill in missing values for numerical variables with the median and categorical variables with the mode.

**4.1 Imputation for Numerical Data (using median):**

I calculated the median for each numerical variable using the PROC MEANS procedure and stored the results in a temporary dataset named "median\_stats."

**4.3 Imputation for Categorical Data (using mode):**

I calculated the mode for each categorical variable using the proc freq procedure. Finally, replaced missing values in categorical variables with the calculated modes.

## Step 2: Descriptive Statistics

In this step, I computed descriptive statistics for the continuous variables in the dataset to gain insights into their central tendency, spread, and distribution. We also utilized summary statistics and created visualizations like histograms and boxplots to better understand the data. The following sections outline the analysis performed for each variable:

I started by computing summary statistics for the following continuous variables: Age, Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phos, SGOT, Triglycerides, Platelets, and Prothrombin. These statistics help us gain insights into the general characteristics of the data.

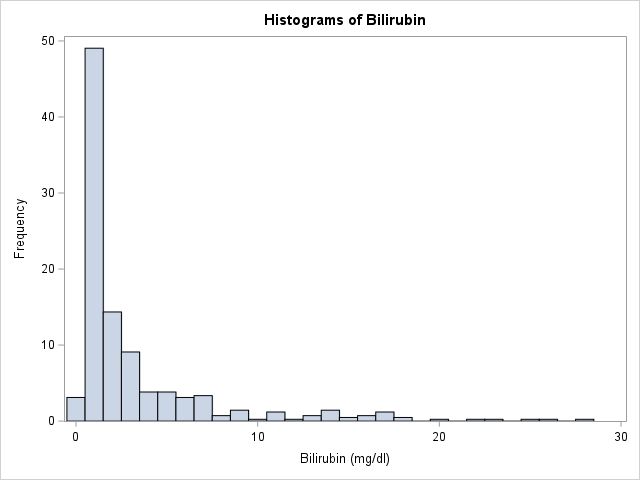
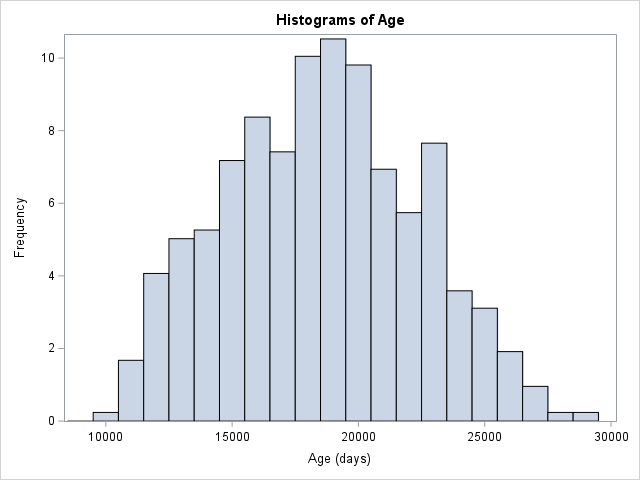
A screenshot of a computer screen

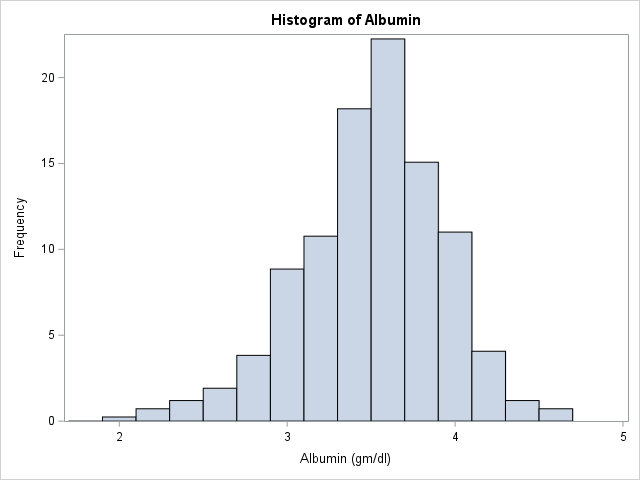
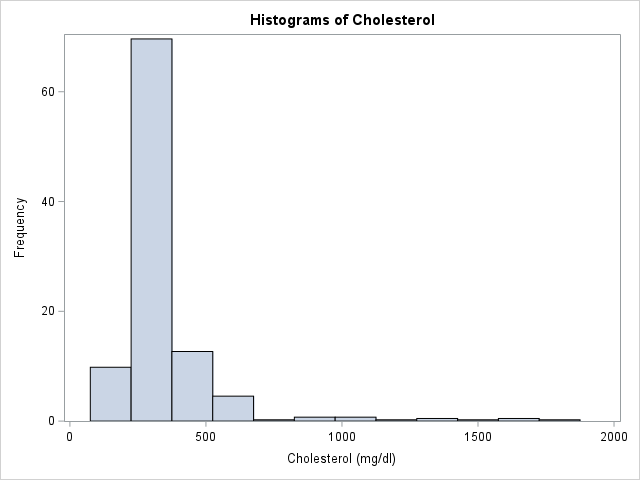
Description automatically generated

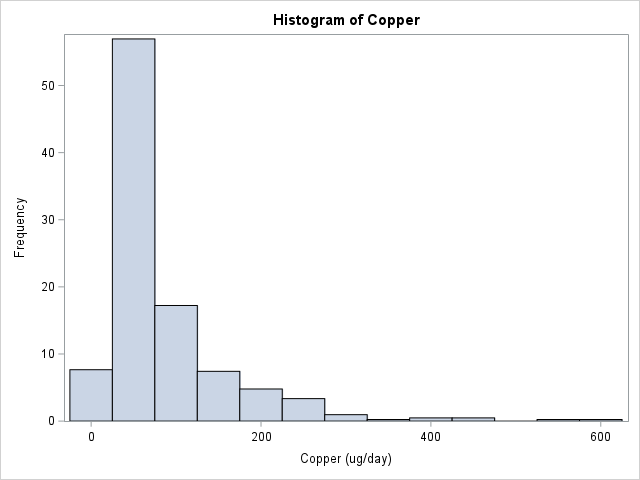
* Age has a relatively large range (9,598 to 28,650 days) with a standard deviation of around ten years (3,815 days).
* Bilirubin levels show a significant difference between the mean and median values (mean = 3.22 and median =1 .4), indicating a skewed distribution.
* Cholesterol levels have a wide range (120 to 1,775), suggesting that there might be outliers in the data.
* Albumin levels have a small standard deviation (0.42), indicating that the values are closely clustered around the mean.
* Copper levels also show a significant difference between the mean and median values (mean =91 .28 and median =73 ), indicating a skewed distribution.
* Alk\_Phos, SGOT, and Triglycerides have relatively large standard deviations compared to their means, suggesting a wide spread of values in the data.

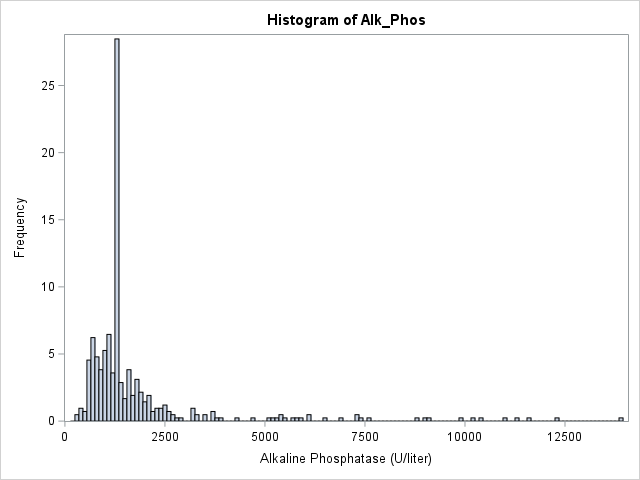
**Visualizations for Continuous Variables:**

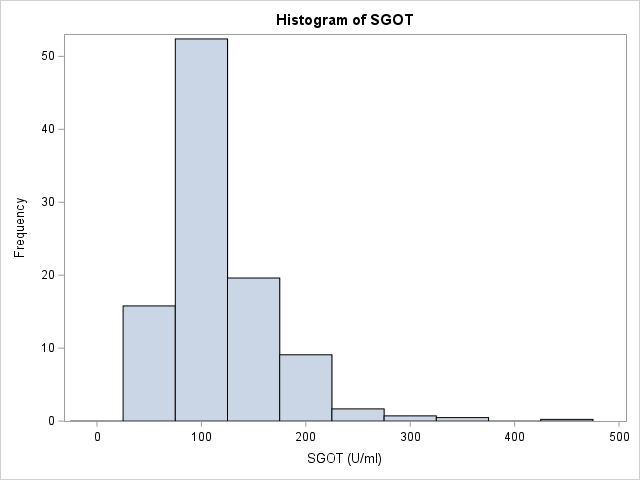
To better understand the distribution of the continuous variables, we created histograms and boxplots for each of them.

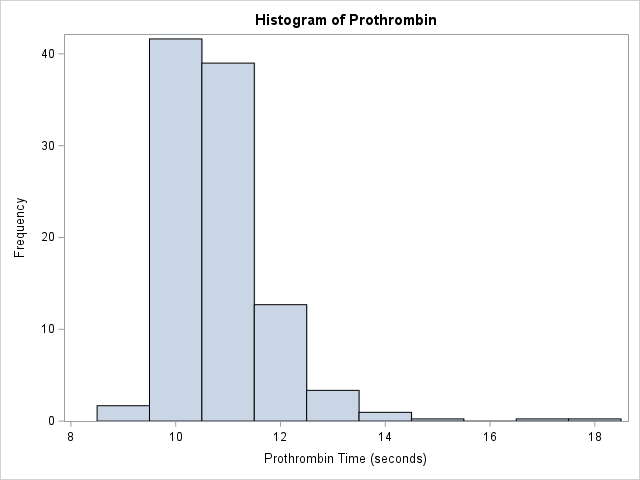


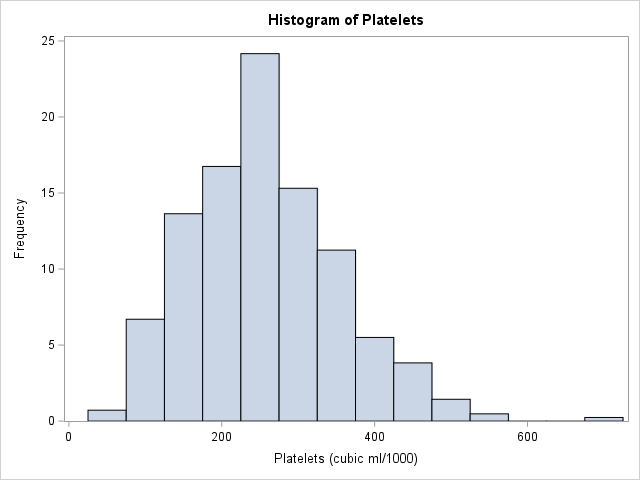
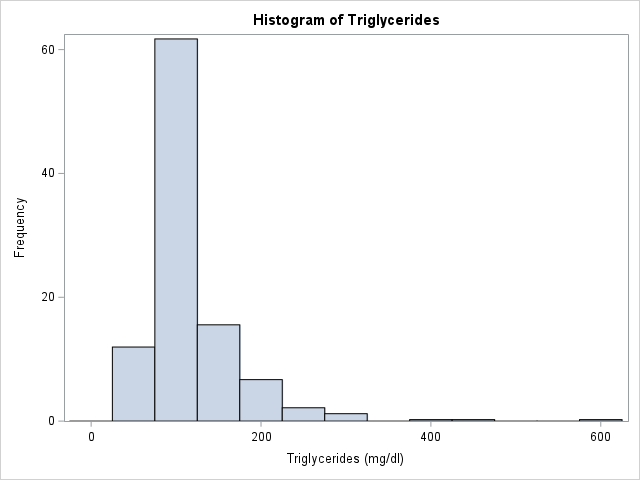








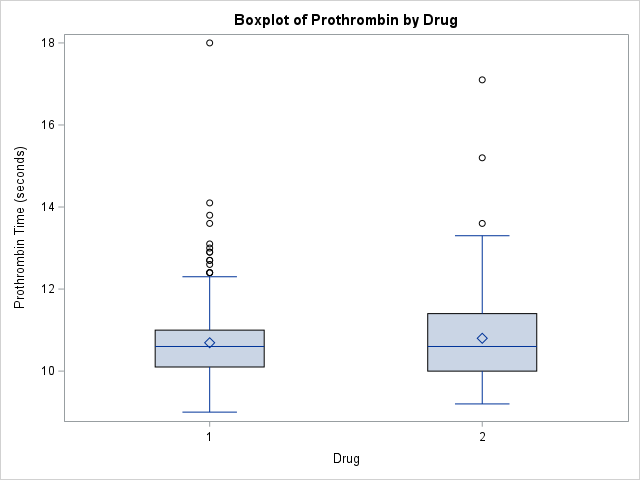
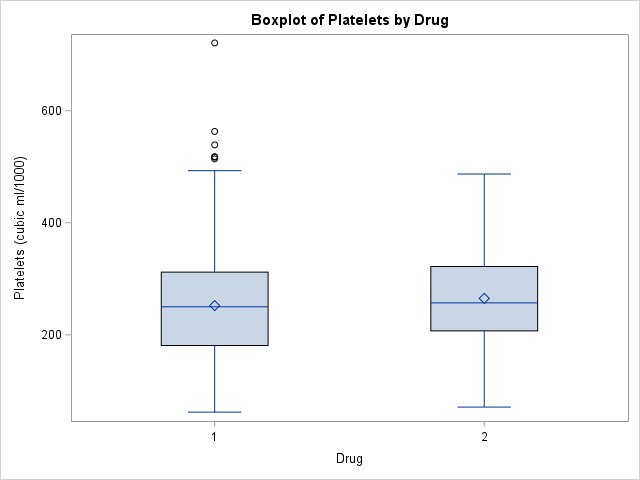
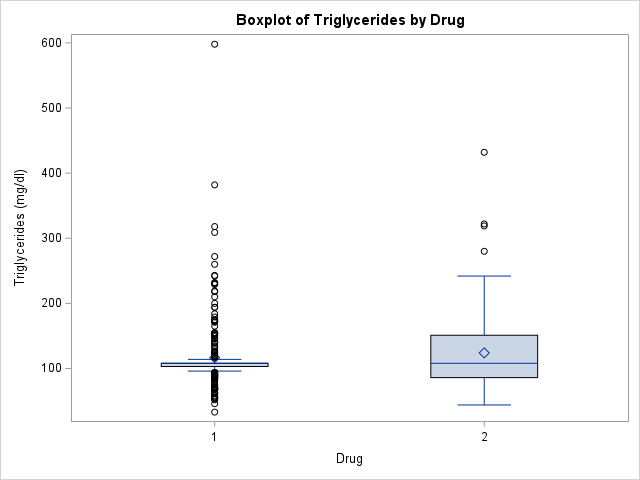
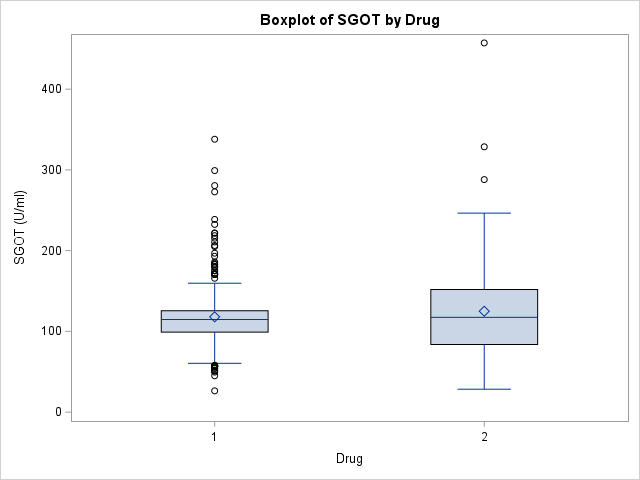
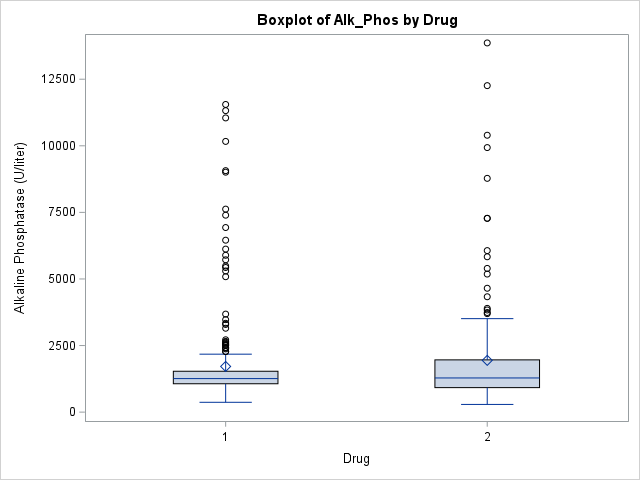
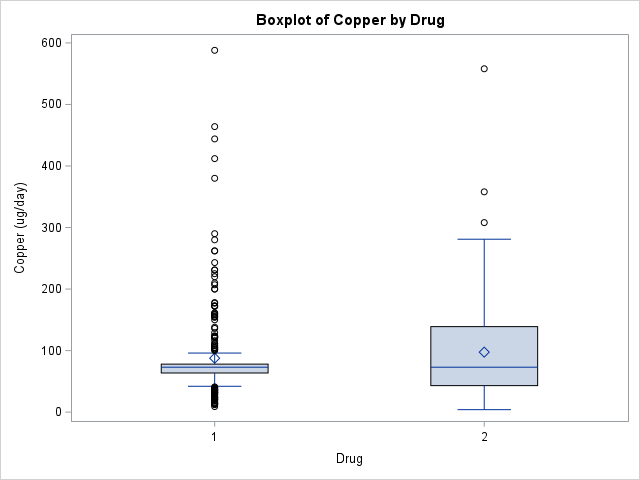
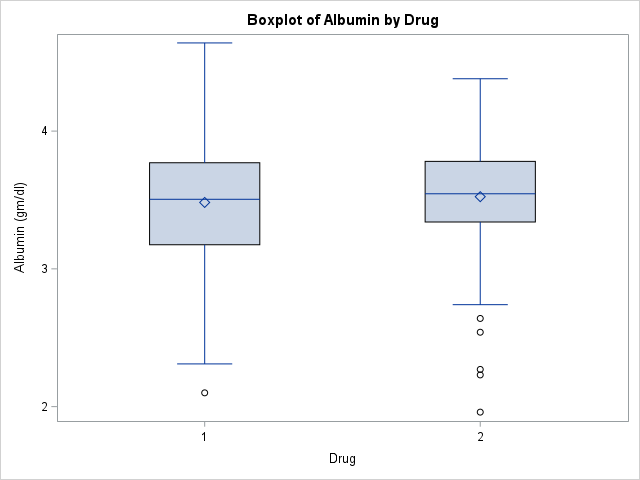
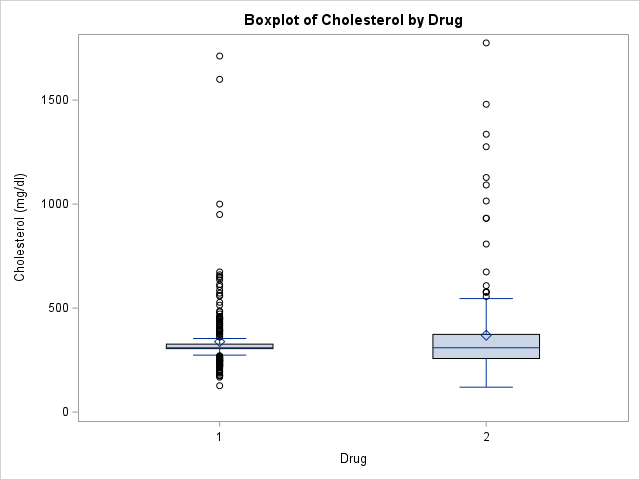
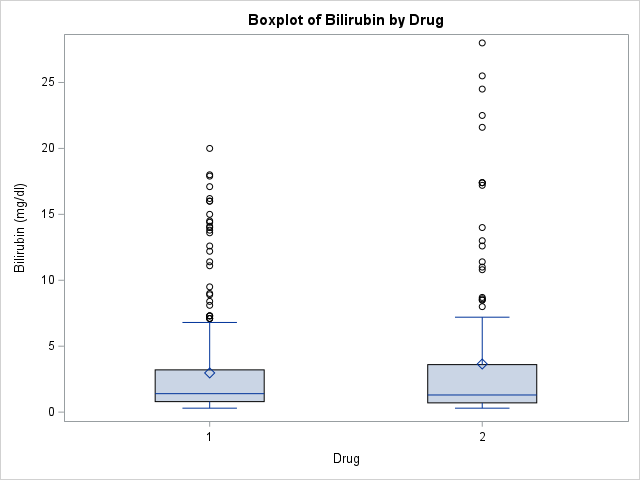
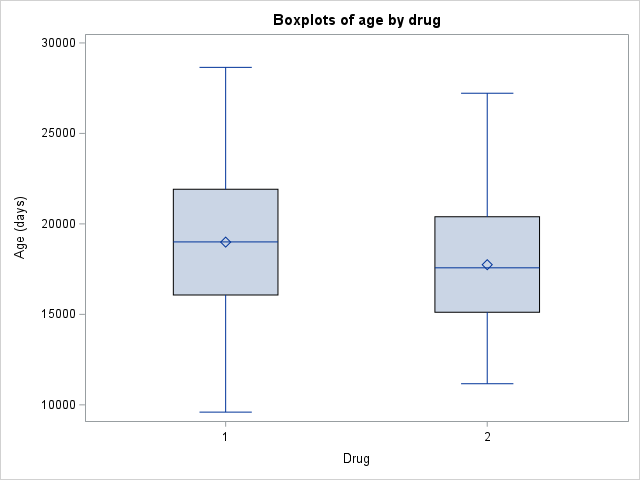




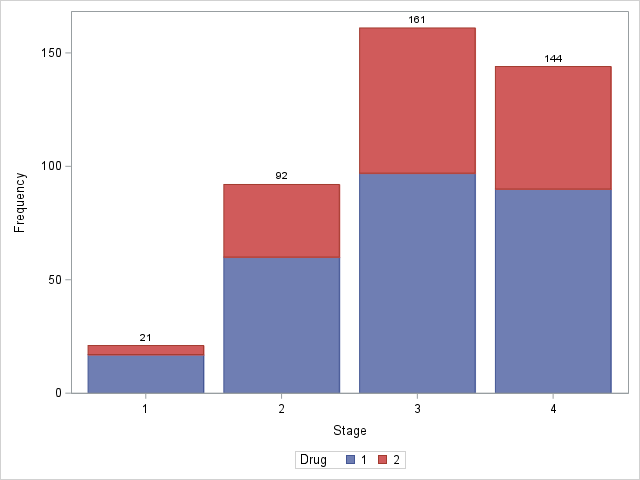
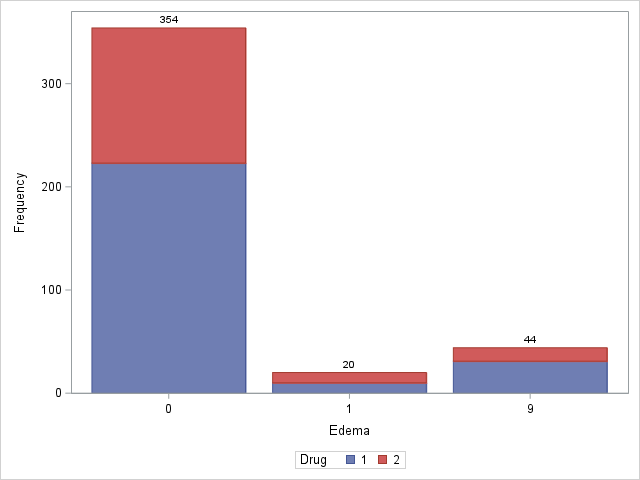
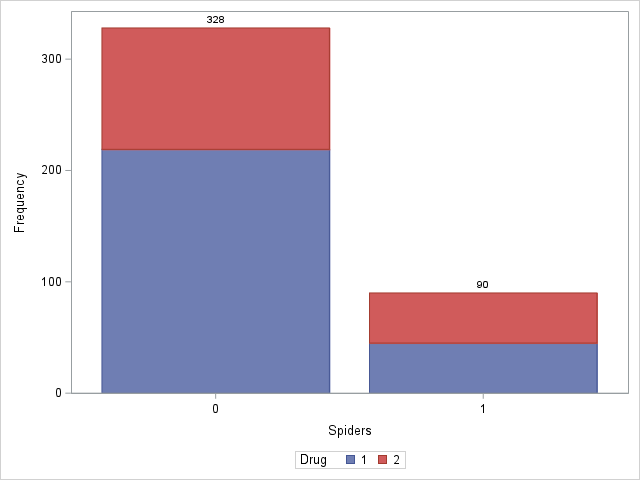
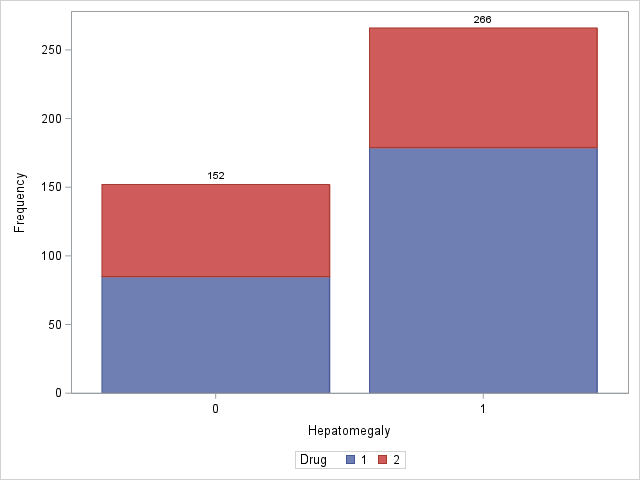
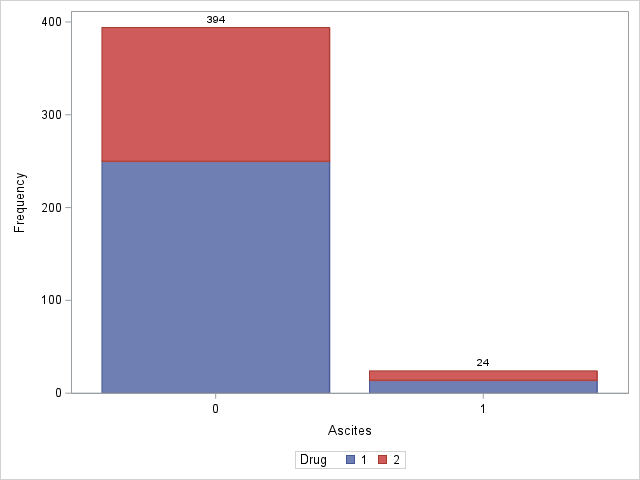
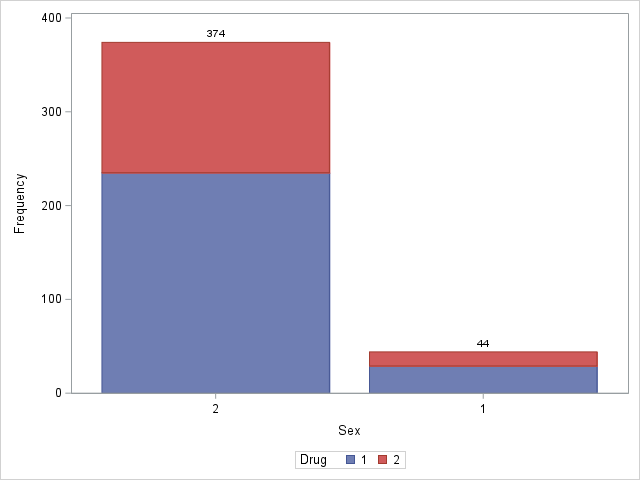
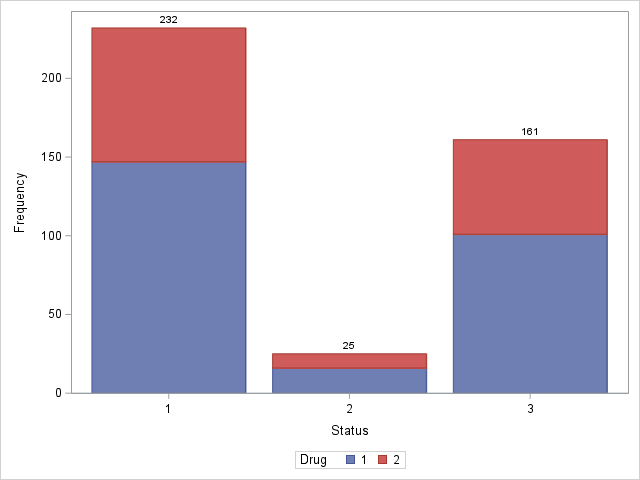
* Age distribution appears to be approximately normal
* Bilirubin, Cholesterol, Copper, Alk\_Phos, SGOT, and Triglycerides all exhibit right-skewed distributions with long tails, indicating that there are some patients with much higher values than the majority.
* Albumin distribution is left-skewed with a long tail on the left side, suggesting that there are some patients with lower albumin levels than the majority.
* Platelets distribution is slightly right-skewed but appears to be closer to a normal distribution compared to other variables.
* Prothrombin time has a relatively symmetric distribution but shows some outliers on the right side.

**Boxplots comparing continuous variables across different drug groups (D-penicillamine and placebo)**

there are no striking differences between the treatment groups for most continuous variables based on visual inspection of boxplots alone. While some variables have a higher number of outliers in one group than the other or slight variations in distribution patterns between groups, this may not necessarily indicate a significant difference in treatment effects.



**Categorical Variable**



* Status: The distribution of patient status (C, CL, and D) is similar between the two drug groups, with no significant differences observed.
* Sex: There are more female patients than male patients in both drug groups, but the proportions are similar between D-penicillamine and placebo.
* Ascites: The majority of patients do not have ascites in both drug groups, with a slightly higher number of cases with ascites in the D-penicillamine group.
* Hepatomegaly: The presence of hepatomegaly is almost evenly distributed between the two drug groups, with a slightly higher number of cases without hepatomegaly in the D-penicillamine group.
* Spiders: The presence of spiders is similar between the two drug groups, with no significant differences observed.
* Edema: The distribution of edema (N, S, and Y) is similar between the two drug groups, with no significant differences observed.
* Stage: The distribution of disease stages (1 to 4) is relatively similar between the two drug groups, with a slightly higher number of stage 3 cases in the placebo group.

## Step 3- Survival Analysis

The LIFETEST procedure is a statistical method used to analyze survival data. In this case, it is testing the homogeneity of survival curves for N\_Days (number of days between registration and the earlier of death, transplantation, or study analysis time) over strata (groups) based on the drug type (D-penicillamine or placebo).

/\*Step 3: Survival Analysis\*/

/\* Create a new binary variable for event status (1 for death, 0 for censored) \*/

**data** Clinical\_trial;

set Clinical\_trial;

/\* Check if the Status variable indicates "Death" (Status = "D") \*/

if Status = "D" then

Event = **1**; /\* Event status is 1 for death \*/

else

Event = **0**; /\* Event status is 0 for censored \*/

**run**;

/\* Perform survival analysis using PROC LIFETEST \*/

**proc** **lifetest** data=Clinical\_trial plots=survival(atrisk cb);

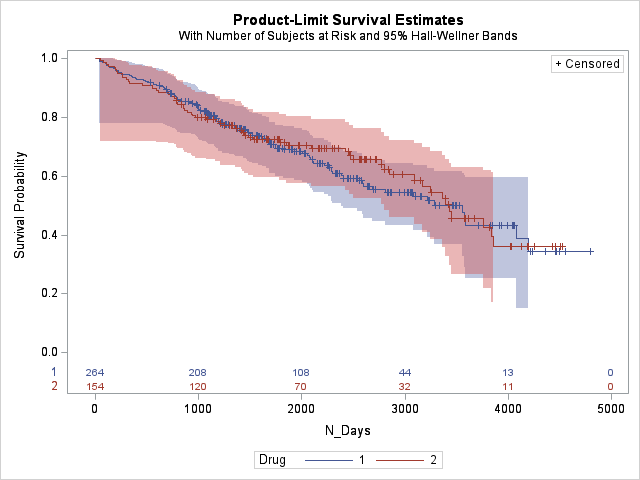
time N\_Days\*Event(**0**);

strata Drug;

**run**;

A screenshot of a computer

Description automatically generated



The output provided includes several test statistics and their corresponding p-values:

1. Log-Rank Test: Chi-Square = 0.1461, DF (degrees of freedom) = 1, Pr > Chi-Square = 0.7023

2. Wilcoxon Test: Chi-Square = 0.0377, DF = 1, Pr > Chi-Square = 0.8460

3. -2Log(LR): Chi-Square = 0.0818, DF = 1, Pr > Chi-Square = 0.7748

All three tests have p-values greater than the common significance level of 0.05, which indicates

that there is no significant difference in survival curves between the two drug groups (D-penicillamine

and placebo). In other words, the treatment with D-penicillamine does not appear to have a significant

impact on patient survival compared to the placebo group based on this analysis.

## Step 4: Cox Proportional Hazards Model

The Cox proportional hazards model was employed to investigate the impact of demographic and clinical factors on the survival time. The Cox model aims to assess the association between various demographic and clinical variables and the risk of an event (death or transplantation) occurring in the patient population.

/\*Step 4: Cox Proportional Hazards Model\*/

**proc** **phreg** data=Clinical\_trial;

class Drug(ref='2') Sex(ref='1') Ascites(ref='0') Hepatomegaly(ref='0') Spiders(ref='0') Edema(ref='0')Stage(ref='1');

model N\_Days\*event(**0**) = Drug Age Sex Ascites Hepatomegaly Spiders Edema Bilirubin Cholesterol Albumin Copper Alk\_Phos SGOT Tryglicerides Platelets Prothrombin Stage;

**run**;

**Model Fit Statistics**

**A table with numbers and text

Description automatically generated**

The model fit statistics for the Cox model are as follows:

* Without Covariates: -2 LOG L = 1746.975, AIC = 1746.975, SBC = 1746.975
* With Covariates: -2 LOG L = 1523.454, AIC = 1563.454, SBC = 1625.082

The model with covariates provided a significantly better fit to the survival data compared to the model without covariates.

**Testing Global Null Hypothesis**

**A screenshot of a graph

Description automatically generated**

The global null hypothesis, indicating the absence of any significant covariate effects on survival, was tested using the likelihood ratio, score, and Wald tests. All tests resulted in very low p-values (p < 0.0001), indicating strong evidence against the null hypothesis. Hence, the model with covariates demonstrates a significant association between the included variables and the risk of events.

**Type 3 Tests and Analysis of Maximum Likelihood Estimates**

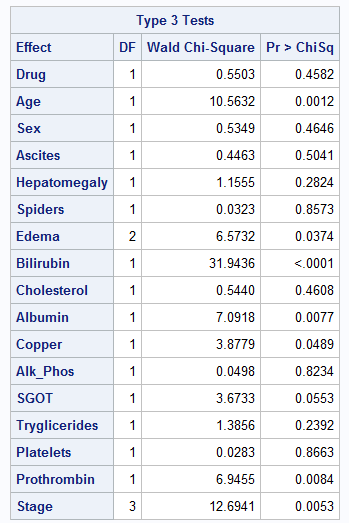
Based on the output of the Cox Proportional Hazards model using the SAS PHREG procedure,

we can identify significant predictors of survival and quantify their impact. The p-values

(Pr > ChiSq) in the "Analysis of Maximum Likelihood Estimates" table indicate the significance

of each predictor. A smaller p-value (typically less than 0.05) suggests a significant relationship

between the predictor and survival outcome.



A screenshot of a data sheet

Description automatically generated

Significant predictors and their hazard ratios (HR) are:

1. Age: HR = 1.000, p-value = 0.0012
2. Edema:

* Edema 1: HR = 2.354, p-value = 0.0116
* Edema 9: HR = 1.240, p-value = 0.3713 (not significant)

1. Bilirubin: HR = 1.109, p-value < .0001
2. Albumin: HR = 0.541, p-value = 0.0077
3. Copper: HR = 1.002, p-value = 0.0489
4. Prothrombin: HR = 1.204, p-value = 0.0084
5. Stage:

* Stage 2: HR=2 .046 ,p-value=0 .3389(not significant)
* Stage3 :HR=3 .012,p- value=0 .1353(not significant)
* Stage4 :HR=4 .819,p- value=0 .0356

A hazard ratio greater than one indicates an increased risk of death, while a hazard ratio less than one indicates a decreased risk.

In summary:

* Older age is associated with a slightly increased risk of death.
* Presence of edema without diuretics (Edema type "S") is associated with an increased risk of death.
* Higher bilirubin levels are associated with an increased risk of death.
* Higher albumin levels are associated with a decreased risk of death.
* Higher copper levels are associated with a slightly increased risk of death.
* Longer prothrombin time is associated with an increased risk of death.
* Stage 4 disease is associated with an increased risk of death compared to stage 1.

Other predictors, such as drug type, sex, ascites, hepatomegaly, spiders, cholesterol, alk\_phos, SGOT, triglycerides, and platelets were not found to be significant predictors in this analysis.

## Step 5: Stage Analysis

This report presents the results of stage-based survival analysis conducted on a dataset of primary biliary cirrhosis (PBC) patients. The aim of this analysis was to assess the homogeneity of survival curves across different histologic stages and investigate whether there are significant differences in survival among the stages.

The PROC LIFETEST procedure in SAS was used to perform the survival analysis. The dataset contained information on the number of days until an event (death or transplantation) occurred (N\_Days) and the censoring status (Event) for each patient. Patients were stratified based on two factors: the type of drug received (Drug) and the histologic stage of the disease (Stage).

/\*Step 5: Stage Analysis\*/

/\* Load the LIFETEST procedure for survival analysis \*/

/\* Perform survival analysis using PROC LIFETEST \*/

**proc** **lifetest** data=Clinical\_trial plots=survival(atrisk cb);

time N\_Days\*Event(**0**);

strata Drug;

strata Stage; /\* Separate patients into different groups based on histologic stage \*/

**run**;

1. Homogeneity of Survival Curves

A screenshot of a test

Description automatically generated

The log-rank and Wilcoxon rank statistics were used to test the homogeneity of survival curves across different histologic stages. The test results are as follows:

* Log-Rank Chi-Square: 68.3025, p < 0.0001
* Wilcoxon Chi-Square: 73.7693, p < 0.0001

Both tests indicate that there are significant differences in survival among the histologic stages. Therefore, the null hypothesis of homogeneity of survival curves is rejected.

1. Rank Statistics:

A screenshot of a graph

Description automatically generated

The rank statistics table shows the Log-Rank and Wilcoxon test statistics for each stage (1, 2, 3, and 4).

These statistics help us understand how different the survival times are between stages. Positive values suggest worse survival, while negative values suggest better survival. In this case, Stage 4 has worse survival compared to Stages 1, 2, and 3.

1. Covariance Matrix for Log-Rank and Wilcoxon Statistics:

A screenshot of a table

Description automatically generated

These matrices show the relationships between different stages in terms of their survival times.

The diagonal elements represent variances for each stage's test statistic, while off-diagonal elements

represent covariances between pairs of stages' test statistics.

1. Test of Equality over Strata:

A screenshot of a test

Description automatically generated

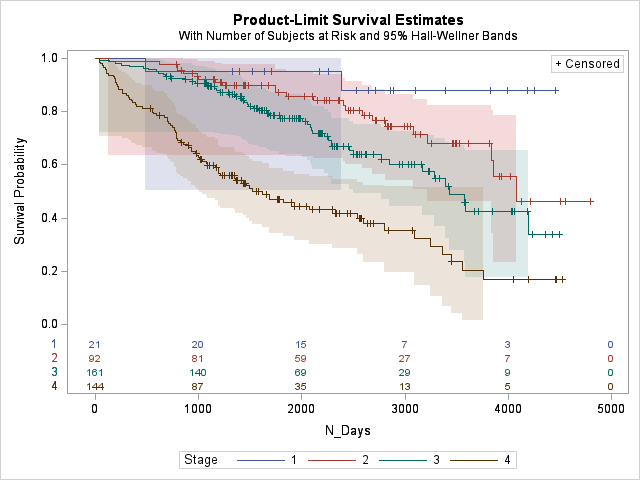
This section provides p-values for three different tests: Log-Rank, Wilcoxon, and -2Log(LR). These p-values help us determine if there is a significant difference in survival times between different stages.

- Log-Rank: This test compares the entire survival curves across all stages.

- Wilcoxon: This test places more weight on early time points in the survival curves.

-2Log(LR): This is a likelihood ratio test that compares nested models.

All three p-values are less than 0.0001, indicating a significant difference in survival times between different stages. This means that the null hypothesis (i.e., no difference in survival curves across histologic stages) can be rejected.



In summary, based on these results from LIFETEST procedure output, there is a significant difference in survival times among patients with primary biliary cirrhosis (PBC) across different histologic stages of disease progression. Stage 4 has worse overall survival compared to Stages 1, 2, and 3.

## Step 6:Correlation Analysis

The aim of this analysis was to explore the relationships between various clinical variables and the number of days until an event (death or transplantation) occurred (N\_Days). The PROC CORR procedure in SAS was used to calculate the Pearson correlation coefficients between N\_Days and the following clinical variables: Age, Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phos (Alkaline Phosphatase), SGOT (Serum Glutamic Oxaloacetic Transaminase), Tryglicerides, Platelets, and Prothrombin.

/\* Load the PROC CORR procedure \*/

**proc** **corr** data=CStep **6**linical\_trial;

var Age Bilirubin Cholesterol Albumin Copper Alk\_Phos SGOT Tryglicerides Platelets Prothrombin;

with N\_Days;

**run**;

A screenshot of a computer

Description automatically generated

1. Age: The correlation coefficient is -0.12593 (p-value = 0.0100), indicating a weak negative correlation between N\_Days and Age. As age increases, survival time (N\_Days) slightly decreases.

2. Bilirubin: The correlation coefficient is -0.40395 (p-value < 0.0001), indicating a moderate negative correlation between N\_Days and Bilirubin levels. As bilirubin levels increase, survival time decreases.

3. Cholesterol: The correlation coefficient is -0.09932 (p-value = 0.0424), indicating a weak negative correlation between N\_Days and Cholesterol levels.

4. Albumin: The correlation coefficient is 0.43083 (p-value < 0.0001), indicating a moderate positive correlation between N\_Days and Albumin levels. As albumin levels increase, survival time increases.

5. Copper: The correlation coefficient is -0.29572 (p-value < 0.0001), indicating a weak to moderate negative correlation between N\_Days and Copper levels.

6. Alk\_Phos: 0.15237, p = 0.0018. Alk\_Phos shows a weak positive correlation with N\_Days, implying that higher alkaline phosphatase levels may be associated with slightly longer survival times.

7. SGOT: -0.18793, p = 0.0001. SGOT exhibits a weak negative correlation with N\_Days, indicating that higher SGOT levels may be associated with slightly shorter survival times.

8. Tryglicerides: -0.11145, p = 0.0227. Tryglicerides show a weak negative correlation with N\_Days, suggesting that higher triglyceride levels may be associated with slightly shorter survival times.

9. Platelets: 0.14614, p = 0.0027. Platelets display a weak positive correlation with N\_Days, indicating that higher platelet counts may be associated with slightly longer survival times.

10. Prothrombin: -0.11068, p = 0.0236. Prothrombin exhibits a weak negative correlation with N\_Days, suggesting that longer prothrombin times may be associated with slightly shorter survival times.

these correlations do not imply causation but only show associations between variables in the dataset.

In summary, some variables show significant correlations with survival time (N\_Days). For example, higher bilirubin levels are associated with shorter survival times, while higher albumin levels are associated with longer survival times.

## Results

The survival analysis and Cox proportional hazards model provided valuable insights into the factors affecting the survival of patients with primary biliary cirrhosis (PBC). The main findings are summarized below:

* Survival Analysis: The LIFETEST procedure revealed no significant difference in survival curves between the two drug groups (D-penicillamine and placebo), indicating that treatment with D-penicillamine does not have a significant impact on patient survival compared to the placebo group.
* Cox Proportional Hazards Model: Several demographic and clinical factors were found to be significant predictors of survival, including age, edema, bilirubin levels, albumin levels, copper levels, prothrombin time, and disease stage. Notably, higher bilirubin levels were associated with an increased risk of death, while higher albumin levels were associated with a decreased risk of death.
* Stage Analysis: A significant difference in survival times was observed among patients across different histologic stages of disease progression. Stage 4 had worse overall survival compared to Stages 1, 2, and 3.
* Correlation Analysis: Some variables showed significant correlations with survival time (N\_Days). For example, higher bilirubin levels were associated with shorter survival times, while higher albumin levels were associated with longer survival times.

## Discussion

**- Interpretation of Results and Clinical Implications**: The survival analysis conducted in this study revealed no significant difference in survival curves between the D-penicillamine and placebo groups, suggesting that D-penicillamine treatment does not have a substantial impact on patient survival in PBC. However, other factors such as age, edema presence without diuretics (Edema type "S"), bilirubin levels, albumin levels, copper levels, and prothrombin time were found to be significant predictors of survival. These findings highlight the importance of considering multiple factors when evaluating treatment options for PBC patients.

- **Strengths and Limitations**: The study has several strengths, including a comprehensive analysis of various demographic and clinical factors affecting patient survival. Additionally, the use of advanced statistical methods such as Cox proportional hazards models allowed for a more nuanced understanding of the relationships between predictors and outcomes.

However, there are also limitations to consider. First, the sample size may not be large enough to detect small differences in survival between treatment groups or other subgroups. Second, potential sources of bias or confounding could affect the results; for example, unmeasured variables or differences in baseline characteristics between groups might influence the observed associations. Third, the study's observational design limits our ability to draw causal conclusions from the findings.

- **Challenges and Future Research Direction**s: During the analysis process, some challenges included interpreting complex output from statistical procedures and ensuring that appropriate methods were used to account for censoring in survival data. In future research directions, it would be beneficial to explore additional treatment options for PBC patients beyond D-penicillamine or placebo treatments. Moreover, investigating potential interactions between demographic factors (e.g., age or sex) and clinical variables could provide further insights into personalized treatment approaches for PBC patients.

## Conclusion

In conclusion, while this study did not find a significant impact of D-penicillamine treatment on PBC patient survival compared to placebo treatment, it identified several factors that significantly influence patient prognosis. Further research is needed to explore alternative treatments and interventions that could improve survival outcomes for PBC patients, taking into account the complex interplay of demographic and clinical factors.

## References

Fedesoriano. (2021, August 2). *Cirrhosis prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset

Purohit, T., & Cappell, M. S. (2015, May 8). *Primary biliary cirrhosis: Pathophysiology, clinical presentation and therapy*. World journal of hepatology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419097/