Course Name: Statistical Modelling and Computation

Course Number and Section: 16:954:567:02

Project: In-depth analysis of Patient Outcomes in Liver Cirrhosis: A Comprehensive Approach Using Missing Value Imputation and Survival Analysis Techniques.

Instructor: Jack Mardekian

Submitted By: Sarthak Singh (ss4767)

Date Submitted: 24.04.2024

Abstract

In this project, we employ advanced statistical techniques to analyze the survival times of patients diagnosed with liver cirrhosis, focusing on the impact of various clinical and demographic factors. Utilizing the cirrhosis dataset from the Mayo Clinic [1], we integrate missing value imputation methods to address data incompleteness and apply survival analysis to predict patient outcomes. The primary aim is to provide a comprehensive model that enhances the understanding of factors influencing survival rates among cirrhosis patients, which can guide clinical decision-making and improve patient management strategies.

Introduction

Liver cirrhosis represents a significant public health issue, characterized by the progression of liver fibrosis leading to liver failure and potentially death [2]. Accurate prediction of survival times in patients diagnosed with cirrhosis is crucial for effective treatment planning and patient counseling. However, clinical datasets often suffer from missing data, which can compromise the validity of survival analyses[3]. Addressing this challenge, our study utilizes robust missing value imputation techniques to prepare a comprehensive dataset for analysis, ensuring that subsequent survival analysis models are both reliable and informative.

This report details the methodology and insights from our analysis of the Mayo Clinic's cirrhosis dataset, which includes data from a randomized placebo-controlled trial testing the drug D-penicillamine along with additional observational data from non-participants. Our approach combines statistical techniques for handling missing data with survival analysis models to explore how different variables such as treatment type, demographic factors, and clinical measurements influence the survival probabilities of cirrhosis patients.

Dataset Description

The dataset[4] under study comprises 424 patients with primary biliary cirrhosis referred to the Mayo Clinic between 1974 and 1984. It includes comprehensive data on 312 participants of a clinical trial and additional data on 112 non-participants who consented to provide basic metrics and undergo survival tracking. The data contains 17 clinical features, including demographic information (e.g., age and sex), clinical measurements (e.g., bilirubin levels, albumin levels), and treatment details (e.g., type of drug administered). The survival states are categorized into three types: 'D' for death, 'C' for censored, and 'CL' for censored due to liver transplantation.

- **N_Days:** number of days between registration and the earlier of death, transplantation, or study analysis time.
- **Drug:** type of drug D-penicillamine or placebo.
- **Age:** The age of the patient (calculated in days).
- **Sex:** The gender of the patient.
- **Ascites**: the presence of ascites, either N (No) or Y (Yes).
- **Hepatomegaly:** the presence of hepatomegaly N (No) or Y (Yes).
- **Spiders:** the presence of spiders N (No) or Y (Yes).
- **Edema:** presence of edema N (no edema), S (edema present without diuretics), or Y (edema despite diuretic therapy)
- **Bilirubin:** serum bilirubin (mg/dl).
- **Cholesterol:** serum cholesterol (mg/dl).
- **Albumin:** albumin (gm/dl).
- **Copper:** urine copper (ug/day).
- **Alk_Phos:** alkaline phosphatase (U/liter).
- **SGOT:** Serum Glutamic-Oxaloacetic Transaminase levels (U/ml).
- **Triglycerides:** a type of fat (lipid) found in the blood that can increase the risk of heart disease (mg/dL).
- **Platelets:** platelets per cubic ml/1000.
- **Prothrombin:** prothrombin time (s).
- **Stage:** histologic stage of disease (1, 2, 3, or 4).

- **Status (Target):** status of the patient C (censored), CL (censored due to liver tx), or D (death).

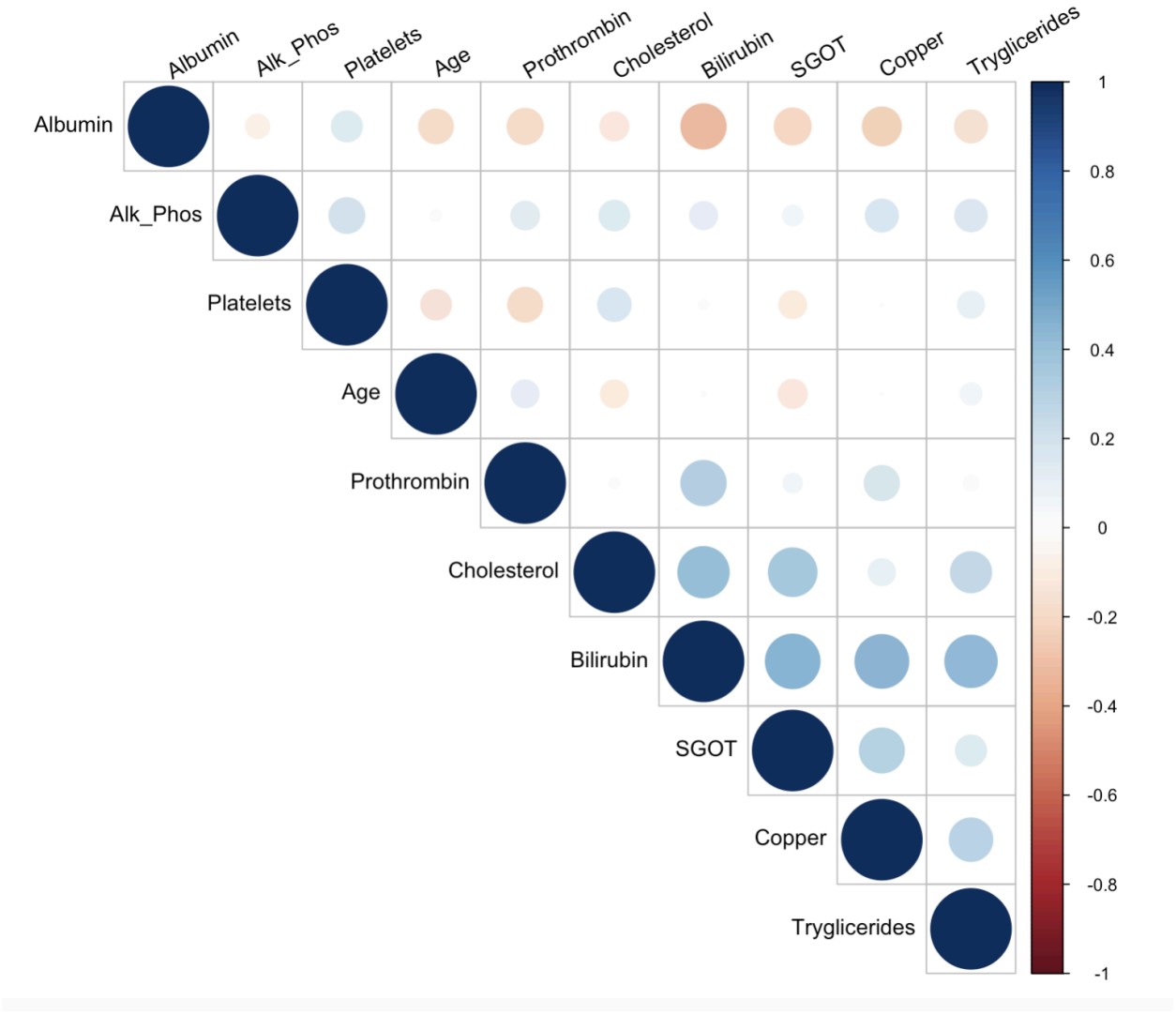

Figure 1: Correlation between Continuous variables.

The correlation analysis of clinical variables in the dataset reveals significant statistical relationships that inform our understanding of liver cirrhosis. Key findings include a strong positive correlation between bilirubin and SGOT (0.4568), indicating that increases in liver enzyme levels accompany higher bilirubin, a marker of liver dysfunction. Similarly, bilirubin is notably correlated with copper (0.4431), suggesting that liver impairment impacts copper metabolism. Conversely, a significant negative correlation exists between albumin and bilirubin (-0.3142), illustrating how worsening liver function is associated with decreased protein synthesis.

Materials and Methods

Missing Value Exploration

For an informed missing value imputation, we should first analyze the pattern of missing data, starting with which variables have how much data missing. From the figure below, we can see that clinical measurements like cholesterol and copper levels are missing the most in our data.

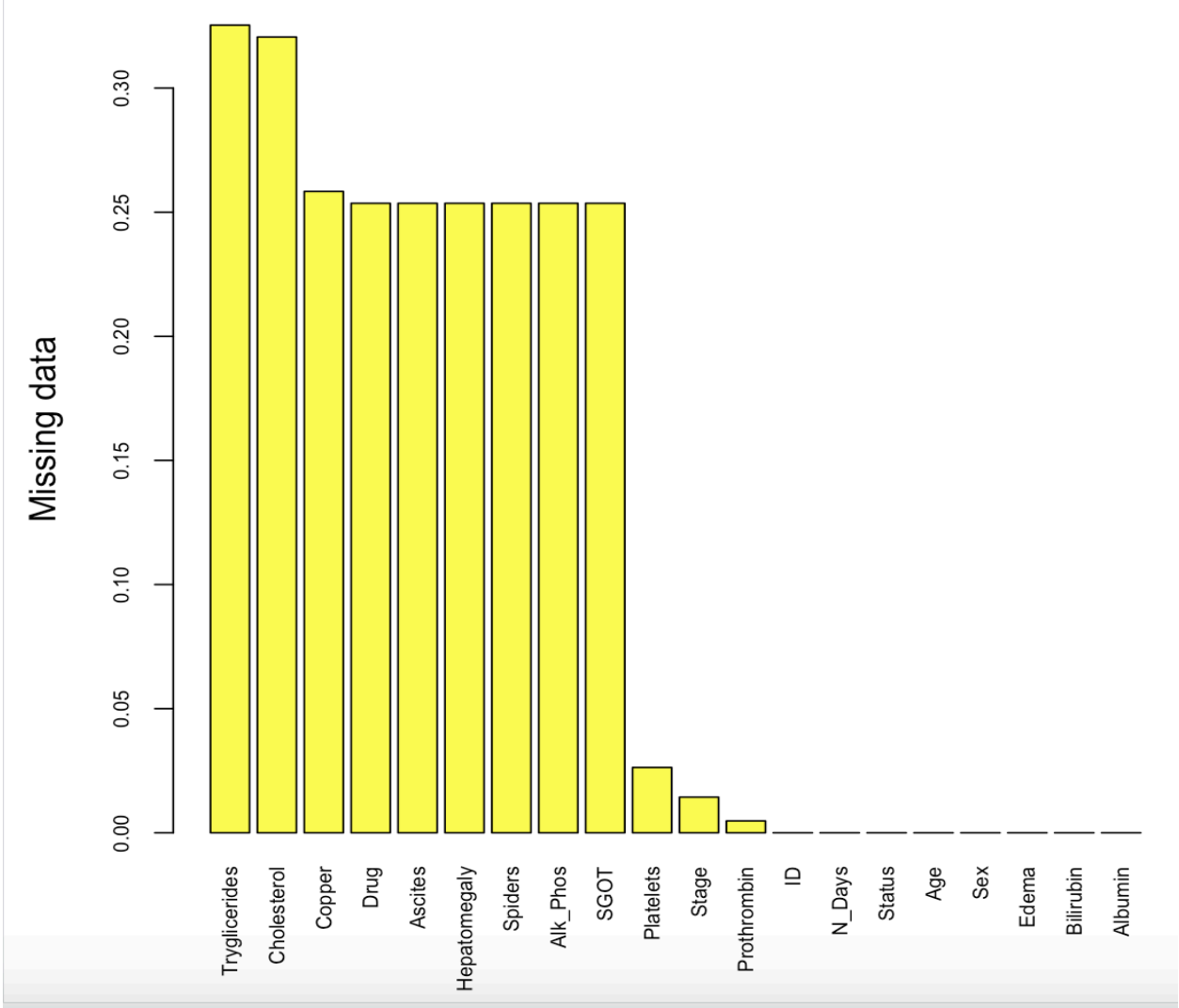

Figure 1: Percentage of Missing Data by Variable

Another thing to consider is the "pattern" of missing data, answering questions like whether some variables are missing together/ following some kind of missing pattern. For this we use the *"mice"* package available in R. From the figure below we can observe **correlated missingness:** The missing data in Drug, Ascites, Hepatomegaly, Spiders, SGOT, and Copper, etc are often missing together, suggesting a possible correlation in how data is missing. This pattern may imply that when one of these is missing, the others are likely to be missing as well.

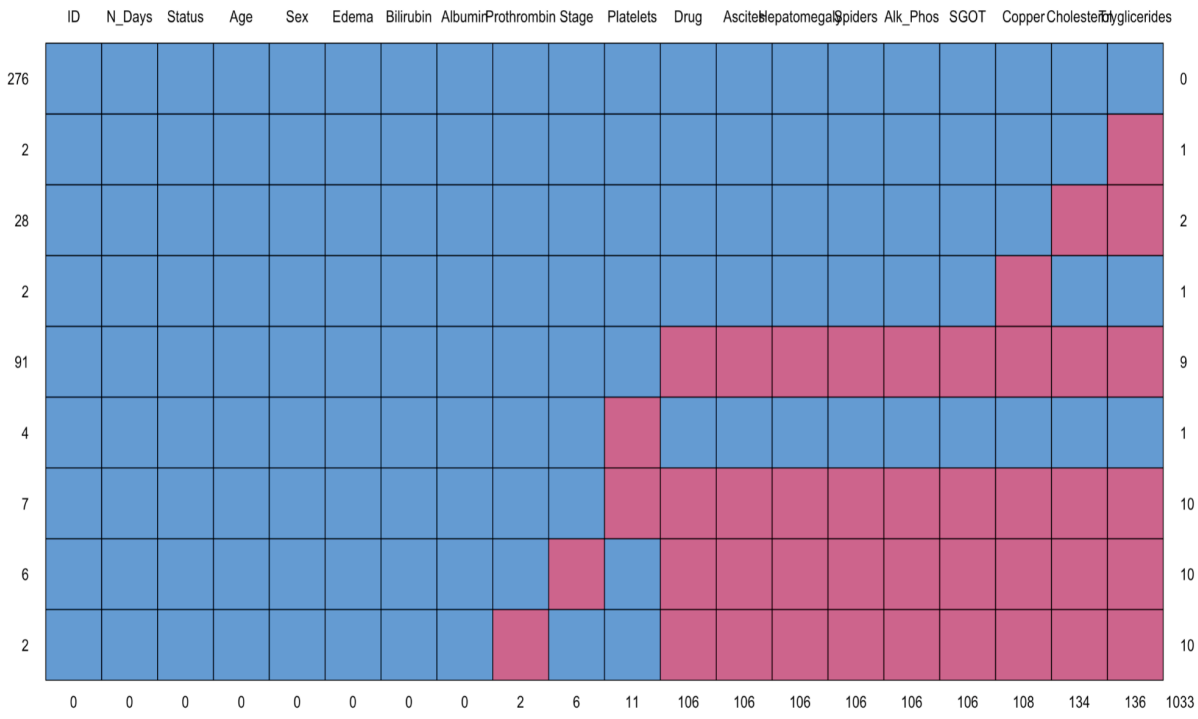

Figure 2: Pattern of missing data

From the figure below, utilizing the "vim" package available in R, we can clearly see that our data only has about 66% of complete cases, which is a small proportion compared to the total dataset, suggesting that simply removing cases with missing data could result in a significant loss of data and thus we must try to implement some kind of missing value imputation technique to better equip ourselves for survival analysis.

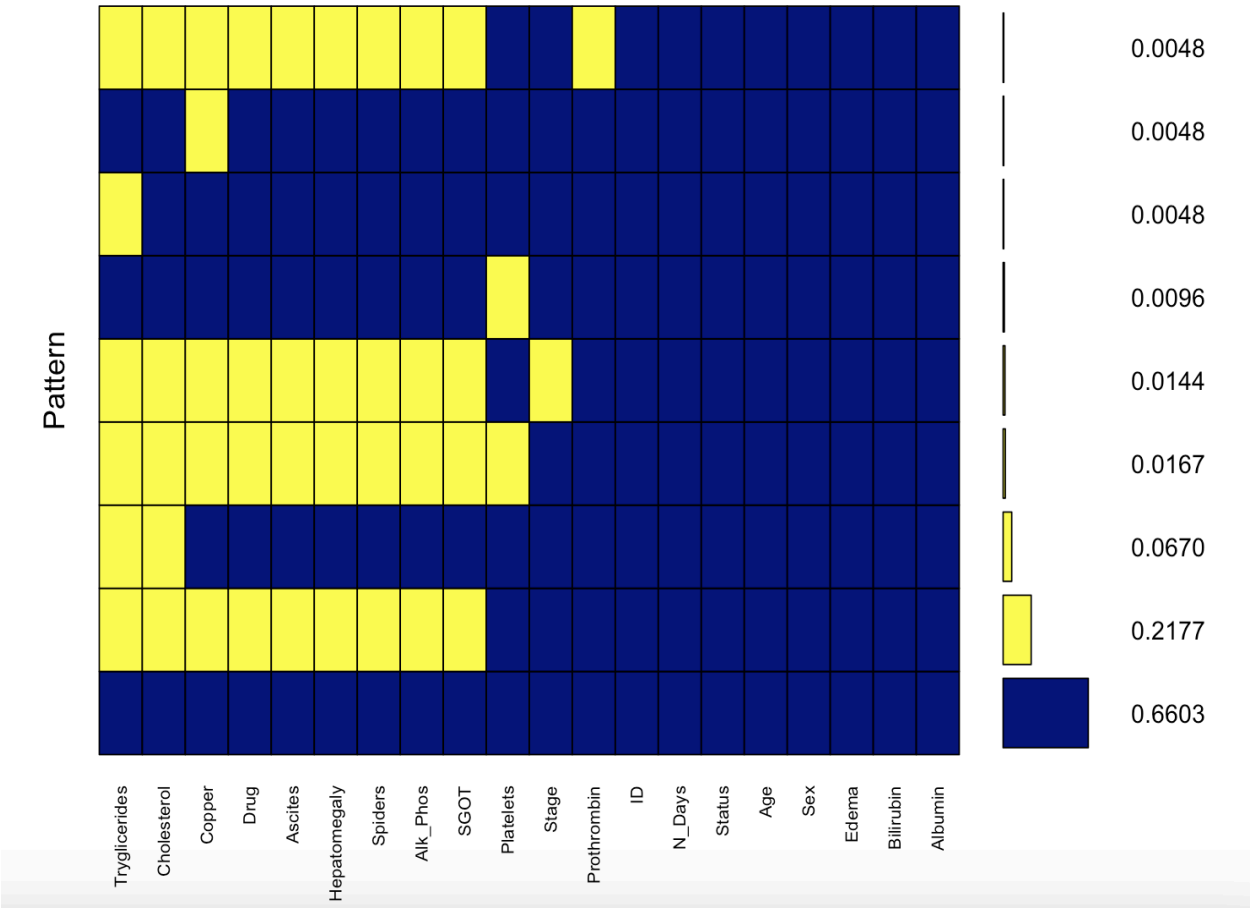

Figure 3: Pattern and Percentage of Missing Data

Little's MCAR Test

Little's MCAR test was first performed for all the continuous variables in the dataset to check for data Missing Completely at Random(MCAR). The results as shown below, (p-value = 0.0626) suggest that the missing data in our dataset could not be definitively ruled out as MCAR, but the p-value is very close to the conventional significance level (0.05). This result does not conclusively confirm that the data is MCAR, but neither does it provide strong evidence against it. There might be a pattern to the missing data that is not completely random but weakly dependent on the observed data.

Given the borderline p-value from Little's MCAR test, **we will proceed with the assumption that the data is Missing at Random (MAR)**, which is a weaker assumption than MCAR and is often considered more realistic in practice. MAR assumes that the propensity for a data point to be missing is related to some of the observed data but not the missing data itself.

```
> mcar_test(data=df_select)
# A tibble: 1 × 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1      55.7    41  0.0626                8
```

Figure 4: Little's MCAR Test

MICE (Multivariate Imputation via Chained Equations)

MICE assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on the observed value and can be predicted using them. It imputes data, on a variable-by-variable basis, by specifying an imputation model per variable.

For example: Suppose we have n1, n2….nk variables. If n1 has missing values, then it will be regressed on other variables n2 to nk. The missing values in n1 will be then replaced by predictive values obtained. Similarly, if n2 has missing values, then n1, n3 to nk variables will be used in the prediction model as independent variables. Later, missing values will be replaced with predicted values.

By default, Linear Regression is used to predict continuous missing values while Logistic Regression is used for categorical missing values. Once this cycle is complete, multiple data sets are generated. These data sets differ only in imputed missing values. Generally, it's considered to be a good practice to build models on these data sets separately and combine their results. [6]

Predictive Mean Matching

While a well-designed parametric model can be effective, PMM imputation is often favored by analysts due to its ability to preserve original data characteristics. PMM respects the discreteness of the data, avoids introducing impossible values, maintains the quantiles' locations, and offers robustness against imputation model misspecification, all without requiring extra effort from the analyst. When the analysis hinges on these data features, using PMM imputation can enhance the quality of inference, ensuring that the results more accurately reflect the underlying data structure.[5]

Analyzing the Distribution of Data

The distribution of Continuous and Categorical Data is given below. One of the ways of checking our MAR assumption and imputation technique is to check the distribution of continuous variables visually.

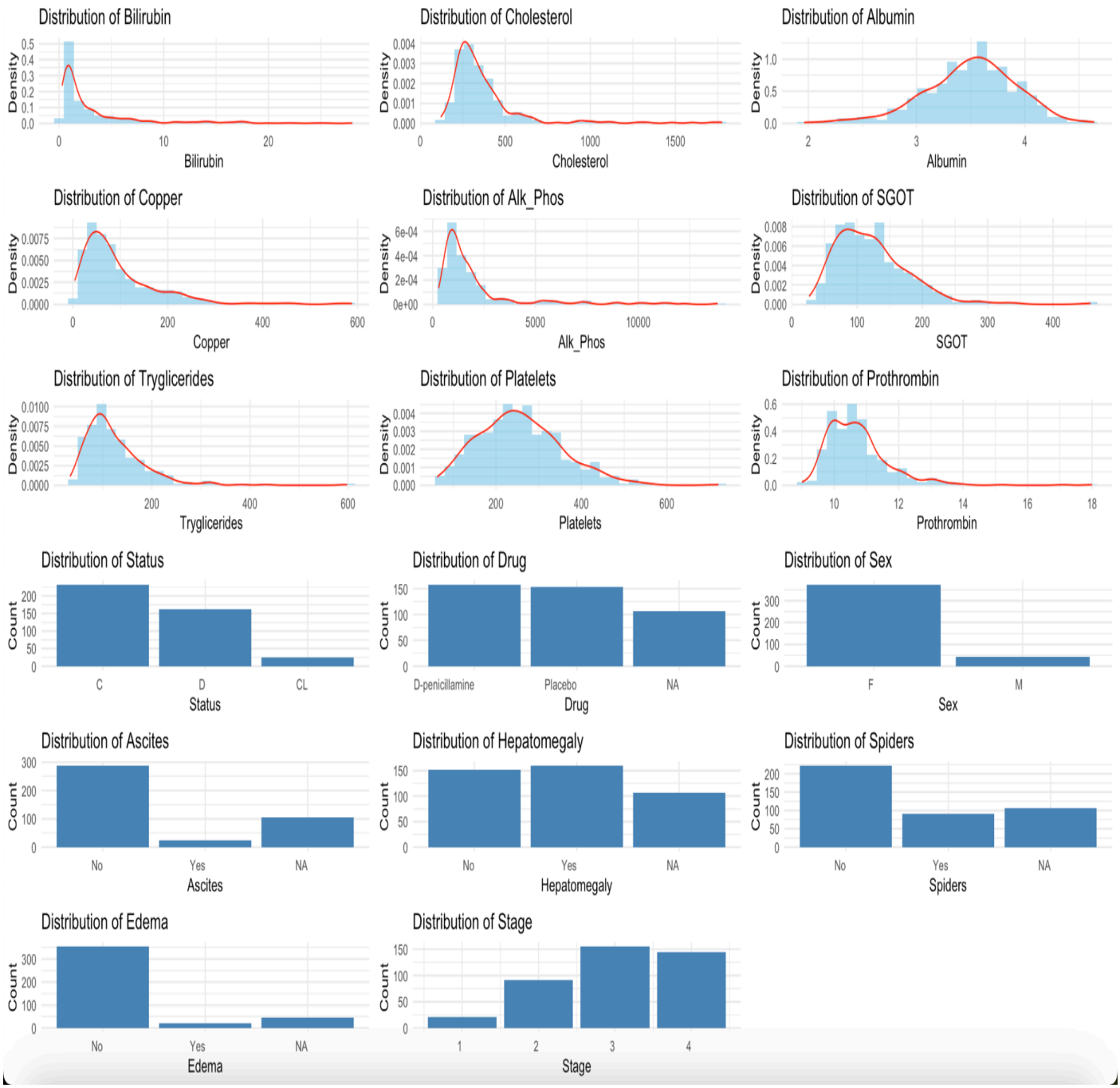

Figure 5: Distribution of Continuous and Non-categorical values Before Imputation

"Densityplot" compares the density of observed data with the ones of imputed data. We expect them to be similar (though not identical) under MAR assumption. [7] That is the case for almost all the continuous variables as can be seen from the figure below.

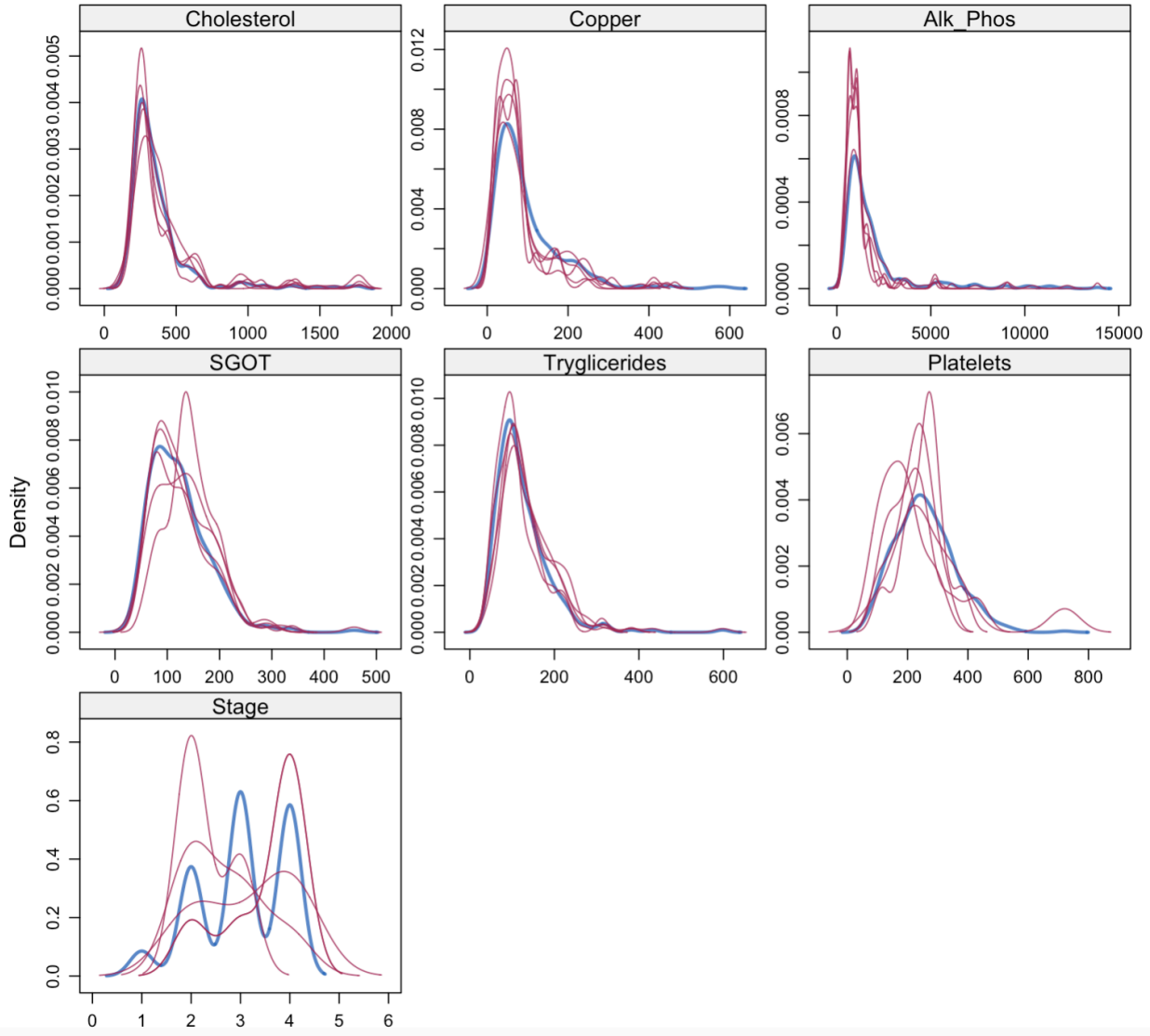

Figure 6: Density Plot of original(blue) and imputed datasets(red).

Survival Analysis

We plan on doing a survival analysis with the following in mind-

- We will treat 'CL' signifying Liver Transplant as Right Censoring while fitting the survival object using the 'Surv' function in R.

- For Kaplan Meier Curve we will use one of our 5 imputed datasets for displaying results in this report. (when we are stratifying by a variable that's been imputed)

Kaplan Meier Curve

A more advanced technique for aggregating survival data is the Kaplan-Meier approach, which uses all of the cases in a series rather than just the ones that are tracked until the chosen cut-off. The method involves breaking up the follow-up period into several smaller time windows and counting the number of cases that were followed up on and events of interest (like deaths) that occurred during each window.

The probability of living until the end of that time period is calculated by multiplying the surviving percentage by the surviving proportions for each of the previous time periods. Next, a temporal plot of this survival probability is created. [8]

From the plot below, we can see the survival curves crossing each other, which is a strong indicator of violation of the proportional hazards assumption. Also, we can see that there are a lot more events and censors (overall samples) present in the case of Female patients leading to a class imbalance. We can also be more confident about the survival curve for Females due to the abundance of data leading to more statistical surety in our inference.

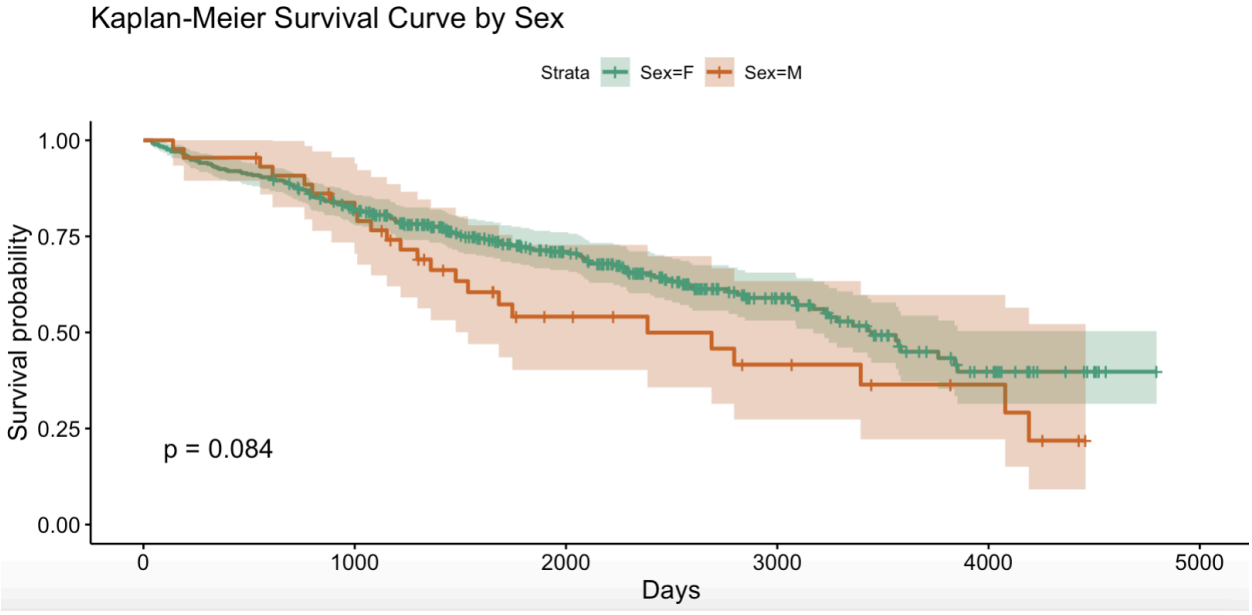

Figure 7: Survival curve stratified by sex.

Even the survival curve stratified by Drugs seems to be crossing each other as seen below, again suggesting that the proportional hazard assumption is violated. This can also suggest that patients who are prescribed D-penicillamine, have higher survival rates initially(up to 5 years)...but can lead to worse survival probabilities after that. This can also be due to the correlation between certain variables in the dataset and needs to be analyzed further.

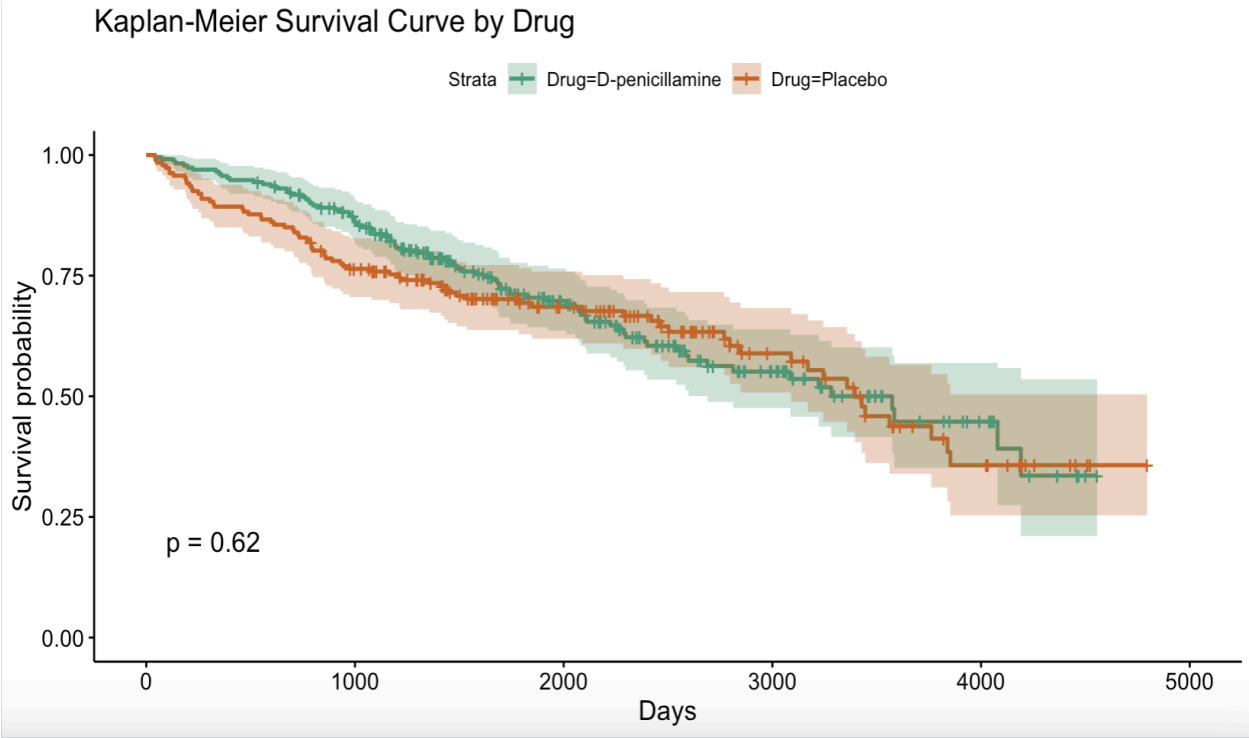

Figure 8: Survival Curve Using One of the Imputed Datasets Stratified by Drug.

In figures 9-12 below, we can see the curves do not cross over and there is a significant difference between patients that suffer from Ascites/Hepatomegaly/Spiders/Edema vs those who do not. The significant difference between the groups highlights the clinical importance of these diseases as a prognostic factor in the studied condition. Patients who suffer from these symptoms have significantly lower survival rates throughout and if we leave the few censored observations, almost all of them died 9 years after being diagnosed(study start time). Also, a significant p-value <0.0001 supports the notion that the survival experiences of the two groups are different.

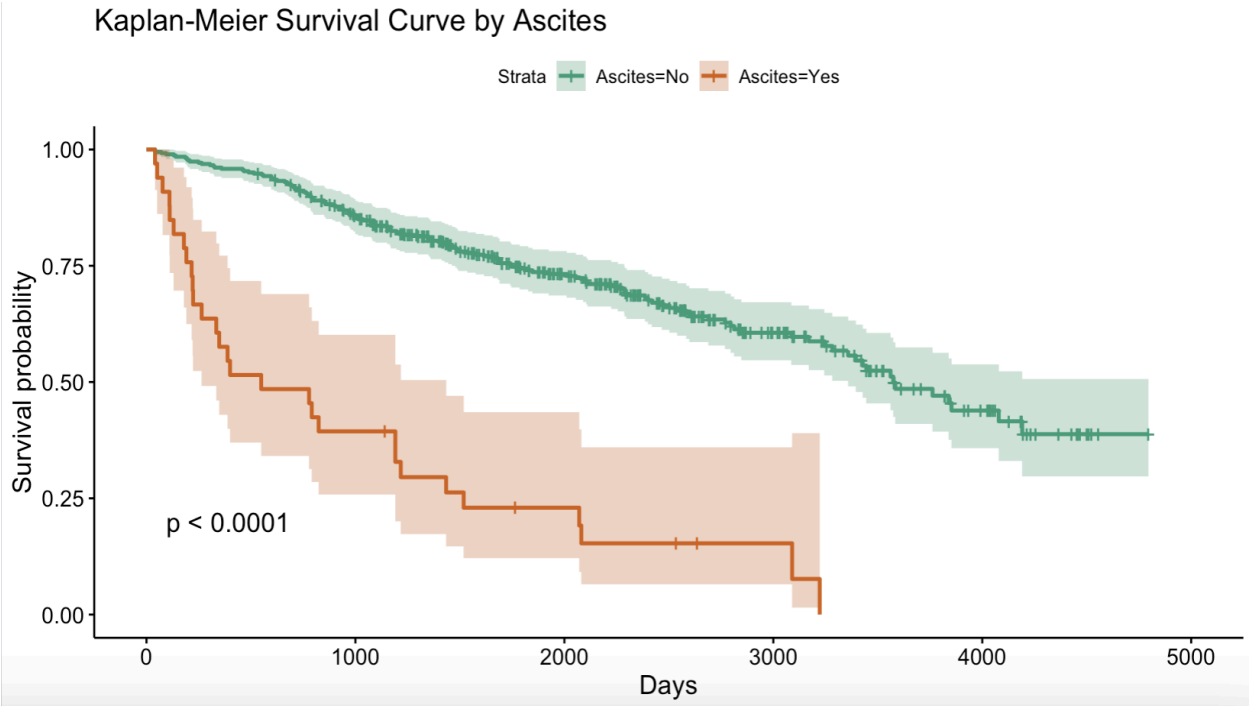

Figure 9: Survival Curve Using One of the Imputed Datasets Stratified by Ascites.

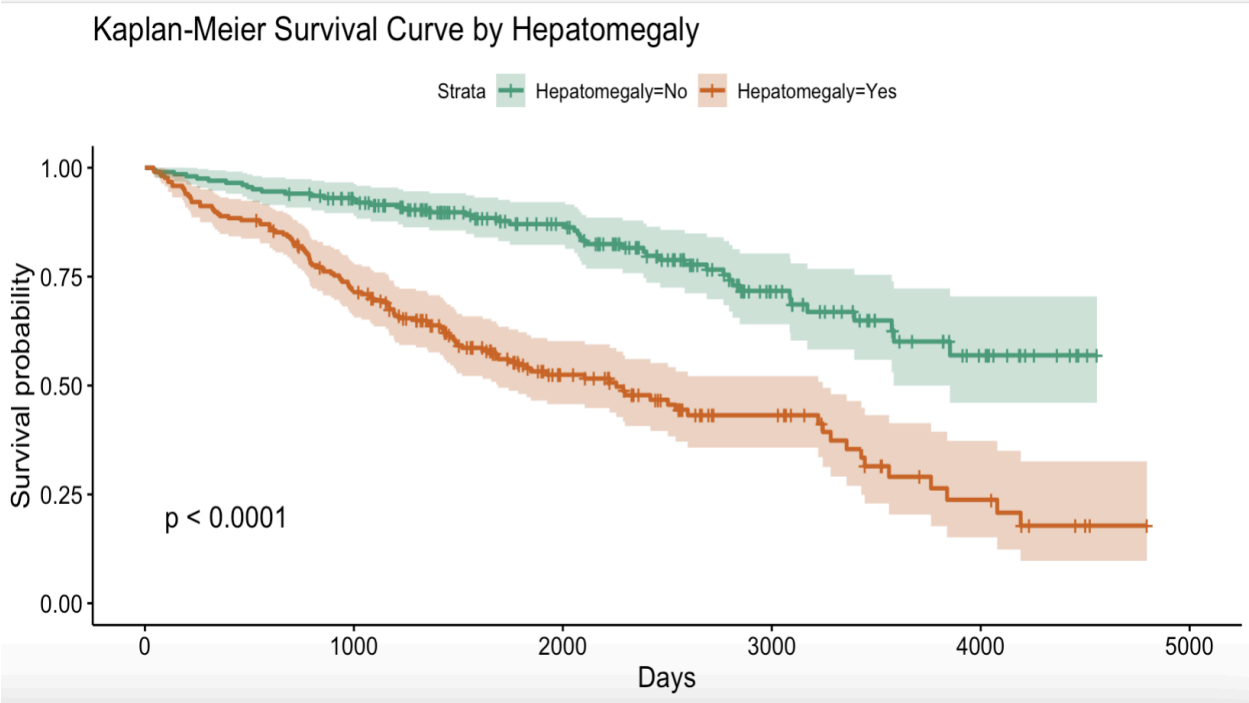

Figure 10: Survival Curve Using One of the Imputed Datasets Stratified by Hepatomegaly.

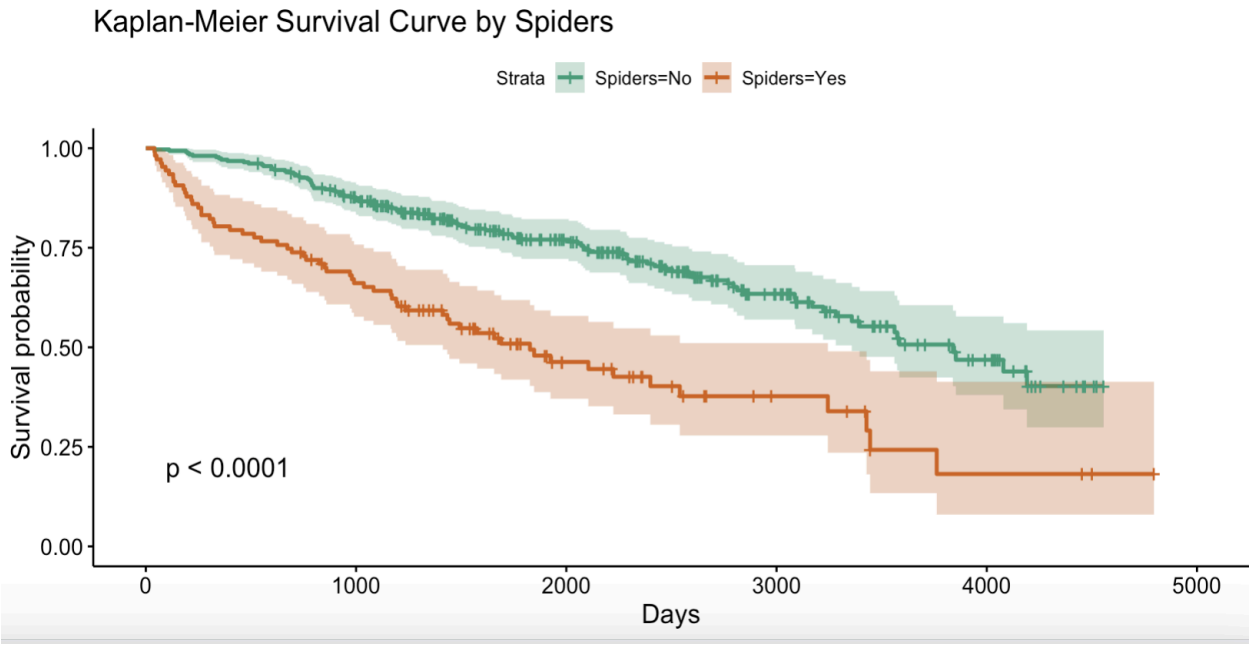

Figure 11: Survival Curve Using One of the Imputed Datasets Stratified by Spiders.

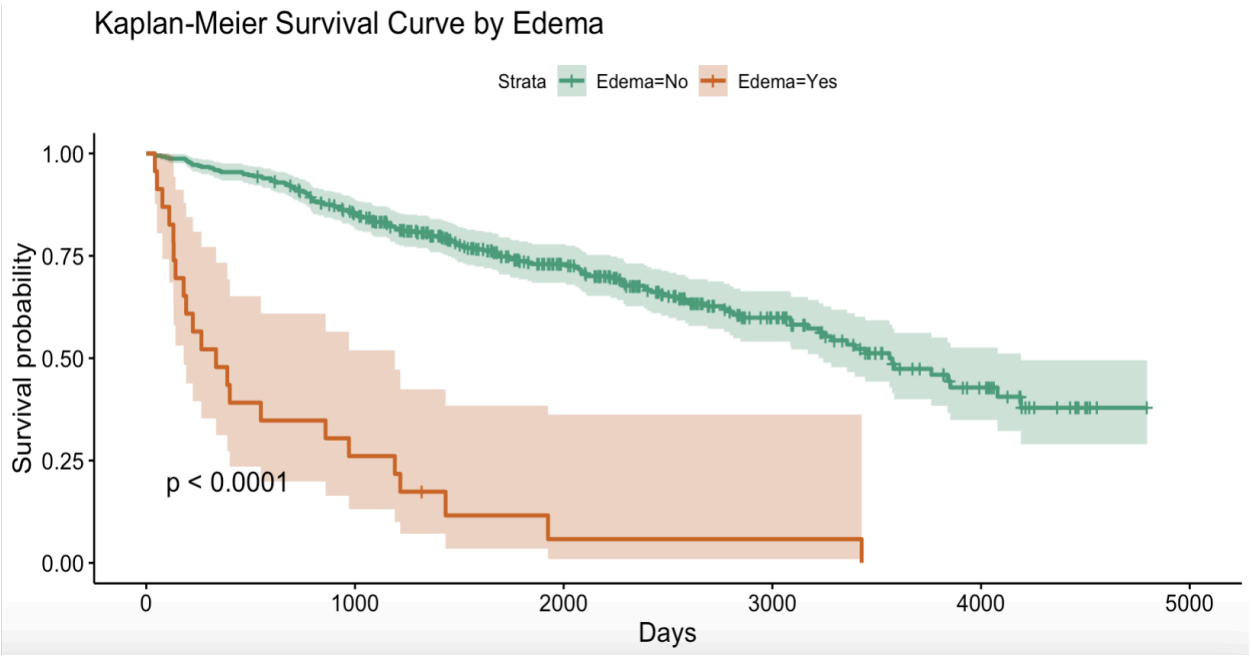

Figure 12: Survival Curve Using One of the Imputed Datasets Stratified by Edema.

Among the 4 diseases above, cases of Edema and Ascites are the most fatal for liver cirrhosis patients as can be inferred from the survival curves. Patients having these two diseases have less than 50% probability of surviving the past three years after diagnosis.

Cox Regression

Eventually, employing a categorical division for K-M plots becomes problematic, and you may also wish to evaluate the interaction of several variables that affect survival. Multivariate effects can be simultaneously modeled and the effects of both continuous and categorical variables can be evaluated using Cox PH regression. [9]

```
Call:
coxph(formula = surv_obj ~ Drug + Ascites + Hepatomegaly + Spiders +
    Age + Sex + Bilirubin + Albumin + Cholesterol + Copper +
    Alk_Phos + SGOT + Trygliceride + Platelets + Prothrombin,
    data = ex_data)

  n= 418, number of events= 161

                     coef  exp(coef)   se(coef)      z Pr(>|z|)
DrugPlacebo      1.494e-01  1.161e+00  1.750e-01  0.853 0.393390
AscitesYes       6.161e-01  1.852e+00  2.748e-01  2.242 0.024957 *
HepatomegalyYes  5.632e-01  1.756e+00  1.982e-01  2.842 0.004484 **
SpidersYes       2.990e-01  1.348e+00  1.991e-01  1.502 0.133121
Age              3.987e-02  1.041e+00  8.841e-03  4.510 6.48e-06 ***
SexM            -1.717e-02  9.830e-01  2.654e-01 -0.065 0.948416
Bilirubin        8.207e-02  1.086e+00  1.890e-02  4.344 1.40e-05 ***
Albumin         -7.218e-01  4.859e-01  2.137e-01 -3.378 0.000729 ***
Cholesterol      2.395e-04  1.000e+00  3.169e-04  0.756 0.449936
Copper           1.955e-03  1.002e+00  1.004e-03  1.948 0.051462 .
Alk_Phos         7.197e-06  1.000e+00  3.252e-05  0.221 0.824853
SGOT             3.770e-03  1.004e+00  1.554e-03  2.426 0.015260 *
Trygliceride    -8.750e-04  9.991e-01  1.091e-03 -0.802 0.422390
Platelets       -1.979e-04  9.998e-01  9.681e-04 -0.204 0.838047
Prothrombin      1.888e-01  1.208e+00  6.445e-02  2.930 0.003387 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 13: Summary statistics on Cox Regression on one of the Imputed Datasets.

The pooled analysis (Figure 13), which incorporates estimates from five different imputed versions of the dataset, showed broader confidence intervals and higher p-values. This approach, by accounting for the variability across multiple plausible datasets, likely offers a more conservative estimate of effect sizes and a realistic assessment of statistical significance. It reflects a comprehensive handling of the data's inherent uncertainties, enhancing the credibility of the findings by mitigating potential biases introduced by any single imputation process.

```
# A tibble: 15 × 7
   term              estimate std.error statistic    df  p.value significance
   <fct>                <dbl>     <dbl>     <dbl> <dbl>    <dbl> <chr>
 1 DrugPlacebo      0.0493     0.202         0.244  35.6 0.808    ""
 2 AscitesYes       0.557      0.329         1.69   41.6 0.0978   ""
 3 HepatomegalyYes  0.608      0.263         2.32   17.1 0.0332   "*"
 4 SpidersYes       0.135      0.246         0.549  23.3 0.589    ""
 5 Age              0.0353     0.00998       3.54   54.0 0.000827 "***"
 6 SexM            -0.0683     0.271        -0.251 140.  0.802    ""
 7 Bilirubin        0.0827     0.0215        3.84   44.5 0.000387 "***"
 8 Albumin         -0.776      0.228        -3.40   95.2 0.000993 "***"
 9 Cholesterol      0.000350   0.000347      1.01   73.3 0.317    ""
10 Copper           0.00268    0.00124       2.16   28.7 0.0396   "*"
11 Alk_Phos        -0.00000995 0.0000404    -0.246  47.0 0.806    ""
12 SGOT             0.00312    0.00227       1.38   10.5 0.198    ""
13 Trygliceride    -0.000844   0.00122      -0.693  82.8 0.490    ""
14 Platelets       -0.000507   0.00102      -0.498  78.0 0.620    ""
15 Prothrombin      0.223      0.0774        2.88   26.8 0.00773  "**"
```

Figure 14: Summary statistics on Cox Regression all of the imputed datasets with results pooled.

Results and Discussion

From the summary of our pooled Cox proportional hazards model, several continuous predictors have shown significant p-values, suggesting that they play a notable role in influencing survival outcomes.

1. **Age**:
    - *Coefficient:* The coefficient for age (0.03534) is positive, which implies that the risk of the event (e.g., death) increases as age increases. This is typical in survival analysis where older age is often associated with higher mortality risk.
    - *Hazard Ratio:* The exponentiated coefficient (exp(0.03534) ≈ 1.036) indicates that each additional year of age increases the risk of the event by approximately 3.6%. This effect is statistically significant, and the small effect size is typical for age where incremental changes accumulate over time.
    - *Implications:* This result underscores the importance of age as a risk factor in the prognosis of liver cirrhosis, guiding clinicians to consider age as a critical factor in their management strategies.

2. **Bilirubin:**
    - *Coefficient:* The positive coefficient (0.08268) for bilirubin suggests that higher bilirubin levels are associated with an increased hazard of death. Bilirubin is a byproduct of the body's metabolism of worn-out red blood cells; high levels can indicate liver dysfunction.
    - *Hazard Ratio:* With an exponentiated coefficient of about 1.086, each unit increase in bilirubin levels increases the risk by 8.6%. This finding is clinically significant as it highlights bilirubin as a powerful prognostic marker.
    - *Implications:* This result is particularly valuable for clinicians monitoring liver function in cirrhosis patients, suggesting that interventions to manage bilirubin levels could be crucial.
3. **Albumin:**
    - *Coefficient:* The negative coefficient (-0.7757) for albumin indicates that higher albumin levels are associated with a reduced risk of the event. Albumin, a major protein made by the liver, is crucial for maintaining fluid balance and nourishing tissues.
    - *Hazard Ratio:* The hazard ratio of about 0.46 means that higher albumin levels are protective, halving the risk approximately. This is a significant protective effect, showing the importance of maintaining good liver function.
    - *Implications:* These findings reinforce the clinical importance of nutritional support and liver function optimization in managing patients with liver disease.
4. **Prothrombin Time:**
    - *Coefficient:* The positive coefficient (0.2229) for prothrombin time suggests that longer times are associated with increased hazards. Prothrombin time is a measure of how quickly blood clots, with longer times indicating potential liver impairment.
    - *Hazard Ratio:* The hazard ratio of about 1.25 indicates that longer clotting times significantly increase risk, which is clinically relevant for assessing liver function and disease progression.

- *Implications:* This result suggests the importance of regular monitoring and possibly interventions to manage coagulation factors in patients with liver cirrhosis.

In the initial univariate Kaplan-Meier analysis, significant differences in survival were observed for patients with Ascites, Hepatomegaly, Spiders, and other conditions, suggesting these factors individually impact patient survival. However, when adjusting for multiple covariates in the Cox proportional hazards model, only Hepatomegaly retained its statistical significance. This shift underscores the importance of considering confounding factors that may influence the apparent effects of individual predictors in univariate settings. The multivariate Cox model provides a more nuanced understanding by accounting for potential interactions and confounders among the predictors, offering a clearer picture of which factors independently affect survival. This distinction is crucial for clinical decision-making, focusing interventions on factors that independently predict outcomes.

References

[1] https://www.mayo.edu/research/documents/pbcseqhtml/doc-10027141

[2] Hepatic Cirrhosis

[3] Goldberg, David*,1,2; Mantero, Alejandro2; Kaplan, David3,4; Delgado, Cindy1; John, Binu1,5; Nuchovich, Nadine1; Emanuel, Ezekiel6; Reese, Peter P.7. Accurate long‐term prediction of death for patients with cirrhosis. Hepatology 76(3):p 700-711, September 2022. | DOI: 10.1002/hep.32457

[4] https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1

[5] *Austin PC, White IR, Lee DS, van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Can J Cardiol. 2021 Sep;37(9):1322-1331. doi: 10.1016/j.cjca.2020.11.010. Epub 2020 Dec 1. PMID: 33276049; PMCID: PMC8499698.*

[6] https://medium.com/analyticgeeks/missing-value-imputation-techniques-in-r-be9ec5c97a07

[7] https://web.maths.unsw.edu.au/~dwarton/missingDataLab.html

[8] Bertil Damato, Azzam Taktak, Chapter 2 - Survival after Treatment of Intraocular Melanoma, Editor(s): Azzam F.G. Taktak, Anthony C. Fisher, Outcome Prediction in Cancer, Elsevier,

2007, Pages 27-41, ISBN 9780444528551
(https://www.sciencedirect.com/science/article/pii/B9780444528551500040)
[9] https://bioconnector.github.io/workshops/r-survival.html

Appendix

Use of ChatGPT-

ChatGPT's help was utilized in writing cleaner code and proofreading the report's content.

R Code-

```
#Loading the required libraries
library(dplyr)
library(tidyr)
library(tidyverse)
library(survival)
library(survminer)
library(mice)
library(gridExtra)
#Loading the Dataset
data <- read.csv("/Users/manasarthak/Downloads/cirrhosis.csv")
head(data)
# Convert days to years for age
data$Age <- data$Age / 365.25

# Calculate the percentage of missing data for each column
missing_data_summary<-data %>%
  summarise(across(everything(), ~sum(is.na(.))/n()*100)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "MissingPercentage")%>%
  mutate(Variable = factor(Variable, levels = Variable[order(MissingPercentage)]))

ggplot(missing_data_summary, aes(x = Variable, y = MissingPercentage)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Percentage of Missing Data by Variable", y = "Percentage Missing", x = "Variables")


md.pattern(data)
#beautification
install.packages("VIM")
library(VIM)
# Aggregated missing data plot
aggr(data, col=c('navyblue', 'yellow'), numbers=TRUE, sortVars=TRUE,
     labels=names(data), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```

```
# Convert 'Drug' to factor with appropriate levels
data$Drug <- factor(data$Drug, levels = c("D-penicillamine", "Placebo"), labels = c("D-penicillamine", "Placebo"))
data$Status <- factor(data$Status, levels = c("C", "D","CL"), labels = c("C", "D","CL"))
data$Sex <- factor(data$Sex, levels = c("F", "M"), labels = c("F", "M"))

# List of true binary variables
binary_vars <- c("Edema","Ascites", "Hepatomegaly", "Spiders")

# Convert binary variables to factors with levels "N" and "Y"
data[binary_vars] <- lapply(data[binary_vars], function(x) {
  factor(x, levels = c("N", "Y"), labels = c("No", "Yes"))
})

# Check the structure to confirm changes
str(data[c("Drug", binary_vars)])

# Continuous Variables
continuous_vars <- c("Bilirubin", "Cholesterol", "Albumin", "Copper", "Alk_Phos", "SGOT", "Tryglicerides", "Platelets", "Prothrombin")
df_select <- data[,continuous_vars]
library(naniar)
mcar_test(data=df_select)

continuous_plots <- lapply(continuous_vars, function(var) {
  ggplot(data, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", alpha = 0.7) +
    geom_density(color = "red") +
    labs(title = paste("Distribution of", var), x = var, y = "Density") +
    theme_minimal()
})

# Categorical Variables
categorical_vars <- c("Status", "Drug", "Sex", "Ascites", "Hepatomegaly", "Spiders", "Edema", "Stage")

# Create plots for each categorical variable
categorical_plots <- lapply(categorical_vars, function(var) {
  ggplot(data, aes_string(x = var)) +
    geom_bar(fill = "steelblue") +
    labs(title = paste("Distribution of", var), x = var, y = "Count") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 0, hjust = 1))
})
# Combine all plots
all_plots <- c(continuous_plots, categorical_plots)

# Use grid.arrange to display all plots together
do.call(grid.arrange, c(all_plots, ncol = 3))

# Initialize the MICE model for PMM
init_data <- mice(data, method = 'pmm', m = 5, seed = 123, printFlag = FALSE)
summary(init_data)

#visualize the "similarity" between
```

```
densityplot(init_data)


#For surival curves we will use one of the imputed datasets
#One of the 5 imputed datasets
ex_data<-complete(init_data,4)

# Convert Status to a survival object, assuming 'D' stands for the event (death)
surv_obj <- Surv(ex_data$N_Days, ex_data$Status == "D")

# Fit Kaplan-Meier survival curves
km_fit <- survfit(surv_obj ~ Sex, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
  title = "Kaplan-Meier Survival Curve by Sex" # Title of the plot
)

print(g)

# Fit Kaplan-Meier survival curve startified by Drug(primary variable for analysis)
km_fit <- survfit(surv_obj ~ Drug, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = ex_data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
  title = "Kaplan-Meier Survival Curve by Drug" # Title of the plot
)

print(g)

# Fit Kaplan-Meier survival curve startified by Ascites
km_fit <- survfit(surv_obj ~ Ascites, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = ex_data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
```

```
  title = "Kaplan-Meier Survival Curve by Ascites" # Title of the plot
)

print(g)

# Fit Kaplan-Meier survival curve startified by Hepatomegaly
km_fit <- survfit(surv_obj ~ Hepatomegaly, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = ex_data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
  title = "Kaplan-Meier Survival Curve by Hepatomegaly" # Title of the plot
)

print(g)

# Fit Kaplan-Meier survival curve startified by Spider
km_fit <- survfit(surv_obj ~ Spiders, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = ex_data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
  title = "Kaplan-Meier Survival Curve by Spiders" # Title of the plot
)

print(g)

# Fit Kaplan-Meier survival curve startified by Spider
km_fit <- survfit(surv_obj ~ Edema, data = ex_data)

# Plot the Kaplan-Meier survival curves
g <- ggsurvplot(
  km_fit,
  data = ex_data,
  pval = TRUE,              # Show p-value of the log-rank test
  conf.int = TRUE,          # Show confidence intervals
  palette = "Dark2",        # Color palette
  xlab = "Days",            # Label for the x-axis
  ylab = "Survival probability", # Label for the y-axis
  title = "Kaplan-Meier Survival Curve by Edema" # Title of the plot
)

print(g)
```

```
#Cox Regression

# Fit a Cox proportional hazards model to each imputed dataset
cox_models <- with(data = init_data, exp = {
  # Ensure the event is correctly coded within the function scope
  event <- ifelse(Status == "D", 1, 0)
  surv_obj <- Surv(time = N_Days, event = event)
   coxph(surv_obj ~ Drug+Ascites+Hepatomegaly+Spiders+Age+Sex+Bilirubin+Albumin +Cholesterol+ Copper+
Alk_Phos+ SGOT+ Tryglicerides+Platelets+Prothrombin)
})
# Pool the results of fitting the model to each imputed dataset
pooled_results <- pool(cox_models)

# Print the summary of the pooled results
df<-as_tibble(summary(pooled_results))
df$significance <- ifelse(df$p.value < 0.001, "***",
                          ifelse(df$p.value < 0.01, "**",
                                 ifelse(df$p.value < 0.05, "*", "")))
df


# Assuming 'ex_data' is your dataset with imputed values
# Ensure the Status column is correctly encoded for the event (1 for event occurred, 0 for censored)
ex_data$event <- ifelse(ex_data$Status == "D", 1, 0)

# Create the survival object
surv_obj <- Surv(time = ex_data$N_Days, event = ex_data$event)

# Fit the Cox proportional hazards model
cox_model <- coxph(surv_obj ~ Drug+Ascites+Hepatomegaly+Spiders+Age+Sex+Bilirubin+Albumin
+Cholesterol+ Copper+ Alk_Phos+ SGOT+ Tryglicerides+Platelets+Prothrombin, data = ex_data)

# Summary of the Cox model
summary(cox_model)

library(corrplot)
continuous_vars <- ex_data[, c("Age", "Bilirubin", "Albumin", "Prothrombin", "Cholesterol", "Copper",
"Alk_Phos", "SGOT", "Tryglicerides", "Platelets")]

# Compute correlation matrix
cor_matrix <- cor(continuous_vars, use = "complete.obs")  # Handling missing values

# Plot the correlation matrix
corrplot(cor_matrix, method = "circle", type = "upper",
         order = "hclust",
         tl.col = "black", tl.srt = 30,  # Text label color and rotation
         title = "Correlation Matrix of Continuous Predictors")
```