Analysis of Time to Failure in Primary Biliary Cirrhosis Clinical Trials

Name: Ronnald Le
Data: Pbc3.txt

Introduction

The PBC3 clinical trial conducted in six European hospitals between 1983 and 1987 with primary biliary cirrhosis (PBC), where patients randomly assigned to receive Cyclosporin A (CyA) or a placebo. The main outcome became the "time to failure of medical treatment," and is defined as death or liver transplantation. This study dives deep into the factors that influence this time-to-failure, taking into account various patient characteristis recorded at the beginning of time.

Methods

- The dataset 'Pbc3' was loaded with the “Import Dataset” method. I managed to give it a quick scan and evaluate any data that may be missing, and/or unique.
- I loaded all of the necessary libraries: tidyr, dplyr, survival, ggplot2, and survminer.
- After loading the libraries, I performed data cleaning by checking for missing values and removing rows with missing data. This step makes sure that my analysis is based on complete records. I also calculated the number of patient records and the mean age of patients.
- The dataset contains 275 patient records after removing rows with missing data with an average age of 54.15 years. The distribution of the *status* variable shows that 205 patients are censored, 24 are in the process of liver transplantation, and 46 patients died during the study.

I provided a summary statistics for clinical measurements for a quick overview of the range and distribution in order to help understand the context in terms of plots:

Clinical Measurement	**Min**	**Max**
Creatinine (crea)	35.00	199.00
Albumin (alb)	24.00	58.00
Bilirubin (bili)	2.333	453.100
Alkaline Phosphatase (alkph)	66.33	5108.00
Aspartate Transaminase (asptr)	10.50	316.30

For survival analysis, I attempted to recode *status* as a binary variable (1 for an event, 0 for censoring, and 2 for dead). I created a survival object *surv_object* for (censored, transplant, and dead) to help analyze how long it takes for events to happen through time-to-event analysis. I provided a visual representation of the Kaplan-Meier survival curve (**Figure 1.**) to help visualize the differences in survival probabilities. The Kaplan-Meier survival curves gives a visual representation of the survival probabilities over time for different patient groups in the dataset. Each curve is meant for specific category of patients, and is differentiated by their status: censored, liver transplantation, or deceased.

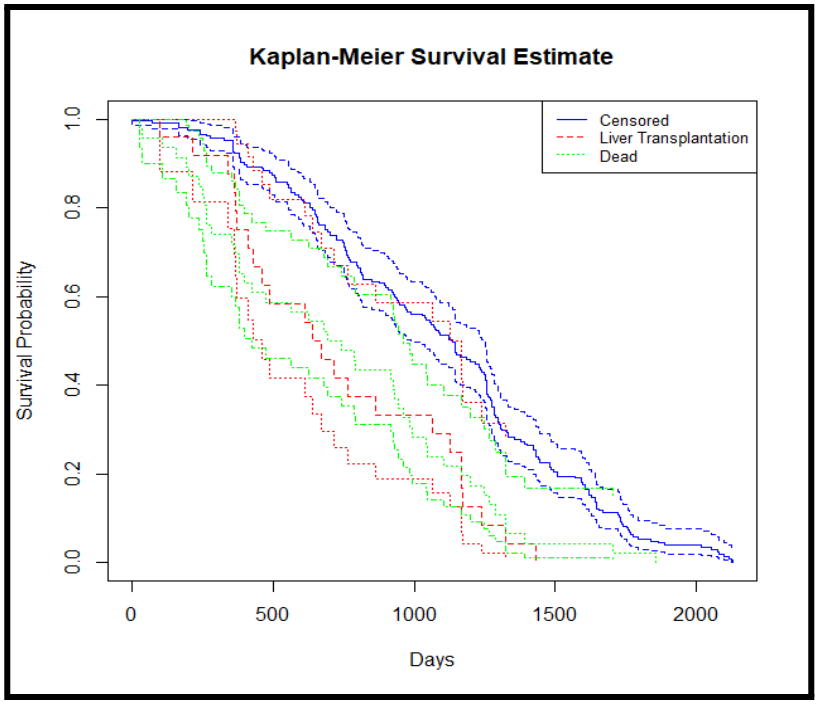

Figure 1.

The Cox proportional hazards regression model used as: "*cox_model*" was fitted to see the impact of various covariates on survival. Based on the output of "*cox_model*" in **Figure 2.,** the concordance statistic is 0.831, indicating a good predictive ability of the model, therefore a value closer to 1.0 suggests better concordance. We can see that Treatment (tment) shows a coefficient of -0.5427 with a p-value of 0.08468. Age (age) has a coefficient of 0.06271 with a p-value of 0.00183. Sex (sex) has a coefficient of 1.3600 with a p-value of 0.00175. Creatinine (crea), alkaline phosphatase (alkph), and aspartate transaminase (asptr) do not show any significant effects since the p-value is high. On the other hand, Albumin (alb) and bilirubin (bili) have significant effects.

```
coxph(formula = surv_object ~ tment + age + sex + crea + alb +
    bili + alkph + asptr + weight, data = data_cleaned)

  n= 275, number of events= 46

             coef  exp(coef)   se(coef)      z Pr(>|z|)
tment   -5.427e-01  5.812e-01  3.147e-01 -1.724  0.08468 .
age      6.271e-02  1.065e+00  2.012e-02  3.116  0.00183 **
sex      1.360e+00  3.895e+00  4.345e-01  3.130  0.00175 **
crea    -1.304e-02  9.870e-01  1.153e-02 -1.131  0.25808
alb     -1.037e-01  9.015e-01  3.414e-02 -3.037  0.00239 **
bili     8.516e-03  1.009e+00  1.898e-03  4.488  7.2e-06 ***
alkph    3.623e-05  1.000e+00  1.878e-04  0.193  0.84702
asptr    3.712e-03  1.004e+00  3.265e-03  1.137  0.25565
weight -4.872e-02  9.524e-01  2.132e-02 -2.285  0.02229 *
```

Figure 2.

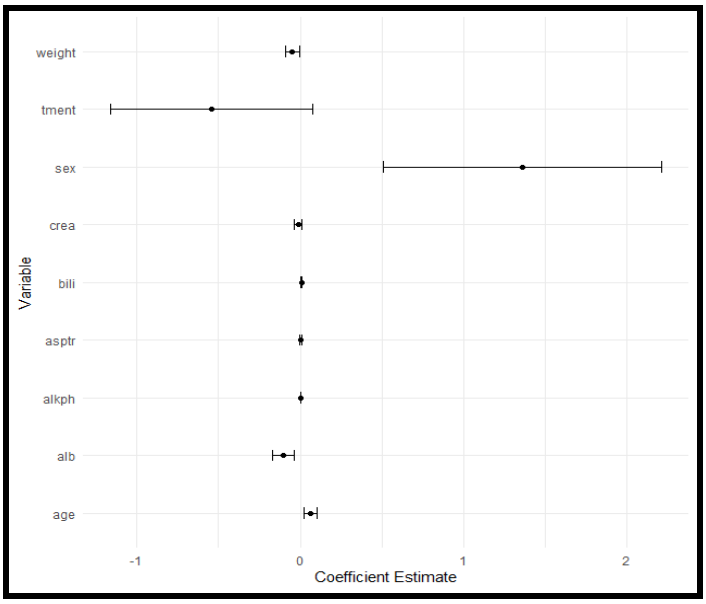

Figure 3.

For **Figure 3.**, this plot shows the coefficients of the Cox proportional hazards model, along with their confidence intervals. Each of the point represents the estimated coefficient for a variable, and the error bars represent the 95% confidence intervals. This plot shows the position of the points along the y-axis, which indicates the direction and size of each variable's effect. A point above zero shows a positive effect (increasing the hazard), while below zero shows a negative effect (decreasing the hazard).

To validate the Cox model's performance in **Figure 4.**, I split the data into training and testing sets. The model was fitted on the training data "*cox_model_train*", and the risk scores are predicted for the testing data. Risk groups are created based on the median risk score. Kaplan-Meier survival plots are generated to compare survival between the risk groups.

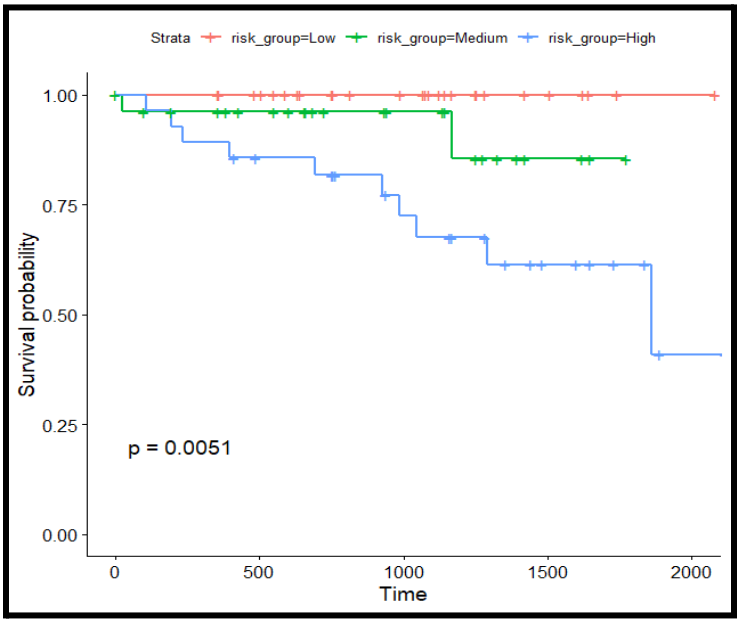

Figure 4.

I created plots for univariate analysis for continuous and categorical variables. For continuous variables (e.g., age, creatinine, albumin), I created histograms to explore their distributions. For categorical variables (e.g., hospital unit, treatment type), I generated bar plots to display their frequencies.

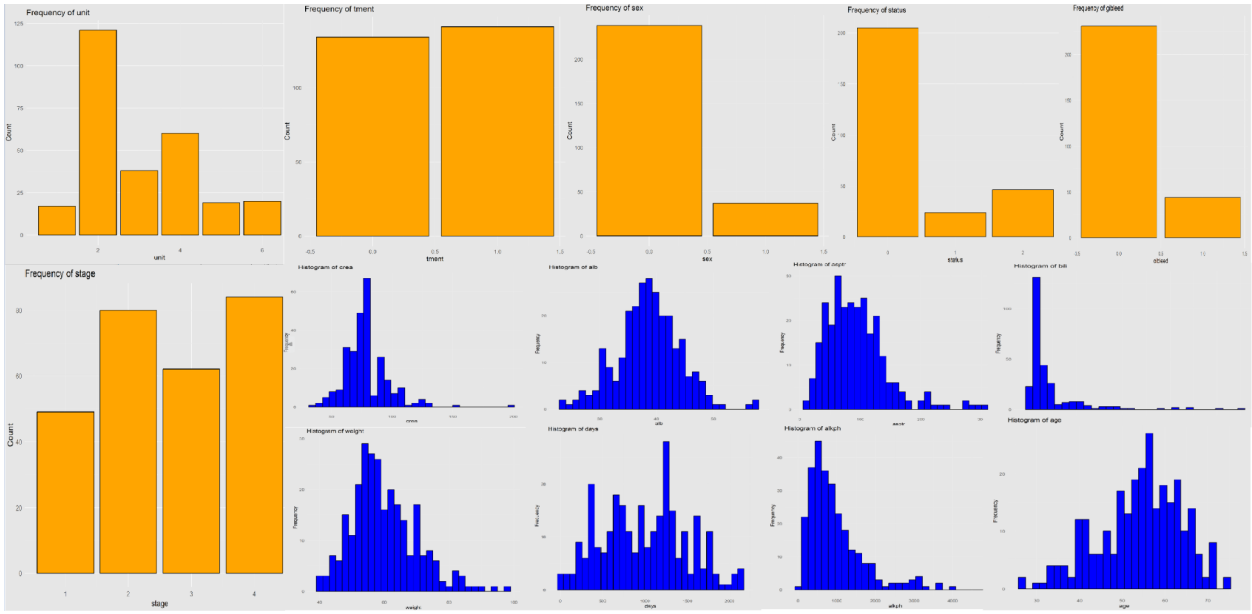

Figure 5.

Results

Figure 1. The plot shows how survival probability decreases over time for each patient group. This trend is important, because of the overall effectiveness of the medical treatments and how the progression of the disease in different patient categories displays. Differences in the survival curves between the groups can be observed. For example, the curve for the liver transplantation shows a steeper decline compared to the others, which indicates a faster decrease in survival probability for this group. This suggests that patients in this group may have a more aggressive form of the disease or are less responsive to treatment.

Figure 2, The model suggests that age, sex, albumin, bilirubin, and weight are significant predictors of the time to treatment failure in PBC patients. Bilirubin levels have a notable impact on increasing hazards, while higher albumin levels and weight seems to deal with decreased hazards.

Figure 3. We can see that the coefficient that are above zero and have confidence intervals not crossing zero indicates that there is a significant positive effect on the hazard (bili, alkph). Those below zero and have confidence interval not crossing zero indicates significant negative effect on the hazard (weight, alb). This plot is important when understanding the factors that influence the time to treatment failure in patients with primary biliary cirrhosis.

Figure 4. This plot visualizes the survival probabilities over time for different groups within the study. Each line in the plot represents a different group's survival probability over time. The

x-axis represents time, and the y-axis represents the survival probability. We can see that the High group drops more steeply suggesting a lower survival probability over time for that group, which will lead to a high risk of event of interest. For the Low group, we can see that they have long term survival rates. We can see “+” like shapes around the survival curves, which is a tick representing censored data points where a patient may exit the study without the event occuring. Confidence intervals provide a range for the true survival curve.

Figure 5. (barplots) The gibleed plot shows a comparison of counts between two categories. The category 0 shows that the majority of patients did not experience gastrointestinal bleeding. On the other hand, the 1 category has a much lower count, indicating fewer patients had this condition. For the sex bar plot,between the two groups of male and female patients. There seems to be more female among the two categories having a much higher count. For the stage bar plot, it displays the count of patients across four different stages of disease progression. The bars show a varied distribution, with stage 2 having the highest count, then stages 3, 1, and 4 respectively. This indicates that most patients are in the middle stages of the disease. For the status bar plots, this indicates that the count of patients across three categories represent different outcomes such as censored, liver transplantation, or death. The highest bar corresponds to 0, suggesting that the majority of patients are censored. For the treatment bar plot, this shows the distribution of patients between two treatment groups. Both groups have a lot of patients, with one group slightly larger than the other. The dataset has a fairly even distribution of patients across the two treatment types to reduce any biases. The unit bar plot represents the count of patients across different units, which corresponds to different hospitals or study centers. The counts ranges differently with one unit having the highest number of patients and the others having fewer. Majority of a treatment location may lead to biases.

Figure 5. (histograms) These histogram showcase key variables, with days displaying right skewness, suggesting shorter follow-up times are common, and weight. Age appears roughly normal but with notable peaks. Alb levels also center around a norm, while alkph and asptr show right skewness, indicating less frequent high values. Bili emphasizing a subset with high levels. Crea levels are predominantly lower, with right skewness.

Discussion

The survival analysis, enhanced by visual aids such as Kaplan-Meier curves, helped illuminate the differences in survival probabilities among patient groups and emphasized the importance of considering censoring in survival estimates.

Pros include thorough utilization of the dataset's details and clear visual aids which helps model robustness and ensured reliable predictive insights through validation techniques. The analysis also addressed a broad range of covariates, showcasing the dataset's richness.
On the other hand, cons involve the multi-center study design potentially leading to unmodeled within the hospital correlations, oversimplified assumptions of linear covariate relationships, and skewness in clinical variables that could skew hazard ratio accuracy. Linear assumptions in the Cox model may not capture all the complexities of the data.

The distributions shows signs of early event occurrences on these models for the data's complexity. The analysis does face some limitations and the possibility of overlooking factors that may have a better overview of the data.

Appendix

```
library(tidyr)
library(dplyr)
library(survival)
library(ggplot2)
library(survminer)

data <- Pbc3

#checking head(data) to see if any NA is present
head(data)
unique(data)

#number of missing values for each column
missing_values_summary <- sapply(data, function(x) sum(is.na(x)))

#remove rows with NA values
data_cleaned <- na.omit(data)

#check head(data_cleaned) to see if NA still persist / check unique(data_cleaned) for unseen
values
head(data_cleaned)
unique(data_cleaned)

#find how many patient records are there
num_records <- nrow(data_cleaned)

#finding mean age of patients
mean_age <- mean(data_cleaned$age, na.rm = TRUE)

#distinct values for status
status_ccounts <- table(data_cleaned$status)

#optional clinical measurements
clinical_measurements_summary <- summary(data_cleaned[, c("crea", "alb", "bili", "alkph",
"asptr")])

num_records
mean_age
status_ccounts
clinical_measurements_summary

#status: status at exit (0: censored, 1: liver transplantation, 2 : dead)
```

```
#data_cleaned$status <- ifelse(data_cleaned$status == 2, 1, 0)
#unique(data_cleaned$status)

#status variable is 1 for event or 0 for censor
#surv_object <- Surv(time = data_cleaned$days, event = data_cleaned$status)
#estimate the survival function w/o covariates
#km_fit <- survfit(surv_object ~ 1)
#plot(km_fit, main = "Kaplan-Meier Survival Estimate", xlab = "Days", ylab = "Survival
Probability")
#add a legend for easier read
#legend("topright", legend=c("Survival Estimate"), col=c("black"), lty=1:1, cex=0.8)
#treatment (tment), age, sex, and relevant clinical measures

#separate survival objects for each status
surv_object_censored <- Surv(time = data_cleaned$days[data_cleaned$status == 0])
surv_object_transplant <- Surv(time = data_cleaned$days[data_cleaned$status == 1])
surv_object_dead <- Surv(time = data_cleaned$days[data_cleaned$status == 2])

#K-M survival curves
km_fit_censored <- survfit(surv_object_censored ~ 1)
km_fit_transplant <- survfit(surv_object_transplant ~ 1)
km_fit_dead <- survfit(surv_object_dead ~ 1)

#plot and legend for survival curve
plot(km_fit_censored, col = "blue", lty = 1, main = "Kaplan-Meier Survival Estimate", xlab =
"Days", ylab = "Survival Probability")
lines(km_fit_transplant, col = "red", lty = 2)
lines(km_fit_dead, col = "green", lty = 3)
legend("topright", legend = c("Censored", "Liver Transplantation", "Dead"), col = c("blue", "red",
"green"), lty = c(1, 2, 3), cex = 0.8)

cox_model <- coxph(surv_object ~ tment + age + sex + crea + alb + bili + alkph + asptr +
weight,
                   data = data_cleaned)
summary(cox_model)

#proportional hazards assumption to test for each covariate and plots
cox.zph_model <- cox.zph(cox_model)
plot(cox.zph_model)

#extract coef, SE, CI
coef_df <- summary(cox_model)$coefficients
coef_df <- as.data.frame(coef_df)
coef_df$Variable <- row.names(coef_df)
```

```
str(coef_df)
#renaming cause giving error problems with geom_errorbar
coef_df$stderror <- coef_df$`se(coef)`


#cox model coefficients
ggplot(coef_df, aes(x = Variable, y = coef)) +
  geom_point() +
  geom_errorbar(aes(ymin = coef - 1.96 * stderror, ymax = coef + 1.96 * stderror), width = 0.2) +
  theme_minimal() +
  coord_flip() +
  xlab("Variable") +
  ylab("Coefficient Estimate")

# split data into training and testing sets
set.seed(123)
training_indices <- sample(1:nrow(data_cleaned), 0.7 * nrow(data_cleaned))
training_data <- data_cleaned[training_indices, ]
testing_data <- data_cleaned[-training_indices, ]

#1 to 0 and 2 to 1 in both training and testing data because dead is basically 0. Main focus is
liver transplantation
training_data$status <- ifelse(training_data$status == 2, 1, 0)
testing_data$status <- ifelse(testing_data$status == 2, 1, 0)

#fitting cox model on the training data
cox_model_train <- coxph(Surv(time = days, event = status) ~ tment + age + sex + crea + alb +
bili + alkph + asptr + weight,
                         data = training_data)

#testing data prediction
surv_pred <- predict(cox_model_train, newdata = testing_data, type = 'risk')

#predicted risks into risk groups
risk_thresholds <- quantile(surv_pred, probs = c(0.33, 0.66))
testing_data$risk_group <- cut(surv_pred, breaks = c(-Inf, risk_thresholds, Inf), labels = c("Low",
"Medium", "High"))

#K-M survival curves for risk groups
km_fit_risk_groups <- survfit(Surv(time = days, event = status) ~ risk_group, data =
testing_data)

#plot of the survival curves
ggsurvplot(km_fit_risk_groups, data = testing_data, pval = TRUE)
```

```
#univariate analysis for continuous variables
continuous_vars <- c("age", "crea", "alb", "bili", "alkph", "asptr", "weight", "days")
for (var in continuous_vars) {
  continuousplot <- ggplot(data_cleaned, aes(x = .data[[var]])) +
    geom_histogram(bins = 30, fill = "blue", color = "black") +
    theme_minimal() +
    xlab(var) +
    ylab("Frequency") +
    ggtitle(paste("Histogram of", var))
  ggsave(paste0("Histogram_of_", var, ".png"), plot = continuousplot)
}


#univariate analysis for categorical variables
categorical_vars <- c("unit", "tment", "sex", "stage", "gibleed", "status")
for (var in categorical_vars) {
  categorical_plot <- ggplot(data_cleaned, aes(x = .data[[var]])) +
    geom_bar(fill = "orange", color = "black") +
    theme_minimal() +
    xlab(var) +
    ylab("Count") +
    ggtitle(paste("Frequency of", var))
  ggsave(paste0("Bar_Plot_of_", var, ".png"), plot = categorical_plot)
}
```