REGULAR ARTICLE

Predicting Survival Rates in Primary Biliary Cholangitis Patients

Syed Shah*

Correspondence:
shah24s2@ncssm.edu
North Carolina School of Science and Mathematics, 1219 Broad St., 27705 Durham, NC
Full list of author information is available at the end of the article
*NCSSM Online Program

Abstract

Primary Biliary Cholangitis (PBC), a chronic autoimmune liver disease, presents a significant challenge in predicting patient survival due to its complex etiology and variable progression. This research paper uses a clinical dataset encompassing a wide array of clinical variables to analyze survival rates among PBC patients. Using a comprehensive array of statistical tools, including logistic regression, clustering, and principal component analysis (PCA), this study aimed to find the significant predictors of mortality. Key clinical variables such as bilirubin, copper, albumin, ascites, and hepatomegaly were looked at specifically. The findings revealed that elevated bilirubin and the presence of ascites are the most significant predictors of mortality, with gender differences also playing a notable role. The clustering analysis stratified patients into distinct groups, reflecting the disease's heterogeneity and enhancing our understanding of PBC. The results can potentially guide future clinical assessments and support healthcare providers in prioritizing care based on individual risk profiles.

Keywords: Cirrhosis; Primary biliary cholangitis; Primary biliary cirrhosis; Cluster analysis; Principal component analysis; Logistic regression

1 Introduction

The liver, an organ pivotal to various metabolic processes, detoxification, and nutrient storage, is vulnerable to a range of diseases impacting its structure and function. Among these, Primary Biliary Cholangitis (PBC), previously known as Primary Biliary Cirrhosis, is a notable chronic autoimmune liver disease (Lindor et al., 2008 [1]). Characterized by the progressive destruction of small bile ducts leading to cirrhosis, PBC primarily affects middle-aged women and displays varying global incidence and prevalence rates (Yoo et al., 2018 [2]).

The etiology of PBC is complex, believed to stem from both genetic and environmental factors (Kowdley & Kaplan, 2012 [3]; Hirschfield & Gershwin, 2013 [4]). Despite its rarity, understanding of PBC has evolved, with epidemiological studies highlighting significant geographic variability in prevalence (Zhou et al., 2020 [5]). For instance, the Marshfield Epidemiologic Study Area (MESA) reported an incidence of 4.9 cases per 100,000 person-years in the United States, with a higher incidence observed in females (Zhou et al., 2020 [5]).

PBC's defining characteristics, including chronic cholestasis, highlight its complexity (Ballestri et al., 2020). Chronic cholestasis results from impaired bile flow due to the destruction of bile ducts, leading to bile accumulation in the liver and contributing to cirrhosis development (Cleveland Clinic, n.d. [6]).

Current treatments for PBC, such as Ursodeoxycholic Acid (UCDA), focus on slowing disease progression and reducing liver transplant needs, guided by prognostic models like the GLOBE and UK-PBC scores (Lindor et al., 2008 [1]).

Dyslipidemia, particularly hypercholesterolemia, occurs in most PBC patients, presenting both a unique aspect of the disease and implications for cardiovascular risk due to PBC's effects on serum lipids (Yoo et al., 2019 [7]). Additionally, alterations in coagulation factors such as prothrombin and platelet counts offer insights into hepatic synthetic function and the state of liver disease, contributing to prognosis and management strategies (Prince et al., 2004 [8]).

Key indicators for assessing PBC severity include alkaline phosphatase and SGOT (aspartate aminotransferase), which signal cholestasis, liver damage, and cellular injury (Cleveland Clinic, n.d.; Hepatitis and Liver Disease, n.d.). Moreover, spider angiomas (SAs) have been proposed as potential prognostic indicators in chronic liver disease, although their exact role requires further research (Ghosh & Mittal, 2020 [9].

Progressively increasing serum bilirubin concentrations have been identified as crucial for prognostication in PBC (Carbone et al., 2015 [10]). A study from 1979 indicated a significantly reduced survival period when total bilirubin levels exceed 10 mg/dL (Dickson et al., 1989). This understanding, alongside the diverse clinical presentation of PBC and its potential overlap with autoimmune hepatitis, underscores the complexity influencing its clinical course (Hirschfield & Gershwin, 2013 [4]).

The variability in disease progression, influenced by factors such as fluctuating serum bilirubin levels and the presence of symptoms mimicking other liver diseases, poses makes it difficult to predict whether a PBC patient will survive.

To be able to predict the survival of a patient, recent studies, such as those done by Yoo et al. (2018 [2], 2019 [7]), have introduced the Neutrophil-to-Lymphocyte Ratio (NLR) as an indicator of inflammation and a predictor of prognosis in PBC patients. Elevated NLR levels have been linked to increased risks of liver transplantation or death, emphasizing inflammation's role in long-term outcomes (Yoo et al., 2019 [7]).

To build on current research, this paper aims to use data science techniques for predicting mortality in PBC patients. The content of this paper is organized as follows. In Sect. 2, we describe the dataset and data preparation. In Sect. 3, we discuss the results of various data visualizations and models. We conclude in Sect. 4 by summarizing the findings. By analyzing a dataset encompassing a broad range of clinical and non-clinical factors, the goal is to identify key predictors of mortality, thereby aiding healthcare professionals in managing this chronic autoimmune liver disease more effectively.

2 Computational Methods

2.1 Dataset and Preparation

In this section, we delve into the computational methods employed for the analysis and prediction of survival states in individuals with liver cirrhosis. We will begin by discussing the dataset and methods used to clean it. The dataset chosen was created as part of a Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver, conducted from 1974 to 1984. This study aimed to investigate cirrhosis resulting from prolonged liver damage, often associated with conditions like hepatitis or

chronic alcohol consumption. The data was funded by the Mayo Clinic and includes individuals who participated in a clinical trial, as well as those who agreed to record basic metrics and undergo survival tracking, resulting in a comprehensive dataset. It consists of 17 clinical features, one of which being survival states, the one we are trying to predict. These survival states are categorized as follows: 0 = D (Death), 1 = C (Censored), and 2 = CL (Censored due to Liver Transplantation).

Our research process began with the setup of our computational environment in R, a programming language and environment widely used in statistical analysis. We loaded essential libraries including tidyverse for data manipulation and visualization, cluster and factoextra for clustering analyses, car for advanced hypothesis testing, and ggplot2 for sophisticated data visualization.

The first step in our analysis involved the importation and cleaning of data. We utilized the read.csv function in R to read data from a CSV file. After loading the data, we began the process of data cleaning. This involved removing the 'ID' column, which was repeating row numbers, and improving the clarity of the 'Status' column by updating it with more descriptive terms. We also addressed the issue of missing data. Specifically, rows 313 to 418 were removed due to their incomplete nature, a decision crucial for maintaining the integrity of our dataset. Furthermore, the 'Age' variable was transformed from days to years and rounded to the nearest whole number for a more intuitive understanding.

The next phase was the management of missing values and the transformation of categorical variables. In dealing with missing data, we used 2 different strategies. For categorical variables, any row with missing data was removed, ensuring the consistency and reliability of our categorical analyses. For continuous variables, missing values were replaced with the mean of the respective variable. Additionally, binary categorical variables, especially those in a 'Yes/No' format, were transformed into a binary 1/0 format. This conversion simplified the representation of these variables, making them easier to use for statistical analysis.

2.2 Exploratory Data Analysis

The exploratory data analysis (EDA) phase was extensive. We started this phase by summarizing the dataset using descriptive statistics. This step provided us with an overview of the central tendencies, dispersion, and distribution shape of the dataset, offering initial insights into its characteristics. We then proceeded to create histograms for each continuous variable. These histograms were instrumental in visualizing the distribution of these variables, enabling us to detect any skewness or outliers that could impact our analysis. For categorical variables, bar graphs were employed to visualize their frequency and distribution within the dataset. These visualizations were pivotal in understanding the categorical data's structure and trends. Moreover, we constructed a correlation heatmap to explore the relationships among continuous variables. This heatmap not only highlighted potential correlations but also aided in identifying patterns and associations that warranted further investigation.

2.3 Principal Component Analysis

A critical component of our analysis was the Principal Component Analysis (PCA) conducted on continuous variables. PCA is a powerful tool in data analysis, used

for dimensionality reduction and the identification of underlying patterns in data. We visualized the results of PCA using a scatter plot and a scree plot. The scatter plot provided a visual representation of how the data points were distributed across the principal components, revealing clusters and outliers that might not be apparent in the high-dimensional space. The scree plot was instrumental in determining the number of principal components to retain, based on their contribution to the variance in the dataset. After this, we would get a list of variables ranked by their loading score, or their contribution to the first principal component, which is the most important. Since this tells us the variables that are most influential to PBC as a whole, it can also give insights into which variables are most influential to mortality as well.

2.4 Cluster Analysis

Using results from the Principal Component Analysis, we used clustering analysis. This allowed us to investigate the existence of potential subgroups within our patient cohort. Clustering helped to highlight patterns within the subgroups in relation to mortality outcomes, giving a deeper understanding of the factors that led to death in liver cirrhosis patients.

Before applying clustering analysis, it was essential to scale the continuous data. This step standardizes the range of independent variables, ensuring that the clustering algorithm treats all variables equally. To determine the optimal number of clusters for k-means clustering, we used various methods including the Elbow Method, Silhouette Method, and Gap Statistic Method. The Elbow Method involved plotting the within-cluster sum of squares against the number of clusters, looking for an 'elbow' point where the rate of decrease sharply changes. The Silhouette Method calculated the average silhouette score for different numbers of clusters, assessing the quality of the clustering. The Gap Statistic Method compared the total within-cluster variation for different numbers of clusters against expected values under a null reference distribution, providing an objective criterion to determine the number of clusters. After identifying the optimal number of clusters, we visualized them using a scatter plot, which depicted the grouping of data points.

2.5 Stacked Bar Plots for Categorical Variables

In the next stage of our analysis, we turned our attention to the influence of categorical variables on liver cirrhosis outcomes. Variables such as Ascites, Hepatomegaly, Spiders, Edema, and Gender were scrutinized. We created stacked bar plots for these categorical variables, which allowed us to observe their distribution and impact concerning liver cirrhosis outcomes.

2.6 Logistic Regression

Building on the insights taken from our data visualizations, we finalized our analysis by conducting a logistic regression, incorporating all the variables identified so far. This regression model served as the final method in the analytical process. With this statistical method, we aimed to quantify the strength of the associations between each categorical variable and liver cirrhosis outcomes.

Overall, our approach was designed to provide a comprehensive understanding of the factors influencing liver cirrhosis outcomes, using the power of statistical and data visualization techniques to uncover significant patterns within the data.

3 Results

3.1 Categorical Data Spread

Beginning with an analysis of the categorical variable spreads within a dataset on primary biliary cirrhosis (PBC) liver disease as seen in Fig. 1, we note that there are more Censored cases than Death cases, pointing to a higher number of patients alive or lost to follow-up, and a minority that has received liver transplants. In the drug distribution, D-penicillamine and Placebo are used equally across the study, showing a controlled clinical trial setting. The sex distribution within the dataset heavily favors women, aligning with the recognized higher incidence of PBC in women. Ascites, a serious complication of liver disease, is absent in the majority of patients, suggesting that many are not in the late stages of the disease. Hepatomegaly, or liver enlargement, is split roughly evenly across the dataset, indicating varied disease progression among patients. Fewer patients exhibit 'Spider nevi', a skin condition linked to liver disease, compared to those who do not, and similarly, edema is less frequently observed. The stage distribution skews towards more advanced stages, with Stage 1 being the least represented, which suggests that the condition is often diagnosed or included in the dataset at a later stage in PBC's progression.

3.2 Continuous Data Spread

After looking at the spread of the continuous variables through histograms in Fig. 2, we saw several trends. The age distribution is roughly normal with a central concentration around the 50s, suggesting a middle-aged patient population. Bilirubin levels are right-skewed, indicating most patients have low to normal levels, with a few showing significantly elevated levels, a typical sign of liver dysfunction. Cholesterol levels also exhibit a right skew, with most patients on the lower end of the spectrum, potentially reflecting liver-related impairments in cholesterol metabolism. Albumin, a liver-synthesized protein, is left-skewed in its distribution, highlighting that while many patients have mid-range levels, high albumin levels are less common, aligning with liver disease's impact on protein synthesis. SGOT, an enzyme indicating liver damage when elevated, shows a right-skewed distribution, where elevated levels are present in a notable subset of patients. Lastly, the prothrombin time is somewhat normally distributed but with a tendency towards higher values in some patients, suggesting varied impacts of liver disease on blood clotting factors. These graphical analyses provide insight into the clinical characteristics and potential complications faced by patients within the PBC dataset.

3.3 Correlation Matrix Analysis

Expanding our analysis from the distribution of individual variables to the interrelations within the PBC liver disease dataset, the correlation matrix in Fig. 3 shows a detailed map. The gradations of color transition from blue to red, each color marking the intensity of correlations — blue for negative and red for positive. The prominent blue at the intersection of 'Albumin' and 'Bilirubin' signifies a strong inverse relationship, a commonality in liver disease where impaired hepatic function causes bilirubin levels to rise as albumin synthesis declines. This is crucial as it may hint at the fact that both variables are important for predicting mortality in PBC.

Next to this, the correlation between SGOT, an enzyme indicative of hepatocellular injury, and 'Alk Phos', associated with cholestasis, is illuminated in red, pointing

to their concurrent elevation in hepatobiliary diseases. Similarly, a positive correlation between 'Cholesterol' and 'Triglycerides' may reflect the liver's compromised ability to metabolize lipids properly. Another area of note is the pairing of 'Copper' and 'Bilirubin', their proximity in the dendrogram suggesting a shared pathway of accumulation in conditions like Wilson's disease or cholestatic liver disorders.

Moreover, the heatmap's dendrogram, with its branches, categorizes variables not just by their strength but also by the pattern of their correlations, offering a hierarchical perspective on variable associations. For example, the clustering of 'Platelets', 'Age', and 'Prothrombin' shows the age-related progression of liver disease, impacting coagulation and platelet count.

3.4 Principal Components Scatterplot Analysis

We now shift to examining the PCA. The scatter plot in Fig. 4 delineates the data points in relation to the first two principal components derived from the PCA of the PBC dataset, with PC1 and PC2 plotted on the x-axis and y-axis respectively. The observed dense central clustering of the data points signifies that while these components capture the most significant variance, they do not reveal distinct subgroups or outliers within this variance space. The greater spread along PC1 as compared to PC2 indicates its dominance in explaining the dataset's variance, likely encapsulating key clinical features or biomarkers of PBC.

3.5 Principal Component 1 Analysis

We now look into the specific variables that are the most important for PBC in general. The output in Fig. 5 ranks the variables based on their loadings on the first principal component. PC1 is the most informative in PCA, as it captures the largest variance in the dataset. The loading values represent the weight or contribution of each variable to PC1, offering insights into their relative importance. Therefore, the output shows the most important variables that cause PBC in general.

At the top of the list, Bilirubin and Copper are the most influential variables on PC1, suggesting they are major factors in the variation of data within the PBC dataset. High levels of both bilirubin and copper are associated with liver dysfunction, which is consistent with their prominence in a dataset focused on liver disease.

Following these, 'N Days', representing the number of days since diagnosis or treatment, and Albumin, a key protein produced by the liver, are also significant. Their ranking indicates a strong association with the clinical course of PBC.

SGOT (a liver enzyme), Triglycerides, and Prothrombin (related to blood clotting) are next, implying that liver damage and altered lipid metabolism are important dimensions in the dataset. Cholesterol also ranks highly, further underscoring the impact of liver disease on lipid regulation.

The presence of 'Alk Phos' (alkaline phosphatase, another liver enzyme), Platelets (which can be affected by liver disease), and Age at the lower end of the ranked variables still indicates their relevance, but with less influence on the variability captured by PC1 compared to the others.

Overall, the ranking demonstrates the multifaceted nature of PBC. It suggests that a combination of liver function tests, measures of metabolic function, and patient demographics like age and treatment duration are integral to understanding

the disease's progression and outcomes. Now, using the 3 most important clinical factors from the PCA (bilirubin, copper, and albumin), we will move forward in analyzing the data.

3.6 Influential Continuous Variables Scatterplot Analaysis

In our analysis of the primary biliary cirrhosis (PBC) dataset, the scatter plot in Fig. 6 depicting bilirubin, copper, and albumin variables offers significant insights into the biochemical factors associated with patient outcomes. Bilirubin and copper, plotted on the x-axis and y-axis respectively, are well-established markers of hepatic dysfunction, with elevated levels often signaling severe liver pathology. The size gradation of the points representing albumin levels further elucidates the clinical picture; larger points indicating higher albumin levels are predominantly clustered in regions of lower bilirubin and copper levels, which correlates with a more favorable prognosis and liver function status.

The data points corresponding to patient mortality, distinguished by the blue color, are notably concentrated in the zone of higher bilirubin and copper levels, coupled with smaller points indicative of lower albumin levels. This aggregation suggests a strong association between increased mortality and compromised liver function, as reflected by these particular biochemical parameters. The interplay of high bilirubin and copper alongside low albumin has been documented in clinical literature as a triad indicative of advanced liver disease, which can culminate in patient demise.

Patients who have undergone liver transplantation, categorized as 'Censored (transplant)' and represented by green points, exhibit a broad spectrum of bilirubin levels but generally high copper levels, showing how liver transplants occur more for those with raised copper levels.

3.7 Cluster Analysis

To further investigate the causative factors of mortality in primary biliary cirrhosis (PBC), we turned to cluster analysis to unravel the complex interactions between patient characteristics and outcomes. To determine the optimal number of clusters within our PBC dataset, we applied three different evaluative methods as seen in Fig. 7. The Silhouette Method, which assesses the fit of data within a cluster compared to other clusters, suggested an optimal cluster count of two. This method focuses on the cohesion and separation of the clusters, proposing that two clusters maximize the average silhouette width. In contrast, the Gap Statistic Method, which compares the total within-cluster variation with expected variation under a null reference distribution, indicated a larger number of clusters, specifically eight, potentially reflecting a more granular subdivision within the patient cohort.

The Elbow Method, which observes the rate of decrease in the total within-cluster sum of squares, pointed to four clusters as the elbow point where additional clusters do not significantly improve the tightness of the clustering. Given that the Elbow Method offers a balance between the extremes suggested by the other two methods, we decided on four as the optimal number of clusters. This decision is a compromise that acknowledges the variability within the data while still allowing for meaningful separation into subgroups. By proceeding with four clusters, we

aim to capture significant distinctions in patient characteristics that correlate with mortality outcomes.

With the number of clusters decided, we move forward. As seen in Fig. 8, we visualized the clustering analysis based on scaled values of Bilirubin, Copper, and Albumin. The k-means algorithm was set to identify four distinct clusters within the data, as shown by the different colors.

The clustering suggests that there are distinct groups within the patient population with different biochemical profiles. The red cluster primarily contains patients with lower levels of both Bilirubin and Copper, and this group shows a range of Albumin levels. This cluster likely represents patients with milder liver conditions or those in an earlier stage of liver disease.

The green cluster, with moderate levels of Bilirubin and Copper and generally smaller-sized points indicating lower Albumin levels, could correspond to a group with more advanced liver dysfunction but not the most severe among the cohort.

The blue and purple clusters, characterized by higher Bilirubin levels, potentially signify patient groups with more severe liver conditions. Notably, the purple cluster shows some patients with exceptionally high Copper levels, which are visually distinct from the rest of the data points. Within these clusters, the variation in point sizes suggests a mix of Albumin levels, with the purple cluster featuring some individuals with higher Albumin levels.

The clustering and the size of the points indicate that as Bilirubin and Copper levels increase, there is a tendency toward higher Albumin levels, especially in the blue and purple clusters. However, the relationship is not strictly linear, and there is considerable variability within each cluster, reflecting different underlying conditions or stages of disease progression.

3.8 Categorical Variables Stacked Bar Plots Analysis

Now, let's focus on analyzing the categorical variables. The set of stacked bar plots in Fig. 9 presents a comparative overview of patient outcomes based on various clinical factors. Ascites show a pronounced association with mortality; patients with ascites have a higher incidence of death compared to those without. This suggests that ascites is a significant risk factor for mortality in this patient group.

For hepatomegaly, the data indicates a more balanced distribution of outcomes among patients with and without this condition. However, there's a slight increase in mortality for those with hepatomegaly, hinting at a possible link to adverse outcomes.

When looking at the presence of spiders, there's an observable increase in deaths among affected patients. This could suggest that spiders are a marker for more severe disease or a contributing factor to increased mortality risk, albeit with less impact than ascites.

Edema's relationship with mortality is less clear from these plots. The similar death counts for patients with and without edema imply that it may not be as strong a predictor of mortality compared to other factors like ascites or spiders.

Lastly, the gender plot reveals a difference in outcomes, with the male gender experiencing fewer transplants and potentially higher mortality.

In summary, these visualizations underscore the potential of certain clinical signs, such as ascites and spiders, to be associated with higher mortality risk in patients. Hepatomegaly's role is less definitive but still noteworthy, while edema's impact on patient outcomes appears ambiguous. Gender differences are evident, with males having a higher death rate. These insights could be valuable for healthcare providers in prioritizing and managing patient care based on individual risk profiles.

3.9 Logistic Regression Analysis

Now, we will discuss the results of the logistic regression as seen in Fig. 10. After using the most crucial variables determined in previous methods, the logistic regression analysis presented in the output provides valuable insights into the relationship between several clinical variables and the likelihood of death. The model's coefficients reveal the direction and magnitude of the association between each predictor and the response variable when all other variables are held constant.

Firstly, the coefficient for Bilirubin is notably significant, with a p-value well below the 0.001 threshold, indicating a strong positive association with death. This suggests that higher Bilirubin levels are significantly associated with an increased likelihood of the outcome being studied, which could be the presence of liver cirrhosis or a related condition.

Sex, represented as a binary variable, also shows a significant relationship with death, as indicated by a p-value of 0.034. The positive coefficient for Sex suggests that males have a higher odds of the outcome compared to females.

Ascites shows a particularly strong association with death, with a coefficient of over 3 and a p-value of 0.0037, which is highly significant. This result underscores the clinical importance of Ascites in the context of the liver disease being studied.

Hepatomegaly's coefficient is positive and statistically significant at the 0.05 level, suggesting a moderate association with death. Although less impactful than Ascites, its significance implies it should not be overlooked in clinical assessments.

The variable for Spiders is not statistically significant at the conventional 0.05 level, with a p-value of 0.179, suggesting that it may not have a strong independent effect on death within this model.

Copper presents a borderline case; its p-value is just above 0.05, which traditionally denotes the cutoff for statistical significance. While it may not be considered a strong predictor for death in this analysis, the p-value close to the threshold suggests that it could be of interest for further study.

The model fit is assessed by the null and residual deviance values, which indicate that the model with predictors better fits the data than the null model. The AIC value is reasonably low, suggesting a good model fit when considering the trade-off between the model's complexity and its ability to fit the data.

4 Conclusion

In conclusion, we were able to use multiple analytical techniques to illuminate the factors contributing to patient outcomes. The heatmap's correlation analysis provided a hierarchical view of biomarker interplays, highlighting the inverse relationship between albumin and bilirubin. The PCA scatter plot, showing the biochemical interrelations of bilirubin, copper, and albumin, showed their collective significance

as indicators of liver function and disease severity. Together, these methodologies painted a detailed picture of the PBC landscape, driving home the importance of a multifaceted approach in assessing liver cirrhosis outcomes and tailoring patient care. The stacked bar plots revealed key clinical signs like ascites and hepatomegaly, hinting at their potential roles in mortality. Clustering analysis, informed by these insights, stratified patients into distinct groups, enhancing our understanding of the heterogeneity within PBC. Finally, with our logistic regression model, we are able to conclude that Bilirubin and Ascites are the most significant predictors of the outcome in this analysis, with Sex and Hepatomegaly also contributing significantly. Spiders do not appear to have a strong independent effect, and Copper, while not significant at the 0.05 level, requires further research.

Availability of data and materials
The data for this work was obtained from
https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

Funding
Funding for this program is provided by the North Carolina School of Science and Mathematics, the University of North Carolina General Administration, and the General Assembly for the State of North Carolina.

Acknowledgements
I would like to express my gratitude to Robert Gotwals, my data science teacher at the North Carolina School of Science and Mathematics. His expertise and insights have significantly contributed to my understanding and application of data science principles in this work.

References

1. Lindor, K.D., Gershwin, M.E., Poupon, R., Kaplan, M., Bergasa, N.V., Heathcote, E.J.: Primary biliary cirrhosis. Orphanet Journal of Rare Diseases **3**(1) (2008)
2. Yoo, J.J., Kim, S.G., Choi, J., Kim, Y.S., Lee, Y.S.: Recent research trends and updates on nonalcoholic fatty liver disease. Molecular Cellular Toxicology **14**(3), 239–252 (2018)
3. Kowdley, K.V., Kaplan, M.M.: Aasld practice guidelines: Primary biliary cirrhosis. Hepatology **55**(6), 2045–2060 (2012)
4. Hirschfield, G.M., Gershwin, M.E.: The immunobiology and pathophysiology of primary biliary cirrhosis. Annual Review of Pathology: Mechanisms of Disease **8**, 303–330 (2013)
5. Zhou, Y., Jia, J., Chen, H., Zhang, X., Cai, Q., Chen, Z., ... You, H.: The prevalence and risk factors of primary biliary cirrhosis in mainland china: A population-based study. Annals of Translational Medicine **8**(17), 1078 (2020)
6. Clinic, C.: Alkaline Phosphatase (ALP) (n.d.). https://my.clevelandclinic.org/health/diagnostics/22029-alkaline-phosphatase-alp
7. Yoo, E.R., Kim, D., Vazquez-Montesino, L.M., Esquivel, C.O., Cheung, R.: Neutrophil-to-lymphocyte ratio in evaluation of inflammation in patients with chronic liver disease. Gastroenterology Research **12**(1), 11–18 (2019)
8. Prince, M.I., Chetwynd, A., Diggle, P., Jarner, M., Metcalf, J.V., James, O.F.W.: The geographical distribution of primary biliary cirrhosis in a well-defined cohort. Hepatology **39**(4), 1084–1090 (2004)
9. Ghosh, S., Mittal, S.: Spider angioma: Number and location as potential prognostic indicators in chronic liver disease – a case report. Cureus **12**(8) (2020)
10. Carbone, M., Mells, G.F., Pells, G., Dawwas, M.F., Newton, J.L., Heneghan, M.A., ... Jones, D.E.: Sex differences in presentation, diagnosis, and progression of primary biliary cirrhosis. American Journal of Gastroenterology **110**(5), 752–760 (2015)

Figures

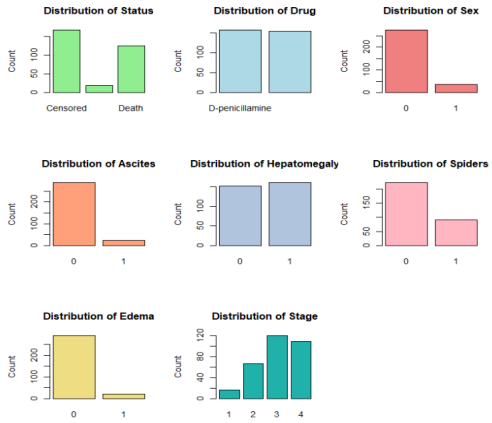

Figure 1 Bar Plots of Categorical Data in Drug Study

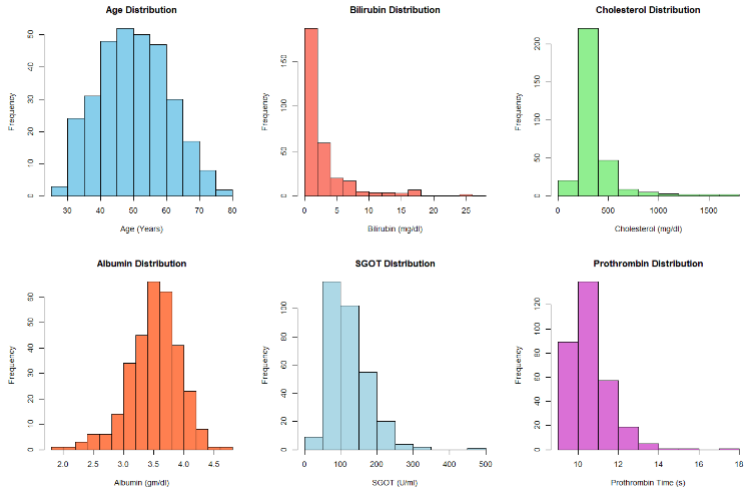

Figure 2 Histograms of Continuous Data

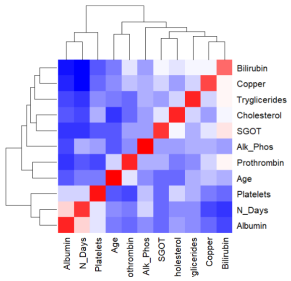

Figure 3 Heatmap of Variable Correlations

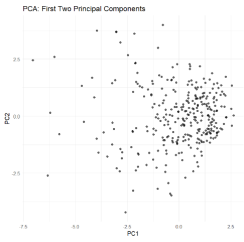

Figure 4 PCA Scatter Plot

Rank	Variable
1	"Bilirubin"
2	"Copper"
3	"N_Days"
4	"Albumin"
5	"SGOT"
6	"Tryglicerides"
7	"Prothrombin"
8	"Cholesterol"
9	"Alk_Phos"
10	"Platelets"
11	"Age"

Figure 5 Output of PCA Variable Rankings

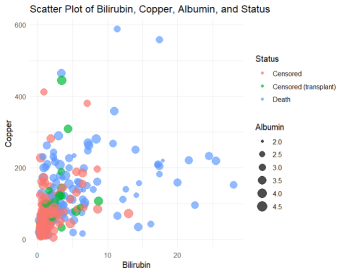

Figure 6 Scatter Plot of Key Variables

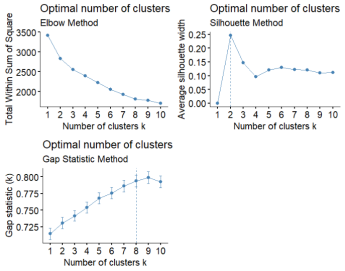

Figure 7 Comparison of Cluster Methods

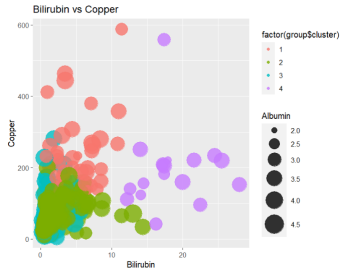

Figure 8 Cluster Plot Visualization

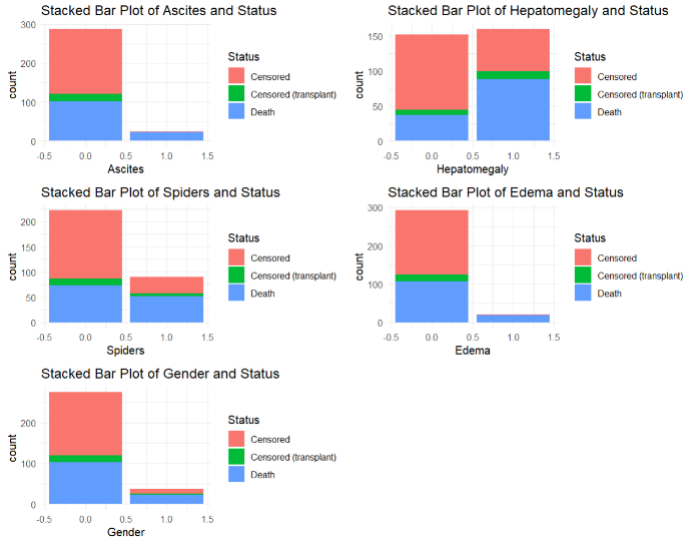

Figure 9 Stacked Bar Plots of Patient Outcomes

Coefficients:	Estimate	Std. Err	z value	Pr(>|z|)	Columr
(Intercept)	-2.241265	0.285279	-7.856	3.95E-15	***
Bilirubin	0.239097	0.065413	3.655	0.000257	***
Copper	0.00397	0.002171	1.829	0.067436	.
as.factor(Sex)1	0.919688	0.43457	2.116	0.034318	*
as.factor(Ascites)1	3.126432	1.077025	2.903	0.003698	**
as.factor(Spiders)1	0.43553	0.324132	1.344	0.179051	
as.factor(Hepatomegaly)1	0.689015	0.296609	2.323	0.020181	*

Signif. codes: 0 '***' 0.001 '**'	0.01 '*' 0	.05 '.'	0.1 ' ' 1		
Null deviance: 420.12 on 311 degrees of freedom					
Residual deviance: 306.67 on 305 degrees of freedom					
AIC: 320.67					
Number of Fisher Scoring iterations:	6				

Figure 10 Output of Logistic Regression Results