Analyzing the Relationships Between Predisposed Factors, Biological Metrics, and Stage in Cirrhosis

S&DS 363 Final Project

Group: Dinesh Bojja and Jonah Bahr
May 6th, 2023
Yale University
S&DS 363: Multivariate Statistics for the Social Sciences
Professor Jonathan Reuning-Scherer

Contents

1 Introduction

Broadly speaking, cirrhosis is a chronic liver disease that results from long-term scarring. Consequently, this scarring replaces healthy liver tissue—an organ that's integral in filtering toxins from blood, producing bile to aid in digestion, and regulating the body's metabolism—and gradually leads to loss of overall liver function over time (Schuppan and Afdhal 2008). In more elaboration, when the liver is damaged by factors such as chronic alcohol abuse, viral hepatitis, or non-alcoholic fatty liver disease, the liver tissue becomes inflamed and injured—ultimately leading to the development of cirrhosis (Hanai et al. 2021). In order to compensate for this liver damage, the organ tries to create new liver cells and form scar tissue, which progressively builds up and replaces healthy liver tissue—leading to a loss of liver function. The liver becomes smaller, harder, and less flexible, and the blood flow through the liver is restricted (Navaeu and Balian 2005). This can cause a range of complications, such as portal hypertension (high blood pressure in the veins that supply the liver), which can lead to the formation of varices (enlarged veins in the esophagus or stomach), as well as an increased risk of liver cancer (Ginès et al. 2021). Additionally, the liver's ability to produce proteins decreases, leading to a range of symptoms, such as fatigue, muscle wasting, and a tendency to develop infections.

Although cirrhosis isn't primarily known as the largest cause of mortality in the United States (compared to say, heart disease, cancer, or unintentional injuries), cirrhosis continues to be a progressively more pressing problem, as from 1999 to 2016, annual cirrhosis deaths increased by 65% (34,174 deaths total)—with the largest increases being related to alcoholic cirrhosis among individuals ages 25-34 (Xu et al. 2021; Tapper and Parikh 2018). This trend of cirrhosis being a deathly burden is also globally reflected, as cirrhosis caused 1.48 million deaths in 2019: an 8.1% increase compared to 2017 (Liu and Chen 2022). Additionally, cirrhosis shows signs of being socioeconomically correlated, as the disease's presence was commonly attributed to minorities (non-Hispanic blacks and Mexican Americans), those living below the poverty level, and those with less than a 12th-grade education; in total, 69% of cirrhosis victims in America were unaware of having a liver disease (Scaglione et. al 2015). With the dataset provided—which contains basic measurements of 424 cirrhosis-having patients from the Mayo Clinic—we set out to better understand cirrhosis and see if we can potentially aid in the effort of using medical information to curb the prevalence of the disease and promote preventive measures.

2 Planned Analysis

In this paper, we set out to examine relationships between individual predisposed characteristics, biological response factors, and cirrhosis stage in order to better learn the relationships between them. In doing so, we hope to answer the following questions:

- How useful are the dependent variables in being a medically predictive and descriptive power for the liver disease?
- How do individual traits (sex, age, etc.) affect the response variables?
- What underlying factors exist that cause differences in the response variables across individuals?

In particular, we will be using MANOVA/GLM in order to figure out which independent variables (sex, the use of cirrhosis-treating drugs, age, and a combination of some categorical variables) are significant in indicating the prevalence of cirrhosis in patients; ordination to investigate common factors that contribute to each of the stages of cirrhosis; and factor analysis to create a clearer understanding of what underlying biological or environmental factors cause changes in the response variables.

3 Data and Description of Variables

Although the dataset provides many different variables to analyze, we primarily focused on several, some of which are independent and some of which are response variables.

Table 1: Descriptions of Independent and Response Variables

Variable	Type	Description
Drug	Independent	Categorical variable that describes whether the individual takes the treatment drug (D-penicillamine) or the placebo. Individuals have only one of those two states.
Age	Independent	Continuous variable for the age of the individual, measured in days.
N_Days	Independent	Continuous variable that tracks the number of days between registration for the study and the earliest of the following outcomes: death, transplantation, and study analysis time.
Stage	Independent	Categorical variable that describes the histologic stage of cirrhosis in the individual. Stages are either 1, 2, 3, or 4. The lowest stage is 1, the highest stage is 4, with higher stages being worse disease.
Bilirubin	Response	Continuous variable that measures the serum bilirubin level in the individual (mg/dL). Bilirubin is a pigment made when red blood cells are degraded.
Cholesterol	Response	Continuous variable that measures the cholesterol level in the individual (mg/dL). Cholesterol is a lipid used in many biological contexts.
Albumin	Response	Continuous variable that measures the albumin level in the individual (mg/dL). Albumin is a globular protein used to move biomolecules.
Copper	Response	Continuous variable that measures the urine copper level in the individual (measured in ug/day). Copper is a molecule that can be toxic at high levels.
Alk_Phos	Response	Continuous variable that measures the alkaline phosphatase level in the individual (U/L). These enzymes are used often in the body to dephosphorylate compounds.
SGOT	Response	Continuous variable that measures the SGOT level in the individual (U/mL). SGOT, or aspartate transaminase, is associated with organ failure at high levels.
Triglycerides	Response	Continuous variable that measures the triglyceride level in the individual (mg/dL). Triglycerides are esters made of fatty acids.
Platelets	Response	Continuous variable that measures the platelet level in the individual (mL/1000). Platelets are important clotting factors in the body.
Prothrombin	Response	Continuous variable that measures the individual's prothrombin time (seconds). PT time measures the rate it takes for clotting.

This data was collected in a Mayo Clinic trial studying primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. 424 PBC patients were included in the trial, and many different metrics were measured for each of the individuals beyond what is listed in the table. There are potential sources of error, as the machinery used to measure the variables was not as accurate and precise when the study was conducted. Additionally, because some individuals had incomplete datasets, they were removed from this analysis, which could also lead to a change in the dataset trends (and why the analysis does not use all 424

studied individuals). However, considering that there are well over two hundred individuals analyzed, we do not expect there to be any major errors by removing those individuals.

4 Plots

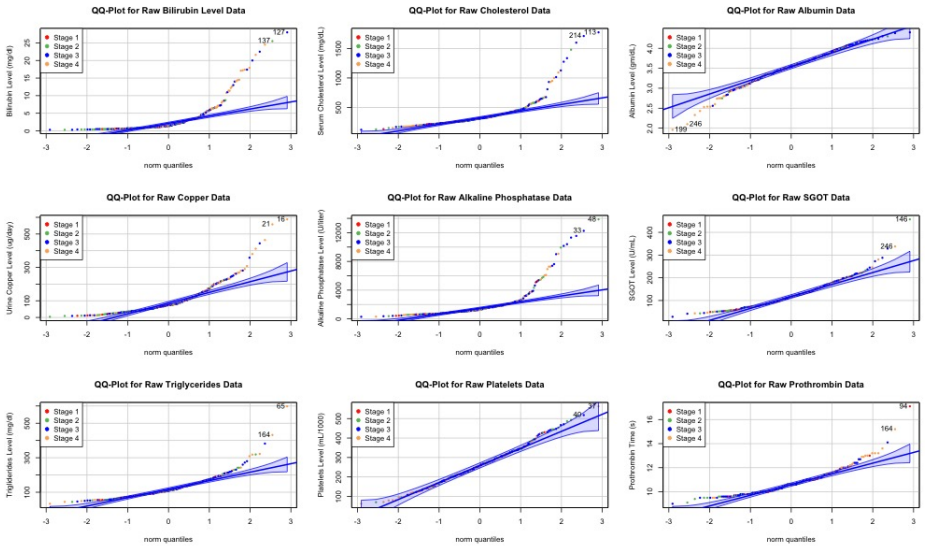

Figure 1: Raw QQ Plots for cirrhosis-related factors, separated by stage

Looking at the raw plots, it is evident that some transformations need to be done to bring the data to multivariate normality. Log regressions were done on all of the variables, which brings each of the quantile-quantile plots closer to normality for each of the variables, as depicted in Figure 2. Since these response variables all measure a concentration or measure of a biological system, it makes sense to do a logarithmic transformation, as the raw data for these types of data tend to be exponential in character.

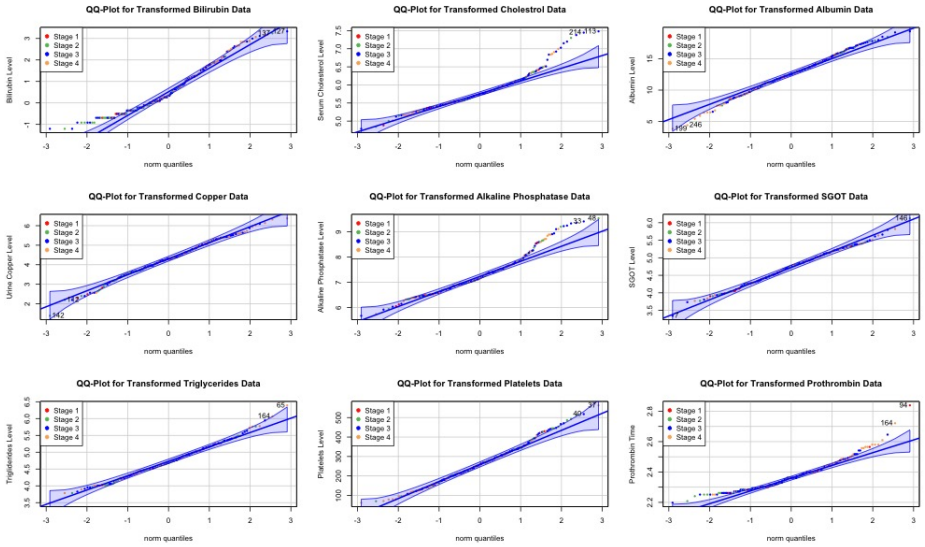

Figure 2: Transformed QQ Plots for cirrhosis-related factors, separated by stage

To confirm if the data is multivariate normal, a Chi-square plot suffices. Although there are some deviations from normality, the data tends to be approximately multivariate normal, or at least close enough to allow for an adequate analysis.

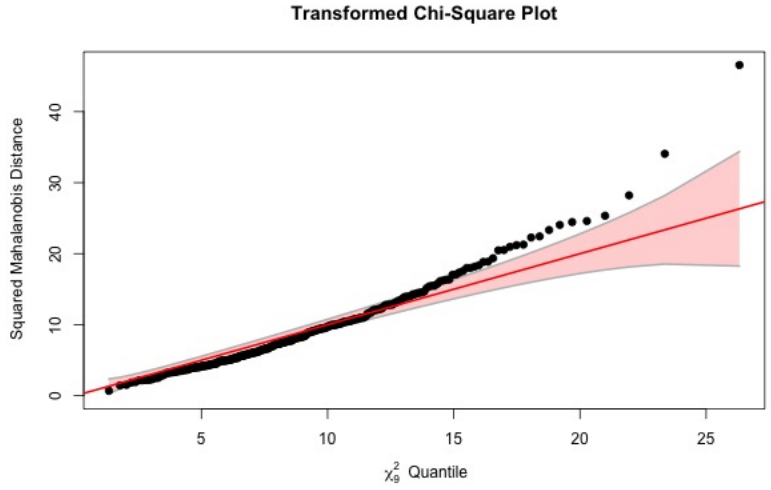

Figure 3: Chi-square plot for transformed cirrhosis data

It is worth noting that there are some points in the raw and transformed data that are outliers, both in the quantile plots and in the Chi-square plot. However, because these data are multivariate in nature, an outlier in one variable may not be an outlier in another. To preserve the original data and prevent from accidentally removing any important findings, we opted to keep all of the points, as none are so extreme that they would need to be removed.

To understand the data better, we also printed boxplots (Figure 4) for each of the response variables, separated by stage. Some of the variables, like Bilirubin and Albumin, seem to have significant differences between the different stages, while others, like Platelets, tend to be relatively more similar across the stages. Because of the sheer number of variables used in this dataset, for brevity, we excluded analysis of the summary statistics, as they did not provide any valuable information beyond what is found in the boxplots and the quantile plots.

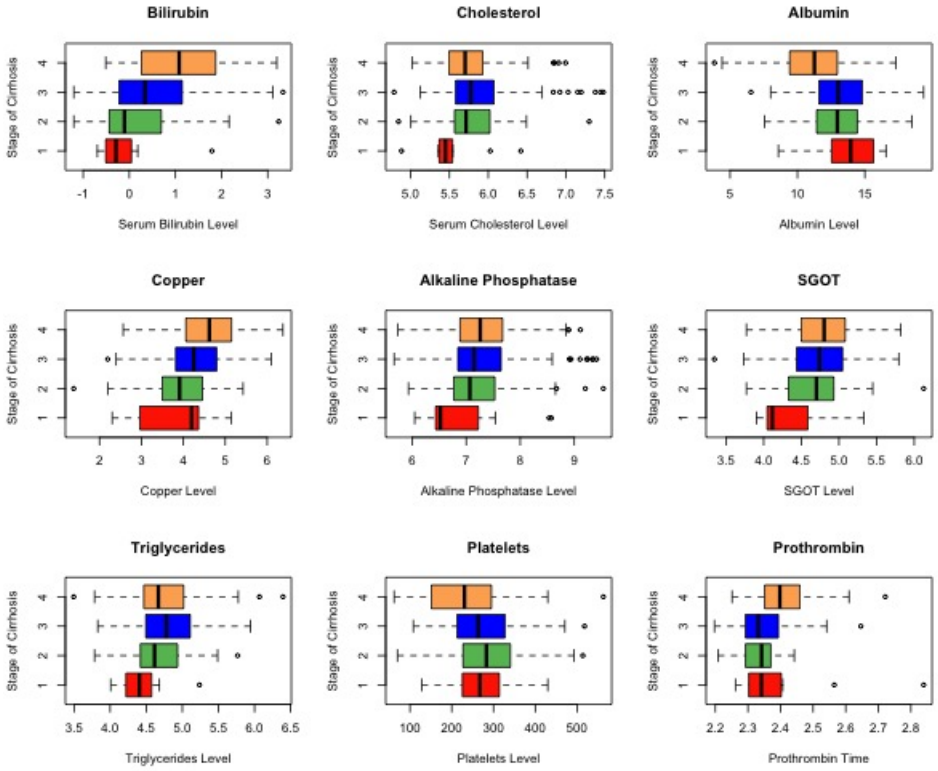

Figure 4: Panel of boxplots for each transformed response variable, separated by stage

5 MANOVA and Multivariate GLM

We begin by running MANOVA and Multivariate GLM. Through these tests, we aim to see if there are any statistically significant differences in the different biological metrics between males and females, between individuals who took a treatment drug (D-penicillamine) or a placebo, and between groups when combining these categorical variables. Considering that there is a large number of continuous variables used in this dataset, it may be worthwhile to remove such overlapping variables before running MANOVA, as it may confound any significant results from the other more differentiated variables.

After creating boxplots for each of the continuous variables stratified by their sex and whether they use the treatment drug or placebo (Figure 5), we can see that some of the groups have significant overlap, which suggests that there are no significant differences between their means for between each of the groups. In particular, the boxplots for SGOT levels and Platelet levels seem to overlap quite a bit across all of the groups, to the point where their inclusion may not contribute anything worthwhile to the analysis. Hence, when it comes time to run the MANOVA and further tests, we will opt to exclude them.

We created set of interaction plots for each of the continuous response variables used in the analysis, separating them by sex and treatment drug status (placebo or treatment) (Figure 6). Since there are only two groups for each of the categorical variables that are used to separate the groups, it is more likely than not that if there is a significant interaction between treatment drug status and sex for each continuous response variable, there will be an intersection (or at the very least opposite slopes) between the lines for the treatment drug and the placebo as they move between males and females.

Most notably, for Platelets and Prothrombin, the slopes are the same for each of the lines and they do

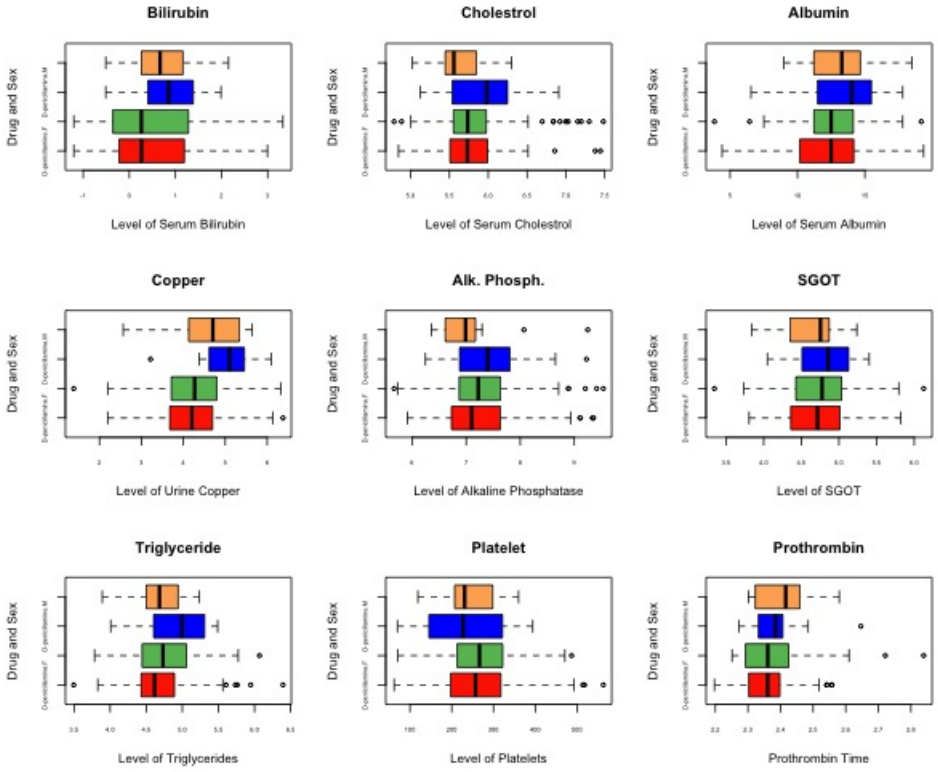

Figure 5: Panel of boxplots for cirrhosis response variables, separated by a combination of drug and sex

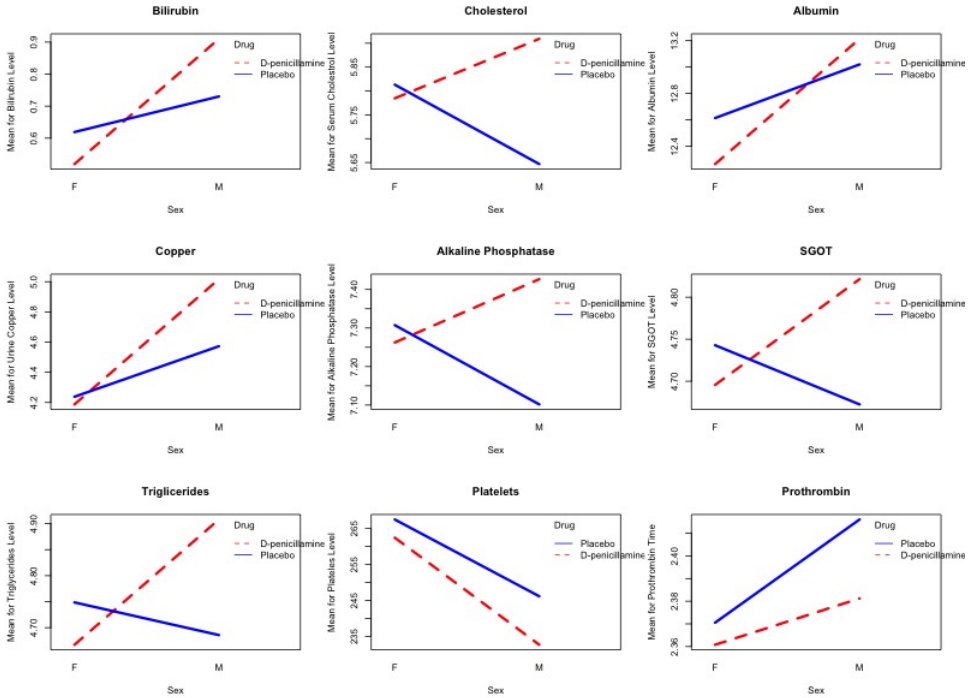

Figure 6: Panel of interaction plots for cirrhosis response variables across drug and sex

not intersect. While they are not perfectly parallel, there is likely not a statistically significant interaction between sex and treatment drug status that affects these continuous variables. Hence, in accordance with the need to reduce the number of continuous response variables, we will opt to remove both Platelets and Prothrombin from MANOVA, Multivariate GLM, and other analyses in this section, to make sure that there is no masking of statistically significant results by these variables.

After variable removal, the Box M-statistic is 0.3467, so we still fail to reject the null hypothesis that the groups are the same. Thus, we can continue with MANOVA with the reduced number of response variables.

5.1 ANOVA/MANOVA

We then checked to see the significance of each categorical variable on the univariate and multivariate level. Although there are many other measures included in the output for these tests, since the main metric that was used to determine significance was the corresponding p-value, for brevity we only included these p-values in the table. All significant p-values are bolded, and analyzed in more detail for their coefficients and interpretation in the context of the dataset.

Table 2: P-values for ANOVA and MANOVA Results

Variable	Drug	Sex	Drug:Sex
Bilirubin	0.4525	0.1189	0.4663
Cholesterol	0.6143	0.2439	0.0767
Albumin	0.3271	0.1571	0.5971
Copper	0.6234	**0.00002813**	0.09859
Alkaline Phos.	0.6272	0.3470	0.1680
Triglycerides	0.1563	**0.02625**	0.06737
Multivariate Data	0.6932	**0.00016**	0.2332

For the six response variables analyzed through ANOVA, only Copper and Triglycerides had a statistically significant difference between two groups. Both of them had a statistically significant difference between the two sexes at a significance value of $\alpha < 0.05$. For Copper, the coefficient for changing the group for Drug is 0.05088, for Sex is 0.82980, and for changing both is 0.38543, with a female on the treatment drug as the baseline. This means that while males tend to have higher levels of copper than females, this difference is no longer significant when the person's drug status changes as well. For Triglycerides, the coefficient for changing the group for Drug is 0.08141, for Sex is 0.24079, and for changing both is 0.01860, also with a female on the treatment drug as the baseline. This means that males also have a higher triglyceride level than females, a trend that is no longer significant when changing the drug status. This is supported by scientific literature, as men are shown to have higher triglyceride levels and copper levels than females in general (Bittner 2008, Buxaderas and Farré-Rovira 1986). When looking at the MANOVA results, we see that when looking at the multivariate data at once, only separating by sex has a statistically significance difference between the groups. This means that when it comes to looking at the different biological metrics, the use of a placebo or treatment drug does not have much of an effect, while the levels of these metrics varies greatly across sexes.

5.2 Contrasts

We also ran contrasts to determine if there are differences between each of the different groups by running different hypotheses to test for each. To test for a statistically significant difference between the placebo and treatment drugs, we see check if the difference between individuals who take the treatment drug and placebo

is equal to zero or not. Similarly, to test if there is a statistically difference between the sexes, we see if the difference between males and females is zero or not. After running these tests, we get the following results:

Table 3: Multivariate and Univariate Contrasts		
Variable	Drug	Sex
Bilirubin	0.8351	0.1915
Cholesterol	0.1546	0.7991
Albumin	0.8801	0.1878
Copper	0.1890	**0.0001229**
Alkaline Phos.	0.2967	0.8776
Triglycerides	0.3952	0.2826
Multivariate Data	0.5437	**0.0007617**

Interestingly, the results differ slightly compared to the results from the univariate and multivariate ANOVA tests ran before. In both, Copper and the multivariate results were significant when differentiating by sex. However, while in univariate ANOVA Triglycerides were significant, in the contrast analysis Triglycerides are **not** significant. This is likely because the contrasts are run using different settings and a different hypothesis than the MANOVA.

5.3 Multivariate GLM

We then ran a Multivariate GLM analysis on the dataset, incorporating Age as the continuous variable into the model. While we considered adding in N_Days as well, after noting that the number of days between registration and some important event (death, transplantation, or study analysis time) would have a very strong correlation with the stage or progression of cirrhosis, and thus probably the response variables, we opted to remove it to prevent it from confounding the effects of any of the other indicator variables.

Table 4: Type III Sum of Squares for Multivariate GLM				
Variable	Drug	Sex	Age	Drug:Sex
Bilirubin	0.855	0.291	0.377	0.435
Cholesterol	0.12665	0.63983	**0.00302**	0.10734
Albumin	0.9733	**0.0189**	**2.15e-05**	0.7800
Copper	0.186139	**0.000136**	0.693274	0.104800
Alkaline Phos.	0.286	0.950	0.378	0.186
Triglycerides	0.4049	0.3647	0.5728	0.0632
Multivariate Data	0.4654	**7.3076e-05**	**4.684e-06**	0.30045

Only some of the values were statistically significant under the multivariate GLM using an $\alpha < 0.05$. First, Cholesterol was statistically significant with age as a predictor. The coefficient for cholesterol was -2.121×10^{-5}, meaning that as age increased for a patient, their cholesterol level tended to decrease. This is proven scientifically, as total cholesterol levels in the body tend to decrease across sexes (Ferrara et al. 1997). Albumin was shown to be statistically significant for both sex and age, with a coefficient of -0.6051 for sex (with male as the base) and -1.888×10^{-4} for age. This means that as age increases, albumin level tends to decrease, but also that females have lower albumin levels than males. Both of these facts were shown to be true in recent scientific literature as well (Weaving et al. 2015). Finally Copper has a statistically significant relationship with sex, having a coefficient of -0.2984. This confirms the MANOVA results, that females have lower copper levels than males. In multivariate space, it seems that there is a statistically

significant difference in the response variables with respect to both sex and age. Interestingly, triglycerides is no longer significant, meaning that there is likely some anomaly with the MANOVA results that caused it to come up as falsely significant.

5.4 Analysis of Residuals

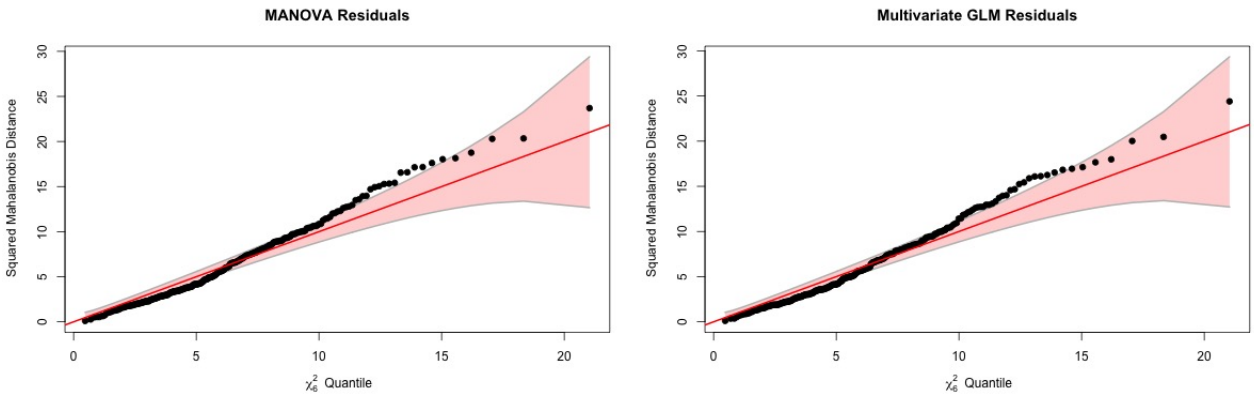

Figure 7: Chi-Square plots for residuals of MANOVA and Multivariate GLM

One of the main conditions for MANOVA and Multivariate GLM is that the residuals are multivariate normal. To show that this is the case, we printed the Chi-square residual plots for each test to see if the residuals are normal or not. While there is some deviation from multivariate normality, since the plot is close to normality on average, we determine that these tests are both applicable and usable for this dataset.

5.5 Multiple Response Permutation Procedure (MRPP) and Conclusions

Finally, we ran a Multiple Response Permutation Procedure, or MRPP, on the variables used in the MANOVA analysis (Bilirubin, Cholesterol, Albumin, Copper, Alkaline Phosphatase, and Triglycerides) to check for significance as well. The delta value for sex was 0.028, meaning that there is a statistically significant difference between males and females. However, there is not a statistically significant difference between people taking the treatment drug and the placebo because the significance of the delta is 0.435.

These conclusions support the ones made in MANOVA and Multivariate GLM as well (that the difference between sexes has a statistically significant difference on the response variables, while treatment or placebo drug status does not), which makes sense; even though the methodology is different, if the groups are truly the same or different they will still give the same result no matter what type of test is run.

6 Ordination

Ordination is a way of visualizing the data on a plot to determine the relationship between individuals and variables. We tried three different methods of ordination to draw conclusions about the data: Correspondence Analysis, Detrended Correspondence Analysis, and Non-Metric Multidimensional Scaling.

6.1 Correspondence Analysis

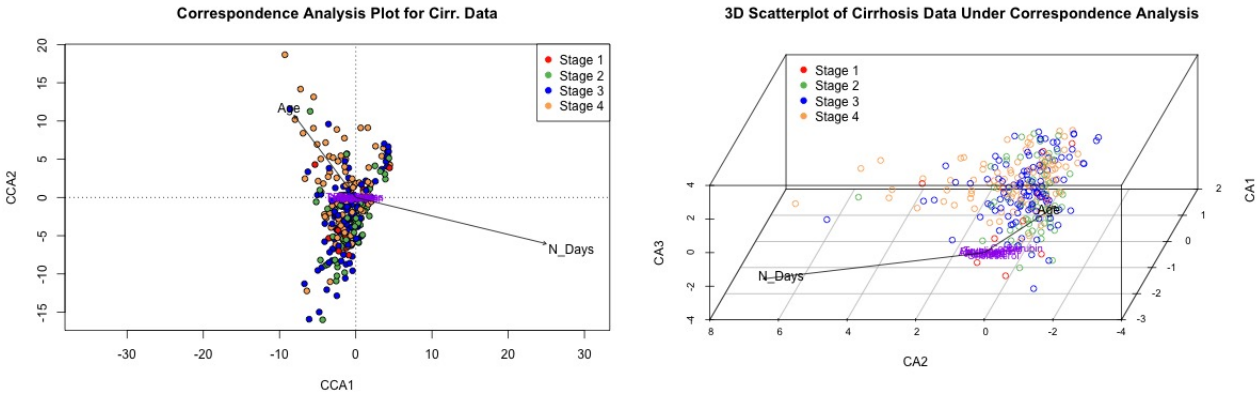

Figure 8: 2D and 3D Plot for Correspondence Analysis

6.1.1 Data Snaking

We can see that there is some snaking of the data when presented in the graphs showing the first two CA directions. The data can be traced from higher stages to lower stages by following the colors of the points plotted on each graph. Starting from the top of the CA plot, the points for Stage 4 start going lower, slowly transitioning into points that mark Stage 3. At the bottom, there is a high proportion of individuals who have Stage 2, and the points go up a little to the red points that have Stage 1 cirrhosis.

We can see this trend even more clearly in a 3D scatterplot of the data; the Stage 4 points increase in the first direction and third direction while decreasing in the second, then reach a cluster of Stage 3 points. These decrease in the first direction, while staying relatively constant in the third direction. Finally, this trajectory reaches the Stage 2 and Stage 1 points.

This pattern of snaking makes sense in the context of the dataset. The data being analyzed in this part of the dataset looks at several different factors in a person's blood and how they are affected by the stage of the individual's cirrhosis. Thus, it would logically follow that as the stage increases or decreases, different factors in the blood will be affected in different ways at different rates, which would lead to the transition between stages seen in these graphs. Additionally, we would expect the data to not look linear in even higher dimensional space, as the rate by which these different variables/factors change across stages does not stay constant, meaning an assumption of linearity is likely flawed. There would likely be movement in different CA directions that correspond to the effect of different variables, which would act differently upon each stage and individual and remove any linear trends in multidimensional space.

6.1.2 Plot Interpretation

Unfortunately, while the CA plots are very useful in understanding the snaking and relationship between the different groups, it does not seem to be very useful to understand the relationship between the different stages and the measures of blood factors. Most of the variables are clustered together closer to the regions associated with Stage 1 and Stage 2 cirrhosis, while two variables (Bilirubin and Copper) trend a little closer to the individuals with higher stages of cirrhosis.

When looking at the environmental variables, we can see that only N_Days is a significant environmental variable (p-value = 0.000999). On the other hand, Age is not a significant environmental variable (p-value = 0.061938). These trends in the p-values and significance is reflected in the length of the vectors seen on the graphs, as N_Days has a much longer vector than Age does.

What is interesting is that N_Days is in a different direction with respect to the Age vector. N_Days is in the opposite direction as the trend for the individuals with Stage 4 cirrhosis, instead pointing in the direction of individuals with a lower stage of cirrhosis. Logically speaking, individuals who have have a higher stage of cirrhosis would need to go to the hospital more often, which is why the N_Days vector points in the opposite direction of those points. Age points roughly in the direction of increasing stage of cirrhosis, which does make sense to some degree, as people who have cirrhosis would likely be older because they have a longer period of time for their liver damage to accumulate (due to alcoholism or disease), which leads to a correlation between age and stage of cirrhosis.

6.2 Inertia of Correspondence Analysis

When looking at the correspondence analysis, the total inertia comes out to be roughly 0.1786. The total inertia is equal to the sum of the sum of square eigenvalues, or the chi-squared statistic divided by the population size. In effect, inertia can be used to measure variance, or how far the data departs from the independence model.

Table 5: Variance Explained by Correspondence Analysis Directions

	CA1	CA2	CA3	CA4
Eigenvalue	0.1099	0.02907	0.02043	0.009552
Prop. Explained	0.6157	0.16279	0.11444	0.053498
Cumulative Prop.	0.6157	0.77847	0.89291	0.946406

We hope to see that the first few CA ordination directions explain most of the inertia. We find that CA1 explains a proportion of 0.6157 of the scaled chi-square (which is directly related to inertia), and CA2 explains a proportion of 0.16279. Together, they explain a proportion of 0.77847 of the inertia, which is already explaining a relatively high proportion of the variance; adding in CA3 (which has a proportion explained of 0.11444) brings the total inertia explained up to 0.89291. This means that most of the variance can be explained by the first two or three CA directions, proving the significance of these directions.

6.3 Detrended Correspondence Analysis

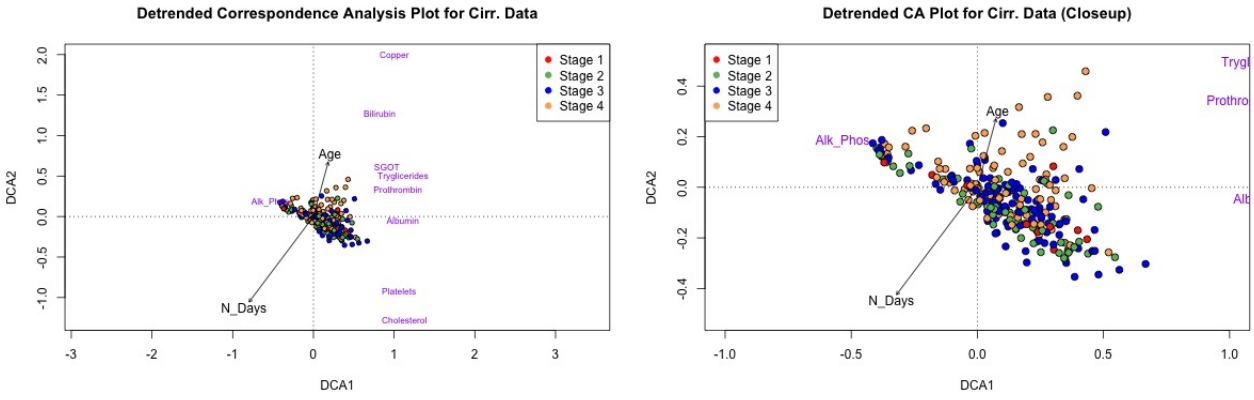

Figure 9: 2D and 3D Plot for Detrended Correspondence Analysis

The relationships between the variables and the individuals are much clearer to see in the DCA, where we can look at the direction of the different variables to understand their relationship with the different species.

DCA2 helps differentiate between most of the different variables, which allows for better understanding of their relationship with the different groups (Alkaline Phosphatase, on the other hand, is differentiated by DCA1). We can see that in both the complete dataset and the subset, Copper, Bilirubin, SGOT, Triglycerides, and Prothrombin all are in the direction marked by the presence of individuals in Stage 4 of cirrhosis; this means that these variables are likely correlated (in decreasing strength, based on how high they are on the DCA2 axis) with end-stage cirrhosis. Albumin, Cholesterol, and Platelets are more likely correlated with Stage 3 of cirrhosis, while Alkaline Phosphatase is likely closely correlated with Stage 1/2 of cirrhosis.

Looking at the DCA plot, we find that since both p-values are below the threshold of 0.05 (0.000999 and 0.047952, respectively), N_Days and Age both have a statistically significant correlation with the data in the two-way table. Since the p-value is lower for N_Days, it is more significant than Age and has a stronger effect on the data, which is also reflected in the lengths of the vectors.

For the same reason as in CA, the N_Days and Age vectors are in opposite directions on the DCA; older individuals, who tend to have higher stages of cirrhosis, have to go to the hospital or may have an important event marked in N_Days due to their cirrhosis, which leads to the vectors being in opposite directions.

6.4 Stress

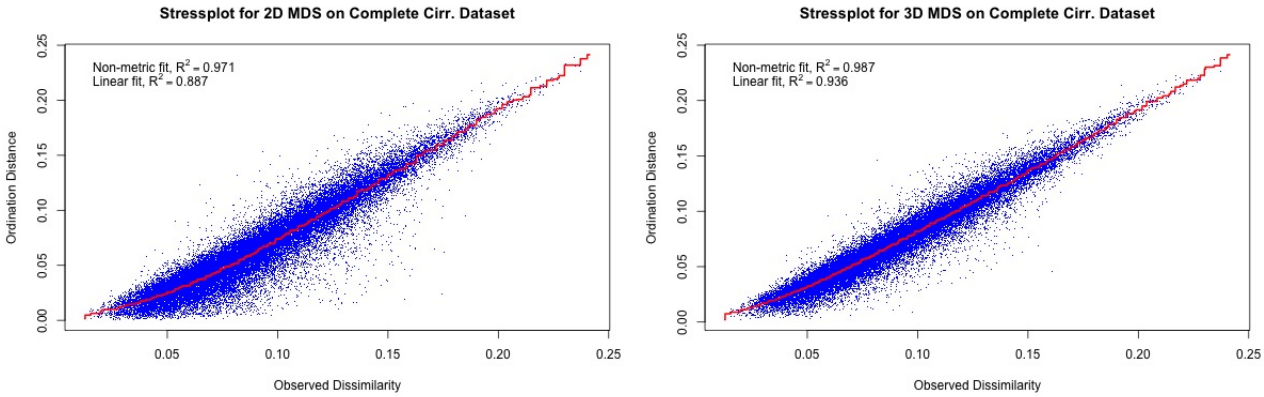

Figure 10: Stressplots for 2D and 3D NMDS analyses

We can also plot stressplots to see how strong the MDS solutions are. All of the solutions are, based on the R-squared values, very strong, when looking at a Non-metric fit. However, it can be said without a doubt that adding a third dimension to the MDS solution increases the strength of the model and making it more accurate and representative of the distances between points in the original dataset. Given that these increases in strength are so minute, a two-dimensional plot would suffice for the sake of understanding and interpreting the data, and for the purpose of distilling the multidimensional data into fewer dimensions.

6.5 Non-Metric Multidimensional Scaling

After plotting the two-dimensional MDS results for the complete dataset, we can see some general trends about how the different variables relate to the different stages of cirrhosis. More individuals with Stage 3 and 4 cirrhosis are near the outskirts of the data (top, bottom, and left), showing that they have some significant deviation from the other stages of cirrhosis.

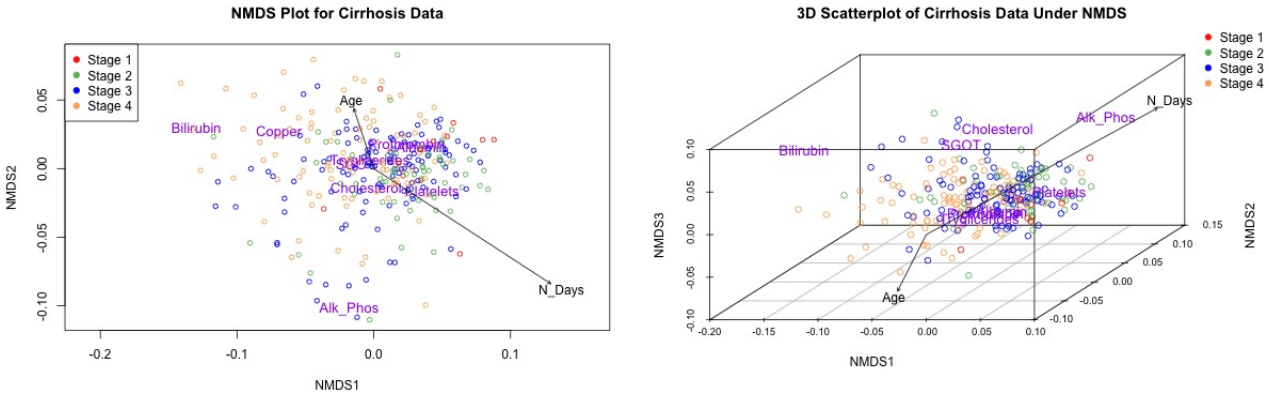

Figure 11: 2D and 3D Plot for non-metric multidimensional scaling

Changes in Bilirubin and Copper levels are more likely to be attributed to individuals with Stage 4 cirrhosis, as the variables are deeper into the regions with Stage 4 cirrhosis (and with some people with Stage 3 cirrhosis). Alkaline Phosphatase levels are more seen with select individuals who have Stage 3 cirrhosis, which is interesting considering that the people with Stage 3 cirrhosis are widely spread throughout the population. This suggests that some individuals in Stage 3 who are at the bottom of the graph have irregularly high Alkaline Phosphatase levels that causes the individuals and the variable itself to be shifted that far. The rest of the variables (Triglycerides, Prothombin, Platelets, SGOT, Albumin, and Cholesterol) are all deep in the cluster of individuals with lower stages of cirrhosis, meaning that these variables are more closely related to individuals with Stage 2 or Stage 1 cirrhosis.

The 3D NMDS plot for the complete dataset helps differentiate between the clusters much better. Here, we can see that Bilirubin is highly correlated with Stage 4 cirrhosis (Copper is as well, but it is off of the graph entirely), while the giant cluster is split across the third NMDS axis. SGOT and Cholesterol are higher up that axis, which is where fewer individuals with low-stage cirrhosis are centered. The rest of the variables are closer to the region where there are more people with Stage 1 and 2 cirrhosis instead.

When looking at the results of the environmental variable overlays on the 3D NMDS plots, we can emulate a similar analysis to what we did for the two-dimensional plots. For the complete dataset, N_Days and Age both have a statistically significant effect on the variables in the dataset as the p-values (0.000999 and 0.041958, respectively) are below the 0.05 threshold.

As in the CA and DCA, N_Days is more significant than Age, and N_Days is in the opposite direction from the individuals with a higher stage of cirrhosis, and instead points in the direction of people with a lower stage of cirrhosis. Age points generally in the direction of people with a higher level of cirrhosis, which also follows the same line of reasoning used in the analyses from before. This is interesting, as even in three dimensions, the same trends with the data are followed, suggesting that even in higher dimensions the same conclusions can be drawn from the dataset.

6.6 Wireplots and Further Analysis

We can produce wireplots to see the effect of these overlaid environmental variables on the dataset. Wireplots use regression splines to model the relationships between the environmental variables and the ordination plot by describing a surface for each one that the points can be plotted upon.

The green lines, which represent Age, are non-linear, as evidenced by the circular nature of the regression

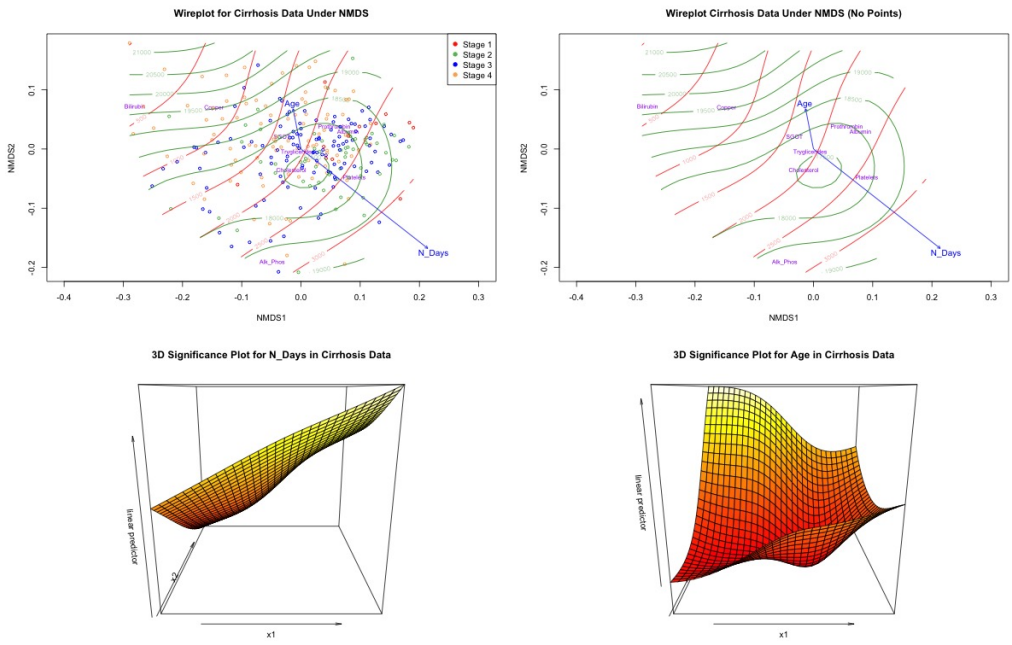

Figure 12: Wireplots for NMDS and 3D Significance Plots for environmental variables (N_Days and Age)

splines. The 3D heatmap reflects this nature, as there is a very prominent dip in the surface at the center of the graph. As the surface starts reaching the corners, the plot becomes more and more of a linear predictor, which is reflected in the more linear nature of the green lines as they trend outward from the center.

The red lines, which represent N_Days, are still non-linear, albeit less so. The 3D heatmap is much more linear than the one for age, which makes sense as the red regression splines appear a lot more linear than the green curves; however, the curvature of the red lines is reflected in the slight dip of the 3D plot.

Looking at the points on the wireplots and the stage they are in, it seems as though an increase in Age tends to lead to an increase in the stage of cirrhosis, but the correlation is not as strong as the relationship between N_Days and the stage. As N_Days increases across the regression splines, the stage of cirrhosis tends to decrease, with Stage 4 at the top left and quickly decreasing wire by wire. This is representative of how the p-values and significance of the N_Days vector was always lower and more significant than that of the Age vector.

6.7 Conclusions

Overall, we ran many different plots to try and understand the relationship between stage of cirrhosis and different blood factors, as well as the effect of environmental variables on the data itself. Since we hope to find the most effective plots to analyze the data, we will restrict our analysis to the plots used in Questions 8 and 9, which incorporate the environmental variables into the study and determine how they have a relationship with the dataset itself. Additionally, since most of the graphs (except for the CCA) have a 3D plot that goes with the standard 2D plot, we will group them together for the sake of explanation.

The CA and DCA plots allow us to see several things: First, we can tell quite effectively the snaking in the model. There is an obvious trend that moves from high-stage cirrhosis to low-stage cirrhosis that works in multiple dimensions, which can be seen in the CA and DCA plots. The main issue, however, is that even though the DCA is able to separate the variables, they both suffer from excessive clustering of the data. In

the CA, all of the variables are so clustered together that there is no way to differentiate between them, while in the DCA, the variables are so far away that it is hard to make detailed conclusions between the clustered individuals and the associated variables.

The NMDS plots allow us to understand all that the CA and DCA provide, and more. We get to see the same trends of clustering in the lower stages of cirrhosis, and trends of high variation away from that major cluster for people with higher stage cirrhosis (stages 3 and 4), which is similar to the snaking we see in the earlier plots. However, these graphs are a lot easier to read, with more spread across the different individuals and variables. This allows us to better understand and make conclusions about the trends in the data. A 2D graph is sufficient for understanding the data through an MDS model, but a 3D plot gives even more context to analyze. We can take the MDS plots further than all of the other analyses, as we can draw wireplots and understand the surfaces through which the environmental variables have an effect on the data.

When taking all of the CA, DCA, and NMDS plots together, we can make some connections between the trajectory of the environmental variable vectors and the blood factor variables. Generally speaking, N_Days and Age seem to go in opposite directions, so an increase in N_Days would correlate with a decrease in Age, and vice versa. All three of the graphs support the idea that Platelets and Cholesterol are more closely linked to higher number of days between hospital visits and lower age, as the variables tend to be more in the direction of N_Days rather than in the direction of Age. Some graphs, like the complete dataset NMDS plot, suggest that Alkaline Phosphatase levels are correlated with N_Days, but others, like the CA and DCA plots suggest that it is more of an even split between N_Days and Age in terms of which variable has more of an impact on it. Interestingly, the subset NMDS plot suggests that Age has more of an effect on Alkaline Phosphatase levels than N_Days, and instead N_Days is correlated with Albumin and Prothrombin levels (though all of the other graphs seem to disagree). What cannot be denied, though, is that Copper, Bilirubin, and Triglyceride levels are all more correlated with an increase in Age than an increase in N_Days, across the different plots.

7 Factor Analysis

7.1 Correlations and Usability of Factor Analysis

Immediately after printing the correlation matrix (Figure 13), it is clear that the variables Platelets and Prothrombin are not highly correlated with the rest of the dataset. This is because several of the paired correlations with the variable and some other factor are incredibly low, meaning that both Platelets and Prothrombin may not have any shared factors with the rest of the data. We will need to consider if we will include these factors in the final factor analysis or not; including them would allow for full analysis of the data but at the risk of less robust results, but removing them may not allow for full analysis and interpretation of the data for the benefit of a more accurate result. We can make a more educated decision based on KMO. The rest of the variables, however, seem to have some consequential (some weak, some strong) relationship with one another.

After running KMO on the complete set of indicators, we get that the KMO value is middling (with a KMO value of 0.72), while we get a similarly middling KMO value after removing Platelets and Prothrombin, albeit a little higher (0.75). Since the KMO value does not increase substantially by removing Platelets and Prothombin, instead giving a KMO value that is roughly the same, we conclude that it would be better to include these two indicators for the sake of data completeness.

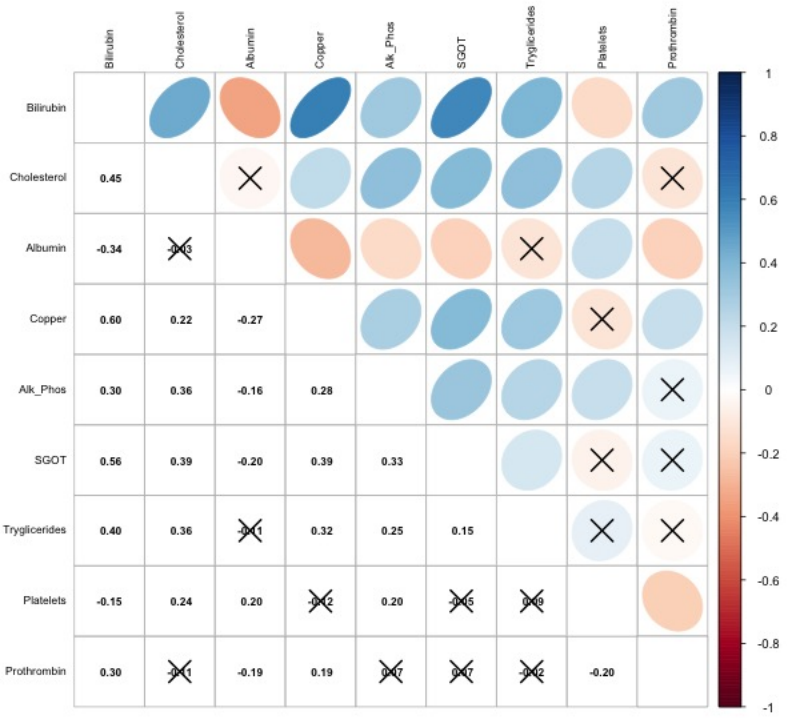

Figure 13: Correlation matrix for all response variables in cirrhosis dataset

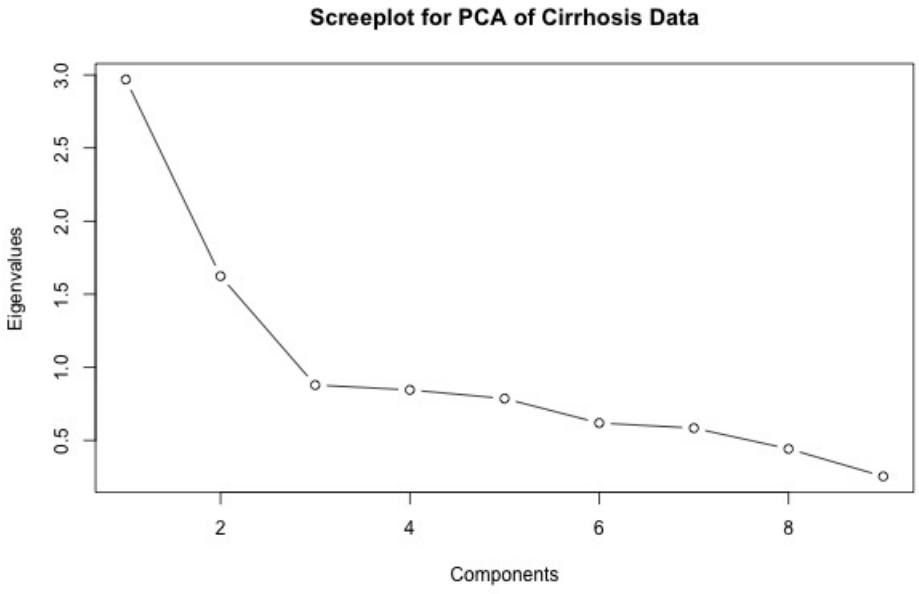

Figure 14: Screeplot for the Principal Components of cirrhosis dataset

7.2 Number of Factors

In the scree plot that shows the Eigenvalues of each component (Figure 14), we can clearly see how many of the components have a variance over 1. Only Components 1 and 2 have eigenvalues greater than 1, meaning that the Eigenvalue > 1 test suggests that we use only two factors for this analysis. Looking at the same scree plot, however, we can see that there is a very strong elbow at the third component. While in clustering analysis it would make sense to choose three clusters to correspond with the elbow, when it comes to PCA and factor analysis, we would actually cut the scree plot above the elbow at three components. This is because the elbow is the point where the variances in the data are residual variances from the random variability of data rather than variability from actual factors, meaning that we would need to cut above the elbow to deduce the number of factors. So, the scree plot suggests that we use two factors.

7.3 Extraction Methods and Residual Correlations

Four different extraction methods were analyzed using two factors, of which the best one was chosen to do rotations: PCA, PAF, Iterative PAF, and Maximum Likelihood analysis. Because the graphs are only really meaningful after rotations, for brevity, we saved the graphs for when the rotations on the best method is discussed. There were four different metrics that were used to determine which extraction method was the best: (1) RMSR, or Root Mean Square Residual, (2) Percent of residuals greater than 0.05, (3) Tucker-Lewis Index, and (4) RMSEA index, or the root mean square error of approximation. A table with all of these values was generated:

Table 6: Comparison of Extraction Methods

Extraction Method	RMSR	% Residuals	TL Index	RMSEA
PCA	0.1040	64%	0.611	0.147
PAF	0.0452	25%	0.859	0.088
Iterative PAF	0.04142555	19%	0.895	0.076
Max. Likelihood	0.04569435	25%	N/A	N/A

Note: Because of the functionality of the factanal() function used for Maximum Likelihood, it was not possible to find the Tucker-Lewis Index nor the RMSEA index for that extraction method.

Overall, we want to choose an extraction method that has: (1) a low RMSR, (2) a low percent of residuals greater than 0.05, (3) a high TL Index, and (4) a low RMSEA. Combined, these mean that the extraction method is able to use two factors to analyze the data and map the underlying factors onto the indicators and explain the variance in the data. Given this, Iterative PAF seems to be the best extraction method for this dataset. After conducting rotations, we hope to get a better interpretation of what the loadings and axes mean with regard to the data.

7.4 Rotations and Conclusions

Unfortunately, due to the limitations of the factor.pa() function in R, the only possible rotations to do are Varimax and Promax. Unlike Varimax and Quartimax (which we originally intended to do), which are both orthogonal rotations, Promax is an oblique rotation. Oblique rotations mean that the factors are correlated, while orthogonal rotations assume that the factors are uncorrelated. In some respects, it may be interesting to see if a rotation assuming the variables are correlated can provide new information about the data that cannot be seen otherwise, so we will try both rotations just for the sake of deeper understanding.

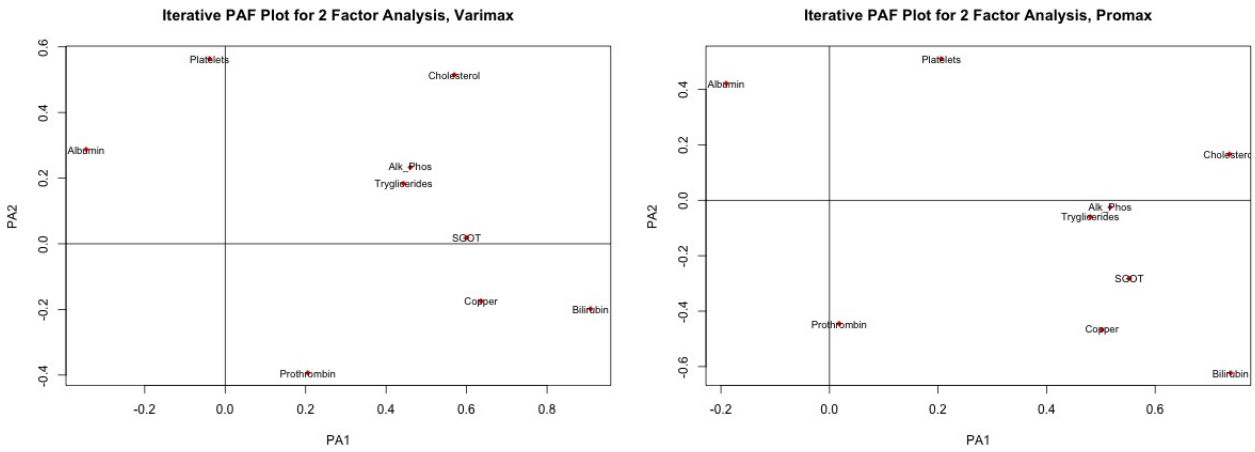

Figure 15: Interative PAF Plots with Varimax and Promax rotations

These two graphs are very different from each other, which shows that the nature of the rotations shows two different factors in the data. Some notable differences include:

- In the Varimax Plot, Platelets is more to the left compared to Prothrombin on the direction of PA1, while in Promax it is the opposite.
- There seems to be a very interesting rotation of the cluster of the indicators Cholesterol, Bilirubin, SGOT, Copper, Triglycerides, Alkaline Phosphatase, as in the Promax rotated plot, the cluster seems to be a clockwise rotation of where it was in the Varimax plot around the origin (or some point in that general region).

Direction 1 likely represents a factor for how far the cirrhosis has progressed in an individual, measured through the stage of cirrhosis. Looking at the boxplots printed at start of the document, we can explain the trends seen in this dataset. On the right side of the graph are the indicators that are correlated with a higher stage of cirrhosis (Stage 3 and 4), as the indicators have much higher values for individuals in a later stage of cirrhosis. Alternatively, the indicators on the left side of the graph are more likely correlated with a lower stage of cirrhosis (Stage 1, 2 and 3). For example, Albumin levels are higher in lower stages of cirrhosis than they are in higher stages which makes sense for why it is on the opposite side of the plot.One thing to note is that the interpretation of "high stage" is somewhat variable, as individuals in Stage 3 can be classified as either in high or low stage due to their status as an intermediate level of cirrhosis.

Direction 2 likely represents a factor related to the coagulation of blood. The upper end of the axis, with Platelets and Cholesterol, are for values that correlate with high rates of coagulation.On the bottom of the axis are indicators related with anti-coagulation. Prothrombin Time measures how long it takes for a blood to clot in a sample, which is why an increased Prothrombin Time leads to anti-coagulation, and puts it far on the opposite side of the graph compared to Platelets.The other factors found at the bottom of the graph, like SGOT, Copper, and Bilirubin, have been correlated with or directly cause a decrease in coagulation. While there are some factors that are in the middle that shift significantly over both rotations (like Alkaline Phosphatase and Triglycerides), after reviewing the literature, they do not have a very strong correlation with clotting, at least not as much as Prothrombin Time and Platelets.

The reason for the differences in these two plots might be because of the nature of the rotation done. Since Promax assumes that there are correlations between the variables and Varimax assumes that there are

no correlations between the variables, it makes sense they give different plots and different loadings when rotated.

8 Conclusion and Discussion

The assorted multivariate analysis methods help describe important trends in the cirrhosis dataset, which gave useful information regarding the relationship between the independent and response variables, as well as the way that the stage of cirrhosis affects them. Each of the three methods gave different results and were used for different purposes:

1. **MANOVA and Multivariate GLM** were used to determine the relationships between individual characteristics/history and the different response variables. Through this method, we were able to find that some response variables, like Copper, Albumin, and Cholesterol, were significantly correlated with a difference in sex or a change in age. The treatment drug, however, did not have a significant effect on these variables. This method allowed for the comparison and co-analysis of many different categorical and environmental variables at once—something that the other methods are not able to do as well.

2. **Ordination** was used to determine the relationships between individual patients tracked in the study and the different response variables, to see if there are any shared characteristics amongst those individuals that are highly correlated with a particular biological measure. We find that the number of days between the individual's registration for the study and a major event (like death, a hospital visit, etc.) goes in the opposite direction of increasing stage, while age goes in the direction of increasing stage. Variables like how Bilirubin and Copper are highly correlated with late stage cirrhosis, while others are closely related with lower stages of cirrhosis. Ordination allowed for the analysis of what stage of cirrhosis was most closely related to which biological metric, and how environmental variables affect these trends.

3. **Factor Analysis** was used to determine if there are any underlying factors in the relationships between the response variables that can explain the trends seen in the data. The main trend seen was the difference in the stage, as the first direction discriminated between whether a variable was correlated with high or low level of cirrhosis. The second major factor is related to the coagulation of blood, with the direction with Platelets representing variables related to coagulation and the direction with Prothrombin Time representing variables related to anti-coagulation. Factor Analysis helps determine trends beyond the data, and see if there are any other factors that can lead to variability in the data.

These analyses supplement the existing scientific literature by providing statistical results that reinforce already-proven biological results. Hopefully, this analysis can reinforce the necessity to understand and interpret cirrhosis population data to understand the dynamics and trends of the disease in a larger sociological and chronic epidemiological context.

One source of future research could be in the efficacy of D-penacillamine as a treatment drug for cirrhosis; given that our analyses suggest that the treatment drug does not have a statistically significant effect on the different factors that were shown to be closely related with cirrhosis, it may be possible that the drug is not actually useful. Most literature on the drug is limited to the late 1900s, which was the same time when this dataset was published by Mayo Clinic, which suggests that perhaps this data or other information in the same timeframe showed the inefficacy of the drug (or a better drug was found). In either case, it would be useful to find conclusive evidence for whether the drug does or does not help with cirrhosis in the long and short-term to better inform future drug creation for this disease.

9 References

Bittner, V. (2008, January 28). Impact of gender and life cycle on Triglyceride Levels. Medscape. Retrieved May 5, 2023, from https://www.medscape.org/viewarticle/569070#::text=In%20young%20adulthood%2C%20men%20tend,dL%20vs%20104%20mg%2FdL.

Buxaderas, S. C., & Farré-Rovira, R. (1986). Whole blood and serum copper levels in relation to sex and age. Revista espanola de fisiologia, 42(2), 213–217.

Ferrara, A., Barrett-Connor, E., & Shan, J. (1997). Total, LDL, and HDL cholesterol decrease with age in older men and women. Circulation, 96(1), 37–43. https://doi.org/10.1161/01.cir.96.1.37

Ginès, P., Krag, A., Abraldes, J. G., Solà, E., Fabrellas, N., & Kamath, P. S. (2021). Liver cirrhosis. Lancet (London, England), 398(10308), 1359–1376. https://doi.org/10.1016/S0140-6736(21)01374-X

Hanai, T., Nishimura, K., Miwa, T., Maeda, T., Ogiso, Y., Imai, K., Suetsugu, A., Takai, K., & Shimizu, M. (2021). Usefulness of nutritional therapy recommended in the Japanese Society of Gastroenterology/Japan Society of Hepatology Evidence-based clinical practice guidelines for liver cirrhosis 2020. Journal of Gastroenterology, 56(10), 928–937. https://doi.org/10.1007/s00535-021-01821-z

Liu, Y., & Chen, M. (2022). Epidemiology of liver cirrhosis and associated complications: Current knowledge and Future Directions. World Journal of Gastroenterology, 28(41), 5910-5930. doi:10.3748/wjg.v28.i41.5910

Naveau, S., Perlemuter, G., & Balian, A. (2005). Epidémiologie et histoire naturalle de la cirrhose [Epidemiology and natural history of cirrhosis]. La Revue du praticien, 55(14), 1527–1532

Schuppan, D., & Afdhal, N. H. (2008). Liver cirrhosis. Lancet (London, England), 371(9615), 838–851. https://doi.org/10.1016/S0140-6736(08)60383-9

Scaglione, S., Kliethermes, S., Cao, G., Shoham, D., Durazo, R., Luke, A., & Volk, M. L. (2015). The Epidemiology of Cirrhosis in the United States: A Population-based Study. Journal of clinical gastroenterology, 49(8), 690–696. https://doi.org/10.1097/MCG.0000000000000208

Tapper, E. B., & Parikh, N. D. (2018). Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: Observational study. BMJ. doi:10.1136/bmj.k2817

Weaving, G., Batstone, G. F., & Jones, R. G. (2015). Age and sex variation in serum albumin concentration: An observational study. Annals of Clinical Biochemistry: International Journal of Laboratory Medicine, 53(1), 106–111. https://doi.org/10.1177/0004563215593561

Xu, J., Murphy, S., Arias, E., & Kochanek, K. (2021). Deaths: Final Data for 2019. doi:10.15620/cdc:106058