

Contents

List of figures	2
1.Introduction to the dataset	3
2. Key Challenges and Problems	5
3.General analysis and summary of the dataset	6
4. Unsupervised Learning-Clustering Using K-Means	10
5.Supervised Learning-Logistic Regression	11
6.Reflection of the chosen methods and Conclusion	13
Software versions, data and included packages	16
Bibliography	16

List of Figures

3.1 Histogram of Bilirubin, Albumin and Age	6
3.2 Boxplots of baseline characteristics by treatment group	7
3.3 Correlation between Age, Bilirubin and Albumin	8
3.4 Scatter plot of Age vs Bilirubin	9
4.1 Cluster of variables	10
4.2 Evaluation summary	11
5.1 The confusion matrix	12
5.2 Evaluation summary (Regression)	13

Chapter 1

1.1 Introduction to the dataset

Liver Cirrhosis is a chronic disease in which the small bile ducts in the liver become damaged, inflamed, and eventually destroyed. When there are no bile ducts, bile builds up and causes liver damage. The symptoms of the disease include fatigue and itching skin. The presentation of these symptoms can take long to be noticed until the liver is significantly damaged. Some of the known common causes of liver cirrhosis is excessive alcohol consumption, some conditions that affect the liver like fatty liver disease and chronic hepatitis.

Diagnosis is based on medical and family history, physical exam, and the markers on medical tests that are often used for diagnosis are bilirubin and albumin. There are no cures for cirrhosis, but medications are available to help slow the progression of the disease and prevent complications. (National Institute of Diabetes, Digestive and Kidney Disease,2023)

The severity of the disease is measured in four stages being:

- . Stage 1- involves scaring of the liver with mild symptoms
- . Stage 2-moderate symptoms
- . Stage 3-involves development of abdominal distention with severe symptoms
- . Stage 4-life threatening. (Jacob and Wedro,2024)

The purpose of this report is to investigate the efficacy of the drug D-Penicillamine in survival outcomes of patients with liver cirrhosis using data from a randomised control trial.

The questions that the report is addressing include:

- a. Does treatment with D-Penicillamine improve survival compared to placebo?
- b. What key disease severity predictors of cirrhosis influence survival of patients? - this will be done using the Logistic regression model
- c. To find out if there are groups of patients with distinct disease patterns- will be achieved using the K-Means clustering model.

1.2 Dataset description

Source: The data is from the ICU Machine Learning Repository:

<https://archive.ics.uci.edu/ml/datasets.php>, the title of the document is Cirrhosis Patient Survival Prediction. This data is about 17 clinical features utilized for predicting survival of patients with liver cirrhosis. The data used was sourced from a Mayo clinic randomised control trial on primary cirrhosis of the liver carried out from 1974 to 1984. The number of participants who took part in the trial were 418 both females and males of various age groups. The features utilized to conduct the trial include:

>Age

> Sex

>Bilirubin

>Albumin

>Hepatomegaly

>Edema

>Ascites

>Status (Binary Outcome-death and survival)

Analysing the dataset will help to find out if D-Penicillamine is effective in improving lifespan of patients with cirrhosis and in that case can be adopted as part of treatment regimens. It will also help discover some patterns in patients with cirrhosis that can be used as markers or risk factors to identify in liver cirrhosis patients.

Chapter 2

Key challenges and Problems

The main challenge that was faced during data exploration is that this dataset contained missing values across several variables even those critical for assessing the patients' progress in the trial. This might have happened because of patients' dropouts, missed follow-up appointments, lack of adherence to treatment or unrecorded data by research staff which consequently resulted in inconsistent data. Patients' dropouts and non-adherence to treatment is always likely during clinical research trials, especially for a trial with a long duration of follow-up, in this case 10-year period. To handle this challenge, the missing values were dropped. This could have reduced the statistical analysis strength and potentially caused bias as well as have had an impact on the final models.

Dropping these missed entries meant that some important information is removed and the dataset size reduced limiting the statistical analysis and the performance of the models. Another challenge was that the units in which some key variables used in analysis were recorded in a non-scalable format, particularly the age and N_days variables. These variables were recorded as days in the dataset which was not suitable to use. To utilize them conversion into years by dividing by 365 was done. This conversion was necessary for a meaningful analysis and for use in visualising the plots.

Chapter 3

General Analysis and Summary of the dataset

The dataset contains 276 rows and 20 columns. The first part of the analysis was to find out if there are missing values and handle them, followed by descriptive statistical analysis and visualisation of the baseline of key variables that are often used as diagnosis for liver cirrhosis. According to Roberts and Torgerson (1998), “In controlled randomised trials, randomisation ensures that allocation of patients to treatment is left purely to chance. The characteristics of patients that may influence outcomes are distributed between groups so that any difference in outcome can be assumed to be due to the intervention. Imbalance between groups in baseline variables that may influence outcome (such as age, disease severity) can bias statistical tests. In reporting clinical trials, it is recommended that variables should be described for each treatment group. This clarifies generalisability of the trial and ascertain that randomisation was done properly”. To assess this dataset for balance the plots below were produced.

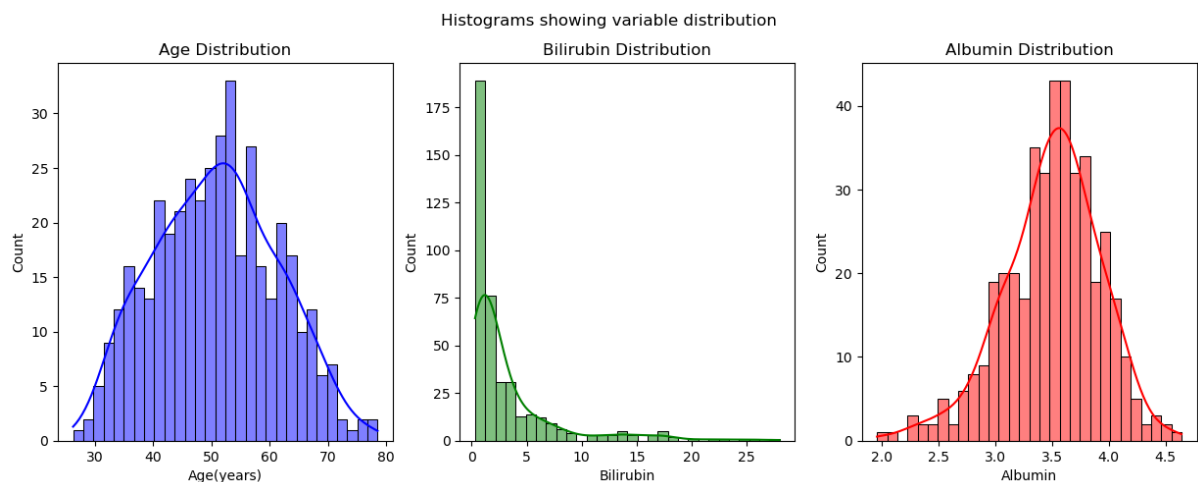


Figure 3.1 Histograms of Bilirubin, Albumin and Age

The above histograms on figure 3.1 lays out a summary of how the key variables mentioned were distributed across all participants at baseline. The age histogram illustrates that most of the participants were between the ages of 50 and 60 years old. This suggests that liver cirrhosis seems to be common as people become older. The bilirubin distribution uncovered that in this cohort, majority of the participants had values of between 0 and 5mg/dl. This indicates that most of the participants had moderate liver cirrhosis. On the other hand, the albumin histogram depicts that there was less variation of baseline values across the groups. General conclusion from all these histograms is that the dataset included a range of patients and confirms that randomisation was done successfully.

Subsequently, boxplots to assess baseline of the key variables by treatment groups was developed to further check how the groups were balanced. The bilirubin boxplot shows that the median of both treatment and placebo group is the same, suggesting that both groups have the same level of disease severity, however it can be seen that the placebo group had the highest deranged values(outliers).The albumin boxplot further reinforces the significant comparability of the two treatment groups as the median is similar with a slight difference noticed in the placebo groups which has more values lower than median . The Age boxplot shows that the median age is almost the same between the two groups, an important factor in randomisation that confirms that research participants were almost the same age therefore comparison was fair.

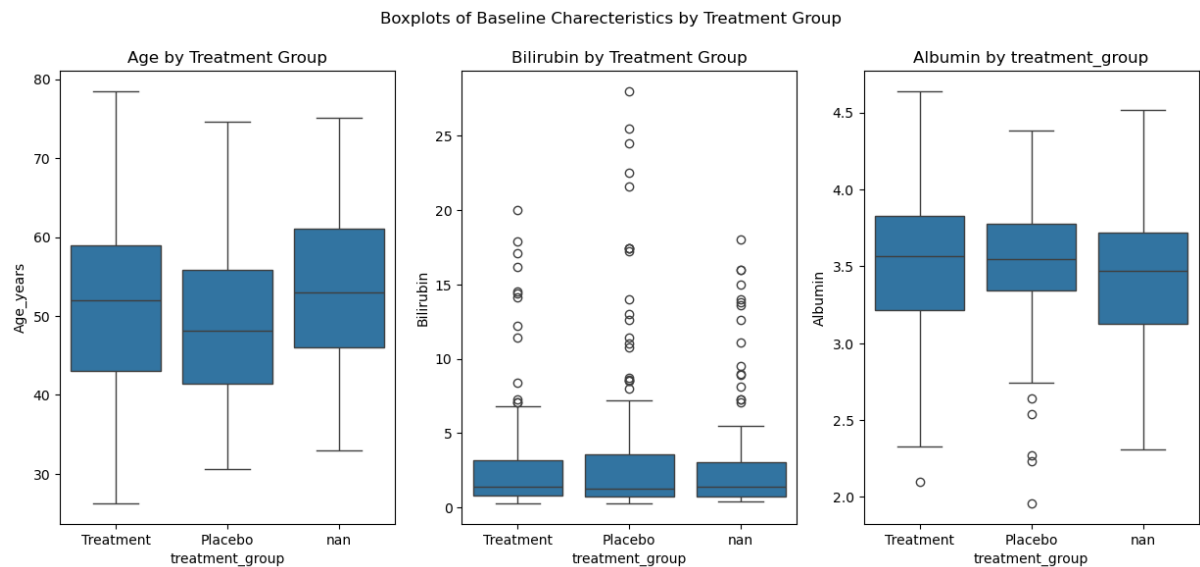


Figure 3.2 Boxplots of Baseline Characteristics by treatment group

Another angle to understand the data set was to produce a correlation heatmap (Figure 3.3) to study how the key variables were related to each other. The heatmap shows a negative correlation between bilirubin and albumin. As the bilirubin increases, the albumin goes down. This affirms the clinical knowledge that is often seen on laboratory results when there is some degree of liver damage. There is a positive correlation between age and bilirubin which explains why most of the participants of this trial were older, as people advance in age their condition worsens. Another factor is that people around this age often take interest in research sometimes in hope to find a solution to their condition or just to be part of development of new evidenced care policies as compared to younger people. It is therefore likely that more people who consented for the trial and adhered to the study requirements were older.

Correlation Heatmap-Identifying correlation between Albumin ,Bilirubin and Age

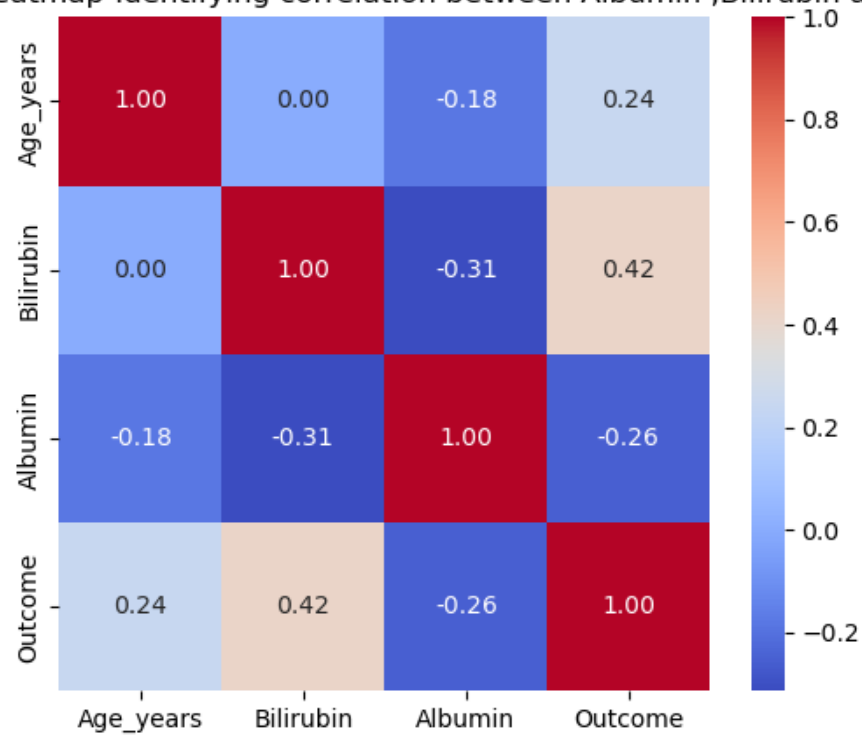


Figure 3.3 Correlation between age, bilirubin and albumin

A scatter plot (figure 3.4) of age vs bilirubin was also developed to see if there is a trend common among the participants. Insights revealed from this is that older patients tend to have higher bilirubin levels and therefore have more severe cases of the disease as compared to younger ones.

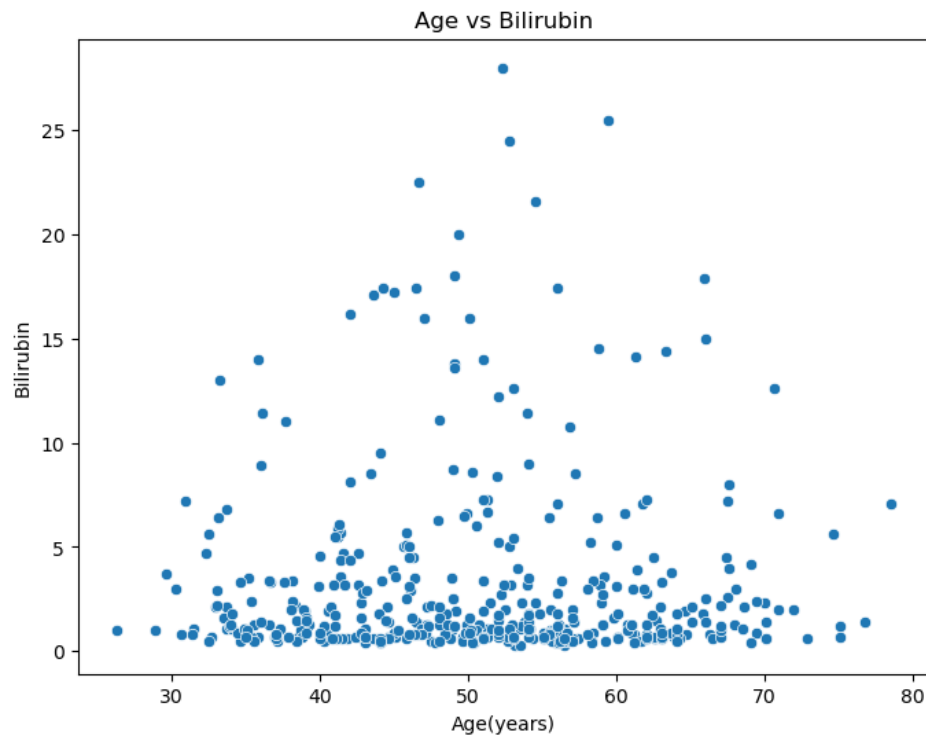


Figure 3.4 Scatter plot of Age vs bilirubin

All these visualisations were important as a baseline for the unsupervised and supervised modelling. They provided proof that the data set was significantly balanced. Building on this information, the next step was to further uncover deeper insights about the dataset. To achieve this, two machine learning models were used as elaborated in the following chapters four and five.

Chapter 4

Unsupervised Learning-Clustering Using K-Means

The aim was to identify if there are underlying patient groups with similar biochemical and clinical features. The 3 features that were scaled then applied to the K-Means cluster include age, albumin and bilirubin. To determine the number of optimal clusters two methods were used i.e. the Elbow and silhouette, which indicated $k=3$ as the best. PCA was applied for a better 2D presentation. The model successfully clustered the variables as illustrated by figure 4.1, however there was some overlapping which is explained by the model evaluation metrics summary on figure 4.2

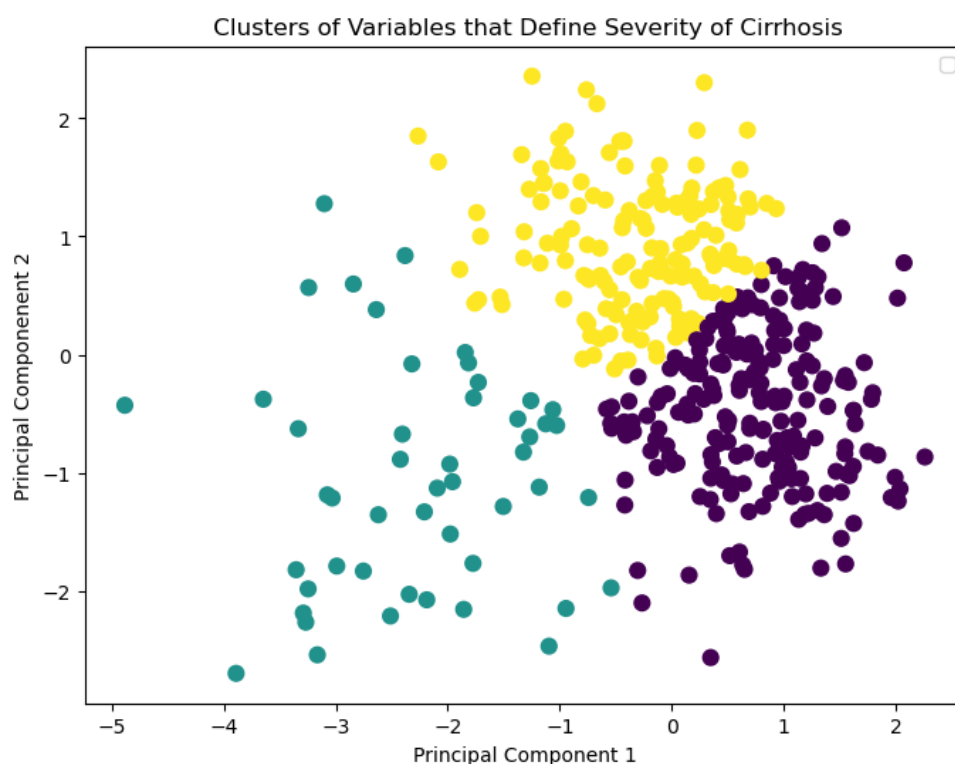


Figure 4.1 Cluster of variables

Figure 4.2 Evaluation summary(K-means model)

Silhouette Coefficient: 0.4054910298762692

Calinski-Harabasz Coefficient: 353.0027228933699

Homogeneity score; 0.13237060235496753

Completeness score: 0.09151394227094996

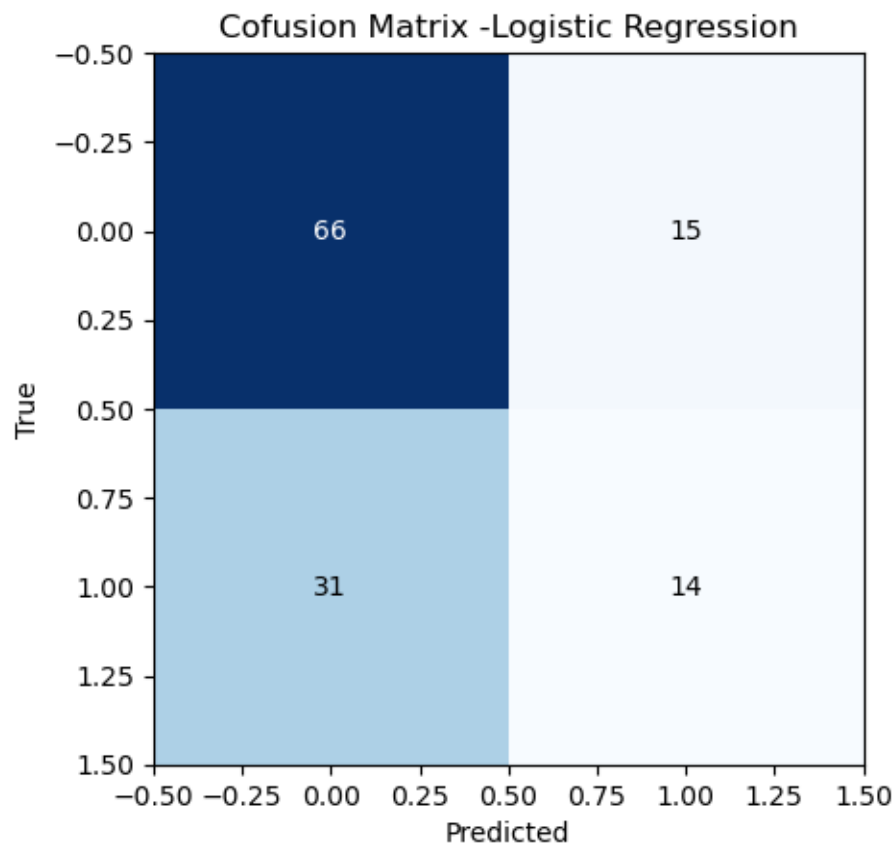
The Silhouette Coefficient of 0.40 suggests that there is very moderate distinct separation of clusters. For further evaluation the homogeneity and completeness scores were carried out, both were low elaborating that despite the capturing the groups of patients with similar characteristics this did not align the features and survival outcome but rather only showed disease severity. This implies that the features used to fit this model are not the only determinants of survival. Factors that can lead to poor prognosis and eventually death among patients with liver cirrhosis are complex and a wide range. Perhaps incorporating more of the features like ascites, hepatomegaly could have improved the sensitivity of the model.

Chapter 5

Supervised Learning-Logistic Regression

Following the clustering analysis, Logistic regression was developed to predict survival outcome. The variables used were age, bilirubin and albumin. Target and dependent variables were assigned, then data was split into 70% training, 30% testing and scaled before fitting into the model. The performance of the model was the evaluated using the confusion matrix fig 5.1 and the classification metrics as shown on the summary

Figure 5.1 The Confusion matrix



The confusion matrix illustrates that the model identified 66 true negatives (participants alive) and 14 true positives (deaths), 10 cases were misclassified as deaths while 36 were misclassified as survivors. This confirms that the model was better at predicting alive patients than deaths. Some deaths might have been wrongly placed in the survivors.

Figure 5.2 Evaluation summary (Regression model)

	Precision	recall	F1-score	support
0	0.63	0.87	0.73	76
1	0.55	0.24	0.33	50
accuracy			0.62	126
Macro avg	0.59	0.55	0.53	126
Weighted avg	0.60	0.62	0.57	126
[[66 10] [38 12]				

The model showed better performance in identifying those who were alive than the death cases as indicated by the recall of 0.87, recall of 0.24 shows that the model did not capture majority of deaths concluding that there might have been some imbalance in the data resulting from residual missing values therefore the model learned to prioritise the majority class. The large gap evident between the two classes of deaths and survivors further confirms the bias of this model towards the majority class. The precision tells us how many predicted cases were correct, in this model 63% of cases were truly predicted as alive and 55% were predicted as dead.

Chapter 6

Reflection of chosen methods and conclusion

For a thorough analysis of any dataset, it is good practise to use more than one model. This approach helps to improve predictive accuracy and can unearth effective applications of the models developed. In that breath for this report two models were chosen to provide two distinct analytical in-depth understanding of the cirrhosis dataset.

The k-means clustering was selected because it is simple to interpret for a beginner in data analysis. It is also effective in identifying patterns in datasets that contain mostly

continuous data as it was the case with cirrhosis dataset. The model performance was relatively moderate in terms of clustering; however, the limitations were exceeding the success as completeness, homogeneity scores were very low indicating that the clusters did not directly align with survival outcomes.

On the other hand, the logistic regression model was chosen because the aim was to predict a binary outcome i.e. survival or death and this type of modelling is well suited for this. The accuracy of the model was 0.62 indicating moderate performance. The limitations are that it lacked sensitivity to recognise the class with fewer cases and leaned more towards the majority. In that case it cannot be adapted for use as it is likely to produce false results.

Finally, at the beginning of this report there was a fundamental question of whether the drug D-Penicillamine was effective in reducing mortality among cirrhosis patients. Gathering from the analysis and performance of the models it shows that the drug has not demonstrated a clear reduction in mortality. The drug was not adapted as treatment for liver cirrhosis after several randomised trials proven that it is not effective.

Software Versions and packages

Python version: Python 3.13.2

Packages used

Pandas

Matplotlib

Seaborn

Sklearn

Bibliography

Jacob, D. and Wedro, B. (2024), "Cirrhosis (Liver)". [Accessed November 1st, 2025].

URL: <https://www.medicinenet.com/cirrhosis/article.htm>

Roberts, C. and Torgerson, D. (1998), "Randomised methods in controlled trials".

[Accessed October 27th, 2025].

URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC1114206>

National Institute of Diabetes, Digestive and Kidney Disease (2023) Cirrhosis. Available from: <https://www.niddk.nih.gov/health-information/liver-disease/primary-biliary-cholangitis>