Data Visualization Techniques for Analyzing Cirrhosis Patient Outcomes Using R

ABSTRACT

In an era driven by data, clear methods for visualization and interpretation are essential. Data Visualization provides graphical representations, making complex datasets understandable through diagrams, charts, and maps. It's widely used across sectors, notably in healthcare. This study introduces Data Visualization techniques for analyzing cirrhosis patient survival rates using R. It illustrates how visuals enhance data accessibility, providing insights into survival factors and disease progression. The paper aims to offer a comprehensive overview of Data Visualization methods for predicting survival in cirrhosis patients, aiding informed decision-making, and advancing knowledge in the field.

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here   • Insert CCS text here

KEYWORDS

Cirrhosis, Predictive modeling, Data visualization, R programming, Clinical outcomes

**1. INTRODUCTION**

Liver cirrhosis, a condition marked by prolonged liver damage and extensive scarring, has significant implications for health. Often resulting from chronic conditions such as hepatitis or long-term alcohol abuse, cirrhosis impairs liver function over time. This study focuses on data sourced from a Mayo Clinic research project on primary biliary cirrhosis (PBC) conducted between 1974 and 1984.

Survival prediction in cirrhosis patients stands as a crucial aspect of medical research, aiming to enhance treatment strategies and patient care. Leveraging the power of data visualization techniques, this study seeks to unravel the complexities of cirrhosis patient data and gain insights into factors influencing survival rates. The dataset, obtained from the UCI Machine Learning Repository, encompasses a rich array of clinical and demographic variables pertaining to cirrhosis patients. Through innovative visualization methods, we aim to elucidate hidden patterns and relationships within the data, providing a deeper understanding of cirrhosis progression and patient outcomes.

Similar to the enigmatic nature of cirrhosis patient survival prediction, predicting patient outcomes in medical research presents significant challenges. While traditional statistical approaches offer valuable insights, they often fall short in capturing the intricate dynamics of disease progression and its impact on patient survival. Thus, there arises a pressing need for sophisticated data visualization methods tailored specifically to unlock the potential insights hidden within medical datasets.

By harnessing the power of data visualization, this research endeavour aims to empower healthcare professionals with the tools and knowledge necessary to make informed decisions in patient management. Through the exploration of influential factors and the identification of predictive trends, we aspire to contribute to the advancement of medical science and the enhancement of patient care practices in the realm of liver cirrhosis.

**2. LITERATURE REVIEW**

This research presents a comprehensive analysis contributing to improved clinical decision-making and personalized patient care strategies when data visualization techniques are used.

[1]. The Model for End-Stage Liver Disease (MELD) serves as a reliable tool for assessing short-term mortality risk in patients with chronic liver disease. Based on serum creatinine, bilirubin, and INR, it offers advantages over the Child-Turcotte-Pugh classification, including greater discriminatory ability and objectivity. Despite variations in laboratory measurements, the MELD scale's continuous nature and statistical derivation provide a more robust predictor of survival. Complications of portal hypertension, such as ascites and encephalopathy, do not significantly enhance its predictive accuracy. The MELD scale's validity extends across diverse etiologies and disease severities, supporting its use in liver transplant allocation decisions.

[2]. Yang et al. (2022) in BMC Gastroenterology highlights the evolving landscape of liver transplantation (LT) and the challenges in predicting short-term survival for acute-on-chronic liver failure (ACLF) patients post-LT. While the Model for End-Stage Liver Disease (MELD) score has traditionally been used for organ allocation, its limitations in predicting ACLF patient outcomes have led to the exploration of alternative models like the Chronic Liver Failure Consortium (CLIF-C) ACLFs. Additionally, machine learning (ML) algorithms, particularly the Random Forest (RF) model, show promise in improving prognostic accuracy. However, further multicenter studies are needed to validate these findings and optimize organ allocation strategies.

[3]. The growing interest in applying machine learning (ML) techniques to medical data, particularly in predicting outcomes in liver transplantation (LT). Traditional statistical models like the Cox model are compared with ML methods such as Random Survival Forests (RSF) and Artificial Neural Networks (NNs). While RSF outperformed Cox models in terms of discrimination, NNs showed better predictive performance. Challenges include interpretation and stability of ML models, but they offer promise for improving decision-making in healthcare. Further research is needed to explore the impact of variable selection and dynamic methods on predictive accuracy.

[4]. The literature review highlights the significant mortality rates among cirrhotic patients admitted to the ICU, ranging from 46% to 64%. Severity of illness was assessed using four mortality prediction systems, which indicated a high severity of illness in the study population. Coma and acute renal failure emerged as independent predictors of mortality. The study also questions the utility of invasive procedures such as pulmonary artery catheterization in cirrhotic patients. Furthermore, all patients admitted post-cardiac arrest died, prompting discussion on the appropriateness of CPR in this population. These findings underscore the need for rationalization of critical care delivery and informed decision-making regarding aggressive interventions.

[5]. The literature on predicting mortality among patients with liver cirrhosis underscores the critical need for accurate prognostic tools due to the condition's high morbidity and mortality rates. While the Model for End Stage Liver Disease (MELD) score has been a standard tool for risk assessment, its limitations in predicting outcomes across various patient populations and timeframes have prompted exploration into alternative approaches. Recent studies have highlighted the potential of machine learning techniques, particularly deep learning algorithms, to leverage electronic health record data for more precise mortality predictions. These advancements aim to improve clinical decision-making and ultimately enhance patient outcomes in cirrhotic populations.

[6]. Novel nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis, utilizing LASSO regression analysis on data from the MIMIC database. It addresses the significant mortality risk associated with sepsis in cirrhosis patients, highlighting the scarcity of research in this area. The nomogram, incorporating nine independent variables including age, heart rate, bilirubin levels, glucose, sodium, anion gap, fungal infections, mechanical ventilation, and vasopressin use, demonstrates superior predictive performance compared to established scoring systems like SOFA, MELD, and ABIC. The study underscores the need for prospective validation of the nomogram in clinical settings.

**[**7**]**. The study by Campbell et al. (2015) evaluated prognostic scoring tools for predicting ICU mortality in patients with cirrhotic liver disease. They found that the RFH and CTP + L scoring tools showed similar performance in predicting ICU mortality, with the latter being more practical due to its simplicity. Multivariable analysis identified lactate, bilirubin, and PaO2/FiO2 ratio as significant predictors of mortality. While hepatic encephalopathy scores did not enhance predictive value, further validation is needed. The study highlights the importance of practical, validated scoring tools for assessing ICU mortality in cirrhotic patients, suggesting potential improvements in patient outcomes.

[8]. Hepatocellular carcinoma (HCC) risk in patients with alcoholic cirrhosis underscores the importance of accurate statistical methods. Studies emphasize the superiority of the cumulative incidence function over the Kaplan-Meier estimator for estimating HCC risk, given its ability to appropriately handle competing events like death without HCC. Multistate disease models are advocated for analyzing the complex clinical course of cirrhosis, delineating transitions between various states. Prognostic factors, both causal and predictive, are of interest in understanding HCC development and patient outcomes. Recommendations for statistical methods, such as Cox regression and Fine and Gray regression, depend on the research question and disease model, ensuring robust analyses and informed decision-making.

[9]. The significance of the MESIAH score in predicting survival outcomes for patients with hepatocellular carcinoma (HCC). It highlights the complexity of prognostic factors in HCC, emphasizing the importance of tumor extent and underlying liver function. The MESIAH score stands out for its reliance on objective variables, such as tumor characteristics and MELD score, facilitating its applicability in diverse clinical settings. Moreover, comparative analyses with existing staging systems underscore its superior performance, suggesting its utility in both epidemiological research and clinical decision-making for patient prognosis and treatment guidance.

[10]. The prognostic nomogram developed by Xu et al. represents a significant advancement in predicting in-hospital mortality in patients with liver cirrhosis and esophageal varices (LCEV). Leveraging data from the MIMIC databases, the study identified key prognostic factors including age, Elixhauser score, AG, sodium, albumin, bilirubin, INR, vasopressor use, and bleeding. By incorporating these factors into a user-friendly nomogram, clinicians can better stratify patients' risk of death and tailor treatment strategies accordingly. Validation analyses demonstrated the nomogram's superiority over existing scoring systems, highlighting its potential to enhance clinical decision-making and improve outcomes in LCEV patients.

**3. DATASET**

A set of clinical data called the Cirrhosis Patient Survival Prediction dataset is used to forecast the prognosis of individuals suffering from liver cirrhosis. It originates from the 1974–1984 Mayo Clinic research on primary biliary cirrhosis (PBC) of the liver and was contributed to the UCI Machine Learning Repository on September 11, 2023.

The dataset in question pertains to the prediction of survival among cirrhosis patients, housed within a tabular format within the Health and Medicine domain. Tailored for classification tasks, it encompasses a blend of real and categorical feature types across 418 instances and 17 attributes. This dataset offers a vital resource for analyzing and forecasting patient outcomes within the context of cirrhosis treatment and management.

The attributes include:

* ID: unique identifier (patient #)
* N\_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
* Status: status of the patient C (censored), CL (censored due to liver Tx), or D (death)
* Drug: type of drug D-penicillamine or placebo
* Age: age in [days]
* Sex: M (male) or F (female)
* Ascites: presence of ascites N (No) or Y (Yes)
* Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)
* Spiders: presence of spiders N (No) or Y (Yes)
* Edema: presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
* Bilirubin: serum bilirubin in [mg/dl]
* Cholesterol: serum cholesterol in [mg/dl]
* Albumin: albumin in [gm/dl]
* Copper: urine copper in [ug/day]
* Alk\_Phos: alkaline phosphatase in [U/liter]
* SGOT: SGOT in [U/ml]
* Triglycerides: triglicerides in [mg/dl]
* Platelets: platelets per cubic [ml/1000]
* Prothrombin: prothrombin time in seconds [s]
* Stage: histologic stage of disease (1, 2, 3, or 4)

Researchers and medical practitioners who want to create prediction models for liver cirrhosis patient prognoses may find this dataset very helpful.

**4. DATA VISUALIZATION TECHNIQUES**

In this section, all the diagrams/graphs are plotted using RStudio. The dataset that we are using for this research (the dataset at a glance)

A group of small colored lines

Description automatically generated with medium confidence

Fig: Cirhosis Patient Servival PredictionDataset

In this dataset, there are missing values. For this study, we used the median and mode functions in RStudio to replace the missing values for data visualization purposes.

A white background with black text

Description automatically generated

Fig-1: Dataset Containing Null Values

A close-up of a computer screen

Description automatically generated

Fig-2: After Replacing the Null Values

The following figures show Univariate Visualization techniques. Examines one attribute at a time.

4.1. **Histograms:**

The dataset’s quantitative data (Albumin, Copper, Alk\_Phos, SGOT, Tryglicerides, Platelets, Prothrombin) is represented graphically by histogram. It depicts the frequency of occurrences within specific ranges to illustrate the distribution of the data.

A group of graphs showing different types of data

Description automatically generated with medium confidence

Fig-3: Histogram

4.2 **Bar Graphs:**ISBN:978-1-4503-0000-0/18/06

The dataset’s numerical and categorical data (Drug, Sex, Status, Ascites, Hepatomegaly, Spiders, Edema and Stage) are displayed using a bar graph. The length of the bars distinguishes between feature values.

A group of blue rectangular shapes

Description automatically generated

Fig-4: Bar Graph

4.3 **Box Plot:**ISBN:978-1-4503-0000-0/18/06

The dataset’s numerical and categorical data (Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phos and SGOT) are displayed using a box plot. The values from the dataset outside the box are considered outliers, while the middle line represents the median value.

A group of graphs showing different types of numbers

Description automatically generated with medium confidence

Fig-5: Box Plot

Year:2018

The following figures show Multivariate Visualization techniques. Examines more than one attribute at the same time.

4.4 **Violin Plot**Date:June

In a violin plot, the median is represented by a black line. The shape of the violin plot is formed where there is a wide distribution of data, with the narrower portion indicating a smaller spread.

The violin plot below represents the distribution of SGOT levels for each status category as a density curve mirrored on either side of the median line. The median SGOT value for each status category is shown as a horizontal line within the violin plot. The width of each violin plot segment indicates the density of SGOT levels within that category. By observing the plot, one can discern the distribution of SGOT levels based on patient status and infer any notable differences in median SGOT values between these groups. The inclusion of boxplots alongside each violin plot segment facilitates a visual comparison of the median and quartiles of SGOT levels among different patient statuses. This combined visualization method provides a detailed insight into how SGOT distributions correlate with patient outcomes in cases of cirrhosis.

A graph of different colored shapes

Description automatically generated with medium confidence

Fig-6: Violin Plot for STOG by Status

In this violin plot, the distribution of Alk\_Phos levels among cirrhosis patients with different statuses is depicted. The plot visually represents how Alk\_Phos levels vary across these categories, showing density curves that highlight the distribution patterns within each group.As mentioned before, the presence of boxplots within the violin plot segments enables a direct comparison of the median and quartiles of Alk\_Phos levels across different patient statuses. This integrated visualization approach provides detailed insights into how Alk\_Phos distributions relate to patient outcomes in cirrhosis cases.

A graph of different colored shapes

Description automatically generated with medium confidence

Fig-7: Violin Plot for Alk\_Phos by Status

The plot below visually displays the density curves, highlighting the distribution patterns of triglyceride levels within each status category. The inclusion of boxplots within the violin plot segments facilitates a clear comparison of the median and quartiles of triglyceride levels across different patient statuses. This combined visualization approach offers detailed insights into the relationship between triglyceride distributions and patient outcomes in cirrhosis cases.

A graph of different colored shapes

Description automatically generated with medium confidence

Fig-8: Violin Plot for Triglyceride by Status

The plot below visually showcases how platelet levels vary within each status category, using density curves to highlight distribution patterns. The presence of boxplots within the violin plot segments allows for a direct comparison of the median and quartiles of platelet levels across different patient statuses. This integrated visualization approach provides detailed insights into how platelet distributions correlate with patient outcomes in cirrhosis cases.

A diagram of a graph

Description automatically generated with medium confidence

Fig-9: Violin Plot for Platelets by Status

4.5 **Scatter Plot**Date:June

Here, the scatter plot showcases the relationship between Age and Serum Bilirubin within the dataset. The x-axis represents Age, and the y-axis represents Serum Bilirubin, with each dot representing an individual data point. Despite potential variations in alignment, a predominant trend along the x-axis suggests a strong positive correlation, where Serum Bilirubin tends to consistently change with Age across the dataset.

A graph with red and blue dots

Description automatically generated

Fig-10: Scatter Plot of Age by Serum Bilirubin

The second scatter plot illustrates the relationship between copper and Alk\_Phosphatase within the dataset. The clustering of points towards the lower left corner indicates an inverse relationship, where higher values of copper correspond to lower values of Alk\_Phosphatase, suggesting a negative correlation between these two variables.

A graph with red blue and green dots

Description automatically generated

Fig-11: Scatter Plot of Copper by Alk\_Phos

The third scatter plot displays individual data points as dots on a two-dimensional plane, with Cholesterol plotted on the x-axis and Albumin on the y-axis. This visualization allows for examining the relationship between Cholesterol and Albumin levels within the dataset. Most points forming a perfect straight line along the y-axis indicate a strong positive correlation, where Albumin levels change linearly with changes in Cholesterol levels across the dataset.

A graph with red blue and green dots

Description automatically generated

Fig-12: Scatter Plot of Cholesterol by Albumin

The fourth scatter plot presents individual data points as dots on a two-dimensional plane, with Triglycerides plotted on the x-axis and Platelets on the y-axis. Most points forming a perfectly straight line starting from the x-axis and extending upwards indicate a strong positive correlation between Triglyceride levels and Platelet counts, where Platelets increase linearly with increasing Triglyceride levels across the dataset.

A graph of a number of red and blue dots

Description automatically generatedPrice:$15.00

Fig-13: Scatter Plot of Triglycerides by Platelets

4.6 **Scatter Plot Matrix**

The scatter plot matrix provides a comprehensive view of multiple attributes and their relationships within a dataset. Each plot in the matrix represents the pairwise relationship between two attributes, showcasing how variables interact and correlate across different combinations.

Here, the scatter plot matrix displays the pairwise relationships among key attributes in the cirrhosis dataset, including Age, Bilirubin, Cholesterol, Albumin, and Status. Each cell in the matrix represents a scatter plot comparing two specific attributes, enabling a comprehensive exploration of their correlations and distributions across different patient statuses.

A screenshot of a graph

Description automatically generated

Fig-14: Scatter Plot Matrix for Age, Bilirubin, Cholesterol, Albumin, and Status

Here, the scatter plot matrix illustrates the pairwise relationships among key attributes in the cirrhosis dataset, including Copper, Alk\_Phos, SGOT, Tryglicerides, Platelets, and Status.

A screenshot of a graph

Description automatically generated

Fig-15: Scatter Plot Matrix for Copper, Alk\_Phos, SGOT, Tryglicerides, Platelets, and Status

4.7 **Radar Chart**

By comparing the radar chart depicting attributes of cirrhosis patients, it becomes evident that certain characteristics vary with patient status. The radar chart reveals that age, bilirubin levels, cholesterol levels, albumin levels, and copper levels exhibit distinct patterns across different patient statuses. For instance, the mean values suggest that certain attributes may be more pronounced or concentrated in specific patient groups, providing insights into how these attributes correlate with varying health conditions among cirrhosis patients. In this plot, Albumin has the greatest number of mean values.

A diagram of a graph

Description automatically generated

Fig-16: Radar Chart for finding mean values of Age, Copper, Albumin, Cholesterol and Bilirubin

This radar chart below shows that stage attribute exhibits the highest number of mean values, indicating its prominence in the dataset and potential significance in understanding cirrhosis patient health conditions (histologic stage of disease specifically).

A diagram of a radiator

Description automatically generated

Fig-17: Radar Chart for finding mean values of Bilirubin, Stage, Prothrombin and Albumin

**5. DISCUSSION**

The utilization of data visualization techniques in medical research, particularly in the context of liver cirrhosis patient survival prediction, represents a significant advancement in understanding complex datasets and extracting actionable insights. This study employed a diverse array of visualization methods, including histograms, bar charts, box plots, violin plots, radar charts, scatter plots, and scatter plot matrices, to unravel the intricate relationships and patterns within the Cirrhosis Patient Survival Prediction dataset.

One of the primary strengths of data visualization lies in its ability to render complex datasets comprehensible and interpretable. Through univariate visualizations such as histograms and bar charts, this study provided insights into the distribution of key clinical indicators, offering valuable perspectives on patient health and disease severity. For instance, the distribution of serum albumin levels, urine copper levels, alkaline phosphatase activity, SGOT levels, triglyceride levels, platelet counts, and prothrombin times could be visualized, shedding light on the variability within the dataset and potential factors influencing patient outcomes.

Moreover, multivariate visualizations such as scatter plots and scatter plot matrices enabled the exploration of relationships between multiple variables simultaneously. By visually representing the interactions between different clinical and demographic attributes, these visualizations facilitated the identification of correlations and dependencies that may impact cirrhosis patient survival. For instance, scatter plots revealed relationships such as platelet counts with patient status, age with serum bilirubin levels, and platelet counts with triglyceride levels. These insights provide valuable understanding of the complex interactions between various factors, offering potential clues to disease progression and patient outcomes.

The incorporation of more advanced visualization techniques, such as radar charts and violin plots, further enhanced the depth of analysis by offering alternative perspectives on the data. Radar charts, for instance, allowed for the comparison of multiple variables across different patients, highlighting individual profiles and potential clusters within the dataset. Similarly, violin plots provided a visual representation of the distribution of a variable while simultaneously depicting summary statistics, offering a comprehensive view of the data distribution.

Overall, the integration of data visualization techniques in this study not only eased the exploration and understanding of the Cirrhosis Patient Survival Prediction dataset but also empowered more informed decision-making in clinical settings. By revealing patterns, relationships, and trends within the data, healthcare practitioners can gain deeper insights into the factors affecting cirrhosis patient survival rates and customize treatment approaches accordingly. Additionally, the findings from this analysis add to the collective knowledge in liver cirrhosis research, propelling advancements in medical science and ultimately enhancing patient care protocols.

**6. CONCLUSION**

In conclusion, this study emphasizes the vital role of data visualization in understanding complex medical data for liver cirrhosis patients. Through diverse visualization methods, we gained valuable insights into prognosis and disease progression. These findings advocate for informed decision-making in clinical practice and contribute to liver cirrhosis research. Integrating data visualization promises to advance patient care and innovate healthcare practices. This study demonstrates the transformative impact of data visualization in enhancing our understanding of diseases and improving patient outcomes.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Mayo Clinic for funding the creation of the dataset used in this study. Special thanks to E. Dickson, P. Grambsch, T. Fleming, L. Fisher, and A. Langworthy for their contributions to the dataset. Additionally, we acknowledge the UCI Machine Learning Repository for hosting and providing access to the dataset. This research would not have been possible without their efforts in data collection, curation, and dissemination.

REFERENCES

1. Kamath, Patrick S., et al. "A model to predict survival in patients with end–stage liver disease." *Hepatology* 33.2 (2001): 464-470.
2. Yang, Min, et al. "Models to predict the short-term survival of acute-on-chronic liver failure patients following liver transplantation." *BMC gastroenterology* 22.1 (2022): 80.
3. Kantidakis, Georgios, et al. "Survival prediction models since liver transplantation-comparisons between Cox models and machine learning techniques." *BMC medical research methodology* 20 (2020): 1-14.
4. Arabi, Yaseen, et al. "Outcome predictors of cirrhosis patients admitted to the intensive care unit." *European journal of gastroenterology & hepatology* 16.3 (2004): 333-339.
5. Guo, Aixia, et al. "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning." *PloS one* 16.8 (2021): e0256428.
6. Lin, Hai-rong, et al. "Development of a nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis." *Scientific Reports* 14.1 (2024): 9759-9759.
7. Campbell, Joseph, et al. "Validation and analysis of prognostic scoring systems for critically ill patients with cirrhosis admitted to ICU." *Critical Care* 19 (2015): 1-9.
8. Jepsen, Peter, Hendrik Vilstrup, and Per Kragh Andersen. "The clinical course of cirrhosis: the importance of multistate models and competing risks analysis." *Hepatology* 62.1 (2015): 292-302.
9. Yang, Ju Dong, et al. "Model to estimate survival in ambulatory patients with hepatocellular carcinoma." *Hepatology* 56.2 (2012): 614-621.
10. Xu, Fengshuo, et al. "A new scoring system for predicting in-hospital death in patients having liver cirrhosis with esophageal varices." *Frontiers in medicine* 8 (2021): 678646.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00