Method and EDA Result

Chen Liang

2024-12-13

Method

Data Description

The dataset used in this analysis originates from a study on primary biliary cirrhosis (PBC) conducted at the Mayo Clinic. It contains data for 276 patients, each characterized by 20 variables reflecting demographic, clinical, and laboratory features. Demographic variables include age and sex. Clinical features include the presence or absence of ascites, hepatomegaly, spiders, and edema, as well as the histologic stage of the disease. Laboratory markers, such as bilirubin, albumin, copper, alkaline phosphatase (alk_phos), SGOT, triglycerides, platelets, and prothrombin time, provide insight into liver function and disease severity. Outcome measures include the number of days from registration to death, liver transplantation, or censoring, and patient status (D: death, C: censored, CL: censored due to liver transplantation). The dataset underwent transformations to clean column names, adjust age to years, and harmonize categorical variables for consistency.

Missing data were handled by removing rows with incomplete values. Out of the original dataset, 142 entries were removed due to missing information, resulting in a final dataset of 276 complete cases. This approach ensures the integrity of statistical analysis by avoiding imputation biases. By excluding records with missing data, the analysis avoids biases introduced by imputation but acknowledges the trade-off between sample size and data quality.

As shown in Table 1, the baseline characteristics of patients reveal significant differences across survival outcomes. Patients who died had the shortest survival time, highest bilirubin, alkaline phosphatase, and SGOT levels, as well as the most advanced disease stage (50% in stage 4). In contrast, younger patients were more likely to undergo liver transplantation, with a mean age of 40.7 years compared to 53.4 years in the death group. Clinical features such as ascites, hepatomegaly, and edema were more prevalent among those who died or underwent transplantation, indicating disease severity. Additionally, lower albumin levels and prolonged prothrombin time in the death group highlight impaired liver function as a key prognostic factor.

Table 1: Baseline Characteristics

Characteristic	**Censored**, N = 147[1]	**Censored due to liver tx**, N = 18[1]	**Death**, N = 111[1]
N_days	2,391.8 / 2,224.0 (984.3)	1,511.6 / 1,368.0 (754.4)	1,508.5 / 1,191.0 (1,110.4)
Drug			
D-penicillamine	70 (48%)	9 (50%)	57 (51%)
Placebo	77 (52%)	9 (50%)	54 (49%)
Age	48.3 / 48.0 (10.3)	40.7 / 40.5 (6.0)	53.4 / 53.0 (10.0)

[1]Mean / Median (SD); n (%)

Table 1: Baseline Characteristics

Characteristic	**Censored**, N = 147[1]	**Censored due to liver tx**, N = 18[1]	**Death**, N = 111[1]
Sex			
Female	137 (93%)	15 (83%)	90 (81%)
Male	10 (6.8%)	3 (17%)	21 (19%)
Ascites	1 (0.7%)	0 (0%)	18 (16%)
Hepatomegaly	55 (37%)	12 (67%)	75 (68%)
Spiders	29 (20%)	5 (28%)	46 (41%)
Edema	8 (5.4%)	2 (11%)	32 (29%)
Bilirubin	1.6 / 0.9 (1.8)	3.2 / 3.3 (2.0)	5.7 / 3.3 (6.2)
Cholesterol	326.9 / 293.0 (168.1)	439.5 / 343.5 (335.5)	418.9 / 344.0 (277.9)
Albumin	3.6 / 3.6 (0.3)	3.6 / 3.6 (0.4)	3.4 / 3.4 (0.5)
Copper	68.1 / 52.0 (58.7)	123.3 / 101.0 (102.9)	140.3 / 121.0 (100.9)
Alk_phos	1,501.1 / 1,120.0 (1,376.8)	1,509.7 / 1,253.5 (854.4)	2,731.8 / 1,794.0 (2,765.3)
SGOT	110.2 / 97.0 (54.4)	130.2 / 123.5 (38.0)	141.5 / 134.9 (57.7)
Tryglicerides	111.1 / 103.0 (47.8)	133.9 / 124.0 (70.5)	141.8 / 124.0 (79.3)
Platelets	267.0 / 265.0 (86.4)	294.8 / 297.5 (79.9)	249.5 / 236.0 (102.1)
Prothrombin	10.4 / 10.2 (0.9)	10.4 / 10.2 (0.6)	11.2 / 11.0 (1.0)
Stage			
1	11 (7.5%)	0 (0%)	1 (0.9%)
2	42 (29%)	3 (17%)	14 (13%)
3	62 (42%)	8 (44%)	41 (37%)
4	32 (22%)	7 (39%)	55 (50%)

[1]Mean / Median (SD); n (%)

Results

EDA

Figure 1 provides insights into the distributions of continuous variables through boxplots, revealing heterogeneity in liver disease severity. Variables such as bilirubin, alkaline phosphatase, SGOT, and prothrombin exhibit highly skewed distributions with significant outliers, reflecting the heterogeneity in liver disease severity among patients. These patterns highlight the diversity in clinical markers and their potential implications for survival outcomes. Figure 2 focuses on categorical variables, showing that most patients are female, lack ascites, and are evenly split in terms of hepatomegaly presence. Most are in stages 2 and 3 of the disease, with a notable proportion in stage 4, indicating disease progression. Drug distribution is balanced between D-penicillamine and placebo groups, supporting comparability in treatment outcomes.

Finally, Figure 3 presents the correlation matrix, which highlights strong positive associations between biliru-

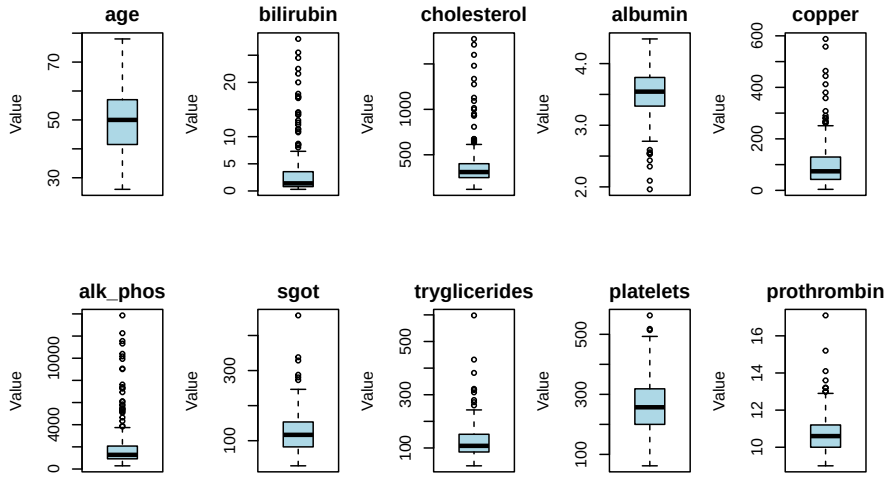

Figure 1: Boxplots for continuous variables

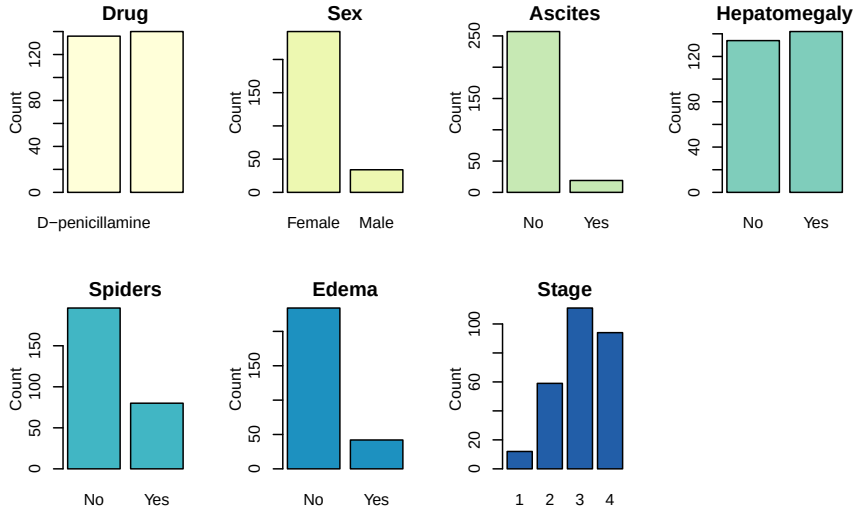

Figure 2: Barplots for categorical variables

bin, alkaline phosphatase, and SGOT, emphasizing their relationship with liver dysfunction. In contrast, albumin and platelet counts negatively correlate with disease stage, indicating their decline as the disease advances. These relationships highlight key biomarkers of cirrhosis progression.

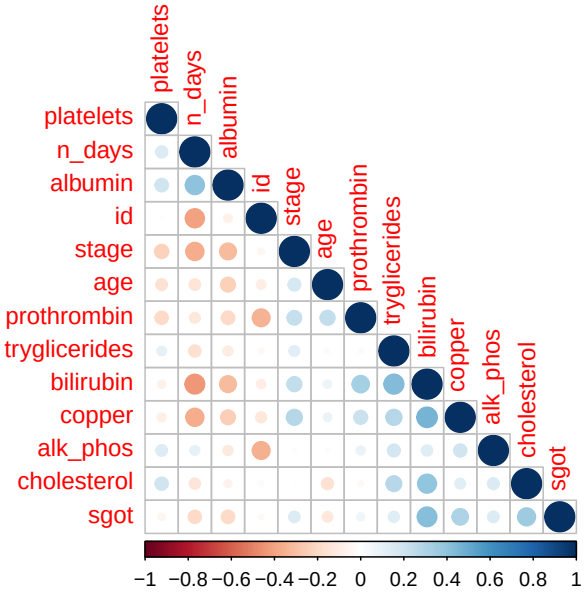

Figure 3: correlation matrix