# Cirrhosis Patient Survival Prediction

Chen Liang, Jessie Li, Xinyi Shang, Ixtaccihuatl Obregon, Jane Ma

## Background

### Cirrhosis Overview

### Related research

### Study Objective

## Methods

### Dataset 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, alsekaline 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. For the purpose of analysis, censored cases (C: censored, CL: censored due to liver transplantation) were grouped together as a single category (C), while death (D) remained as a separate outcome.

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) |
| 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) |

Table 1: Baseline Characteristics

| **Characteristic** | **Censored** N = 147[1] | **Censored due to liver tx** N = 18[1] | **Death** N = 111[1] |
|---|---|---|---|
| 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 (%)

## Survival Analysis

### Kaplan-Meier Estimates

The Kaplan-Meier estimator was used to estimate survival probabilities. This method provides a non-parametric estimation of the survival function $S(t)$, defined as the probability of survival beyond time $t$. For a set of ordered survival times $t_1, t_2, \ldots, t_k$, the Kaplan-Meier survival estimate is computed as:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where:

- $d_i$ is the number of events (deaths or censoring) at time $t_i$,
- $n_i$ is the number of individuals at risk just before time $t_i$.

The Kaplan-Meier survival curves were plotted for overall survival, as well as stratified by key variables such as treatment type (D-penicillamine vs. placebo), presence of edema, and histological stage of disease. Confidence intervals were calculated for survival probabilities to assess the precision of estimates.

**Log-Rank Test**

The log-rank test was employed to compare survival distributions between groups, such as those defined by drug treatment, edema status, and disease stage. The log-rank test statistic is computed as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

- $O_i$ is the observed number of events in group $i$,
- $E_i$ is the expected number of events in group $i$, under the null hypothesis.

This statistic follows a chi-squared distribution with degrees of freedom equal to the number of groups minus one. P-values were calculated to assess the significance of differences in survival.

**Cox PH**

(Feature selection, model assumptions, model evaluation)

We used a cox proportional hazards model to assess the effect of the covariates on the hazard of developing cirrhosis. The assumption of the model is that hazard ratios between groups is constant over time. In addition, the effects of the covariates on the hazard are assumed

to be proportional. To determine if the Cox PH model violated the proportional hazard assumptions we used the cox.zph() function in R.

After the proportional hazards assumption is met, interaction terms between model covariates are considered using the likelihood ratio test. Suppose there are $p$ covariates in the model. First, each one of the $\frac{p(p-1)}{2}$ interaction terms is added to the model to obtain $\frac{p(p-1)}{2}$ $p$ value, and the interaction term with the lowest $p$ value is added to the model. Then, the rest of the interaction terms are added to the new model one by one to obtain the $p$ value. The process is repeated until $p > 0.05$ holds for all the likelihood ratio tests. The model with the added interaction terms is the final model.

With the final model, model evaluation is conducted. Deviance residuals and Cox-Snell residuals are used to evaluate model fit. For models with good fit, deviance residuals should be randomly distributed around 0, and for KM survival estimates using Cox-Snell residuals as the pseudo survival time, $log(-log(S(t)))$ should be approximately equal to $log(t)$. Influence diagnostics with LD option is used to identify influential individuals. It evaluates how much the log-likelihood would change if the $i^{th}$ person was removed from the sample. After identifying outliers and influential individuals, these subjects are removed from the sample and the model is re-fit to see if the results would change.

# Results

## Descriptive Statistics (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.

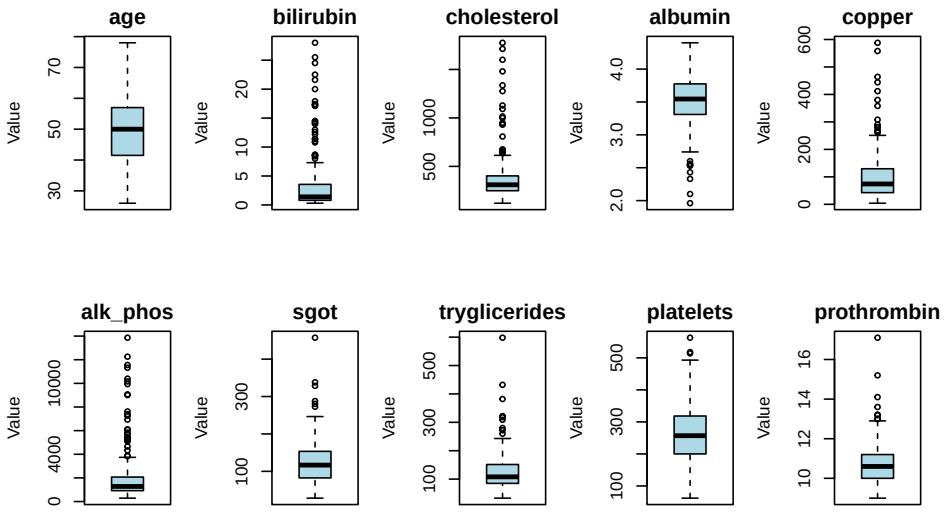


Figure 1: Boxplots for Continuous Variables

As shown in the Figure 2, the majority of patients are female, lack ascites, and are evenly distributed regarding hepatomegaly. 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.

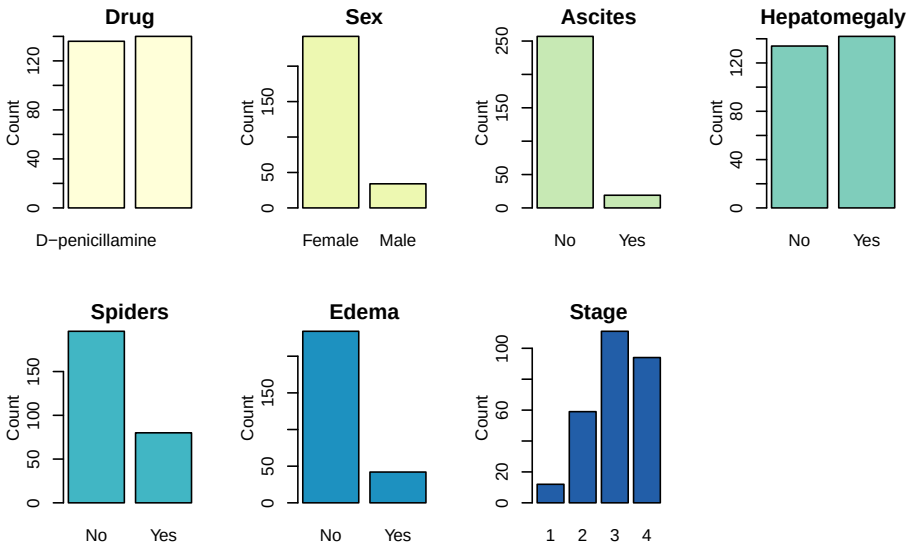


Figure 2: Barplots for Categorical Variables

Finally, Figure 3 presents the correlation matrix, which highlights strong positive associations between bilirubin, 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.

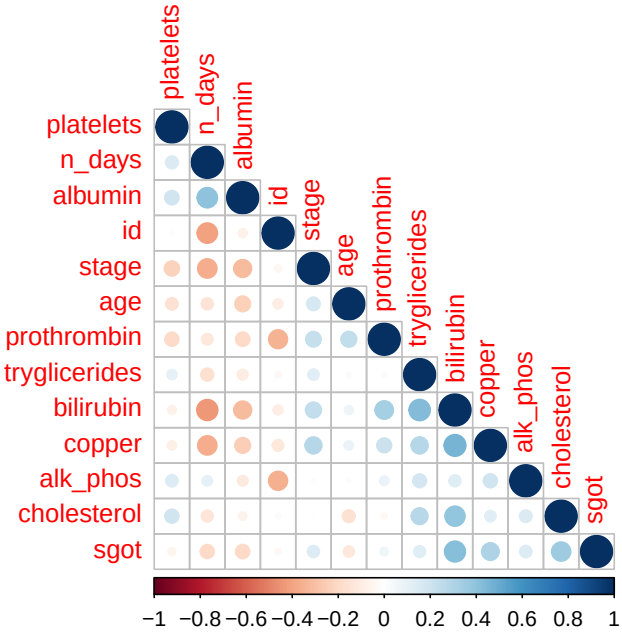


Figure 3: Correlation Matrix

## KM and Log-Rank Test

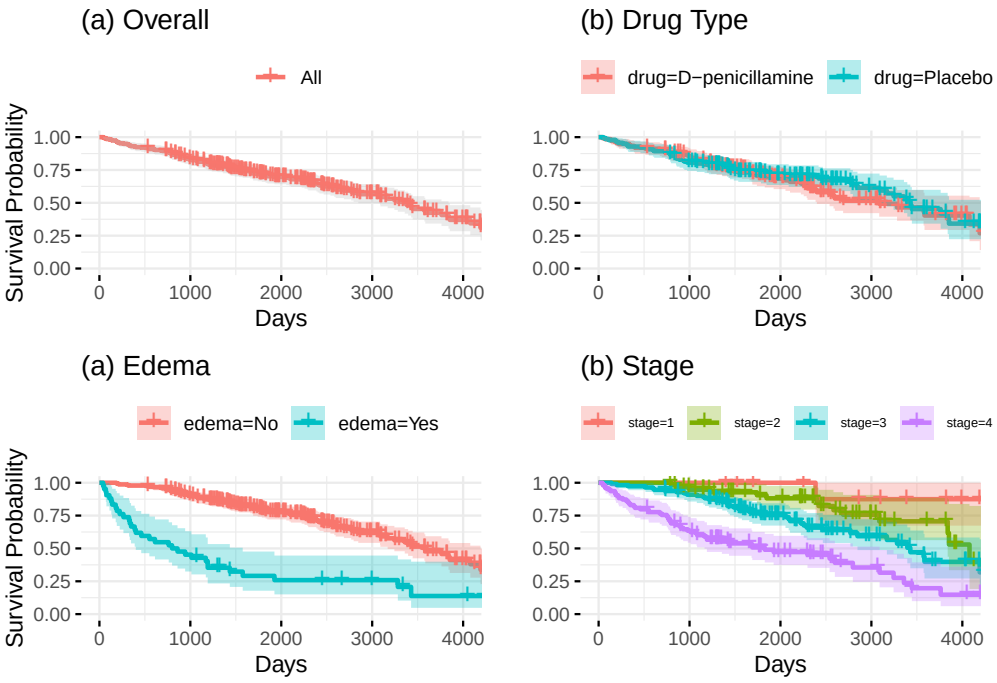


Figure 4: Kaplan-Meier Survival Curve

Table 2: Kaplan-Meier Survival Summary in Years

| Time Interval (Years) | At Risk | Events | Censored | Survival Probability | Lower CI | Upper CI |
|---|---|---|---|---|---|---|
| [0, 1) | 276 | 19 | 0 | 0.93 | 0.90 | 0.96 |
| [1, 2) | 257 | 10 | 1 | 0.89 | 0.86 | 0.93 |
| [2, 3) | 246 | 22 | 12 | 0.81 | 0.77 | 0.86 |
| [3, 4) | 212 | 14 | 29 | 0.76 | 0.71 | 0.81 |
| [4, 5) | 169 | 9 | 24 | 0.71 | 0.66 | 0.77 |
| [5, 6) | 136 | 6 | 18 | 0.68 | 0.62 | 0.74 |
| [6, 7) | 112 | 9 | 23 | 0.62 | 0.55 | 0.69 |
| [7, 8) | 80 | 6 | 15 | 0.57 | 0.50 | 0.64 |
| [8, 9) | 59 | 5 | 13 | 0.51 | 0.43 | 0.60 |
| [9, 10) | 41 | 6 | 8 | 0.42 | 0.34 | 0.53 |
| [10, 11) | 27 | 3 | 7 | 0.37 | 0.28 | 0.48 |
| [11, 12) | 17 | 2 | 10 | 0.31 | 0.21 | 0.45 |

The Kaplan-Meier survival analysis provides valuable insights into overall survival probabilities over time. The survival curve shows a consistent decline in survival probability Figure 4 (a), with a median survival probability of 9 years and a 75% survival probability of 4 years. Key survival probabilities at yearly intervals are summarized in Table 2.

Table 3: Log-Rank Test Results

| Chi-Squared Statistic | Degrees of Freedom | P-Value |
|---|---|---|
| **Drug** | | |
| 0.4049 | 1 | 0.5246 |
| **Edema** | | |
| 53.0933 | 1 | <0.0001 |
| **Stage** | | |
| 44.6499 | 3 | <0.0001 |

A comparison of survival probabilities between the two treatment groups—D-penicillamine and placebo—demonstrated no statistically significant differences. The log-rank test produced a p-value of 0.5246, far above the significance threshold of 0.05. The Kaplan-Meier curves for these groups overlap substantially, indicating that D-penicillamine does not significantly improve survival outcomes compared to placebo (Table 3, Figure 4 (b)).

Edema shows as a critical factor influencing survival, as highlighted by the log-rank test

Table 4: Multivariate Cox Proportional Hazards Analysis - Stepwise Selection Model

| **Characteristic** | **HR**[1] | **95% CI**[1] | **p-value** |
|---|---|---|---|
| drug | | | |
| D-penicillamine | — | — | |
| Placebo | 0.94 | 0.63 to 1.40 | 0.75 |
| age | 1.03 | 1.01 to 1.05 | 0.004 |
| edema | | | |
| No | — | — | |
| Yes | 1.47 | 0.88 to 2.47 | 0.14 |
| bilirubin | 1.09 | 1.05 to 1.13 | <0.001 |
| albumin | 0.47 | 0.28 to 0.82 | 0.007 |
| copper | 1.00 | 1.00 to 1.00 | 0.002 |
| sgot | 1.00 | 1.00 to 1.01 | 0.015 |
| prothrombin | 1.33 | 1.07 to 1.64 | 0.010 |
| stage | | | |
| 1 | — | — | |
| 2 | 3.88 | 0.47 to 32.1 | 0.21 |
| 3 | 5.29 | 0.68 to 41.1 | 0.11 |
| 4 | 8.02 | 1.04 to 61.8 | 0.046 |

[1]HR = Hazard Ratio, CI = Confidence Interval

(p-value $< 0.0001$). Patients without edema exhibit significantly better survival probabilities than those with edema (Table 3, Figure 4 (c)). The stark contrast in survival curves underscores edema as a key predictor of survival and its importance in risk stratification.

The histologic stage of the disease also plays a significant role in survival outcomes. The log-rank test for stage groups yielded a highly significant p-value ($< 0.0001$), indicating marked differences in survival curves across stages (Table 3, Figure 4 (d)). Patients in advanced stages (3 and 4) have substantially lower survival probabilities compared to those in earlier stages (1 and 2).

## Cox Model

(Feature selection, model assumptions, model evaluation)

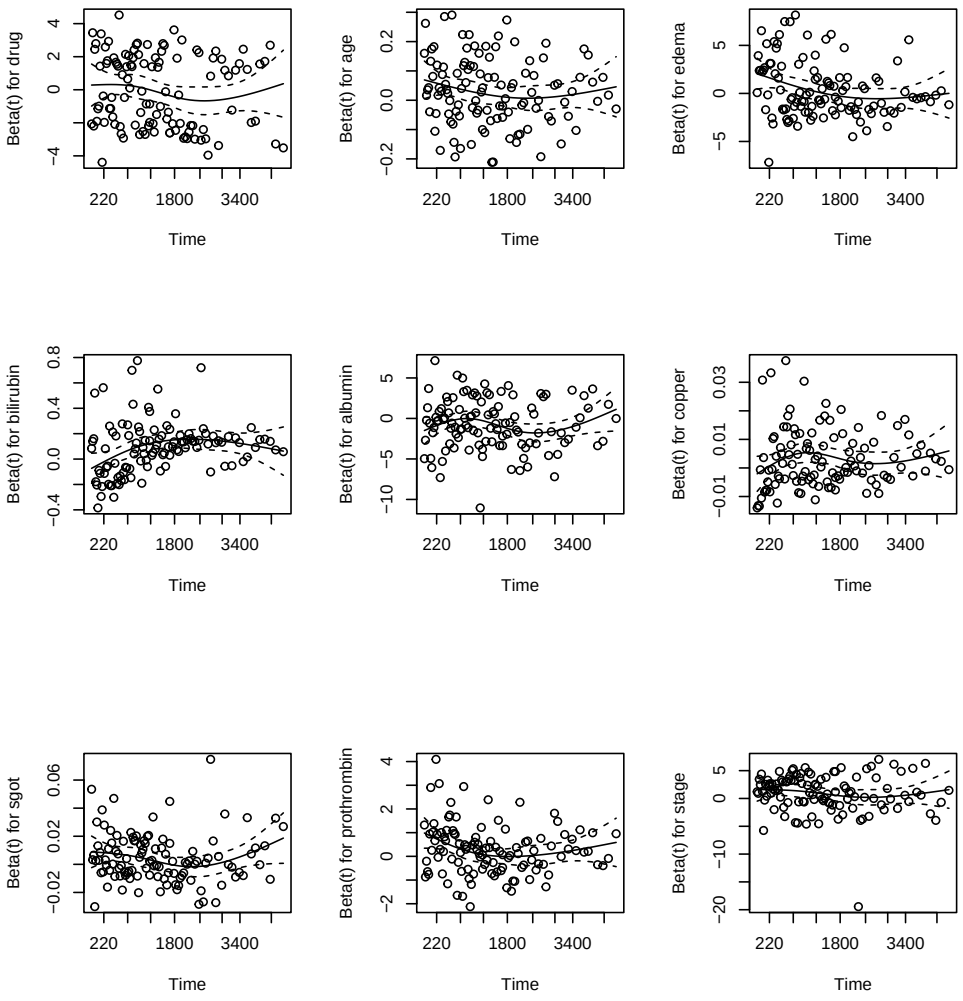


Table 5: Proportional Hazards Assumption Test for Cox PH Model - Stepwise Selection

| | chisq | df | p |
|---|---|---|---|
| drug | 0.1600772 | 1 | 0.6890854 |

Table 6: Multivariate Cox Proportional Hazards Analysis - Stratification of Edema Model

| **Characteristic** | **HR**[1] | **95% CI**[1] | **p-value** |
|---|---|---|---|
| drug | | | |
| D-penicillamine | — | — | |
| Placebo | 0.88 | 0.59 to 1.31 | 0.53 |
| age | 1.03 | 1.01 to 1.05 | 0.005 |
| bilirubin | 1.08 | 1.04 to 1.13 | <0.001 |
| albumin | 0.50 | 0.29 to 0.86 | 0.011 |
| copper | 1.00 | 1.00 to 1.00 | 0.004 |
| sgot | 1.00 | 1.00 to 1.01 | 0.033 |
| prothrombin | 1.36 | 1.09 to 1.70 | 0.006 |
| stage | | | |
| 1 | — | — | |
| 2 | 4.64 | 0.55 to 39.4 | 0.16 |
| 3 | 6.55 | 0.83 to 52.0 | 0.075 |
| 4 | 9.43 | 1.20 to 74.3 | 0.033 |

[1]HR = Hazard Ratio, CI = Confidence Interval

| | chisq | df | p |
|---|---|---|---|
| age | 2.6909476 | 1 | 0.1009198 |
| edema | 6.1134319 | 1 | 0.0134158 |
| bilirubin | 8.3071868 | 1 | 0.0039489 |
| albumin | 0.6258766 | 1 | 0.4288719 |
| copper | 0.1021024 | 1 | 0.7493211 |
| sgot | 1.3384725 | 1 | 0.2473035 |
| prothrombin | 5.0189196 | 1 | 0.0250718 |
| stage | 4.5185052 | 3 | 0.2106456 |
| GLOBAL | 24.6087203 | 11 | 0.0103973 |

We can see from Table **??** that edema(p=0.013), bilirubin (p=0.003), and prothrombin (p=0.025) violate the ph assumptions. To reduce bias, stratification was conducted.

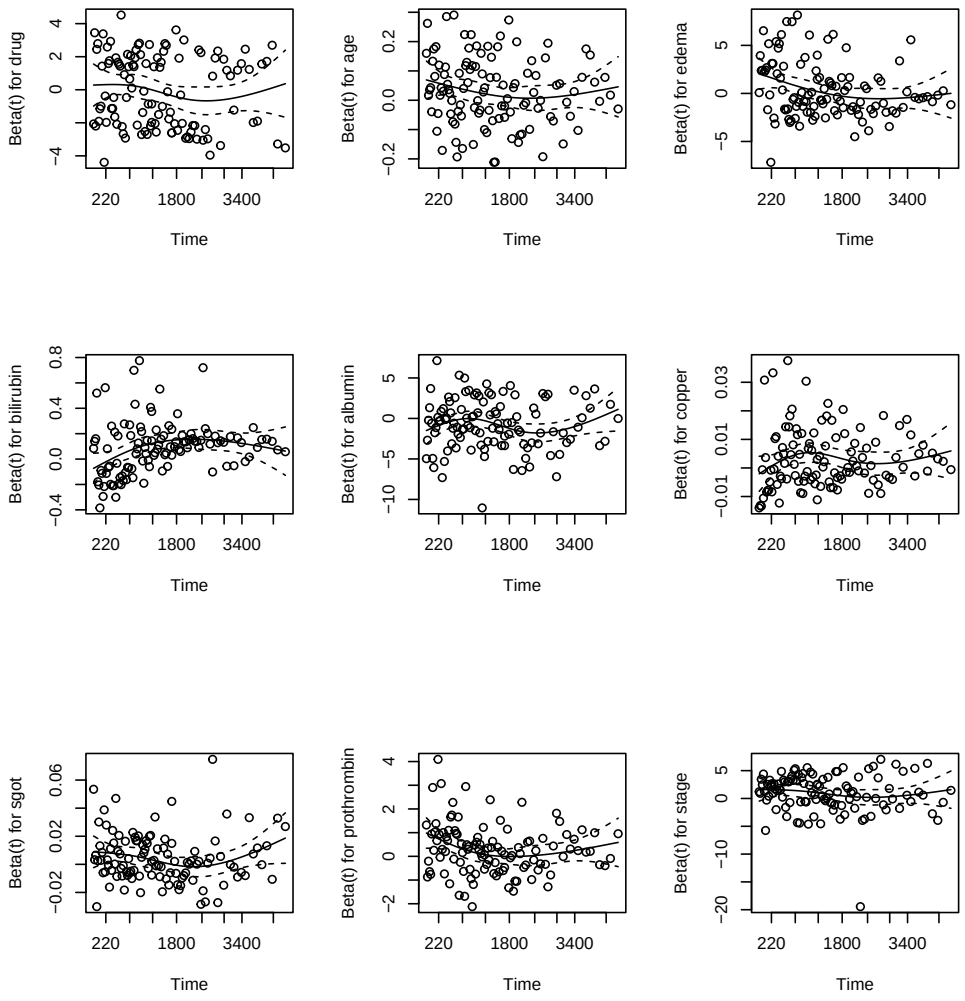


Table 7: Proportional Hazards Assumption Test COX PH Model Stratified for Edema

| | chisq | df | p |
|---|---|---|---|
| drug | 1.4344763 | 1 | 0.2310353 |

| | chisq | df | p |
|---|---|---|---|
| age | 1.8937409 | 1 | 0.1687806 |
| bilirubin | 11.5075689 | 1 | 0.0006931 |
| albumin | 0.0049180 | 1 | 0.9440915 |
| copper | 0.3914633 | 1 | 0.5315312 |
| sgot | 1.0484233 | 1 | 0.3058705 |
| prothrombin | 2.6117691 | 1 | 0.1060734 |
| stage | 3.0850010 | 3 | 0.3787045 |
| GLOBAL | 18.9514038 | 10 | 0.0408843 |

In Table **??**, after the straification of edema, bilirubin (p=0.0006) is the only variable violating PH assumptions. A time interactin is added to bilirubin to make the model appropriate.

Now, interaction terms are considered between the covariates in the model. During the first iteration, five interaction terms, including interaction for Copper with age, Albumin, SGOT, Prothrombin, and diease stage are selected using criteria $p < 0.05$, and the interaction between Albumin and Copper is added to the model as it has the lowest p value. During the second iteration, none of the interaction terms gets selected, and the iteration ends. The final model can then be specified as:

$$\begin{aligned} log(\text{HR}) =& \beta_1 I(\text{drug=D-penicillamine}) + \beta_2 age + \beta_3 \text{billirubin} + \beta_4 \text{albumin} + \beta_5 \text{copper} \\ &+ \beta_6 \text{sgot} + \beta_7 \text{prothrombin} + \beta_8 I(\text{stage=2}) + \beta_9 I(\text{stage=3}) + \beta_{10} I(\text{stage=4}) \\ &+ \beta_{11} \text{bilirubin : n\_days} + \beta_{12} \text{albumin : copper} \end{aligned}$$

Table 8 shows the model estimates. It can be concluded that, holding other covariates constant:

- The primary variable of interest, drug, has a negative yet insignificant impact on survival. Other covariates that impose a negative significant impact include age, Bilirubin, SGOT, Prothrombin, Stage 4 (compared with Stage 1), and interaction between Al-

bumin and Copper. Albumin, Copper, and interaction between Bilirubin and number of days instead have a protective significant impact.

- The hazard for PBC patients treated with D-penicillamine is 1.2720 times that of PBC patients treated with Placebo.
- The hazard ratio for PBC patients with one year increase in age is 1.0343.
- The hazard ratio for PBC patients with 1 mg/dl increase in Bilirubin is 1.2798.
- The hazard ratio for PBC patients with 1 gm/dl increase in Albumin is 0.2316.
- The hazard ratio for PBC patients with 1 ug/day increase in Copper is 0.9779.
- The hazard ratio for PBC patients with 1U/ml in SGOT is 1.0065.
- The hazard ratio for PBC patients with 1s increase in Prothrombin is 1.3257.
- The hazard for PBC patients at Stage 2, 3, and, 4 is 3.7034, 5.4604, and 8.0139 times that of PBC patients at Stage 1.
- For the same level of Bilirubin, a unit increase in survive time results in $-0.0002$Bilirubin change in the effect of Bilirubin on log hazard ratio for PBC patients.
- For the same level of Albumin, a unit increase in Copper results in 0.0076Albumin change in the effect of Albumin on log hazard ratio for PBC patients.

Model evaluation is then conducted on this final model. Figures 5 and 6 show the deviance residuals distribution and the KM estimates using Cox-Snell residuals as pseudo survival time respectively. As there is no obvious trend in the deviance plots and the line is close to the reference line in Cox-Snell plot, it can be concluded that the model is a relatively good fit.

For influence diagnostics, the individuals that provide the 5 largest absolute differences for the LD option are selected. After removing the outliers (identified by a deviance residual larger than 3, 2 are selected) and the 5 influential individuals, the model is re-fit. Table 9 compares the model estimates for the two models. There are subtle differences between model estimates, but the direction of impact stays the same for all the variables, thus resulting in similar conclusions.

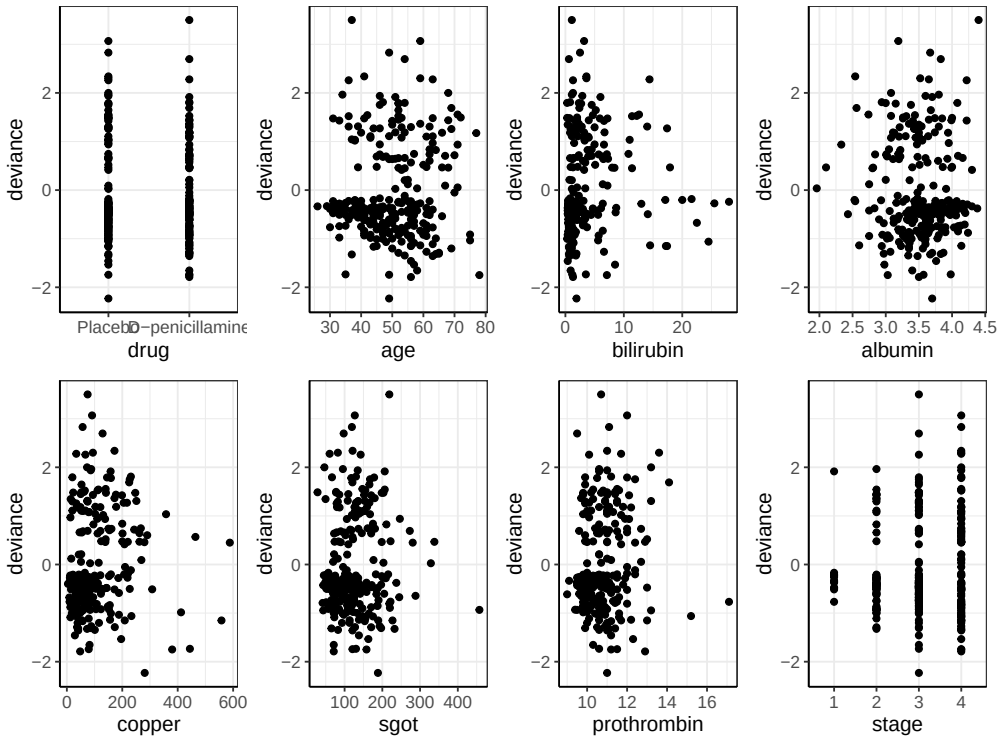


Figure 5: Deviance Residuals Scatterplot for Individual Variable

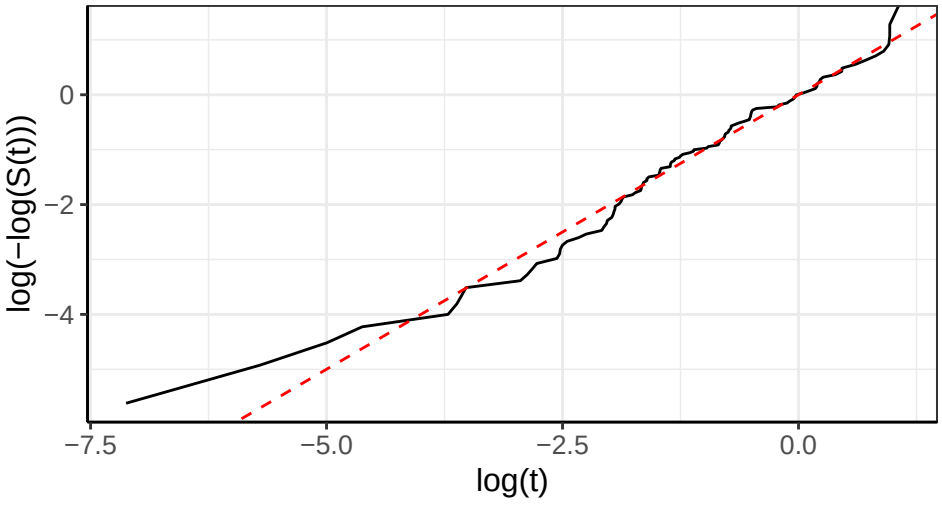


Figure 6: KM Estimates Using Cox-Snell Residuals

Table 8: Final Model Hazard Ratio Estimates

| **Characteristic** | **HR**[1] | **95% CI**[1] | **p-value** |
|---|---|---|---|
| drug | | | |
| drug.L | 1.2720 | 0.9451 to 1.7118 | 0.1124 |
| age | 1.0343 | 1.0119 to 1.0572 | 0.0025 |
| bilirubin | 1.2798 | 1.1927 to 1.3731 | 0.0000 |
| albumin | 0.2316 | 0.1089 to 0.4927 | 0.0001 |
| copper | 0.9779 | 0.9641 to 0.9919 | 0.0020 |
| sgot | 1.0065 | 1.0026 to 1.0104 | 0.0010 |
| prothrombin | 1.3257 | 1.0521 to 1.6704 | 0.0168 |
| stage | | | |
| 1 | — | — | |
| 2 | 3.7034 | 0.4513 to 30.392 | 0.2228 |
| 3 | 5.4604 | 0.6981 to 42.713 | 0.1058 |
| 4 | 8.0139 | 1.0305 to 62.324 | 0.0467 |
| bilirubin * n_days | 0.9998 | 0.9997 to 0.9999 | 0.0000 |
| albumin * copper | 1.0076 | 1.0034 to 1.0119 | 0.0004 |

[1]HR = Hazard Ratio, CI = Confidence Interval

Table 9: Model Parameter Estimates Comparison

| | Original Model | | | New Model | | |
|---|---|---|---|---|---|---|
| | Estimate | Hazard Ratio | p value | Estimate | Hazard Ratio | p value |
| drug.L | 0.2406 | 1.2720 | 0.1124 | 0.2907 | 1.3374 | 0.0662 |
| age | 0.0337 | 1.0343 | 0.0025 | 0.0303 | 1.0308 | 0.0098 |
| bilirubin | 0.2467 | 1.2798 | 0.0000 | 0.3033 | 1.3543 | 0.0000 |
| albumin | -1.4627 | 0.2316 | 0.0001 | -1.5639 | 0.2093 | 0.0001 |
| copper | -0.0224 | 0.9779 | 0.0020 | -0.0226 | 0.9777 | 0.0021 |
| sgot | 0.0065 | 1.0065 | 0.0010 | 0.0063 | 1.0063 | 0.0024 |
| prothrombin | 0.2819 | 1.3257 | 0.0168 | 0.3428 | 1.4089 | 0.0063 |
| stage2 | 1.3093 | 3.7034 | 0.2228 | 1.4232 | 4.1505 | 0.1921 |
| stage3 | 1.6975 | 5.4604 | 0.1058 | 1.7935 | 6.0104 | 0.0930 |
| stage4 | 2.0812 | 8.0139 | 0.0467 | 2.1404 | 8.5030 | 0.0440 |
| bilirubin:n_days | -0.0002 | 0.9998 | 0.0000 | -0.0002 | 0.9998 | 0.0000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| albumin:copper | 0.0076 | 1.0076 | 0.0004 | 0.0078 | 1.0078 | 0.0003 |

# Discussion

The analysis demonstrates that D-penicillamine is inefficient in improving survival outcomes for patients with primary biliary cirrhosis (PBC). The hazard ratio for those treated with the drug is 1.2720 compared to placebo, indicating no survival benefit and a potential negative effect. This finding suggests the need to reconsider its use and focus on alternative therapies that may offer greater efficacy.

Kaplan-Meier analysis further confirmed the inefficacy of D-penicillamine, as survival probabilities between the drug and placebo groups showed no statistically significant difference. The overlapping survival curves reinforce the finding that D-penicillamine does not confer a survival advantage and may require reevaluation as a treatment option. Additionally, Kaplan-Meier analysis highlighted the importance of edema, as patients without edema showed significantly better survival probabilities, making it a critical factor in risk stratification.

Age and disease stage emerged as critical determinants of survival, underscoring the importance of early detection and stage-specific care. The hazard of mortality increases significantly with each advancing stage, with Stage 4 patients facing an eightfold higher risk compared to Stage 1. Kaplan-Meier survival curves demonstrated markedly reduced survival probabilities in advanced stages, with a sharp decline observed in stages 3 and 4. Early interventions to halt disease progression are vital, as is tailoring treatment strategies to the patient's disease stage.

Liver function indicators such as Bilirubin, Copper, SGOT, Prothrombin, and Albumin are vital in survival outcomes. Elevated Bilirubin, SGOT, and Prothrombin levels are associated with higher hazards, reflecting liver damage and dysfunction. Conversely, higher albumin levels provide a strong protective effect, emphasizing the importance of maintaining good nutritional and synthetic liver function. The modest protective impact of Copper is proved

by previous studies where Copper deficiency is identified as a risk factor for mortality in advanced liver disease (Yu et al. 2019). These findings highlight the necessity of monitoring liver function and metabolic health closely to identify high-risk patients and address reversible factors.

The analysis also revealed key interactions affecting survival. A negative Albumin-Copper interaction was observed, indicating a synergistic detrimental effect. This underscores the complexity of metabolic interactions in liver disease and the need to address deficiencies or toxicities in a balanced manner. Furthermore, continued investigation of other potential interactions is needed to uncover personalized treatment strategies.

To improve patient outcomes, it is crucial to monitor high-risk patients routinely, focusing on those with poor liver function indicators or advanced disease stages. Additionally, conducting biological investigations into the protective role of copper and its interactions with other variables could provide valuable insights. These efforts can pave the way for personalized therapies that address the unique risk profiles of individual patients and enhance survival outcomes.

There were several limitations within the project including missing data, imbalanced data, and a high censoring rate. Regarding missing data, 147 observation had missing values, which may require the application of imputations techniques to address potential bias. Furthermore, the data imbalance was attributed to the distribution of sex, where we had about 80-90% female participants and the right-skwedness of bilirubin. Lastly, with more than 50% of data being censored, high censoring data was a major limitation on the survival analysis and the robustness of the results.

# Conclusion

# References

Yu, L., I. W. Liou, S. W. Biggins, M. Yeh, F. Jalikis, L. N. Chan, and J. Burkhead. 2019. "Copper Deficiency in Liver Diseases: A Case Series and Pathophysiological Considerations." *Hepatology Communications* 3 (8): 1159–65. https://doi.org/10.1002/hep4.1393.

# Appendix

## Code

```
knitr::opts_chunk$set(echo = FALSE, message = F, warning = F, out.width = "80%", fig.ali
options(knitr.kable.NA = '')
library(tidyverse)
library(RColorBrewer)
library(corrplot)
library(gtsummary)
library(flextable)
library(stringr)
library(survival)
library(survminer)
library(kableExtra)
library(ggplot2)
library(ggpubr)
library(gridExtra)

write_matex <- function(x) {
  begin <- "$$\\begin{bmatrix}"
  end <- "\\end{bmatrix}$$"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  writeLines(c(begin, X, end))
}
theme_set(
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )
)
cirrhosis <- read_csv("data/cirrhosis.csv")|>
  janitor::clean_names() |>
```

```
  mutate(age = round(age / 365),
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes"))|>
  na.omit()
theme_gtsummary_journal(journal = "nejm")

cirrhosis_df <- cirrhosis |>
  mutate(
    status = case_when(
      status == "C" ~ "Censored",
      status == "CL" ~ "Censored due to liver tx",
      status == "D" ~ "Death",
      TRUE ~ status))

table_1 <- cirrhosis_df |>
  select(-id) |>
  tbl_summary(
    by = status,
    statistic = list(
      all_continuous() ~ "{mean} / {median} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    digits = all_continuous() ~ 1,
    missing = "no",
    label = list(
    n_days ~ "N_days",
    drug ~ "Drug",
    age ~ "Age",
    sex ~ "Sex",
    ascites ~ "Ascites",
    hepatomegaly ~ "Hepatomegaly",
    spiders ~ "Spiders",
    edema ~ "Edema",
    bilirubin ~ "Bilirubin",
    cholesterol ~ "Cholesterol",
    albumin ~ "Albumin",
    copper ~ "Copper",
    alk_phos ~ "Alk_phos",
    sgot ~ "SGOT",
    tryglicerides ~ "Tryglicerides",
    platelets ~ "Platelets",
```

```
    prothrombin ~ "Prothrombin",
    stage ~ "Stage"
  )) |>
  modify_caption("Baseline Characteristics") |>
  as_flex_table() |>
  set_table_properties(width = 0.8, layout = "autofit") # Set width to 80%

table_1
conti_vars = cirrhosis |>
  select(age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, tryglicerides, pl

# Boxplot for all continuous variables
par(mfrow = c(2, 5), oma = c(2, 2, 3, 1), mar = c(4, 4, 2, 1))
conti_names <- names(conti_vars)

p1 <- for (i in seq_along(conti_names)) {
  boxplot(conti_vars[[conti_names[i]]],
          main = conti_names[i],
          ylab = "Value",
          col = "lightblue",
          outline = TRUE)  # Show outliers
}
cate_vars = cirrhosis |>
  select(drug, sex, ascites, hepatomegaly, spiders, edema, stage)

par(mfrow = c(2, 4),  # 2 rows, 5 columns
    oma = c(2, 2, 3, 1),  # Outer margins
    mar = c(4, 4, 2, 1),  # Inner margins for individual plots
    mgp = c(2, 1, 0))     # Margins for axis labels and titles

colors <- c(brewer.pal(9, "YlGnBu"), "darkblue")

barplot(table(cate_vars$drug), main = "Drug", ylab = "Count",col = colors[1])
barplot(table(cate_vars$sex), main = "Sex", ylab = "Count",col = colors[2])
barplot(table(cate_vars$ascites), main = "Ascites", ylab = "Count", col = colors[3])
barplot(table(cate_vars$hepatomegaly), main = "Hepatomegaly", ylab = "Count", col = colo
barplot(table(cate_vars$spiders), main = "Spiders", ylab = "Count", col = colors[5])
barplot(table(cate_vars$edema), main = "Edema", ylab = "Count", col = colors[6])
barplot(table(cate_vars$stage), main = "Stage", ylab = "Count", col = colors[7])
numeric_cirr <- cirrhosis |>
  select_if(is.numeric)

cor_matrix <- cor(numeric_cirr, use = "complete.obs")
```

```r
corrplot(cor_matrix, method = "circle", type = "lower", order = "hclust")
cirrhosis <- read_csv("../data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365),
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes")) |>
  drop_na()


cirrhosis$event <- ifelse(cirrhosis$status == "D", 1, 0)

surv_object <- Surv(time = cirrhosis$n_days, event = cirrhosis$event)

km_fit <- survfit(surv_object ~ 1, data = cirrhosis)

plot_all = ggsurvplot(km_fit, conf.int = TRUE,
           title = "(a) Overall",
           xlab = "Days", ylab = "Survival Probability",
           legend.title = "",
           ggtheme = theme_minimal())

km_fit_drug <- survfit(surv_object ~ drug, data = cirrhosis)

plot_drug = ggsurvplot(km_fit_drug, conf.int = TRUE,
           title = "(b) Drug Type",
           xlab = "Days", ylab = " ",
           legend.title = "",
           ggtheme = theme_minimal() )


# Fit survival curves by edema
km_fit_edema <- survfit(surv_object ~ edema, data = cirrhosis)
plot_edema <- ggsurvplot(
  km_fit_edema, conf.int = TRUE,
  title = "(a) Edema",
  xlab = "Days", ylab = "Survival Probability",
  legend.title = "",
  ggtheme = theme_minimal()
)

# Fit survival curves by stage
```

```r
km_fit_stage <- survfit(surv_object ~ stage, data = cirrhosis)
plot_stage <- ggsurvplot(
  km_fit_stage, conf.int = TRUE,
  title = "(b) Stage",
  xlab = "Days", ylab = " ",
  legend.title = "",
  ggtheme = theme_minimal() +
    theme(legend.text = element_text(size = 6)) # Adjust legend text size

)

# Arrange the plots side by side with adjusted spacing
grid.arrange(
  plot_all$plot,
  plot_drug$plot,
  plot_edema$plot,
  plot_stage$plot,
  ncol = 2,
  widths = c(1, 1) # Equal sizing for both plots
)

max_time <- max(cirrhosis$n_days, na.rm = TRUE)
max_years <- floor(max_time / 365)
yearly_times <- seq(0, max_years * 365, by = 365)

km_summary_yearly <- summary(km_fit, times = yearly_times)

# Create the data frame from the KM summary
surv_yearly_table <- data.frame(
  years = yearly_times / 365,
  n_risk = km_summary_yearly$n.risk,
  n_event = km_summary_yearly$n.event,
  n_censor = km_summary_yearly$n.censor,
  survival = km_summary_yearly$surv,
  lower_ci = km_summary_yearly$lower,
  upper_ci = km_summary_yearly$upper
)

# If time=0 row does not exist, add it
if (!any(yearly_times == 0)) {
  surv_yearly_table <- rbind(
    data.frame(
      years = 0,
      n_risk = km_fit$n.risk[1],
```

```
      n_event = 0,
      n_censor = 0,
      survival = 1,
      lower_ci = 1,
      upper_ci = 1
    ),
    surv_yearly_table
  )
}

surv_yearly_table <- surv_yearly_table[order(surv_yearly_table$years), ]

interval_labels <- sapply(2:nrow(surv_yearly_table), function(i) {
  paste0("[", surv_yearly_table$years[i-1], ", ", surv_yearly_table$years[i], ")")
  })

surv_yearly_intervals <- surv_yearly_table[-1, ] # Remove the first row if needed

surv_yearly_intervals$interval <- interval_labels

surv_yearly_intervals$n_risk[1] <- surv_yearly_table$n_risk[1]

for (i in 2:nrow(surv_yearly_intervals)) {
  surv_yearly_intervals$n_risk[i] <- surv_yearly_intervals$n_risk[i-1] -
                                     surv_yearly_intervals$n_event[i-1] -
                                     surv_yearly_intervals$n_censor[i-1]
}

surv_table = surv_yearly_intervals %>%
  rownames_to_column() %>%          # Convert any existing row names to a column
  select(-rowname) %>%              # Remove the converted row names column
  select(interval, n_risk, n_event, n_censor, survival, lower_ci, upper_ci)|>
  as.data.frame()
kable(
  surv_table,
  caption = "Kaplan-Meier Survival Summary in Years",
      col.names = c("Time Interval (Years)", "At Risk", "Events", "Censored", "Survival
  digits = 2,
  booktabs = TRUE
) %>%
  kable_styling(full_width = FALSE)
log_rank_test <- survdiff(surv_object ~ drug, data = cirrhosis)

log_rank_results_drug <- data.frame(
```

```
  Group = "Drug",
  Statistic = log_rank_test$chisq,
  Degrees_of_Freedom = 1,
  P_Value = log_rank_test$pvalue
)

log_rank_results_drug[, -1] <- log_rank_results_drug[, -1] %>%
  mutate(across(where(is.numeric), ~ round(., 4)))


# Log-rank test for edema
log_rank_test_edema <- survdiff(surv_object ~ edema, data = cirrhosis)
log_rank_results_edema <- data.frame(
  Group = "Edema",
  Statistic = log_rank_test_edema$chisq,
  Degrees_of_Freedom = 1,
  P_Value = ifelse(log_rank_test_edema$pvalue < 0.0001, "<0.0001", log_rank_test_edema$p
)

# Log-rank test for stage
log_rank_test_stage <- survdiff(surv_object ~ stage, data = cirrhosis)
log_rank_results_stage <- data.frame(
  Group = "Stage",
  Statistic = log_rank_test_stage$chisq,
  Degrees_of_Freedom = 3,
  P_Value = ifelse(log_rank_test_stage$pvalue < 0.0001, "<0.0001", log_rank_test_stage$p
)

log_rank_results_combined <- rbind(log_rank_results_drug,
                                   log_rank_results_edema,
                                   log_rank_results_stage)

kable(
  log_rank_results_combined[, -1], # Exclude the "Group" column for main table
  digits = 4,
  col.names = c("Chi-Squared Statistic", "Degrees of Freedom", "P-Value"),
  caption = "Log-Rank Test Results"
) %>%
  pack_rows("Drug", 1, 1) |>
  pack_rows("Edema", 2, 2) |>
  pack_rows("Stage", 3, 3)
cirrhosis = read_csv("data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365),
```

```
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes")) |>
  na.omit()
cirrhosis = cirrhosis |>
  mutate(
    status = case_when(
      status == "D" ~ 1,  # Event of interest (death)
      status == "C" | status == "CL" ~ 0,  # Censored data
      TRUE ~ as.numeric(status)),
      stage = factor(stage)  # Convert 'stage' to a factor
)

# cox model based on stepwise selection varibales above (ixta)
cox_model_a = coxph(Surv(n_days, status) ~ drug + age + edema +
                          bilirubin + albumin + copper + sgot +
                          prothrombin + stage,
                    id=id,
                   data = cirrhosis)

# Summarize the results
cox_summary_a = tbl_regression(cox_model_a, exponentiate = TRUE) |>
  modify_caption("Multivariate Cox Proportional Hazards Analysis - Stepwise Selection Mo
cox_summary_a
ph_assumption_a = cox.zph(cox_model_a)
par(mfrow = c(2, 3))
plot(ph_assumption_a)
ph_assumption_df = as.data.frame(ph_assumption_a$table)
knitr::kable(ph_assumption_df, caption = "Proportional Hazards Assumption Test for Cox P
cox_model_edema_strat = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                                bilirubin + albumin + copper + sgot +
                                prothrombin + stage,
                                id = id,
                                data = cirrhosis)
# Summarize the results
cox_summary_b = tbl_regression(cox_model_edema_strat, exponentiate = TRUE) |>
  modify_caption("Multivariate Cox Proportional Hazards Analysis - Stratification  of Ed
cox_summary_b
ph_assumption_b = cox.zph(cox_model_edema_strat)
par(mfrow = c(2, 3))
plot(ph_assumption_a)
ph_assumption_edema_strat = cox.zph(cox_model_edema_strat)
```

```r
ph_assumption_edema = as.data.frame(ph_assumption_edema_strat$table)
knitr::kable(ph_assumption_edema, caption = "Proportional Hazards Assumption Test COX PH
cirrhosis = read_csv("../data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365),
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes"),
         stage = factor(stage),
         drug = factor(drug, levels = c("Placebo", "D-penicillamine"), order = T)) |>
  na.omit()
cirrhosis = cirrhosis |>
  mutate(
    status = case_when(
      status == "D" ~ 1,  # Event of interest (death)
      status == "C" | status == "CL" ~ 0,  # Censored data
      TRUE ~ as.numeric(status)))

# Interaction between Convariates
cox_init = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                                  bilirubin + albumin + copper + sgot +
                                  prothrombin + stage + bilirubin : n_days,
                           id = id,
                          data = cirrhosis |> na.omit())
variables = c("drug", "age", "albumin", "copper", "sgot",
              "prothrombin", "stage")
vars_df = tibble()
for(var in variables[1 : (length(variables) - 1)])
{
  left_vars = variables[(which(variables == var) + 1) : length(variables)]
  for(var2 in left_vars)
  {
    cox_fit = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                                  bilirubin + albumin + copper + sgot +
                                  prothrombin + stage + bilirubin : n_days +
                        eval(parse(text = var2)) : eval(parse(text = var)),
                           id = id,
                          data = cirrhosis |> na.omit())
     # aic_vec= c(aic_vec, AIC(model_four))
     chisq_stat=-2 * (logLik(cox_init)-logLik(cox_fit))
     p_val = 1 - pchisq(chisq_stat,
                        attr(logLik(cox_fit), "df") -
```

```
                          attr(logLik(cox_init),"df"))
    if(p_val < 0.05)
    {
      vars_df = vars_df |> rbind(c(round(p_val, 4), var, var2))
    }
  }
}

colnames(vars_df) = c("p_value", "variable1", "variable2")
# vars_df |>
#   mutate(interaction = paste0(variable1, " * ", variable2)) |>
#   select(interaction, p_value) |>
#   knitr::kable(col.names = c("Interaction Term", "P Value"),
#     caption = "Siginificant Interaction term")

# We first add the albumin*copper term into the model and evaluate again.

cox_fit2 = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                   bilirubin + albumin + copper + sgot + prothrombin + stage +
                   bilirubin : n_days + albumin * copper,
                 id = id, data = cirrhosis)
vars_df = tibble()
for(var in variables[1 : (length(variables) - 1)])
{
  left_vars = variables[(which(variables == var) + 1) : length(variables)]
  for(var2 in left_vars)
  {
    cox_fit = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                      bilirubin + albumin + copper + sgot + prothrombin +
                      stage + bilirubin : n_days + albumin * copper +
                      eval(parse(text = var2)) : eval(parse(text = var)),
                         id = id,
                        data = cirrhosis)
    # aic_vec= c(aic_vec, AIC(model_four))
    chisq_stat=-2 * (logLik(cox_fit2)-logLik(cox_fit))
    p_val = 1 - pchisq(chisq_stat,
                       attr(logLik(cox_fit), "df") -
                         attr(logLik(cox_fit2),"df"))
    if(p_val < 0.05)
    {
      vars_df = vars_df |> rbind(c(round(p_val, 4), var, var2))
    }
  }
}
```

```r
# This is our final model.
cox_final = cox_fit2
# summary(cox_final)$coefficient %>% .[, c(1, 2, 5)] |>
#   data.frame() |> mutate(significance = c("", "**", "***", "***", "**", "***", "*",
#                                           "", "", "*", "***", "***")) |>
#   knitr::kable(col.names = c(" ", "Estimate", "Hazard Ratio", "p value", "Sig."),
#                digits = 4, caption = "Final Model Parameter Results")
cox_final |> tbl_regression(
  exponentiate = T,
  estimate_fun = purrr::partial(style_ratio, digits = 4),
  pvalue_fun = purrr::partial(style_sigfig, digits = 4)) |>
  modify_caption("Final Model Hazard Ratio Estimates")
deviance_res = residuals(cox_final, type = "deviance", var = stage)

dev_drug = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = drug, y = deviance)) +
  geom_point()
dev_age = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = age, y = deviance)) +
  geom_point()
dev_bili = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = bilirubin, y = deviance)) +
  geom_point()
dev_albu = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = albumin, y = deviance)) +
  geom_point()
dev_copper = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = copper, y = deviance)) +
  geom_point()
dev_sgot = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = sgot, y = deviance)) +
  geom_point()
dev_proth = cirrhosis |>
  mutate(deviance = deviance_res) |>
  ggplot(aes(x = prothrombin, y = deviance)) +
  geom_point()
dev_stage = cirrhosis |>
  mutate(deviance = deviance_res) |>
```

```
  ggplot(aes(x = stage, y = deviance)) +
  geom_point()

ggarrange(dev_drug, dev_age, dev_bili, dev_albu, dev_copper,
          dev_sgot, dev_proth, dev_stage, ncol = 4, nrow = 2)

# plot(deviance_res, ylab = "Deviance Residuals", xlab = "Index",
#      main = "Deviance Residuals Scatterplot")
# abline(h = c(-3, 3), col = "red", lty = 2) # Flag large residuals
# which(deviance_res > 3)
coxsnell_res = - (predict(cox_final, type = "survival") |> log())
# hist(coxsnell_res, main = "Cox-Snell Residuals Histogram", freq = F, breaks = 15)
# curve(exp(- x), add = T, col = "red")
# plot(coxsnell_res, ylab = "Cox-Snell Residuals", xlab = "Index",
#      main = "Cox-Snell Residuals Scatterplot")
km_fit = cirrhosis |> mutate(pseudo_time = coxsnell_res) |>
  survfit(Surv(pseudo_time, status) ~ 1, id = id, data = _)
km_summary = summary(km_fit)
tibble(
  t = km_summary$time,
  survival = km_summary$surv
) |>
  mutate(y = log(- log(survival))) |>
  ggplot(aes(x = log(t), y = y)) +
  geom_line() +
  geom_abline(intercept = 0, slope = 1, color = "red", lty = 2) +
  labs(y = "log(-log(S(t)))", title = "")
ld_res = c()
for(i in 1 : nrow(cirrhosis))
{
  dat = cirrhosis |> slice(- i)
  model_ld = coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                   bilirubin + albumin + copper + sgot + prothrombin + stage +
                   bilirubin : n_days + albumin * copper,
                 id = id, data = dat)
  ld_res = c(ld_res, 2 * abs(logLik(model_ld) - logLik(cox_final)))
}
cox_after = cirrhosis |>
  slice(c(- 77, - 143, - 82, - 100, - 108, - 129, - 210)) |>
  coxph(Surv(n_days, status) ~ drug + age + strata(edema) +
                   bilirubin + albumin + copper + sgot + prothrombin + stage +
                   bilirubin : n_days + albumin * copper,
                 id = id, data = _)
summary(cox_final)$coefficient %>% .[, c(1, 2, 5)] |>
```

```r
cbind(summary(cox_after)$coefficient %>% .[, c(1, 2, 5)]) |>
knitr::kable(col.names = c(" ", rep(c("Estimate", "Hazard Ratio", "p value"), 2)),
             digits = 4, caption = "Model Parameter Estimates Comparison") |>
add_header_above(header = c(" " = 1, "Original Model" = 3, "New Model" = 3))
```