

# Report 2: Survival Analysis of Primary Biliary Cirrhosis

Guanqun LIU

2022-05-26

## 1. Introduction

Primary biliary cirrhosis is a relatively rare autoimmune liver disease that leading slow but progressive destruction of small bile ducts, which will result in permanent cirrhosis and liver decompensation. Patients may also suffer abdominal pain, diarrhea, and an increased risk of cancer. The dataset is originally collected from the Mayo Clinic trial of primary biliary cirrhosis (PBC) of the liver from 1974 to 1984. 424 PBC patients met the eligibility criteria for a randomized placebo-controlled trial of the drug D-penicillamine (DPCA). The first 312 data entries represent participants of the above trial and their data is mostly complete. The rest 112 patients declined to participate but agreed to record basic measurements and track survival. 6 of these cases were discarded due to track lost shortly after diagnosis. The purpose of this analysis is to investigate the effect of DPCA on the lifetime of patients with PBC, along with what covariates significantly influence patients' risk of death.

## 2. EDA

The dataset contains 418 data entries, corresponding to 312 trial participants and 106 additional cases in order. Each entry has 3 basic columns recording their survival time in days, status at endpoint (alive, transplant, dead), and treatments by DPCA/placebo. Besides, each entry has 15 covariates describing a patient's personal information (age, sex), clinical assay indexes, and other complications. Regarding all variable columns, the dataset contains 276 complete cases without missing values among 312 formal participants. The sex ratio of the dataset is at least 9 : 1 (female to male) and 125 of the 312 formal participants died at the endpoint and the rest were alive (censored/transplant). To formalize the analysis, both censored and transplant status are considered alive. The defined variable table is shown in Appx.1 and the following analysis will be conducted on the 276 complete cases as stated above.

### 2.1 Univariate Analysis

#### Quantitative Variables

Combining the results in Table 1 and Figure 1, we can clearly see that covariates *bili*, *chol*, *copper*, *alk.phos*, *ast*, *trig*, and *protime* have typical right-skewed distributions, which I will apply the log transformation to them to compensate for a normal distribution in survival

analysis. The median survival time and patient ages are 1788 days after registration and 49.71 years respectively.

	time	age	bili	chol	albumin	copper	alk.phos	ast	trig	platelet	protime
mean	1979.167	49.800	3.334	371.261	3.517	100.768	1996.612	124.119	124.978	261.772	10.736
sd	1112.380	10.523	4.601	234.788	0.405	88.269	2115.478	56.720	65.281	93.129	1.008
median	1788.000	49.710	1.400	310.000	3.545	74.000	1277.500	116.625	108.000	257.000	10.600
min	41.000	26.278	0.300	120.000	1.960	4.000	289.000	28.380	33.000	62.000	9.000
max	4556.000	78.439	28.000	1775.000	4.400	588.000	13862.400	457.250	598.000	563.000	17.100
1st Qu.	1185.750	41.513	0.800	249.500	3.310	42.750	922.500	82.458	85.000	200.000	10.000
3rd Qu.	2689.750	56.585	3.525	401.000	3.772	129.250	2068.250	153.450	151.250	318.250	11.200
IQR	1504.000	15.072	2.725	151.500	0.462	86.500	1145.750	70.992	66.250	118.250	1.200
MAD	1131.224	10.633	1.186	106.747	0.348	53.374	756.867	52.299	45.961	87.473	0.890

Table 1: Summary statistics of numerical variables

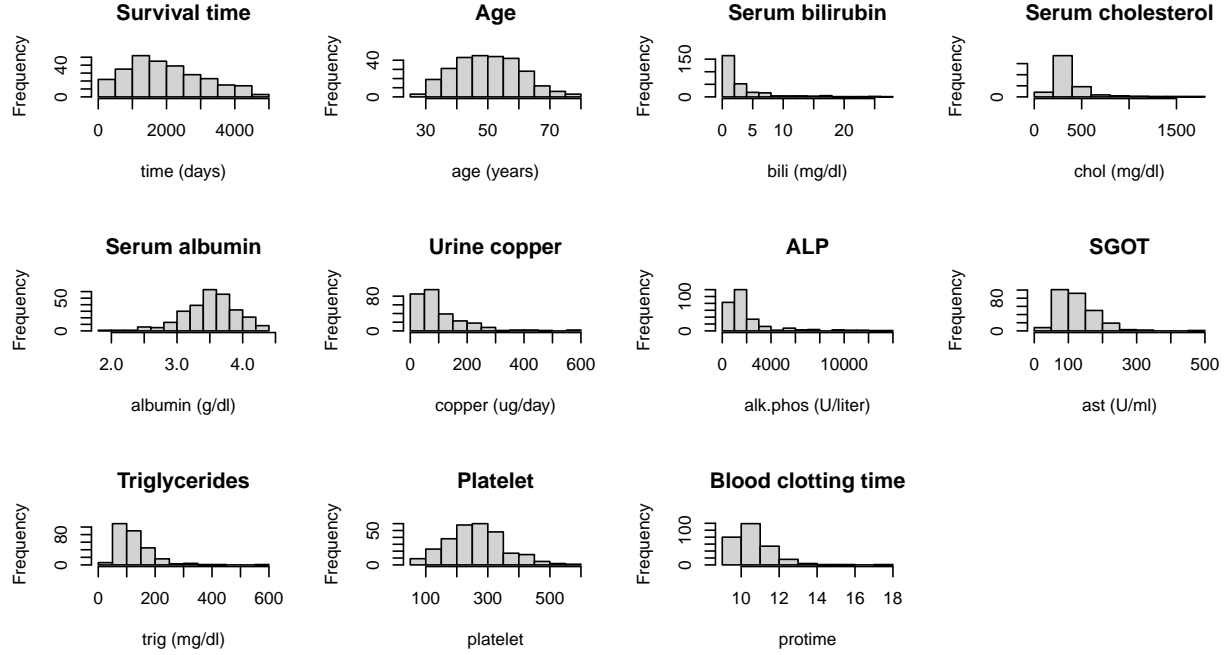


Figure 1: Histograms of numerical variables

## Qualitative Variables

From the frequency tables, the death rate of complete cases (40.22%) is fairly close to the original trial data (40.06%) and the number of patients treated with DPCA and placebo is nearly identical. And the sex ratio is 7.11:1 (female to male).

## 2.2 Bivariate Analysis

We test the correlation matrix for the variable pairs. we have the following key strong linear correlations (signif. level 0.05):

- *time* and *status*, *time* and *stage* has a negative PCC of -0.35 and -0.36 respectively, which confirms our data setting and clinical observations.

Table 2: Frequency table (a)

Var	class	freq.	%
status	alive	165	59.78
	dead	111	40.22
trt	DPCA	136	49.28
	placebo	140	50.72
sex	male	34	12.32
	female	242	87.68
ascites	no	257	93.12
	yes	19	6.88
hepato	no	134	48.55
	yes	142	51.45

Table 3: Frequency table (b)

Var	class	freq.	%
spiders	no	196	71.01
	yes	80	28.99
edema	no	234	84.78
	treated	25	9.06
	yes	17	6.16
stage	1	12	4.35
	2	59	21.38
	3	111	40.22
	4	94	34.06

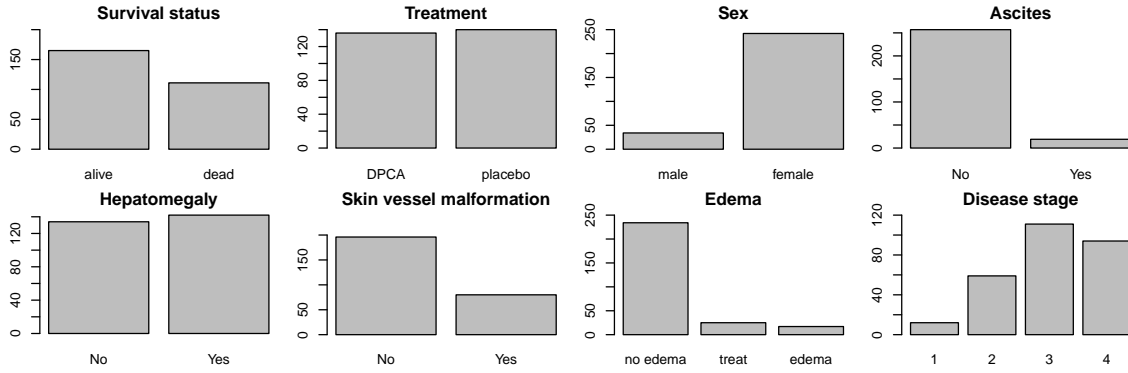


Figure 2: Frequency distribution of categorical variables

- *bili* and *time* have a negative PCC of -0.43, which is reasonable since an elevated level of serum bilirubin is a sign of liver damage or disease that the liver is not clearing the toxic chemical properly.
- *ascites* and *edema* have a positive PCC of 0.63, which is reasonable since they are both typical symptoms of liver diseases.

Besides, we notice that *status* have linear correlations with almost all clinical indicators and symptoms (Fig.7). Therefore, we will consider all covariates at this stage.

### 3. Survival Analysis

#### 3.1 Kaplan-Meier (KM) Estimator

The definition of Kaplan-Meier estimator of survive function:  $\hat{S}(t) = \prod_{t_i < t} (1 - \frac{d_i}{r_i})$ , where  $r_i$  is the number of individuals at risk just before  $t_i$  (including censored individuals at  $t_i$ ), and  $d_i$  is the number of individuals experiencing the event at time  $t_i$ .

First, we investigate whether there is a significant difference in survival curves between two groups of patients treated with DPCA and placebo. The null and alternative hypothesis are:  $H_0 : S_{DPCA}(t) = S_p(t)$ ,  $H_1 : S_{DPCA}(t) \neq S_p(t)$ , where  $H_0$  states there is no significant difference in survival time between patients who receive DPCA and placebo,  $H_1$  states there is a significant difference between the above groups. We consider  $p < 0.05$  to indicate statistical

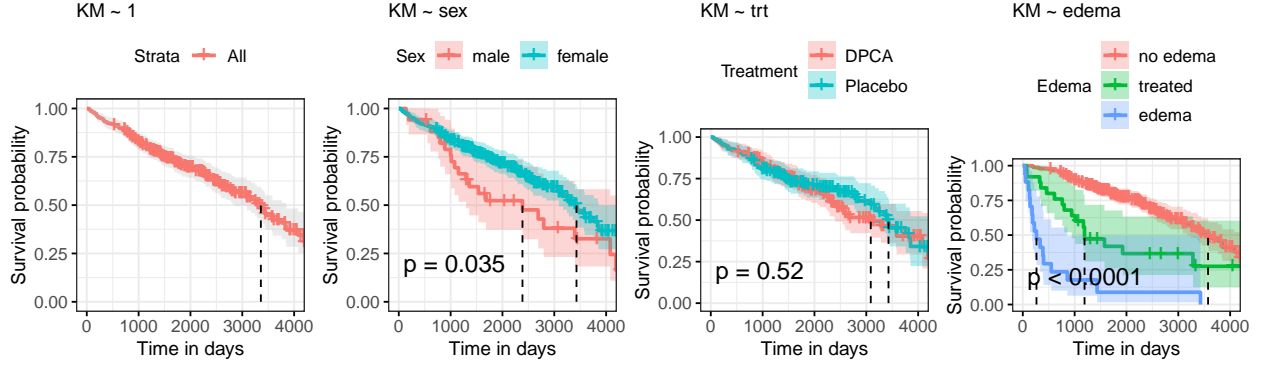


Figure 3: Survival plots by KM estimator

Var.	Cat.	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$	Chisq.	p
trt	DPCA	136	57	53.7	0.209	0.405	0.4	0.5
	placebo	140	54	57.3	0.195	0.405		
sex	male	34	21	13.7	3.878	4.47	4.5	0.03
	female	242	90	97.3	0.546	4.47		
edema	no edema	234	79	100.95	4.77	53.09	113	<2e-16
	treated	25	16	8.06	7.83	8.46		
	edema	17	16	1.99	98.74	101.59		

Table 4: Log-rank stats table

significance. The KM survival curve and log-rank test statistics are shown in Figure 3 and Table 4 separately. Since  $p = 0.5 > 0.05$ , we do not reject the null hypothesis, which means there is no significant difference in survival curves between DPCA and placebo treatment for complete patient cases. Then we investigate the survival curve difference in patients' sex. The null and alternative hypothesis are:  $H_0 : S_m(t) = S_f(t)$ ,  $H_1 : S_m(t) \neq S_f(t)$ , where  $H_0$  states there is no significant difference in survival time between male and female patients,  $H_1$  states there is a significant difference. since  $p = 0.03 < 0.05$ , we can reject the null hypothesis and state that survival time indeed has a significant difference between sex groups. This difference can be clearly captured as the survival curve of the female group is generally above the male group, indicating a higher survival probability at large  $t$ . A similar analysis conducts on the edema symptom also rejects the null hypothesis, revealing a significant difference in survival curves.

### 3.2 Cox Proportional Hazards (Cox PH) Model

Var.	Chisq.	exp(coef)	se(coef)	z	p	C-index	likelihood	log-rank	Wald
age	0.0323	1.0328	0.0092	3.509	<0.001	0.849	173.8, p<2e-16	249.7, p<2e-16	173, p<2e-16
edema:treated	0.1887	1.2076	0.2904	0.650	0.516				
edema:edema	0.9053	2.4727	0.3386	2.674	0.008				
bili	0.7343	2.0839	0.1220	6.017	<2e-09				
albumin	-0.7948	0.4517	0.2582	-3.078	0.002				
copper	0.3834	1.4673	0.1468	2.612	0.009				
protime	2.6860	14.6730	1.1867	2.263	0.024				

Table 5: AIC model selection result

Since we will consider the effect of covariates have on the risk of death (or survival time in reverse) and some of them are quantitative, we fit a Cox PH model that measure

the hazard function instead of the survival function in KM estimator. The hazard function can be expressed as  $h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$ , where  $t$  represents the survival time,  $h(t)$  is the hazard function determined by  $p$  covariates  $(x_1, x_2, \dots, x_p)$  and the coefficients  $(\beta_1, \beta_2, \dots, \beta_p)$  measure the impact of them.  $h_0$  is the baseline hazard equals to  $h(t)|_{x_i=0, i \in [p]}$ . The estimated Cox PH model should satisfy the following assumptions:

Var.	Chisq.	df	p
age	0.954	1	0.329
edema	5.769	2	0.056
bili	0.782	1	0.377
albumin	0.312	1	0.577
copper	0.776	1	0.378
protime	3.805	1	0.051
GLOBAL	2.6860	7	0.068

Table 6: Test for proportional hazards

- $\beta_i, i \in [p]$  is constant over time (proportional hazard)
- Linear form of the covariates
- No outliers which are poorly predicted by the model

Global Schoenfeld Test p: 0.06777

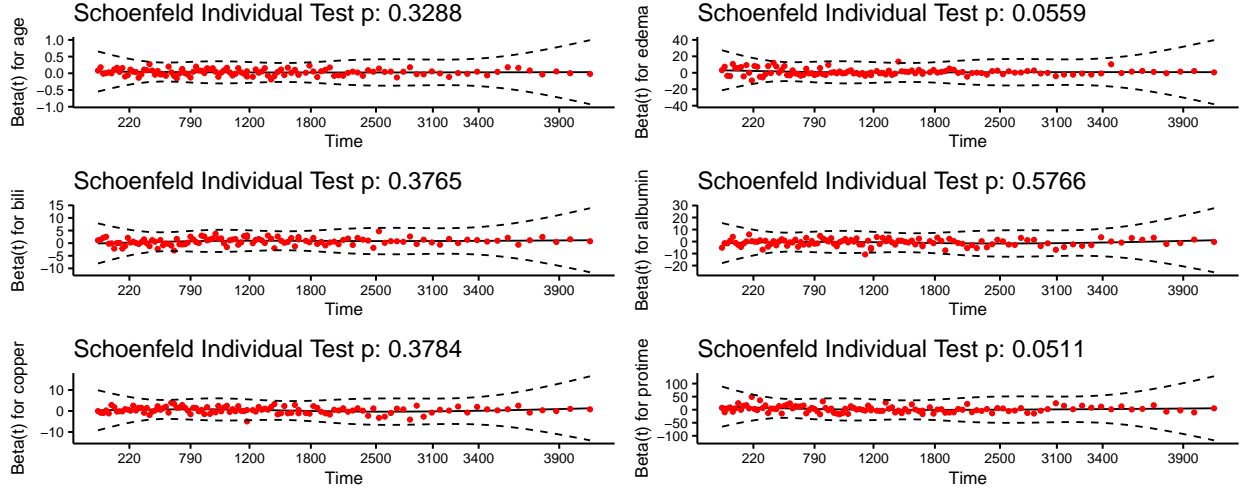


Figure 4: Schoenfeld residual plots

We apply backward elimination to select the best model regarding its explainability and complexity (number of covariates in analysis). Starting from the full model including all variables, we minimize the AIC and each time eliminate 1 covariate which gives the minimum value if deleted. The Cox PH model is fitted by maximum partial likelihood. The selection result is shown in Table 5, with its goodness-of-fit test passed and concordance of 0.849 indicating its strong robustness. Statistical and graphical assessments are shown in Table 6 and Figures 4 and 5. For all covariates selected, we cannot reject the null hypothesis that they follow the proportional hazard pattern.  $\beta(t)$  has no time-related pattern in Schoenfeld plots and also ensures proportional hazard. The martingale residual plots regarding all

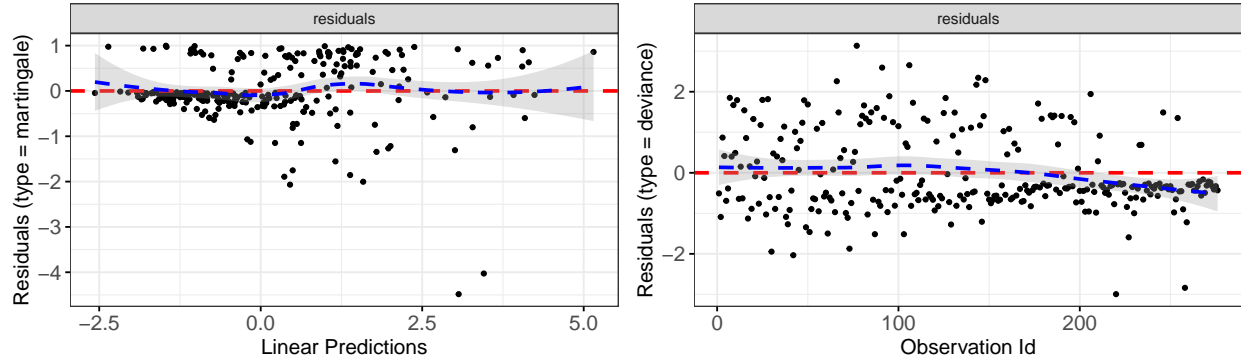


Figure 5: Martingale residual and deviance plot

included covariates generally follow a linear pattern, and the linear prediction of deviance is fairly symmetric around 0 except at the very end of observations. The model satisfies the assumptions and explains the complete patient cases properly.

### 3.3 Cox PH Model Estimation

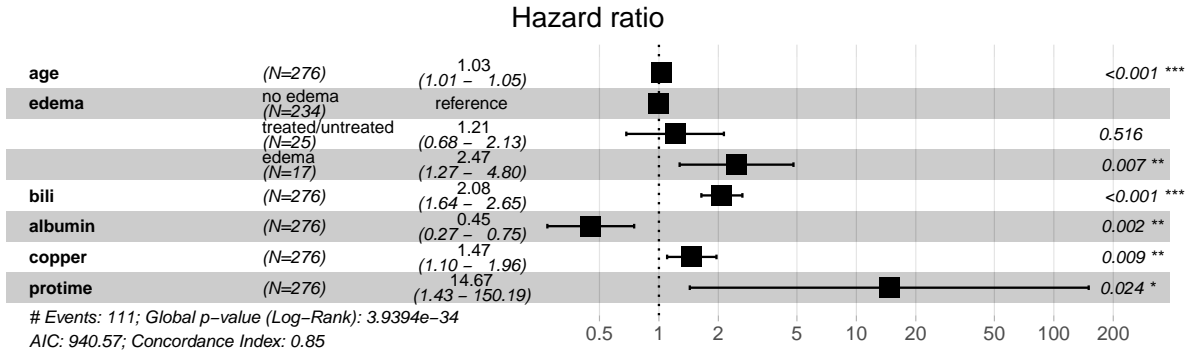


Figure 6: Forest plot of estimated model

Estimated model (0.05 signif. level, after log transformation):

$$\hat{h}(t) = \hat{h}_0(t) \times \exp(0.03 \times \text{age} + 0.91 \times \text{edema} + 0.73 \times \text{bili} - 0.79 \times \text{albumin} + 0.38 \times \text{copper} + 2.69 \times \text{protime})$$

- Blood clotting time (*protime*) has an HR of 14.67, indicating bleeding will considerably increase the risk of death for PBC patients.
- A higher level of urine copper and serum bilirubin, edema, and older age will all increase the risk of death; whereas a higher level of serum albumin will decrease the risk (Fig.6), which corresponds well to the current clinical study.
- The estimated model excludes treatment method and sex as they all failed the significance test in the full model. This confirms the KM estimator of DPCA's irrelevance with PBC patients' survival time but delivers a different result for sex since in the KM estimator the survival curves have a significant difference.

## 4. Summary

The survival analysis focuses on 276 complete cases of PBC patients and applies both the KM estimator and Cox PH model to estimate the survival function and the hazard function. The result shows that treatment with DPCA does not have a significant effect on patients' survival time. In addition, the model selected by backward AIC explains that higher age, a higher level of urine copper and serum bilirubin, having edema will increase the risk of death and a higher level of serum bilirubin will decrease the risk on the opposite. To further expand data modeling, we can train a random survival forest or construct parametric models to compare with the Cox PH model and select the one with the best explainability.

## Appendix

Variable	Type	Description
id	N/A	case number, discard in analysis
time	Quant.	<b>response variable</b> , number of days between registration & earlier status
status	Qual.	0=alive, 1=dead
trt	Qual.	treatment, 0=DPCA, 1=placebo
age	Quant.	patient's age in years at registration
sex	Qual.	patient's sex, m=male, f=female
ascites	Qual.	presence of ascites, 0=No, 1=Yes
hepato	Qual.	presence of hepatomegaly, 0=No, 1=Yes
spiders	Qual.	blood vessel malformations in the skin, 0=No, 1=Yes
edema	Qual.	0=no edema, 0.5=untreated/successfully treated, 1=edema despite diuretic therapy
bili	Quant.	serum bilirunbin (mg/dl)
copper	Quant.	urine copper (ug/day)
chol	Quant.	serum cholesterol (mg/dl)
albumin	Quant.	serum albumin (g/dl)
alk.phos	Quant.	alkaline phosphatase (U/liter)
ast	Quant.	aspartate aminotransferase, or SGOT (U/ml)
trig	Quant.	triglycerides (mg/dl)
platelet	Quant.	platelet count
protime	Quant.	standardized blood clotting time
stage	Qual.	histologic stage of disease (biopsy)

Table 7: Table of variables

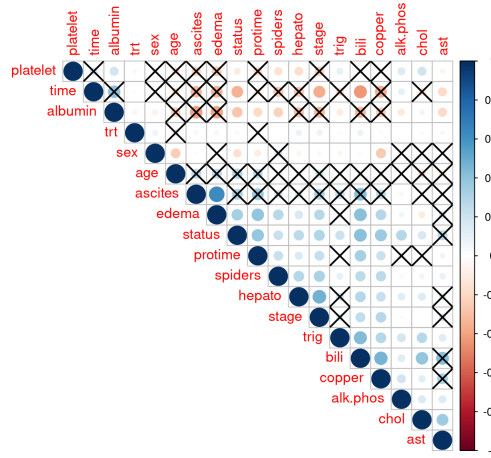


Figure 7: Visualized correlation matrix