Survival Analysis of Primary Biliary Cirrhosis Patients

P.B.W.S.R. Kumarasinghe*, E.M.P. Ekanayake, N.A.D.N. Napagoda
Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka
*Email: rasadarikumarasi@gmail.com

Abstract: Primary Biliary Cirrhosis (PBC) is a dreadful rare disease as it affects an important human organ, the liver. This study was conducted by survival analysis on Primary Biliary Cirrhosis patients in the United States using prerecorded data from 310 patients who were suffering from PBC. The Follow-up time, Status at the end, and seventeen other factors were analyzed statistically to identify the risk factors for Primary Biliary Cirrhosis. Baseline characteristics were calculated and a nonparametric model (Kaplan Meier), a semi-parametric model (Cox Proportional Hazards), and two parametric models (Exponential and Weibull) were developed for analyzing mortality and survival appropriately. Area under the Receiver Operator Characteristics curve was used for comparing the four models and the model which produced the highest Area under the Receiver Operator Characteristics curve was identified as the best fitted model. A nomogram was constructed for the graphical prediction of mortality or survival appropriately. Baseline characteristics showed that Age, Serum Bilirubin, Albumin, Urine Copper, Alkaline Phosphatase, SGOT level, Platelets, Prothrombin Time, Presence of Ascites, Presence of Hepatomegaly, Presence of Spiders, Presence of Edema, and Histologic Stage of Disease are significant factors on PBC. The Cox Proportional Hazards model could be identified as the best-fitted model from among the four fitted models, according to which Age, Serum Bilirubin, Albumin, and Urine Copper were the risk factors of Primary Biliary Cirrhosis. Moreover, Age and Serum Bilirubin were the most significant risk factors as they emerged in all fitted models.

Keywords: *Primary Biliary Cirrhosis, Risk factors, Survival models, Model selection criteria, Nomogram*

1. INTRODUCTION

The liver is one of the most important organs in the human body. Primary Biliary Cirrhosis (PBC) is an inflammatory chronic disease of the liver and it is known that slow and progressive destruction of the small bile ducts of the liver is caused by PBC [1]. Further, slow damage to the liver tissue can lead to liver cancer. Researchers believe that environmental factors like infections, smoking, and toxic chemicals are responsible along with genetic characteristics to activate PBC [1].

Though PBC is a relatively rare disease, it occurs worldwide. Numerically, PBC is reported to have affected up to one in 3000 – 4000 people while North America and Europe recorded the highest worldwide prevalence of PBC [2].

Researchers have already paid their attention to investigating PBC incidence recently [3], for identifying and assessing the clinical variables presumed to influence the survival of PBC patients [4] in the North American Continent and the rest of the world. Nevertheless, there is room for constructing survival (or mortality) models to predict survival rates (or mortality rates) and also for identifying possible links among major risk factors. Hence, the present study will focus on the survival of PBC patients using the survival analysis statistical technique.

It is intended to achieve the following objectives from this work, using an online database of PBC patients from Mayo Clinic in the United States: to construct nonparametric, semi-parametric, and parametric survival (or mortality) models for PBC patients, to trace the best fitting model among them, to estimate the death rates (or survival rates) due to PBC, and to identify major risk factors affecting PBC after which public awareness could be enhanced to mitigate them.

For this investigation of survival analysis, a credible and well-established dataset, which incorporated both recent and older data on several variables, from a recognized source was required. Though primary data from a local source were preferred initially, the absence of essential information for a comprehensive study of this nature made it necessary to look for a

widely used source recognized among the researchers. Despite its frequent application in various research domains, surprisingly, this dataset is yet to be utilized in survival analysis studies. Notably, the dataset is exceptionally suitable for such analyses, as it aligns with the potential to generate novel insights into the existing scope of research.

Hereafter this paper is organized as follows viz. Section 2 provides literature and theories related to the study, Section 3 describes the data used and the statistical methods applied, Section 4 is arranged into four parts that present principal findings with a discussion, and Section 5 summarizes the conclusions drawn from the findings.

2. LITERATURE REVIEW & THEORETICAL BACKGROUND

2.1 Literature related to the study

This sub-section shall review the key facts in previous studies on Primary Biliary Cholangitis (PBC) patients.

Galoosian *et al.* (2020) [2] studied the present congruency in the epidemiology, diagnosis, and management of PBC patients in San Francisco, United States of America. They discussed and reviewed confessed medical management and new therapeutics for PBC management. As conclusions of their study, information on the promising therapeutics landscape of PBC (including agents under clinical trials) was provided, promising outlooks for cures of PBC were provided by new pharmacologic agents, and combination therapy was indicated by early emerging data.

Kanth *et al.* (2017) [3] investigated the incidence of PBC in Midwestern Wisconsin, USA from June 1992 to June 2011. They identified patients from the electronic medical records initially and later verified them with the American Association for the Study of Liver Disease criteria for PBC. A KM analysis was also performed. 79 PBC patients were found for 20 years and it was an incidence of 4.9 cases per 100,000 people. The incidence of females was higher than that of males but changes over time were negligible. The probability of mortality of PBC patients was 0.29 and the estimated 10-year survival probability was 0.76. They concluded that the overall incidence of PBC in Midwestern Wisconsin was relatively stable, patients had a better prognosis, and survival of PBC had increased.

Jackson *et al.* (1971) [4] attempted to determine whether portocaval shunts (a therapy to prevent PBC) would prevent rehemorrhage from esophagogastric varices in cirrhotic patients and determine if it would develop survival of PBC. They also desired to find out the clinical factors affecting PBC. A sample of 155 bleeding PBC patients was chosen randomly and divided into two therapy groups viz, medical and shunt. The Standard t-test, $\mathcal{X}^2$ test, or Fisher's Exact test (where suitable) were used for analyzing the differences between means and proportions. Life tables and nonparametric tests were used for the comparison of survival curves. They found that age, varying values of standard liver function tests, histological changes in the liver, the threat of peptic ulcer, hepatic failure, and post-shunt encephalopathy affected PBC. However, they could not support the claim that the incidence of shunted ones is significantly greater than medically treated ones. Hypersplenism was not considered a factor affecting PBC. They concluded shunting was the recommended therapy method.

2.2 Theories related to the study

This sub-section presents the previous studies related to Survival Analysis of patients who suffered from various medical problems and the statistical tools adopted for the present study from such work.

Lai *et al.* (2021) [5] investigated the factors affecting mortality in psycho–cardiac disease with a Chinese sample by analyzing clinical characteristics, clinical outcomes, and survival characteristics. Their KM analysis showed that dysphagia, limb function, self-care ability, percutaneous coronary intervention, low density

lipoprotein, total cholesterol, pro–brain natriuretic peptide, and high-sensitivity troponin had significant effects on the survival of cardiac patients. Further, Cox Regression analysis showed that total cholesterol, troponin, and percutaneous coronary intervention caused a significant effect on survival independently. In the preliminary stage of the present work, baseline characteristics of PBC patients were analyzed in line with the above study while applying nonparametric and semi-parametric survival methods (KM model and CPH model).

Chakraborthy and Tsokos (2021) [6] focused on finding any significant difference in the survival probabilities of Acute Myeloid Leukemia (AML) patients according to gender in the United States. They also looked for the best-fitting parametric model for gender-wise survival and compared the survival probabilities with nonparametric models. Both parametric and nonparametric models were used in the survival analysis of 2015 AML patients. Their findings showed that there was a significant difference in the mean survival time of males and females and that the Generalized Extreme Value (GEV) model was the best-fitted parametric model. The survival probabilities estimated with the GEV model were compared with KM model probabilities. The comparison of estimated survival probabilities of two methods (GEV and KM) revealed that the parametric GEV model was the best-fitted model, as it produced the highest estimated survival probabilities and the estimated survival probability at time zero was very close to one. This technique was also adopted in the present work to compare the parametric model with the nonparametric (or semi-parametric) model using estimated survival probabilities.

Giolo *et al.* (2012) [7] designed several models to predict the survival of patients with heart failure in Sao Paulo, Brazil. Survival data were analyzed by using the proportional hazards Cox model, variations of Cox's model, and Aalen's additive model. Their findings revealed that age, serum sodium, hemoglobin, serum creatinine, and left ventricular ejection fraction were significantly associated with mortality. In addition, high hemoglobin and high left ventricular ejection fraction were associated with reduced risk. The impact of age and sodium on mortality remained constant over time. This paper provided useful insights into building up the KM model with the Log Rank test of the present study.

Adeboye *et al.* (2020) [8] investigated the success of treatments given to asthmatic patients over time. The factors affecting the response to treatments administered to patients were studied with parametric survival models and semi-parametric models. Data from 464 asthmatic patients from Ogun State, South Western Region of Nigeria were used in their study. Four parametric models (Weibull, Gompertz, Log-normal, and Log-logistic) and a CPH model were fitted for asthma patients' data. The parametric Log-normal regression model (with affecting factors: Smoking, Obesity, Environmental pollution, and Respiratory illness) was identified as the best-fitted model for asthma patients, as it had the least AIC value and the least Negative Log Likelihood value. Fitting two different parametric models for survival data and the model selection technique (with AIC values and Negative Log Likelihood values) in the present study resemble those in [8].

Ahamad *et al.* (2017) [9] conducted a study on survival analysis of heart patients in Faisalabad, Pakistan by applying Cox Regression to model mortality for which they considered risk factors and used a KM Plot to figure out the pattern of survival in the analysis. Their results showed that age, renal dysfunction, blood pressure, ejection fraction, and anemia were the significant risk factors for mortality of patients with heart failure. The survival curves, building up the CPH model, comparison of models using CC and ROC curves, and prediction with the Nomogram of the present study are congruent with those of [9].

Table 1: Description of continuous variables

Continuous Variable	Abbreviation	Measurement
Follow up time	FU_DAYS	in days
Age of the patient	AGE	in days
Serum Bilirubin level	BILI	in mg/dl
Serum Cholesterol level	CHOL	in mg/dl
Albumin level	ALBUMIN	in gm/dl
Urine Copper level	COPPER	in ug/day
Alkaline Phosphatase level	ALK_PHOS	in U/liter
SGOT level	SGOT	in U/ml
Triglicerides level	TRIG	in mg/dl
Platelets amount	PLATELET	per cubic ml/1000
Prothrombin time	PROTIME	in seconds

Table 2: Description of categorical variables

Categorical Variable	Abbreviation	Categories
Status at the end	STATUS	0 = censored, 1 = censored due to liver tx, 2 = death
Type of drug used	DRUG	1 = D-penicillamine, 2 = placebo
Gender of the patient	GENDER	0 = male, 1 = female
Presence of Ascites	ASICTES	0 = no, 1 = yes
Presence of Hepatomegaly	HEPATOM	0 = no, 1 = yes
Presence of Spiders	SPIDERS	0 = no, 1 = yes
Presence of Edema	EDEMA	0 = no edema, 0.5 = edema resolved by diuretics, 1 = edema despite diuretic therapy
Histologic stage of disease	STAGE	1,2,3,4

3. METHODOLOGY

3.1 Description of Data

For the present study, prerecorded data of 310 PBC patients who attended the Mayo Clinic in the United States were collected from the http://lib.stat.cmu.edu/datasets/pbc online database. The Follow-up time, Status at the end, and seventeen other factors were analyzed to isolate the major risk factors for PBC. Table 1 shows the abbreviations and units of measurements of variables on the continuous scales while Table 2 presents the abbreviations and sub-categories of the categorical variables in the above database. For the convenience of calculations, FU_DAYS and AGE were converted from days to years and renamed as FU_TIME and AGE_YRS.

3.2 Statistical Analysis

In this section, descriptive statistics are explained first [5] and then a nonparametric model [6], a semi-parametric model [7], and two parametric models [8] are developed for analyzing mortality and survival appropriately. The best-fitted model from among the four models is recognized with the aid of the ROC curve [9]. Finally, a nomogram is constructed for the graphical prediction of mortality or the survival of the patients affected by PBC [9].

Continuous variables are tabulated around their mean values with standard deviation and the 2-sample t-test will be used to find out the significant factors [5]. Categorical variables are presented as percentages and the Chi-square test will be used to find out the significant factors in them [5]. The level of significance is chosen as

0.01 and the significance will be defined as $p < 0.01$. The following hypotheses were adopted for the respective tests.

Hypotheses for the 2-sample t-test:

- H01: There is no significant difference between alive and dead PBC patients
- H02: There is a significant difference between alive and dead PBC patients

Hypotheses for the Chi-square test:

- H01: Status (Alive or Dead) is independent from the factor considered
- H02: Status (Alive or Dead) is dependent on the factor considered

Univariate survival analysis was carried out with the mostly used non-parametric statistical technique called the Kaplan Meier (KM) method in which the medical and general factors likely to be associated with the survival time of PBC patients were investigated [6]. Further, the KM Estimator described below was used in constructing survival functions.

Let $T_1 < T_2 < \ldots < T_d$ be the ordered event times in the sample ($j - 1 < j < d$), D_j represents the total number of failures occurring at T_j, and N_j represents the total number at risk at time T_j. Then the estimate of survival time $\left(\hat{S}(T)\right)$ can be written as follows [8].

$$\hat{S}(T) = \prod_{j:T_j \leq T} \left(1 - \frac{D_j}{N_j}\right) \qquad (1)$$

The Log Rank Test was used to compare the survival curves [6] [10]. KM method and the Log Rank Test were performed at 0.01 level of significance and the significance would be defined as $p < 0.01$. KM plots were used to illustrate the survival probabilities according to the KM Estimator and the Log Rank Test [6] [10].

The widely used semi-parametric survival statistical technique known as the Cox Proportional Hazards (CPH) Model could be applied [7] at a 0.01 level of significance while defining the significance as $p < 0.01$. The CPH model investigated the relationship between the survival time of PBC patients and the variables (or medical and general factors) considered. This model usually does not follow any specified distribution for survival time as in parametric models [7]. The CPH model was constructed subject to two assumptions viz. hazard for two individuals is proportional and the variables are in linear form [7]. Accordingly, the CPH model can be written as follows [8].

$$h(t) = h_0(t) \times \exp[\, \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \,] \qquad (2)$$

Where; $h_0(t)$ – baseline hazard, $h(t)$ – expected hazards, $\beta_1, \beta_2, \ldots, \beta_n$ – coefficients of X_1, X_2, …, X_n respectively, and X_1, X_2, …, X_n – variables.

The EXP distribution is one of the mostly used parametric survival models of patients with dangerous diseases such as PBC, which is specified by a single rate parameter. EXP distribution only supports a hazard that is constant over time. In other words, the hazard is simply equal to the rate parameter [11]. For the construction of the EXP model, the level of significance was chosen as 0.01 and the significance was defined as $p < 0.01$. This EXP model can be written as follows [11]:

$$S(t, X_1, X_2, \ldots, X_n) = \exp\{-t\,[\, \exp(-\alpha_0 - \alpha_1 X_1 - \alpha_2 X_2 - \ldots - \alpha_n X_n)\,]\} \qquad (3)$$

Where; $S(t, X_1, X_2, \ldots, X_n)$ – survival function, t – time, α_0 – intercept, and $\alpha_1, \alpha_2, \ldots, \alpha_n$ – are the coefficients of the variables $X_1, X_2, \ldots, X_n$ respectively.

Weibull distribution is another parametric survival model appropriate for addressing medical issues because it supports monotonically increasing and decreasing hazards as well as in Gompertz and Gamma distributions [11]. For this WEI model building, the level of significance was chosen as 0.01 and the significance was defined as $p < 0.01$. The WEI model can be presented as follows [11]:

$$S(t, X_1, X_2, \ldots, X_n) = \exp\{\, t^K\,[\, -\exp(-\alpha_0 - \alpha_1 X_1 - \alpha_2 X_2 - \ldots - \alpha_n X_n)^K]\} \qquad (4)$$

Where; $S(t, X_1, X_2, \dots, X_n)$ – survival function, t – time, K – scale, α_0 – intercept, and $\alpha_1, \alpha_2, \dots, \alpha_n$ – are the coefficients of the variables $X_1, X_2, \dots, X_n$ respectively.

ROC Curve is a visualization tool for evaluating and checking the discrimination ability of a fitted survival model for PBC patients, which is quantified in terms of the area under the curve (AUC) [9] [12]. ROC curves were drawn for all four models and the areas under them were calculated as part of the criteria for selecting the best model.

A nomogram is a graphical prediction method for assessing the survival probability or risk probability of an individual [9]. After assigning appropriate points to all independent prognosis factors (according to their level of significance in the constructed model), the nomogram results in graphical predictions of survival probability (or risk probability). The total of allocated points will give an estimated value for the survival probability (or mortality probability) for a particular patient [9].

4. RESULTS AND DISCUSSION

This section is arranged into four parts as follows. Part A explains the baseline characteristics of patient details and Part B presents results from the non-parametric analysis with the KM model, semi-parametric analysis with the CPH model, and the parametric analysis with the EXP model and the WEI models. Part C presents comparative statistics among the KM, CPH, EXP, and WEI models while Part D illustrates the graphical prediction of survival probability or risk probability complied by the nomogram.

PART A

4.1. Baseline Characteristics of Dead and Alive Patients

In Tables 3 and 4, baseline characteristics of observed Primary Biliary Cirrhosis (PBC) patients are presented for continuous variables and categorical variables respectively. A clinical and demographic data sample of 310 patients who suffered from PBC could be used, wherein 36 patients (11.6%) were male and the rest (274 patients) were female. This accounted for 88.4% of the tested sample. Out of the 310 patients, 124 patients died at the end of the follow-up time and only 186 patients survived. The average age of patients who died within the follow-up time was 53.44 years and it was 47.97 years for the patients who survived within the observation period.

Numerical figures in Tables 3 and 4 are the mean values and percentages of all categories in continuous and categorical variables of both dead and alive patients. Asterisks (*) indicate the significant factors that emerged from appropriate tests. The 2-sample t-test and the Chi-squared test showed that there are significant differences between dead and alive groups in the variables of Age-Yrs, Bili, Albumin, Copper, Alk – phos, Sgot, Platelets, Protime, Ascites, Hepatom, Spiders, Edema, and Stage.

PART B

4.2 Kaplan Meier Model and Log Rank Test

The univariate survival analysis was conducted by employing the widely used non-parametric statistical technique of the Kaplan-Meier (KM) method. The factors likely to be associated with the survival time were investigated using the Log Rank test that compared the survival curves [7]. Both the KM method and the Log Rank Test were performed at a 0.01 level of significance while defining the significance as $p<0.01$.

The continuous variable Age-yrs was divided into two categories, as in Part A, for the KM analysis. All other continuous variables with medical characteristics were categorized according to their standard values or ranges as follows: Bili – (low, normal, high), Chol – (healthy, unhealthy), Albumin – (normal, abnormal), Copper – (low, normal, high), Alk_phos – (low, normal, high), Sgot – (low, normal, high), Trig – (low, normal, high), Platelets – (low, normal, high) and Protime – (low, normal, high).

Table 3: Baseline characteristics for dead and alive patients (Continuous Variables)

Continuous Variables	Dead (n= 124)	Alive (n=186)	Test Statistics (*t*)	*p-value*
Age-Yrs *	53.43621	47.9736	-4.6148	$5.782e^{-06}$
Bili *	5.62	1.67	4768.5	$< 2.2e^{-16}$
Chol	406.81	343.80	9977	0.04436
Albumin *	3.36	3.63	5.3317	$2.576e^{-07}$
Copper *	134.98	72.72	6104.5	$2.221e^{-12}$
Alk – phos *	2605.48	1535.03	7864	$2.099e^{-06}$
Sgot *	141.85	109.89	7444	$1.243e^{-07}$
Trig	137.75	115.84	9545	0.01018
Platelets *	241.69	274.8	3.0364	0.002599
Protime *	11.24	10.38	5361	$1.336e^{-15}$

Asterisks (*) in Table 3 indicate the significant factors that emerged from the 2 sample t-test.

Table 4: Baseline characteristics for dead and alive patients (Categorical Variables)

Categorical Variables		Dead (n= 124)	Alive (n=186)	Test Statistics (X^2)	*p*-value
Drug	D penicillamine	64 (51.6%)	93 (50%)	0.026348	0.8711
	Placebo	60 (48.4%)	93 (50%)		
Gender	Male	22 (17.7%)	14 (7.5%)	6.6011	0.01019
	Female	102 (82.3%)	172 (92.5%)		
Ascites *	No	101 (81.5%)	185 (99.5%)	31.315	2.194e-08
	Yes	23 (18.5%)	1 (0.5%)		
Hepatom *	No	37 (29.8%)	115 (61.8%)	29.199	6.533e-08
	Yes	87 (70.2%)	71 (38.2%)		
Spiders *	No	72 (58.1%)	149 (80.1%)	16.602	4.61e-05
	Yes	52 (41.9%)	37 (19.9%)		
Edema *	0	88 (71.0%)	173 (93.0%)	33.692	4.83e-08
	0.5	17 (13.7%)	12 (6.5%)		
	1	19 (15.3%)	1 (0.5%)		
Stage *	1	1 (0.8%)	15 (8.1%)	33.217	2.898e-07
	2	16 (12.9%)	51 (27.4%)		
	3	43 (34.7%)	77 (41.4%)		
	4	64 (51.6%)	43 (23.1%)		

Asterisks (*) in Table 4 indicate the significant factors that emerged from the Chi-squared test.

Table 5: Results of KM model using Log Rank test and final models for CPH, EXP, and WEI

	Statistics		Coefficients		
Variable	**KM**	**CPH**		**EXP**	**WEI**
		Model A	Model B		
Age_yrs	16.8	0.0315447	0.0315447	-0.042721364	-0.02530799
Bili	127	0.1179812	0.1179812	-0.093940582	-0.08328034
Chol					
Albumin	42	-1.3347825	-1.3347825		
Copper	59.2	0.0038807	0.0038807		
Alk – phos					
Sgot				-0.005496734	
Trig					
Platelets	17				
Protime	49.2			-0.310553626	-0.17386850
Drug					
Gender					
Ascites	104				
Hepatom	41.4				
Spiders	34.2				
Edema	126				-1.01127728
Stage	56.3				
Intercept				9.216356405	5.97996905

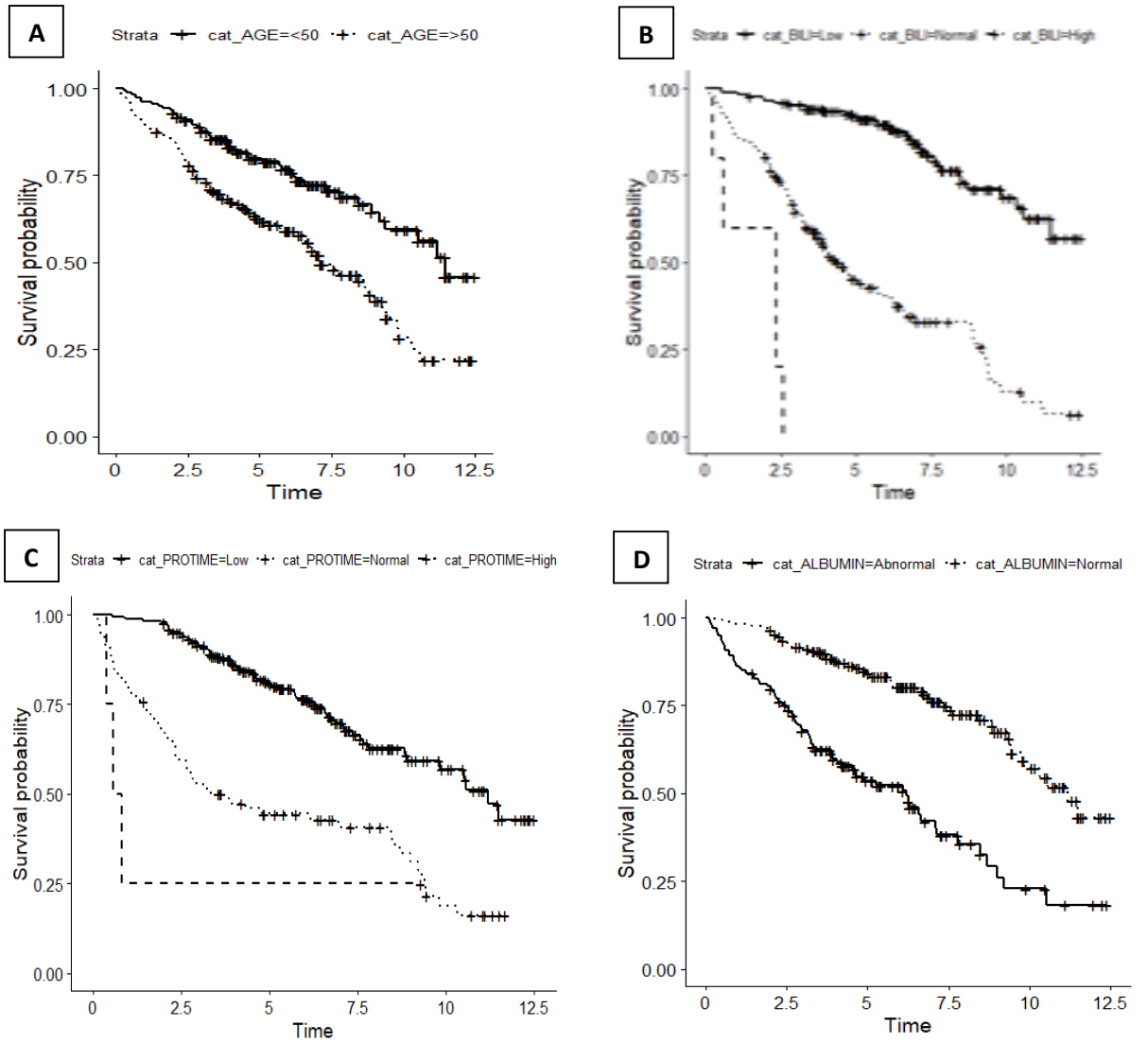

Figure 1: KM plots for (A) Age-Yrs, (B) Serum Bilirubin (Bili), (C) Albumin (Albumin), and (D) Prothrombin Time (Protime) which are medically significant in the Kaplan Meier model.

KM analysis and the Log Rank Test used for comparing survival curves showed that Age-Yrs, Bili, Albumin, Copper, Platelets, Protime, Ascites, Hepatom, Spiders, Edema, and Stage (Table 5) had a significant association with the cumulative survival time, as their *p* values were less than 0.01.

After fitting the KM model, it was clear that Age, Bili, Albumin, Copper, Platelets, Protime, Ascites, Hepatom, Spiders, Edema, and Stage affect the survival time of PBC patients.

4.2.1 Kaplan Meier Plots

According to KM analysis, the Kaplan Meier (KM) plots for the factors, that were identified as significant from the Log Rank Test, are illustrated in Figure 1. After fitting the KM model and performing the Log Rank test, four factors with medical importance [1], [13], [14], [15] viz. (Age-Yrs, Bili, Albumin and Protime) emerged significant in the KM model. Hence, KM curves were plotted for them (Figure 1) [6].

Figure 1A shows that patients younger than 50 years of age have a better survival advantage compared to patients older than 50 years.

Bili was categorized into three standard groups (low, normal, high) [13] whose survival curves are depicted in Figure 1B. It can be inferred that there is a survival advantage for those having

low level of Bili compared to patients with normal and high levels.

Albumin was also categorized as normal and abnormal based on its standard normal range, 3.5mg/dl - 5.5mg/dl [14]. KM curves (Figure 1C) clearly show that patients in the normal range of Albumin level have a better survival benefit compared to patients in the abnormal group.

Protime was categorized into three categories (low, normal, high) as the normal Protime, according to medical definitions, lies in the range of 11s – 13.5s [15]. The KM curve of Protime (Figure 1D) indicates that patients in the low range of Protime have a significant survival benefit over those in the other two categories.

After fitting the KM Model and performing the Log Rank Test, the following observations could be made. If the patients are below 50 years of age and have low-level of Serum Bilirubin and Albumin in the normal range, and Prothrombin Time in the low range, they have a greater survival probability, whereas the patients above 50 years of age with a high level of Serum Bilirubin, Albumin in the abnormal range, and Prothrombin Time in the high range had a lower survival probability.

4.3 Cox Proportional Hazards Model

Extending the analysis to a semi-parametric level, the statistical technique of the Cox Proportional Hazards (CPH) Model was performed at a 0.01 level of significance while defining the significance as $p < 0.01$.

Two models were considered at the beginning of the analysis such that all variables were included in Model A but only the significant factors in the KM Model were included in Model B. Model A was reduced to the four factors of Age-Yrs, Bili, Albumin, and Copper which were associated with survival. Model B was also reduced to the same four factors which were also associated with survival (Table 5). The factor-reduced Model A and Model B with only the significant factors were the same and the corresponding coefficients of all factors in them were also the same. Hence, the final CPH Model was considered with Age-Yrs, Bili, Albumin, and Copper.

COX PROPORTIONAL HAZARDS MODEL (INITIAL):

$$(2) \implies h(t) = h_0(t) \times \exp[0.0315447(\text{Age}_{\text{yrs}}) + 0.1179812(\text{Bili}) - 1.3347825(\text{Albumin}) + 0.0038807(\text{Copper})]$$

According to the fitted CPH model; Age-Yrs, Bili, and Copper were positively related to Hazards but Albumin had a negative relationship with Hazards.

The proportional hazards assumption was valid for the fitted model. For validation of the linearity assumption, log transformation should be used for Serum Bilirubin and Urine Copper. Hence, the final CPH Model was valid with Age-Yrs, log[Bili], Albumin, and log[Copper].

COX PROPORTIONAL HAZARDS MODEL (FINAL):

$$(2) \implies h(t) = h_0(t) \times \exp[0.0315447(\text{Age}_{\text{yrs}}) + 0.1179812(\log(\text{Bili})) - 1.3347825(\text{Albumin}) + 0.0038807(\log(\text{Copper}))]$$

According to this modified CPH model, the Age-Yrs, log[Bili], and log[Copper] were positively related to Hazards but Albumin had a negative relationship with Hazards.

4.4 Exponential Model

Advancing the analysis to the parametric level, an Exponential (EXP) model parameterized by a single rate parameter was also performed [11]. EXP model is known to support a hazard that is constant over time. In other words, the hazard is simply equal to the rate parameter [11]. The level of significance for the EXP model was chosen as 0.01 and the significance was defined as $p < 0.01$ corresponding to the levels of significance adopted previously.

All independent variables were included at the beginning of constructing the EXP model, which was reduced to the following four factors with an intercept at the final stage: Age-Yrs, Bili, Sgot, and Protime, all of which were associated with survival (Table 5).

Thus, the final EXP model was considered with Age-Yrs, Bili, Sgot, Protime, and the Intercept.

EXPONENTIAL MODEL:
(3) $\Longrightarrow$ S(t) = exp{−t [exp(−9.216356405 + 0.042721364(Age-Yrs) + 0.093940582(Bili) + 0.005496734(Sgot) + 0.310553626(Protime))]}

Accordingly, all the variables of Age-Yrs, Bili, Sgot, and Protime were found to be positively associated with survival.

4.5 Weibull Model

Weibull (WEI) model is a well-known parametric survival model that supports monotonically increasing and decreasing hazards as in Gompertz and Gamma models [11]. The level of significance for WEI model building was retained at 0.01 and the significance was $p < 0.01$.

As before, all independent variables were included at the beginning of building the WEI model, which also eventually reduced to the four factors with an intercept: Age-Yrs, Bili, Edema, Protime, and Intercept, showing associations with survival (Table 5). The scale for the final WEI model was 0.7.

Hence, the final EXP model could be considered with Age-Yrs, Bili, Edema, Protime, and the Intercept.

WEIBULL MODEL:
(4) $\Longrightarrow$ S(t) = exp{ $t^{0.7}$ [- exp(−5.97996905+1.01127728 (Edema)+0.02530799 (Age-Yrs)+0.08328034 (Bili)+0.17386850 (Protime))$^{0.7}$]}

According to the fitted WEI model, Age-Yrs, Bili, Edema, and Protime were positively associated with survival.

PART C

4.6 Comparison of Models

An ROC Curve is a visualization tool for evaluating and checking the discrimination ability (the percentage of ability to correctly identify the event of death) of a fitted model. The numerical value for checking the discrimination ability of a fitted model is the AUC [9].

The AUC value of the ROC Curve for the KM model was 0.72 (Figure 2A, Table 6) and it was 0.864 for the CPH model (Figure 2B, Table 6). This implies that the KM model is capable of identifying the event of death 72% correctly and the CPH model 86.4% correctly.

The AUC value of ROC in the EXP model was 0.82 (Figure 2C, Table 6) while it was 0.84 for the WEI model (Figure 2D, Table 6). Though this difference is just marginal, the values convey that the EXP model is good at identifying the event of death with 82% correctly and the WEI model with 84% accuracy.

The next step was to find out comparatively the best-fitting model from among the KM, CPH, EXP, and WEI models. AUC values of the ROC curve were calculated for this purpose, as summarized in Table 6.

The numerical figures in Table 6 imply that the discrimination ability of the CPH model was the best, which corresponds to its superiority compared to other models.

COX PROPORTIONAL HAZARDS MODEL:

$$(2) \Longrightarrow h(t) = h_0(t) \times \exp[0.0315447(\text{Age}_{\text{yrs}}) + 0.1179812(\log(\text{Bili})) - 1.3347825(\text{Albumin}) + 0.0038807(\log(\text{Copper}))]$$

Thus, according to the finalized CPH model, Age (Age-Yrs), log[Serum Bilirubin (Bili)], and log[Urine Copper (Copper)] were positively related to Hazards while Albumin (Albumin) had a negative relationship with Hazards.

Since the factors of Age (Age-Yrs) and Serum Bilirubin (Bili) emerged significant in all the four models considered, they could be determined as the most influential from among all the factors.

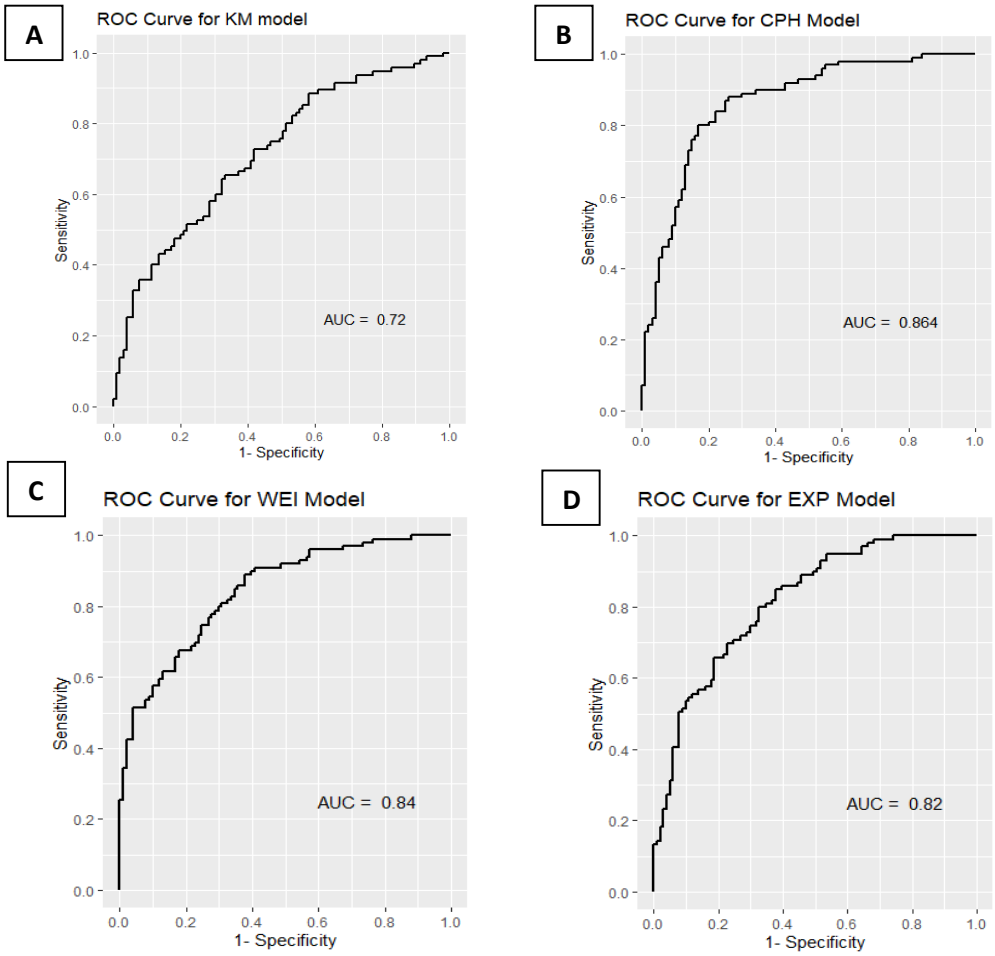

Figure 2: ROC curves for (A) KM and (B) CPH (C) EXP and (D) WEI models

Table 6: Model selection criteria

Model Selection Criteria	Models			
	KM	**CPH ***	**EXP**	**WEI**
AUC of ROC	0.72	0.864	0.82	0.84

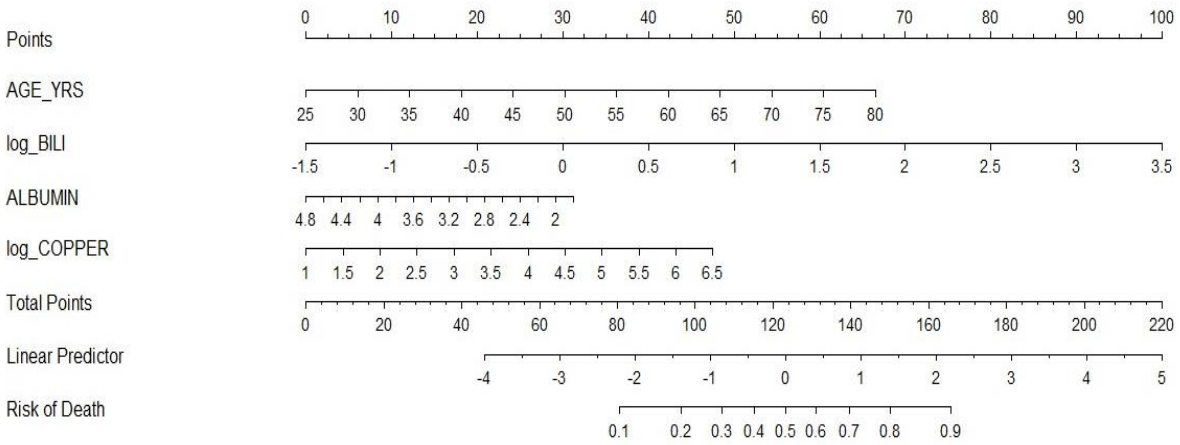

Figure 3: Nomogram for prediction in the CPH model

PART D

4.7 Prediction

A nomogram is a graphical prediction method of assessing survival probability or risk probability, wherein the independent variables will be assigned points [9] as shown in Figure 3. According to the points accumulated, the overall risk probability of death (mortality probability) is calculated by the nomogram to a single value [9].

The final CPH model was used for constructing the nomogram (Figure 3) of the risk of death by predicting mortality directly and survival probabilities indirectly.

For instance, consider an arbitrary patient with the following characteristics: AGE-YRS = 50, log(BILI) = 0.5, ALBUMIN= 3.8, log(COPPER) = 3.3. These values are matched with the Point Scale, 0-100, which can then be added as 30+40+10+20 = 100 on the Total Point Scale, 0-220, of the nomogram. According to the total points (100) obtained by this patient, his status corresponds to having a death risk of 0.24 probability (mortality probability). In conclusion, he has a (1 – 0.24) = 0.76 survival probability.

5. CONCLUSION

Despite its rarity in the current generation, Primary Biliary Cirrhosis (PBC) poses a significant threat due to its harmful impact on the liver. This study was aimed at modeling the survival outcomes of PBC patients through a comprehensive statistical approach, which would be extremely useful for understanding mortality rates and identifying influential factors to recommend effective mitigative measures.

In pursuit of choosing appropriate data commensurate with research objectives, a well-established and widely utilized dataset, albeit one not extensively employed in survival analysis, had to be identified. Despite its popularity in various research domains, the dataset selected for this research remains untapped for survival analysis applications and its inherent characteristics make it amply suitable for addressing the objectives of this study.

Accordingly, pre-existing data from 310 PBC patients recorded between 1974 and 1984, sourced from the Mayo Clinic's website were used with the analysis focusing on factors such as Follow-up time, Status at the end, and seventeen additional variables to identify risk factors associated with PBC.

A non-parametric model (KM model) and a semi-parametric model (CPH model) were constructed in the first stage. Two parametric models (EXP and WEI models) were developed in the second stage. According to the AIC values, the CPH model was found to be superior to all the other three models. The nomogram illustrated the prediction of mortality based on the results produced by the CPH model. The analysis revealed that Age and Serum Bilirubin level were consistently significant factors across all four models, establishing themselves as the primary influencers on Primary Biliary Cholangitis (PBC). Additionally, Albumin level and Urine Copper level demonstrated noteworthy relevance in the optimal Cox Proportional-Hazards (CPH) model. Conversely, the remaining prognostic factors did not exhibit significant effects on PBC outcomes.

While achieving the objectives stated under the introduction, it should be stated that no methods could be found in the literature that had attempted to identify the risk factors of PBC and forecast the death (survival) rate using the best-fitted model in this study, the CPH.

6. REFERENCES

[1]. Primary Biliary Cholangitis – Symptoms & Causes, retrieved from https://www.mayoclinic.org/diseases-conditions/primary-biliary-cholangitis/symptoms-causes/syc-20376874 on August, 2022

[2]. Galoosian, A., Hanlon, C., Zhang, J., Holt, E.W. and Yimam, K.K., 2020. Clinical updates in primary biliary cholangitis: trends, epidemiology, diagnostics, and new therapeutic approaches. *Journal of clinical and translational hepatology*, *8*(1), p.49.

[3]. Kanth, R., Shrestha, R.B., Rai, I., VanWormer, J.J. and Roy, P.K., 2017. Incidence of primary biliary cholangitis in a rural Midwestern population. *Clinical Medicine & Research*, *15*(1-2), pp.13-18.

[4]. Jackson, F.C., Perrin, E.B., Felix, W.R. and Smith, A.G., 1971. A clinical investigation of the portacaval shunt. V. Survival analysis of the therapeutic operation. *Annals of surgery*, *174*(4), p.672.

[5]. Lai, M., Shen, T., Cui, H., Lin, L., Ran, P., Huo, P., Chen, L. and Li, J., 2021. Clinical outcomes and survival analysis in patients with psycho-cardiological disease: a retrospective analysis of 132 cases. *Journal of International Medical Research*, *49*(3), p.0300060521990984.

[6]. Chakraborty, A., & Tsokos, C. P. (2021). Parametric and non-parametric survival analysis of patients with acute myeloid leukemia (AML). *Open Journal of Applied Sciences*, *11*(01), 126.

[7]. Giolo, S. R., Krieger, J. E., Mansur, A. J., & Pereira, A. C. (2012). Survival analysis of patients with heart failure: implications of time-varying regression effects in modeling mortality. *PLoS One*, *7*(6), e37392.

[8]. Adeboye, N.O., Ajibode, I.A. and Aako, O.L., 2020. On the Survival Assessment of Asthmatic Patients Using Parametric and Semi-Parametric Survival Models. *Occupational Diseases and Environmental Medicine*, pp.50-63.

[9]. Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M. and Raza, M.A., 2017. Survival analysis of heart failure patients: A case study. *PloS one*, *12*(7), p.e0181001.

[10]. Survival Analysis Basics-Easy Guide, retrieved from: http://www.sthda.com/english/wiki/survival-analysis-basics on August, 2022

[11]. Parametric Survival, retrieved from www.devinincerti.com/2019/06/18/parametric_survival.html on August, 2022

[12]. A Guide to Model Selection for Survival Analysis, retrieved from www.towardsdatascience.com/a-guide-to-model-selection-for-survival-analysis-2500b211c733 on August, 2022

[13]. Sodium Bilirubin Test, retrieved from www.mayoclinic.org/tests-procedures/bilirubin/about/pac-20393041 - 11 on August, 2022

[14]. Albumin Blood Serum Test, retrieved from www.mountsinai.org/health-library/tests/albumin-blood-serum-test on August, 2022

[15]. Prothrombin Time, retrieved from https://medlineplus.gov/ency/article/003652 on August, 2022