A SURVIVAL ANALYSIS PROJECT OF TWO BIOMEDICAL DISEASES: IDENTIFYING FACTORS ASSOCIATED WITH MORTALITY OF PATIENTS WITH PRIMARY BILIARY CHOLANGITIS (PBC) AND CORONAVIRUS DISEASE 2019 (COVID-2019)

By

Osvaldo Ramirez

A capstone project submitted for Graduation with University Honors

MAY 06, 2022

University Honors
University of California, Riverside

APPROVED

Dr. Esra Kurum and Dr. Analisa Flores
Department of Statistics

Dr. Richard Cardullo, Howard H Hays Jr. Chair
University Honors

Abstract

The focus of this research project is to perform survival analysis on two medical data sets - the Mayo Clinic Primary Biliary Cholangitis (PBC) data and COVID-19 data collected by the Open COVID-19 Data Working Group. The PBC data was collected between 1974 and 1984 to identify factors (such as age, sex, and other comorbidities) affecting the risk of death for patients with PBC, a disease that destroys small bile ducts in the liver. The COVID-19 data, collected between February 2020 and March 2021, provides patient-level information that is used to understand factors (such as demographics and initial symptoms) contributing to the risk of death from COVID-19. The initial step in the analyses is data wrangling to format and prepare the data to be analyzed. Exploratory data analysis methods such as descriptive statistics and statistical graphics are employed to summarize and visualize the data. The survival data analysis methods include obtaining Kaplan-Meier estimators to explore how the risk of the event of interest, which is death for our data applications, changes over time and applying Cox proportional hazards models to determine the effect of each factor on the risk of death. The analysis is performed in R statistical software and the R code is publicly available.

Acknowledgements

This project would not have been possible without the help and support of many people and I would like to express my gratitude towards everyone.

I am thankful for Dr. Kurum taking the time to share her knowledge and resources about survival analysis with me, where I was able to learn to a great extent. I am also grateful for her suggestions when revising my document.

I would also like to express my gratitude to Dr. Flores for revising my writing and creating to-do lists to keep me updated on what to get done by the next week. This project would not have been possible without her. She was the person who introduced me to Dr. Kurum to help start this project.

Lastly, I would like to acknowledge my family, especially my parents and siblings, because they are my motivation to never give up.

Contents

List of Tables

List of Figures

1 Introduction

Survival data analysis methods are widely used in engineering and medical fields to identify factors associated with time to an event of interest such as death, occurrence of a disease, and failure of a machine. Traditional regression models are not suitable for this type of data for two reasons. First, the outcome, that is, time to an event, should always be positive and usually has a skewed distribution; therefore, methods that rely on normality are not directly applicable. Second, in this type of analysis, we need to take censoring into account. The defining feature of censoring is that the time to an event is not observable for all subjects; in other words, some subjects never developed a disease or experienced death during the study, so their time to event is missing in the data set. Traditional statistical models assume that we have complete information on all subjects, and when applied to this type of outcome, they would produce biased estimates. Due to these reasons, special methods for survival data analysis have been developed. Among these methods, in this project, we discuss Kaplan-Meier (KM) estimators, Cox proportional hazards models, and parametric survival models.

Our motivation for this project comes from two data sets: the Primary Biliary Cholangitis (PBC) data, which is gathered by Mayo Clinic on patients with this condition and the COVID-19 data gathered by the Open COVID-19 Working Group. PBC is a disease that slowly destroys the small bile ducts in the liver. The PBC data consists of 424 PBC patients and was administered for 10 years (1974-1984). The first 312 subjects were part of a randomized clinical trial and the remaining 112 subjects did not participate in the trial but still had their measurements recorded for the study. COVID-19 is an infectious disease that began to spread in December 2019. The COVID-19 data consists of over two million observations and was collected from 147 countries around the world. The goal of this project is to understand factors that are significant to risk of death from PBC and COVID-19.

Before applying the survival data analysis methods to the motivating data sets, we discuss the basic quantities for analysis of time-to-event outcomes. The four quantities for analysis of time-to-event outcomes are the survival function, hazard (risk) function, probability density function, and the cumulative distribution function. In particular, the survival function is the probability of an individual surviving until the time of interest. The hazard (risk) function is the probability that a subject who survived the time of interest might experience the event in the next instant. The probability density function is used to quantify probabilities linked to the time of interest. Finally, the cumulative distribution function is the probability that a subject survives at most until the time of interest.

In terms of advanced survival analysis methods, we first analyze the motivating data sets with KM estimator. In practice, this method is mostly utilized to investigate the effect of categorical variables on time to an event, in other words, to compare the survival probabilities among different groups. The limitation of KM estimators is that it can only be used on categorical covariates. Therefore, in order to identify the effects of quantitative covariates, in addition to the categorical covariates, we employ Cox proportional hazards model. To test the validity of the Cox proportional hazards model, diagnostics are performed using Schoenfeld and Martingale residuals. More specifically, the Schoenfeld residuals are employed to test the proportional hazards (PH) assumption, that is, the hazard/risk of an event does not change over time. In the condition that the PH assumption is not satisfied, one solution is to apply parametric survival models. The parametric survival models are applied to examine the significance and the effect of categorical and quantitative covariates on risk of an event through a pre-defined probability distribution, such as the Weibull or log-logistic distributions. The diagnostics for these models involve evaluating the validity of the probability distribution employed in the modeling scheme. We demonstrate an application of these models for both data sets.

The remainder of this project is organized as follows. A brief review of the main features of survival data, along with advanced statistical methods for this type of data, are presented in Section 2. In Section 3, we provide the analysis of PBC and COVID datasets via advanced survival data analysis methods. We conclude with a brief discussion in Section 4.

2 Survival Analysis

Survival analysis is an essential statistical tool, primarily used on clinical data when the main focus is to analyze time until a prespecified event of interest occurs. In these studies, the variable of interest is the time until that event, that is, survival time, which is a random variable denoted as T. The distribution of T is characterized by survival ($S(t)$), hazard ($h(t)$), cumulative hazard ($H(t)$), probability density $(f(t))$, and cumulative distribution $(F(t))$ functions. The survival function is the probability of a subject surviving to time t and is defined as

$$S(t) = Pr(T > t),$$

where $S(t)$ denotes the survival probability of the individual and $Pr(T > t)$ is the probability that the subject surpasses time t. Another function in survival data analysis is the hazard (risk) function, which is the probability that an individual who survived until time t might experience the event in the next instant. This function is given as follows,

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t},$$

where $h(t)\Delta t$ can be viewed as the "approximate" probability of a person experiencing the event in the next instant. A related measure is the cumulative hazard function, $H(t)$, which represents the accumulated probability of experiencing the event until time t and is defined by

$$H(t) = \int_0^t h(u)du = -\log[S(t)],$$

where log represents the natural logarithm. Next, the probability density function, $f(t)$, can be utilized to calculate probabilities associated with time t. Lastly, the cumulative distribution function, $F(t) = Pr(T \leq t)$, is the probability that a patient survives at most to time t. In practice, these functions are used to showcase different aspects of the distribution of the time-to-event outcome

T, and once one of these functions is known, the rest can be uniquely determined. In particular, the survival function can be found as the complement of the cumulative distribution function, that is, $S(t) = 1 - F(t)$. The survival function is also the integral of the probability density function, that is, $S(t) = Pr(T > t) = \int_t^\infty f(x)dt$; thus, $f(t) = -\frac{dS(t)}{dt}$ (Klein and Moeschberger, 2003).

An important consideration of time-to-event data is censoring, which creates a missing data challenge in the analysis. In particular, due to censoring, we do not observe the "true event time", denoted as T^*, for all subjects, instead, for some subjects, we only observe the censoring time C. The censoring time is usually recorded as the end of the study; in other words, the last time we know that the subject has not experienced the event. Let T_i represent the "observed event time" for the ith subject. If the subject has experienced the event during the study, $T_i = T_i^*$, since the true event time for the ith subject is known. However, if the subject has not experienced the event during the study, $T_i = C_i$. That is, the subject is censored and their true event time is unknown, only the last time they were free of the event has been recorded.

The incorporation of censoring in survival data analysis depends on the type of censoring. In particular, we have the following categorization: right, left, and interval. Right censoring occurs when not all subjects experience the event during the study time. In other words, there are subjects that survive the event during the study time and experience the event after the study. In this type of censoring, all subjects are free of the event at the beginning of the study. Left censoring happens when the event of interest has already occurred for some of the individuals before they are observed in the study. Interval censoring is when the time until an event of interest is not known accurately and is only known to fall into a particular interval (Radke, 2003).

2.1 Kaplan-Meier Estimator

Kaplan-Meier (KM) estimation is the first method we discuss in terms of survival data analysis methods. In this nonparametric approach, our interest is in estimating the survival function. Specifically, we estimate the survival function at the event times and observe how the survival probabilities change over time. The KM estimator is formulated as follows,

$$\hat{S}(t) = \prod_{t_i \leq t} 1 - \frac{d_i}{Y_i},$$

where t_i is the time where at least one event happened, d_i is the number of events that happened at time t_i, Y_i is the number of subjects that are at risk at time t_i, and $i = 1, \ldots, n$, with n as the number of subjects. To draw inference on the survival function $S(t)$, we use the $(1 - \alpha)100\%$ pointwise confidence interval

$$\hat{S}(t) \pm z_{\alpha/2} \times \{\hat{V}[\hat{S}(t)]\}^{1/2},$$

where $\hat{S}(t)$ is the estimated survival function, $z_{\alpha/2}$ is the $100 \times (1 - \alpha/2)$ percentile of the standard normal distribution, and $\{\hat{V}[\hat{S}(t)]\}^{1/2}$ is the estimated standard error of the KM estimator obtained using Greenwood's formula

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

Our goal in using the KM estimator is to illustrate the effect of variables of interest on survival time and investigate their significance.

2.2 Cox Proportional Hazards Model

The second method we consider in this project is the Cox proportional hazards model. This approach allows researchers to investigate the effects of predictors (risk factors) on survival time. The advantage of this method over the KM estimator is it can include multiple predictors, which can be both categorical and numeric, whereas KM estimation is limited to observing the effects of categorical predictors only. Let $h(t)$ be the hazard/risk of an event at time t. That is, as previously defined, $h(t)$ is the probability that an individual might experience the event at the next instant. The Cox proportional hazards model (Cox, 1972) is given as

$$h(t) = h_0(t)\exp(Z^{\mathrm{T}}\beta), \quad (1)$$

where $h_0(t)$ is the baseline risk/hazard of the event indicating the hazard/risk of an event when all covariates/factors in the model are equal to zero, $Z = (Z_1, \dots, Z_p)^{\mathrm{T}}$ denotes the covariates with $\beta = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ as the vector of corresponding coefficients and p as the number of covariates. Our goal in this model is to estimate the coefficients in β, which quantify the effect of each factor on the risk of the event, via the partial likelihood approach proposed by Cox (1972). Let $t_1 < t_2 < \dots < t_D$ denote the ordered event times and $R(t_D)$ be the risk set at the Dth event time t_D, that is, the set of all individuals in the study that have not experienced the event until t_D. In order to construct the partial likelihood, we start with calculating the probability of failure at the ith event time, t_i, using the following formula:

$$\frac{\exp\left[\sum_{k=1}^{p}\beta_k Z_{(i)k}\right]}{\sum_{j\in R(t_i)}\exp\left[\sum_{k=1}^{p}\beta_k Z_{jk}\right]},$$

where $Z_{(i)k}$ and Z_{jk} are the covariates measured at time t_i for ith subject whose failure time is t_i and jth subject who is in the risk set at time t_i, respectively, with the corresponding regression coefficient β_k, $i = 1, \dots, n$ with n as the total number of subjects, $k = 1, \dots, p$ with p as the total number of covariates, and j is the index for the subjects within the risk set $R(t_i)$ at time t_i. Once these probabilities are multiplied over all individuals, we obtain the partial likelihood as follows:

$$L(\beta) = \prod_{i=1}^{D}\frac{\exp\left[\sum_{k=1}^{p}\beta_k Z_{(i)k}\right]}{\sum_{j\in R(t_i)}\exp\left[\sum_{k=1}^{p}\beta_k Z_{jk}\right]},$$

where D is the total number of events. We obtain the estimated coefficients $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^{\mathrm{T}}$ by maximizing the partial log-likelihood,

$$LL(\beta) = \log L(\beta) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k Z_{(i)k} - \sum_{i=1}^{D} \log \left[\sum_{j \in R(t_i)} \exp \left(\sum_{k=1}^{p} \beta_k Z_{jk} \right) \right].$$

Similar to all statistical models, Cox proportional hazards model also has a set of assumptions about the data generating process, and estimation and inference would be misleading if these assumptions do not hold for the data. We employ the Schoenfeld, Martingale, deviance, and score residuals to check the assumptions of the Cox proportional hazards model and test the validity of these models for our data sets. The first residual we discuss is the Schoenfeld residuals (Schoenfeld, 1982). These residuals are calculated for each subject i, $i = 1, \dots, n$, at each failure time. For any subject i, who is in the risk set at a failure time t_m, the Schoenfeld residual is the difference between the covariate values for that subject, Z_i, and the weighted average of covariates for all subjects that are in the risk set $R(t_m)$. The sum of the Schoenfeld residuals over all subjects that fail at time t_m gives the Schoenfeld residuals corresponding to time t_m

$$r_{s,m} = \sum_{i \in R(t_m)} \delta_{im} \{Z_i - \bar{Z}(t_m)\}, \quad (2)$$

where δ_{im} equals to one if the subject fails at t_m and zero otherwise, and $\bar{Z}(t_m)$, which is the weighted average of the covariate values for individuals in the risk set $R(t_m)$ with weights $w_j(t_m)$, is defined as

$$\bar{Z}(t_m) = \sum_{j \in R(t_m)} Z_j w_j(t_m) \text{ with } w_j(t_m) = \frac{\exp(\hat{\beta}^T Z_j)}{\sum_{l \in R(t_m)} \exp(\hat{\beta}^T Z_l)}.$$

The main idea behind the formulation of the Schoenfeld residuals is the same as the construction of the partial likelihood, that is, comparing the covariates of a subject that experienced the event at a certain time (t_m) to the rest of the subjects that are in the risk set at that time point. The goal of the Schoenfeld residuals is to check for the proportional hazards (PH) assumption, which is the assumption that each covariate has a multiplicative effect in the hazards function that does not change through time. When the PH assumption holds, the Schoenfeld residuals are uncorrelated and have a mean of zero.

The second residuals we focus on are Martingale residuals (Lagakos, 1981; Barlow and Prentice, 1988; Therneau et al., 1990, Fleming and Harrington, 1991) which are defined as

$$\hat{M}_i = \delta_i - \hat{H}_0(t_i) \exp \left(\sum_{k=1}^{p} Z_{ik} \hat{\beta}_k \right),$$

where for the ith subject, δ_i is the event indicator, Z_{ik} represents the kth covariate with the corresponding estimated coefficient $\hat{\beta}_k$, and $\hat{H}_0(t_i)$ is the cumulative baseline hazard rate calculated using Breslow's estimator at the event time t_i (Breslow, 1972) , with $k = 1, \dots, p$, and $i = 1, \dots, n$. The Martingale residuals are used to examine the overall model fit and whether transformations are required in covariates after the rest of the covariates have already been included in the model. Under the correct model formulation, the Martingale residuals exhibit a linear pattern.

The next step in our model diagnostics is the identification of outliers. In terms of survival data analysis, an outlier is an unusual failure-time observation given the covariate values; in other words, these are subjects that our model does not fit appropriately. To determine potential outliers in our analysis, we utilize deviance residuals (Therneau et al., 1990)

$$D_i = sign[\hat{M}_i]\{-2[\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)]\}^{1/2},$$

where for the ith subject, $sign[\hat{M}_i]$ is the sign of the Martingale residual, δ_i is the event indicator, and $i = 1, \dots, n$. Note that although Martingale residuals are used to calculate the deviance residuals, they cannot be directly used to identify outliers as their distribution is heavily skewed. Deviance residuals can be seen as standardized Martingale residuals, which are more symmetrical around zero, and therefore, make them a better measure of outliers. In addition, the distribution of deviance residuals can be approximated by the Gaussian distribution; thus, we can assess an observation as an outlier if their deviance residual is outside the range of (-3, 3) or even (-2.5, 2.5), to be more conservative.

The last step in model diagnostics is to establish influential observations. An observation is identified as influential for a covariate if it strongly influences the estimated regression coefficient corresponding to that covariate. In order to determine these observations, we can either use delta-beta values or score residuals. Delta-beta values (Belsley, 1980) are obtained as follows,

$$\Delta_{ik} = \hat{\beta}_k - \hat{\beta}_k^{(i)}$$

where Δ_{ik} is the delta-beta value for the ith subject kth coefficient, $\hat{\beta}_k$ is the estimate of β_k from the whole data set, and $\hat{\beta}_k^{(i)}$ is the estimate of β_k from the data set with the ith subject removed. Delta-beta values indicate which subject or subjects are influential for the kth covariate. An ideal plot has delta-beta values symmetric around zero; influential observations deviate from the symmetry.

The score residual for the ith subject and the kth covariate is computed as follows (Cain and Lange, 1984; Reid and Crépeau, 1985)

$$S_{ik} = \delta_i[Z_{ik} - \bar{Z}_k(t_i)] - \sum_{t_b \leq t_i} [Z_{ik} - \bar{Z}_k(t_b)] \exp(\hat{\beta}^{\mathbf{T}} Z_i)[\hat{H}_0(t_b) - \hat{H}_0(t_{b-1})],$$

where $\delta_i[Z_{ik} - \bar{Z}_k(t_i)]$ is the Schoenfeld residual for the kth covariate (see equation 2), that is, the difference between covariate value Z_{ik} at the failure time t_i and the expected value of the covariates at t_i, $\hat{\beta}$ are the estimated coefficients, Z_i is all the covariates for subject i, t_b is any time point before the event time t_i, and $\hat{H}_0(.)$ is the estimated cumulative baseline hazard with $i = 1, \dots, n$ and $k = 1, \dots, p$. We calculate the score residuals with respect to every covariate in the Cox model. A large magnitude of the score residual of an individual with respect to a particular covariate indicates heavy influence of that individual in the estimation of the regression coefficient of that covariate. Score residuals must be symmetric around zero to indicate that there are no influential observations.

2.3 Parametric Survival Model with Weibull distribution

The third method is parametric survival models. Parametric survival models are used to estimate survival probability for censored data through a specified probability distribution (generally the Weibull or the log-logistic distribution). In practice, these models are often utilized when the proportional hazard assumption for the Cox models is not satisfied. In our data applications, we employ the Weibull distribution as under this distribution we can flexibly model the hazard function, that is, the hazard can be allowed to increase, decrease, or stay constant over time. In other words, the parameter of the Weibull distribution allows us to model multiple situations in our practical world.

The result is a log linear model with parameters $\lambda = exp(-\mu/\sigma)$ and $\sigma = 1/\alpha$. Under this framework, the time to event, T, is modeled as

$$Y = \log T = \mu + \sigma W,$$

where W is the apex value distribution with the probability density function, $f_W(w) = exp(w - e^w)$, and survival function, $S_W(w) = exp(-e^w)$. Therefore, the probability density and survival functions of Y, respectively, are obtained as

$$f_Y(y) = (1/\sigma) exp\left[(y-\mu)/\sigma - e^{[(y-\mu)/\sigma]}\right],$$

$$S_Y(y) = exp\left(-e^{[(y-\mu)/\sigma]}\right).$$

The estimation of the parameters μ and σ are performed via the maximum likelihood approach.The likelihood function for a parametric model with Weibull distribution is calculated as follows,

$$L(\beta) = \prod_{i=1}^{n} [f_Y(y_i)]^{\delta_i} [S_Y(y)]^{(1-\delta_i)},$$

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{1}{\sigma} f_W \left(\frac{y_i - \mu}{\sigma} \right) \right]^{\delta_i} \left[S_W \left(\frac{y_i - \mu}{\sigma} \right) \right]^{(1-\delta_i)} .$$

The parameters μ and σ are estimated using numerical algorithms such as the Newton-Raphson algorithm. Although the modeling scheme presented above does not include predictors, it is possible to incorporate them as follows

$$Y = \mu + \gamma^{\mathrm{T}} Z + \sigma W,$$

where γ^{T} is a vector of regression covariates, Z is the baseline time to event divided by the acceleration factor and W is the conventional extreme value distribution, which leads to a proportional hazards model with Weibull distribution, calculated as follows,

$$h(t|Z) = (\alpha \lambda t^{\alpha-1}) exp(\beta^{\mathrm{T}} Z),$$

where $h(t|z)$ is the baseline risk of the event indicating the hazard of an event when all covariates in the model are equal to zero, $\alpha = 1/\sigma$ at time t, $\lambda = exp(-\mu/\sigma)$, $\beta = -\gamma/\sigma$, and $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ indicates the covariates with $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ as the vector of corresponding coefficients and p as the number of covariates. To verify that the Weibull distribution is the correct distribution to use for our data set, we first fit an accelerated failure time (AFT) model with Weibull distribution (Wei, 1992), which explains the relationship between $Y = logT$ and covariates as follws,

$$Y_i = Z_i^{\mathrm{T}} \beta + W_i, \quad (3)$$

where Y_i is the log of T_i, Z_i is a vector of covariates with β as the corresponding coefficients, and W_i are the residuals that are assumed to follow a Weibull distribution. After residuals from the AFT model (see equation 3) are obtained, we compare them to the residuals obtained via a KM estimator. If the KM plot demonstrates that the AFT residuals obtained from equation 3 are covered by the 95% confidence interval of the KM estimate, the Weibull distribution fits the data strongly.

3 Data Analysis

3.1 Data Description

3.1.1 Mayo Clinic Primary Biliary Cholangitis (PBC) Data

The Mayo Clinic Primary Biliary Cholangitis (PBC) data focuses on patients who have been diagnosed with (PBC), a disease that affects the small bile ducts in the liver. The study followed patients for a ten-year period from 1974 to 1984. Of 424 total patients that participated in the study, 312 cases in the data set participated in the randomized trial. The rest of the patients, 112 cases, did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. In this report, we study 393 cases out of the 424 cases as the patients that had a transplant were excluded from the analysis. We aim to examine the effect of risk factors such as existing comorbidities and demographics on the survival of patients with PBC. This data set is publicly available in the survival package of R statistical software (R Core Team, 2020).

In the PBC study, among the 312 patients that were on the randomized trial, 148 were given D-penicillamine, 145 were given a placebo, and the rest were not randomized. According to Figure 1, we observe a small difference between the survival rates of patients in treatment (56%) and placebo (59%) groups.

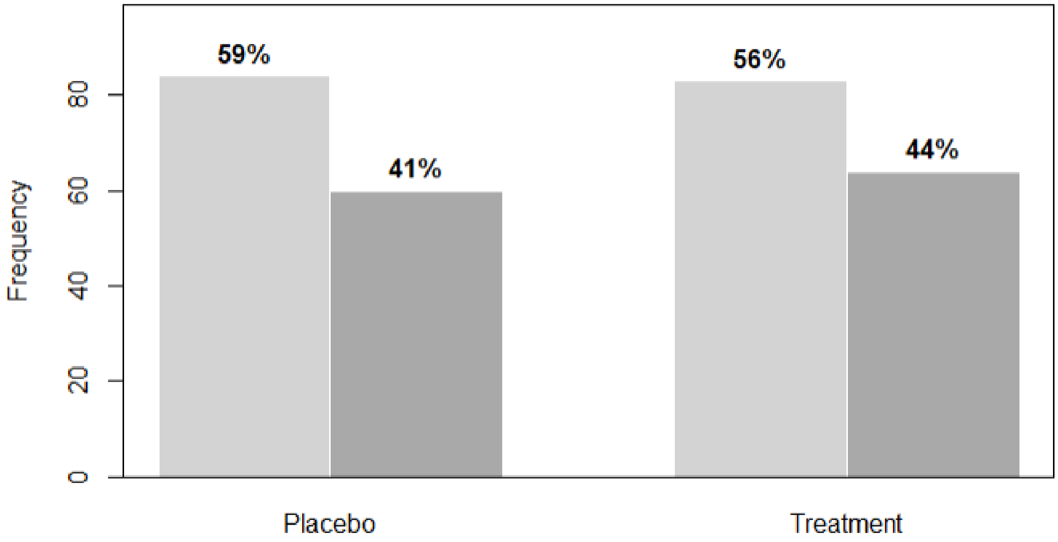

Figure 1: The light grey color represents patients who survived or were censored and the dark grey color represents patients who died during the study. The percentages refer to relative frequency within each group.

Patient-level covariate information for this study is summarized in Table 1. According to this table, the study cohort included patients with a mean age of 50.528 (SD 10.556), where 88.7% were females. The average levels of albumin and AST, which are proteins made by the liver, were calculated as 3.519 (SD 0.424) and 121.893 (SD 57.898), respectively. The mean bilirubin (a yellow pigment that is caused by red blood cells breaking down) and copper levels are 3.280 (SD 4.660) and 95.928 (SD 84.498), respectively. Finally, another important factor in liver health, that is, the standardized blood clotting time (protime) is recorded as 10.751 (SD 1.024) on average.

Table 1: Descriptive statistics of PBC data.

Variable	Mean / Count	SD / Percent
Age	50.528	10.556
Female	258	0.887

Variable	Mean / Count	SD / Percent
Albumin (g/dl)	3.519	0.424
AST (U/ml)	121.893	57.898
Bilirubin (mg/dl)	3.280	4.660
Copper (ug/day)	95.928	84.498
Protime	10.751	1.024

Ascites, which is the accumulation of fluid in the peritoneal cavity causing abdominal swelling, was also observed in patients with PBC. Figure 2 (a) illustrates that patients with ascites (4%) have a lower survival rate than patients without ascites (62%). PBC has four stages - 1, 2, 3, 4 - based on the amount of damage present in the liver. More specifically, stage 1 causes inflammation to the portal areas of the liver, stage 2 causes inflammation and fibrosis to the portal and periportal areas of the liver, stage 3 is bridging fibrosis, and stage 4 is cirrhosis. Figure 2 (b) highlights that stages 1-3 have more survivors (94%, 74%, 62%) than deaths (6%, 26%, 38%) but an increase in deaths can be seen as the PBC stages progress from 1 to 3. On the other hand, we observe more deaths (64%) than survivors (36%) in stage 4.

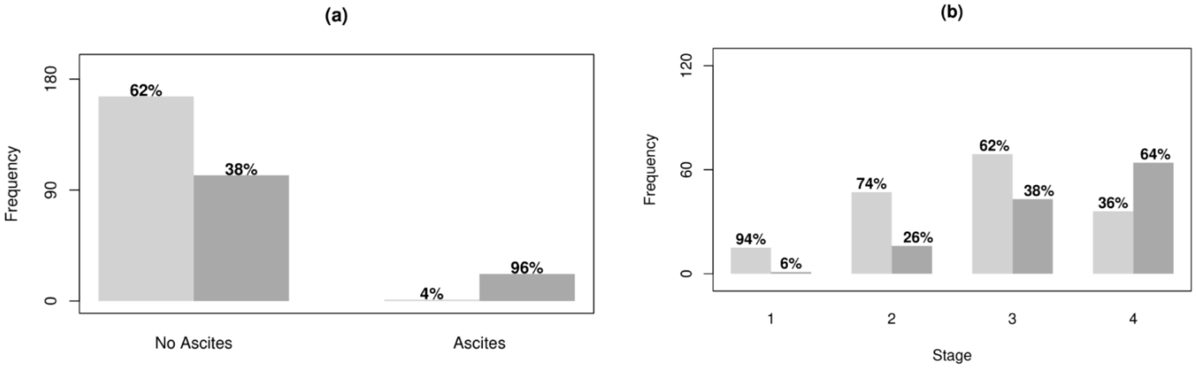

Figure 2: These plots display the frequency of patients (a) with/without ascites and (b) in each stage of PBC. The light grey color represents patients who survived or have been censored while the dark grey color represents patients who have died in the study. The percentages refer to relative frequency within each group.

3.1.2 COVID-19 Data

The COVID-19 data set is collected by the Open COVID-19 Data Working Group (Xu et al., 2020). COVID-19 is an infectious disease that began to spread in the year 2020. One of the most notable COVID-19 complications is that infected people have respiratory problems. The COVID-19 data, collected between February 2020 and March 2021, provides patient-level information including age and sex. The data consists of subjects from 140 countries and includes information on the initial symptoms of the patient and locations they traveled. The goal of our analysis is to identify significant factors that contribute to death caused by COVID-19 across various nations.

Table 2 provides descriptive statistics of age, sex, and chronic disease for patients with COVID-19. The mean age in the study was recorded as 58.720 (SD 18.434). Of all patients in the study, 61.2% were males and 6.3% had a chronic disease present.

Table 2: Descriptive statistics of COVID-19 data

Variable	Mean / Count	SD / Percent
Age	58.720	18.434
Male	477	0.612
Chronic Disease Present	49	0.063

Chronic disease refers to any illness that lasts for 1 year or longer and is in constant need of medication such as diabetes, cancer, and stroke. According to Figure 3 (a), 63% of patients with chronic disease and 79% of patients without a chronic disease died during the study.

In the examination of the COVID-19 data, the difference in survival rates of females and males is negligible with 22% and 21% among males and females, respectively. Figure 3 (b) shows that for both the female and male sample populations, there are more deaths than there are survivors.

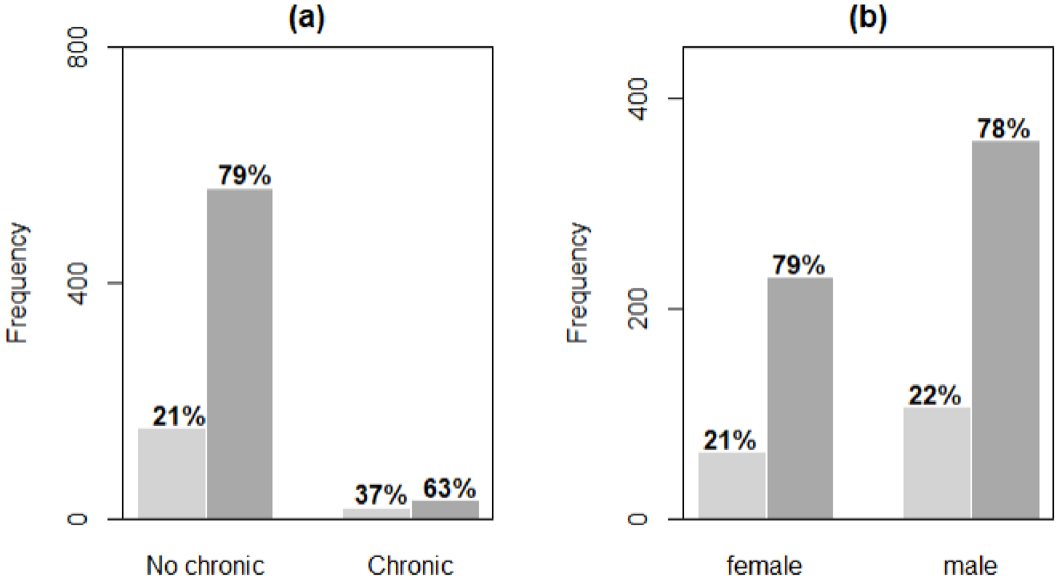

Figure 3: In barplot (a) and (b) the light grey color represents patients who survived or have been discharged while the dark grey color represents patients who have died in the study. The percentages refer to relative frequency within each group.

3.2 Data Application

3.2.1 Mayo Clinic Primary Biliary Cholangitis (PBC) Data

3.2.1.1 Kaplan-Meier (KM) Estimators In this section, we perform survival data analysis of the PBC data using the KM estimator. As discussed in Section 2.1, the KM estimator can be used to explore the difference in survival probabilities between the levels of a categorical variable such as stage of the disease, ascites, sex, and treatment in the PBC data. Figure 4 shows the KM estimates for (a) stage, (b) ascites, (c) sex, and (d) treatment. According to Figure 4 (a), the survival probability of patients decreases as the stage in PBC increases. Next, in Figure 4 (b), the survival rate of patients with ascites dramatically decreases as time progresses compared to patients that do not have ascites. None of the patients with ascites survived until the end of the study. Lastly, Figure 4 (c) and (d) illustrate that survival probabilities of sex and treatment vs. placebo are not different because they intersect throughout the years.

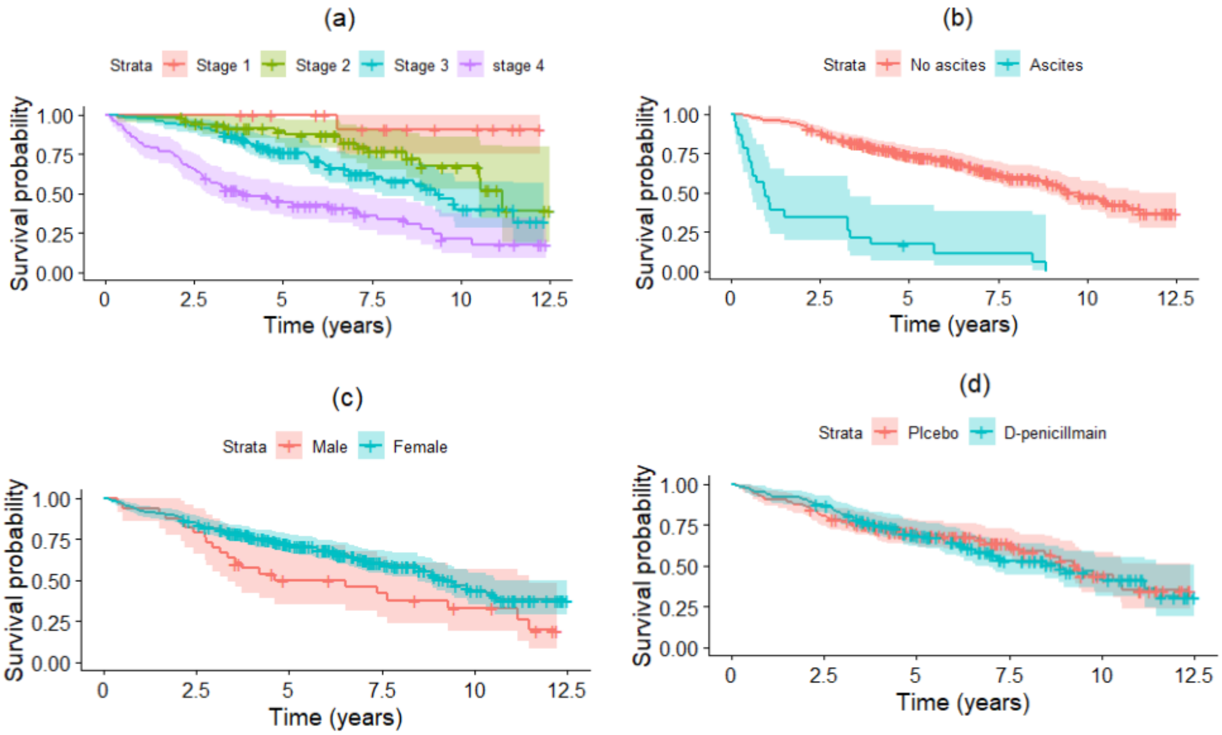

Figure 4: KM estimator for (a) stage, (b) ascites,(c) sex, and (d) treatment.

3.2.1.2 Cox Proportional Hazards Model Although the KM estimator is informative in identifying the significance of categorical variables, in order to explore the effect of continuous and categorical variables while controlling for other risk factors, we employ Cox proportional hazards model. The PBC study originally contained 17 predictor variables including sex, amount of albumin, amount of copper, triglycerides, and cholesterol. Predictors were systematically removed to find the optimal Cox proportional hazards model, that is, the model with the smallest Akaike Information Criterion (AIC). The final set of predictor variables included in the optimal model are albumin, aspartate aminotransferase (AST), bilirubin, copper, standardised blood clotting time (protime), age, ascites and PBC stage. The optimal model is given as:

$$h(t) = h_0(t)exp(z_1\beta_1+z_2\beta_2+z_3\beta_3+z_4\beta_4+z_5\beta_5+z_6\beta_6+z_7\beta_7+z_8\beta_8+z_4{}^*z_5\beta_9+z_4{}^*z_6\beta_{10}+z_7{}^*z_8\beta_{11}), \quad (4$$

where $h_0(t)$ is the baseline hazard function, z_1 = AST, z_2 = bilirubin, z_3 = protime, z_4 = age, z_5 = copper, z_6 = albumin, z_7 = stage, z_8 = ascites, and β_k are the corresponding regression coefficients with $k = 1, ..., 11$. Table 3 presents the results of our model fit. We observe that all covariates in the model, with the exception of albumin, are statistically significant (p-value<0.05). Albumin is marginally significant with a p-value between 0.05 and 0.10.

Table 3: Results of the Cox proportional hazards model (4)

Variable	estimate($\hat{\beta}$)	exp($\hat{\beta}$)	SE($\hat{\beta}$)	Z	P-value
AST	0.005	1.005	0.002	3.216	0.001
Bilirubin	0.095	1.100	0.018	5.162	0.000
Protime	0.311	1.364	0.090	3.441	0.001
Age	0.282	1.326	0.091	3.114	0.002
Copper	0.021	1.021	0.005	4.257	0.000

Variable	estimate($\widehat{\beta}$)	exp($\widehat{\beta}$)	SE($\widehat{\beta}$)	Z	P-value
Albumin	2.453	11.627	1.368	1.794	0.073
Stage	0.486	1.626	0.141	3.450	0.001
Ascites	3.532	34.193	1.570	2.250	0.024
Age*Copper	-3.138e-04	1.000	0.000	-3.583	0.000
Age*Albumin	-0.061	0.941	0.025	-2.461	0.014
Stage*Ascites	-0.860	0.423	0.430	-2.003	0.045

3.2.1.3 Diagnostics The diagnostics for the Cox Proportional Hazards presented in model (4) are performed via the residuals as described below.

3.2.1.3.1 Schoenfeld Residuals The Schoenfeld residuals proportional hazards assumption table (Table 4) identifies multiple violations in the Cox proportional hazards model (4). The variables bilirubin, protime, and stage have a p-values below the significance level, 0.05. Thus, the null hypothesis, which indicates the assumption is valid, is rejected.

Table 4: Results of Schoenfeld test for model (4)

Variable	Chi.sq	df	P-value
AST	2.2384	1	0.1346
Bilirubin	8.3838	1	0.0038
Protime	5.0019	1	0.0253
Age	1.2896	1	0.2561
Copper	0.0006	1	0.9805
Albumin	0.5787	1	0.4468

Variable	Chi.sq	df	P-value
Stage	4.5812	1	0.0323
Ascites	1.1418	1	0.2853
Age*Copper	0.0471	1	0.8281
Age*Albumin	0.0753	1	0.7838
Stage*Ascites	1.2022	1	0.2729
Global	26.3125	11	0.0058

3.2.1.3.2 Martingale Residuals Martingale residuals were employed to check for the true functional form for the quantitative covariates, which is only age in this application. To satisfy the true functional form for each particular covariate, the plotted residuals must show linearity. According to Figure 5, our model meets the assumption of linearity because the Martingale residuals are close to zero and illustrate linearity.

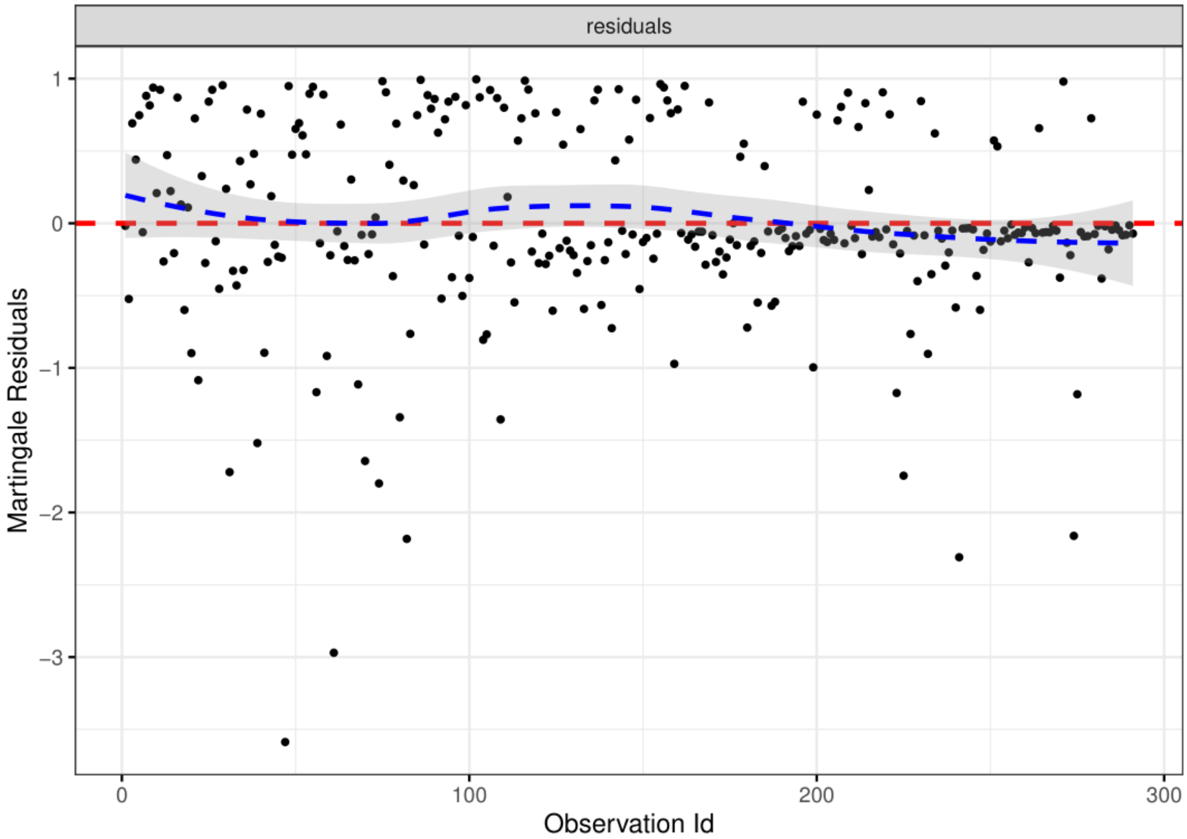

Figure 5: Martingale residuals for the covariate age in model (4). The red dotted line is a horizontal line at zero and the blue dotted line is line of best fit of the Martingale residuals.

3.2.1.4 Parametric Survival Model with Weibull Distribution for PBC data application In the previous Cox proportional hazards model (4), multiple covariates failed to satisfy the proportional hazards (PH) assumption (refer to the p-values smaller than 0.05 in Table 4). The two most common approaches to handling non-proportionality when quantitative variables fail to satisfy the PH assumption is to either include interactions between these covariates and time or employ a different model such as the parametric survival models. We preferred the latter as the former approach would be computationally intensive given that we would need to include several interaction terms in the Cox model. In the parametric survival models we fit, we assumed the baseline hazard function follows a Weibull distribution. This parametric survival model takes the following form,

$$h(t|Z) = \alpha\lambda t^{\alpha-1} exp(z_1\beta_1+z_2\beta_2+z_3\beta_3+z_4\beta_4+z_5\beta_5+z_6\beta_6+z_7\beta_7+z_8\beta_8+z_4{}^*z_5\beta_9+z_4{}^*z_6\beta_{10}+z_7{}^*z_8\beta_{11}),$$

where $h(t|Z)$ is the expected hazard at time t, $\alpha\lambda t^{\alpha-1}$ represents the hazard when all predictors are equal to zero, z_1 = AST, z_2 = bilirubin, z_3 = protime, z_4 = age, z_5 = copper, z_6 = albumin, z_7 = stage, z_8 = ascites, and $\beta_1, \dots, \beta_{11}$ are the regression coefficients, respectively. Table 5 indicates that AST, bilirubin, protime, stage, ascites, age, copper, the interaction of age * copper, and the interaction of age * albumin are significant since the p-values are below the significance level, 0.05. Also, albumin and the interaction between stage and ascites are marginally significant since their p-values are between 0.05 and 0.10. It can be concluded that AST, bilirubin, and protime contribute to a lower relative risk of death by 0.3%, 5.5%, and 18.4% respectively (when controlling for other factors). To calculate the effect of copper, albumin, and ascites both the interaction and the main effect are used for calculation. Controlling for all other covariates in the model, copper levels increase by one unit, copper leads to a decreased relative risk of death by 1.3% (exp(1.92e-04 - 0.013) = 0.987), if albumin levels increase by one unit, albumin contributes to a decreased risk of death by 78.4%. (exp(-1.570 + 0.038) = 0.216). Furthermore, as a patient's condition in terms of stage of the disease gets worse, we estimate that the hazard of death increases by 20.6% (exp(0.485-0.298) = 1.206). Finally, subjects who have ascites have a decreased hazard of 78.9% (exp(-2.040 + 0.485) = 0.211).

Table 5: Results of parametric survival model (5)

Variable	estimate($\hat{\beta}$)	exp($\hat{\beta}$)	SE($\hat{\beta}$)	Z	P-value
AST	-0.003	0.997	0.001	-3.32	0.001
Bilirubin	-0.057	0.945	0.010	-5.58	0.000
Protime	-0.203	0.816	0.055	-3.71	0.000

Variable	estimate($\widehat{\beta}$)	exp($\widehat{\beta}$)	SE($\widehat{\beta}$)	Z	P-value
Age	-0.175	0.839	0.054	-3.27	0.001
Copper	-0.013	0.987	0.003	-4.46	0.000
Albumin	-1.570	0.208	0.809	-1.94	0.052
Stage	-0.298	0.742	0.086	-3.46	0.001
Ascites	-2.040	0.130	0.970	-2.10	0.035
Age*Copper	1.92e-4	1.000	0.000	3.72	0.000
Age*Albumin	0.038	1.039	0.015	2.60	0.009
Stage*Ascites	0.485	1.624	0.265	1.83	0.067

3.2.1.5 Diagnostics The diagnostics for the parametric survival model with Weibull distribution (5) is performed as described in Section 2.3. Figure 6 demonstrates a comparison of a survival function corresponding to residuals obtained via an AFT model (solid red line) versus the KM estimate (solid black line). According to this figure, as the KM estimate and the survival function of the extreme value distribution align very well, the Weibull distribution was an appropriate choice for our parametric survival model.

3.2.2 COVID-19 Data

3.2.2.1 Kaplan-Meier Estimators In parallel to the previous section, KM estimators are also employed in the COVID-19 data to further investigate binary covariates sex and chronic disease. Both plots in Figure 7 illustrate that there is not a difference of the survival probabilities between males and females (a), nor between no chronic disease and chronic disease (b). Since neither covariate is a significant predictor for survival, further investigation is required using the Cox model.

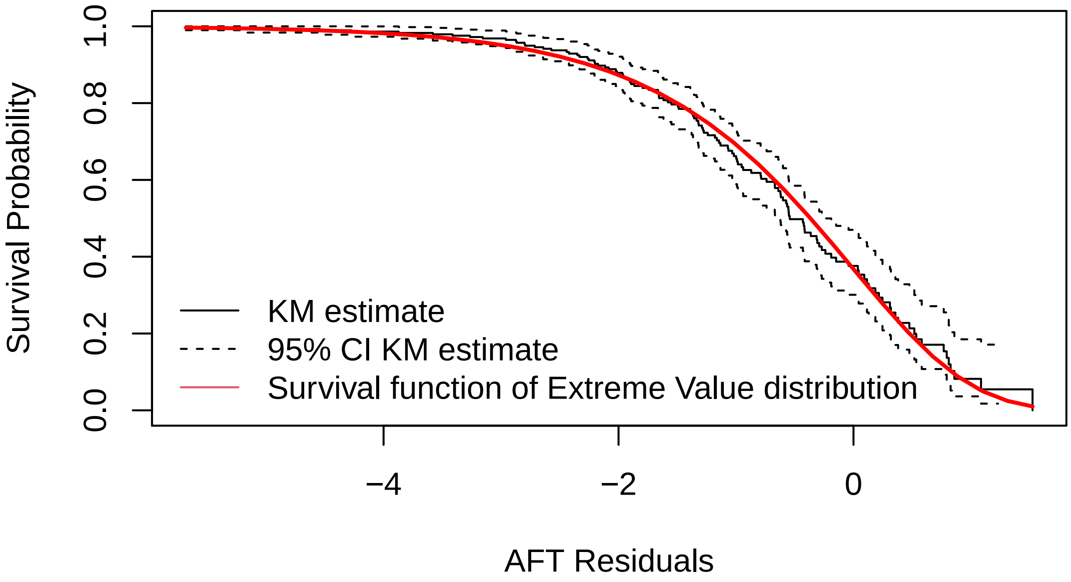

Figure 6: KM plot estimating AFT residuals of parametric survival model (5).

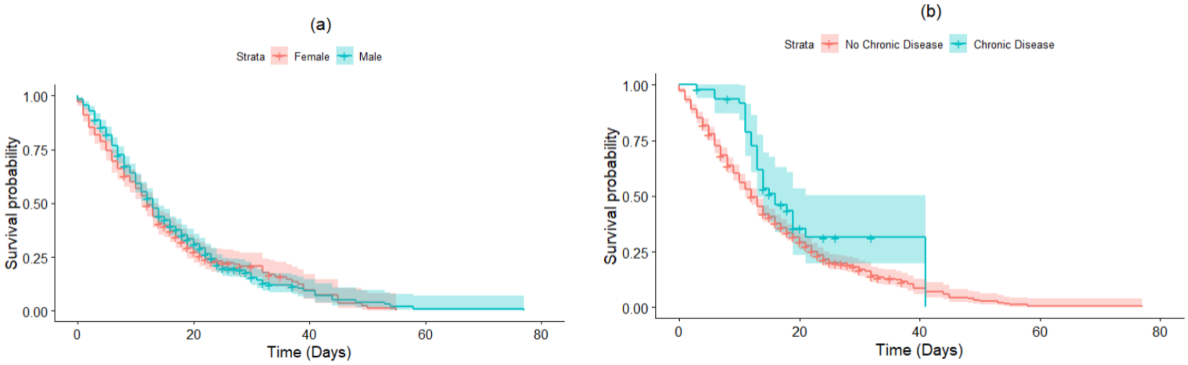

Figure 7: KM estimator for (a) sex and (b) chronic disease.

3.2.2.2 Cox Proportional Hazards Model Patients in the COVID-19 data provided information on their age, symptoms, and demographics. To get the most favorable model, we performed the same methods that were used to obtain the Cox proportional hazard model for the PBC data and chose the model with the lowest AIC value. Our final model consists of predictors age, sex, and chronic disease, stated as follows:

$$h(t) = h_0(t)exp(z_1\beta_1 + z_2\beta_2 + z_3\beta_3), \quad (6)$$

where $h(t)$ is the expected hazard at time t, $h_0(t)$ represents the baseline hazard, z_1 = age, z_2 = sex, z_3 = chronic disease, and β are the corresponding coefficients. According to Table 6, the covariates age and chronic disease are significant because they have a p-value below 0.05. Although the covariate sex is not significant, it is kept in the model to control for its effect.

Table 6: Results of Cox proportional hazards model (6)

	estimate($\widehat{\beta}$)	exp($\widehat{\beta}$)	SE($\widehat{\beta}$)	Z	P-value
Age	0.023	1.023	0.003	8.808	0.000
Sex	0.026	1.026	0.085	0.303	0.762
Chronic Disease	-0.601	0.548	0.186	-3.229	0.001

3.2.2.3 Diagnostics The diagnostics for the Cox Proportional Hazards presented in model (6) are performed via the residuals as described below.

3.2.2.3.1 Schoenfeld Residuals The Schoenfeld residuals were utilized to check the proportional hazards (PH) assumption, that is, each covariate has a multiplicative effect in the hazards function that does not change through time. In Table 7, all of the p-values from the covariates are below 0.05 indicating that the PH assumption is not satisfied.

Table 7: Results of Schoenfeld test for model (6)

Variable	Chi.sq	df	P-value
Age	16.3581	1	0.0001
Sex	4.2521	1	0.0392
Chronic Disease	10.3311	1	0.0013
Global	32.3064	3	0.0000

3.2.2.3.2 Martingale Residuals Similar to the PBC data, Martingale residuals were also used to check for the true functional form for the quantitative covariate, which is age for this data. According to Figure 8, our model satisfies the assumption of linearity since the Martingale residuals are close to zero and illustrate linearity.

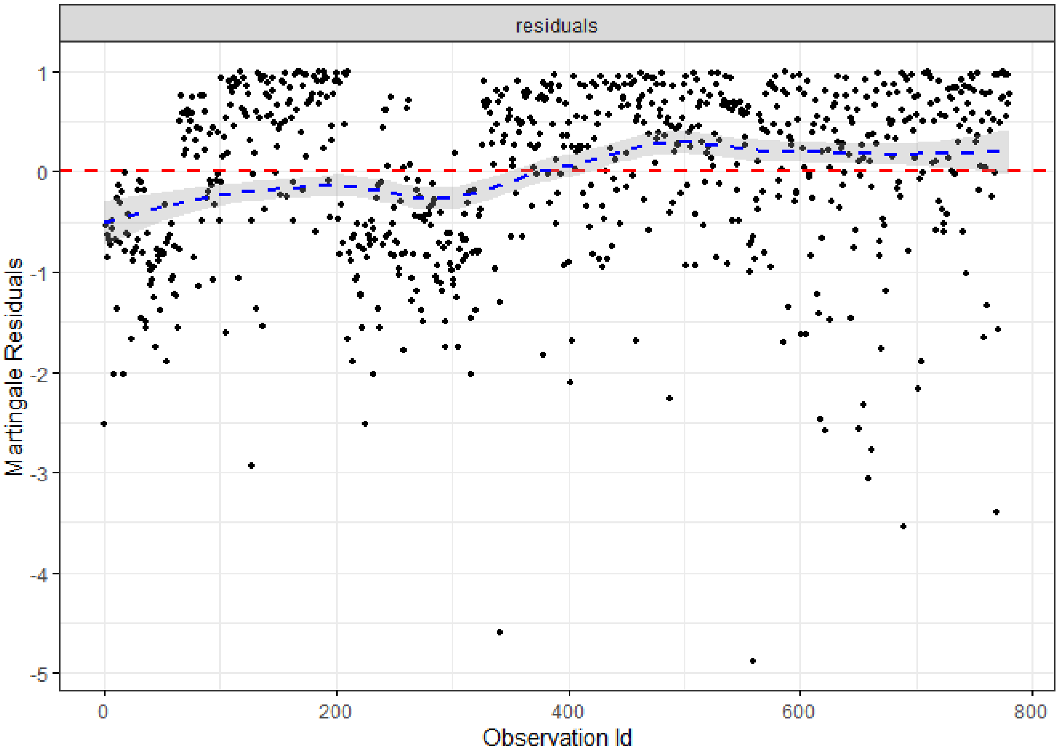

Figure 8: Martingale residuals for the covariate age in model (6). The red dotted line is a horizontal line at zero and the blue dotted line is line of best fit of the Martingale residuals.

3.2.2.4 Parametric Survival model with Weibull Distribution for COVID-19 data application In the Cox proportional hazards model in (6), multiple covariates failed to satisfy the proportional hazards (PH) assumption through the Schoenfeld test because the p-values were below 0.05. Since the Cox proportional hazards model is not valid, we utilized a parametric survival model with Weibull distribution. The model is as follows,

$$h(t|Z) = \alpha\lambda t^{\alpha-1} exp(z_1\beta_1 + z_2\beta_2 + z_3\beta_3), \quad (7)$$

where $h(t|Z)$ is the expected hazard at time t, $\alpha\lambda t^{\alpha-1}$ represents the hazard when all predictors are equal to zero, z_1 = age, z_2 = sex, z_3 = chronic disease. In Table 8, chronic disease and age are significant as they have p-values below 0.05. Although the covariate sex is insignificant, we kept it in our model to control for its effect. The results in Table 8 show that as patients get one year older, they have a decreased relative risk of death by 1.8% and patients who have a chronic disease have an increased relative risk of death by 55.6% compared to subjects who do not have a chronic disease (when controlling for other factors).

Table 8: Results of parametric survival model in (7)

Variable	estimate($\hat{\beta}$)	exp($\hat{\beta}$)	SE($\hat{\beta}$)	Z	P-value
Age	-0.018	0.982	0.002	-8.529	0.000
Sex	-0.025	0.975	0.066	-0.383	0.702
Chronic Disease	0.442	1.556	0.145	3.051	0.002

3.2.2.5 Diagnostics As previously discussed, the diagnostics for the parametric survival model with Weibull Distribution (7) is achieved by comparing a survival function corresponding to residuals from an AFT model (solid red line) against the KM estimate (solid black line). Figure 9 illustrates that the Weibull distribution fits the COVID-19 data well because the survival function of the extreme value distribution strongly aligns with the KM estimate.

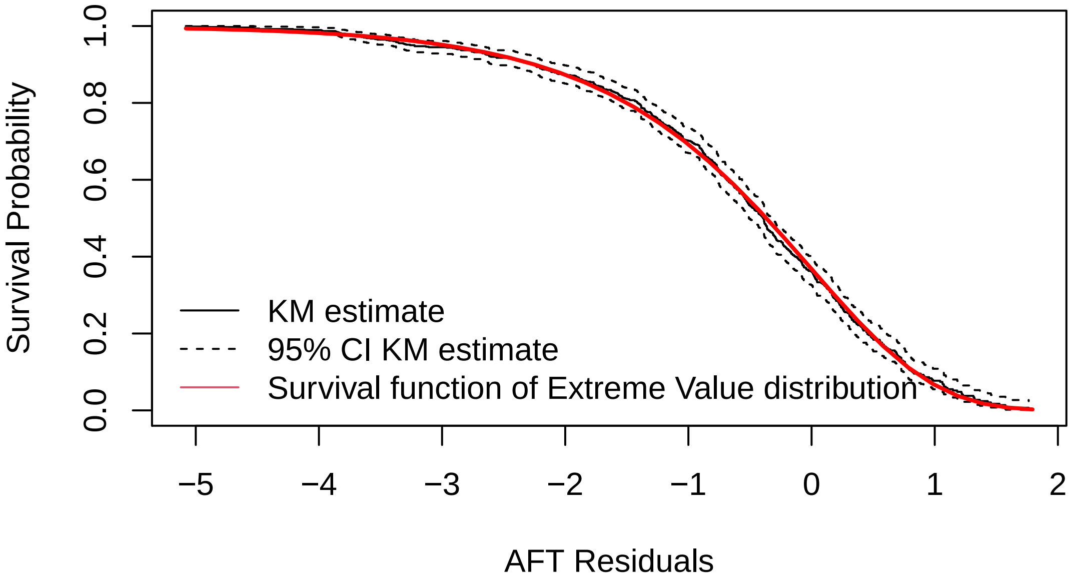

Figure 9: KM plot estimating AFT residuals of parametric survival model in (7).

4 Conclusion

In this project, we presented a brief overview of survival data analysis methods motivated by two medical data sets. In particular, we first applied KM estimation to study the estimated survival probabilities over time in both data sets. Our results for the PBC data indicated that the survival probability of patients decrease with later stages of PBC, patients with ascites have a significantly lower survival probability compared to patients with no ascites, survival probabilities of males are statistically different from females, and the survival probabilities of patients that were treated with D-penicillamine are statistically different from patients that were treated with a placebo. In terms of the COVID-19 data analysis, we observed that the survival probabilities of males are statistically different from females and the survival probabilities of patients with chronic disease are statistically different from patients without a chronic disease. However, due to the limitation of the KM estimation, that is, this method can only be used to examine the effects of categorical predictors, we also applied Cox proportional hazards model to each data. The main drawback we experienced in terms of the Cox model application is that our analysis failed to satisfy the proportional hazards assumption. We utilized parametric survival models with Weibull distribution to overcome this challenge and the fitted models satisfied the assumptions. As a result of our analysis of PBC data via these models, we concluded that albumin, AST, bilirubin, copper, protime, age, ascites, PBC stage, the interaction of age * copper, and age * albumin are significant factors to death rate of PBC, while albumin and stage * ascites are marginally significant. In particular, all of the covariates decreased the risk of death for those with PBC, except for the stage of the disease, which contributed to an increase in the hazard of death. For the COVID-19 data, we discovered that chronic disease and age are significant covariates in the death rate of COVID-19 patients. Specifically, COVID-19 patients have an increased risk of death due to chronic disease. The R codes for this project are publicly available at https://drive.google.com/drive/folders/0AChOwKbOm59dUk9PVA; the PBC data is available in the R survival package, and COVID-19 data can be accessed via https://www.nature.com/articles/s41597-020-0448-0.

5 References

Barlow, W. E., & Prentice, R. L. (1988). Residuals for relative risk regression. Biometrika, 75(1), 65-74.

Belsley, D. A. (1980). On the efficient computation of the nonlinear full-information maximum-likelihood estimator. Journal of Econometrics, 14(2), 203-225.

Breheny, P. (n.d.). Accelerated Failure Time Models. Iowa; The University of Iowa.

Breheny, P. (n.d.). Residuals and model diagnostics. Iowa; The University of Iowa.

Breslow, N. E. (1972). Discussion of Professor Cox's paper. J Royal Stat Soc B, 34, 216-217

Cain, K. C., & Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data. Biometrics, 493-499.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.

Fleming, T. R., & Harrington, D. P. (2011). Counting processes and survival analysis. John Wiley & Sons.

Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data (Vol. 2, pp. 3-5). New York: Springer.

Lagakos, S. W. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. Biometrika, 68(1), 93-98.

Reid, N., & Crépeau, H. (1985). Influence functions for proportional hazards regression. Biometrika, 72(1), 1-9.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239-241.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. Biometrika, 77(1), 147-160.

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In Modeling survival data: extending the Cox model (pp. 39-77). Springer, New York, NY.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 11(14-15), 1871-1879.

Xu et al. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. Scientific data, 7(1), 1-6.