Survival Analysis with Semi-parametric Mixture Cure Model

基于半参数混合治愈模型的生存分析

Yue Ju

ID: 1718235

Supervisor: Mu He

May 5, 2021

Abstract:

A mixture cure model is a survival model which combines cured probability and survival probability. In this study, we used a semi-parametric method to analyze the impact of risk factors on mortality of patients with COVID-19 in Wuhan or heart disease in Pakistan based on the mixture cure model and show the performance of the model by R package smcure.

摘要:

混合治愈模型是一种将治愈概率和生存概率相结合的生存模型。在本研究中，我们基于混合治愈模型，采用半参数方法分析一些危险因素对于武汉地区感染新冠肺炎的患者以及巴基斯坦地区心脏病患者死亡率的影响，并通过R包smcure展示模型。

Keywords: survival analysis, mixture cure model, COVID-19, heart failure.

Contents

I. Introduction

In survival data analysis, the traditional method generally holds that the observation object with the cut-off survival time, like those who obtain the full survival time, have the same chance to experience death at any observation point. They believed that each observed patient will die sooner or later, which means each of them will have the end event sooner or later. The end event "death" is not a death in a sense, it may be relapse, alleviate, or infectious disease infection. For example, in some medical disease researches, a recurrence of disease is usually the end event, especially in breast cancer. However, with the advancement of medical technology and the improvement of medical quality, the phenomenon of "long-term survivors" which refers to the experimenter's disease without recurrence during the observation period widely exists in the field of clinic researches. Many models are constantly coming up in the context of actual problems and various estimation methods have been constructed. A mixture cure model which combines the cured probability and survival distribution part is well known used in these researches. Due to the different expression of the survival function, the mixture cure model can be divided into different categories, there are two well-known models: proportional hazard(PH) mixture cure model and accelerated failure time(AFT) mixture cure model. Proportional hazard model is also called cox model which mainly focuses on the hazard ratio. Under the assumption, the hazard ratio in this model is proportional so that it can estimate the coefficient with the unknown distribution of hazard ratio at time 0. Accelerated failure time model mainly pays attention to the survival function based on an acceleration factor because some indicators may have an effect on reducing the time of survival and this can be considered as a speed of time in the model. In this paper, we mainly use the R package called smcure to briefly applying these two models to analyze two cases respectively: the COVID-19 case and the heart disease case.

The organization of the paper is as follows:

In the next section, we outline the literature about the mixture model and the background on the two cases. In section 3, we expound the related method and principle of survival function, cox model, EM algorithm estimator and the baseline on both parametric and non-parametric. We do the data analysis in section 4.

II. Literature Review

1. Mixture Cure Model

Many scholars like Taylor (1995), and Li *et al.* (2019) pointed out that an implicit assumption of the traditional survival analysis method is that given an ample follow-up time, all individuals will eventually experience the event of interest, which is considered as ”death”. The end event “death” actually is not a real death in a sense, it may be relapse, alleviate, or infectious disease infection. For example, in some medical disease researches, a recurrence of disease is usually the end event especially in the breast cancer. However, with the advances in medical care, a considerable proportion of patients have a chance to be cured and may never experience the event of interest. The application of cure models are particularly important for cancer research and in clinical studies of breast cancer, it is found that some patients no longer have recrudescence of the disease, said by Amico *et al.* (2019). According to Jiang *et al.* (2017), it is very likely that a part of patients responds well to the treatment that can be cured and does not experience the event of interest in the studies of prostate cancer and Hodgkin’s lymphoma. Li and Taylor (2002) also pointed that in the study of tonsil cancer, the Kaplan–Meier cure does not closed to zero which means a part of patients were cured after radiation therapy. These suggests that a fraction of the objects of observation will not or never experience the events even if the follow-up time is sufficient. Thus, the single survival model is not feasible to analyze such data, to respond to this situation more accurately, the researches on survival analysis model theory and methods have been continuously promoted and deepened. Early Boag (1949) and Berkson and Gage (1952) introduced the mixture cure model which can popular used to deal with such situations. In this model, the population can be divided into two parts, the “cure” part and the “uncure” part with survival function. And also it is widely accepted a mixture model with covariate effects, originally introduced by Boag (1949) and Farewell (1982). The covariates like x and z are independent that z can only have an impact on incidence(the proportion of cure) and x can only have an effect on latency. The latency part is a primary interest and a few various parametric distributions have been proposed to describe the survival function, such as Yamaguchi (1992) and Peng *et al.* (1998). If it can be expressed by several unknown parameters, the model will be estimated using different methods, for instance, Taylor (1995) a logistic regression model can be used in incidence part, a Kaplan-Meier method and EM algorithm method described by Larson and Dinse (1985) can be estimating the latency distribution. It is also common to use the maximum likelihood to get the estimators of S(t) and the details can be found in Maller and Zhou (1996). Although it is appealing, there are still several problems associated with it that need to be focused on. For example, Farewell (1986) and Laska and Meisner (1992) pointed out the identifiability problem and also Taylor (1995) supported that maximum likelihood estimators may be infinite and for the logistic-Weibull mixture model is too restrictively parametric in nature. In addition, Laska and Meisner (1992) presented a non-parametric generalized maximum likelihood (GML) method to estimate the mixture cure model and De Iorio *et al.* (2009) proposed a flexible nonparametric model based on the Dirichlet process model. However, these methods have no assumption of the population and need a large number of data so that they are at a great computational cost. Thus, a semi-parametric model is feasible and valid to reduce the model's dependence on the assumption and the burden of computing.

1.1 The proportional hazards (PH) mixture cure model

It is an improvement based on the Cox PH model, which adds the cured function in the model and in the literature, a couple of methods to estimate the model has been discussed. Fine and Gray (1999) introduced a new semi-parametric proportional hazard model by using a partial likelihood principle and weighting techniques based on the cumulative incidence and they compared the PH model with the data of breast cancer clinical trials. Also for the breast cancer cases, in order to deal with the situation met in practice more flexibly, Amico *et al.* (2019) proposed a single-index mixture cox model which can reduce hypothetical constraints and at the same time it can also avoid curse of dimensionality problems. It is worth noting that how to appropriately relax the restriction of the assumption and improve the model is still a problem that many studies focus on. For instance, Peng (2003) discussed about semi-parametric method which combines the cox model and EM algorithm approach, so that it can be applied in a broader area. They also showed an illustrative example of the patients with lymphoma. Apart from this, Wang and Zhou (2018) used simulation studies based on the population of Canada to illustrate that the semiparametric maximum likelihood estimators are better than any other estimators mentioned in the literature.

1.2 The accelerated failure time (AFT) mixture model

This model is also widely used in researches. It is a method of modeling failure time instead of a proportion hazard ratio. AFT model is mainly a model of survival time. It analyzes the patient's survival rate curve based on the habit of physics, and finds that the decline in survival rate is downward convex , so it may exist an acceleration, so people call it an accelerated failure time model. It is first applied to the accelerated life experiments by Pieruschka (1961) in 1961. The semi-parameter accelerated failure time cure model is more complex and its development is immature so that there is less research literature about the estimation methods on this model and there are also a lot of controversies in different areas. Taylor (1995) conducted a simulation study based on the datasets of radioactivity experiment and suggested that the semi-parametric is equivalently efficient as parametric in estimating the regression coefficient of incidence, and less efficient in latency part. Li and Taylor (2002) tried to develop an estimation method to determine an AFT model with unspecified baseline distribution. They pointed out the maximum likelihood estimation is still problematic for models that do not have specific distribution assumptions, so they proposed an EM algorithm method based on likelihood methods in their semi-parametric model. Additionally, for the AFT part, Zeng and Lin (2007) introduced a kernel-smoothed profile likelihood method to handle time-dependent covariates and by providing computationally feasible and statistically valid reasoning procedures, they made the accelerated failure time model as a more feasible alternative to the proportional hazard model. Besides, Zhang and Peng (2007) proposed a new estimation method which combined the EM algorithm and the rank-like estimator of the AFT model in M-step, and applying it to the study of bone marrow transplant patients.

2. Data background

The mixture cure model is widely used in many areas, including the treatment of cancer and other chronic diseases in recent years. Heart disease is a typical field of research. According to WHO, the number of people die from cardiovascular disease every year is more than any other cause of death. Almost all cardiovascular diseases eventually lead to heart failure, which is a a series of clinical symptoms and signs arising from a variety of reasons. Therefore, to find the causes of heart disease in order to obtain the effective treatment to control or cure the disease arouse wide concern around the world. According to Kumar *et al.* (2020), a study based on patients with heart failure (HF) in Pakistan's potential shrinkage period showed that ADHF was in hospital for one year, and the admission rate was 37.2%, with a total mortality rate of 27.5%. Ahmad *et al.* (2017) has also mentioned that in the population over 45 years old in Pakistan, 33% of them have high blood pressure, 25% have diabetes, and the number of coronary heart disease has reached more than 200,000 per year reported by Al-Shifa hospital. There are many factors will cause heart diseases, such as smoking, alcohol, high blood pressure, diabetes and hyperlipidemia. In Ahmad *et al.*'s paper, they used cox model to exclude some irrelevant factors like smoking and diabetes that is not the reason caused the heart failure based on a case study in area of Faisalabad. Moreover, Zahid *et al.* (2019) had made another analysis of that case, they pointed out that the reasons of heart failure is based on gender and used a lasso approach to select the informative predictors for the semi-parametric cox model. For example, anaemia is not considered a risk factor for male and ejection fraction, sodium, and platelets count are not considered for female. This paper, we retrospectively analysed this case by using mixture cure model to decide the related risk factors and estimate the survival probability.

In addition, the outbreak of Coronavirus Disease 2019 in late December 2019 has attracted worldwide attention. This is a new field but also a new challenge. The COVID-19 epidemic spreads suddenly with a high fatality rate at the early stage and until now there is still not an effective ways to detect, prevent and control in advance. According to data from the WHO, as of April 27, there were more than 147 million confirmed cases and 3,116,444 deaths worldwide, and the number is also increasing now. Moreover, the newly reported cases in the day is around 702,752, this large number of population and the uncertainty of the COVID-19 treatment process making the new crown pneumonia pandemic as a great pressure on medical infrastructure and health services. According to Salinas-Escudero *et al.* (2020), they use the survival analysis to survey the impact of COVID-19 for people in Mexico and they found that the death rate is higher for males and old people, and patients hospitalized in public health services due to Poor liquidity and other complications. However, to be encouraging, with the continuous deepening of research, the diagnosis and treatment plan for COVID-19 is constantly updated and improved. Yan *et al.* (2020) developed an XGBoost model based on machine learning to determine three indicators, which are LDH, hs-CRP and lymphocytes and also in their results, this model can predict the death rate of patients for ten days early and the accuracy exceeds 90%. Furthermore, according to the status of the cases admitted so far, most patients have a good prognosis, and a few patients are in critical condition. It is still important to find some predictable risk factors in order to provide opportunities for timely medical intervention. This analysis attempts to provide data support on COVID-19 mortality by using the mixture model to further analyze the death rate between gender and age group.

III. Methodology

1. Survival Model

Survival function is typically used to represent the probability of failure or death of some time-based systems. Let T represent a non-negative random variable, indicating the time until an event of interest occurs, then the **survival function** can be written as:

$$S(t) = Pr(T > t) = 1 - F(t).$$

The **hazard function** is the instantaneous rate at which events occur of individuals who survive at time t, given no previous events. It can be written as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Let T denote the expiration time of events and $S(t)$ be the total survival function of T, then the formula of **mixture cure model** is:

$$S(t) = 1 - \pi + \pi S_u(t). \tag{1}$$

where π is cured rate and $S_u(t)$ is a survival function modeled by the uncured patients.

Next, to better describe the cured rate model, the possible covariate variables which are denoted the effects on the probability of cure x and the effects on the latency distribution z can expand into the model. Then the equation (1) can be written as:

$$S(t|x, z) = 1 - \pi(z) + \pi(z)S_u(t|x). \tag{2}$$

where $\pi(z)$ is the probability of uncured patients depending on z, and $S_u(t|x)$ is the survival function of uncured subjects at time t depending on x, which is related to the age, gender, or other factors.

If the latency part is modeled by Cox PH model, then the **PHMC model** can be expressed by:

$$S(t|x, z) = 1 - \pi(z) + \pi(z)S_0(t|x)^{\exp(\beta x)}.$$

where $S_0(t|x)$ is a baseline survival function.

If the latency part is modeled by AFH model, then the **AFTMC model** can be expressed by:

$$S(t|x, z) = 1 - \pi(z) + \pi(z)S_0(te^{\beta x}).$$

2. Cox Model

Cox model, also known as 'proportional hazard model', is a semiparametric regression model proposed by the British statistician D.R.Cox (1972). The model use hazard ratio to investigate the link between time and the risk factors, the formula is:

$$h(t, \mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n).$$

The function has two parts, $h_0(t)$ is the orignal hazard function at time 0 which we also called 'baseline hazard'; the rest is a regression model of the covariates $\mathbf{x}$, which can be considered as age, gender and other factors and β is the estimated coefficients of x. Since the Cox regression model does not make any assumptions about $h_0(t)$, it is also a semi-parametric model.

If x_i is the value of each factor of the observation object in one group, and x_j is the value of each factor of the other group, the relative risk RR of these two groups can be calculated by formula:

$$RR = \frac{h_i(t,x)}{h_i(t,x)} = \frac{h_0(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{h_0(t) exp(\beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_p x_{jp})},$$

$$RR = exp(\beta_1(x_{i1} - x_{j1}) + \beta_2(x_{i2} - x_{j2}) + \cdots + \beta_1(x_{i3} - x_{j3})),$$

for i≠j, i,j=1,2,$\cdots$,n

It can be seen from the above formula that RR satisfies a constant ratio and has unrelated to time, so it is also called a proportional hazard model.

$$log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}),$$

where the PH model is also a linear regression for the log of hazard ratio. There are also many methods to estimating the coefficient of random variables. Partial likelihood approach is proposed by Cox (1972) to estimate β without considering $h_0(t)$. Let $(X_i\ Z_i\ \delta_i)$ represent the set of an individual i where: X_i is a possibly censored failure time random variable, Z_i is a set of covariates,

δ_i is a censored indicator and r_i denotes the risk set at time i, then:

If the person is censored at X_i,

$$\mathcal{L}_i(\beta) = S_i(X_i),$$

if the person is dead at X_i,

$$\mathcal{L}_i(\beta) = S_i(X_i)h_i(X_i).$$

The total likelihood can be expressed as:

$$\begin{aligned}\mathcal{L}(\beta) &= \prod_{i=1}^{n} h_i(X_i)^{\delta_i} S_i(X_i) \\ &= \prod_{i=1}^{n} \left[\frac{h_i(X_i)}{\sum_{j\in r_i} h_j(X_i)}\right]^{\delta_i} \left[\sum_{j\in r_i} h_j(X_i)\right]^{\delta_i} S_i(X_i).\end{aligned}$$

If we only focus on the fist term, it is:

$$\prod_{i=1}^{n} \left[\frac{h_0(X_i)exp(\beta Z_i)}{\sum_{j\in r_i} h_0(X_i)exp(\beta Z_i)}\right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{exp(\beta Z_i)}{\sum_{j\in r_i} exp(\beta Z_i)}\right]^{\delta_i},$$

it can ignore the effect of the underlying hazard function $h_0(t)$.

3. Baseline

3.1 Parametric

Many popular distribution can be applied in the survival baseline of the latency part, such as gamma, weibull and lognormal. The following is some examples:

1. **Gamma Distribution**

   This step models the hazard function. Assume that the baseline survival function S(t) follows gamma distribution:

$$S_0(t) = 1 - F_0(t) = 1 - \frac{\Gamma_{\frac{t}{\beta}}(\alpha)}{\Gamma(\alpha)}, \quad \beta > 0, \alpha > 0,$$

where Γ is defined as:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx,$$

and the incomplete gamma function $\Gamma_{\frac{t}{\beta}}(\alpha)$ has the formula:

$$\Gamma_{\frac{t}{\beta}}(\alpha) = \int_0^{\frac{t}{\beta}} x^{\alpha-1} e^{-x} dx.$$

The derivative of $S_0(t)$ is:

$$f_0(t) = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{\Gamma(\alpha)\beta^\alpha},$$

so according to the definition of hazard function, we can get:

$$h_0(t) = \frac{f_0(t)}{S_0(t)} = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{[\Gamma(\alpha) - \Gamma_{\frac{t}{\beta}}(\alpha)]\beta^\alpha}.$$

2. **Weibull Distribution**

   If the survival function follows weibull distribution, then the baseline can be described as:

   The survival function is :

$$S_0(t) = 1 - F_0(t) = e^{-(\frac{x}{\lambda})^k}, \quad k > 0, \lambda > 0,$$

   where x is a random variable, λ is scale parameter, k is the shape parameter. Weibull distribution has relation ship with many distribution, for example if k=1, it is an exponential distribution. The derivative function is :

$$f_0(t) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(\frac{x}{\lambda})^k},$$

   and the hazard function is:

$$h_0(t) = \frac{f_0(t)}{S_0(t)} = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}.$$

3. **Lognormal Distribution**

   If the logarithm of a random variable is normally distributed, then it is called lognormal distribution. In the short run, it is very close to a normal distribution, but in the long run, the lognormal distribution is going to be a little bit more upward. If the random variable follows lognormal distribution, then the survival function is :

$$f_0(t) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\dfrac{(lnx-\mu)^2}{2\sigma^2}},$$

which is the expression of $lnx \sim (\mu, \delta^2)$,

the derivative function is :

$$S_0(t) = 1 - F_0(t) = 1 - \Phi(\frac{lnx-\mu}{\sigma}),$$

then the hazard function is:

$$h_0(t) = \frac{f_0(t)}{S_0(t)} = \frac{e^{-\dfrac{(lnx-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}(1-\Phi(\frac{lnx-\mu}{\sigma}))}.$$

3.2 Nonparametric

1. **Nelson-Aalen Estimator**

The Nelson–Aalen estimator is used to estimate the cumulative hazard rate function H(t) and the details can be found in Nelson (1969, 1972) and Aalen (1978). The assumption is that the observed data of sample survival times is right censoring and the censorings are independent of time t without any additional information, then the formula can be written as:

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j},$$

where d_j denotes the number of death at t_j, r_j denotes the risk set such as alive and uncensored individuals before time t.

2. **Kaplan-Meier Estimator**

The Kaplan–Meier estimator acquainted by Kaplan and Meier (1958) is a non-parametric statistic used to estimate the survival function from censored data, especially right-censoring. Assume that the survival times of some observed individuals are subject to right-censoring which means their survival times are longer than the experimental period or they withdraw from the research, then the function can be written as:

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{r_j}\right),$$

where d_j denotes the number of death at t_j, r_j denotes the risk set such as alive and uncensored individuals.

4. EM algorithm method

The EM algorithm method is used to estimate the parameters in the latency part and it first begins with the E-step. In some studies, a lot of censored data will be obtained and they cannot be considered exactly whether they will experience the time of interest or not. So in this step, we will get the conditional expectation of the complete log-likelihood with the given data, which is the of expectation of y_i. Let Y be an indicator, if an individual will be uncured(experience the event of interest), Y = 1, otherwise Y=0 and y_i presents the indicator of the ith individual.

An assumption of the estimation is that the censoring is independent and noninformative. Given i=1,$\cdots$,n, let set $\boldsymbol{O_1} = (\boldsymbol{z_i}, \boldsymbol{x_i}, t_i, \delta_i)$ denote the observed data for the ith individual, $\boldsymbol{z_i}$ and $\boldsymbol{x_i}$ respectively refer to the covariates in the incidence and latency parts which is mentioned above, t_i denotes observed survival time and δ_i is a binary indicator of censoring that $\delta_i = 0$ for the censored time and $\delta_i = 1$ for the uncensored time.

Let set $\boldsymbol{O_2 = (b, \beta, S_0(t))}$ denote the unknown parameters, where **b** is the coefficient of z, $\boldsymbol{\beta}$ is the parameter of x in the exponential function of the latency part, $\boldsymbol{S_0(t)}$ is the original survival function without transform.

The complete likelihood function can be expressed as:

$$\prod_{i=1}^{n}[1-\boldsymbol{\pi(z_i)}]^{1-y_i}\boldsymbol{\pi(z_i)}^{y_i}h(t_i|Y=1,\boldsymbol{x_i})^{\delta_i y_i}\mathbf{S}(t_i|Y=1,\boldsymbol{x_i})^{y_i}. \tag{3}$$

From this function, we can see that if y_i=1, the individual is uncured and its function is $\boldsymbol{\pi(z_i)}h(t_i|\boldsymbol{x_i})^{\delta_i}\mathbf{S}(t_i|\boldsymbol{x_i})$, if the individual is cured (y_i=0), its function is $1-\boldsymbol{\pi(z_i)}$.

As $\mathbf{b}$ and $\boldsymbol{\beta}$ refer respectively to the parameters of the incidence and latency parts, the loglikelihood function expression can also be written as:

$$\boldsymbol{l(b,\beta;O_2,y)=l_1(b;O_2,y)+l_2(\beta;O_2,y)},$$

$\boldsymbol{l_1}$ is the log function of $\prod_{i=1}^{n}[1-\boldsymbol{\pi(z_i)}]^{1-y_i}\boldsymbol{\pi(z_i)}^{y_i}$, where

$$\boldsymbol{l_1(b;O_2,y)}=\sum_{i=1}^{n}y_i\log[\boldsymbol{\pi(z_i)}]+(1-y_i)\log[1-\boldsymbol{\pi(z_i)}], \tag{4}$$

$\boldsymbol{l_2}$ is the log function of $\prod_{i=1}^{n}h(t_i|Y=1,\boldsymbol{x_i})^{\delta_i y_i}\mathbf{S}(t_i|Y=1,\boldsymbol{x_i})^{y_i}$, where

$$\boldsymbol{l_2(\beta;O_2,y)}=\sum_{i=1}^{n}\delta_i y_i\log[h(t_i|Y=1,\boldsymbol{x_i})]+y_i\log[\mathbf{S}(t_i|Y=1,\boldsymbol{x_i})]. \tag{5}$$

Given the observed data set $\boldsymbol{O_1}$ and the estimator set $\boldsymbol{O_2^{(k)}}=(\boldsymbol{b^{(k)}},\beta^{(k)},S_0(t)^{(k)})$, the conditional expectation of y_i can be calculated as:

$$E(y_i|\boldsymbol{O_1},\boldsymbol{O_2}^{(k)})=\delta_i+(1-\delta_i)\frac{\boldsymbol{\pi(z_i)}\mathbf{S}(t_i|Y=1,\boldsymbol{x_i})}{1-\boldsymbol{\pi(z_i)}+\boldsymbol{\pi(z_i)}\mathbf{S}(t_i|Y=1,\boldsymbol{x_i})}.$$

It is easy to understand, if the data is uncensored, then the paitent is uncuted that $E(y_i)=1$ (which $\delta_i=1$); if the data is censored, then the paitent has a probability of uncuted and the expectation is the rest part. Thus, let $w_i^{(k)}$ denote $E(y_i|\boldsymbol{O_1},\boldsymbol{O_2}^{(k)})$, the expections of (5) and (6) can be written as:

$$\boldsymbol{E(l_1(b;O_2,y))}=\sum_{i=1}^{n}w_i^{(k)}\log[\boldsymbol{\pi(z_i)}]+(1-w_i^{(k)})\log[1-\boldsymbol{\pi(z_i)}], \tag{6}$$

$$\boldsymbol{E(l_2(\beta; O_2, y))} = \sum_{i=1}^{n} \delta_i w_i^{(k)} \log[h(t_i|Y=1, \boldsymbol{x_i})] + w_i^{(k)} \log[\mathbf{S}(t_i|Y=1, \boldsymbol{x_i})], \tag{7}$$

where $\delta_i w_i^{(k)} = \delta_i$.

The M-step is the second part of the EM algorithm after E-step, which aims to estimate the value of unknown parameters based on the expectation value of y_i obtained at step E. In this step, we are try to maximize (7) and (8) in the previous steps, and we need to demonstrate the approach separately by PHMC model and AFTMC model because the assumption of latency part in equation (8) is different.

PHMC Model

According to Peng *et al.* (1998), they have used a type of partial likelihood method to explain a very exhaustive process about how to estimate β without a parametric baseline. If we also employ the PH assumption, i.e.the hazard function is $h_0(t_i|x_i) = h_0(t_i)\exp(\beta x_i)$, and the survival function is $S_0(t_i|x_i) = S_0(t_i)^{\exp(\beta x_i)}$. Then the estimating equation (8) based on the standard PH model with an additional offset variable $log(w_i^{(m)})$can be rewritten as:

$$log \prod_{i=1}^{n} [h_0(t_i)\exp(\beta x_i) + log(w_i^{(m)})]^{\delta_i} S_0(t_i)^{\exp(\beta x_i +) log(w_i^{(m)})}, \tag{8}$$

In order to continue the E-step, we need to update the estimated survival function by using the Nelson Aalon approach. Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ be the distinct uncensored failure times, d_j denotes the number of uncensored failures and R_j denotes the risk set at time $t_{(j)}$, the Nelson-Aalon-type (also called Breslow-type) estimator for the original survival function with the condition that the individual experiences the event can be expressed as:

$$\widehat{S_0}(t|Y=1) = \left(- \sum_{j:t_{(j)} \leq t}^{n} \frac{d_j}{\sum_{i \in R_j} w_i^{(m)} e^{\hat{\beta} x_i}} \right). \tag{9}$$

Since the estimator $\hat{S_0}(t|Y=1)$ may not close to zero when the time t goes to infinite, we denote $\hat{S_0}(t|Y=1) = 0$ if $t \geq t_{(k)}$, and also the latency part of semiparametric PH model can be expresses as: $\hat{S}(t|Y=1) = \hat{S_0}(t|Y=1)^{\exp(\hat{\beta}x)}$

AFTMC Model

In the paper of Zhang and Peng (2007), they presented an estimation method based on the rank estimator for semiparametric ATFMC model. Similar to the PHMC model, the equation (8) can be rewritten as the following form by appling ATF baseline:

$$log\prod_{i=1}^{n}[w_i^{(m)}h(log((t_i)-\beta x_i)^{\delta_i}[S(log((t_i)-\beta x_i^{w_i^{(m)}})]).$$

And Zhang and Peng (2007) also proposed the gradient of a convex function to maximizing the estimator, which can be written as:

$$L_G(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i w_i^{(m)}|\varepsilon_i - \varepsilon_j|I(\varepsilon_i - \varepsilon_j), \tag{10}$$

where $\varepsilon_i = log((t_i) - \beta x_i$. Therefore, we can using the linear programming method in R to maximized the equation (11). Same to the assumption in the PHMC model, let $\tau_1 < \tau_2 < \cdots < \tau_k$ be the distinct uncensored failure times, d_j denotes the number of uncensored failures and R_j denotes the risk set at time $\tau_{(j)}$, the estimator for the original survival function $S_0(\varepsilon|Y=1)$ can be expressed as:

$$\widehat{S_0}(\varepsilon|Y=1) = \exp\left(-\sum_{j:\tau_{(j)}<\varepsilon}\frac{d_j}{\sum_{i\in R_j} w_i^{(m)}}\right). \tag{11}$$

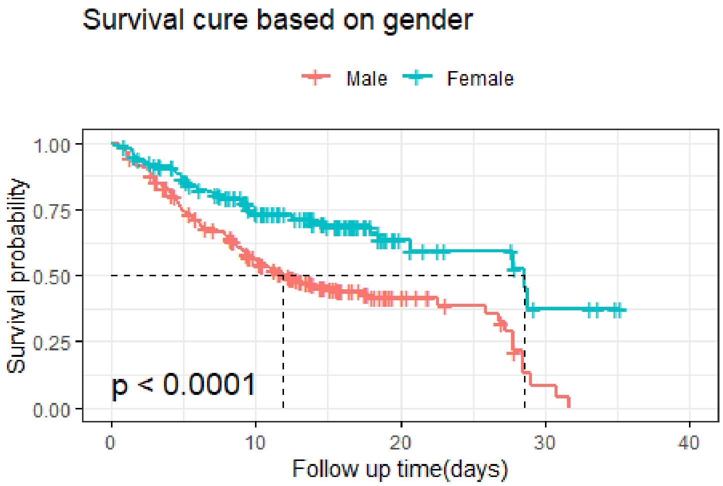
And also we set $S_0(\varepsilon|Y=1) = 0$ for $\varepsilon > \tau_k$, then $S(t|Y=1) \hat{=} S(\varepsilon|\hat{Y}=1)$

IV. Analysis

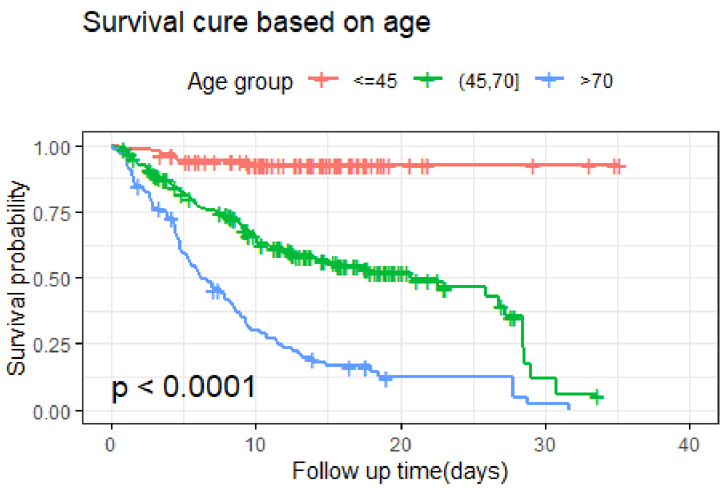
1. COVID-19

The data is collected all patients in the area of Wuhan older than 18 years old, excluding the patients who's recorded data less than 80% or pregnant women and breastfeeding women between 10 January and 18 February 2020. A total of 361 patients were selected in this analysis and the data collected include admission time, discharge time, the status of patients, their blood samples and laboratory results, including electrolytes, inflammatory factors, liver function, kidney function, coagulation function and so on. We mainly focus on the survival time, outcome, age and gender. For those cases, 195 patients were dead during the observation period, 166 recovered and were treated from the hospital. The age of those people is from 18 years old to 95 years old with the average age of 58.9 years old; the follow up time is from 0.085 days to 35.17 days with the mean of 11.22 days. There are 149 female and 212 male.

First, we used Kaplan Meier survival curve to draw two visualizations based on age and gender respectively. The following graph shows information about survival curve of the given data based on gender, where y-axis denotes survival rate, x-axis denotes follow up time. From the graph, we can see that males mortality rate is generally higher than that females. The median time of female is closed to 30 days, however for male, it is a little longer than 10 days.



The following graph shows information about survival curve of the given data based on age, we divide the patients into 3 groups according to their age (i.e. age≤45, 45<age≤70, 70<age). From the graph, we can see that the death rate of the age less than 45 years old is very low, that means those people have a great probability prevented and cured from COVID-19; for those who are older than 45 years old, they are very likely to be died eventually. There is also a great different between the two group (45<age<=70, 70<age), for 45<age<=70, the survival probability is decreased slidely, at about 28 days it declines rapidly and closed to zero but still great than zero which means that there might be some patients survived from COVID-19; for 70<age, people will die quickly before 20 days, after that it may be stable, but at last almost all the patients will die.



Next, we used R package smcure to fit the data by semiparametric accelerated failure time mixture model. Due to the restriction of the package, we first deal with the data. The gender is male or female and the age we divided into three group, so we combine them and reclassify into six groups. For the male, who is younger than 45 years old is classified as group one, who's age between 45 and 70 is in group two, who is older than 70 years old is in group three; for female, who is younger than 45 years old is in group four, who's age between 45 and 70 is in group five, who is older than 70 years old is in group six. Then model can be fitted as following:

```
    Cure probability model:

            Estimate

(Intercept) -1.788636

group        3.638773

Failure time distribution model:

            Estimate

(Intercept)  2.4381655
```

```
group2      -1.5402071

group3      -2.3204165

group4       0.7227988

group5      -0.8930412

group6      -2.0421930
```

Based on the results from Cure probability model part, we can get the cure rate. Because there is only one variable, so the cure rate for the group is 13.6 percent, which is calculated by $1-\pi(z) = 1- \frac{e^{(} - 1.788636 + 3.638773)}{1 + e^{(} - 1.788636 + 3.638773)}$. When we focus on failure time distribution, the survival probability can be obtained. For example, considered group1 and group2, the failure time of a male who is in group two is 1.5402071 times less than that in group one. Also it shows that as the age going up, the coefficient declines which means older people will die quickly than younger one and compared with gender, the value of group4, group5, group6 is more than group2 and group3, which means that a male dies quicker than female.

2. Heart Failure

The data is admitted to Faisalabad Cardiology Institute or Faisalabad United Hospital from April to December 2015, includeing 299 patients with NYHA III and IV left ventricular systolic dysfunction confirmed by heart ultrasound report or physician's record. In this collected data, there are 194 males and 105 females with the age from 40 years old to 95 years old and the average follow up time is 130 days with the minimum time of 4 days and maximum time of 285 days. The risk factors considered to explain the mortality caused by coronary heart disease are age, gender, smoking, diabetes, Blood Pressure (BP), Ejection Fraction (EF),serum sodium, serum creatinine, anemia, platelets and Creatinine Phosphokinase (CPK).

The time is consulting in days which an individual live or be recorded, the event '0' repersents censored data and '1' repersents death. Gender is marked as 1 if the patient is a male, otherwise is a female. According to Ahmad *et al.* (2017), anemia is the evaluation of patients' hematocrit level. If their hematocrit is less than 36, they would be considered as anemic. The informaion of those related risk factors is taken from their blood report and the situation of smooking and blood pressure is obtained from the doctor's records. There are six continuous variables which are age, Ejection.Fraction, serum sodium, serum creatinine, platelets and CPK and five categorical variables which are gender, smoking, diabetes, Blood Pressure and anaemia. For the male, the survival time is from 4 days to 285 days with the average time of 129.3711 days, there are 62 people dead and 132 censored; for the female, the survival time is from 8 days to 278 days with the average time of 131.9048 days, there are 34 people dead and 70 censored.

Following is a table summarizes the information of risk factors:

Table I: Descriptive for the risk factors

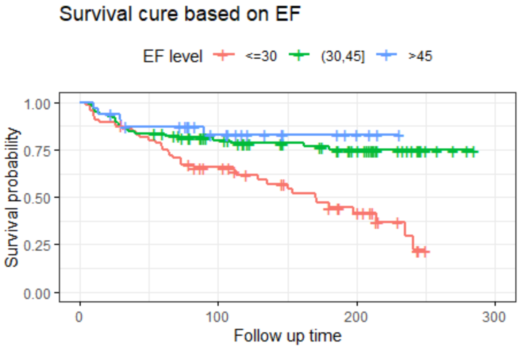
Variable	Describation	male	female	Total
Age	mean	61.4055	59.77778	60.83389
Gender		194	104	298
Smoking	Yes	92(47.42%)	4(3.85%)	96(49.48%)
	No	102(52.58%)	100(96.15%)	202(104.12%)
Diabetes	Yes	70(36.08%)	55(52.88%)	125(64.43%)
	No	124(63.92%)	49(47.12%)	173(89.18%)
BP	Yes	61(31.44%)	44(42.31%)	105(54.12%)
	No	13(6.7%)	60(57.69%)	193(99.48%)
Anaemia	Yes	77(39.69)	52(50)	129(66.49%)
	No	117(60.31)	52(50)	169(87.11%)
Ejection.Fraction	mean	36.79381	40.46667	38.08361
Serum Sodium	mean	136.5361	136.7905	136.6254
Serum Creatinine	mean	1.399175	1.384095	1.39388
Pletelets	mean	254370.2	279964	263358
CPK	mean	638.701	476.781	581.8395

The given risk factors are a little more if they are directly used the semi-parametric mixed cure model because it may cause the iterative process in the EM algorithm approach fail to converge, resulting in a phenomenon that the parameters can be defined. Therefore, we first use the cox model to filter the non-significant factors. According to Zahid *et al.* (2019), they mentioned that the heart failure case we studied is actually based on sex, and the risk factors for different gender are also different. Therefore, in this study, we continue to divide the data into men and women for further research. Then the regression result of men shows as follow:

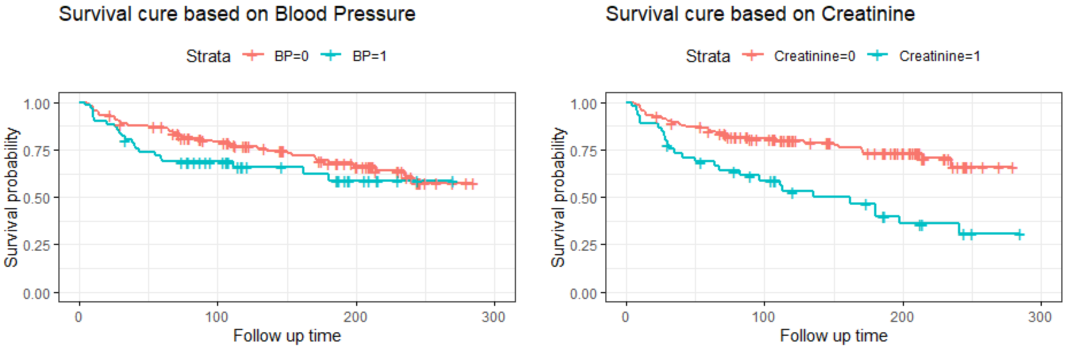
Significance of variables under Cox regression for male:

```
                    coef       exp(coef)   se(coef)  z      Pr(>|z|)

Smoking1          -8.336e-02  9.200e-01  2.641e-01 -0.316   0.7523

Diabetes1         -1.490e-01  8.616e-01  2.861e-01 -0.521   0.6025

BP1                4.213e-01  1.524e+00  2.775e-01  1.518   0.1290

Anaemia1           1.676e-01  1.182e+00  2.755e-01  0.608   0.5430

Age                5.088e-02  1.052e+00  1.148e-02  4.431 9.37e-06 ***

Ejection.Fraction -6.200e-02  9.399e-01  1.440e-02 -4.304 1.67e-05 ***

Sodium            -3.059e-02  9.699e-01  3.272e-02 -0.935   0.3498

Creatinine         2.668e-01  1.306e+00  1.072e-01  2.489   0.0128 *

Pletelets          9.975e-07  1.000e+00  1.402e-06  0.711   0.4769

CPK                1.801e-04  1.000e+00  1.081e-04  1.665   0.0958 .

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the output, age is the most significant variable with the p-value of 9.37×10^{-6}. The age-related coefficient indicates that the chance of death from coronary heart disease increases with age, and if the age increases one year, the risk of death is increased by 5.2 percent. Ejection.Fraction is also a significant variable with the p-value of 1.67×10^{-5}. The patient with a unit increased in EF has 93.99% more chances of death. Another significant factor is Serum creatinine with the p-value of 0.0128, the hazard of death increases by 30.6% for every additional units of Serum creatinine. The p-value of CPK is 0.0958, but in this case the level of 5% is considered, so it can not be included in the following test. According to results, smoking, diabetes, Serum Sodium, anaemial and platelets are found to be non-significant. Additionally, Kaplan Meier survival curve is supported to directly see the difference between different levels. The curve is provided for Blood Pressure (BP), Ejection.Fraction(EF) and Serum creatinine. EF is divided into three levels (i.e. $EF \leq 30$, $30 < EF \leq 45$ and $45 < EF$) and serum creatinine is considered to be divided into two levels (more than 1.5 denote as 1, others denote as 0) due to the effective on mortality.



(a) EF



(b) BP

(c) Serum creatinine

Figure 1:

The EF curve shows that there is slight difference of survival probability between the patients with EF>30, and there is an obvious decline for the patients with EF<30. In the BP curve, it is different for the survival patients before 100 days, however, the mortality seems equal for whether the individuals have high blood pressure eventually. So we cancel it from the model. In Serum creatinine curve, the patients with higher serum creatinine are easier to be dead due to CHD. In the end, the factors selected are age, EF and Serum creatinine for males.

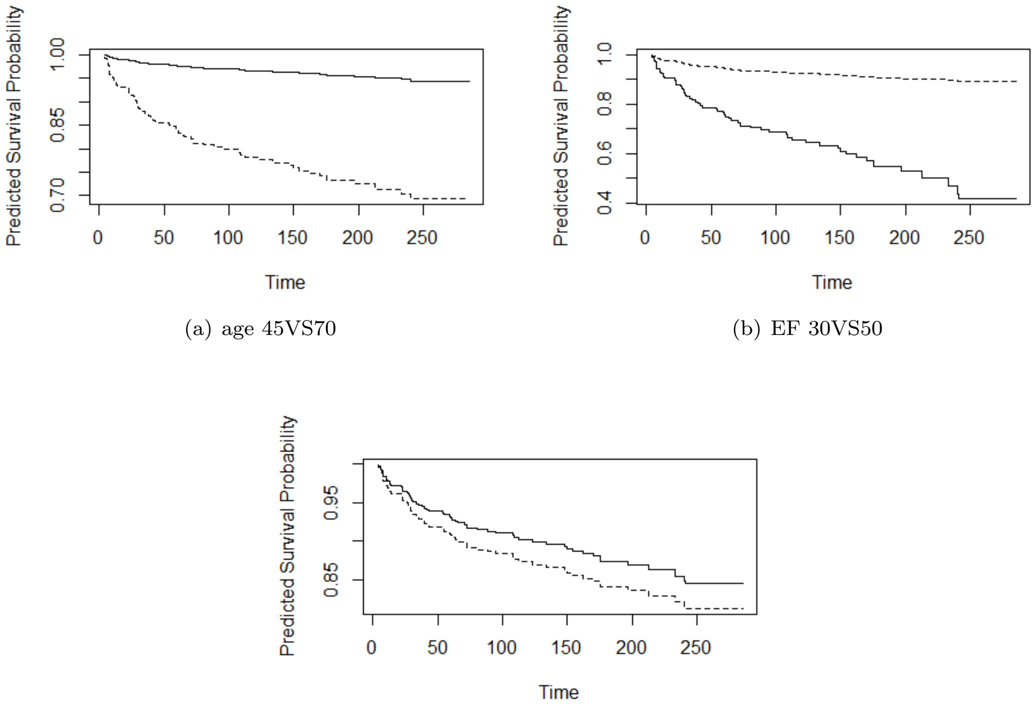
Next we add cured part to the cox model, then it changes to a mixcure model. An R package smcure is still be presented here and it is to estimate the semiparametric PH mixcure model. The output is:

```
    Cure probability model:

                     Estimate
(Intercept)        -1.4752961
Age                 0.0794932
Ejection.Fraction -0.1218163
Creatinine          0.3805140
```

```
Failure time distribution model:

                    Estimate

Age               0.01377726

Ejection.Fraction 0.01422122

Creatinine        0.21033348
```

The failure time distribution model estimator shows that the log hazard ratio increases 1.38% if the patient's age increase one year; it increases 1.42% if the patient's EF increase a unit and it increases about 21% if the patient's serum creatinine increase a unit.

The fitted survival curves are also shown following. The are there prediction picture, the standard we set is age of 61.4 (the median centered age of male), EF of 45 and serum creatinine of 1.5 according the above level. The solid line represents lower level and the dotted line denots higher level, then we can see that the male with lower EF, higher age and serum creatinine has a higher mortality.



(a) age 45VS70

(b) EF 30VS50

(c) Serum creatinine 1vs1.6

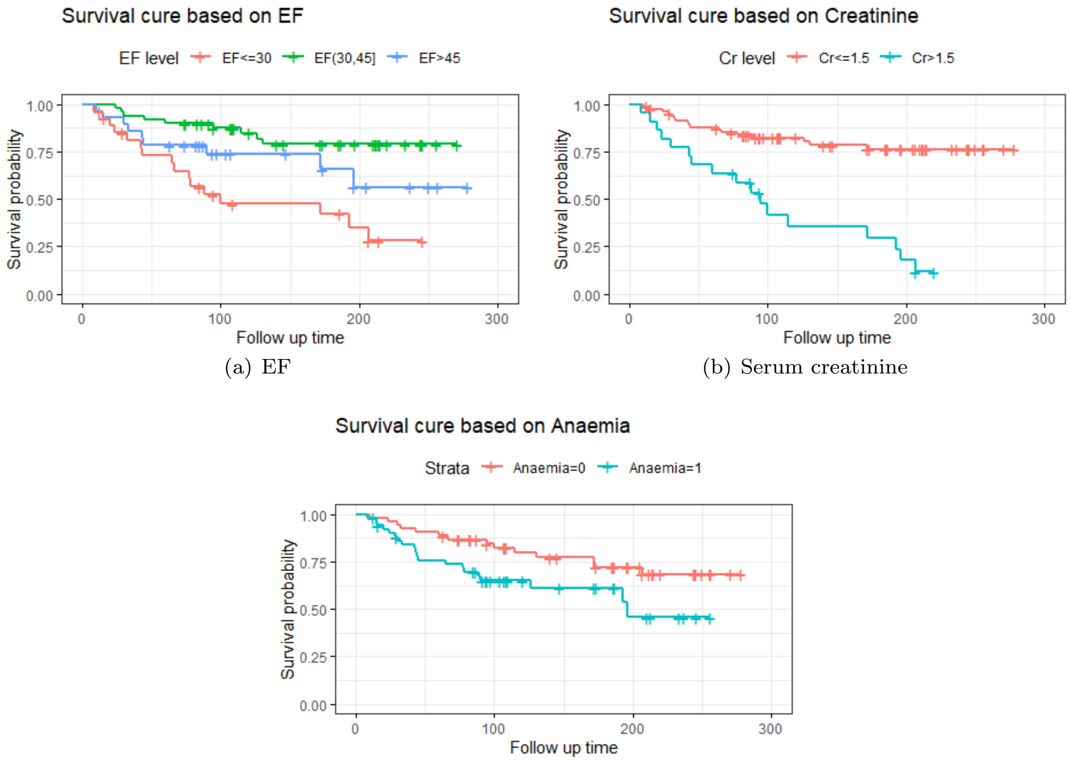
Figure 2:

When we focus on the female, the analysis process is the same as the male, in order to simplify the semi-parametric mixture cure model, first step is to get rid of irrelevant variables by the cox regression. The result is presented following: **Significance of variables under Cox regression for the female:**

```
                    coef   exp(coef)   se(coef) z      Pr(>|z|)

Smoking           1.369e+00  3.930e+00  6.910e-01  1.981  0.04764 *

Diabetes          5.227e-01  1.687e+00  4.124e-01  1.268  0.20493

BP                4.044e-01  1.498e+00  3.561e-01  1.136  0.25607
```

```
Anaemia             8.602e-01  2.364e+00  3.925e-01  2.191  0.02842 *
Age                 4.553e-02  1.047e+00  1.861e-02  2.447  0.01441 *
Ejection.Fraction -3.780e-02  9.629e-01  1.680e-02 -2.250  0.02447 *
Sodium             -6.860e-02  9.337e-01  4.653e-02 -1.475  0.14034
Creatinine          3.166e-01  1.372e+00  1.114e-01  2.842  0.00448 **
Pletelets          -2.289e-06  1.000e+00  1.967e-06 -1.164  0.24457
CPK                 3.565e-04  1.000e+00  3.320e-04  1.074  0.28298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the output, for females, the most significant variable is serum creatinine with the p-value of 0.00448 and the estimator coefficient indicates that the the chance of death is 37.2% for every additional unit. The second one is age, which p-value is 0.01441 and the risk will increase 12% for the female patient growing one year old. Ejection.Fraction is also a significant variable, however, compared with male, anaemia and smoking is significant variables in this test. The p-value of Anaemia is 0.02842 and an anemic woman has more than double chances of death as compared to non-anaemia woman. Also the hazard of an smoking women is 3.93 times more than non-smoking women. Diabetes, BP, serum sodium, pletelets and CPK are found to be non-significant variables. With the Kaplan Meier survival plots, the difference can be shown directly. Different from the plots shown by male, figure (g) shows that the survival probability is not closed to each other between the group of EF≤30 and 30<EF≤45. Figure (h) shows that the effect of Serum creatinine on the mortality of the female heart failure is significantly greater than that of male, the death rate of women with creatinine is higher than that of a men.



(a) EF

(b) Serum creatinine

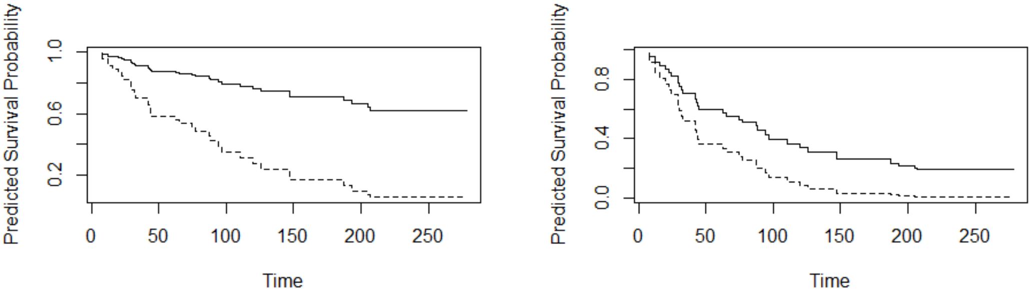
(c) Anaemia

Figure 3:

Next, we focus on the semi-parametric model. The model is included smoking, anaemia, age, EF and creatinine. If considering the female with the median centered age of 59.8, EF equals to 45, creatinine equals to 1.5, we can draw the fitted survival curves controlled by smoking and anaemia respectively.

```
Cure probability model:
                           Estimate
(Intercept)             -7.65144217
Smoking                  3.25212842
```

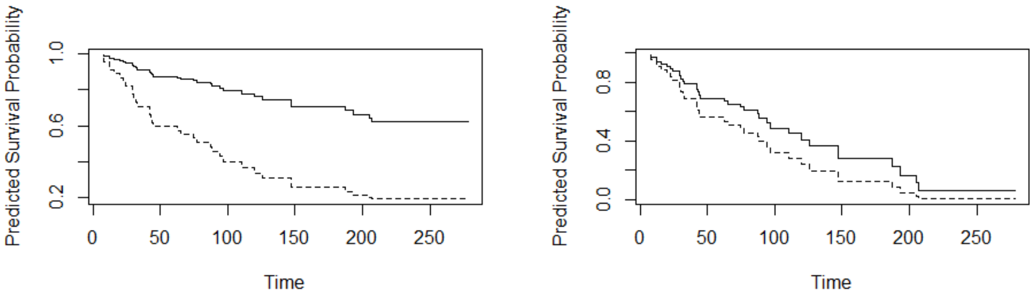
```
Anaemia            1.90277187

Age                0.03467673

Ejection.Fraction -0.03932558

Creatinine         4.56672502

Failure time distribution model:

                       Estimate

Smoking            0.3938178798

Anaemia            0.5628239362

Age                0.0138061690

Ejection.Fraction  0.0004350185

Creatinine        -0.1713402627
```



(a) non-anaemia

(b) anaemia

Figure 4: Fitted survival curve for smoking between non-anaemia and anaemia



(a) non-smoking

(b) smoking

Figure 5: Fitted survival curve for anaemia between non-smoking and smoking

From the failure time distribution, we can see that the coefficient of EF is very small. Due to the limits of the package and data, it is fail to obtain the standard errors of estimated parameters so that it cannot be decided whether EF is a significant variable in the semi-parametric model. However, the value is 0.000435 which is closed to zero, so it might be an irrelevant variable. Besides, the following four figures show that smoking can influence the survival probability obviously on non-anaemia patients of female and anaemia can have an obviously impact on non-smoking patients of female.

V. Conclusion

This article mainly studies two cases, COVID-19 and heart disease based on the semi-parametric mixture cure model. For the COVID-19 case, we use the AFT model to study the impact of age and gender on mortality rate of the patients in the region of Wuhan. The analysis shows that a male dies quickly than a female and the growing age accelerates the process of death, so that means the man with advanced age will die in a short time due to COVID-19. For the heart failure case, we study the influence of different factors on the mortality of male and female patients in Pakistan respectively. The analysis shows that age, Ejection.Fraction and Serum creatinine are significance for both of man and women, but smoking and anaemia are only significance for women. It can be concluded that growing age, lower values of Ejection Fraction, higher serum creatinine which can cause renal dysfunction are the main indicators contributing to increasing the hazard ratio of death among heart failure patients, smoking and higher level of anaemia are contributing to increasing the risk of death among female patients. In addition, during the study, there are also problems need to be considered. For example, the selection of sample data is not large and the coverage is small, this might caused the non-significance of some indicators. And the collected data cause the failure of estimating the parameters in the semi-parametric mixture cure model and also the design of the model need to be improved for further study.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

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), e0181001.

Amico, M., Van Keilegom, I., and Legrand, C. (2019). The single-index/cox mixture cure model. *Biometrics*, **75**(2), 452–462.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53.

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

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, **65**(3), 762–771.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**(446), 496–509.

Jiang, W., Sun, H., and Peng, Y. (2017). Prediction accuracy for the cure probabilities in mixture cure models. *Statistical Methods in Medical Research*, **26**(5), 2029–2041.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457–481.

Kumar, V., Aijaz, S., Sattar, S., and Pathan, A. (2020). Frequency, predictors and prognosis of worsening renal function in patients admitted with acute heart failure. *JPMA*, **2020**.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **34**(3), 201–211.

Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, pages 1223–1234.

Li, C.-S. and Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in medicine*, **21**(21), 3235–3247.

Li, P., Peng, Y., Jiang, P., and Dong, Q. (2019). A support vector machine based semiparametric mixture cure model. *Computational Statistics*, pages 1–15.

Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. John Wiley & Sons.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, **1**(1), 27–52.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**(4), 945–966.

Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, **41**(3-4), 481–490.

Peng, Y., Dear, K. B., and Denham, J. (1998). A generalized f mixture model for cure rate estimation. *Statistics in Medicine*, **17**(8), 813–830.

Pieruschka, E. (1961). Relation between lifetime distribution and the stress level causing the failures. Technical report, LOCKHEED MISSILES AND SPACE CO SUNNYVALE CALIF.

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F., and Garduño-Espinosa, J. (2020). A survival analysis of covid-19 in the mexican population. *BMC public health*, **20**(1), 1–8.

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, pages 899–907.

Wang, X. and Zhou, X.-H. (2018). Semiparametric maximum likelihood estimation for the cox model with length-biased survival data. *Journal of Statistical Planning and Inference*, **196**, 163–173.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in japan. *Journal of the American Statistical Association*, **87**(418), 284–292.

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., *et al.* (2020). An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, pages 1–6.

Zahid, F. M., Ramzan, S., Faisal, S., and Hussain, I. (2019). Gender based survival prediction models for heart failure patients: a case study in pakistan. *PloS one*, **14**(2), e0210602.

Zeng, D. and Lin, D. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, **102**(480), 1387–1396.

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in medicine*, **26**(16), 3157–3171.