Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 265–277

Modeling and Analysis of Competing Risks Cure Rate Regression Model with Weibull Distribution

P. G. Sankaran[1] **and P. P. Rejani**[1,2]
[1]*Department of Statistics, Cochin University of Science and Technology, Kerala, India*
[2]*Department of Community Medicine, Govt.Medical College, Kerala, India*

Received: 04 September 2021; Revised: 08 December 2021; Accepted: 30 December 2021

Abstract

Cure rate models have been widely applied in the analysis of lifetime data in the presence of cured fractions. Regression models need more attention when investigators are interested to study the effects of given treatments. The presence of competing risks is an additional challenge for researchers to analyze lifetime data with cured proportion. In this paper, we propose a parametric cure rate regression model incorporating competing risks for the analysis of survival data. The parameters of the model are estimated by the maximum likelihood estimation procedure via EM algorithm. A simulation study is carried out to evaluate the performance of the proposed model. The practical relevance of the model is illustrated by applying the model to a dataset on heart transplantation.

Key words: Cure rate model; Competing risks; Maximum likelihood; EM algorithm; Weibull distribution

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The recent advancements in diagnostic and other drug design experiments resulted increased rate of favorable response of patients to their received treatments and a good proportion of patients have become free from diseases. These disease-free individuals in a set of survival data are said to be immunes and the proportion of immunes that exists in the data is called cured proportion. The presence of immunes in survival data influences the outcome measures in survival studies. While analysing such data, it can be seen that the survival curve does not taper off to zero at the end of the study period. Hence ordinary survival analysis techniques are not suitable to analyse such data and new models have been developed incorporating cured proportions. Such models are said to be cure rate models in survival analysis. Boag (1949) first proposed cure rate model to estimate the cured proportion of breast cancer patients. Cure rate models have been extended its applicability in several areas like financial, criminology, demography, and industrial reliability. Nelson (1982) explained the life expectancy of electric motors with cure rate model. Yamaguchi (1992) applied the cure rate model to describe inter-firm job mobility in Japan. For further reading one can refer to Maller and Zhou (1996), Sy and Tailor (2000), Ortega *et al.* (2014), Shen *et al.* (2019), and Sreedevi and Sankaran (2021).

Corresponding Author: P.P. Rejani
Email: rejstat@gmail.com

Competing risks occur when the study subjects experience more than one events that compete the event of interest. For example, when a researcher observing peritoneal dialysis patients until they develop peritonitis, kidney transplantation can be regarded as a competing cause because the chance of occurrence of peritonitis is very less among patients who have undergone kidney transplantation. The competing risks aspect seeks more attention in the analysis and interpretation of survival data. It is known that age-related mortality is high among older people than others. Also, the probability of death due to the disease is found to be low in clinical trials with the desired effect. In both cases, deaths occur due to other competing causes rather than the event of interest. Hence failure to consider competing risks in the analysis of such data yields reporting of inaccurate and misleading results. The competing risks models are discussed by many authors. Crowder (2001) and Kalbfleisch and Prentice (2011) are prime among them. Wright *et al.* (2020) and Papastefanou *et al.* (2021) are two recent works that draw out the significance of competing risks in the medical field.

In survival studies carry out in the field of medicine and epidemiology, the investigators focus on determining the effect of factors associated with the time to occurrence of the event such as death or disease recurrence. Regression models such as Cox proportional hazards models or parametric models are usually used to study the effect of covariates present in the data. The presence of competing risks, immune proportions and covariates altogether enhance the complexity of data and burden of analysis. All of these prominent scenarios are encountered by formulating competing risks cure rate regression models. Development of such models needs special attention and less available in literature.

In cure rate models, parametric or semiparametric proportional hazards assumptions can be made for lifetime distribution in latency. In recent times, some semiparametric models are proposed for the analysis of competing risks data in the presence of cured proportions. The interested readers can refer to Choi *et al.* (2018) and Rejani and Sankaran (2020). If a particular probability distribution of survival data can be identified and validated, statistical inference based on a parametric regression perceptive will be considered as more efficient and precise than those derived from survival models in the absence of an explicit distributional function (Collett, 2015). Yusuf *et al.* (2016) discussed Weibull distribution as a suitable distribution for the analysis of data in the presence of cured proportion.

In this paper, we introduce a parametric cure rate regression model based on Weibull distribution for the analysis of survival data in competing risks setting. The model and methods focus on the estimation of regression parameters and the probability of cure in the presence of competing risks. The innovative feature of the proposed model is the proficiency to explain the impact of covariates on the survival time of a group of subjects in the presence of immunes and at the same time, the influence of competing causes is also taken into account.

Heart transplantation is the gold standard for the treatment of end-stage heart failure. Rejection and infection are the two major causes of mortality among patients undergoing heart transplantation. Larson and Dince (1985) considered 65 transplant recipient data, there were 29 (45%) rejection deaths, 12 (18%) deaths from other causes, and 24 (37%) censored observations. They analyzed data by mixture model approach without considering the chance of occurrence of cured proportion. A cure rate regression model separates short and long-term survival of patients. It is useful to determine the proportion of cured patients and to identify the associated factors on survival of patients under study. It helps the public

health professionals in decision making. In this context, we use data on heart transplantation in Section 5 for an illustration of our proposed model.

The rest of the paper is structured as follows. We introduce the parametric competing risks cure rate regression model in Section 2. The likelihood function formulation and estimation procedures are explained in Section 3. In Section 4, we report the results of simulation work to explain the bias of estimators on variations in samples size. Section 5 illustrates the application of the proposed model to real data set. Some concluding remarks are given in Section 6.

2. The Model

Suppose that population consists of two groups of subjects say, susceptibles and immunes. Let T be the time to occurrence of the event. Define the indicator variable function to define the status of cure

$$Y = \begin{cases} 1, & \text{if the individual eventually experience the event of interest} \\ 0, & \text{otherwise.} \end{cases}$$

Let p be the probability of occurrence of the event. The survival function of the uncured population at time t is $S(t|Y=1) = P(T > t|Y=1)$. Then survival function of cure rate model is

$$S(t) = (1-p) + pS(t|Y=1) \tag{1}$$

where $t < \infty$. Note that $S(t)$ tends to $(1-p)$ as $t \to \infty$. Let C = cause of death and the probability of uncured subjects $p_j = Pr(Y=1, C=j)$, $j = 1, 2, \ldots, k$. Assume that the time to occurrence of the event T is defined only when $Y=1$ and $C=j$, $j=1,2,\ldots,k$. Let $f_j(t|Y=1)$ be the probability density function and $S_j(t|Y=1)$ be the sub-survival function (Carriere and Kochar (2000)),of the random variable t due to jth cause, $j=1,2,\ldots,k$. For a censored individual, Y is not observed.

In the presence of competing risks, the survival function of cure rate model is

$$S(t) = 1 - \sum_{j=1}^{k} p_j + \sum_{j=1}^{k} p_j S_j(t|Y=1) \tag{2}$$

Let X be a $p+1 \times 1$ vector of covariates at incidence part and Z be a $p \times 1$ covariate vector at latency part of the model that is independent of X. In practical situations, the covariates X and Z can be same or may share common elements between them. Let $b_j = (b_{0j}, b_{1j}, \ldots, b_{pj})'$ be a vector of regression coefficients with $b = (b_1, b_2, \ldots, b_k)'$ for $j = 1, 2, \ldots, k$.

Then, in a competing risks Weibull regression model, the sub-survival function of t due to jth cause of failure with probability density function

$$f_j(t|Y=1,\theta,Z) = \alpha \exp(\beta_j Z)\, t^{\alpha-1} \exp(-t^{\alpha} \exp(\beta_j Z)) \tag{3}$$

is

$$S_j(t|Y=1,\theta,Z) = \exp(-t^{\alpha} \exp(\beta_j Z)) \tag{4}$$

where $\alpha > 0$, $\beta_j = (\beta_{j0}, \beta_{j1}, ...\beta_{jp})'$ is the vector of regression coefficients associated with the covariate Z, $\theta = (\alpha, \beta_j)$ and $\beta = (\beta_1, \beta_2, ...\beta_k)'$ for $j = 1, 2, \ldots, k$.

Under logistic regression model assumption, the probability of occurrence of the event due to jth cause is

$$p_j(b) = Pr(Y = 1, X) = \frac{\exp(b_j' X)}{1 + \sum_{j=1}^{k} \exp(b_j' X)} \tag{5}$$

for $j = 1, 2, \ldots, k$

Let $F_j(t) = \Pr(T \leq t, C = j)$ be the cumulative incidence function due to jth cause which measures the probability of occurrence of the event before time t due to cause j, $j = 1, 2, \ldots, k$.

Now, the cumulative incidence function due to jth cause in the presence of covariates X and Z and in the presence of $Y = 1$ is

$$\begin{aligned} \Pr(T \leq t, C = j | X, Z, Y = 1) &= \Pr(T \leq t | Z, Y = 1, C = j) \Pr(C = j, Y = 1 | X) \\ &= p_j(b)(1 - S_j(t | Y = 1, \theta, Z)) \end{aligned}$$

Now, the survival function of competing risks cure rate regression model is defined as

$$S(t, \Theta) = p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t | Y = 1, \theta, Z) \tag{6}$$

where $\Theta = (b, \theta)$ denotes the entire set of parameters and $p_0(b) = 1 - \sum_{j=1}^{k} p_j(b)$, the probability of immunes in the model. Suppose that the model parameters are linked to a single covariate Z. (ie., we use the assumption $X = Z$ throughout the paper). We also assume that an independent, non-informative, random censoring model and the censoring variable is statistically independent of Y. Inference procedure of the proposed model is given in the next Section .

3. Inference Procedures

Suppose we have data in the form $(t_{ij}, \delta_{ij}, z_i)$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ and $i \neq j$ where the notations

t_{ij} = the observed event or censoring time due to jth cause and the n distinct event times be $t_{1j} < t_{2j} < \cdots < t_{nj}$.

$$\delta_{ij} = \begin{cases} 1, & t_{ij} \text{ is uncensored} \\ 0, & \text{otherwise.} \end{cases}$$

and z_i = a vector of covariates.

The likelihood equation under multiple modes of failures is

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} (f_j(t_i))^{\delta_{ij}} (S(t_i))^{1-\delta_{ij}} \tag{7}$$

Under the model assumptions made, likelihood function of the cure rate regression model is

$$L(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b) f_j(t_i|Y=1,\theta,Z))^{\delta_{ij}} \left(p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t_i|Y=1,\theta,Z) \right)^{1-\delta_{ij}} \tag{8}$$

Let the complete data be $(t_{ij}, \delta_{ij}, z_i, y_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ which includes the observed data and the unobserved y_{ij}'s, where y_{ij} be the value taken by the random variable Y_i for jth cause. If $\delta_{ij} = 1$, $y_{ij} = 1$ and if $\delta_{ij} = 0$, y_{ij} is unobserved. Then the complete - data full likelihood is

$$\begin{aligned} L_c(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b) f_j(t_i|Y=1,\theta,Z))^{\delta_{ij} y_{ij}} (p_0(b))^{(1-\delta_{ij})(1-\sum_{j=1}^{k} y_{ij})} \\ (p_j(b) S_j(t_i|Y=1,\theta,Z))^{(1-\delta_{ij}) y_{ij}} \end{aligned} \tag{9}$$

By substituting the probability density function and the survival function given in (3) and (4), the above likelihood equation can be expressed as a product of two likelihood functions as

$$L_c(\Theta) = L_1(b) L_2(\beta, \theta) \tag{10}$$

where

$$L_1(b) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j(b))^{y_{ij}} (p_0(b))^{(1-\delta_{ij})(1-\sum_{j=1}^{k} y_{ij})}$$

and

$$L_2(\beta, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left(e^{\beta_j z_i} \alpha {t_i}^{\alpha-1} \right)^{\delta_{ij} y_{ij}} e^{(-{t_i}^{\alpha} \exp(\beta_j z_i) y_{ij})}$$

The likelihood function (10) contains missing observations since partial information of random variable Y is missing. Hence we employ EM Algorithm (Dempster *et al.* (1977)) to estimate the parameters of the model.

3.1. EM algorithm

E-Step : The expectation step (E-step) in the EM algorithm compute the conditional expectation of the complete data log-likelihood function $l(\Theta; y)$ with respect to y_{ij}'s, given the observed data and current estimates of the parameters.

Let the observed data be $\{O = (\text{Observed } y_{ij}\text{'s, } t_{ij}, \delta_{ij}, z_i);\ i = 1, ..., n\}$. Now we have to compute $\pi_j^{(m)} = E(y_{ij}|\Theta^{(m)}, O)$ where $\Theta^{(m)}$ denotes the values of parameters Θ at the mth iteration step. For uncensored i, $E(y_{ij}|\Theta^{(m)}, O) = y_{ij} = 1$. Now for the i'th censored observation, we compute

$$\pi_j^{(m)} = \Pr(Y_i = 1, |T_{ij} > t_{ij}, \delta_{ij} = 0, z_i; \Theta^{(m)})$$

$$= \left[\frac{p_j(b) S_j(t_i|Y=1,\theta,z_i)}{p_0(b) + \sum_{j=1}^{k} p_j(b) S_j(t_i|Y=1,\theta,z_i)} \right]_{|\Theta^{(m)}|}$$

i.e., at the m th iteration, the E-step value of y_{ij} is

$$w_{ij}^{(m)} = \begin{cases} 1, & \text{if the } i^{th} \text{ individual is uncensored} \\ \pi_{ij}{}^{(m)}, & \text{if censored} \end{cases} \tag{11}$$

$$l_c(\Theta; w^{(m)}) = l_1(b; w^{(m)}) + l_2(\theta; w^{(m)}) \tag{12}$$

denote the conditional expectation of the complete data log-likelihood function, where $w^{(m)}$ denote the vector of $w_{ij}^{(m)}$ values.

M-Step : In M-Step, we maximise the conditional expectation of the complete data log-likelihood function $l_c(\Theta; w^{(m)})$ with respect to each parameter in $\Theta = (b, \theta)$ given w_{ij} to obtain an improved estimate $\Theta^{(m+1)}$ at the $(m+1)$th iteration.

The procedures in E-step and M-step are then continued iteratively until we meet the convergence criteria to obtain maximum likelihood estimators of each parameter in the parameter set $\Theta = (b, \theta)$.

3.2. Asymptotic property of estimators

Let $\hat{\Theta} = (\hat{b}, \hat{\theta})$ denote the maximum likelihood estimates of $\Theta = (b, \theta)$, where $\hat{b} = \hat{b}_j$ and $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_j)$, $j = 1, 2, \ldots, k$. Now consider the following regularity conditions.

(a) The first and second order derivatives of the log-likelihood function l with respect to Θ viz., $\frac{\partial l}{\partial \Theta}$ and $\frac{\partial^2 l}{\partial \Theta^2}$ exist and are continuous functions of Θ in a range R (including the true value Θ_0 of the parameter) for almost all t. For every Θ in R

$\left|\frac{\partial l}{\partial \Theta}\right| < H_1(t)$ and $\left|\frac{\partial^2 l}{\partial \Theta^2}\right| < H_2(t)$ where $H_1(t)$ and $H_2(t)$ are integrable functions over $(-\infty, \infty)$.

(b) The third order derivative with respect to Θ, $\frac{\partial^3 l}{\partial \Theta^3}$ exists such that $\left|\frac{\partial^3 l}{\partial \Theta^3}\right| < M(t)$ where $E[M(t)] < Q$, a positive quantity.

(c) For every Θ in R,

$$E\left(-\frac{\partial^2 l}{\partial \Theta^2}\right) = \int_{-\infty}^{\infty} \left(-\frac{\partial^2 l}{\partial \Theta^2}\right) L dt = I(\Theta)$$

is finite and non-zero.

(d) The range of integration is independent of Θ. This assumption is to make differentiation under the integral sign valid.

Under the above mentioned regularity conditions, as $n \to \infty$,

$\sqrt{n}(\Theta - \hat{\Theta}) \to N_8(0, I^{-1}(\Theta))$, where the Fisher information matrix $I(\Theta)$ can be replaced by a consistent estimate $I(\hat{\Theta}) = \left(\frac{-\partial^2 l}{\partial \Theta_i \partial \Theta_j}\right)_{\Theta=\hat{\Theta}}$. The observed information matrix is obtained by applying Louis (1982) method. The variance of the estimates can be determined from diognal elements of $I^{-1}(\hat{\Theta})$. The asymptotic normality property of maximum likelihood estimates is useful to determine the $(1-\alpha) \times 100\%$ confidence interval of each parameter in the parametric set $\Theta = (b, \theta)$. Let $\hat{b}_j$ is the maximum likelihood estimator (MLE) of b_j. Then MLE of cured proportion $1 - p_j$ is $1 - \hat{p}_j = g(\hat{b}_j)$ is also asymptotically normally distributed by the invariance property of maximum likelihood estimators.

4. Simulation Studies

Simulation studies are conducted to evaluate the performance of the proposed model. Let C be the cause of failure and we assumed that there are two causes of failure. We consider a single covariate Z, which is generated from a uniform distribution over the interval (0,1). The censoring variable K is generated from uniform distribution over the interval (0,k) where k chosen in such a way that the lifetimes are mildly or heavily censored. The observations are followed up to a maximum time $\tau = 10$. The data for each observation be (t, δ, Z, C), where $t = \min(T, K, \tau)$ and δ be the event indicator. The data generated from the model with incidence probabilities

$$p_j(b) = \frac{\exp(b_{0j} + b_{1j} Z)}{1 + \sum_{j=1}^{2} \exp(b_{0j} + b_{1j} Z)} \tag{13}$$

for jth cause of failure, $j = 1, 2$. The cause specific survival functions are generated at random using the following sub distribution functions suggested by Dewan and Kulathinal (2007). Let,

$$\begin{aligned} F_1(t) &= P(T \le t, C = 1) = \phi F^a(t) \\ F_2(t) &= P(T \le t, C = 2) = F(t) - \phi F^a(t) \end{aligned} \tag{14}$$

where $1 \leq a \leq 2$, $0 \leq \phi \leq 0.5$ and $F(t)$ is the distribution function at time T. Note that $\phi = P(C = 1)$ and for $a = 1$, T and C are independent. The variables T and C are dependent for other choices of a. The nonnegative condition of cause specific density function of T is maintained by imposing these restrictions on the parameters. We choose the values $\phi = 0.25$ and $a = 1.5$ for simulating data. We fix the initial values $b_{0j} = 2$, $b_{1j} = -1$, $\beta_{0j} = 1.5, j = 1, 2$, $\beta_1 = -0.3$, $\beta_2 = 0.2$ and $\alpha = 0.2$. The initial values of the estimates are chosen using Kaplan-Meier estimate of cured proportion and log-likelihood equation of proposed the model (Balakrishnan and Pal (2012)). We generated random samples of sizes $n = 50$, 100 and 200 and maximum likelihood estimation of the parameters is carried out for the proposed model. The effect of censoring was studied in two situations viz, mild censoring (on an average, 20% of the observations are censored) and heavy censoring (40% of the observations are censored at average level). For the described configuration, 1000 replications are made. The results of absolute bias and MSE of the estimates are reported. The coverage probabilities (CP) of the 95% confidence intervals based on the asymptotic normality of the estimators are also reported. Table 1 shows the average absolute bias and MSE of estimates at different censoring levels. It seems that the proposed model and method work well. The parameters of the model are estimated with lower bias and MSE. There is a slight increase in bias and MSE as the censoring scheme changes from mild to heavy. The coverage probabilities of the asymptotic confidence intervals are also close to the pre-determined levels and it is found to be better for samples of increased size.

Table 1: Absolute Bias and MSE of estimators of parameters

			20% Censored			40% Censored		
Sample size	Parameter	True value	Bias	MSE	CP	Bias	MSE	CP
50	b_{01}	2.0	0.05822	0.013144	95.27273	0.090346	0.016547	95.03546
	b_{11}	-1.0	0.04495	0.01547	95.43568	0.05786	0.003348	95.19231
	b_{02}	2.0	0.07299	0.010114	95.00000	0.07698	0.013841	94.35028
	b_{12}	-1.0	0.05181	0.013698	95.50562	0.05601	0.014436	95.00000
	β_{01}	1.5	0.09006	0.027625	95.23810	0.099391	0.028006	94.83568
	β_{02}	1.5	0.09772	0.211936	95.70896	0.12413	0.242311	95.06173
	β_1	-0.3	0.09006	0.027625	95.23810	0.099391	0.028006	94.83568
	β_2	0.2	0.09772	0.211936	95.70896	0.20013	0.242311	95.06173
	α	0.2	0.00756	0.000681	95.84463	0.00979	0.001025	95.3125
100	b_{01}	2.0	0.02504	0.00976	96.29630	0.05485	0.01495	95.29220
	b_{11}	-1.0	0.04387	0.00360	97.72727	0.04486	0.00360	96.96970
	b_{02}	2.0	0.03678	0.00850	95.13880	0.04312	0.00931	96.23552
	b_{12}	-1.0	0.04452	0.00476	95.74468	0.04532	0.008003	95.55556
	β_{01}	1.5	0.07145	0.027625	95.23810	0.08236	0.028006	94.83568
	β_{02}	-1.5	0.08320	0.211936	95.70896	0.09008	0.242311	95.06173
	β_1	-0.3	0.05586	0.02350	95.58854	0.07920	0.01083	95.45455
	β_2	0.2	0.07491	0.06091	96.31902	0.08921	0.08549	96.31512
	α	0.2	0.00515	0.00049	95.94229	0.00594	0.00034	95.83333
200	b_{01}	2.0	0.02044	0.00112	98.53000	0.02144	0.01180	97.59450
	b_{11}	-1.0	0.01842	0.00335	98.54369	0.02253	0.00356	97.80220
	b_{02}	2.0	0.02052	0.00277	96.90000	0.02385	0.00622	95.45455
	b_{12}	-1.0	0.02765	0.00308	96.50000	0.03839	0.00479	96.35036
	β_{01}	1.5	0.04431	0.027625	95.23810	0.05319	0.028006	94.83568
	β_{02}	1.5	0.04749	0.211936	95.70896	0.05283	0.242311	95.06173
	β_1	-0.3	0.04488	0.01006	97.16981	0.05049	0.01329	95.50000
	β_2	0.2	0.06634	0.03085	97.29730	0.07233	0.05392	96.22302
	α	0.2	0.00254	0.00025	97.73960	0.00437	0.00023	96.51452

5. Data Analysis

To illustrate the applicability of the proposed model, we consider the data set from the Stanford Heart Transplant Program. The data contains the details of 103 patients selected for cardiac transplantation. A detailed description of data is available in Crowley and Hu (1977). We consider a subset of this data set with 63 patients who received the transplant to explain the application of the model. Out of these 63 transplant recipients, there were 27 (43%) deaths that occurred that due to rejection, 12 (19%) deaths from other causes and, the remaining 24 (38%) were censored observations. Survival time was measured in days from the date of transplant surgery. There are nine covariates in the original data set. We select only one covariate, the mismatch score, a key factor that influences survival of patients after heart transplantation (Miller (1976), Opelz, G. and Wujciak, T. (1994), Osorio-Jaramillo *et al.* (2020)) for the analysis of data. The mismatch score measures the degree of dissimilarity between the donor and recipient tissue concerning HLA antigens, and it is therefore related to the phenomenon of rejection of the donor heart by the recipient's immune mechanisms. If the mismatch score is less than one, it is a sign of good match, and if the score is high, greater than one represents a poor match (Miller (1976)). Hence we transform the selected continuous covariate mismatch score into a categorical variable of two categories with cut-off value one as per aforesaid classification criteria of matching and considered for the analysis of data. There are two causes of failure in the data. The cause of death attributable to rejection of the donor heart is labeled as cause 1 and cause of death due to other reasons such as surgical, kidney failure, hepatitis, etc, and not due to rejection of the new heart is labeled as cause 2.

As an initial step of the analysis, Kaplan- Meier plot is drawn for the data and displayed in Figure 1. The plateau in the given survival curve confirms the presence of immunes in the data. Hence the selected data is suitable for the analysis of cure rate models.

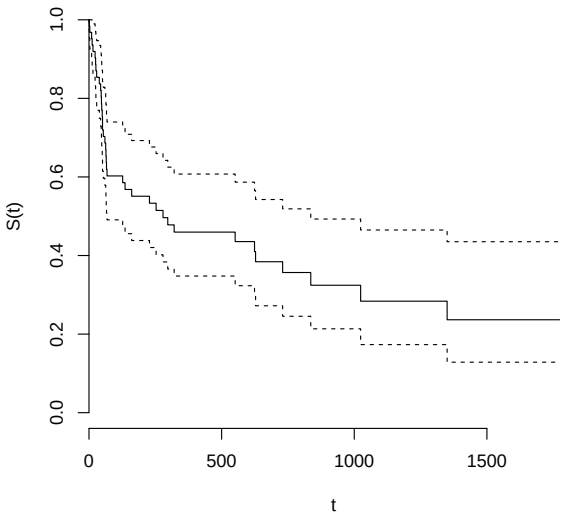

Figure 1: Kaplan-Meier survival curve of heart transplant data

In the present work, we are interested to study the effect of the covariate mismatch score on the survival of patients who have undergone heart transplantation in competing risks setting. The maximum likelihood estimators of regression coefficients are found out using (12) under the given model assumptions. The statistical significance of the regression coefficients is tested by the likelihood ratio test procedure. The estimates of regression coefficients with corresponding standard errors are reported in Table 2. The result shows that the higher mismatch score has a significant effect on rejection-related mortality among patients after heart transplant ($p = 0.013$) but may not affect the survival of patients ($p = 0.137$). The role of mismatch score is negligible on rejection related mortality of patients who died of competing causes.

Figure 2 displays plots of the estimated cumulative incidence rates for mismatch categories. From the Figure, it is obvious that the difference between cause specific failure rates is more in high score (> 1) category of mismatch score compared to low score (< 1) category. This is due to the variations in the influence of mismatch score on mortality of patients due to two causes of failure.

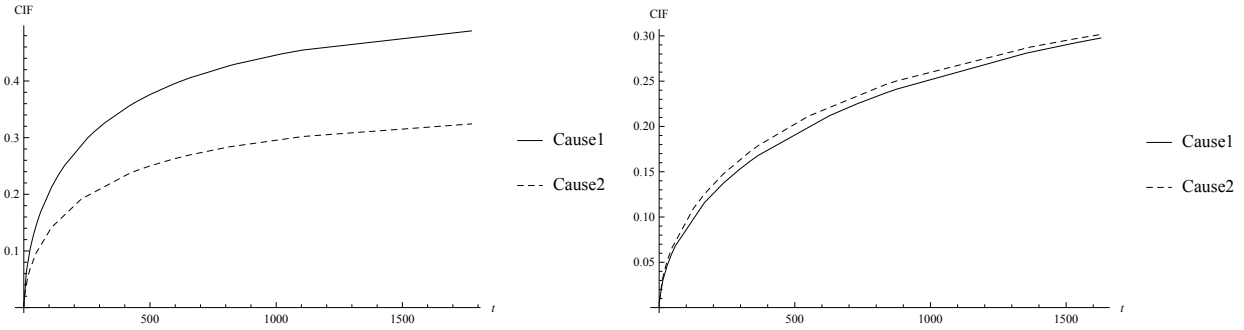

Figure 2: Cumulative incidence curve of mismatch score categories high score (left) and low score (right)

The estimated cured proportion among low score category (17.99%) is greater than that of high score category (12.41%). It bring out the influence of the selected covariate mismatch score on the survival of study subjects. The estimates and 95% Confidence interval of the probability of cure due to rejection and due to other causes obtained from the model are 0.47 (0.34, 0.61) and 0.65 (0.50, 0.81) respectively. The estimated values of cured proportion reveals the presence of cured individuals in the data and confirm the importance of the proposed model. The goodness of fit of latency part of the model is tested using Cox-Snell residuals with the modifications suggested by Peng and Tailor (2017). We consider the Cox-Snell residuals $r_i = -\log S_j(t|Y = 1, \theta, Z)$ using (4). The residuals for each cause of failure estimated with different weights as given in (11) for the censored and uncensored observations. The Kolmogorov - Smirnov test is performed to assess the unit exponentiality of the data and p values obtained as $p = 0.25$ for rejection and $p = 0.12$ for other cause of failure. The values indicate that the model fits well for the given data to explain each cause of failure.

Table 2: Estimates of parameters and Standard Error (SE)

	Rejection ($j=1$)			Other Causes ($j=2$)		
Estimates	Est	SE	P value	Est	SE	P value
b_{0j}	0.861	2.53×10^{-3}	-	0.784	2.59×10^{-3}	-
b_{1j}	0.585	5.14×10^{-3}	0.013	0.251	5.46×10^{-3}	0.654
β_{0j}	-4.135	3.93×10^{-3}	-	-3.950	3.87×10^{-3}	-
β_{1j}	0.730	2.11×10^{-3}	0.137	0.555	2.48×10^{-3}	0.002

6. Conclusion

In this paper, we proposed a regression model with Weibull distribution for the analysis of competing risks data with long term survivors. Maximum likelihood inference via EM algorithm was implemented to estimate the parameters of the model. The goodness of fit of the latency model checked using modified Cox-Snell residuals. The model was illustrated with a real lifetime data on Stanford Heart Transplant Program and distinguished the effect of covariate on short and long term survival of patients after heart transplant in competing risks scenario. This article aimed to evaluate the effect of covariates such as clinico-social variables, different treatment regimens and other prognostic factors on survival of patients suffering from diseases when there is a chance of cure in the presence of competing risks and expected to be useful for investigators in the field of survival analysis. The regression analysis of interval censored data with cured proportion is also challenging in the field of survival analysis. The work in this direction is under progress and it will be communicated in a future paper.

Acknowledgements

The authors are thankful to the referee and editor for the constructive comments and suggestions on earlier version of this manuscript that appreciably improved the article.

References

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.

Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice*, **6(4)**, 698-724.

Carriere, K. C. and Kochar, S. C. (2000). Comparing sub-survival functions in a competing risks model. *Lifetime Data Analysis*, **6(1)**, 85-97.

Choi, S., Zhu, L. and Huang, X. (2018). Semiparametric accelerated failure time cure rate mixture models with competing risks. *Statistics in Medicine*, **37(1)**, 48-59.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC Press, London.

Crowder, M. J. (2001). *Classical Competing Risks*. CRC Press, London.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72(357)**, 27-36.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, **39(1)**, 1-22.

Dewan, I. and Kulathinal, S. (2007). On testing dependence between time to failure and cause of failure when causes of failure are missing, *PLoS One*, **2(12)**, e1255.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. **360**, John Wiley & Sons, New York.

Larson, M. and Dinse, G. (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.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. John Wiley& Sons, New York.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63(3)**, 449-464.

Nelson, W. B. (2003). *Applied Life Data Analysis*. John Wiley& Sons, New York.

Opelz, G. and Wujciak, T. (1994). The influence of HLA compatibility on graft survival after heart transplantation. *New England Journal of Medicine*, **330(12)**, 816-819.

Ortega, E. M., Cordeiro, G. M. and Hashimoto, E. M. (2011). A log-linear regression model for the Beta-Weibull distribution. *Communications in Statistics: Simulation and Computation*, **40(8)**, 1206-1235.

Osorio-Jaramillo, E., Haasnoot, G. W., Kaider, A., Schaefer, A. K., Haberl, T., Goekler, J.,et al. (2020). Molecular-level HLA mismatch is associated with rejection and worsened graft survival in heart transplant recipients–a retrospective study. *Transplant International*, **33(9)**, 1078-1088.

Papastefanou, I., Nowacka, U., Syngelaki, A., Dragoi, V., Karamanis, G., Wright, D. and Nicolaides, K. H. (2021). Competing-risks model for prediction of small-for-gestational-age neonate from estimated fetal weight at 19–24 weeks' gestation. *Ultrasound in Obstetrics & Gynecology*, **57(6)**, 917-924.

Peng, Y. and Taylor, J. M. (2017). Residual-based model diagnosis methods for mixture cure models. *Biometrics*, **73(2)**, 495-505.

Rejani, P. P. and Sankaran, P. G. (2020). Modeling and Analysis of Proportional Hazards Competing Risks Cure Rate Model. *Journal of the Indian Society for Probability and Statistics*, **21(1)**, 175-185.

Shen, P. S., Chen, H. J., Pan, W. H. and Chen, C. M. (2019). Semiparametric regression analysis for left-truncated and interval-censored data without or with a cure fraction. *Computational Statistics & Data Analysis*, **140**, 74-87.

Sreedevi, E. P. and Sankaran, P. G. (2021). Statistical methods for estimating cure fraction of COVID-19 patients in India. *Model Assisted Statistics and Applications*, **16(1)**, 59-64.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model, *Biometrics*, **56(1)**, 227-236.

Wright, D., Wright, A. and Nicolaides, K. H. (2020). The competing risk approach for prediction of preeclampsia. *American Journal of Obstetrics and Gynecology*, **223(1)**, 12-23.

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.

Yusuf, M. U. and Bakar, M. R. A. (2016). Cure models based on Weibull distribution with and without covariates using right censored data. *Indian Journal of Science and Technology*, **9(28)**, 1-12.

Algorithm for maximum likelihood estimation of parameters of the model

1. Determine the parameter values b_j, β_j, and α for $j = 1, 2$; (Select the initial values of the parameters and input these values in first stage).
2. For the ith subject, generate the covariate X_i from Uniform(0,1);
3. Find out the probability of incidence $p_j(b)$ for $\forall$ X_i and $j = 1, 2$;
4. Generate censoring variable K_i from Uniform(0,k), where k is set to control the proportion of censored observations;
5. Generate a random variable u_i from Uniform(0,1);
6. Take v_i as the root of $F(t) - u_i = 0$, where $F(t)$ is the distribution function corresponding to the model;
7. Find $t_i = \min(v_i, K_i, \tau)$, τ=10 (assumed). If $t_i < K_i$, set $\delta_i = 1$, otherwise $\delta_i = 0$;
8. Find out survival functions $S_j(t)$ for $j = 1, 2$ and $S(t)$;
9. Find out $\psi_i = 1 - \phi a(1 - S(t_i))^{a-1}$ for $i = 1, 2, \ldots, n$.; (Dewan and Kulathinal (2007))
10. Generate g_i from Uniform(0,1);
11. If $g_i < \psi_i$, set cause = 1, otherwise cause = 2;
12. Now the data set for the ith subject is $(y_{ij}, t_{ij}, \delta_{ij}, X_i)$, $i = 1, 2, \ldots, n$, $j = 1, 2$;
13. Find out the expected value π_j for $\delta_{ij} = 0$, $j = 1, 2$;
14. Assign y_{ij} =1, if $\delta_{ij} = 1$. Otherwise y_{ij} =π_j according to cause j. ($y_{ij} = w_{ij}$);
15. Maximize the complete data log-likelihood function and estimate the parameters;
16. Repeat the procedure of Expectation-Maximization till the convergence criteria is met to get improved estimate (say, $\lambda - \hat{\lambda} < \delta$, a pre defined small quantity for parameter λ)
17. Replicate the required number of data sets.