Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v15n1p95

Mixture cure survival analysis model for cardiovascular disease in Sulaimani, Iraq
By Ahmad, Ahmed

Published: 20 May 2022

Electronic Journal of Applied Statistical Analysis
Vol. 15, Issue 01, May 2022, 95-109
DOI: 10.1285/i20705948v15n1p95

Mixture cure survival analysis model for cardiovascular disease in Sulaimani, Iraq

Shokh Mukhtar Ahmad*[b] and Nawzad Muhammed Ahmed[a]

[a]*Department of Statistics and Informatics College of Administration and Economics Sulaimani University, Kurdistan*
[b]*Department of Medical Laboratory Sciences, Komar University of Science and Technology, Kurdistan*

Published: 20 May 2022

Cardiovascular disease(CVDs) is one of the leading causes of death worldwide. Iraq ranks 20th in the age adjusted Death Rate due to CDVs. In recent years, the treatment of many diseases, especially heart disease, has significantly improved, so the number of patients who do not experience the desired outcome, including death, has increased. In statistical analysis of this type of diseases, cure models are used instead of the usual survival models. In this paper, a sample include 919 patients referred to Sulaimani Hospital with heart disease (including 365 female and 554 male) were followed up for a maximum of 650 days, during the years 2020 to 2022. Of these, 162 people, or 17.6%, have died. Since the Maller-Zhou test was significant ($P < 0.01$) and considering the cured fraction in this population, the mixture cure model with some statistical distributions was fitted to the data. Based on the results and comparing AIC and BIC, it was observed that the healed model combined with Weibull distribution for survival time and Poisson distribution for the number of deaths with the AIC=1972.54 , BIC=2092.985 was the best model.

keywords: mixture cure survival,cardiovascular disease, Weibull distribution, Iraq.

*Corresponding author: shokh.mukhtar@gmail.com.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

Unfortunately, in recent years, cardiovascular disease (CVDs) has been one of the leading causes of death worldwide. According to the World Health Organization, published in 2018, about 17.9 million people die annually. The countries of the Middle East, especially Iraq, are in a worse position than the rest of the world. According to the latest WHO report, about 19 percent of all deaths in Iraq are due to coronary heart disease. According to these statistics, Iraq ranks 20th in the age adjusted Death Rate 230.27 per 100,000 people.

In recent years, the treatment of many diseases, especially heart disease, has significantly improved, so the number of patients who do not experience the desired outcome, including death, has increased. In statistical analysis of this type of diseases, cure models are used instead of the usual survival models.

Survival analysis models are common models in statistics in order to estimate the probability of diseasers' survival and to investigate the factors affecting them. Cox regression models are the most widely used of these methods in such situations by fitting a variety of survival time distribution functions, such as exponential distribution. However, as the treatment of many diseases, especially heart disease, has progressed significantly in recent years, the number of patients who have received the desired results and mostly do not experience death has increased.

Ordinal survival models assume that all subjects are prone to the event (death) and will eventually experience it. It is not appropriate because the fraction of people who have not experienced the event is not considered in the model and therefore incorrect and sometimes misleading estimates will be obtained from the model parameters (Farewell, 1982; Maller and Zhou, 1996). The use of cure models eliminates these problems. One of the advantages of the cure model is that in these models both the factors affecting the survival functions and the cure ratio are considered (Tsodikov A, 2003).

In the recent 20 years, several studies have been performed on cure models and significant advances have been made in this field. But these models are not common in all areas, and most of them focus on diseases with high chance to cure related issues. There are various statistical methods for estimating and evaluating the factors affecting the incidence of death in such patient, which are mostly done by standard survival models (Cox model, parametric models) and without considering high censorship in these models.They focus on the event in question and do not pay attention to the high percentage of censorship in the study so data analysis is biased (Klein and Moeschberger, 2005; Kleinbaum and Klein, 2012).

cure models are divided into two general types: mixture cure model and non mixture cure model. The mixture Cure model can be an alternative to cox proportional hazard models in these situations when we have significant fraction of cured subject in the cohort (John P. Klein and Scheike, 2014).

The purpose of this study was to identify the factors affecting patient cure Cardiovascular using Weibull and Poisson models.

In the northern part of Iraq, the Kurdistan Region, many medical facilities and specialized doctors help these diseasers return to normal life. When a significant fraction of

the studied cohort are improved and returned to the community, alternative statistical models called mixed cured models are used in the analysis of these diseases instead of the ordinary survival models. In the section 2, we will examine these types of models and their differences with normal survival analysis models. In the section 3, we will analysis the follow-up data from 919 diseaser identified at the Sulaimani Hospital in Iraqi Kurdistan.
In this paper, a sample include 919 patients referred to Sulaimani hospital with heart disease (including 365 female and 554 male) were followed up for a maximum of 650 days in 2020 to 2022. Of these, 162 people, or 17.6%, have died.Since by Maller and Zhou test, observed the significance cured fraction in this cohort, the mixture cure model with some statistical distributions was fitted to the data (Maller and Zhou, 1996). Based on the results and comparing AIC and BIC, it was observed that the cured model combined with Weibull distribution for survival time and Poisson distribution for the number of deaths has the less AIC and BIC as the others models.

2 Cure survival model

For the first time cure models introduced by Boag (1949), and further study was done by Berkson and Gage (1952). The fraction that survives the event is called the cured or immune individuals and the other fraction of the cohort that is prone to the event is called the uncured individuals. Generally in parametric survival analysis there are two kinds of cure models as follows are divided into mixture and non mixture models.

2.1 Mixture models

The survival function of the community in the mixture models is defined as follows:

$$S(t) = P(T > t) = P(T > t|B = 1)P(B = 1)+P(T > t|B = 0)P(B = 0) = \pi+(1-\pi)S_u(t) \quad (1)$$

In the above equation, S(t) is the probability function of survival over time t for any case in the cohort, π is the probability of curing and $S_u(t)$ is the probability function of survival over time t for event-prone individuals with a parametric distribution. In this model, the π as the ratio of cured or immune individuals can be modeled by logistics regression. It should be noted that the cured models in the absence of cure individuals are the same as the standard models of survival. where in parametric methods to its modeling, lifetime distributions such as exponential, Weibull, etc are Used.

2.2 Non mixture models

The non mixture model introdused first time for modeling of tumor recurrence, where the cure fraction is the probability that no clonogenic cancer cells remain (Tsodikov A, 2003). However, non mixture model can be considered a useful mathematical function with an asymptote that can be applied to estimate the cure fraction and also is useful for

data that do not fit the above biological definition as long as assuming cure is reasonable (Chen and Sinha, 2001). In the non mixture cure model, the survival function is:

$$S(t) = e^{(ln(\pi)F_z(t))} \tag{2}$$

where π is the cure probability and $F_z(t)$ is a distribution function to be same $1-S_z(t)$, here $S_z(t)$ is a standard parametric survival function, such as as the Weibull or the others distribution function.

2.3 Maller and Zhou test for existence of cured fraction

Cure survival models is based on existence significant fraction of cure or immune individuals in the cohort. In (1) this proportion shown by π. Maller and Zhou (1992) proposed an estimator for this proportion when a sample of censored failure times is available. This is to use one minus the maximum observed value of the Kaplan-Meier empirical distribution function. By data simulation they shown this estimator is to be consistent and asymptotically normal, under modest conditions on the censoring mechanism. Since the estimator is approximately normal for a small sample size, provided the immune proportion is not too close to zero. This is a non parametric statistic to test whether the assumptions of the cure fraction analysis are likely to be valid.

2.4 Parametric distributions in cure models

A wide range of time-based statistical distributions can be implemented in survival analysis to model survival time. Also these parametric distributions used for too the mixture and non mixture cure models. In This paper for cardiovascular disease(CVDs), the some distribution such as exponential, Weibull, log-logistic, extreme value, gamma, log-normal, Marshall–Olkin exponential, generaliesd gamma and generalised F distributions are all implemented and compered results.

Choice of parametric distribution can be effective on the estimate of the cure fraction. Lambert (2007) found that in his experience, the Weibull distribution works well for most examples, except when there is a high cure fraction (e.g., $> 80\%$) or a high excess mortality rate in the first few weeks of follow-up. This problem may be often occurs in a cohort with a aged patients. The log-normal distribution rarely provides a good estimate of the cure fraction in cancer studies because of its having a long tail and an imposed rise and fall of the (excess) hazard function. The (generalized) gamma distribution is potentially useful because it has the Weibull, exponential, log-normal, and standard gamma distributions as special cases. About cardiovascular disease we observed a same result.

2.5 Akaike and Bayesian Information criterion

Model evaluation is an important step in statistical analysis and modeling, specially in survival analysis. There are many criterion for model selection. Although most of them are based on errors, but information criteria provide an attractive basis this aim.

The Akaike Information Criterion (AIC), is an index for scoring and selecting a model. Collett (2015) introdused the AIC for survival models as follows:

$$AIC = -2log(likelihood) + 2(p + 2 + k) \tag{3}$$

where $k = 0$ for the exponential model, $k = 1$ for the Weibull, log-logistic and log-normal models, and $k = 2$ for the generalized gamma model.
Liang and Zou (2008) suggested following formula for survival model:

$$AIC_{SUR} = AIC + \frac{2(p+2)(p+3)}{n-p-3} \tag{4}$$

This criterion indicates the amount of information loss due to model acceptance instead of data, so the model with the lowest AIC is better and will be selected. To use AIC for model selection, we simply choose the model giving smallest AIC over the set of models considered. (Hastie et al., 2001)
Bayesian Information Criterion (BIC) Like AIC, it is appropriate for models fit under the maximum likelihood estimation framework. BIC is a close approximation to the Bayes factor when a unit prior information about the parameter space is used. Volinsky CT (2000) introduced a revision of the penalty term in BIC so that it is defined in terms of the number of uncensored events instead of the number of observations. AIC, Compared to the BIC, penalizes complex models less, meaning that AIC may be select more complex models. (Murphy, 2013)

2.6 Cox proportional hazards cure regression model

Cox (1972) introduced proportional hazards (PH) model based on regression model which have been widely used in survival analysis. When these models are specified parametrically, the underlying assumption is that the event of interest will eventually occur. This assumption is not appropriate for a cohort with cured fraction. The Cox PH model can be used potentially for the cure information by setting the survival function to 0 after a time threshold. But, when modeling this way, long-term survivors cannot be distinguished from cured fraction.(Wu et al., 2014)
The Cox PH cure model can be written as a mixture model in terms of the survival function such as formula (1) with an additional covariate vector:

$$S(t|X,Z) = P(T > t|X,Z) = \pi(Z) + (1 - \pi(Z))S_u(t|X) \tag{5}$$

and

$$F(t|X,Z) = 1 - S(t|X,Z) \tag{6}$$

then

$$f(t|X,Z) = \frac{-dS(t|X,Z)}{dt} = -(1 - \pi(Z))f_u(t|X) \tag{7}$$

where X and Z are the covariate vectors, $\pi(Z)$ is the cured probability for an individual, and $S(t|X,Z)$ is the survival function. Let $f_u(t|X)$ and $S_u(t|X)$ be the probability density function and the survival function for uncured fraction.

The "incidence" part $(1 - \pi(Z))$ is modeled by logistic regression. The "latency" part $f_u(t|X)$ or $S_u(t|X)$ is modeled by the Cox PH model.

In multiple regression model with X and Z vector covariats, let β and γ be the parameter vectors related to X and Z, respectively. If we model this follwoup data by using the Cox PH cure model specified in (5), $\pi(Z) = \frac{1}{1+exp(\gamma' Z)}$, and $h_u(t|X) = h_0(t)exp(\beta' X)$ and $S_u(t|X) = S_0(t)^{exp(\beta\prime X)}$ are the hazard function and survival function of uncured individual, where $h_0(t)$ and $S_0(t) = exp(-\int_0^t h_0(u)\, du\ )$ are unspecified basic hazard and survival functions, respectively.

When we have some covariates to a better control of survival and hazard function, the conditional survival function of the cohort can be modeled by using cure model which is based on the probability of being uncured (incidence) and the conditional survival function of the uncured individual (latency), and a mixed of logistic regression and Cox proportional hazards (PH) regression is used to model the incidence and latency. Mohammad et al. (2020) in their paper have shown the asymptotic normality of the profile likelihood estimator via asymptotic expansion of the profile likelihood and obtain the explicit form of the variance estimator with an implicit function in the profile likelihood.

In many studies, there are different methods available to provide additional information about whether an individual is cured. However, the further information about cured status may not be available for all individuals, and all procedures are likely associated with a certain degree of accuracy in terms of sensitivity and specificity. Complete separation of cured and uncured individuals in the censored fraction can be difficult to achieve. Hence the PH cure model that incorporates the further information also needs to take into account the sensitivity and specificity of the diagnostic procedure that produces this further information. (Sy and Taylor, 2000)

3 Data Analysis

As mentioned in the introduction, this study was conducted to model and analyze the survival time of cardiovascular patients recognized in Sulaimani Hospital. The sample includes 919 cardiovascular patients, who were followed up in a maximum of 650 days between 2020 to 2022. Demographic characteristics of patients such as gender, age, place of residence, job, etc., as well as interventions performed by the medical team, such as the type of surgery, the treating physician, etc., have been recorded as covariate variables. All data in the form of survival setting is entered in the R program and all relevant analyzes are performed in this environment.

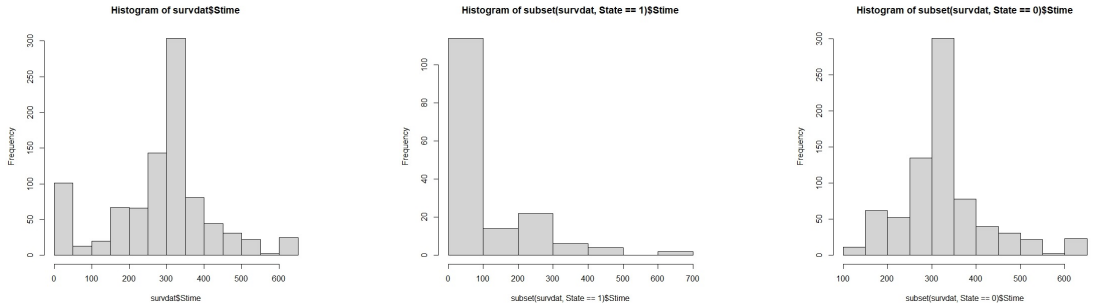

Figure 1: Histograms of survival time for three different parts of cohort, respectively for all, failed and censored patients

According to the histograms (Figure 1) it is seen that unfortunately some patients died only a few days after entering the hospital. However, a closer look at the 3rd histogram, the time of censoring of non-failed patients, existence of a cured fraction is obvious from those who have failed.

3.1 Survival Analysis

In a standard survival setting, without considering the cured and uncured fractions, the results of the Kaplan-Meier model are shown in Table 1 and Figure 2. The smoothing of the tail in the Kaplan Meier plot also confirms the existence of a significant cured fraction in this cohort.

Table 1: Kaplan-Meier life table

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
30	826	96	0.896	0.0101	0.876	0.916
60	814	9	0.886	0.0105	0.865	0.907
90	808	7	0.878	0.0108	0.857	0.900
180	751	15	0.862	0.0114	0.840	0.884
270	593	19	0.838	0.0123	0.814	0.862
360	181	7	0.825	0.0130	0.800	0.851
450	81	7	0.779	0.0211	0.739	0.822
540	27	0	0.779	0.0211	0.739	0.822
630	9	1	0.692	0.0837	0.546	0.878

As can be seen from the table and Kaplan-Meier plots, for the probability of survival in the first days, there is a break down, that is due to the presence of emergency patients.

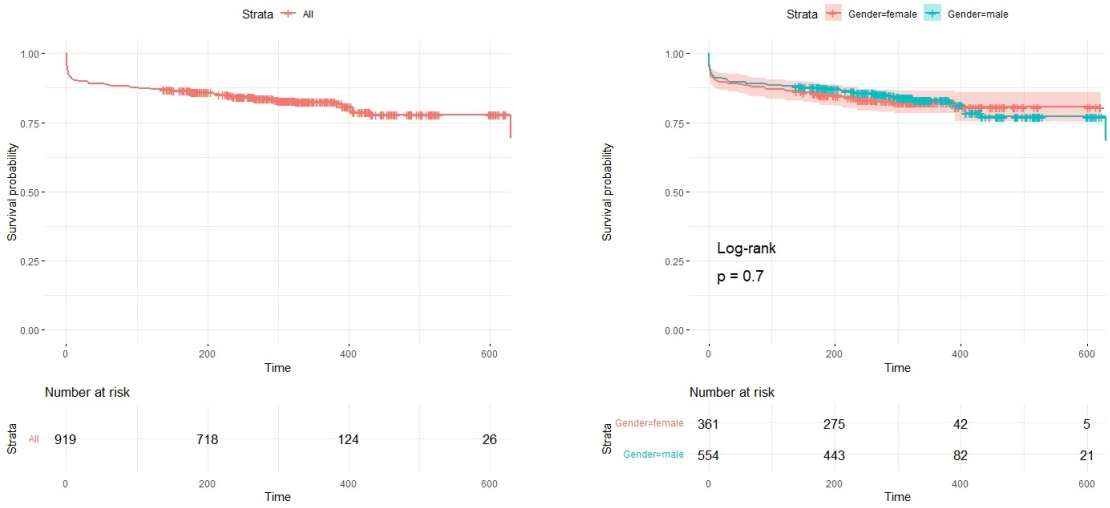

Figure 2: Kaplan-Meier plot

So that the probability of survival for more than 30 days is $S(30) = P(T > 30) = 0.896$. But gradually we will reach patients who are in a more stable condition, as it can be seen at the end of the table that the probability of survival for more than 630 days is $S(30) = P(T > 30) = 0.692$. In Kaplan-Meier plot in Figure 1, this can be seen as smoothing on the right tail of the plot. This in fact reinforces the formation of the idea of cure survival analysis in this study.

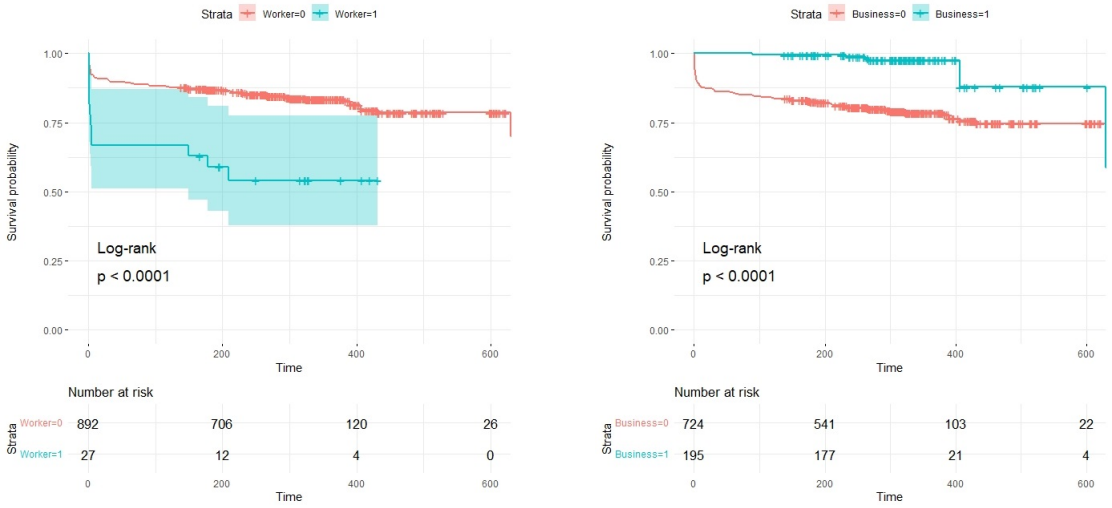

Figure 3: Kaplan-Meier plot

Also, according to Kaplan-Meier plots, based on dummy (binary) variables: gender (Figure 2), occupation (Figure 3), location (Figure 4), type of medical intervention (Figure 5) and physician (Figure 6), the probability of patients' survival in different levels of these variables can be compared.

The comparison of the survival plot of male and female in Figure 2, shows that men had a higher probability of survival in the first months, but this trend was changed after the about 400th day onwards. In other words, women have better conditions, after stabilization. Although this difference is not significant ($P > 0.05$).

Regarding jobs, in Figure 3, business as a job with a high level of welfare and labor as a job with a low level of welfare was choices. It is clear that those who were businessmen were more higher survival, but those who worked as laborers were had lower survival probability than others. These differences are significant ($P < 0.001$).

In Figure 4, the habitation of the patients is the factor that separates the plots. In order to understand the conditions of cardiovascular patients referred to Sulaimani city hospital, in two separate plots, Sulaimani city as the host city and its Kurdish neighbor province, but outside the Kurdistan region, ie Kerkuk has been selected. Patients admitted from Kerkuk are more likely to survive than others, as they are largely non-emergency due to their relatively long journey. Although this difference is not significant (P¿0.05). However, in the case of the city of Sulaimani, where the hospital is located, it can be seen that patients coming from the city themselves are less likely to survive than others who have been transferred to more distant places, because they also include the emergency patients ($P < 0.001$).

As mentioned earlier, significant advances in the treatment of cardiovascular disease have played a significant role in increasing the likelihood of survival of this type of patient. Coronary angiography, especially if associated with PCI, as well as Coronary Artery By-

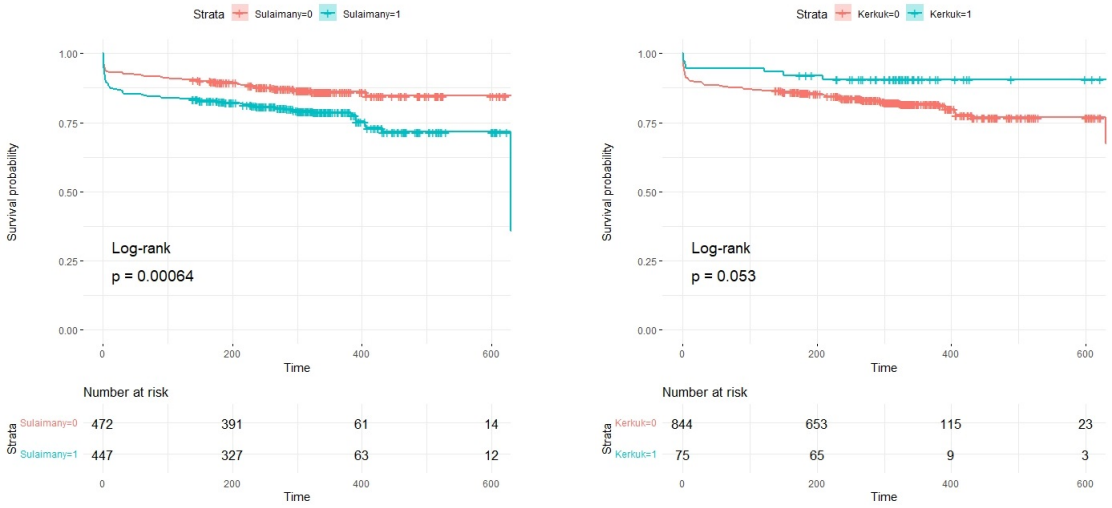

Figure 4: Kaplan-Meier plot

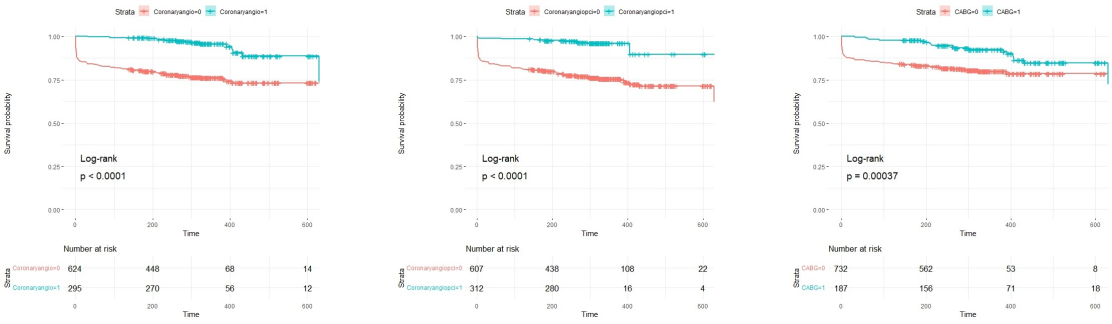

Figure 5: Kaplan-Meier plot

pass Graft (CABG) surgery are very effective in maintaining the survival of this cohort of patients. Figure 5 shows that all three types of interventions significantly increased the probability of survival ($P < 0.001$).

The role of the human factor in the treatment sector, especially doctors, is very important. To investigate this role, Kaplan-Meier plot of patients with three different physicians is shown in Figure 6. To protect medical ethics, in this article, treating physicians are named A, B and C. Looking at the plots in Figure 6, it is clear that patients treated by these physicians were more likely to survive than others. This difference was significant for A at $P < 0.001$, B at $P < 0.01$ and for C at $P < 0.05$ level.

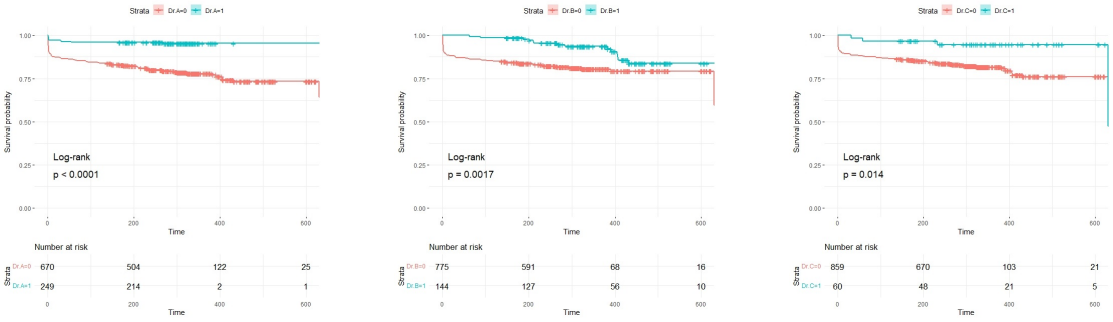

Figure 6: Kaplan-Meier plot

3.2 Mixture Cure Survival Analysis

First, Maller and Zhou test for testing of existence of cured fraction was applied.

Table 2: Maller-Zhou test

statistic	n	p.value
6	919	0.002430466

As the results of the non-parametric Maller and Zhou test show in Table 2, the null hypothesis is rejected and the existence of a cure fraction is accepted ($P < 0.01$). So appling a cure model can be useful for this situation.

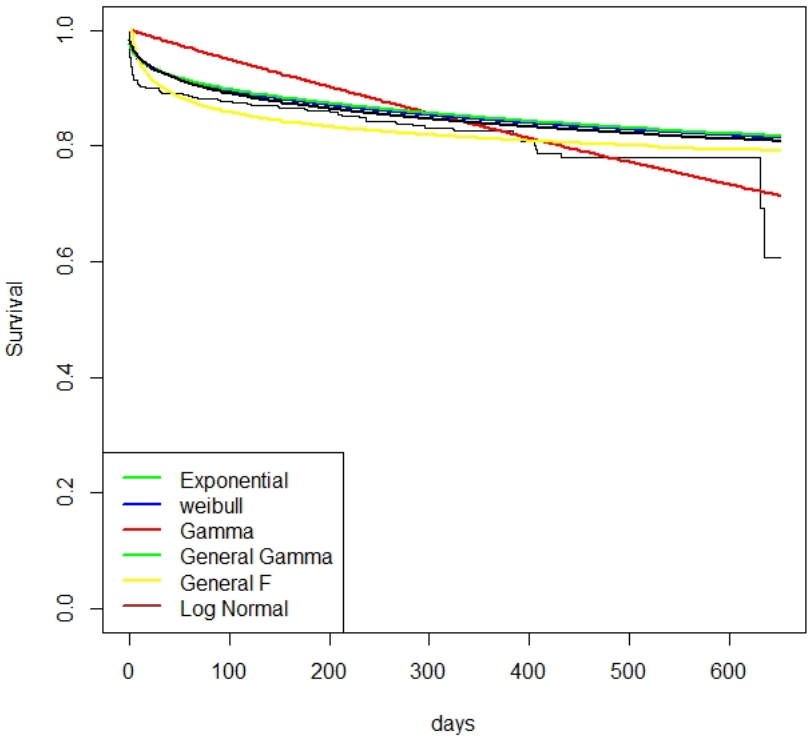

Figure 7: Fitting Some Distributions on the Kaplan-Meier Survival Plot

Based on the results of fitting the cure regression model with the covariate variables and comparing the Akaike and Bayesian information criteria (AIC and BIC) shown in Table 3 and Figure 8, it can be seen that the mixture cure model with the Weibull distribution fitting over the survival time is the best model for the data. In this case, AIC, BIC and Log Liklihood criteria provide the same results. Therefore, for the data analysis of this from cardiovascular patients cohort , the best survival model is mixture cure Weibull. The results of the model is summarized in Table 4 and 5.

Table 3: AIC, BIC and Log-Likelihood of the Cure Regression Models with Some Survival Time Distribution

	Exponential	Weibull	Gamma	Gen. Gamma	Log-normal	Gen. F
AIC	2207.647	1965.864	2145.791	1995.768	1972.198	1990.205
BIC	2323.275	2086.31	2266.236	2121.031	2092.644	2120.287
Log-Likelihood	-1079.824	-957.932	-1047.895	-971.8839	-961.0991	-968.1027

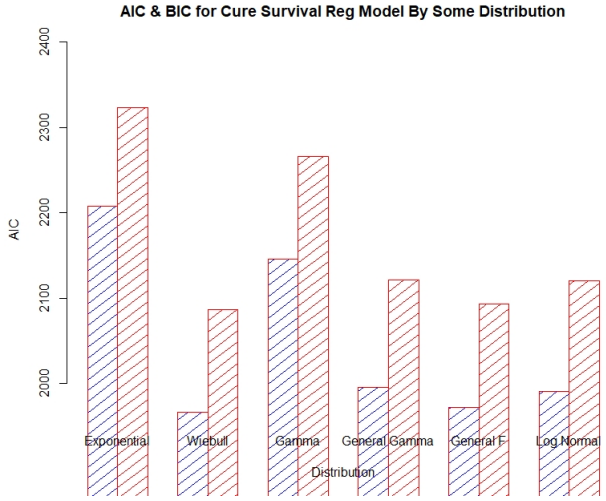

Figure 8: AIC, BIC and Log-Likelihood of the Cure Regression Models with Some Survival Time Distribution

n = 914, Events: 158, Censored: 756 Log-likelihood = -957.932 AIC = 1965.864

4 Conclusions

Analysis of 650-day follow-up data from 919 cardiovascular patients referred to a hospital in the city of Sulaimani in the Iraqi Kurdistan region, showed that there was a significant cure fraction among them. Maller and Zhou nonparametric test also confirmed this conjecture. By testing the variables correlated with patient survival, mixture cure several regression models with some different survival probability distributions were fitted to the data. Comparisons of AIC, BIC, and log of the likelihood function all identified the Weibull distribution as the best distribution. So we chose the Weibel distribution model to fit. Age, occupational, place of residence, medical interventions and physician were the identified variables and affected the survival of patients.

Table 4: Estimation of Coefficient in Cure Survival Regression Model, Distribution: Weibull promotion time model

Covariates	Cure probability model	Failure time distribution model
(Intercept)	-2.9172	4.4315
Age	0.4520	0.9510
Labor	0.1087	-1.5406
Business	-0.0270	1.6741
Sulaimani	0.0331	-0.0377
Kerkuk	-1.3694	-0.3082
Dr.A	-0.9897	1.9312
Dr.B	-2.0322	-2.3022
Dr.C	-1.3374	-0.4380
Coronaryangio	-1.9118	1.0772
Coronaryangio&pci	-2.9989	-0.5095
CABG	-1.3922	-1.5868
Log(shape)	-	-0.8488

Acknowledgement

First and foremost, we would like to sincerely thank the respected doctors and staff of Sulaimani Center For Heart Disease for their cordial guidance and constant supervision in enabling us with the access to the necessary information related this research.

References

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

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

Chen, J. G. I. M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the royal statistical society series b-methodological*, 34:187–220.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

John P. Klein, Hans C. van Houwelingen, J. G. I. and Scheike, T. H. (2014). *Handbook of Survival Analysis*. Boca Raton: CRC Press.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-learning Text*. Springer Science & Business Media.

Lambert, P. C. (2007). Modeling of the cure fraction in survival studies. *Stata Journal*, 7(3):351–375.

Liang, H. and Zou, G. (2008). Improved aic selection strategy for survival analysis. *Comput Stat Data Anal*, 52(5):2538–2548.

Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79(4):731–739.

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

Mohammad, K. A., Hirose, Y., Surya, B., and Yao, Y. (2020). Efficient estimation for the cox proportional hazards cure model.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

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

Tsodikov A, Joseph G. Ibrahim, Y. A. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *J Am Statist Ass.*, 98(464):1063–1078.

Volinsky CT, R. A. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1)::256–262.

Wu, Y., Lin, Y., Lu, S.-E., Li, C.-S., and Shih, W. J. (2014). Extension of a cox proportional hazards cure model when cure information is partially known. *Biostatistics*, 15 3:540–54.