Bayesian Analysis: MTH 535A

Project report

April 28, 2022

A Bayesian Approach to Assess Survival Time of Heart Failure Patients

---

Submitted by : **Bayes Squad**

Abir Naha(201257),
Koyel Pramanick(201333),
Sayan Bhowmik(201409),
Soumyadip Sarkar(201431)

Supervised by :
Dr. Arnab Hazra

Contents

1 Introduction

1.1 Background of the project

Heart failure basically means that the heart muscle doesn't pump blood as well as it should. When this happens, blood often backs up and fluid can build up in the lungs, heart rate can increase, causing shortness of breath, wheezing, ankle swelling etc. Heart Failure(HF) is defined as a clinical syndrome not a disease.

Heart failure is a growing problem in the world and the overall prevalence of HF in the adult population in developing countries is 7%-10% with an exponential rise with age ([3]). Though, heart failure is more common in people above 60 but that does not mean children and younger are protected from it. Heart failure patients in India are almost 10 years younger than patients in the US and Europe.[10] India houses about 40 percent of the world's 2.6 crore heart failure patients. As per projections. [7]

In most of the cases descriptive statistics or logistic regression are used to study the heart failure patients in hospitals. But these methods usually not consider the survival rate of patients.

Most of the medical studies use cox-regression model to assess the survival distribution of the Heart Failure patients, but parametric survival models like exponential, Weibull, log-normal, log-logistic (Accelerated Failure Time (AFT) models) can provide more description about survival data if one can identify the survival time because of having more realistic interpretation. It can provide more informative results than the Cox Proportional Hazard (Cox-PH) model. There are several risk factors for HF, such as age, hypertension, smoking, anemia, which can increase risk of mortality among HF patients. Parametric survival models play an important role in Bayesian survival analysis since many Bayesian analyses in practice are carried out using parametric AFT models and provide computational advantages by using the Markov Chain Monte Carlo (MCMC) method. Since MCMC methods have some limitations, like the burden of time in approximating the posterior and convergence problem, we use Integrated Nested Laplace Approximation (INLA) method of estimation. It provides fast and accurate approximation to the posterior marginal distributions of the parameters in the model. So, we chose the Bayesian parametric survival models using the INLA method to analyze HF data-set. Thus, considering the advantages of Bayesian application is the key for the motivation to apply it for the HF data-set in this project. Our objectives in this project are to answer-

- which factors significantly affect the survival time of HF patients,
- the estimated the survival time of HF patients,
- which parametric models are most appropriate for this data-set

So the main aim of this project to assess the survival time of heart failure patients using Bayesian approach and investigate the Bayesian accelerated failure time (AFT) models using the INLA method.

1.2 Significance of the project

After studying the survival time of HF patients we can overcome the problem of health in society by identifying the serious factors associated with death. We can improve our awareness on those factors which trigger the death of HF patients. It may helps policymakers to enhance awareness of society about those factors which increases the chance of death due to HF, which are easily protect-able if taken care in advance.

2 Methodology

2.1 Data Collection

The data is available at UCI Machine Learning Repository in the link `https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records` The data-set was elaborated by Davide Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International (CC BY 4.0) copyright in January 2020. For detailed information check reference [8].

2.2 Data Description

This data set contains the medical records of 299 patients who had heart failure, collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during their follow-up period April–December 2015, where each patient profile has 13 clinical features which report clinical, body, and lifestyle information. Some features anaemia, high blood pressure, diabetes, sex, and smoking are binary. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. The descriptive results of the study indicated that a total of 299 HF patients were included in our data. The minimum and maximum event time observed from HF patient's follow-up were 4 and 285 days respectively.

2.3 Variables in the data

This data-set contains the following 13 clinical features
- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (Boolean)
- high blood pressure: if the patient has hypertension (Boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (Boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction
- platelets: platelets in the blood (kilo platelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)

- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (Boolean, 0 or 1)
- time: follow-up period (days)
- [target variable] death event: if the patient deceased during the follow-up period (Boolean, 0 or 1)

2.4 Method of analysis

Survival data are censored in the sense that they did not provide complete information since subjects of the study may not have experienced the event of interest. Survival Analysis is a suitable method for heart failure data-set which are commonly used in medical research since studies in medical areas have a special feature that follow-up studies could start at a certain observation time and could end before all experimental units had experienced an event. Right censoring occurs to the right of the last known survival time and the observation of the patient is terminated before the event occurs. This type of censoring is commonly recognized in survival analysis.

2.5 Comparison of survival function

The Kaplan Meier plots and log-rank test were used for comparison of survival functions. To check the consistency of Laplace or lapse we used the Kaplan-Meirer (KM) estimates.

2.5.1 Kaplan-Meier estimator:

The Kaplan-Meirer estimator is a non-parametric statistic used to estimate the survival function from lifetime data. For a survival function $S(t)$, the estimate of KM (the probability that life time is longer than t)

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Where, t_i: is time when at least one event (death) happened,
d_i: The number of deaths at time t_i, and
n_i: the individuals known to have survived (have not yet had an event or been censored) up to time t_i. [12]

So, to compare the survival times between groups of features we used the Keplan-Meier(KM) plots. But we can not decide the survival times of HF patients for each individual co-variates based on KM plot.

2.5.2 Log rank test

For this purpose we have used log-rank test. Statistically we want to test,
H_0: There is no difference between the survival curves (*or* $S_{1t} = S_{2t}$). Vs
H_1: There is a significant difference between the survival curves (*or* $S_{1t} \neq S_{2t}$)

at a time t.

The log rank statistic is approximately distributed as a chi-square test statistic. There are several forms of the test statistic, and they vary in terms of how they are computed. We use the following:

$$\chi^2 = \sum_j \frac{\left(\sum_t O_{jt} - \sum_t E_{jt}\right)^2}{\sum_t E_{jt}} \tag{1}$$

where $\sum_t O_{jt}$ represents the sum of the observed number of events in the j^{th} group over time (e.g., $j = 1, 2$) and $\sum_t E_{jt}$ represents the sum of the expected number of events in the j^{th} group over time.[1]

The sums of the observed and expected numbers of events are computed for each event time and summed for each comparison group. The log rank statistic has degrees of freedom is equal to $k - 1$, where k represents the number of comparison groups.

2.6 Analysis of survival time

The Cox-PH model and the Bayesian AFT models were used to analyze the survival time of heart failure patients using R-software. The description of survival data utilizes non-parametric methods to compare the survival functions of two or more groups.

2.6.1 The Cox Proportional Hazard (Cox-PH) model

Cox proportional-hazards (Cox-PH) model is essentially a regression model commonly used in statistical medical research for investigating the association between the survival time of patients and one or more predictor variables. The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., death) at a particular point in time. This rate is commonly referred as the hazard rate.

As the name suggests, the model is expressed by the hazard function $h(t)$. Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + ... + b_px_p) \tag{2}$$

Where,
t denotes survival times,
$h_0(t)$ is called the baseline hazard. It corresponds to the value of the hazard if all the x_i are zero ($exp(0) = 1$).
$h(t)$ is the hazard function determined by a set of p covariates $(x_1, x_2, ..., x_p)$. The t in $h(t)$ reminds us that the hazard may vary over time.
$(b_1, b_2, ..., b_p)$ are the coeficients measure the impact (i.e., the effect size) of covariates.

$exp(b_i)$'s are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.([6])

If we take log of hazard function, it becomes linear with respect to the covariates $x_1, x_2, ..., x_p$.

$$log[h(t)] = log[h_0(t)] + \sum_{i=1}^{n} \beta_i x_i$$

So the analysis for this model is analogous to analysis of linear regression. We estimate the train values of h(t) by using,

$$h(t) = \frac{d_t}{n_t}$$

d_t:= number of events up-to time point t
n_t:= no of observations up-to time point t.
Now taking $log[h(t)]$ as response value we fit linear regression model with respect to covariates and take exponent of both sides to get cox PH model. The exponent of the intercept in the linear model turns out to be the baseline hazard $h_0(t)$. The error in estimation is given by,

$$\epsilon(t) = log[h(t)] - log[\hat{h}(t)]$$

2.6.2 Bayesian Survival analysis:

In Bayesian approach we assume that the observed data is fixed and those model parameters are random with have some prior distribution which helps to gain information from previous studies. So the Bayesian methods combine objective prior knowledge with the information acquired from the data by using the Bayes theorem. Since Bayesian approach has more better power of information than the frequentist approach in survival analysis, this approach is preferred. It assumes a prior distribution of the parameters, not only the likelihood of the data. Even if we have not a large enough sample size, it can provide us good estimation than the frequentist approach. So, Bayesian statistics allow for the incorporation of uncertainty about a parameter and the updating of this knowledge via the prior distribution. In a Bayesian set up we have to have the following to draw some inference about a parameter.

Prior: The prior probability, generally denoted by $\pi(\theta)$, express the uncertainty about the unknown parameters by assuming a probability distribution which have some prior information associated with our parameter of interest.

Likelihood: It is generally denoted by $L(\theta|data)$ which is nothing but a likelihood function of the unknown parameters in the presence of independent sample.

$$L(\theta|data) = \prod_{i=1}^{n} f(x_i; \theta)$$

But in this case of our Survival analysis, it is expressed as,

$$L(\theta|data) = \prod_{i=1}^{n} \left[f(t_i/x_i;\theta)^{\sigma_i} \times S(t_i/x_i;\theta)^{1-\sigma_i} \right]$$

Where σ_i is a censor indicator (0=censor, 1=death), $f(t_i/x_i;\theta)$ is probability density and $S(t_i/x_i;\theta)$ is the survival distribution.

Posterior: It is obtained by applying Bayes rule with combining the Likelihood $L(\theta|data)$ and prior $\pi(\theta)$. It is obtained by multiplying the prior distribution over all parameters, θ, by the full likelihood function. That is,

$$Posterior = \frac{Likelihod \times Prior}{\int Likelihod \times Prior \; d\theta}$$

So if we denote the posterior distribution as $\pi(\theta|X)$

$$\pi(\theta|X) = \frac{L(X|\theta) \times \pi(\theta)}{\int L(X|\theta) \times \pi(\theta) d\theta}$$

Since the denominator of the posterior expression does not depends on θ, it is only a function of data X, say $m(x) = L(X|\theta) \times \pi(\theta) d\theta$. We called it a marginal distribution of the data. So we can say, The prior is only proportional to the likelihood times the prior,

$$\pi(\theta|X) \propto L(X|\theta) \times \pi(\theta)$$

Parametric survival models like Exponential, Weibull, Log-Normal, and Log-Logistic etc., play an important role In Bayesian survival analysis since parametric modeling offers straightforward modeling and analysis techniques.

2.6.3 Accelerated Failure Time (AFT) models

In the statistical area of survival analysis, an AFT model is a parametric model that provides an alternative to the Cox-PH model. The AFT models have more realistic interpretation and it can provide more informative results than the Cox-PH model. In general, the AFT can be specified as-

$$\lambda(t|\theta) = \theta \lambda_0(\theta t)$$

where θ denotes the joint effect of covariates. typically $\theta = \exp(-[\beta_1 X_1 + \cdots + \beta_p X_p])$.[2]

The distribution including log-logistic, log-normal, exponential, Weibull, gamma and inverse Gaussian distributions are all suitable for AFT models. However, we have used four of them viz. exponential, log-normal, weibull and log-logistic distribution.

Exponential distribution: A random variable X is said to have an exponential distribution, if it has the probability density function (pdf)

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Here $\lambda > 0$ is the rate parameter parameter of the distribution. The distribution is supported on the interval $[0, \infty)$

Log-normal distribution: A log-normal distribution is continuous distribution of that random variable whose logarithm follow a normal distribution. X is said to follow log-normal(μ, σ^2) if the pdf the distribution is given by,

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$$

Where, parameters $\mu \in (-\infty, +\infty), \sigma > 0$ and support of $x \in [0, \infty)$

Weibull distribution: The probability density function of a Weibull random variable is-

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution. The distribution is supported on the interval $[0, \infty)$

Log-logistic distribution: The pdf of log-logistic distribution is given by,

$$f(x;\alpha,\beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2}$$

Where $\alpha > 0$ scale and $\beta > 0$ shape parameter. Support of $x \in [0, \infty)$.

2.7 Integrated Nested Laplace Approximation (INLA) method

With the use of the Gibbs sampler and other MCMC techniques, fitting of complex survival models become straightforward. But due to some limitations of MCMC like the burden of time in approximating the posterior and convergence problem if there is a complex model with too many parameters we have used Integrated Nested Laplace Approximation (INLA). So the Bayesian approach with the INLA method is focused on providing a good approximation to the posterior marginal distributions of the parameters in the model.

So estimate the parameters in the Bayesian parametric survival models, we used INLA method. INLA combines Gaussian approximations with numerical integration. This works well if most of the parameters are approximately normal and only a few are non-Gaussian and require numerical integration. It computes posterior marginal for each component in the model. Then we can find posterior mean and Standard deviation from that. The survival models can be expressed as a latent Gaussian model on which INLA can be applied. Also by clever using of Laplace approximation and advanced numerical methods, INLA gives extremely fast and very accurate result ([4]). So it is reasonable to fit a survival model using INLA.

2.8 Bayesian model selection criterion

To select the best effective model in our analysis, we may use Akaike information criteria(AIC) or Bayesian information criteria (BIC). Where, with a number of covariates p, and $\hat{\theta}$ as an MLE of θ

$$AIC = -2ln[L(X|\hat{\theta})] + 2p \tag{3}$$

and,

$$BIC = -2ln[L(X|\hat{\theta})] + ln(n)p \tag{4}$$

In general the model with smaller AIC or BIC value is preferred.

However, these has some limitations, e.g., in case of AIC, if a model is too complex it will may have small deviance but it can over-fit our data, and in case of BIC, it is only a asymptotic approximation with large n. To overcome this issue in Bayesian analog here we use Deviance information criteria (DIC) and Watanabe-Akaike information criteria (WAIC).

DIC: If $\bar{D}$ be the posterior mean of the deviance, $D(X|\theta) = -2ln[L(X|\theta)]$, and $\hat{\theta}$ is the posterior mean of θ, the effective number of parameters is

$$p_{_D} = \bar{D} - D(X|\hat{\theta}) \tag{5}$$

Then DIC can be written as,

$$DIC = \bar{D} + p_{_D} = D(X|\hat{\theta}) + 2p_{_D} \tag{6}$$

If there are strong prior, $p_{_D} << p$, intuitively, for an uninformative prior $p_{_D} \approx p$.

WAIC: It is an alternative method to DIC. WAIC is written in terms of the posterior of the likelihood rather than parameters. If m_i and v_i be the posterior mean and variance of $[f(X_i|\theta)]$, then WAIC is defined by,

$$WAIC = -2\sum_{i=1}^{n} m_i + 2p_{_W} \tag{7}$$

Where, $p_{_W}$ denote the effective model size defined by, $p_{_W} = \sum_{i=1}^{n} v_i$.

DIC and the alternative WAIC are commonly used in Bayesian parametric survival model comparison. We choose that model as the best model which have **lowest DIC or WAIC** value.

2.9 Statistical Software

To summarize the data, analyse the data and interpret all the results we have used R software.

3 Results and Analysis

3.1 Descriptive Summaries

From the table 6 in the appendix, we get some brief summaries about the data. The data for this study was done with 299 patients who received treatments for HF at least once in that period of time. Among those HF patients, about 67.89% were censored (right censored), and the remaining 32.11% have died 60% of HF patients who received treatment survived 8 months or above.

In our data there were 64.88% male and remaining 35.12% female who were treated and the survival time of both male and female patients are similar.

Observing the age of the patients, we had divided our data into 4 different age group. The survival time of HF patients decreases as they get older. 39.58% of patients who are older than 70 years could not survived. Among the remaining the survival time was unexpectedly little bit higher (12.50% died) for the age group 51 *to* 60 than the younger age group of below 50 (19.79% died).

By observing the smoking status of HF patients, most of the HF patients 67.49% non-smokers and unexpectedly, the death proportion seems highest for those HF patients who were non-smokers, which was 68.75% compared to smokers which were only 31.25%.

Moreover, about $6.02\%, 32.44\%, 19.73\%$ and 41.81% of HF patients have low to high Creatinine Phosphokinase level respectively. In addition, the survival time of HF patients who have low to high Creatinine Phosphokinase level were $5.21\%, 34.52\%, 19.79\%$ and 39.58% respectively.

About 41.67% of HF patients have no diabetes, and 58.33% have diabetes. Observing the Ejection Fraction of HF patients, about 26.76% and 19.73% were HF patients with high Ejection Fraction (≥ 45) and low Ejection Fraction ≤ 30 respectively, in which HF patients with low Ejection Fraction seem to have lower survival time as the percentage of deaths is much higher (39.58%) than the other intervals ($20.83\%, 19.79\%$ & 19.79%).

Most of the HF patients have high blood pressure, 64.88%, and the remaining 35.12% have normal blood pressure.

About 43.14% of the HF patients were suffering with anemia and the rests 56.86% had no anemia and they have almost same survival time of having or not having anemia.

Regarding the platelets count, the patients with very lower platelets ($<$ 212500) or very higher platelets ($\geq$ 303500) had comparatively less survival time (25% and 32.29% died respectively) than the other two groups (18.75% and 23.96% died) although all the groups were of almost same number of patients.

The patients having high Serum Creatinine (≥ 1.400) had less survival time. About half (48.96%) of the patients with high Serum Creatinine had died.

The patients who have serum sodium within 134 to 137 had died more than the other serum sodium quantity groups.

3.2 The Kaplan-Meier Estimate for the Covariates

Below we plotted the Kaplan–Meier (KM plot) estimates of the survival curves of HF data set for the overall survival function without considering any covariates. The figure 1 below shows the overall survival rate and at the end of 8 months was about 65%.

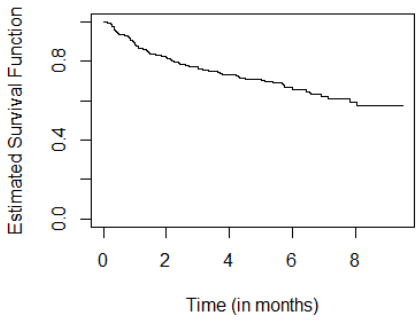
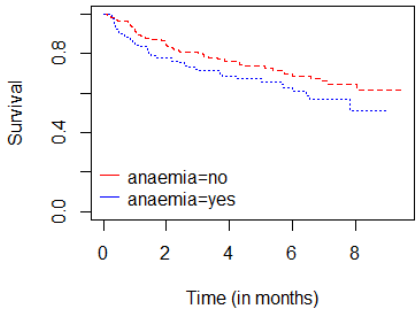


Figure 1: Overall KM plot

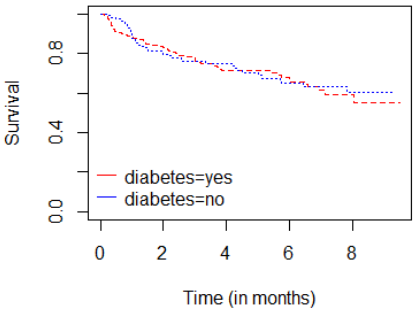
3.2.1 KM plot for Categorical Variables

In the next figure we plotted all the KM plots with respect to all the categorical covariates like anaemia, diabetes, high blood pressure, sex, smoking respectively. By observing the plots we see that, the patients with anemia had less survival time. But, diabetes did not play any major role to the survival time since the curves in the KM plot almost overlaps to each other.

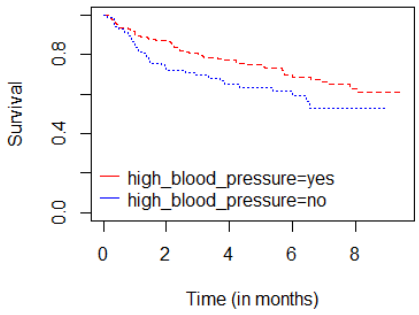
The KM plot of High blood pressure shows that the patients with high blood pressure has more survival time than the other patients. But here again sex and unexpectedly smoking does not play any important role to the survival time as the curves in the KM plot almost overlaps to each other. However who were smoker, they have slightly more survival time after heart failure during the censoring period.



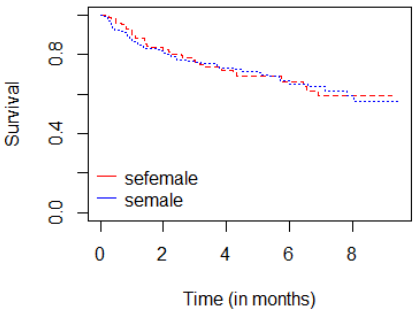
(a) Anaemia



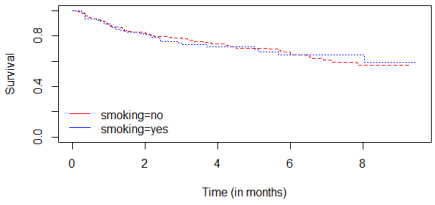
(b) Diabetes



(c) High Blood Pressure



(d) Sex



(e) Smoking

Figure 2: Kaplan-Meier Plots for Categorical Covariates

3.2.2 KM plot for Continuous Variables

In the figure 3 we plotted all the KM plots with respect to all the continuous variables like age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine and Serum Sodium respectively. Clearly, in figure 3

1. the KM plot (a) of Age indicated that HF patients whose age was below 50 years were unexpectedly at a lower probability of surviving than patients whose age was 51 to 60 years. The probability of surviving becomes very less for patients whose age was greater than or equal to 70 years.

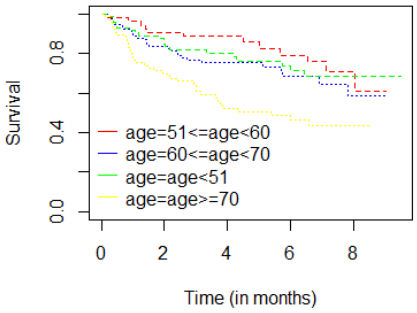
2. From plot (b) it is not that much clear how the different groups affect the survival time as they almost overlap to each other. We need further tests.

3. Plot (c) in figure 3 shows that HF patients with lower Ejection Fraction had very lower chance of survival. The other survival curves of Ejection Fraction more than 30 for the patients were very similar.

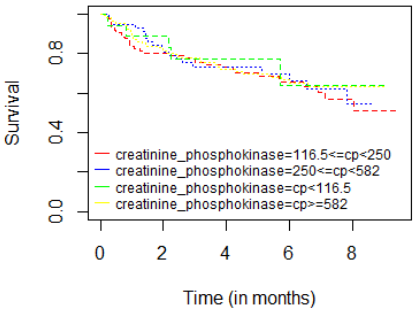
4. In plod (d) it shows almost similar survival time for any counts of platelets as the survival curves overlap. However larger platelet count indicates little bit lower survival time.

5. In Plot (e) the HF patients with high Serum Creatinine (≥ 1.400) had very lower chance of survival and with low value of Serum Creatinine (≥ 1.400) had very dramatically high chance of survival. The other two survival curves of Serum Creatinine were similar.

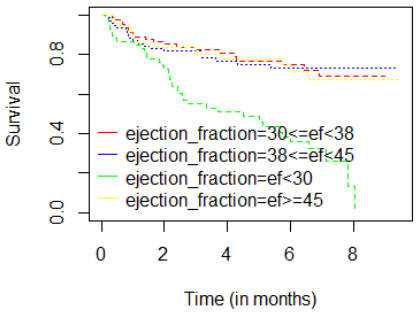
6. Plot (f) of figure 3 shows that HF patients with lower Serum Sodium (≤ 134) and mid-lower (134 *to* 137) had slightly lower chance of survival as compare to others curves.



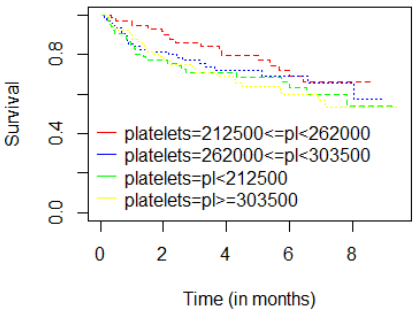
(a) Age



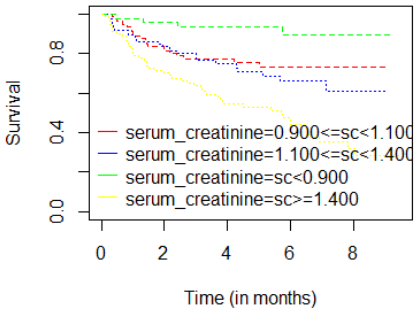
(b) Creatinine Phosphokinase



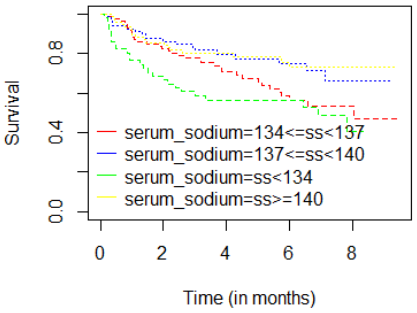
(c) Ejection Fraction



(d) Platelets



(e) Serum Creatinine



(f) Serum Sodium

Figure 3: Kaplan-Meier Plots for Continuos Covariates (group wise

3.3 Results of log-rank test

Visually we can guess which of these factors are significant for survival time by looking at to what extent the curves for different categories of a particular covariate overlaps. Statistically we can have this idea using log rank test as mentioned earlier. For our HF dataset following table indicates results for log-rank test for different covariates and as the table indiactes HBP, age, Ejection fraction, Serum creatinine and serum sodium are the significant factors that effect survival time of HF patients.

Covariate	χ^2 **values**	**p-value**
Anaemia	2.7	0.1
Diabetes	0	0.8
High Blood Pressure (HBP)	4.4	$.04^*$
Sex	0	0.9
Smoking	0	1
Age	17.2	$6 \times 10^{-4*}$
Creatinine Phosphokinase	0.4	0.9
Ejection Fraction	35.6	$9 \times 10^{-8*}$
Platelets	3.4	0.3
Serum Creatinine	34.5	$2 \times 10^{-7*}$
Serum Sodium	12.7	0.005^*

* indicates statistically significant

Table 1: Table showing χ^2 values and p-values of log-rank test considering different factors

3.4 Cox-pH Model

Now we have fitted Cox Proportional Hazard model on our data and the table summarises the findings of the test. From this model we find that Anaemia, HBP, Age, Creatinine phosphokinase, Ejection fraction, Serum creatinine are the major factors of the hazard function.

Covariate	Coefficient	exp(Coefficient)	Sd	p-value
Anaemia	0.46	1.584	0.217	0.0338*
Diabetes	0.14	1.150	0.223	0.5307
High Blood Pressure (HBP)	0.48	1.609	0.216	0.0278*
Sex	-0.24	0.789	0.252	0.3452
Smoking	0.13	1.138	0.251	0.6078
Age	0.05	1.048	0.009	6.45×10^{-7}*
Creatinine Phosphokinase	2.2×10^{-4}	1	9.9×10^{-5}	0.0260*
Ejection Fraction	0.05	0.952	0.010	2.98×10^{-6}*
Platelets	-4.6×10^{-7}	1	1.13×10^{-6}	0.6806
Serum Creatinine	0.32	1.379	0.07	4.8×10^{-6}*
Serum Sodium	0.04	0.957	0.023	0.0575

* indicates statistically significant

Table 2: Cox-PH Table

3.5 Checking the Assumptions of Cox-PH

3.5.1 Linearity Assumption

We have plotted fitted residuals in X-axis against Martingale residual in Y-axis and obtained the residual plot. Upon approximating the residuals with a line (black line in the plot) it coincides with Y=0 (red line in the plot) which indicates that linearity assumption is satisfied for our dataset.

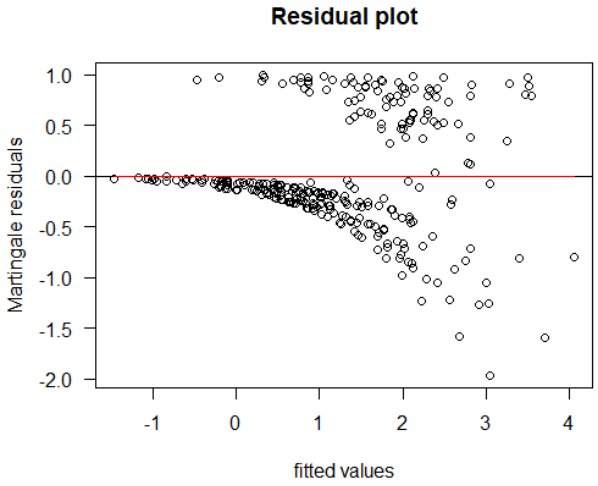


Figure 4: Martingale Residual Plot

3.5.2 Proportionality Assumption

According to the following table, using the Schoenfeld's test, the p-value for ejection fraction is less than the 0.05 i.e., at 5% level of significance, we reject the null hypothesis of presence of proportionality. Which indicates that the Cox-PH model for our dataset was not valid.

Table 3: Test of Assumption in Cox-PH Model

Covariates	**Chi-square**	**P-value**
age	0.103	0.75
anaemia	0.0169	0.90
creatinine phosphokinase	1.02	0.31
diabetes	0.192	0.66
ejection fraction	**4.69**	**0.03**
high blood pressure	0.00823	0.93
platelets	0.00001	1.00
serum creatinine	1.52	0.22
sex	0.0763	0.78
smoking	0.479	0.49
serum_sodium	0.11	0.74
GLOBAL	11.02527	0.39

3.6 Bayesian Survival Analysis

Table 4: The Comparisons of Bayesian AFT models Using INLA Method

Distributions	**Pd**	**DIC**	**WAIC**
Exponential	22.63	650.1391	657.7969
Log-Normal	25.04	654.7185	655.5272
Weibull	23.55	650.9338	657.3042
Log-logistic	**23.66**	**648.5148**	**653.1089**

As it can be shown in Table 3, the assumption of the Cox-PH model was not fulfilled for our HF data set; for this reason, parametric AFT models were used for the HF data set. For the HF data set, the time t_i where $i = 1, 2, \ldots, 299$ of HF patients. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_P)'$ is the vector of coefficients of the covariates considered for analysis, where β_0 is the intercept and p the number of covariates (*here* $p = 11$), we assume that all these coefficients have a **normal prior** with mean 0 and variance of 1000. We also assume that a gamma prior was applied to the Weibull, Log-normal, and Log-logistic distributions with shape parameter 1 and inverse scale parameter 0.001. In table 4, we can see the analysis of the HF data set for model comparison using INLA method. To compare the efficiency of these various models, DIC and WAIC were used, and

the one with the **smallest value** and the best fit is considered. Accordingly, the Bayesian log-logistic AFT model ($DIC = 647.5; WAIC = 650.9$) was found to be the best model for survival time of HF patients data-set from the given alternative because the bold values are smallest DIC/WAIC values.

Table 5: Indicating the results for Bayesian log-logistic AFT model using INLA method

Covariates	**Categories**	**Pmean**	**Sd**	**Median**	**Credible Interval**	**Mode**	**Kld**
	intercept	5.043	0.796	5.030	$[3.518, 6.641]^*$	5.005	0
Age	51-60	ref					
	< 51	-0.867	0.494	-0.862	$[-1.850, 0.088]$	-0.853	0
	60-70	-0.570	0.452	-0.565	$[-1.473, 0.304]$	-0.555	0
	> 70	-1.696	0.456	-1.689	$[-2.614, -0.822]^*$	-1.673	0
Creatinine Phosphokinase	116.5-250	ref					
	< 116.5	0.572	0.619	0.554	$[-0.591, 1.842]$	0.516	0
	250-582	0.392	0.387	0.388	$[-.358, 1.162]$	0.38	0
	> 582	0.333	0.330	0.332	$[-0.313, 0.981]$	0.330	0
Ejection Fraction	30-38	ref					
	< 30	-1.568	0.388	-1.564	$[-2.343, -0.817]^*$	-1.556	0
	38-45	-0.337	0.411	-0.337	$[-1.144, 0.468]$	-0.336	0
	> 45	-0.119	0.408	-0.119	$[-0.919, .682]$	-0.120	0
Platelets	212500-262000	ref					
	< 212500	-0.620	0.421	-0.618	$[-1.453, 0.201]$	-0.614	0
	262000-303500	-0.570	0.408	-0.568	$[-1.376, .227]$	-0.565	0
	> 303500	-0.796	0.394	-0.793	$[-1.578, -0.032]^*$	-0.787	0
Serum Creatinine	0.9-1.1	ref					
	< 0.9	1.266	0.621	1.238	$[0.124, 2.566]^*$	1.181	0
	1.1-1.4	-0.181	0.391	-0.179	$[-0.954, 0.581]$	-0.176	0
	> 1.4	-0.851	0.381	-0.848	$[-1.611, -0.113]^*$	-0.840	0
Serum Sodium	134-137	ref					
	< 134	-0.541	0.384	-0.542	$[-1.293, 0.214]$	-0.544	0
	137-140	0.284	0.4	0.282	$[-0.492, 1.075]$	0.276	0
	> 140	0.275	0.403	0.273	$[-0.509, 1.072]$	0.268	0
Anaemia	no	ref					
	yes	-0.720	0.301	-0.719	$[-1.312, -0.132]^*$	-0.717	0
Diabetes	yes	ref					
	no	-0.178	0.299	-0.178	$[-0.763, 0.409]$	-0.179	0
High Blood Pressure	yes	ref					
	no	-0.506	0.293	-0.507	$[-1.079, 0.071]$	-0.508	0
Smoking	no	ref					
	yes	-0.097	0.326	-0.098	$[-0.735, 0.545]$	-0.100	0
Sex	female	ref					
	male	0.060	0.328	0.060	$[-0.584, 0.703]$	0.060	0
Model Hyperparameters	For log-logistic	1.10	0.087	1.10	$[0.938, 1.28]^*$	1.10	-

* indicates statistically significant

We have shown the final results for the Bayesian log-logistic AFT model, and as this result shows, the survival time of HF patients is statistically significantly affected by age, ejection fraction, serum creatinines, anaemia.
From table 5, the final model was interpreted using acceleration factor, 95% credible interval of Bayesian accelerated failure time estimated values. The estimated acceleration factor is defined as $\gamma = exp\left(\hat{\beta}\right) = exp\,(\text{posterior mean})$

1. Under Bayesian log-logistic AFT model, keeping the effect of other factors constant, the estimated acceleration factor for the age group less than 51, from 60 to 70 and above 70 are estimated to be 0.42 [with CI $(0.157, 1.09)$], 0.565 [with CI $(0.23, 1.36)$] and 0.183 [with CI $(0.07, 0.44)$] respectively. So the expected survival time of HF patients decrease by $58\%, 43.5\%$ and 81.7% for these age groups respectively with respect to HF patients in the age group 51 to 60 (reference).
   The 95% CI of the accelerated factor for the age groups of below 51 and more than 70 do not include 1 implying that both of them have significant effect on patient's survival time.

2. Keeping the effect of other factors constant, the estimated acceleration factor for the creatinine phosphokinase groups less than 116.5, from 250 to 582 and above 582 are estimated to be 1.77 [with CI $(0.554, 6.31)$], 1.48 [with CI $(0.699, 3.196)$] and 1.395 [with CI $(0.73, 2.67)$] respectively. So the expected survival time of HF patients increase by $77\%, 48\%$ and 39.5% for these CP groups respectively with respect to HF patients in the CP group 116.5 to 250 (reference).
   The $95\%CI$ of the accelerated factor for all the CP groups include 1 implying that Creatinine phosphokinase should not be adjudged a significant factor here.

3. Keeping the effect of other factors constant, the estimated acceleration factor for the ejection fraction groups less than 30, from 38 to 45 and above 45 are estimated to be 0.21 [with CI $(0.096, 0.44)$], 0.714 [with CI $(0.318, 1.6)$] and 0.888 [with CI $(0.4, 1.978)$] respectively. So the expected survival time of HF patients decrease by $79\%, 28.6\%$ and 11.2% for these EF groups respectively with respect to HF patients in the EF group 30 to 38 (reference).
   Now the $95\%CI$ of acceleration factor for the EF group less than 30 does not include 1 hence it has significant effect on patient's survival time but for the rest 2 categories they are insignificant as their CI include 1 in them.

4. Keeping the effect of other factors constant, the estimated acceleration factor for the Platelets groups less than 212500, from 262000 to 303500 and above 303500 are estimated to be 0.538 [with CI $(0.234, 1.223)$], 0.565 [with CI $(0.252, 1.255)$] and 0.451 [with CI $(0.206, 0.97)$] respectively. So the expected survival time of HF patients decrease by $46.2\%, 43.5\%$ and 54.9% for these platelets groups respectively with respect to HF patients in the platelets group 212500 to 262000 (reference).

The 95%CI of acceleration factor for the platelets group greater than 303500 does not include 1 hence it has significant effect on patient's survival time but for the rest 2 categories they are insignificant as their CI include 1 in them.

5. Keeping the effect of other factors constant, the estimated acceleration factor for the serum creatinine groups less than 0.9, from 1.1 to 1.4 and above 1.4 are estimated to be 3.55 [with CI $(1.132, 13.01)$], 0.834 [with CI $(0.385, 1.788)$] and 0.427 [with CI $(0.2, 0.893)$] respectively. So the expected survival time of HF patients decrease by $16.6\%, 57.3\%$ for the last 2 SC groups respectively and increase by 250% for the first SC group with respect to HF patients in the SC group 0.9 to 1.1 (reference).
The 95%CI of acceleration factor for the SC group greater than 1.4 and less than 0.9 do not include 1 hence it has significant effect on patient's survival time but for the other category it is insignificant as its CI include 1 in them.

6. Keeping the effect of other factors constant, the estimated acceleration factor for the serum sodium groups less than 134, from 137 to 140 and above 140 are estimated to be 0.582 [with CI $(0.274, 1.24)$], 1.328 [with CI $(0.611, 2.93)$] and 1.32 [with CI $(0.601, 2.92)$] respectively. So the expected survival time of HF patients increase by $32.8\%, 32\%$ for the last 2 SS groups respectively and decrease by 41.8% for the first SS group with respect to HF patients in the SS group 134 to 137 (reference).
Since for all the categories of Serum Sodium covariate, the acceleration factor includes 1 in them none of the categories have significant effect on the HF patient's survival time.

7. Keeping all the other factor constant, the estimated acceleration factor for the anaemia positive category is 0.487 [with CI $(0.269, 1.141)$]. Therefore the expected survival time for the anaemia positive HF patients is 51.3% less than anaemia negative HF patients(reference). Since 95% CI of acceleration factor for anaemia positive category does not include 1, the covariate anaemia is a significant factor in survival time of HF patients.

8. Keeping all the other factor constant, the estimated acceleration factor for the patients without diabetes category is 0.837 [with CI $(0.466, 1.505)$]. Therefore the expected survival time for the HF patients without diabetes is 16.3% less than HF patients with diabetes(reference). Since 95% CI of acceleration factor for category of the patients without diabetes includes 1, the covariate Diabetes is not a significant factor in survival time of HF patients.

9. Keeping all the other factor constant, the estimated acceleration factor for the patients without High blood pressure category is 0.603 [with CI $(0.34, 1.073)$]. Therefore the expected survival time for the HF patients without high blood pressure is 39.7% less than HF patients with high blood pressure(reference). Since 95% CI of acceleration factor for

category of the patients without high blood pressure includes 1, the covariate high blood pressure is not a significant factor in survival time of HF patients.

10. Keeping all the other factor constant, the estimated acceleration factor for smoker category is 0.91 [with CI $(0.48, 1.725)$]. Therefore the expected survival time for the HF patients who smoke is 9% less than HF patients who do not smoke(reference). Since 95% CI of acceleration factor for smoker category of the patients includes 1, the covariate smoking is not a significant factor in survival time of HF patients. Keeping all the other factor constant, the estimated acceleration factor for the male patients category is 1.062 [with CI $(0.558, 2.02)$]. Therefore the expected survival time for the male HF patients is 6.2% more than female HF patients (reference). Since 95% CI of acceleration factor for male patients includes 1, the covariate sex is not a significant factor in survival time of HF patients.

11. From Table 5, the Kullback–Leibler divergence values for all signifcant parameters in the Bayesian log-logistic AFT model were 0, and thus, small values indicate that the posterior distribution was well approximated by a normal distribution.

4 Discussion

From our analysis we saw that, 60% percent of HF patients who received treatment survived more than 8 months. In this study, among those HF patients, about 67.89% were censored (right censored) and the remaining 32.11% died.

We plotted KM curves for different continuous and discrete covariates to have an insight about significant factors visually.Then we performed log rank test to find the statistically significant factors effecting survival function. We also built cox pH model to find out main factors that has significant effect on hazard function. But proportionality assumption was violated for the data we have got.

So the Bayesian parametric survival models were applied to this data set. Then we compared the efficiency of different AFT models using DIC and WAIC. It is found that the Bayesian log-logistic AFT model was the best model to describe the HF data set from the other given alternative. However, the results of this Bayesian log-logistic AFT model using the INLA method shows that the factors **age, ejection fraction, platelets, serum creatinine, anemia have a significant effect** on the survival time of HF patients.

The kullback–leibler divergence (kld) measures the accuracy of the INLA approximation. In this study, the values of kld for all significant parameters in the Bayesian log-logistic AFT model were 0. This indicates that the Bayesian log-logistic AFT model was faster and has higher accuracy.

5 Conclusion

The objective of this project was to find the factors significantly affecting the survival time of heart failure patient. By using different methods throughout the project, we can draw the following conclusions-

1. Firstly, we used KM plots and the log-rank test to identify the significant factors, and they turned out to be: **High Blood Pressure, Age, Ejection Fraction, Serum Creatinine, Serum Sodium.**

2. But this univariate approach doesn't take into account the other factors except its own. So we have introduced the cox-PH model. The significant variables we got from Cox-PH models are: **Anaemia, High Blood Pressure, Age, Ejection Fraction, Serum Creatinine, Creatinine Phosphokinase.**

3. As the proportionality assumption of the cox-PH model was violated, we introduced the **Bayesian log-logistic AFT** model to measure the significant factors as it performed better than various parametric models with baseline distribution (Exponential, Weibull, and Log-normal). Then we found that the significant variables are: **Age, Ejection Fraction, Platelets, Serum Creatinine, Anaemia.**

So as a measure of precaution, the Government and other health organizations can raise awareness about these significant factors among common people and can suggest to follow a better lifestyle and some precautions to reduce the risk of heart failure also the number of heart failure events that are taking place.

6 Appendix

Table 6: Descriptive Summaries of Patient's for Heart Failure data

Covariates	**Categories**	**Number of Censored (%)**	**Number of Deaths (%)**	**Total(%)**
Age	< 51	55 (27.09)	19 (19.79)	74 (24.75)
	[51,60)	43 (21.18)	12 (12.50)	55 (18.39)
	[60,70)	66 (32.51)	27 (28.12)	93 (31.10)
	≥ 70	39 (19.21)	38 (39.58)	77 (25.75)
Anaemia	no	120 (59.11)	50 (52.08)	170 (56.86)
	yes	83 (40.89)	46 (47.92)	129 (43.14)
Creatinine Phosphokinase	< 116.5	13 (6.40)	5 (5.21)	18 (6.02)
	[116.5,250)	63 (31.03)	34 (34.52)	97 (32.44)
	[250,582)	40 (19.70)	19 (19.79)	59 (19.73)
	≥ 582	87 (42.86)	38 (39.58)	125 (41.81)
Diabetes	no	85 (41.87)	40 (41.67)	125 (41.81)
	yes	118 (58.13)	56 (58.33)	174 (58.19)
Ejection Fraction	< 30	21 (10.34)	38 (39.58)	59 (19.73)
	[30,38)	63 (31.03)	20 (20.83)	83 (22.76)
	[38,45)	58 (28.57)	19 (19.79)	77 (25.75)
	≥ 45	61 (30.05)	19 (19.79)	80 (26.76)
High Blood Pressure	no	66 (32.51)	39 (40.62)	105 (35.12)
	yes	137 (67.49)	57 (59.38)	194 (64.88)
Platelets	< 212500	42 (20.69)	24 (25.00)	66 (22.07)
	[212500,262000)	56 (27.59)	18 (18.75)	74 (24.75)
	[262000,303500)	53 (26.11)	23 (23.96)	76 (25.42)
	≥ 303500	52 (25.62)	31 (32.29)	83 (27.76)
Serum Creatinine	< 0.900	45 (22.17)	4 (4.17)	49 (16.39)
	[0.900,1.100)	62 (30.54)	20 (20.83)	82 (27.42)
	[1.100,1.400)	62 (30.54)	25 (26.04)	87 (29.10)
	≥ 1.400	34 (16.75)	47 (48.96)	81 (27.09)
Serum Sodium	< 134	26 (12.81)	25 (26.04)	51 (17.06)
	[134,137)	54 (26.60)	34 (35.42)	88 (29.43)
	[137,140)	64 (31.33)	19 (19.79)	83 (27.76)
	≥ 140	59 (29.06)	18 (18.75)	77 (25.75)
Sex	female	71 (34.98)	34 (35.42)	105 (35.12)
	male	132 (65.02)	62 (64.58)	194 (64.88)
Smoking	no	137 (67.49)	66 (68.75)	203 (67.89)
	yes	66 (32.51)	30 (31.25)	96 (32.11)

Table 7: Counts for Patient Status

Patient Status	**Number of Patients (%)**
Censoring	203 (67.89)
Death	96 (32.11)

7 Bibliography

Acknowledgement:

Real learning comes from a practical work. We are very grateful to our instructor Dr. Arnab Hazra of MTH 535A: Bayesian Analysis for providing us such an opportunity to be engaged in a practical project work based on our area of interests under the context of Bayesian Analysis. We are sincerely thankful to our instructor and project guide Dr. Arnab Hazra Sir for his continuous guidance, monitoring and constant support throughout the course and project.

Authors' contribution:

1. Project selection, theory of the project, methodologies - Sayan Bhowmik (201409) & Soumyadip Sarkar(201431)
2. Implementation of results, plots, tables by coding - Abir Naha(201257) & Koyel Pramanick(201333)
3. Project supervision - Dr. Arnab Hazra

References

[1] https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival5.html#:~:text=The%20log%20rank%20test%20is,identical%20(overlapping)%20or%20not..

[2] https://en.wikipedia.org/wiki/Accelerated_failure_time_model#cite_note-1.

[3] Saheed O Adebayo et al. "Heart failure: Definition, classification, and pathophysiology–A mini-review". In: *Nigerian Journal of Cardiology* 14.1 (2017), p. 9.

[4] Rupali Akerkar, Sara Martino, and Håvard Rue. "Implementing approximate Bayesian inference for survival analysis using integrated nested Laplace approximations". In: *Preprint Statistics, Norwegian University of Science and Technology* 1 (2010), pp. 1–38.

[5] Tafese Ashine, Geremew Muleta, and Kenenisa Tadesse. "Assessing survival time of heart failure patients: using Bayesian approach". In: *Journal of Big Data* 8.1 (2021), pp. 1–18.

[6] Mike J Bradburn et al. "Survival analysis part II: multivariate data analysis–an introduction to concepts and methods". In: *British journal of cancer* 89.3 (2003), pp. 431–436.

[7] Vivek Chaturvedi et al. "Heart failure in India: the INDUS (INDia ukieri study) study". In: *Journal of the Practice of Cardiovascular Sciences* 2.1 (2016), pp. 28–35.

[8] Davide Chicco and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". In: *BMC medical informatics and decision making* 20.1 (2020), pp. 1–16.

[9] Taane G Clark et al. "Survival analysis part I: basic concepts and first analyses". In: *British journal of cancer* 89.2 (2003), pp. 232–238.

[10] Hisham Dokainish et al. "Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study". In: *The Lancet Global Health* 5.7 (2017), e665–e672.

[11] Azmera Hailay, Essey Kebede, and Kasim Mohammed. "Survival during treatment period of patients with severe heart failure admitted to intensive care unit (ICU) at gondar university hospital (GUH), gondar, Ethiopia". In: *Am J Health Res* 3.5 (2015), pp. 257–269.

[12] Stanford. *Kaplan-Meier (KM) Estimator*. STAT331: Unit 3.

[5] [11] [4] [9]