Heart Failure Analysis
using Logistic Regression and Survival Analysis

Author	Raven Chen, Wenxu Liu, Ziming Wei
Group	15
Date	December 13, 2021
Subject	Applied Regression Analysis

Contents

Abstract

In this study focusing on leveraging anthropometric and individual lifestyle data in heart failure prediction, we analyzed electronic health records (EHR) data of 299 patients who were admitted to the Institute of Cardiology and Allied hospital Faisalabad, Pakistan from April to December 2015, had left ventricular systolic dysfunction, and belonged to NYHA class III and IV [1]. Patients included in the study averaged at 60 years old, and 32% of patients died by the end of the 9-month follow-up period. Traditional biostatistics methods, logistic regression, and Cox regression, were utilized to model mortality prediction and identify key risk factors without or with the influence of time. Linear regression model with flexible modeling techniques was adopted to identify key features contributing to EF. Our logistic model showed that the increase in the level of serum creatinine, which indicates an unhealthier kidney, was not only significant in predicting heart failure but also the most important predictor, the survival prediction model suggested the increase in the age and serum creatinine has a harmful effect on heart failure, while increased ejection fraction has a protective effect on heart failure.

Introduction

Cardiovascular disease is still the top one leading cause of death worldwide and in the United States. Heart failure is one kind of cardiovascular disease in which the heart cannot pump or fill enough blood for the body's needs. In 2018, heart failure accounted for 13.4% of all death certificates in the United States, and about 6.2 million adults experienced heart failure during 2013-2016 [3]. Heart failure is categorized into two types based on ejection fraction value, which is the percentage of blood pumped out of the heart during single contraction. One category of heart failure is caused by reduced ejection fraction (HFrEF), or defined as systolic heart failure, with an ejection fraction smaller than 40%; the other is heart failure with preserved ejection fraction (HFpEF), or diastolic heart failure, with a normal ejection fraction [2]. Critical evaluations for risk factors associated with heart failure is of clinical importance, and identifying risk factors most predictive for such events and forecasting heart-failure related events has become priorities for both medical professionals and for public health surveillance. In this study, we are investigating possible anthropometric and lifestyle variables that help predict heart failure, and the survival outcome of these patients.

Literature Review

There have been several published papers and analyses on heart failure using this particular dataset. Ahmad and colleagues, who collected the data set originally, analyzed the survival of the study participants using Cox regression model and Kaplan-Meier survival curve, and found that age, EF, creatinine level, sodium level, anemia, and BP are significant features affecting mortality likelihood [1]. Zahid et al. further analyzed the survival stratifying on gender, however did not find significant difference in survival likelihood between female and male [4]. Chicco and Jurman leveraged different machine learning methods, including survival prediction classifiers (linear regression, tree-based methods such as random forest, SVM, KNN, Naïve Bayes) and feature ranking (Pearson correlation test, chi-square test, random forest), to identify the most important features for high prediction accuracy of the survival model [2]. Interestingly, the authors found that two variables alone, ejection fraction and serum creatinine, were sufficient for the model to reach the highest prediction performance. Risk factors identified by previous studies on the prediction of survival likelihood differ. In this study, we aim to explore the risk factors most predictive for the death event of patients in the dataset with or without the influence of time (whether the aim was survival or not) by leveraging logistic and the Cox model under GLM family.

Methods

The analysis is carried out in two parts, as our outcome variables can be utilized in two ways. Our outcome variables of interest are Death, a binary variable describing whether the patient died within the follow-up period, and time, which is the follow-up time for each patient. First, we ignore the variable time, and try to predict whether the event Death occurs using all other variables in Logistic Regression. Second, we focus on the variable time and perform a survival analysis.

Data Characteristics

Exploratory Data Analysis

The outcomes of interest of this study were death event (DEATH) and follow-up time. Of the explanatory variables, 6 were continuous variables: Age, Serum Creatinine, Serum Sodium, Creatinine Phosphokinase (CPK), Ejection Fraction (EF), Platelets, and 5 were categorical variables: Gender, Smoking, Diabetes, High Blood Pressure (HBP), Anemia. The follow-up time ranged from 4 days to 285 days. Patients included in the study averaged at 60 years old, 65% women, had a mean heart ejection fraction (EF) of 38%, and 32% of patients died by the end of the 9-month follow-up period *(Table 1)*.

Of the numerical explanatory variables, most are approximately normally distributed and meet the assumption of various regression models. The one exception is CPK, which has a distribution that is remarkably right skewed with many large outliers. Therefore, we will use the log-transformed values, log(CPK), as the parameter in the subsequent analysis.

Table 1. Data explanation and characteristics

Variable Name	Full Sample (n=299)	Meaning	Range
Age, mean (SD)	60.83 (11.90)	Patient age	[40,…,95]
Serum Creatinine, mean (SD)	1.39 (1.03)	Level of creatinine in the blood (mg/dL)	[0.50,..., 9.40]
Serum Sodium, mean (SD)	136.62 (4.41)	Level of sodium in the blood (mEq/L)	[114,..., 148]
Creatinine Phosphokinase (CPK), mean (SD)	581.84 (970.29)	Level of CPK enzyme in the blood (mcg/L)	[23,..., 7861]
Ejection Fraction (EF), mean (SD)	38.08 (11.83)	Percentage of blood leaving the heart per contraction (%)	[14,..., 80]
Platelets, mean (SD)	263358 (97804.24)	Platelets in the blood (kiloplatelets/mL)	[25.01,..., 850.00]
Gender (1), n (%)	194 (64.88%)	Patient gender (female=0)	0, 1
Smoking (1), n (%)	96 (32.11%)	If the patient smokes	0, 1
Diabetes (1), n (%)	125 (41.81%)	if the patient has diabete	0, 1
High Blood Pressure (HBP) (1), n (%)	105 (35.12%)	if the patient has hypertension	0, 1
Anemia (1), n (%)	129 (43.14%)	Decrease of RBC or hemoglobin	0, 1

Missing Data

Missing values from the data were imputed with average values. We report 25 missing values for Platelets, 19 for Serum Creatinine, 47 for CPK, and none for other variables. The suspected missing data mechanism is Missing at Random, as the missingness do not seem to be correlated with the outcome (Table 2), but we suspect are related to the condition of the patient and whether the doctor deemed it necessary or practical to collect the data. As the number of missing values is larger than 5%, we did not think it's appropriate to eliminate the entries with missing data. Instead, we imputed the missing values with the averages for variables Platelets (mean = 263358.03) and CPK (mean = 582). For Serum Creatinine, we note that the levels for dead and survived patients apparently have very different distributions (Table 3), thus we used two values for imputation instead of one, depending on patient outcome (mean = 1.18 for survived patients, mean = 1.83 for dead patients)

Table 2. Missing values characteristics.

	No Missing Values		Missing Values	
	DEATH = 0	DEATH = 1	DEATH = 0	DEATH = 1
Platelets	188 (68.6%)	86 (31.4%)	15 (60.0%)	10 (40.0%)
CPK	173 (68.6%)	79 (31.4%)	30 (63.8%)	17 (36.2%)
Serum Creatinine	192 (68.6%)	88 (31.4%)	11 (57.9%)	8 (42.1%)

Table 3. Characteristics of variables Platelets, CPK and Serum Creatinine for the dead and survived patients.

	Full sample			DEATH = 1			DEATH = 0		
Variable	Median	Mean	σ	Median	Mean	σ	Median	Mean	σ
Platelets	262.00	263.36	97.80	258.50	256.38	98.53	263.00	266.66	97.53
CPK	250.00	581.80	970.29	259.00	670.20	1316.58	245.00	540.10	753.80
Serum creatinine	1.10	1.39	1.03	1.30	1.84	1.47	1.00	1.19	0.65

Logistic Regression

When the outcome variable, i.e., whether the event death occurred or not, is a binary variable, logistic regression becomes the most appropriate regression method. Let the outcome death event depend on p explanatory variables $X_1, X_1, \dots, X_p$, then the multiple logistic regression model can be defined as:

$$logit(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \widehat{\beta_0} + \widehat{\beta_1}X_{i1} + \widehat{\beta_2}X_{i2} + \cdots + \widehat{\beta_{11}}X_{ip}, where\ i = 1, \dots 299$$

$$logit(DEATH) = \widehat{\beta_0} + \widehat{\beta_1}Age_{i1}$$

First, we identified variables of interest for the logistic regression. All eleven variables were considered possible covariates of domain interest or statistical importance and are included in univariate screening. A simple logistic regression model is fit for each variable, e.g., $logit(DEATH) = \widehat{\beta_0} + \widehat{\beta_1}Age_{i1}$ for the variable age, and the p-value for each variable was recorded. Next, a full regression model with all eleven covariates is fitted, namely $logit(DEATH) = \widehat{\beta_0} + \widehat{\beta_1}Age_{i1} + \widehat{\beta_2}Anemia_{i1} + \cdots + \widehat{\beta_1}Smoking_{i1}$, and the p-values are again recorded. Variables with p-values less than 0.05 from both models are the ones that are most likely to contribute to prediction power of the model and would be our focus for subsequent analysis. We also applied automated algorithms on forward, backward and stepwise model

selection, and any covariates identified by the algorithms but not elsewhere would also be taken into consideration in the finalizing steps of the final model but will not be extensively focused on.

Next, we considered possible confounding and effect modification among the previously identified variables, and possible transformations of the variables that might improve the model fit, such as polynomials. To assess confounding, we checked the classical definition of confounders for each couplet of variables and moved on to statistical definition where appropriate. To assess effect modification, a logistic regression model with an interaction term was fit for each couplet of variables, and the strength of the interaction term is recorded. After adjusting for above elements and including relevant polynomial terms, we had a full version of our final model. Then we trimmed down covariates on the basis of that in order to have the simplest model that would still achieve similar or better prediction powers.

Survival Analysis

Survival Analysis is a widely used method to analyze a time series data to find out the survival rate of a certain population at the exposure of different variates. In this study, survival analysis is used to see the survival rate of Heart Failure based on age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, and smoking data. Kaplan Meier Analysis and Cox Analysis were conducted to look at the overall survival rate of Heart Failure based on these variables.

Analysis and Results

Logistic Regression

To identify potential variables of interest, a simple logistic regression model was fitted for each of the eleven variables against the outcome DEATH. Table 4 summarizes the results for each model. Next, a full model with all eleven variables was fitted, where Table 5 summarizes the results. Variables with p-value < 0.05 are highlighted in red.

Table 4. Logistic Regression for each variable respectively

Variable	Estimate	Std. Error	P-value
Age	0.04694987	0.01107111	$2.227688 \cdot 10^{-5}$
Anaemia	0.2852695	0.2492401	0.2523926
log(CPK)	0.06043951	0.10897796	0.5791662
Diabetes	-0.008438869	0.251189948	0.9731997
Ejection Fraction	-0.05619986	0.01257730	$7.882363 \cdot 10^{-6}$
HBP	0.3508366	0.2561945	0.1708697
Platelets	$-1.115232 \cdot 10^{-6}$	$1.316095 \cdot 10^{-6}$	0.3967834
Serum Creatinine	0.8241855	0.1971648	$2.912706 \cdot 10^{-5}$
Serum Sodium	-0.096386117	0.029894074	0.001263009
Sex	-0.01934819	0.25923161	0.1708697
Smoking	-0.05813118	0.26633736	0.8272254

From these results, we have identified these four variables of interest: Age, EF, serum creatinine and serum sodium. The automated forward, backward and stepwise model selection algorithms

identified an additional two besides the four, which are anaemia and HBP, and all three algorithms gave the same set of variables.

Table 5. Logistic Regression Full Model

Variable	Estimate	Std. Error	P-value
Age	0.05443564	1.297127	$2.709138 \cdot 10^{-5}$
Anaemia	0.4063030	0.3035079	0.1806710
log(CPK)	0.2010330	1.332018	0.1312386
Diabetes	0.1309434	-0.2967406	0.6590160
Ejection Fraction	0.01480204	0.01480204	$2.729820 \cdot 10^{-6}$
HBP	0.4227126	0.3059557	0.1670903
Platelets	$-5.817674 \cdot 10^{-7}$	$1.600034 \cdot 10^{-6}$	0.7161590
Serum Creatinine	0.6768479	0.1740534	$1.007671 \cdot 10^{-4}$
Serum Sodium	-0.05324029	0.03295177	0.1061584
Sex	-0.3554407	0.3483163	0.3075133
Smoking	0.1291356	0.3473023	0.7100233

Next, we considered possible confounding and effect modification among the four identified variables, and found the relationships summarized in Table 6 and 7 in the Appendix and revisited in the Discussion. None of the variables is an effect modifier of another. We then included polynomials of the variables in the model to identify ones that may improve the model fit, which are found to be $(EF)^2$ and $(\text{Serum Creatinine})^2$.

Table 6. Checking the classical definition of confounding between the variables

Is (column name) a potential confounder of (row name)?	age	ejection fraction	serum creatinine	serum sodium
age	-	No	No	No
ejection fraction	Yes	-	Unsure-check literature	Unsure-check literature
serum creatinine	Yes	Unsure-check literature	-	Unsure-check literature
serum sodium	Yes	Unsure-check literature	Unsure-check literature	-

Table 7. Checking the statistical definition of confounding between the variables.

Is (column name) a potential confounder of (row name)?	age	ejection fraction	serum creatinine	serum sodium
age	-	No	No	No
ejection fraction	Yes	-	No	No
serum creatinine	Yes	No	-	Yes
serum sodium	No	Yes	Yes	-

After trimming covariates, we have determined that our final model would take form of

$$logit(DEATH) = \widehat{\beta_0} + \widehat{\beta_1}Age + \widehat{\beta_2}EF + \widehat{\beta_3}EF^2 + \widehat{\beta_4}Serum\ Creatinine$$

Of the four potentially impactful variables identified before, serum sodium was the weakest and constantly was associated with non-significant p-values across various models. As we have

determined it to not be a confounder of the remaining variables, it is removed from the final model. We also tried including the two variables identified only by automated model selection algorithms, anaemia, and HBP, in the final model to see if they improve model performance, but they did not and in fact led to an increase in AIC. Therefore, the final model is designed to only depend on the three variables age, EF and serum creatinine.

After fitting the model to the dataset, our final logistic regression model is as below.

$$logit(DEATH) = 1.9387 + 0.0590Age - 0.3265EF + 0.0032EF^2 + 0.6036Serum\_Creatinine$$

β_0 represents log odds of death in the unexposed group, on average.

$\beta_i, i = 1,2,3,4$ represents log(OR) of the death associated with 1-unit change in each covariate among the exposed vs. unexposed group, on average, holding all other covariates constant.

Exponentiating the coefficient we got:

$e^{\beta_0} = 6.9499$; this represents odds of death in the unexposed group across all covariates, on average.

e^{β_i} represents OR of the death associated with 1-unit change in each covariate among the exposed vs. unexposed group, on average, holding all other covariates constant (Table 8).

Table 8. Estimates for parameters of the logistic regression model

Variable	Estimate	e^{β}	Std. Error	Z-score	P-value
(Intercept)	1.9387305	6.949922	1.3787168	1.406	0.159669
Age	0.0590040	1.060779	0.0130444	4.523	$6.088074 \cdot 10^{-6}$
EF	-0.3264504	0.7214801	0.0682338	-4.784	$1.715901 \cdot 10^{-6}$
$(EF)^2$	0.0032282	1.003233	0.0008068	4.001	$6.298837 \cdot 10^{-5}$
Serum Creatinine	0.6035551	1.828608	0.1740540	3.468	$5.250677 \cdot 10^{-4}$

Survival Analysis

Kaplan Meier Analysis

A complete Kaplan Meier plot of Heart Failure based on all the variates was created to illustrate the overall survival rate of the cohort.

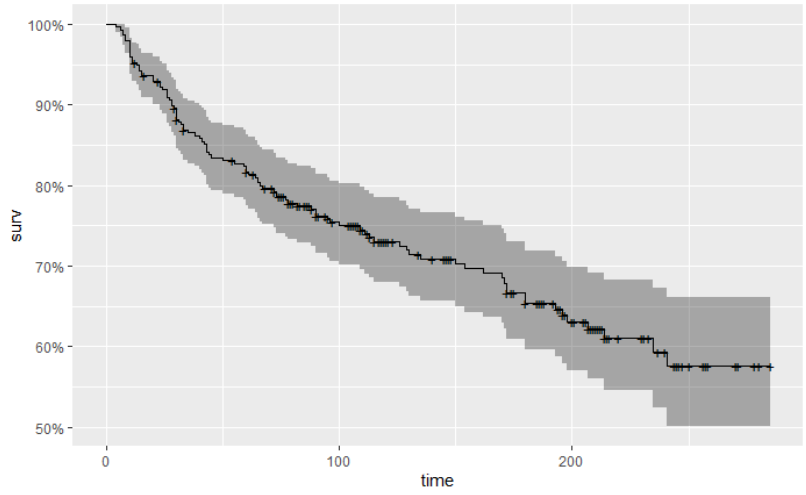

Figure 1 Kaplan Meier Plot

We could see from the here that survival rate keeps decreasing to 57.6% at a gradual rate, and the survival rate reaches constant at 57.6% after 241 days, which indicates that the remaining 57.6% of the cohort didn't develop a lethal heart failure during the 241 days following up period *(Figure 1)* thus we could say that Heart Failure has a relatively high survival rate in a short term period, and is more likely to be a chronic disease.

To get a deeper understanding on what would possibly be a risky factor of heart failure, we considered to do a Kaplan Meier plot based on selected variables to see the effect of certain variables on the survival rate. Age, ejection fraction, serum creatinine, high blood pressure, anaemia, and creatinine phosphokinase data were plotted to see if different levels of variable would lead to a significant difference in the survival rate.

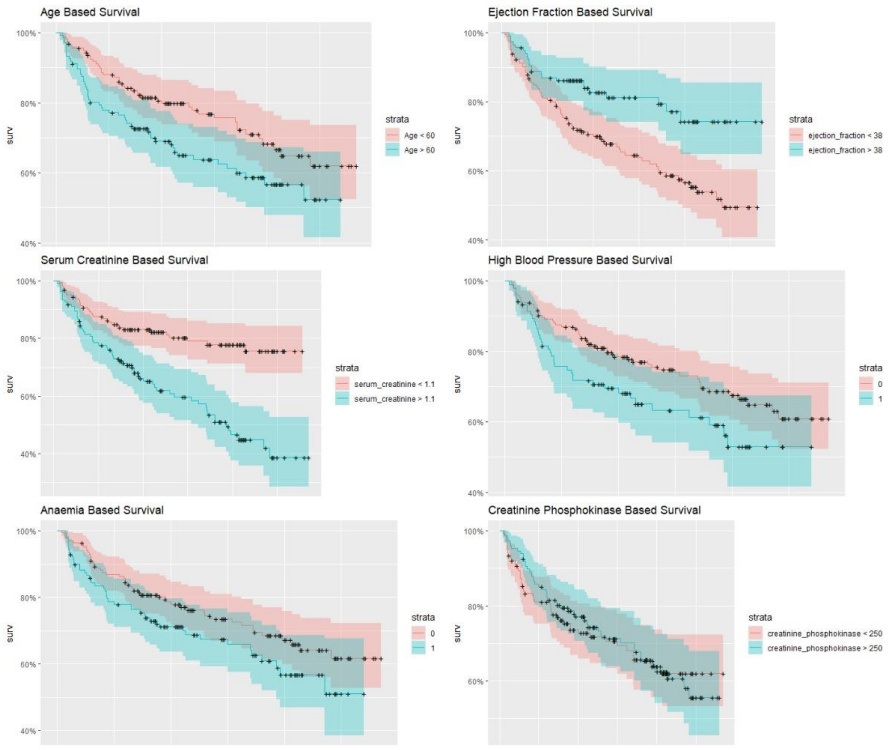

Figure 2 Variable-wise Kaplan Meier Plot

From the variable-wise Kaplan Meier Plot above, we could see that there is almost no overlapping in average survival rate and 95% Confidence Interval over time between higher and lower Ejection Fraction and Serum Creatinine level, which indicates that ejection fraction and serum creatinine level may have a significant effect on estimating the survival rate of heart failure. The average survival rate based on Age, High Blood Pressure, and Anaemia differs between higher and lower level, however, their confidence intervals overlap, which may fail to reject the null hypothesis or their effect on survival rate may be insignificant. The Kaplan Meier plot based on Creatinine Phosphokinase show no difference between higher level and lower level, which indicates that Creatinine Phosphokinase level may have a minor influence on estimating the survival rate of heart failure.

Cox Analysis

Cox Analysis (or Cox Proportional Hazard Analysis) is used to find out the relationship between the variables and the survival rate. Proportional Hazard is the most important assumption we need to verify before doing this analysis. The r package *survival* was used to test for the proportional hazard assumption. The p-values we get from the testing showed that the global p-value is 0.39 (> 0.05), which indicates that there is no significant non-proportional hazard in this data set. *(Table 9)* However, we could see in the table that the p-value for ejection fraction is 0.03 (< 0.05), which means that there may exist non-proportional hazard of ejection fraction on the survival rate.

Table 9 Proportional Hazard Testing

Variable	p-value
Age	0.75
Anaemia	0.90
Creatinine Phosphokinase	0.31
Diabetes	0.66
Ejection Fraction	0.03
High Blood Pressure	0.93
Platelets	1.00
Serum Creatinine	0.22
Serum Sodium	0.74
Sex	0.78
Smoking	0.49
Global	0.39

Since the global p-value is larger than 0.05, we could assume that the overall hazard in this data set is proportional. Thus, a cox proportional hazard analysis could be done to find out the relationship between the variables and the survival rate. *survival* R package was introduced to

conduct the cox analysis. The result of cox analysis suggested that age, ejection fraction, and serum creatinine have significant influences on the survival rate of heart failure.

According to the result of Cox analysis *(Table 10)*, we could get the following conclusions that:

- One-unit increase vs current unit in age leads to 1.05 times increase of hazard ratio of heart failure, on average, holding other covariates constant, so there's an increased risk of heart failure as age increases.
- One-unit increase vs current unit in ejection fraction leads to 0.952 times decrease of hazard ratio of heart failure, on average, holding other covariates constant, so there's a decreased risk of heart failure as age increases.
- One-unit increase vs current unit in serum creatinine leads to 1.38 times increase of hazard ratio of heart failure, on average, holding other covariates constant, so there's an increased risk of heart failure as age increases.

Table 10 Cox Analysis

Variable	Coefficient	exp(Coefficient)
Age	0.0464	1.05
Ejection Fraction	-0.0489	0.952
Serum Creatinine	0.321	1.38

Discussion

Our final logistic regression model returned 3 covariates significant for heart failure prediction. It seems intuitive that age and ejection fraction are common predictors for coronary heart disease; creatinine is a chemical product left over from energy-producing processes in the body filtered out by the kidney. A measurement of serum creatinine helps determine how well the kidney functions. Our study results suggested that the increase in the level of serum creatinine, which indicates an unhealthier kidney, was not only significant in predicting heart failure but also the most important predictor, with the odds ratio to be 1.82 times higher for patients with 1 mg/dL increase in the serum creatinine level on average.

Survival Analysis gives an estimation of survival rate of heart failure rate against time. The result is quite informative despite the relative low hazard rate. Among the variables from the data set, age, ejection fraction, and serum creatinine were the three most significant variables that could predict the survival rate of heart failure. The increase in the age and serum creatinine has a harmful effect on heart failure, while increased ejection fraction has a protective effect on heart failure.

Limitation

The number of samples in the data set is relatively small compared to the high dimension of variable, thus the generalization of the data set needed to be tested with more samples. There might also be unmeasured confounding related to our covariates but not included in our study. Although we did confounding and effect measure modification in the study, the relationship between serum sodium and serum creatinine was still inconclusive due to the small number of studies published assessing the relationship between the two. Therefore, it is hard to assess the confounding effect between the two variables. If these two variables are not consequences of

each other, then they do not meet the definition of confounders, and will be considered as such even though they both satisfy the statistical definition. There is also little literature on the relationship between ejection fraction and serum creatinine/sodium; most discuss their effect on some other variables, but results from testing the statistical definition have conveniently ruled out many possibilities. Finally, publicly available datasets are prone to unreliable data manipulation and data collection methods prior to when they are published.

Future Scope

Assessing collinearity in survival analysis would be the next step of follow up research. After analyzing more detailed in the collinearity of the variables, a more accurate regression model could be built to predict the occurrence of death event.

For future research, we could focus more on Investigate ML methods and how the literature was able to achieve good predictions with one less variable than us (age), since Machine Learning Algorithms are nowadays more and more efficient predicting either continuous or categorical output. Algorithms like random forest, support vector machine, or even neural network could be tested on the data set to see if could produce a higher prediction power.

References

1. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. *PLoS ONE.* 2017; 12(7): e0181001, https://doi.org/10.1371/journal.pone.0181001.
2. Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak.* 2020; 20, 16. https://doi.org/10.1186/s12911-020-1023-5
3. Heart failure. cdc.gov. Updated September 8, 2020. Accessed December 10, 2021. https://www.cdc.gov/heartdisease/heart_failure.htm
4. Zhid FM, Ramzan S, Faisal S, Hussain I. Gender based survival prediction models for heart failure patients: A case study in Pakistan. *PLoS ONE.* 2019. https://doi.org/10.1371/journal.pone.0210602