Survival Analysis of Heart Failure Patients

Alexander E

Industrial, Manufacturing and Systems Engineering

Reliability Theory IE5345

November 27, 2022

Table of Contents

Abstract

Objective: This Project is a study of the patient's survival rate due to heart failure condition. One of the premises of this study is that it was based on other researches on cardiovascular diseases of the heart, which has become very common in medical profession. This project will aim to investigate the most at risk features to help predict heart failure conditions.

Methods: The data used for this study was collected from Faisalabad Institute of Cardiology during the periods April to December 2015. It comprised of 13 features and 299 observations and was subjected under different distribution models through lifelines library in python programming software technology to determine reliability and failure conditions of the heart. Other concepts used were Kaplan Meier nonparametric estimator for censored data to model survival time outcomes on one or more predictors (features) and cox proportional hazards model to train the data for model accuracy.

Result: Among the different features, we observed the alpha (lambda_) of each feature, serum-creatinine with alpha value of **2.063559** and ejection-fraction with alpha value of **37.613387** which is very low compared with others. low alpha values indicate features are at risk because the alpha tells us about the characteristic life of the model and how long it will last before it fails. Also, the corresponding p-values for age (P=6.4e-07), creatinine phosphokinase (P=0.03), ejection fraction (P=3.0e-06) and serum creatinine (P=4.8e-06) were the features that were statistically significant below $P<0.05$

Conclusion: In conclusion, when comparing the significant features, Ejection fraction seems to be the feature most associated with heart failure of the patients in this study, with the lowest survival rate of 0.01%. Anaemia and High blood pressure begins to worsen and reduce the survival rate of patients if they are not diagnosed and treated properly along the follow time for

treatment. While Serum creatinine levels can be managed as critical diagnosis still have a 50% survival rate.

Keywords – *Kaplan Meier, Weibull, Cox Proportional Hazard (Cox), Ejection fraction, Serum creatinine, Anaemia, Model Assumptions, Reliability, Survival rate, Hazard rate, Heart failure, P-value, Alpha*

Introduction

The data used for this study was collected from Faisalabad Institute of Cardiology during the periods April to December 2015. It comprised 13 features and 299 observations, including censored data of heart failure patients. Below is a display of the data and the features used in the study.

Features: age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high-blood-pressure, platelets, serum creatinine, serum sodium sex, smoking, time, and death event.

Fig 1

cardiovascular dataset

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	2	582	2	20	1	265000.00	1.9	130	1	2	4	1
1	55.0	2	7861	2	38	2	263358.03	1.1	136	1	2	6	1
2	65.0	2	146	2	20	2	162000.00	1.3	129	1	1	7	1
3	50.0	1	111	2	20	2	210000.00	1.9	137	1	2	7	1
4	65.0	1	160	1	20	2	327000.00	2.7	116	2	2	8	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Cleaning

The data gathered was then examined to determine the data structure and cleaned to remove any unwanted element, ensuring there are no missing values and replacing these missing values with options. There are some basic methods to do this; A good method will be replacing missing items in the data with the median value of the variables (features) in the data set. After

which the cleaned data was subjected to validation. A false response indicates there are no missing items as shown in fig. 2 below.

Fig 2

Cleaned dataset

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False

Data Description and Basic Statistics

The data comprised of both continuous and categorical features, and the table below shows the basic statistics of the different features which gives the mean, std, quantile ranges and maximum and minimum values of each variable. It also helps with an indication of how the data is distributed.

Continuous Features:

- Age - The length of time a person has lived before, during and death by diagnosis of heart failure disease. The mean age is 60 with maximum life expectancy of 95 and std of 11.9.
- Creatinine phosphokinase (CPK) – This is an enzyme found in the body. When CPK is high, it means there has been injury or stress to the muscle tissue, the heart, or the brain because of CPK flowing into the blood. The patients in this study recorded CPK levels with reference to injury to the heart, and the ranges were between 23 to 7861 mcg/L, while normal CPK range is between 10 to 120mcg/L. High levels CPK can lead to heart failures like heart attacks, heart muscle inflammation etc.

- Ejection Fraction – This measures the percentage proportion of blood pumped out of the heart through the left ventricle of the heart. In this study, the mean amount of blood is 38% across the patients, with the highest at 80%. The normal range should be between 55 to 70%, while a range less than 55% is considered below normal heart function and indicates signs of previous damage to the heart. On the other hand, a higher range of 75% and above or 40% and below confirms a patient is diagnosed with heart failure.
- Platelets – The measure of platelets in the blood shows the ability of your blood cells to clump together to form blood clots to seal wounds and stop bleeding after an injury has occurred both internal and external. Low platelet counts may be life threatening if you are unable to stop the bleeding, especially when it's internal. The normal range is between 150,000 to 450,000 kilo-platelets/ml, while below 150,000 indicates low count levels. The patients in the study have platelets counts ranging from 25,100 to 850,000 kilo-platelets/mL with the mean platelets count of 263,358 kilo-platelets/ml.
- Serum creatinine – This is waste substance that occurs when the muscles in the body breakdown through the functions of the kidney. It gives doctors an indication of the amount of creatinine levels in the blood. Creatine levels differ in males and females, normal range is 0.7 – 1.3mg/dl and 0.6 – 1.1mg/dl respectively. High serum creatinine indicates poor functionality of the kidneys and occurs too often with patients diagnosed with heart failure conditions. The patients in this study had a mean level of 1.3mg/dl with the highest at 9.4mg/dl.
- Serum sodium – This measures the level of sodium in the blood, there needs to be a balance between sodium and water for normal functioning ability of the muscles and nerves in the blood. Sodium level in the blood should have a normal range of 136 –

145mEq/L. If the sodium level falls below 136mEq/L, it indicates excessive water over sodium and can lead to seizures, convulsions etc. It also increases death rate of patients diagnosed with heart failure. The patients in this study had a mean sodium level of 136.6mEq/l with the lowest level at 113mEq/L.

- Time – In this study time refers to the follow up period after patients have been diagnosed and started treatment measured in days of survival. Among the 299 patients, the mean number of days is 130days and the shortest follow up period was 4 days.

Fig 3

Continuous Features Leading to Heart Failure.

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000
mean	60.833893	581.839465	38.083612	263358.029264	1.39388	136.625418	130.260870
std	11.894809	970.287881	11.834841	97804.236869	1.03451	4.412477	77.614208
min	40.000000	23.000000	14.000000	25100.000000	0.50000	113.000000	4.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.90000	134.000000	73.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.10000	137.000000	115.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.40000	140.000000	203.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.40000	148.000000	285.000000

Categorical features:

The categorical features in the study represent variables that could only be measured as Boolean, hence can only assume 2 positions or binary values, either true or false, existence or nonexistence, 1 or 2. In this study there were 5 categorical features as shown below in fig 4.

- Anaemia – This is an indication of the decrease of red blood cells in the body. A value of 1 indicates presence of aneamia in the patient, while 0 indicates otherwise. Out of 299 patients, 129, 43.14% indicated presence of aneamia characterized as patients with red blood cell levels lower than 36%.

- High blood Pressure – A condition when blood flows through your arteries at a higher-than-normal pressure. It's made of 2 numbers systolic pressure when the ventricles pump blood out of the heart and diastolic is the pressure between heartbeats when the heart is filled with blood. Normal blood pressure level is less than 120/80mmHg. Levels from 130/80mmHg are considered high. In the data, 64% of the patients show no sign of high blood pressure.
- Diabetes – A patient is diagnosed as diabetic when the body doesn't make enough insulin and in some case is unable to make use of its insulin. This inevitably leads to high glucose levels in the blood which makes the body not function properly. Insulin levels above 126mg/dl or higher. From the data 41% of the patients who died indicated presence of diabetes.
- Sex – This describes if patient in the study is male or female. Further analysis can be done to analyze the patients into groups to discover which group is prone to each feature leading to heart failure which will help in discovery of new treatments to prevent heart failure.
- Smoking – An indication of if the patient in the study smokes or doesn't smoke.

Fig 4

Categorical Features of Heart Failure

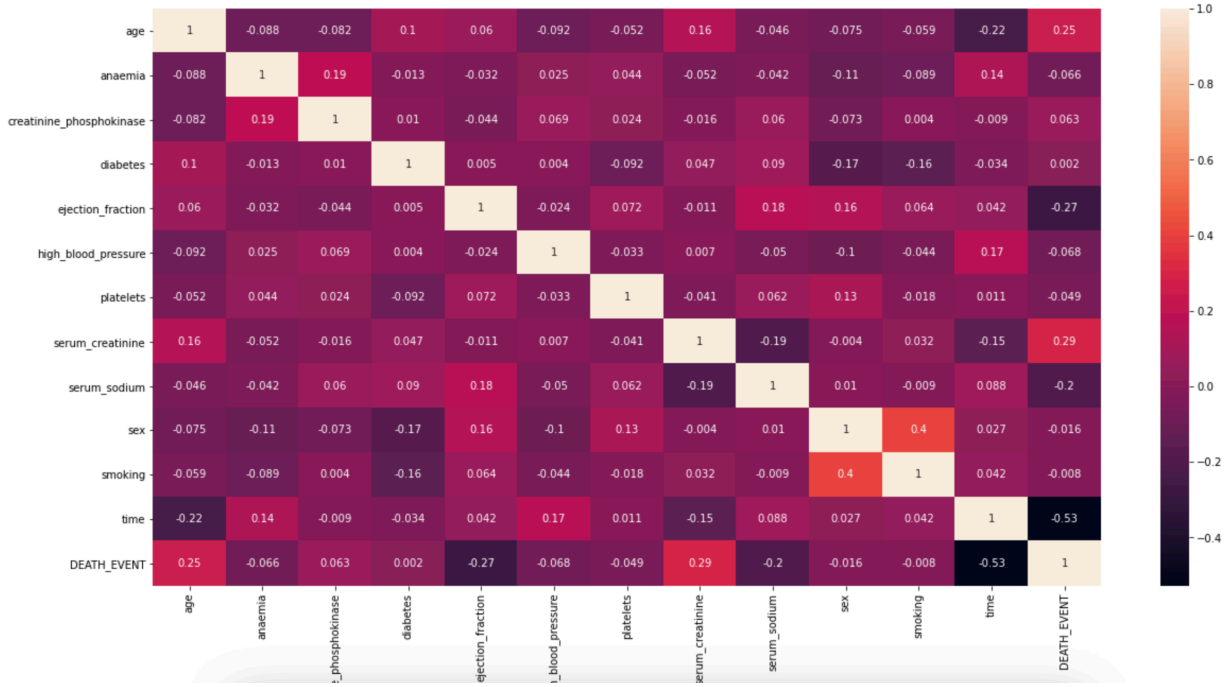
	Category feature	Total	%	Dead Total	%.1	Survived Total	%.2
0	Anaemia_nopresence	170	56.86	50	52.08	120	59.11
1	Anaemia_presence	129	43.14	46	47.92	3	40.89
2	High blood pressure_nopresence	194	64.88	57	59.38	137	67.49
3	High blood pressure_presence	105	35.12	39	40.62	66	32.51
4	Diabetes_nopresence	174	58.19	56	58.33	118	58.13
5	Diabetes_presence	125	41.81	40	41.67	85	41.87
6	Sex_women	105	35.12	34	35.42	71	34.98
7	Sex_man	194	64.88	62	64.58	132	65.02
8	Smoking_nopresence	203	67.89	66	68.75	137	67.49
9	Smoking_presence	96	32.11	30	31.25	66	32.51

Correlation

Because this study focuses on multivariate features in patients' survival rate (death rate) from heart failure, we tried to get an overview of how the features are associated by doing a correlation to determine if there is any other relationship that exist between them.

Fig 5

Correlation Coefficients and Heatmap of Heart Failure Features



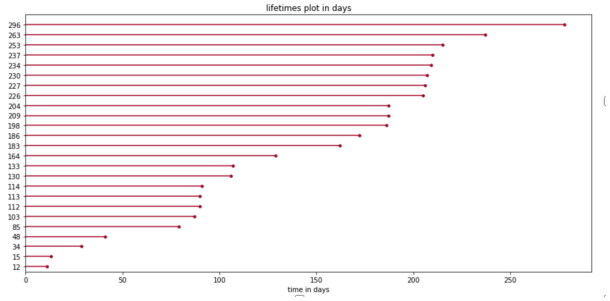
The correlation plots help us identify other relationships between the features. From the correlation plot, Time (follow-up period) has a high inverse relationship with death event. This is understandable as patients' recovery is associated with recuperation time. There also seems to be a 40% relationship between sex and smoking. These results can be used for future studies with other multivariate analysis techniques.

Lifeline plots

Lifeline plots is used to visualize the lifetime plots of interval censored data and failure events. Due to the number of observations, a random selection of 25 samples of the data was used to develop the lifeline plots.

Fig 6

Lifeline Plots of Heart Failure



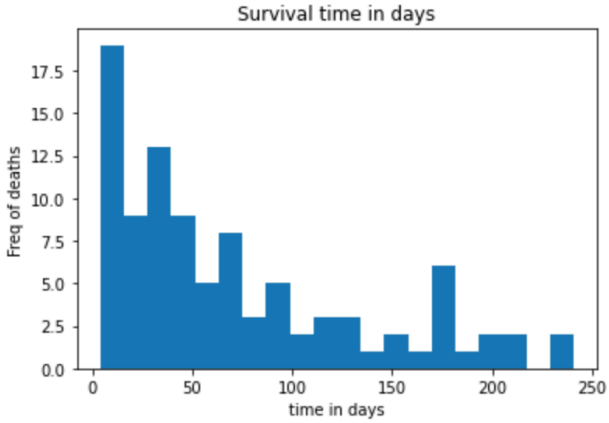
From the plot we can observe patient 296 as censored data, meaning we are unable to determine the heart failure event of that patient, hence the patient probably survived treatment after 250days.

Histogram to show frequency of death patients against survived time in days.

The death rate feature in the data is binary which assumes only 2 values. 1 means a death event occurred while 0 means the patient survived, we can also represent this as a histogram plot to show the frequency of failure events (deaths) over a period.

Fig 7

Histogram of heart failure



From the plot there are more death occurrence at the early stage of treatment which reduces over time. This can indicate prevention of heart failure if the patient responds to treatment. Since in this project we are concerned with survival analysis of the heart, we used reliability concept to

quantify heart failure by measuring the failure probability of the heart while its functioning under different conditions by making inference to features we can find in the blood.

Reliability

Reliability is the probability that an item will perform its intended function for a specified period of time ***(t)*** under a given set of conditions. When reliability is 1 this represents success and 0 represents failure. Hence R(t) is the probability that a random unit drawn from a population will still be operating (surviving) after $\boldsymbol{t}$ hours. That is R(t) is the fraction of all units in the population that will survive after $\boldsymbol{t}$ hrs.

E (number *of failures*) = *nF*(*t*), *E* (number *of survivors*) = *nR*(*t*)

If n identical units are operating and $F(t)$ describes the population they come from, then

nF(t) is the expected or average number of failures up to time t,

nR(t) is the expected or average number of survivors expected to still be operational.

Kaplan Meier

Kaplan Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. For example, the fraction of patients living for a certain amount of time after treatment, the time-to-failure of machine parts etc.

The estimator of the survival function *S(t)* (the probability that life is longer than t) is given by:

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

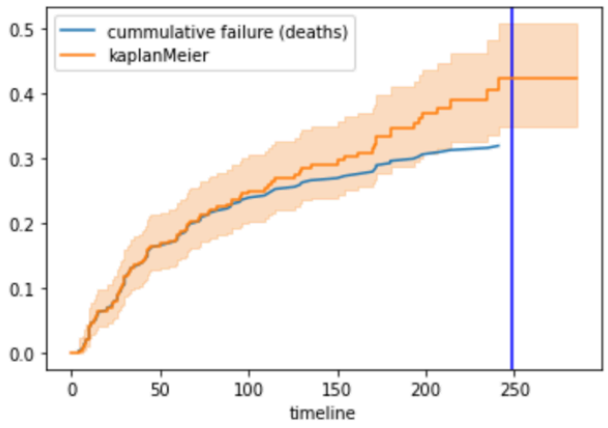
The Kaplan Meier estimator considers censored data while performing its analysis, unlike using a distribution model which only works with data that have established failure points, that is data with failure events. Hence, we could generate a failure Kaplan Meier estimate curve and compare this with the cumulative failure events of the patients to get a picture of the survival situation using the death rate of the patients over a period.

Cumulative failure plots to describe death rate

There are various methods of getting cumulative failure rate, using sum of cumulative failure over the sample size, or using rank method. For this project we considered the rank method and generated a plot to compare the failure rate with Kaplan Meier estimator as shown below:

Fig 8

Comparison of Kaplan Meier Estimator and Cumulative Failure Plot of Heart Failure



From the plot, there is a slight difference in the projection of the curves between the Kaplan Meier and cumulative failure rate. However, the curves also indicate increasing death rate overtime, but stops before the 250th day of the follow-up time, where there seems to be no more deaths (heart failure).

Distribution Models for Measuring Failure Analysis:

Next, we worked on the development of a model which we can use to predict the survival rate of patients with heart failure. There are several distribution models for measuring failure probability, however for this experiment we considered only 2 distribution models to check

which one will be a better fit, and the distributions we considered were Weibull distribution and Exponential distribution.

- Exponential distribution, which is a continuous distribution model, is used to measure the expected time for an event to occur, the event here represents a failure event. That is, it's used to measure the time to failure of an event. Exponential distribution also has the assumption of constant failure rate lambda (λ) which is the average number of events in a unit time interval. Therefore, it is a good distribution for modeling time to failure between events in the lifetime distribution of a process under a constant failure rate.

  Hence time to failure is denoted as: $F(t) = 1 - e^{-\lambda t}$

The exponential distribution applies only under the constant failure rate assumption. But what do we do if the failure rate is increasing or decreasing?

- Weibull distribution is a continuous distribution model and can also be used to measure time to failure of an event. However, it is most appropriate for measuring lifetime data of a process because it doesn't assume constant failure rate. A typical example is the bathtub curve model which assumes a machine or product lifetime can be estimated with its failure rate which comprises of various stages in engineering analysis. Therefore, the failure rate of a machine or product decreases in its infant stage, becomes constant in its normal life (exponential – λ) and increases in its wear out stage during its lifetime usage. Weibull distribution model is the only model that can measure this phenomenon and it makes use of 2 parameters, α is a scale parameter (characteristic life) and β is shape parameter. Hence time to failure is denoted as: $F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$

First, we started with comparing the model estimates, considered maximum log-likelihood and AIC of both models. A model with a smaller AIC value is a better fit, while the model with a higher(maximum) log-likelihood is a good fit.

```
############################## exponential ####################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	405.708301	41.407427	324.551236	486.865366	0.0	9.79796	1.148826e-22	72.882258

```
AIC of Exponential model is:  1347.081826041116
The loglikelihood for exponential dist is -672.54
```

```
############################## weibull ####################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	491.735790	80.564012	333.833227	649.638352	1.0	6.091253	1.120303e-09	29.733464
rho_	0.833304	0.076913	0.682557	0.984051	1.0	-2.167328	3.020983e-02	5.048838

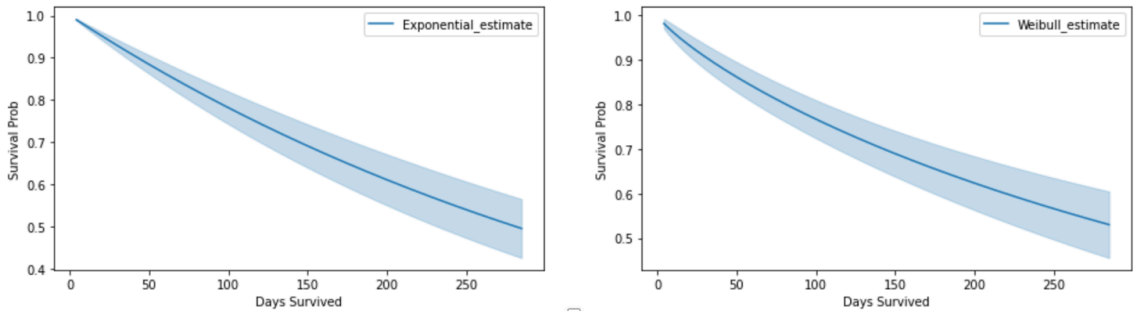
```
AIC of Weibull model is:  1344.8757115822702
The loglikelihood for weibull dist is -670.44
```

After running initial analysis, its observed that Weibull distribution has a higher loglikelihood of -672.54 and the smallest AIC of 1344.86 when compared with exponential model estimates. Therefore, Weibull is a better fit for the model because in statistics, you would usually want a higher loglikelihood and lower AIC for model fits and predictors.

Below is also a visual distribution of both models.

Fig 10

Survival Plots for Distribution Models of heart failure

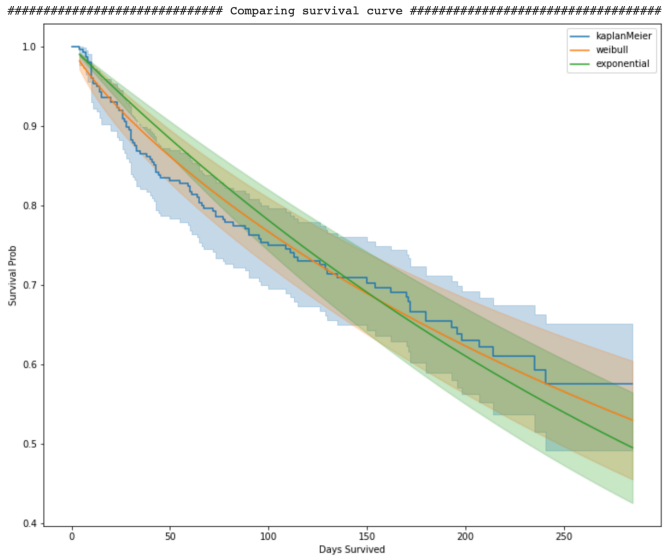


Other things considered was Kaplan Meier survival plot against the survival plots of both models to visualize the direction of the curve. Therefore, if we model the data using Kaplan Meier estimator function, we can deduce a visual representation of the patients who survived under

certain heart conditions over a period of time and compare the direction of the curve with the other distribution models to ascertain which curves portrays a better representation of the Kaplan Meier estimator.

Fig 10

Kaplan Meier Estimator of Heart Failure



From careful observation, we can detect a steady decline in patients' survival rate below 50% after 250 days. This means most patients who died of heart failure happened in the latter days of treatment. This might be due to other underlying ailments that led to the cause of the heart failure, or the patients were unable to get adequate treatment.

And finally, the survival plot of the 2 models also reveals which model follows the Kaplan Meier curve. The green curve which represents exponential distribution survival plot seems to be a little bit deviated from the blue curve, which is the Kaplan Meier survival plot, while the orange plot is line with it.

Reliability data plot:

To further explore which model is a better fit, a comparison of the data probability plot can be used to check the clustering of observations along a slope line. To do this we would run a regression analysis on the data for the 2 distributions, that is Exponential and Weibull distribution. This method also helps to validate the maximum likelihood estimation.

Setting up the regression analysis requires different parameters based on the distribution model. While Weibull distribution requires a constant because it has a slope and intercept due to its model parameters, α - scale parameter (characteristic life) and β - shape parameter, the reverse is the case for exponential distribution when performing regression analysis because it considers only failure rate λ lambda which is a constant.

Therefore, assuming the lifetime of the patients follows exponential distribution we can fit the model in ordinary least square regression to estimate the coefficients. The Input and output are derived from reliability exponential distribution formula

$$-ln(1-Ft) = lambda\ (\lambda) \text{ x time } (t)$$

```
############################## exponential ###################################
                            OLS Regression Results
==============================================================================
Dep. Variable:            DEATH_EVENT   R-squared (uncentered):                   0.935
Model:                            OLS   Adj. R-squared (uncentered):              0.934
Method:                 Least Squares   F-statistic:                              1367.
Date:                Mon, 12 Dec 2022   Prob (F-statistic):                    3.43e-58
Time:                        16:55:42   Log-Likelihood:                          143.43
No. Observations:                  96   AIC:                                     -284.9
Df Residuals:                      95   BIC:                                     -282.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
time           0.0022   5.91e-05     36.971      0.000       0.002       0.002
==============================================================================
Omnibus:                       29.210   Durbin-Watson:                   0.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.060
Skew:                          -1.394   Prob(JB):                     2.71e-10
Kurtosis:                       4.800   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard s assume that the covariance matrix of the s is correctly specified.
lambda is: 0.0021867574032586176
```

On the other hand, if the lifetime of patients follows a Weibull distribution, it requires a constant to fit the model in regression analysis to estimate the coefficients alpha α and beta β. In reliability the Weibull distribution formula is derived as

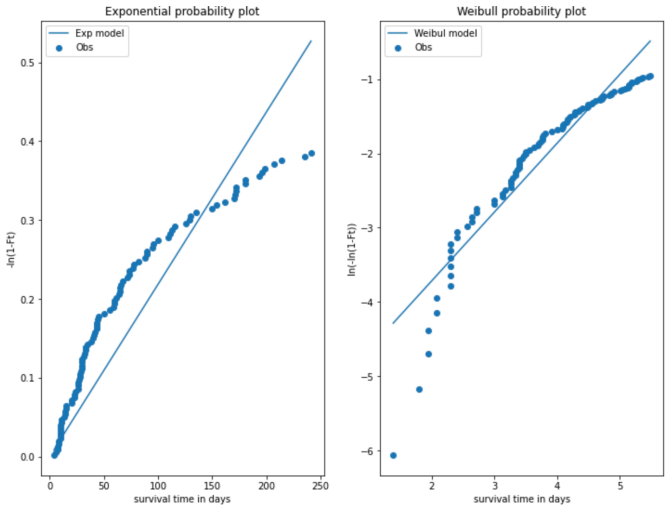
$$ln(-ln(1-Ft)) = beta\ (\beta)\ x\ ln\ (time\ (t)) - beta\ (\beta)\ x\ ln\ (alpha\ (\alpha))$$

```
############################## weibull ###################################
                            OLS Regression Results
==============================================================================
Dep. Variable:             DEATH_EVENT   R-squared:                       0.884
Model:                             OLS   Adj. R-squared:                  0.883
Method:                  Least Squares   F-statistic:                     716.7
Date:                 Mon, 12 Dec 2022   Prob (F-statistic):           9.14e-46
Time:                         16:55:42   Log-Likelihood:                -32.772
No. Observations:                   96   AIC:                             69.54
Df Residuals:                       94   BIC:                             74.67
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.5710      0.137    -40.732      0.000      -5.843      -5.299
time           0.9267      0.035     26.771      0.000       0.858       0.995
==============================================================================
Omnibus:                       66.287   Durbin-Watson:                   0.065
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              320.339
Skew:                          -2.297   Prob(JB):                     2.75e-70
Kurtosis:                      10.680   Cond. No.                         16.3
==============================================================================

Notes:
[1] Standard s assume that the covariance matrix of the s is correctly specified.
alpha is: 408.2465413413826
beta is: 1.0791364876436285
```

Fig 11

Reliability Data Plot of Exponential and Weibull Distribution



By careful observation between the 2 plots, it shows that Weibull is a better fit for the model as most of the data points (observations) seem to be clustered along the slope line. Hence, we would use the Weibull distribution model to predict other features affecting heart failure of the patients, this will help us determine which features are highly associated with patients' survival.

Modeling of Continuous Features

While modeling for the continuous features, censored data was removed, that is patients in which we were unable to determine a failure event (death) because we are of the assumption that they survived treatment from the data we obtained.

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
0	75.0	582	20	265000.00	1.90	130
1	55.0	7861	38	263358.03	1.10	136
2	65.0	146	20	162000.00	1.30	129
3	50.0	111	20	210000.00	1.90	137
4	65.0	160	20	327000.00	2.70	116
...	...	...	...	...	...	...
220	73.0	582	20	263358.03	1.83	134
230	60.0	166	30	62000.00	1.70	127
246	55.0	2017	25	314000.00	1.10	138
262	65.0	258	25	198000.00	1.40	129
266	55.0	1199	20	263358.03	1.83	134

96 rows × 6 columns

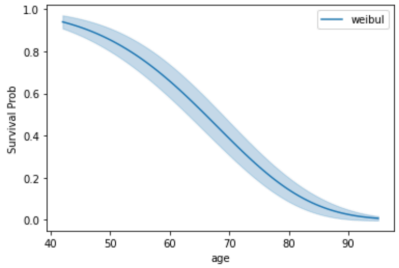
Hence 96 patients experienced heart failure and did not survive treatment, the features of these patients were modeled using Weibull distribution to determine the survival analysis in which we considered the Weibull parameters and the survival curve of each feature. The results in this test will help determine which feature can be used to predict patient's heart failure and survival rate for future studies or experiments.

Fig 12

Continuous Features Model Analysis

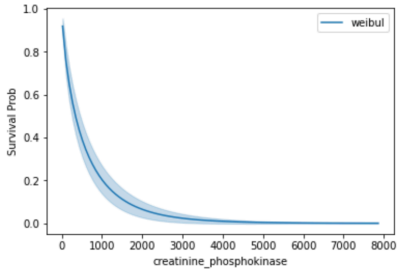
```
############################## age ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	70.651805	1.428920	67.851174	73.452436	1.0	48.744382	0.000000e+00	inf
rho_	5.345640	0.411925	4.538281	6.152999	1.0	10.549579	5.102532e-26	84.018917



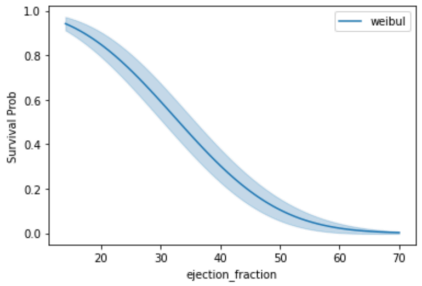

```
############################## creatinine_phosphokinase ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	551.824846	76.796256	401.306950	702.342742	1.0	7.172548	7.361450e-13	40.305075
rho_	0.780300	0.054459	0.673562	0.887038	1.0	-4.034208	5.478686e-05	14.155811



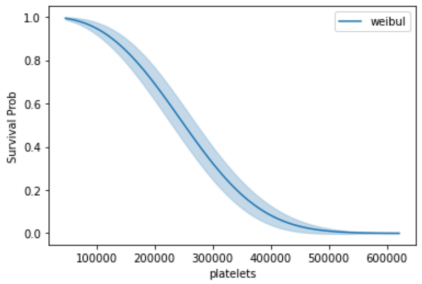
```
############################## ejection_fraction ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	37.613387	1.429357	34.811899	40.414875	1.0	25.615286	1.030769e-144	478.313924
rho_	2.848726	0.215262	2.426819	3.270632	1.0	8.588244	8.830884e-18	56.652148



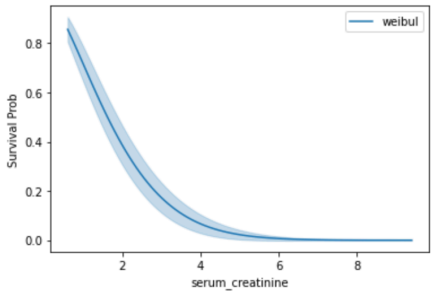

```
############################## platelets ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	287538.155810	11229.168689	265529.389604	309546.922017	1.0	25.606273	1.298882e-144	477.980375
rho_	2.756795	0.208681	2.347788	3.165802	1.0	8.418574	3.810943e-17	54.542630



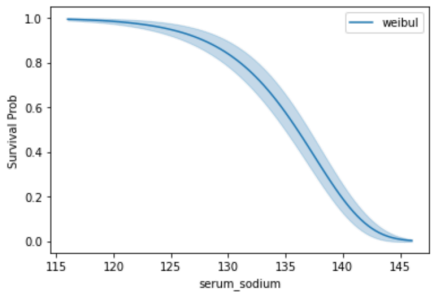

```
############################## serum_creatinine ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	2.063559	0.148934	1.771653	2.355465	1.0	7.141130	9.256674e-13	39.974571
rho_	1.507325	0.100534	1.310282	1.704368	1.0	5.046303	4.504418e-07	21.082156




```
############################## serum_sodium ##################################
```

	coef	se(coef)	coef lower 95%	coef upper 95%	cmp to	z	p	-log2(p)
lambda_	137.668670	0.489708	136.708860	138.628480	1.0	279.082022	0.000000e+00	inf
rho_	30.354604	2.280551	25.884806	34.824402	1.0	12.871716	6.494321e-38	123.534089



Among the different features, we observed the alpha (lambda_) of each feature, serum-creatinine with alpha value of **2.063559** and ejection-fraction with alpha value of **37.613387** which is exceptionally low compared with others. low alpha values indicate features are at risk because the alpha tells us about the characteristic life of the model and how long it will last before it fails. Hence, we want alphas of high values. Age is another indicator; it is the feature with the 3rd least alpha value of **70.651805**. When reviewing the age curve, it is likely that patients within that age range from 70 and upwards have low survival probability. In similar research on targeting age-related pathways in hearth failure (Haobo L, et al, 2020) concluded that during aging, there is a chance of deterioration in cardiac structure and function which leads to increased vulnerability to heart failure. The study also highlights 1% of people ages 50 and above are likely to experience this symptom.

Model Accuracy – Cox proportional hazards (CPH)

To train the model we used CoxPHFitter from lifeline to check model accuracy and fitting. Cox proportional hazards is a regression baseline model used to predict hazard ratio on one or more features. In our study hazard ratio refers to the probability of the death event over a period of time. It's also important that the model operates under certain assumptions; the (1) assumption is that the hazards are proportional and most remain constant over time; and (2) The relationship between log hazard and the covariates are linear. Below is the test result table

Fig 13

Summary Table for Cox Proportional Hazard Rates

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.046	1.048	0.009	0.028	0.065	1.029	1.067	0.000	4.977	<0.0005	20.564
anaemia	0.460	1.584	0.217	0.035	0.885	1.036	2.423	0.000	2.122	0.034	4.885
creatinine_phosphokinase	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	2.226	0.026	5.263
diabetes	0.140	1.150	0.223	-0.297	0.577	0.743	1.781	0.000	0.627	0.531	0.914
ejection_fraction	-0.049	0.952	0.010	-0.069	-0.028	0.933	0.972	0.000	-4.672	<0.0005	18.354
high_blood_pressure	0.476	1.609	0.216	0.052	0.899	1.053	2.458	0.000	2.201	0.028	5.170
platelets	-0.000	1.000	0.000	-0.000	0.000	1.000	1.000	0.000	-0.412	0.681	0.555
serum_creatinine	0.321	1.379	0.070	0.184	0.459	1.201	1.582	0.000	4.575	<0.0005	17.681
serum_sodium	-0.044	0.957	0.023	-0.090	0.001	0.914	1.001	0.000	-1.899	0.058	4.120
sex	-0.238	0.789	0.252	-0.731	0.256	0.482	1.291	0.000	-0.944	0.345	1.535
smoking	0.129	1.138	0.251	-0.363	0.621	0.695	1.861	0.000	0.513	0.608	0.718

The summary table provides useful information that helps identify at risk features highly associated with heart failure disease in the patients. In statistical analysis, the P-value is an indication that the probability under the assumption of null hypothesis (no effect) is correct when the test results are equal to or at least as extreme as the result observed. In Cox (CPH) model, a P-value equal to or greater than 0.05 is statistically insignificant assuming a 95% confidence threshold is used to test the features. From the table, diabetes, platelets, sex, smoking, and serum sodium are statistically insignificant with P>0.05. While features *anaemia, age, creatine phosphokinase, ejection fraction, high blood pressure, and serum creatine have P<0.05* and are statistically significant. Hence these features are highly associated with heart failure.

To check the model assumptions of Cox (CPH), a chi squared test was done with p-value threshold set at 0.05, in which each feature was also ranked to test its relation to proportional hazard ratio.

Fig 14

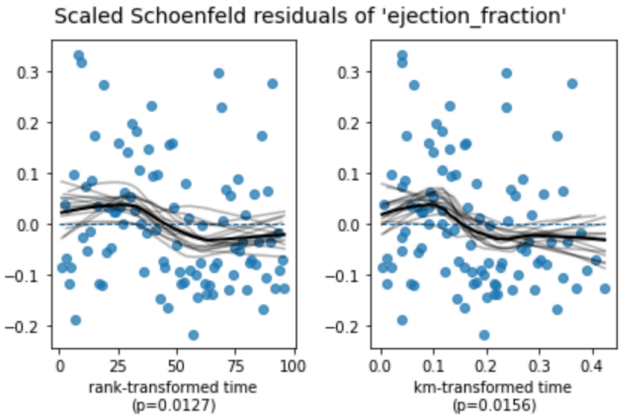
Summary Table for Cox Proportional Hazard Rates

				chi squared
degrees_of_freedom				1
model		<lifelines.CoxPHFitter: fitted with 299 total ...		
test_name				proportional_hazard_test
		test_statistic	**p**	**-log2(p)**
age	**km**	0.07	0.79	0.34
	rank	0.02	0.90	0.16
anaemia	**km**	0.00	0.99	0.02
	rank	0.01	0.93	0.10
creatinine_phosphokinase	**km**	1.10	0.29	1.77
	rank	1.07	0.30	1.73
diabetes	**km**	0.06	0.81	0.31
	rank	0.00	0.96	0.06
ejection_fraction	**km**	5.68	0.02	5.87
	rank	5.99	0.01	6.12
high_blood_pressure	**km**	0.20	0.66	0.61
	rank	0.18	0.67	0.58
platelets	**km**	0.05	0.82	0.29
	rank	0.16	0.69	0.53
serum_creatinine	**km**	3.23	0.07	3.79
	rank	3.51	0.06	4.03
serum_sodium	**km**	1.11	0.29	1.77
	rank	1.76	0.19	2.43
sex	**km**	0.06	0.80	0.32
	rank	0.15	0.70	0.52
smoking	**km**	0.77	0.38	1.40
	rank	0.52	0.47	1.08

From the table, ejection fraction was the highest ranked at 5.99, and serum creatinine at 3.51. This further indicates features that are most predictable for survival analysis associated with heart failure. Also, ejection fraction was the only feature that failed the chi squared non-proportional test with P-value of 0.0127 below the 0.05 threshold. To further investigate the violation of proportional hazards, visual plots of the scaled Schoenfeld residuals of ejection fraction were generated. Shown in fig 15.

Fig 15

Schoenfeld Residual plots



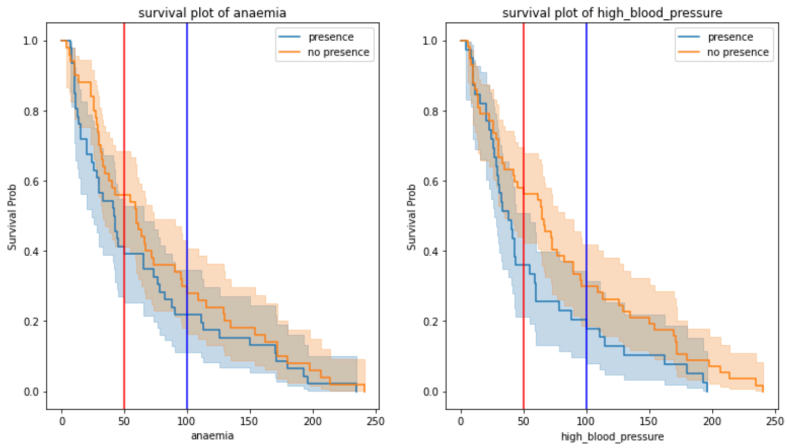
From the plots its evident that ejection fraction feature has not violated the proportional hazard assumption, the lowess line (thick line) appears to nearly represent horizontal line. Other means of checking violations of the assumption is to randomly sample the data and retrain the sample to test the assumptions. However, the case may be, several statisticians have documented in their research that light violations of the cox (CPH) assumptions may still be valid if the hazard ratios were computed based on their averages.

Kaplan Meier Curve and Analysis on Categorical Features

For categorical features, anaemia (P=0.034) and high blood pressure (P=0.028) were the features that were statistically significant below P<0.05, hence they are highly associated with patient's survival.

Fig 16

Plots for at Risk Categorical Features



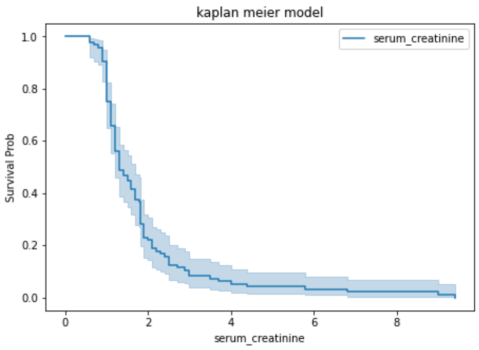
From the model plots, we used 50days to predict the survival rate for each feature. As a result, we can observe that patients with high_blood_pressure are slightly more at risk with heart failure when compared with anaemia because the survival rate curve falls below 40% on the 50th day during follow-up periods. The blue vertical line extends follow time further to the 100th day and highlights an even worse survival rate of 20%. On the 200th follow up day for high blood pressure patients, the survival rate is 0%. However, at the start of the research project, earlier indications from the data collected informed us that out of the total sample size of 299 only 43.14% of the patients had presence of anaemia (low red blood cells) and 36% had presence of high blood pressure. With this information we could apply further statistical analysis like Bayes theorem for future studies and research on medical diagnosis to reduce mortality rate of heart failure patients. The case is evident in research by (Thomas L, et al, 1998) on diagnosis with dependent symptoms: Bayes theorem and the analytical hierarchy process.

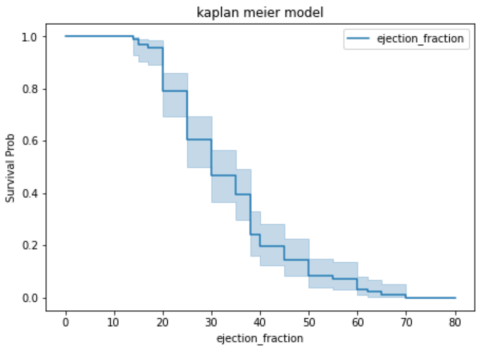
Kaplan Meier Curve and Analysis on Continuous Features

For continuous features, age (P=6.4e-07), creatinine phosphokinase (P=0.03), ejection fraction (P=3.0e-06) and serum creatinine (P=4.8e-06) were the features that were statistically significant below P<0.05, hence they are highly associated with patient's survival for heart failure. However, more emphasis was placed on Serum creatinine and Ejection fraction. Serum creatinine measures the level of creatine in the blood and creatinine is the waste product in the blood that comes from the muscles. Healthy kidneys filter creatinine out of your blood through urine, and a higher creatinine level means the kidneys are not functioning properly which may lead to hypertension and heart failure. Based on the Kaplan Meier curves in *fig 17*, there is a likelihood that patients at high creatinine levels of 0.7 - 1.3mg have a 50% chance of survival and earlier data survey indicated that 75% of the patients diagnosed with high serum creatinine had levels of 1.4 with a mean score of 1.39. Ejection fraction measures the volume of fluid ejected from heart chamber with contraction of the left ventricle and its measured is % of blood pumped out. The normal heart readings are 70%< x >55%, low readings 55%< x >40% and high readings >75% which might indicate the patient is diagnosed with heart failure conditions like arrhythmia, and hypertrophic cardiomyopathy a cause of cardiac arrest. From the Kaplan Meier curve generated, normal ejection fraction readings are barely surviving with a low rate of 10% and higher readings > 70 have a 0.01% chance survival rate.

Fig 17

Kaplan Meier Curve for at Risk Continuous Features





Conclusion

In conclusion, when comparing the significant features, Ejection fraction is the feature most associated with heart failure of the patients in this study, with the lowest survival rate of 0.01%. Additionally, Anaemia and High blood pressure begins to worsen and reduce the survival rate of patients if they are not diagnosed and treated properly along the follow time for treatment. While Serum creatinine levels can be managed as critical diagnosed patients still have a 50% survival rate.

The methods used in performing the analysis were precise and accurate; Kaplan Meier estimator and Weibull distribution used to model and predict the patient's survival rate were explicit in identifying the feature that highly associates with patient's heart failure and cox proportional hazard to check model accuracy and validation. Although there was an assumption violation while using cox proportional hazard model to train the data; the violation was investigated with Schoenfeld residuals plot visualization and rendered as insignificant.

Although previous research has not been able to establish a direct link for platelet to cause heart failure, other studies by (Delcea, et al, 2019) on cholesterol and platelets have focused on blood clot formation inside the blood vessels which can cause blockage in the blood vessels and the

inability of the vessels to supply blood to the heart or the brain leading to heart attacks, strokes, and heart failure.

References and sources of article are listed below:

Tanvir Ahmad, Et al. PLoS ONE 12(7), 0181001 (2017): *Survival analysis of heart failure patients: a case study.* https://www.researchgate.net/publication/318595531_Survival_analysis_of_heart_failure_patients_A_case_study

Delcea C., Et al (October 2019). *Low Platelets in Heart Failure: Small Cells, Important Impact on all Causes Long-term Mortality.* https://academic.oup.com/eurheartj/article/40/Supplement_1/ehz747.0353/5594427

Horacio A. (March 2017). National heart, Lung, and Blood Institute. *Hyponatremia in Heart Failure* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5385798/

Prof. Khalid Naseem (February 2017). British Heart Foundation. *Characterizing The Thrombo-inflammatory Roles of Platelet.* https://www.bhf.org.uk/research-projects/characterising-the-thromboinflammatory-roles-of-platelet-cd36

Suneel U., Et al (August 2011). National Library of Medicine. *The Effects of Heart Failure on Renal Function*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904358/

National heart, Lung, and Blood Institute (2022). *What is High Blood Pressure, Also Known as Hypertension.* https://www.nhlbi.nih.gov/health/high-blood-pressure

Penn Heart and Vascular (2022). *Ejection Fraction: What the Numbers Mean* https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/april/ejection-fraction-what-the-numbers-mean

National heart, Lung, and Blood Institute (2022). *Platelet Disorders Thrombocytopenia* https://www.nhlbi.nih.gov/health/thrombocytopenia

Richard H., Et al. (June 2021). *Overview of The Treatment of Hyponatremia in Adults* https://www.uptodate.com/contents/overview-of-the-treatment-of-hyponatremia-in-adults?search=serum%20sodium&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1

Thomas S., Luis V. (August 1998). *Diagnosis with Dependent Symptoms: Bayes Theorem and Analytic Hierarchy Process*. https://pubsonline.informs.org/doi/abs/10.1287/opre.46.4.491

Haobo Li, et al. (February 2013). *Targeting Age-Related pathways in Heart Failure.* https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.119.315889#d3616873e1

P. Connor, A. Kleyner. (2012). *Practical Reliability Engineering* Wiley 5th Edition, Southern Gate, UK. Published by Wiley & Sons Ltd