Introduction

Cardiac disease is the most common cause of death in the United States. Cardiovascular disorders include both the heart muscle and blood vessels, which are classified into coronary heart disease, cerebrovascular disease, and heart failure, among others. This paper focuses on heart failure, lifestyle choices, and associated biomarkers. There are two main types of heart failure based on ejection fraction value. Ejection fraction (EF), defined by $EF = \frac{SV}{EDV} * 100$, where (SV) stands for the blood volume pumped out of the ventricle with each contraction (the stroke volume). It is divided by the end-diastolic volume (EDV), which is the total amount of blood in the ventricle. Normal ejection fraction levels are between 50% and 75%. An ejection fraction less than 40% is known as heart failure due to reduced efficiency. When heart failure is present, but ejection fraction is normal, it is known as heart failure with preserved ejection fraction. In this study, ejection fraction and serum creatine were the most important variables predicting death. Additionally, medical data has individual variability among clinical measurements of some biological markers. This could explain the insignificance of some variables in the models; regardless, some data was dismissed.

Dataset Description

The data set contains 299 individual profiles with 13 clinical features from the Faisalabad Institute of Cardiology in Punjab, Pakistan. All 299 patients had left ventricular systolic dysfunction. Some clinical features are binary, and some are not properly defined such as the high blood pressure binary feature. Kidney problems may mask or be associated with heart dysfunction, but the study omits this data. CPK is a relevant indicator of heart tissue damage. High levels of serum creatine may indicate kidney problems, which tend to be associated with heart problems. Low levels of sodium in the body may indicate heart failure.

Feature definition

Age: In years, the age of the patient. Integer
Creatine phosphokinase (CPK): The level of the CPK enzyme in the blood (mcg/L). Integer
Ejection fraction (EF): Percentage of blood leaving the heart at each contraction. Numeric
Platelets: Measurement of the number of platelets in the blood (kiloplatelets/mL). Numeric
Time: In days, a patient's follow-up period length. Integer
Serum creatinine: The level of serum creatinine in the blood (mg/dL). Numeric
Serum sodium: the level of serum sodium in the blood (mEq/L). Integer
Anemia: The decrease of red blood cells or hemoglobin. A factor with levels No (0) and Yes (1)
High blood pressure (HBP): If a patient has hypertension. A factor with levels No (0) and Yes (1)
Diabetes: Presence of diabetes in the patient; no distinction between Type I and Type II. A factor with two levels No (0) and Yes (1)
Sex: Woman or man. A factor with two levels Woman (0) Man (1)
Smoking: If a patient smokes. A factor with two levels NoSmoke (0) YesSmoke (1)
Death event: If the patient deceased during the follow-up period. A factor with two levels Not-Deceased (0) Deceased (1)

Research Questions

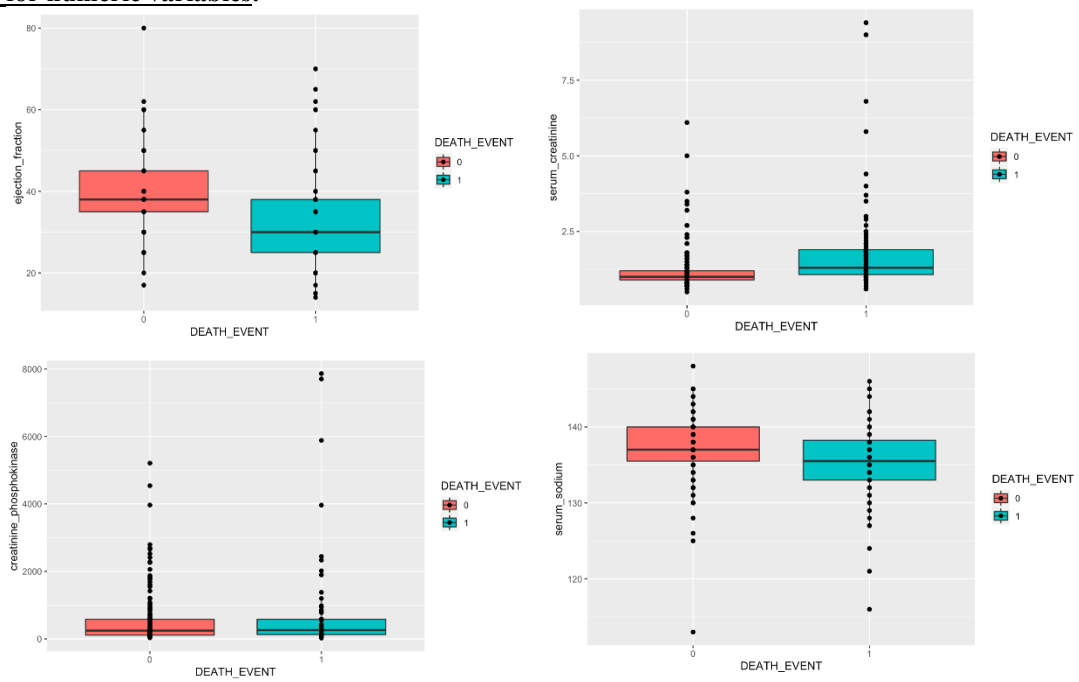
Given the 13 clinical features that contained in the data set, each feature must be analyzed to find the best indicators that associated with early death. The questions intended to be answered are:
(1) What is the relationship between each of these variables and a death event, which of these variables are associated with heart failure?
(2) If there are strong enough relationships between some of these variables and death events, can an accurate model be built to describe the relationships? Which variable is the most relevant in association with death events?
(3) Additionally, can we build a classification model to help to predict the possibility of fatality given patients' health data? And which model gives us the most accurate result?

Methodology

This paper will use multiple methodologies to answer the previous questions. Firstly, hierarchical clustering will be employed to see any patterns in data. Secondly, logistic models will be used to address the relationships found within heart failure, biomarkers, and ejection fraction. For logistic regressions, confusion matrices will be used to check for predictability, type I, and type II errors. Additionally, models with the lowest Akaike Information Criterion (AIC) will be given preference. Thirdly, linear discriminant analysis will be used to present additional ways of classification.

Data preview

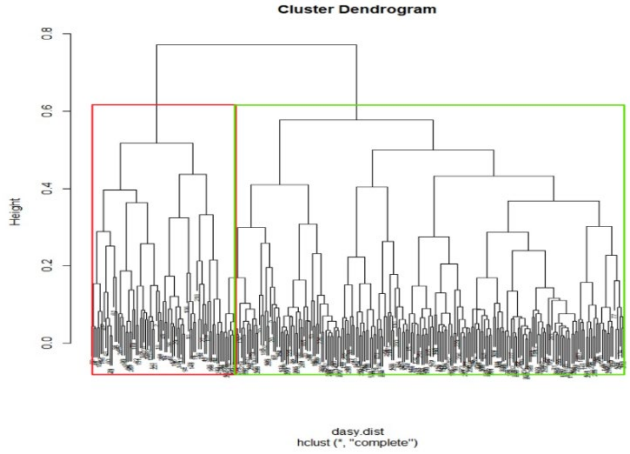
Boxplot for numeric variables:



The significance level of a death event seems to have different variables: Ejection fraction, serum creatine, creatine phosphokinase, and serum sodium. The significance of these variables will be checked later with regression models.

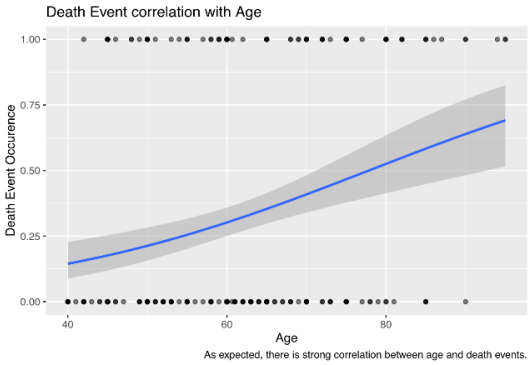
Hierarchical Clustering

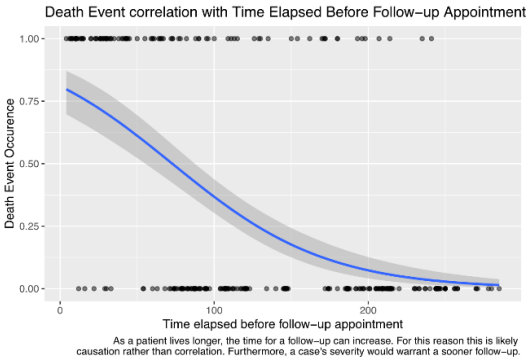
Hierarchical clustering allowed for exploration of patterns for data used in this study. Since the dataset contained both numerical and categorical variables, the distance was measured using the *daisy* built-in function in *R* that uses the metric *Gower Distance*, which standardizes the data and stratifies categorical variables. Hierarchical clustering of the data was used to discriminate the features of the data. Thus, it found the numerical differences in the variables for patients who survived heart failure and patients who did not. The output of hierarchical clustering is a dendrogram shown below. The dendrogram elucidates two main clusters in the dataset, which are enclosed in red, and green boxes. Unsurprisingly, these two clusters show the patients who did not die from heart failure and those that did. It is important to note that hierarchical clustering had an accuracy of 93.15 %, implying an error rate of 6.85%. The features showing a clear numerical difference between the two clusters were ejection fraction and serum creatinine, suggesting their importance in predicting the occurrence of death of a patient who experiences heart failure. The patients who did not survive had roughly a 9% lower ejection fraction as the patients that did survive



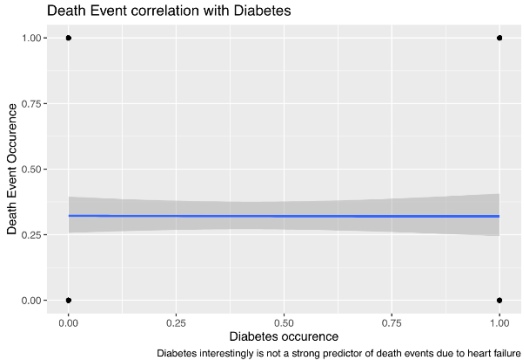
General Linear Models per individual variable:

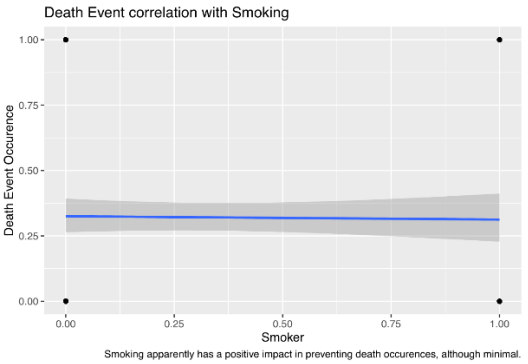
A main objective of this analysis is to predict death related to heart failure. Hence, it is valuable to plot individual variables against death occurrences to see general trends between them. Firstly, an obvious prediction is that as a patient's age increases, so do their chances of death. This is confirmed by the general linear model graph of death event in correlation with age. A second obvious correlation is time elapsed before a follow up appointment. As expected, as a patient's condition is less severe, their follow up appointments can be postponed to a later date. Thus, some of these variables weren't as tested as they provided not enough relevancy.



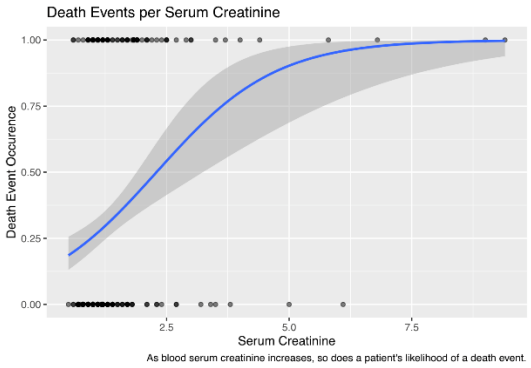


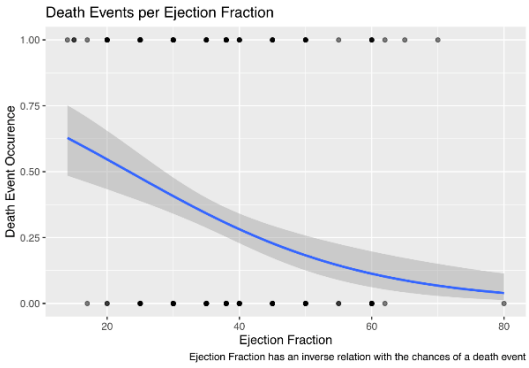
Two observations were made comparing diabetes and smoking respectively with fatal heart failure. Both seemed to have little to no effect on the mortality of heart failure. Patients who smoked had a small decrease in the fatality of heart failure probably due to the high individual variability in medical studies.





The two most significant variables, serum creatinine and ejection fraction, both provide a reliable basis for predicting fatality of heart failure.





Supervised Learning Classification

Functions

It is repetitive and time consuming to calculate misclassification error rates and confusion matrices for each model. Instead, a function that calculates a confusion matrix and accuracy of the model was made. All the confusion matrices and models' performances are provided by this function which takes both training data and test data.

Logistic Regression Model

After previewing the data in the plots above, a logistic regression was made to create a logistical model. Logistic regressions have the form of $e^{\mathbf{b_0}+\mathbf{b_1x_1}+\ldots+\mathbf{b_n}x_n}$, whose probability is between 0 and 1. The data was subset into an 80% to 20% training-to-testing sampling, proportionally to the population in terms of dead to living. The estimated parameters and significant levels for each variable are:

```
##
## Call:
## glm(formula = DEATH_EVENT ~ ., family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1861  -0.4524  -0.1911   0.3474   2.7597
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.520e+00  7.123e+00   0.635 0.525699
## age                       5.370e-02  1.963e-02   2.735 0.006239 **
## anaemia                  -1.339e-01  4.428e-01  -0.302 0.762288
## creatinine_phosphokinase  1.552e-04  2.383e-04   0.651 0.514914
## diabetes                  2.043e-01  4.236e-01   0.482 0.629531
## ejection_fraction        -9.339e-02  2.174e-02  -4.295 1.74e-05 ***
## high_blood_pressure       4.771e-02  4.343e-01   0.110 0.912517
## platelets                -1.932e-06  2.173e-06  -0.889 0.373978
## serum_creatinine          6.916e-01  2.006e-01   3.448 0.000565 ***
## serum_sodium             -1.483e-02  5.039e-02  -0.294 0.768549
## sex                      -4.512e-01  5.033e-01  -0.897 0.369964
## smoking                  -1.234e-01  5.113e-01  -0.241 0.809312
## time                     -2.642e-02  3.935e-03  -6.713 1.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 301.20  on 239  degrees of freedom
## Residual deviance: 155.08  on 227  degrees of freedom
## AIC: 181.08
##
## Number of Fisher Scoring iterations: 6
```

Confusion Matrix:	True Alive (0)	True Dead (1)
Predicted Alive (0^)	20	1
Predicted Dead (1^)	20	18

The misclassification rate for this model is 0.3559322, with an accuracy of 0.6441. Another logistic regression is made with the two most significant variables, serum creatinine and ejection fraction. This produced an even more accurate model. An ANOVA test was run on both models, indicating that using just serum creatinine and ejection fraction as predictors provides a more precise model. The model is as follows:

```
## Call:
## glm(formula = DEATH_EVENT ~ ejection_fraction + serum_creatinine,
##     family = binomial, data = data_train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.299  -0.801  -0.628   1.054   2.294
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.61435    0.61560   0.998    0.318
## ejection_fraction -0.06606    0.01613  -4.096 4.21e-05 ***
## serum_creatinine   0.72109    0.18376   3.924 8.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 301.20  on 239  degrees of freedom
## Residual deviance: 260.44  on 237  degrees of freedom
## AIC: 266.44
##
## Number of Fisher Scoring iterations: 4
```

```
##                      True Death Events
## Predicted Death Events  0  1
##                  ALIVE 38 11
##                  DEAD   2  8
```

```
## [1] "Misclassification Error Rate"
```

```
## [1] 0.220339
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	227	155.0841	NA	NA	NA
2	237	260.4401	-10	-105.356	4.593313e-18

Confusion Matrix:	True Alive (0)	True Dead (1)
Predicted Alive (0^)	38	11
Predicted Dead (1^)	2	8

The misclassification rate for this model is 0.2203, with an accuracy of 0.7797, and an AIC of 266.44.

The ANOVA test for the new model versus the original full model shows that there is a significant improvement in the new model.

```
## Call:
## glm(formula = DEATH_EVENT ~ ejection_fraction + serum_creatinine,
##     family = binomial, data = data_train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.299  -0.801  -0.628   1.054   2.294
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.61435    0.61560   0.998    0.318
## ejection_fraction -0.06606    0.01613  -4.096 4.21e-05 ***
## serum_creatinine   0.72109    0.18376   3.924 8.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 301.20  on 239  degrees of freedom
## Residual deviance: 260.44  on 237  degrees of freedom
## AIC: 266.44
##
## Number of Fisher Scoring iterations: 4
```

```
##                      True Death Events
## Predicted Death Events  0  1
##                 ALIVE 38 11
##                 DEAD   2  8
```

```
## [1] "Misclassification Error Rate"
```

```
## [1] 0.220339
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	227	155.0841	*NA*	*NA*	*NA*
2	237	260.4401	-10	-105.356	4.593313e-18

A third logistic regression was performed using only ejection fraction, and it had the same misclassification rate as the previous model of 0.2203, but it had a greater AIC of 286.48. Thus, the previous model with a lower AIC was favored. The final most accurate prediction model with the lowest AIC only used ejection fraction and serum creatinine as significant predictors: $e^{\mathbf{0.61435-0.06606x_1+\ 0.72109}x_2}$.

Linear Discriminant Analysis (LDA) :

In addition to logistic regressions, a LDA method can be used to classify death events from the dataset.
Firstly, a *full model* including all variables is made:

```
Call:
lda(DEATH_EVENT ~ ., data = heartdata_train)

Prior probabilities of groups:
        0         1
0.6791667 0.3208333

Group means:
       age  anaemia1 creatinine_phosphokinase diabetes1 ejection_fraction high_blood_pressure1 platelets
0 58.49489 0.4171779                563.9816 0.4233129          39.77301            0.3006135  267221.1
1 65.15152 0.4285714                666.4026 0.4805195          33.33766            0.4155844  261258.0
  serum_creatinine serum_sodium      sex1  smoking1
0         1.192699     137.0123 0.6441718 0.3312883
1         1.805714     135.9481 0.6103896 0.3116883
```

	Prediction "0"	Prediction "1"
In fact "0"	33	7
In fact "1"	11	8

From the model, a confusion matrix is made with an accurate classification rate of 0.6949153

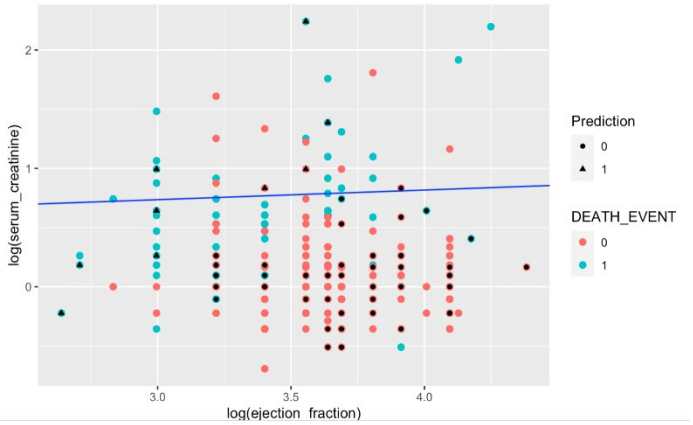
A second LDA is made with only ejection fraction and serum creatine; the *2s model* is:

```
Call:
lda(DEATH_EVENT ~ ejection_fraction + serum_creatinine, data = heartdata_train)

Prior probabilities of groups:
        0         1
0.6791667 0.3208333

Group means:
  ejection_fraction serum_creatinine
0          39.77301         1.192699
1          33.33766         1.805714
```

	Prediction "0"	Prediction "1"
In fact "0"	38	2
In fact "1"	11	18

The accuracy classification rate is 0.779661.

The last model has the best performance. The *2s model* is plotted below.



Logistical values are applied to the plot for an easier view. The decision boundary for the model has a slope of 0.081513 and an intercept of 0.4911746. The graph shows the prediction performance around the boundary area is less than the performance in areas where there are more clustering patterns of predictions interlaced with death events.

K-Fold Cross Validation for LDA:

Since we only use one data set to train and test the previous LDA models, it is not enough to get a general performance of the models. A K-fold cross validation can help view the average performance of each model and since the proportion of death events is around one-to-four for dead and alive, a 5-fold cross validation for the *2s model* is constructed.

The model in which death event is predicted by ejection fraction and serum creatinine:

```
Linear Discriminant Analysis

299 samples
  2 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 239, 238, 240, 239, 240
Resampling results:

  Accuracy   Kappa
  0.7460137  0.3141002
```

For the full model:

```
Linear Discriminant Analysis

299 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 239, 240, 239, 240, 238
Resampling results:

  Accuracy   Kappa
  0.7219533  0.2782093
```

By comparing the performance of both models, the model with only serum creatine and ejection fraction produces the higher classification rates and a higher interrater reliability (kappa index is higher).

Interpretation of results

(1) What is the relationship between each of these variables and a death event, which of these variables are associated with heart failure?

Briefly analyzing data with box plots and hierarchical clustering, it is shown that there are some strong relationships between death, serum creatine and low ejection fraction.

(2) If there are strong enough relationships between some of these variables and death events, can an accurate model be built to describe the relationships? Which variable is the most relevant in association with death events?

By discarding age and the time for the next medical appointment, serum creatine and ejection fraction help were shown to have strong correlation with death. Such relationships were shown to be model using logistic regressions and LDAs, where it was shows they were the best predictors.

(3) Additionally, can we build a classification model to help to predict the possibility of fatality given patients' health data? And which model gives us the most accurate result?

Given that serum creatine and ejection fraction were the strongest predictors for death, the logistic model containing both creatine and ejection fraction, $e^{\mathbf{0.61435 - 0.06606x_1 + \ 0.72109}x_2}$, was shown to be better compared to the logistic model using all variables. Likewise, the LDA using only serum creatine and ejection fraction was shown to better than the LDA containing all possible variables from the data set.

Conclusion

Heart disease, and specifically, heart failure is a highly deadly disease as shown by the number of deaths associated with a low ejection fraction and serum creatine. Ejection fraction and serum creatine were found to be the strongest biomarkers that predicted death in the patients. While age and time for medical appointments also had strong death predictors, it was of little use to have them be part of the models as context is important. An aging body can have multiple diseases, besides cardiac problems, that ultimately leads to death. Medical appointment time is very unlikely to be the cause of death. Rather, it probably implies people who more urgently need earlier medical appointments as their condition is grave.

Further complications from diabetes, blood pressure, and smoking data showed very weak correlations. This is in part due to how these variables were classified as binary categories instead of measured data. Additionally, there is a high degree of variability in individual medical data which could lead to weak predictive models.

After using data boxes and clustering to see data patterns, serum creatine and ejection fraction were coming up as the most prominent variables. Running logistic models and LDAs that exclusively only used serum creatine and ejection fraction were shown to be the best at predicting death.

Heart disease will continue to be a major cause of death in the US and abroad in both the short and long term. Predictive models that can show which patients are more at risk will be highly valuable to assets. It will help doctors determine a patient's overall risk of death. It also can potentially help researchers to create treatments that target the main reasons behind heart failure.

References

Heart failure clinical records. (2020). UCI Machine Learning Repository.