

Egypt University of Informatics

Computer and Information Systems

Data Analysis Course

Analyzing Heart Failure Clinical Records

Submitted by: Mohamed Mohamed Hussien Ahmed [22-101158]  
Seif Eldin Amr Mahmoud Aly Elzeiny [22-101213]  
Mohamed Wael Mohamed Habib [22-101129]  
Karim Alaa Salama Aly Salama [22-101268]  
Abdelkareem Ahmed Abdelkareem Mohamed Soliman [22-101251]

Under the Supervision of: Dr. Mohamed Taher Alrefaie

5/24/2024

# Introduction

Understanding the determinants of death events in patients is crucial for improving clinical outcomes and public health strategies. This study focuses on analyzing various clinical parameters and demographic factors to identify significant predictors of mortality. Using a robust dataset and statistical techniques, we aim to uncover meaningful insights that can guide healthcare decision-making and intervention planning.

# Research Question

1-Serum Creatinine Levels and Death Event: What is the nature of the relationship between serum creatinine levels and the likelihood of experiencing a death event? Can higher serum creatinine levels predict the likelihood of experiencing a death event?

2-Ejection Fraction and Death Event: How does the ejection fraction relate to the likelihood of experiencing a death event? Can lower ejection fractions predict the likelihood of experiencing a death event?

3-Sex and Smoking Habits: How do smoking habits vary between sexes, and are there significant differences that impact heart failure outcomes? Are smoking habits significantly different between sexes, and do these differences impact heart failure outcomes?

# Hypothesis

1-There is a correlation between higher serum creatinine levels and the likelihood of experiencing a death event.

Question: Can higher serum creatinine levels predict the likelihood of experiencing a death event?

2-There is a correlation between lower ejection fraction (a measure of heart function) and the likelihood of experiencing a death event.

Question: Can lower ejection fractions predict the likelihood of experiencing a death event?

3-There is a correlation between sex and smoking habits, with differences in smoking prevalence and behavior potentially impacting heart failure outcomes.

Question: Are smoking habits significantly different between sexes, and do these differences impact heart failure outcomes?

Population of Interest:

The population of interest includes individuals from the dataset who have various health and lifestyle attributes recorded, such as age, anemia status, creatinine phosphokinase levels, diabetes status, ejection fraction, high blood pressure status, platelet count, serum creatinine levels, serum sodium levels, sex, smoking status, and whether they experienced a death event.

Sampling Method:

The dataset used for this analysis was obtained from Kaggle and includes a comprehensive collection of clinical records related to heart failure. Each record represents an individual with various health metrics and outcomes. The dataset includes 299 samples, which provides a diverse and representative sample of individuals with different health conditions and lifestyles.

Bias Identification:

If we were to make any inferences from this dataset, since it is limited by a single year, it would not take into consideration any occurring changes through the years. Additionally, the dataset might have inherent biases due to the demographic or geographic distribution of the subjects, which could affect the generalizability of the findings.

Survey Questions/Collected Data/Dataset:

We utilized a Kaggle dataset that includes detailed clinical records for each subject, covering attributes such as age, anemia status, creatinine phosphokinase levels, diabetes status, ejection fraction, high blood pressure status, platelet count, serum creatinine levels, serum sodium levels, sex, smoking status, and whether they experienced a death event.

Each row represents a singular individual, their health metrics, and their outcome regarding heart failure. This dataset served as the basis for our analysis.

Number of samples used: 299

Kaggle Dataset Link: [Heart Failure Clinical Records Dataset](https://www.kaggle.com/datasets/nimapourmoradi/heart-failure-clinical-records/code)

Improvements:  
Null Data: We used a dataset that did not have any null data, ensuring that all records were complete and usable without the need for imputation or exclusion of samples.

Data Cleaning: We cleaned up the dataset by removing columns that were deemed unnecessary for our analysis. For example, we excluded the 'time' column, which represented the follow-up period, as it was not directly relevant to our research questions and hypotheses.

Analysis:

|  |
| --- |
|  |
| **This heatmap was used to identify possible correlations between our columns we can see the different correlations between (ejection fraction) and (serum creatinine) with (Death event)** |

|  |
| --- |
|  |
| **This box plot shows distribution of serum creatinine levels for individuals who experienced a death event vs. those who did not.** |

|  |
| --- |
|  |
| **This box plot Distribution of ejection fractions (**percentage of blood leaving the heart at each contraction**) for individuals who experienced a death event vs. those who did not.** |

|  |
| --- |
|  |
| **This bar chart shows the comparison of smoking status (number of smokers vs. non-smokers) between males and females.** |

|  |  |
| --- | --- |
|  |  |
| **A positive slope in the scatter plot indicates that as serum creatinine levels increase, the likelihood of experiencing a death event also increases.** | **A negative slope in the scatter plot indicates that as ejection fraction decreases, the likelihood of experiencing a death event increases.** |

**Interpretation of Correlation Scores and P-Values**

**Hypothesis 1: Serum Creatinine and Death Event**  
**Point-Biserial Correlation Coefficient**: 0.294  
This value indicates a moderate positive correlation between serum creatinine levels and the likelihood of experiencing a death event. A positive value means that higher serum creatinine levels are associated with a higher likelihood of death events.  
**P-value**: 2.19e-07  
This extremely low p-value (much less than 0.05) indicates that the correlation observed is statistically significant. There is very strong evidence to reject the null hypothesis (which states there is no correlation), suggesting that higher serum creatinine levels are indeed significantly associated with a higher likelihood of death events.

**Hypothesis 2: Ejection Fraction and Death Event**  
**Point-Biserial Correlation Coefficient**: -0.269  
This value indicates a moderate negative correlation between ejection fraction and the likelihood of experiencing a death event. A negative value means that lower ejection fractions are associated with a higher likelihood of death events.  
**P-value**: 2.45e-06  
This very low p-value (again, much less than 0.05) indicates that the correlation observed is statistically significant. There is strong evidence to reject the null hypothesis, suggesting that lower ejection fractions are indeed significantly associated with a higher likelihood of death events.

**Hypothesis 3: Sex and Smoking**  
**Correlation Coefficient**: 0.45  
This value indicates a moderate to strong positive correlation between sex and smoking habits. This suggests that there is a notable association between a person's sex and their smoking behavior.  
**P-value**: 5.18e-16  
This extremely low p-value (far less than 0.05) indicates that the correlation observed is highly statistically significant. There is overwhelming evidence to reject the null hypothesis, indicating that smoking habits are indeed significantly different between sexes.

### Summary:

From our analysis, the correlation scores and p-values indicate the following:

**Serum Creatinine and Death Event**:

* **Finding**: Higher serum creatinine levels are moderately associated with an increased likelihood of death events.
* **Statistical Significance**: This correlation is highly statistically significant (p-value ≈ 0.000000219), supporting the hypothesis that serum creatinine levels can predict death events.

**Ejection Fraction and Death Event**:

* **Finding**: Lower ejection fractions are moderately associated with an increased likelihood of death events.
* **Statistical Significance**: This correlation is highly statistically significant (p-value ≈ 0.00000245), supporting the hypothesis that ejection fraction levels can predict death events.

**Sex and Smoking Habits**:

* **Finding**: There is a moderate to strong association between sex and smoking habits, suggesting that smoking behavior significantly differs between sexes.
* **Statistical Significance**: This correlation is highly statistically significant (p-value ≈ 0.0000000000000518), supporting the hypothesis that smoking habits differ between sexes and these differences may impact heart failure outcomes.

These findings indicate robust statistical evidence supporting the proposed hypotheses, with potential implications for public health strategies targeting heart failure risk factors.

# Conclusions

Our analysis reveals significant correlations between key health indicators and heart failure outcomes, with implications for public health interventions. By focusing on modifiable risk factors such as serum creatinine levels, ejection fractions, and smoking behaviors, particularly across different sexes, health policies can be better tailored to prevent heart failure and improve patient outcomes.

# Potential Issues

1. **Dataset Limitations**: The dataset is limited to a single year, which might not account for changes over time. Longitudinal data could provide more robust insights.
2. **Bias**: The dataset may contain demographic or geographic biases that affect the generalizability of the findings.
3. **Confounding Variables**: There may be other confounding variables not accounted for in the dataset, which could influence the observed correlations.