1. What is the general domain/subject area of this project?

This project will focus on the area of cardiovascular disease (CVD). In this study, we aim to study the association between serum biomarkers and mortality rates among patient who has been diagnosed heart failure.

2. What data will you use, and what is the source?

This project will be based on medical records of 299 patients of heart failure, with left ventricular systolic dysfunction. The dataset was collected and analyzed in by Ahmad et al.’s. Then in January 2020, the dataset was elaborated and donated to the University of California Irvine (UCI) Machine Learning Repository by Davide Chicco under the same Attribution 4.0 International (CC BY 4.0) copyright. We downloaded this dataset from the UCI Machine Learning Repository and will use it under the same Attribution 4.0 International (CC BY 4.0) copyright. The dataset can be accessed through <https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>.

3. What primary questions will you seek to answer?

The primary question we aim to answer in this project is the association between serum creatine, a biomarker indicating renal dysfunction, and mortality risks for patients with heart failure.

4. What secondary questions will you seek to answer

We also consider several secondary questions in this project:

1. The association between serum sodium and mortality risks for patient with heart failure.
2. The association between platelets value and mortality risks for patient with heart failure.

5. What outcome(s)/endpoint(s) will you use?

* The primary outcome in this study is the survival time of the patients with heart failure.
* We also consider the endpoint of survival proportion 30 days after following up (or Death 30-day in the below texts). The outcome is defined as death (0 for alive patient, and 1 for dead patient) at the 30 days after following up, which is a binary outcome.

6. Statistical Analysis Plan

1. Data cleaning: although the dataset has been elaborated by Davide Chicco, we will check any missing data in the outcome, exposure, and covariates. We will report the number of missing data and if the number is less than 10%, we will consider including a “missing” indicator for categorical variables and imputing continuous variables. We will report the final number of patients included in our study.
2. Checking primary exposure and secondary exposures:

The primary exposure, serum creatine, which is continuous, is usually categorized into two different levels (≤ 1.5 normal level vs, > 1.5 abnormal level). In this project will first treat the serum creatine as a continuous variable and calculate its sample mean, standard deviation, median, and range. Then we will treat it as a binary variable and calculate the proportion of normal and abnormal level in patients.

The secondary exposure, serum sodium will be firstly treated as a continuous variable with sample mean, standard deviation, median, and range. Then we will categorize it into four groups based on quantiles.

The secondary exposure, palettes will be firstly treated as a continuous variable with sample mean, standard deviation, median, and range. Then we will categorize it into four groups based on quantiles.

1. Checking outcomes:

We will calculate the average person-time until death in normal serum creatinine group and abnormal serum creatinine group. We will also calculate the Death 30-day in these two groups. The average person-time until death as well as the Death 30-day in serum sodium groups and platelet groups will also be calculated and reported. Survival plots will be made to visualize the mortality rates in different groups. Chis-squared test will be used to test the difference of Death 30-day in these groups.

1. Checking other covariates.

In this project, age group (categorical), sex (male vs. female), c (yes vs. no), diabetes (yes vs. no), ejection fraction (≤ 30, 31-44, and ≥ 45), and smoking (yes vs. no) will be considered as covariates. High blood pressure (yes vs. no) is in the datasets, but the diagnose criteria is unclear, so we decide not to include this into our analysis. Creatinine phosphokinase value which the author called as CPK is also included in the dataset. However, after reading other literature, the CPK usually refers to creatine phosphokinase. Due to this inconsistency, we decide not to include in variable into our analysis, either.

The proportions of this categorical covariates in normal serum creatinine group and abnormal serum creatinine group will be calculate and compared using chi-squared tests.

1. Modeling

Firstly, we will do a crude model to assess the association between serum creatine (normal vs. abnormal) and mortality rates using Cox proportional-hazards model. The outcome is the survival time with event and the only predictor is serum creatine (normal vs. abnormal). Secondly, based on the crude model we include other covariates to check if the confounding exists. Last, we will include the other two secondary variables and check the association between serum creatine (normal vs. abnormal) and mortality rates. The same process will be conducted again by replacing serum creatine (normal vs. abnormal) with continuous serum creatine.

Similar model strategies will be utilized for secondary exposures by replacing serum creatine in the main analysis with serum sodium and platelets in turns.

1. Checking effect modification.

To check whether effect modification exists, we will include an interaction term between serum creatine (normal vs. abnormal) and age (<65 vs. ≥65), sex, diabetes, ejection fraction and smoking in turns into the model. The p-value of the interaction term will be used to determine if there are any effects modification. The same process will be applied for continuous serum creatine and the other two secondary exposures.

1. Checking nonlinearity

To check whether the relationship between serum creatine and mortality risks is linear, we will use replace the linear term in the fully adjusted model with a natural spline of serum creatine. The knots and degrees of freedom of the spline will be determined during the following analysis. The same process will be applied to the other two secondary exposures

7. What are the biggest challenges you foresee in answering your proposed questions and completing this project.

We foresee the biggest problem may be the violation to the proportional hazards assumption. Also, we are concern that the limited number of sample size may not allow us to include such a number of predicators in the model.

8. Will you seek domain expertise? Why or why not? If so, from whom?

Yes. Because this project is very close linked to the clinical medicine especially the CVD and the renal diseases, we need to discuss with experts in this area. However, given the limited time of this project and the limited resources, we would like to read related articles and to learn background knowledge first. Then, we will put unresolved questions into the discussion board on Canvas to see if any classmates can help us. We believe our class incorporates so many students from different discipline related to human health. There must be somebody who can help us.

9. What software package(s) will you use to complete this project? (It is absolutely fine for different group members to use different packages; in fact, some tasks are easier in some packages over others and vice versa.).

* R 4.1.1 (R core team) will be used to manipulate data and to do model analysis
* Tidyverse package will be loaded to make out work easier.
* Survival package will be used to do Cox proportional-hazard model.
* ggfortify and gglot2 will be used to visualize our results.