HEART FAILURE DATA ANALYSIS

Akhil A. Naik

Contents

1 Introduction

Globally every year cardiovascular diseases are responsible for killing over 17 million people. When the heart is unable to pump enough blood to meet the needs of the body, heart failures occur. The modern technology has enabled hospitals to record medical data of patients to quantify symptoms, clinical test values, body features etc. and present it to analysts and scientists for analysing patterns and correlations which would otherwise be undetectable by doctors. Through analysing medical records for these features, survival of a patient predicted or effect of a disease on a patient can be studied among many other uses. Due to the stressful modern life and less health consciousness among people along with many such reasons, heart failures have been on the rise. To prepare for the adverse affects and knowing how patients can prolong their lives has been the hour of need among the scientists.

Knowing what can effect the survival of a patient who has been diagnosed with heart failure, can impact the survival/life of a patient significantly, allowing them to prepare themselves to avoid things which can have adverse effect on their health and allowing them to live longer life which otherwise would not be possible. With motivation of studying features that can affect survival of a patient who has been diagnosed with heart failure, the objective is to analyse heart failure clinical records data set present in the UCI Machine Repository (Davide Chicco 2020) which was donated in 2020. Finding the significant features and performing survival analysis using relevant techniques is the key task ahead.

2 Exploratory Data Analysis

The data set contains the medical records of 299 patients (observations) who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features/attributes (Davide Chicco 2020).

2.1 Attribute Information

1. age: age of the patient (years)
2. anaemia: decrease of red blood cells or hemoglobin (boolean) (1-true,0-false)
3. high blood pressure: if the patient has hypertension (boolean) (1-true,0-false)
4. creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
5. diabetes: if the patient has diabetes (boolean) (1-true,0-false)
6. ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
7. platelets: platelets in the blood (kiloplatelets/mL)
8. sex: woman or man (binary)
9. serum creatinine: level of serum creatinine in the blood (mg/dL)
10. serum sodium: level of serum sodium in the blood (mEq/L)
11. smoking: if the patient smokes or not (boolean) (1-true,0-false)
12. time: follow-up period (days)
13. [target] death event: if the patient deceased during the follow-up period (boolean) (1-died,0-survived)

2.2 Data Manipulation and Visualisations

Our first step is to import the data set and prepare it for further analysis. The data is available in the *heart_failure_clinical_records_dataset.csv* file, with comma separated values. There are a total of 6 categorical features with rest being continuous. Note that categorical variables are present as characters, so must be converted to factors for getting levels. There are no missing values present in the data. Our response variable is Death_event which indicates patient died (1) and patient survived (0) during the follow up period after patient was diagnosed with heart failure. There are few outliers present however are considered for the analysis.

We need to analyse how our target feature Death Event is affected by all the rest of the explanatory features. The Correlation Heatmap (Fig 1) gives us a general overview of how the features are correlated with one another. However note that since the categorical features were converted to numerical for correlation purpose, we will have to analyse the categorical feaures individually. However we can say death of patient during the follow up period is related to age of the patient, ejection fraction, level of serum creatinine in the blood and the follow up period.

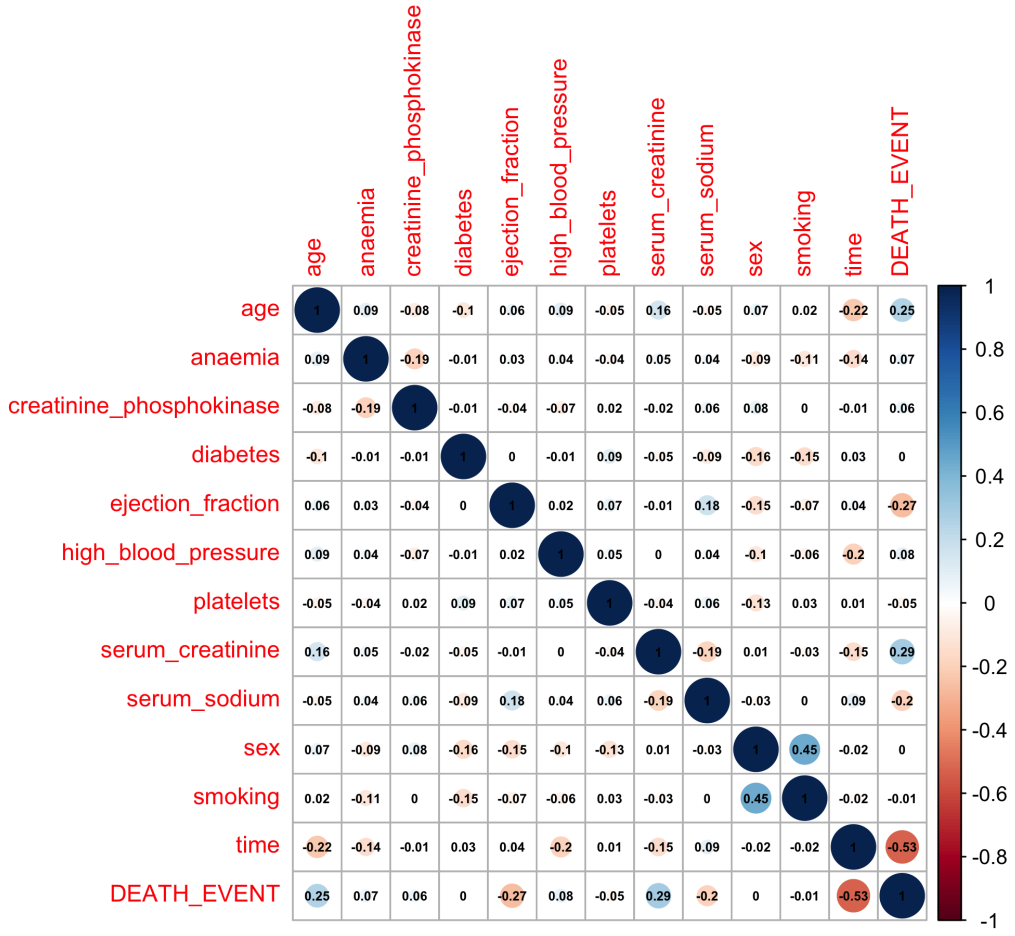

Figure 1: Correlation Heatmap

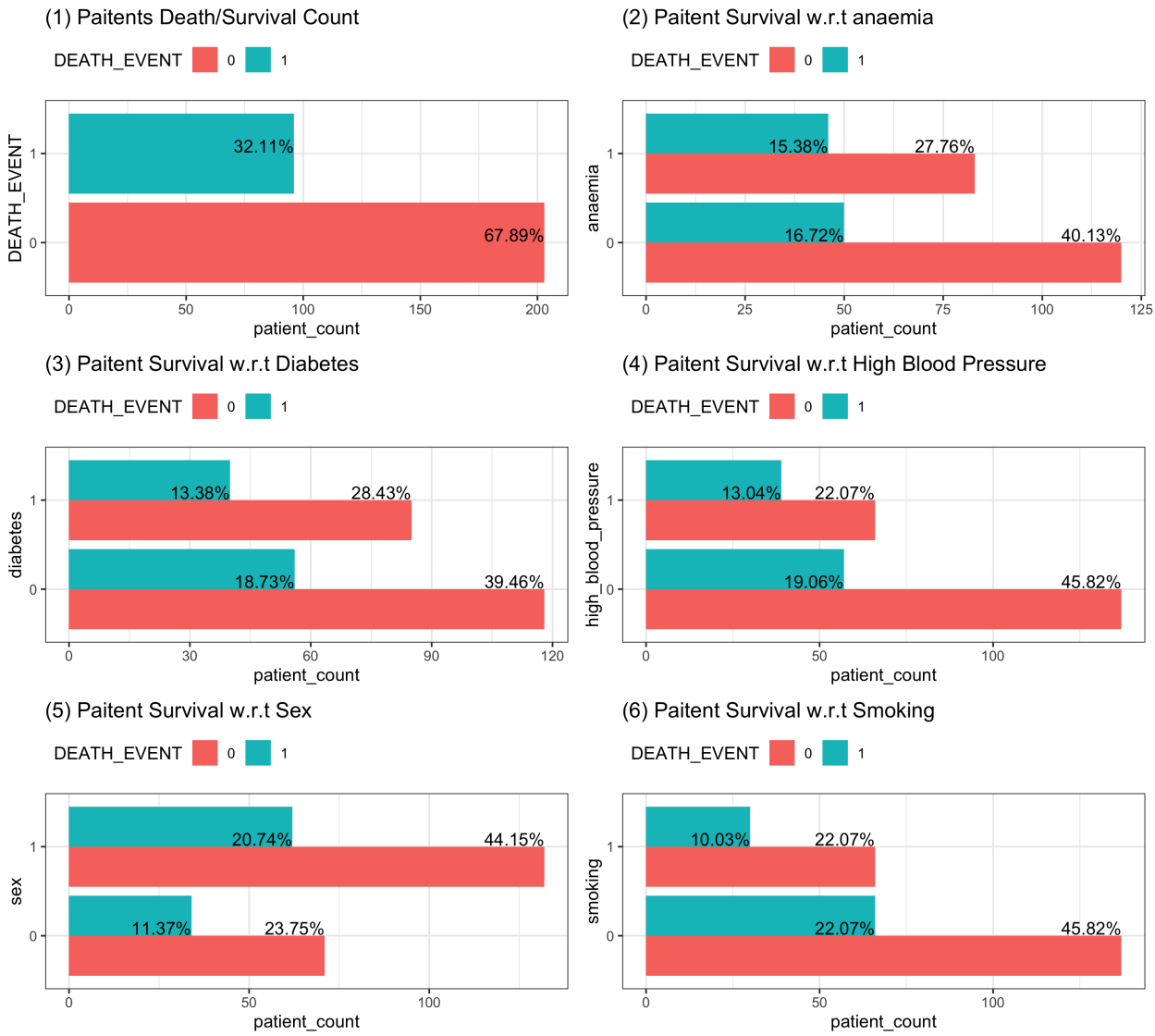

Figure 2: Patients Survival/Death variations w.r.t the categorical features

As we can see from the above figure (Fig 2), of the total patients, 32% didn't survive during the follow up period. Out of the total who didn't survive, 15% were suffering with anaemia, while 13% of patients had diabetes. Out of the total, 10% who died were smokers. Also 65% of the total patients were male and out of the total, patients who didn't survive constituted of more male than female. Let us see how age of a patient also affect the death event.

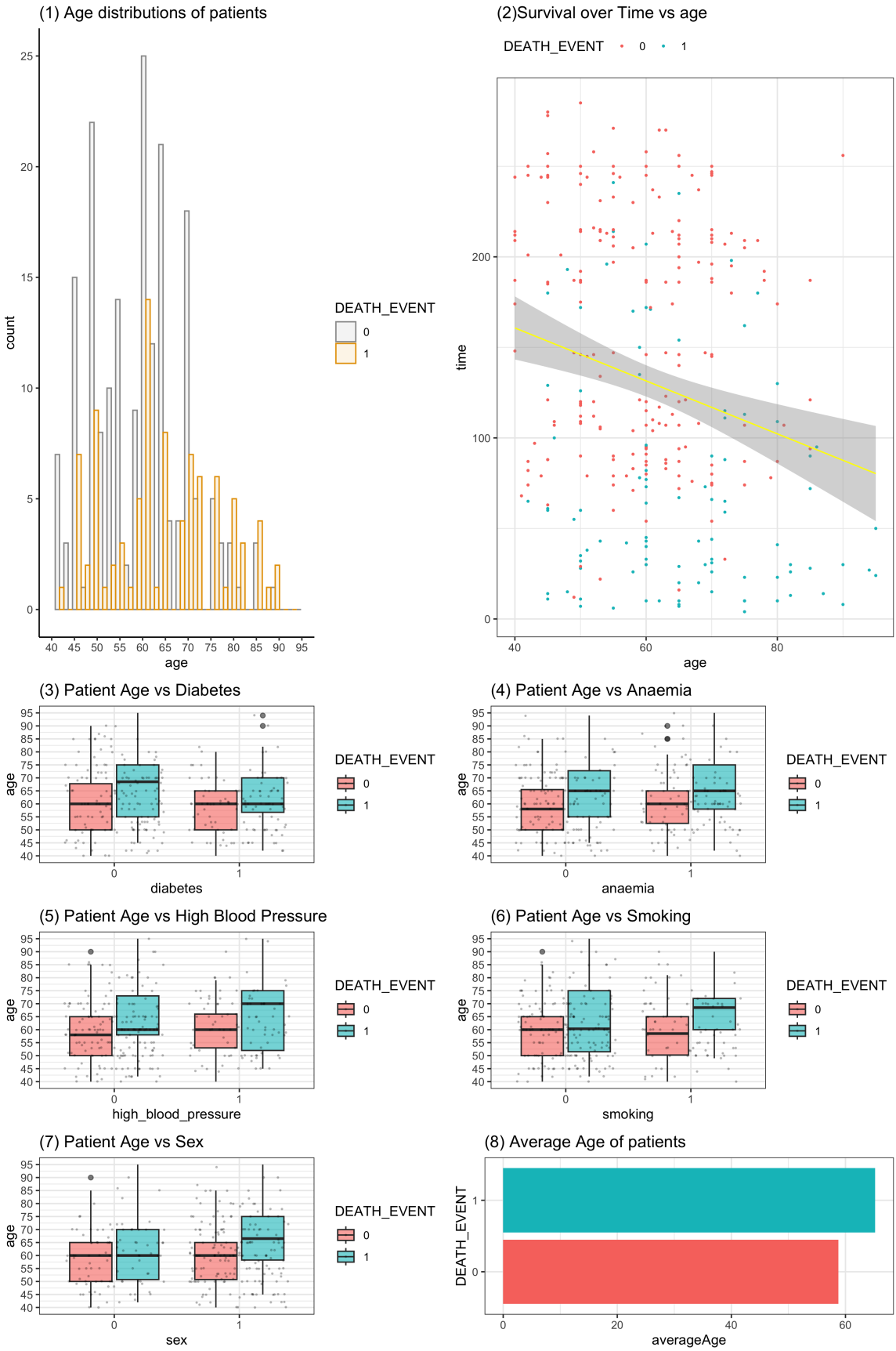

Figure 3: Patients Survival/Death variations w.r.t the categorical features associated with pateints age along with age vs time.

As observed from above figure (Fig 3), majority of patients who died, fall between age group of 50-75 with an average age of above 60. Age along with the other categorical feature play an important role in determining survival factor of a patient. For example person with a heart failure, above 60 years old, who is a smoker or has anemia or is diabetic is likely to die where as a patient below age 60 who is a smoker or diabetic is likely to survive. But these visual observations on their own can't give us a better sense in to estimating what can affect the survival of a patient.

Since we know the time feature i.e. the follow on period after patient is diagnosed with heart failure, we can use Kaplan-Meier estimator (Zabor 2019) which is a survival analysis technique through which we can estimate if an individual in a population can survive beyond a given time. It is often used to evaluate treatment effectiveness or disease risks over time. Let us estimate this based on each category to find if there is sufficient difference between the survival probabilities for each level of each category.

2.3 Survival Analysis

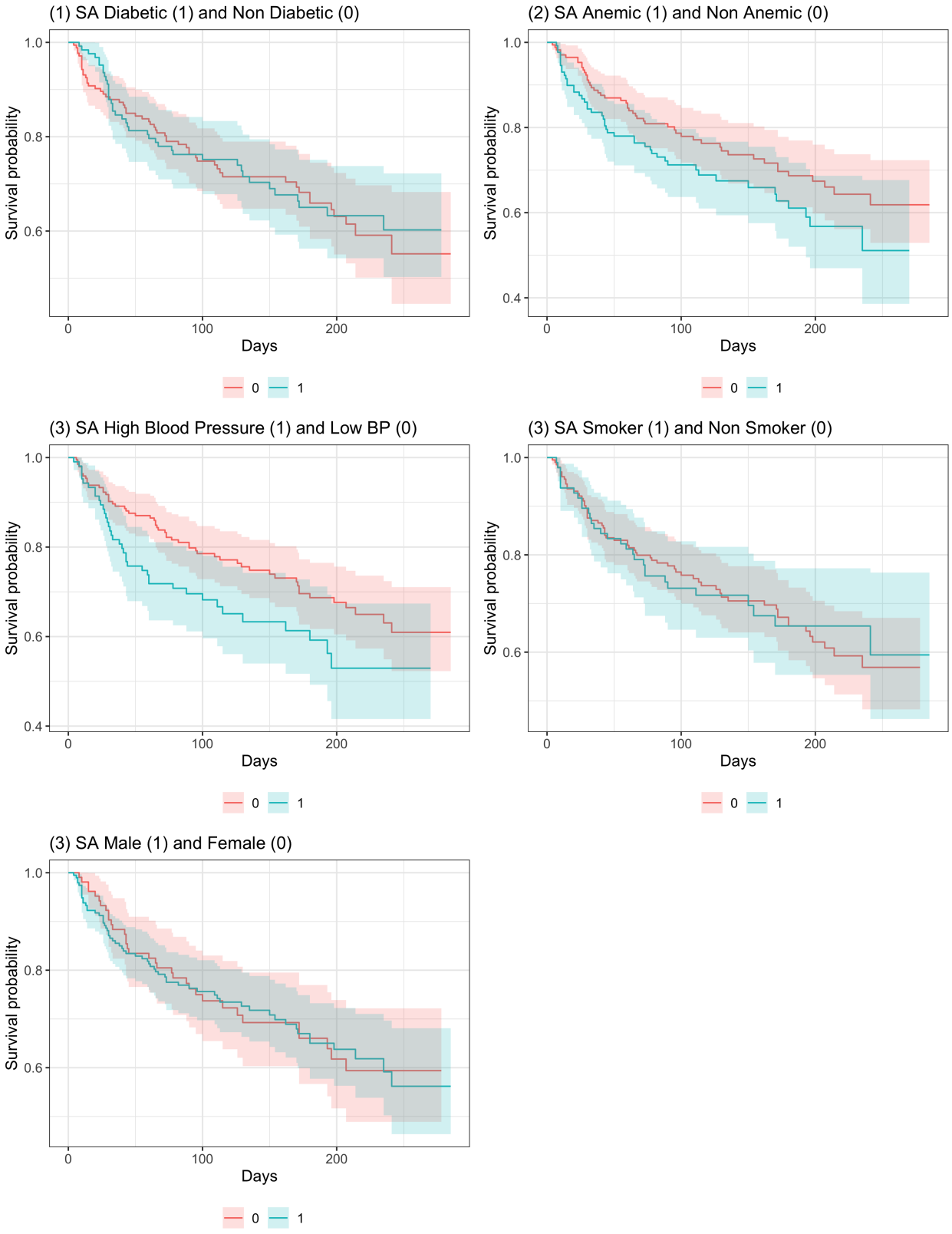

Figure 4: Surval Analysis (SA) of Patients over time based on each categorical feature

As observed from above figure (Fig 4), the survival rate over time curves appear relatively similar for Diabetics, Smoker and Sex category levels/factors hence we can deem them to be not significantly affecting the survival rate. However the survival curves appear to be different for each category level for High blood pressure and anaemic feature. It appears that patients with high blood pressure or anemic patients have low survival probability. We can also perform additional statistical tests to confirm these findings.

We use Cox regression model (Zabor 2019) which will allow us to analyse effect of multiple features or combined effect of several features (quantify an effect size for a single variable, or include more than one variable into a regression model to account for the effects of multiple variables) on survival outcome for further analysis. The following table shows the statistical summary obtained by using the cox-regression model.

Characteristic	**HR**	**95% CI**	**p-value**
age	1.05	1.03, 1.07	<0.001
anaemia	1.58	1.04, 2.42	0.034
creatinine_phosphokinase	1.00	1.00, 1.00	0.026
diabetes	1.15	0.74, 1.78	0.5
ejection_fraction	0.95	0.93, 0.97	<0.001
high_blood_pressure	1.61	1.05, 2.46	0.028
platelets	1.00	1.00, 1.00	0.7
serum_creatinine	1.38	1.20, 1.58	<0.001
serum_sodium	0.96	0.91, 1.00	0.058
sex	0.79	0.48, 1.29	0.3
smoking	1.14	0.70, 1.86	0.6

From the p-value (<0.05) itself, we can confirm some of the significant features which we analysed visually. Age, anaemia, high_blood_pressure are confirmed to be significant and can effect the survival of patient. Additional features like creatinine_phosphokinase, ejection_fraction and serum_creatinine also seems to be significant in predicting the survival of patient.

The following figure (Fig 5) show how creatinine_phosphokinase (CPK enzyme), ejection_fraction and serum_creatinine are associated with time and age. As we can see, the outliers affect the data for serum_creatinine and CPK hence observations seem to squeez at one end of the axis. Majority of patients died within fewer days of the follow up period mainly due to low level of the CPK enzyme in the blood which can thereby confirm its significance for survival. Similarly patients with high level of serum_creatinine (which can affect kidney function as well) appear to die in fewer days compared to thoes who survie.

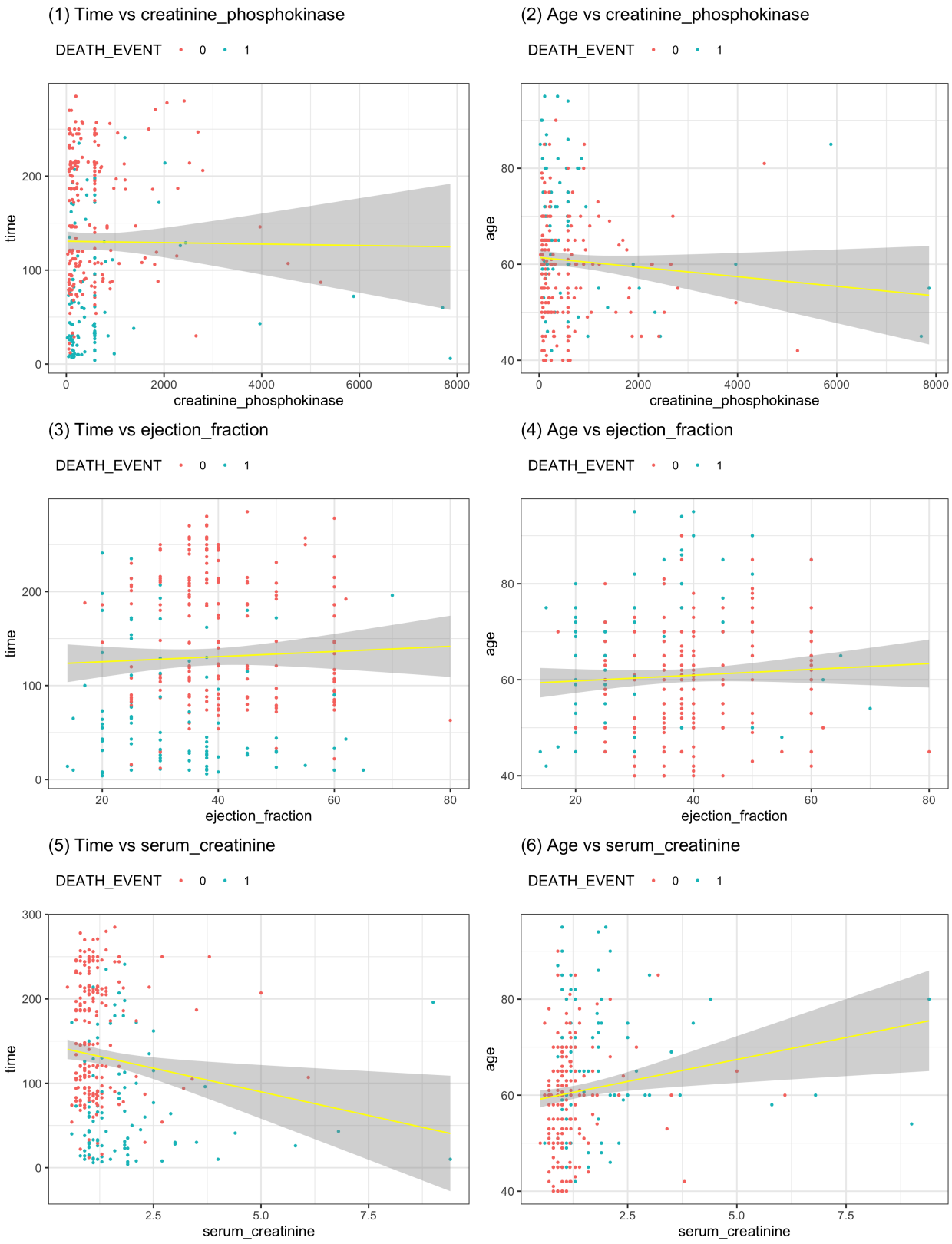

Figure 5: Continuous features compared over Age and Time.

3 Conclusion

After performing survival analysis and visually analysing the plots, we can confirm that Age, anaemia, high_blood_pressure, creatinine_phosphokinase, ejection_fraction and serum_creatinine can significantly affect the survival of a patient who was diagnosed with a heart failure. Using these features and model like the cox regression model, we can generate the hazard ratio, which is the risk or probability of occurrence of an event of interest i.e. the survival chance. However there are many other machine learning techniques available which can be used to predict the survival of patients with heart failure with better accuracy.

4 Code Snippets

```
# Importing data
heart_df = read.csv("heart_failure_clinical_records_dataset.csv",
                    header = TRUE, as.is = FALSE)
# Correlation heatmap
library(corrplot)
corrplot(cor(heart_df),addCoef.col = 1,
    number.cex = 0.5,tl.cex = 0.9)
```

```
# finding patient count for high_blood_pressure grouped by DEATH_EVENT
df1 = df %>% group_by(high_blood_pressure,DEATH_EVENT) %>%
summarise(patient_count = length(DEATH_EVENT))
s = sum(df1$patient_count)
#finding percentage and plotting bargraph
p4 = df1 %>%
  mutate(freq = paste((100 * round(patient_count / s, 4)),"%",sep = "")) %>%
ggplot(aes(x = high_blood_pressure, y=patient_count, fill = DEATH_EVENT)) +
geom_bar(stat="identity", position=position_dodge())+
  geom_text(aes(label = freq), vjust = -0.1, hjust=1)+
theme_bw()+ ggtitle("(4) Paitent Survival w.r.t High Blood Pressure")+
theme(legend.position='top', legend.justification='left',
      legend.direction='horizontal')+coord_flip()
```

```
# Plotting Patient Age vs Anaemia box plot w.r.t DEATH_EVENT
p22 = ggplot(heart_df, aes(x=anaemia, y=age, fill=DEATH_EVENT)) +
    geom_boxplot(alpha=0.6, show.legend = TRUE)+
 geom_jitter(alpha=0.2, size=0.3)+theme_bw()+
  ggtitle("(4) Patient Age vs Anaemia")+
  ylim(40,95)+
  scale_y_continuous(breaks=seq(40, 95, 5), limits=c(40, 95))
#...
```

```r
# Survival analysis using Kaplan Mier estimator and plotting
s1 = survfit2(Surv(time, DEATH_EVENT) ~ diabetes, data = num_df) %>%
  ggsurvfit() +
  labs(
    x = "Days",
    y = "Survival probability"
  )+
  add_confidence_interval()+ggtitle("(1) SA Diabetic (1) and Non Diabetic (0)")
#..
```

```r
# Cox regression model and table display
# . indicated all the features
t = coxph(Surv(time, DEATH_EVENT) ~ ., data = num_df) %>%
  tbl_regression(exp = TRUE)
```

```r
# Continuous explainatory variables variation with time
c3 = ggplot(data=df, aes(x = ejection_fraction, y = time))+
    geom_point(size = 0.5,aes(col = DEATH_EVENT))+
  stat_smooth(method = "lm", col = "yellow", size=0.5) +
  ggtitle("(3) Time vs ejection_fraction")+
    theme_bw()+
  theme(legend.position='top',
        legend.justification='left',
        legend.direction='horizontal')
#..
```

References

Davide Chicco, Giuseppe Jurman. 2020. *Heart Failure Clinical Records Data Set. UCI Machine Learning Repository.* https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

Zabor, Emily C. 2019. *Survival Analysis in r. UCI Machine Learning Repository.* https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html.