Heart Failure Survival Study

Mailman School of Public Health at Columbia University

Huanyu Chen, Runze Cui, Jiahe Deng, Yi Huang, Xuesen Zhao
P8108 Final Project Report: Group 6

Dec 7, 2023

Abstract

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Contents

1 Introduction

1.1 Background

Heart failure (HF) occurs when the muscles in the heart wall weaken and enlarge, impairing the heart's ability to pump blood effectively. This condition can cause the heart's ventricles to become stiff, hindering their ability to fill properly between beats. Over time, the heart becomes less capable of meeting the body's demand for blood, leading to symptoms like difficulty in breathing as the heart struggles to function efficiently.(Ahmad et al., 2017).

According to the statistics, heart failure affects 1-2% of adults in the general population and is more common in older individuals, with over 10% of those aged over 70 years being diagnosed. The actual prevalence might be as high as 4%, as heart failure is often undiagnosed or misdiagnosed, especially in the elderly. Since 2002, the prevalence of heart failure has increased by nearly 25%, driven by factors such as an aging population, better survival rates post-coronary events, and a rise in risk factors like hypertension and atrial fibrillation. (Jones et al., 2019).

1.2 Objective

The goal of our course project is to perform a re-analysis of the survival data from the paper entitled "Survival Analysis of Heart Failure Patients: A Case Study"[1]. This paper focused on heart failure patients admitted to the Institute of Cardiology and Allied Hospital in Faisalabad, Pakistan between April and December 2015. The original study highlighted key risk factors, including age, renal impairment, blood pressure, and ejection fraction, which significantly contribute to mortality in heart failure patients. Through this re-analysis, we aim to further investigate and validate the identified risk factors associated with mortality in heart failure patients. Going beyond the foundation established by this paper and their study, our project aims to re-evaluate and potentially refine the predictive model from other new points of view by employing a comprehensive survival analysis procedure that includes nonparametric methods, semi-parametric methods, parametric methods, and validation methods.

2 Methods

2.1 Exploratory Data Analysis (EDA)

The present study focuses on 299 heart failure patients, including 105 women and 194 men. All participants were over 40 years old and diagnosed with left ventricular systolic dysfunction, classified under NYHA classes III and IV. The follow-up duration ranged from 4 to 285 days, with an average of 130 days. Diagnosis of the disease was confirmed through cardiac echocardiogram reports or physician's notes. A brief description of variables in the dataset is shown below:

- age: Age in years
- time: Survival time in days
- event: Event binary indicator (0 = Censored, 1 = Event)
- gender: Sex binary indicator (0 = Female, 1 = Male)
- smoking: Smoking status (0 = No smoking, 1 = Smoking)
- diabetes: Diabetes status (0 = No diabetes, 1 = Diabetes)
- bp: Blood pressure status (0 = Normal, 1 = Hypertension)
- anemia: Anemia status (0 = No anemia, 1 = Anemia: patients with haematocrit < 36)
- EF: Ejection fraction (Low: EF ≤ 30, Medium: $30 <$ EF ≤ 45 and High: EF > 45)
- sodium: Sodium in mEq/L

[1]For more information and details, see **References** section

- creatinine: Serum creatinine in mg/dL
- platelets: Platelets in mcL
- cpk: Creatinine phosphokinase in U/L

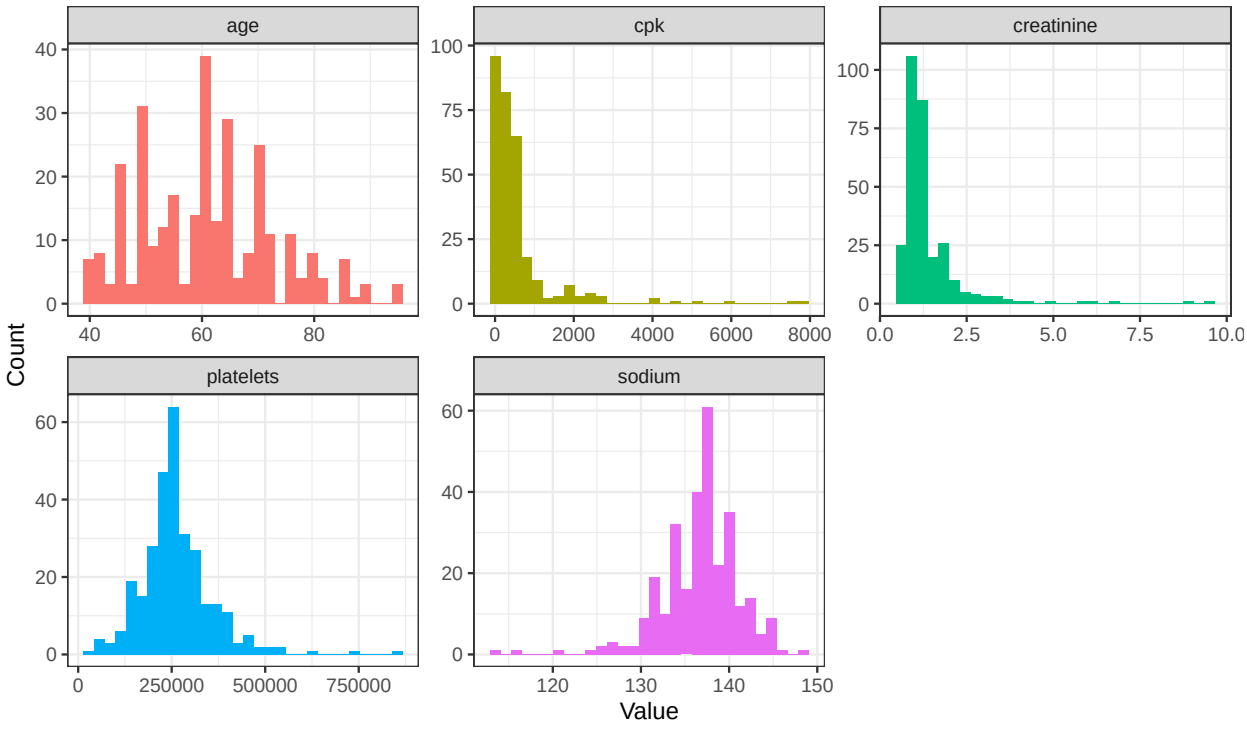
This study falls into the category of Overall Survival (OS), where `event = 1` indicates the death of the subject and is the endpoint of survival. Specifically, 203 subjects were right-censored and 96 subjects have event. Detailed descriptive statistics table stratified by survival status are presented below.

	Censored	Event	Overall	P-value
	(N=203)	(N=96)	(N=299)	
Survival time (days)				
Mean (SD)	158 (67.7)	70.9 (62.4)	130 (77.6)	<0.001
Median [Min, Max]	172 [12.0, 285]	44.5 [4.00, 241]	115 [4.00, 285]	
Age (years)				
Mean (SD)	58.8 (10.6)	65.2 (13.2)	60.8 (11.9)	<0.001
Median [Min, Max]	60.0 [40.0, 90.0]	65.0 [42.0, 95.0]	60.0 [40.0, 95.0]	
Gender				
0	71 (35.0%)	34 (35.4%)	105 (35.1%)	1
1	132 (65.0%)	62 (64.6%)	194 (64.9%)	
Smoking status				
0	137 (67.5%)	66 (68.8%)	203 (67.9%)	0.932
1	66 (32.5%)	30 (31.3%)	96 (32.1%)	
Diabetes				
0	118 (58.1%)	56 (58.3%)	174 (58.2%)	1
1	85 (41.9%)	40 (41.7%)	125 (41.8%)	
Blood Pressure				
0	137 (67.5%)	57 (59.4%)	194 (64.9%)	0.214
1	66 (32.5%)	39 (40.6%)	105 (35.1%)	
Ejection Fraction (EF)				
Low	42 (20.7%)	51 (53.1%)	93 (31.1%)	<0.001
Medium	115 (56.7%)	31 (32.3%)	146 (48.8%)	
High	46 (22.7%)	14 (14.6%)	60 (20.1%)	
Anemia				
0	120 (59.1%)	50 (52.1%)	170 (56.9%)	0.307
1	83 (40.9%)	46 (47.9%)	129 (43.1%)	
Serum Sodium (mEq/L)				
Mean (SD)	137 (3.98)	135 (5.00)	137 (4.41)	0.00187
Median [Min, Max]	137 [113, 148]	136 [116, 146]	137 [113, 148]	
Serum creatinine (mg/dL)				
Mean (SD)	1.18 (0.654)	1.84 (1.47)	1.39 (1.03)	<0.001
Median [Min, Max]	1.00 [0.500, 6.10]	1.30 [0.600, 9.40]	1.10 [0.500, 9.40]	
Creatinine phosphokinase (U/L)				
Mean (SD)	540 (754)	670 (1320)	582 (970)	0.369
Median [Min, Max]	245 [30.0, 5210]	259 [23.0, 7860]	250 [23.0, 7860]	
Plateletes (mcL)				
Mean (SD)	267000 (97500)	256000 (98500)	263000 (97800)	0.399
Median [Min, Max]	263000 [25100, 850000]	259000 [47000, 621000]	262000 [25100, 850000]	

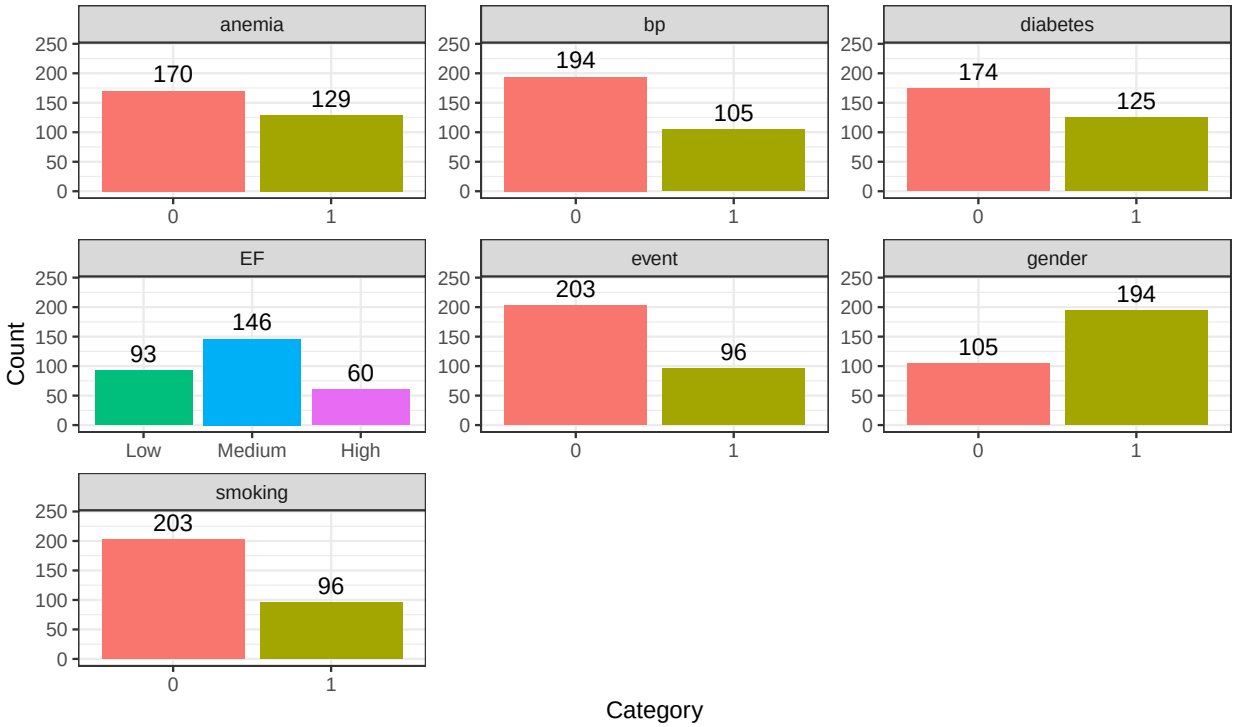
According to the descriptive table, we can observe that the mean survival times for censored and event are 158 and 70.9 days, respectively. Since our dataset is completed, we do not need to concern the missingness issue. The p-values of the variables are also presented in the table. Some variables have relatively larger p-values. However, we still need to check distribution of each variable[2] and perform model selection to determine which variables to ultimately use for the analysis.

[2]Use histograms for continuous variables and bar charts for categorical variables

Figure 1Histograms of Continuous Covariates



Bar Charts of Categorical Covariates



3 References

1. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. PLOS ONE, 12(7), e0181001. https://doi.org/10.1371/journal.pone.0181001

2. Jones, N., Ak, R., Adoki, I., Fdr, H., & Cj, T. (2019). Survival of patients with chronic heart failure in the community: a systematic review and meta-analysis. European Journal of Heart Failure, 21(11), 1306–1325. https://doi.org/10.1002/ejhf.1594

4 Appendix

4.1 Code

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
library(tidyverse)
library(knitr)
library(kableExtra)
library(survival)
library(writexl)
library(readxl)
library(table1)
library(rmarkdown)
dat <- read_csv("./data/heart_failure.csv") %>%
  arrange(TIME) %>% janitor::clean_names() %>%
  mutate(ejection_fraction = case_when(ejection_fraction <= 30 ~ "Low",
                                       ejection_fraction > 30 & ejection_fraction <= 45 ~ "Medium",
                                       ejection_fraction > 45 ~ "High")) %>%
  mutate(gender = factor(gender),
         smoking = factor(smoking),
         diabetes = factor(diabetes),
         bp = factor(bp),
         event = factor(event),
         anaemia = factor(anaemia),
         ejection_fraction = factor(ejection_fraction, levels = c("Low", "Medium", "High"))) %>%
  rename(platelets = pletelets,
         anemia = anaemia,
         EF = ejection_fraction)

# Calculate the number of right-censored:
number_censored <- sum(dat$event == 0)
# Calculate the number of event:
number_event <- sum(dat$event == 1)
dat_table = dat
label(dat_table$time) = "Survival time (days)"
label(dat_table$gender) = "Gender"
label(dat_table$smoking) = "Smoking status"
label(dat_table$diabetes) = "Diabetes"
label(dat_table$bp) = "Blood Pressure"
label(dat_table$anemia) = "Anemia"
label(dat_table$age) = "Age (years)"
label(dat_table$EF) = "Ejection Fraction (EF)"
label(dat_table$sodium) = "Serum Sodium (mEq/L)"
label(dat_table$creatinine) = "Serum creatinine (mg/dL)"
label(dat_table$platelets) = "Plateletes (mcL)"
label(dat_table$cpk) = "Creatinine phosphokinase (U/L)"
dat_table$event <- factor(dat$event, levels = c(0, 1), labels = c("Censored", "Event"))

pvalue <- function(x, ...) {
  # Remove the "overall" column
    x <- x[names(x) != "overall"]
    # Construct vectors of data y, and groups (strata) g
```

```
    y <- unlist(x)
    g <- factor(rep(1:length(x), times = sapply(x, length)))
    if (is.numeric(y)) {
        # For numeric variables, perform a standard 2-sample t-test
        p <- t.test(y ~ g)$p.value
    } else {
        # For categorical variables, perform a chi-squared test of independence
        p <- chisq.test(table(y, g))$p.value
    }
    # Format the p-value, using an HTML entity for the less-than sign.
    # The initial empty string places the output on the line below the variable label.
    c("", sub("<", "<", format.pval(p, digits = 3, eps = 0.001)))
}


cat("\\begingroup\\small")  # Adjust the font size command as needed
table1(~ time + age + gender + smoking + diabetes + bp + EF + anemia + sodium + creatinine + cpk + plat
                  data = dat_table, extra.col = list(`P-value` = pvalue))
cat("\\endgroup")
# Data contains the continuous vars only
cont_dat = dat %>%
  select(age, sodium, creatinine, platelets, cpk)
# Long format
cont_dat.long = cont_dat %>%
  pivot_longer(cols = c(age, sodium, creatinine, platelets, cpk))
# Plot the continuous variable histograms
cont_hist = ggplot(data = cont_dat.long, aes(x = value)) +
  geom_histogram(aes(fill = name), bins = 30) +
  facet_wrap(~name, scales = "free") +
  labs(x = "Value", y = "Count", title = "Figure 1Histograms of Continuous Covariates") +
  theme_bw() +
  theme(legend.position = "none")
cont_hist

# Data contains the categorical vars only
cate_data = dat %>%
  select(event, gender, smoking, diabetes, bp, anemia, EF)
# Long format
cate_dat.long = cate_data %>%
  pivot_longer(cols = c(event, gender, smoking, diabetes, bp, anemia, EF))
# Plot the categorical variable barplots
cate_barplot = ggplot(cate_dat.long, aes(x = value, fill = value)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = ..count..), vjust = -0.5) +
  facet_wrap(~name, scales = "free") +
  labs(x = "Category", y = "Count", fill = "Category", title = "Bar Charts of Categorical Covariates") +
  theme_bw() +
  theme(legend.position = "none") +
  ylim(0, 240)
cate_barplot
```