Heart-Failure-Survival Analysis-Project

Martial_Mallika_Lakshmi

2024-08-24

Survival Analysis model for predicting mortality caused by Heart Failure.

Introduction : Cardiovascular diseases claim the lives of approximately 17 million people worldwide each year, primarily manifesting as heart attacks and heart failure. Heart failure (HF) occurs when the heart is unable to pump sufficient blood to meet the body's needs.

Electronic medical records (EMRs) of patients provide a wealth of data, including symptoms, physical characteristics, and clinical laboratory test results. This data can be analyzed using biostatistics to uncover patterns and correlations that might be missed by medical professionals.

Machine learning, in particular, offers powerful tools for predicting patient survival based on their medical data. It can also identify the most critical features within these records, providing valuable insights for healthcare providers.

Dataset : The dataset includes cardiovascular medical records from 299 patients, consisting of 105 women and 194 men aged between 40 and 95 years. All patients were diagnosed with systolic dysfunction of the left ventricle and had a history of heart failure. Consequently, each patient was classified into either class III or class IV of the New York Heart Association (NYHA) classification, indicating various stages of heart failure.

#Load survival library

```
suppressMessages(library(survival))
```

```
## Warning: package 'survival' was built under R version 4.3.3
```

```
suppressMessages(library(tidyverse))
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
suppressMessages(library(lubridate))
suppressMessages(library(broom))
suppressMessages(library(ggfortify))
```

```
## Warning: package 'ggfortify' was built under R version 4.3.3
```

```r
suppressMessages(library(survminer))
```

```
## Warning: package 'survminer' was built under R version 4.3.3
```

```
## Warning: package 'ggpubr' was built under R version 4.3.3
```

#Load the dataset, EDA : This section aims to explore dataset data before performing any kind of analysis.

```r
setwd("C:/Users/selva/Documents/DSTI/Survival Analysis")
data<-read.csv("heart_failure_clinical_records_dataset.csv")
summary(data)
```

```
##       age           anaemia       creatinine_phosphokinase    diabetes
##  Min.   :40.00   Min.   :0.0000   Min.   :  23.0           Min.   :0.0000
##  1st Qu.:51.00   1st Qu.:0.0000   1st Qu.: 116.5           1st Qu.:0.0000
##  Median :60.00   Median :0.0000   Median : 250.0           Median :0.0000
##  Mean   :60.83   Mean   :0.4314   Mean   : 581.8           Mean   :0.4181
##  3rd Qu.:70.00   3rd Qu.:1.0000   3rd Qu.: 582.0           3rd Qu.:1.0000
##  Max.   :95.00   Max.   :1.0000   Max.   :7861.0           Max.   :1.0000
##  ejection_fraction high_blood_pressure   platelets      serum_creatinine
##  Min.   :14.00     Min.   :0.0000      Min.   : 25100   Min.   :0.500
##  1st Qu.:30.00     1st Qu.:0.0000      1st Qu.:212500   1st Qu.:0.900
##  Median :38.00     Median :0.0000      Median :262000   Median :1.100
##  Mean   :38.08     Mean   :0.3512      Mean   :263358   Mean   :1.394
##  3rd Qu.:45.00     3rd Qu.:1.0000      3rd Qu.:303500   3rd Qu.:1.400
##  Max.   :80.00     Max.   :1.0000      Max.   :850000   Max.   :9.400
##   serum_sodium        sex            smoking            time
##  Min.   :113.0   Min.   :0.0000   Min.   :0.0000   Min.   :  4.0
##  1st Qu.:134.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 73.0
##  Median :137.0   Median :1.0000   Median :0.0000   Median :115.0
##  Mean   :136.6   Mean   :0.6488   Mean   :0.3211   Mean   :130.3
##  3rd Qu.:140.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:203.0
##  Max.   :148.0   Max.   :1.0000   Max.   :1.0000   Max.   :285.0
##   DEATH_EVENT
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3211
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

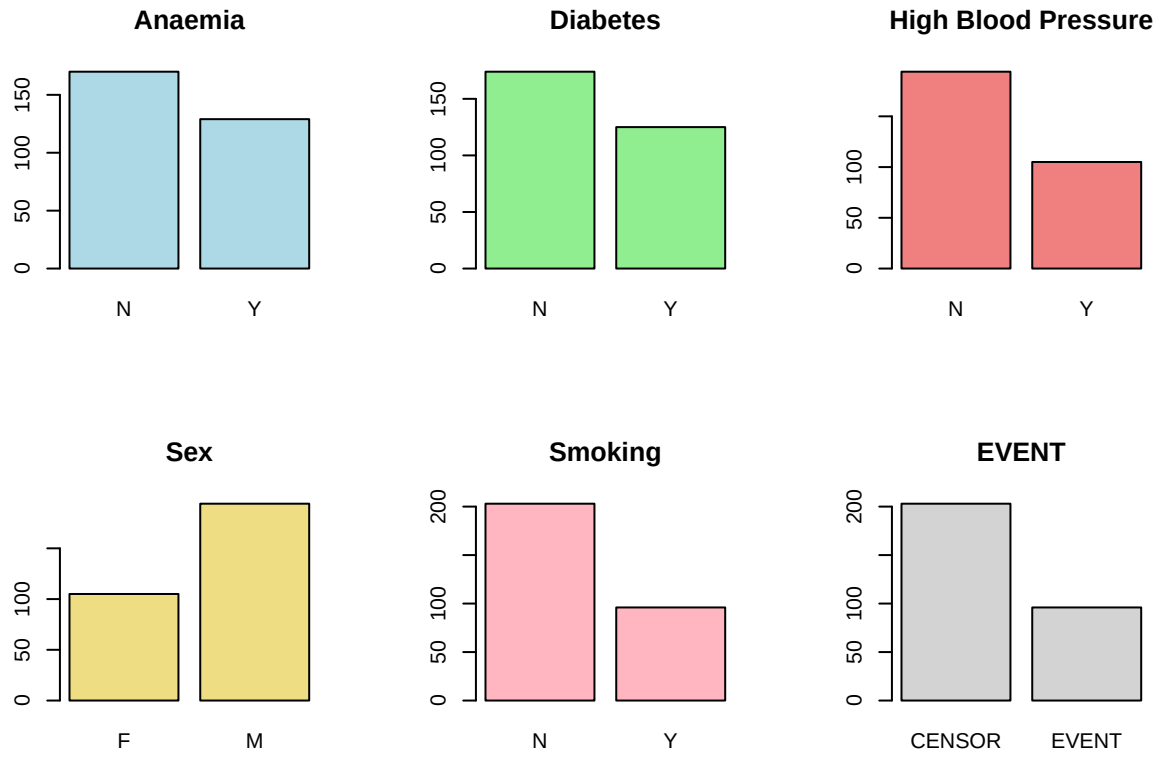
we can see that there are some continous variable that should be categorical like anaemia, diabetes, high_blood_pressure, sex , smoking and DEATH_EVENT Let's categorize them

Categorize variables

Anaemia

```r
df= data %>%
  mutate(anaemia = factor(anaemia, levels = c(0, 1), labels = c("N", "Y")),
         diabetes =factor(diabetes, levels = c(0, 1), labels = c("N", "Y")),
         high_blood_pressure = factor(high_blood_pressure, levels = c(0, 1), labels = c("N", "Y")),
         sex =factor(sex, levels = c(0, 1), labels = c("F", "M")),
         smoking=factor(smoking, levels = c(0, 1), labels = c("N", "Y")),
         DEATH_EVENT=factor(DEATH_EVENT, levels = c(0, 1), labels = c("CENSOR", "EVENT")))

par(mfrow =c(2,3))
barplot(table(df$anaemia), main = "Anaemia", col = "lightblue")
barplot(table(df$diabetes), main = "Diabetes", col = "lightgreen")
barplot(table(df$high_blood_pressure), main = "High Blood Pressure", col = "lightcoral")
barplot(table(df$sex), main = "Sex", col = "lightgoldenrod")
barplot(table(df$smoking), main = "Smoking", col = "lightpink")
barplot(table(df$DEATH_EVENT), main = "EVENT", col = "lightgray")
```




```r
str(df)
```

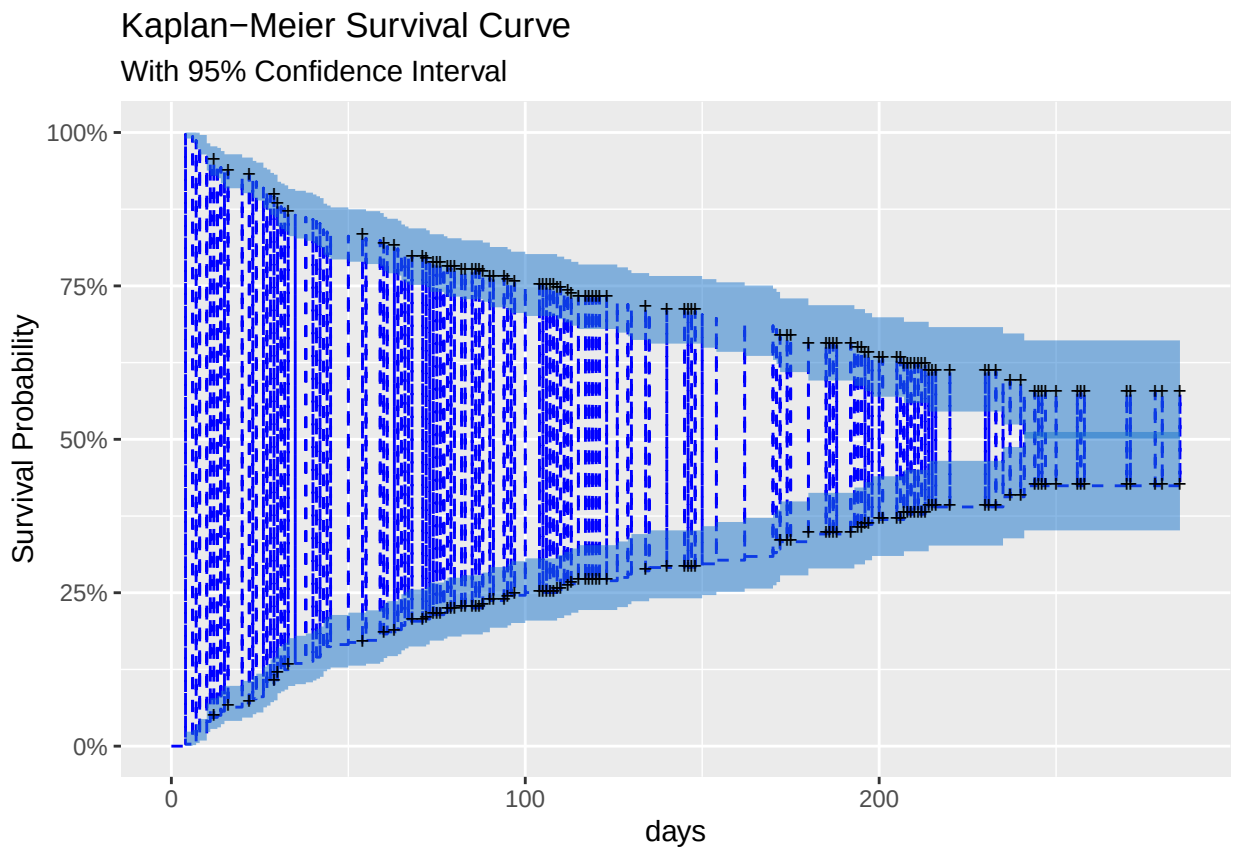
```
## 'data.frame':    299 obs. of  13 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : Factor w/ 2 levels "N","Y": 1 1 1 2 2 2 2 2 1 2 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : Factor w/ 2 levels "N","Y": 1 1 1 1 2 1 1 2 1 1 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
```

```
##  $ high_blood_pressure     : Factor w/ 2 levels "N","Y": 2 1 1 1 1 2 1 1 1 2 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : Factor w/ 2 levels "F","M": 2 2 2 2 1 2 2 2 1 2 ...
##  $ smoking                 : Factor w/ 2 levels "N","Y": 1 1 2 1 1 2 1 2 1 2 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : Factor w/ 2 levels "CENSOR","EVENT": 2 2 2 2 2 2 2 2 2 2 ...
```

Basic Non-Parametric estimation of survival : Here we are performing basic analysis using Kaplan-Meier methods.

```
kmsurvival = survfit(Surv(df$time, df$DEATH_EVENT) ~ 1)
#summary(kmsurvival)
```

```
autoplot(kmsurvival, xlab="days", ylab="Survival Probability", surv.linetype = 'dashed',
         surv.colour = 'blue',
         conf.int.fill = 'dodgerblue3', conf.int.alpha = 0.5)+
  labs(title = "Kaplan-Meier Survival Curve", subtitle = "With 95% Confidence Interval")
```



The survival curve starts at 100% and gradually declines, indicating that the probability of survival decreases over time as more events (deaths) occurs.

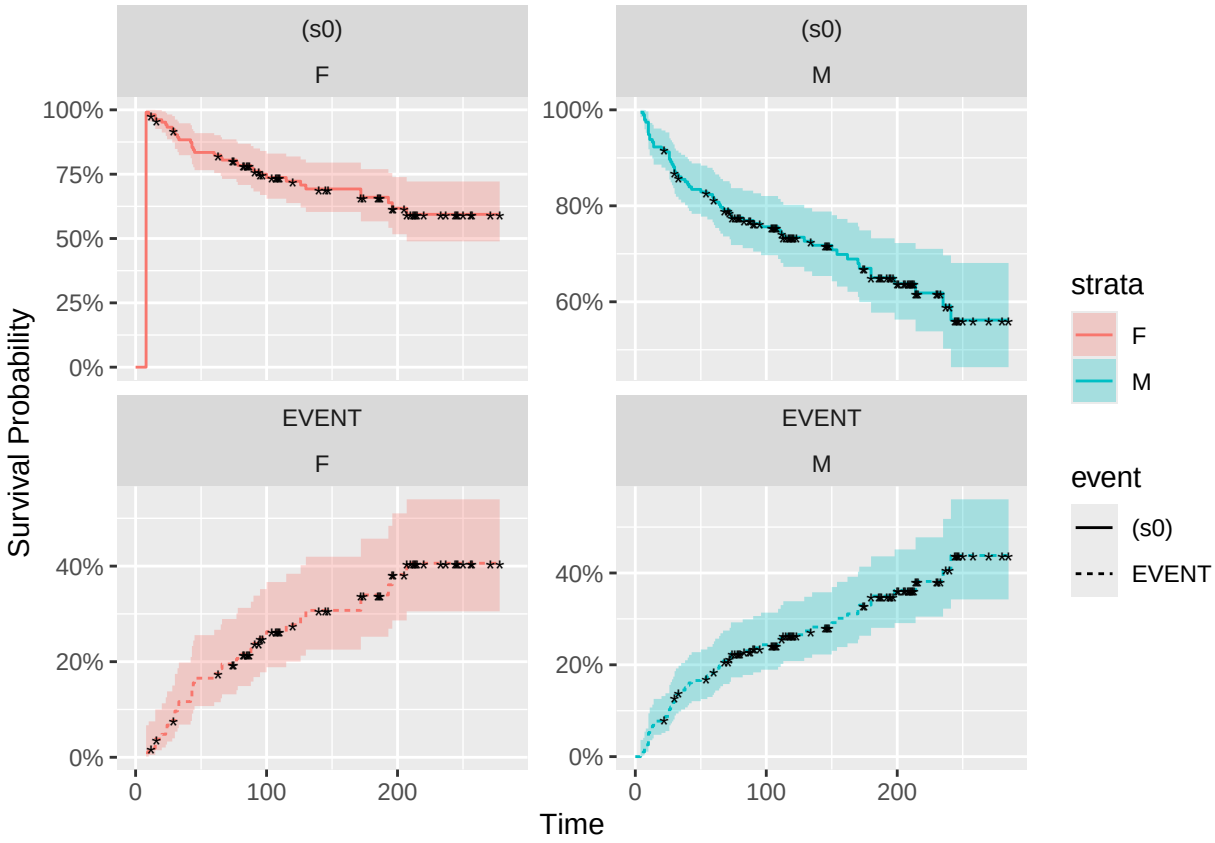
Non-Parametric Groups Analysis : This section aims to perform different group analysis on the dataset.

Gender Analysis It seems from the analysis that males are living longer than females

```
kmsurvival_gender = survfit(Surv(df$time, df$DEATH_EVENT) ~ df$sex)
#summary(kmsurvival_gender)
```

```
autoplot(kmsurvival_gender,
         censor.shape = '*', facets = TRUE, ncol = 2, xlab="Time", ylab="Survival Probability")
```

```
## Warning: Using formula(x) is deprecated when x is a character vector of length > 1.
##   Consider formula(paste(x, collapse = " ")) instead.
```



Initially, both groups have high survival probabilities, but as time progresses, differences become more apparent. Males generally have a higher survival probability compared to females at the early stages but this might change at later stages depending on the number of events and sample size. The number of events (deaths) varies over time. For example, at time 10, females had 1 death while males had 5 deaths, showing that more deaths occurred in males at this specific time point.

```r
# Perform the Log-rank test
logrank_test <- survdiff(Surv(time, DEATH_EVENT == "EVENT") ~ sex, data = df)

# Print the results
logrank_test
```

```
## Call:
## survdiff(formula = Surv(time, DEATH_EVENT == "EVENT") ~ sex,
##     data = df)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 105       34     34.3   0.00254   0.00397
## sex=M 194       62     61.7   0.00141   0.00397
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

The log-rank test suggests that there is no evidence to reject the null hypothesis that the survival curves for males and females are the same. This is consistent with the findings from the Cox proportional hazards model, where the effect of sex on survival was also not statistically significant.

```r
# Perform the Log-rank test for the DEATH_EVENT effect
logrank_test_death <- survdiff(Surv(time, DEATH_EVENT == "EVENT") ~ DEATH_EVENT, data = df)
# Print the log-rank test results
print(logrank_test_death)
```

```
## Call:
## survdiff(formula = Surv(time, DEATH_EVENT == "EVENT") ~ DEATH_EVENT,
##     data = df)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## DEATH_EVENT=CENSOR 203        0     76.8      76.8       399
## DEATH_EVENT=EVENT   96       96     19.2     306.0       399
##
##  Chisq= 399  on 1 degrees of freedom, p= <2e-16
```

This test doesn't provide new insights because it compares the survival outcomes based on the outcome itself, which is inherently circular. The extremely small p-value confirms this obvious distinction between the two groups.

Semi-parametric Cox regression

```r
# Fit the Cox Proportional Hazards model
cox_model <- coxph(Surv(time, DEATH_EVENT == "EVENT") ~ anaemia + diabetes + high_blood_pressure
                   + sex + smoking, data = df)
# Summarize the Cox model
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT == "EVENT") ~ anaemia +
```

```
##     diabetes + high_blood_pressure + sex + smoking, data = df)
##
##   n= 299, number of events= 96
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## anaemiaY              0.33929   1.40395  0.20554  1.651   0.0988 .
## diabetesY            -0.03146   0.96903  0.21072 -0.149   0.8813
## high_blood_pressureY  0.44189   1.55564  0.21114  2.093   0.0364 *
## sexM                  0.05601   1.05761  0.24121  0.232   0.8164
## smokingY              0.02380   1.02408  0.24878  0.096   0.9238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## anaemiaY                  1.404     0.7123    0.9384     2.100
## diabetesY                 0.969     1.0320    0.6412     1.465
## high_blood_pressureY      1.556     0.6428    1.0284     2.353
## sexM                      1.058     0.9455    0.6592     1.697
## smokingY                  1.024     0.9765    0.6289     1.668
##
## Concordance= 0.583  (se = 0.031 )
## Likelihood ratio test= 6.99  on 5 df,   p=0.2
## Wald test            = 7.15  on 5 df,   p=0.2
## Score (logrank) test = 7.25  on 5 df,   p=0.2
```

High Blood Pressure (high_blood_pressureY) is the only predictor with a statistically significant effect on the hazard of the event, with a hazard ratio of approximately 1.56. This suggests that individuals with high blood pressure have a 56% higher risk of the event (DEATH_EVENT) compared to those without, controlling for other factors in the model.

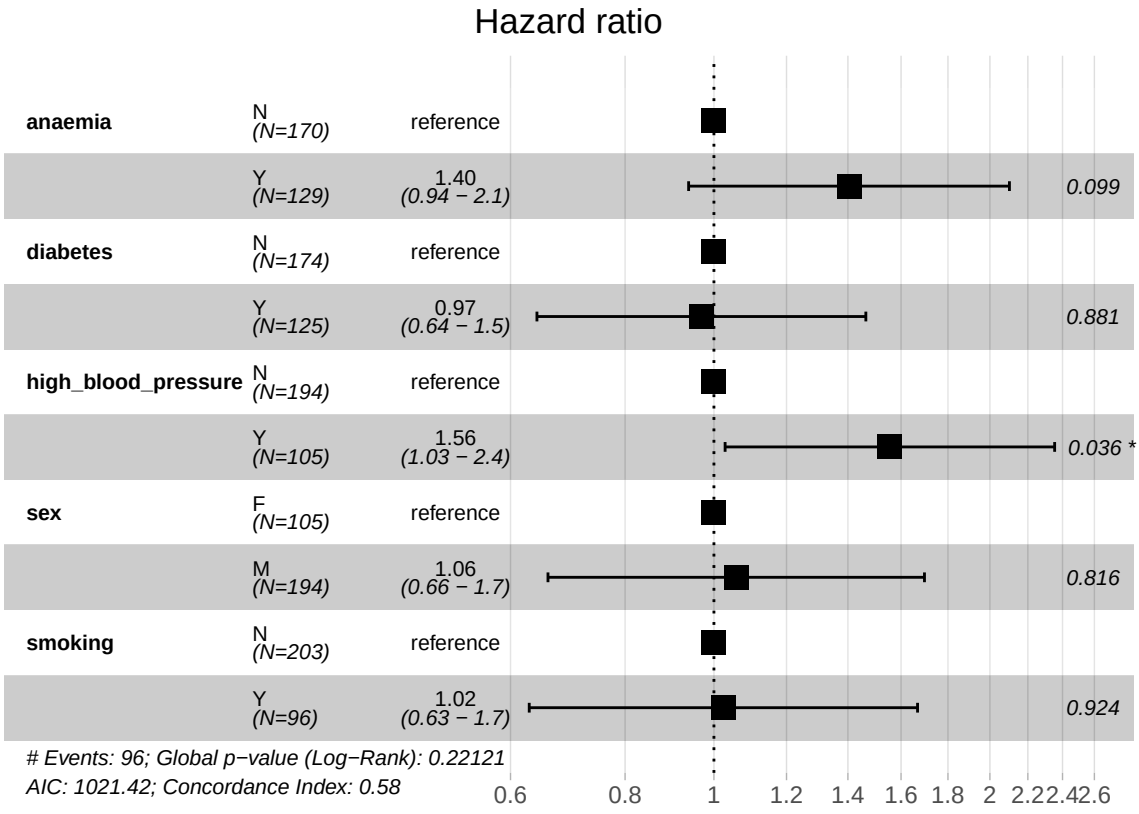
Other predictors (anaemia, diabetes, sex, smoking) did not show statistically significant associations with the event at the conventional 0.05 significance level. The model's concordance index (0.583) indicates modest predictive ability.

```r
# Extract the terms from the Cox model
model_terms <- terms(cox_model)
# Get the variable names
variable_names <- attr(model_terms, "term.labels")
print(variable_names)
```

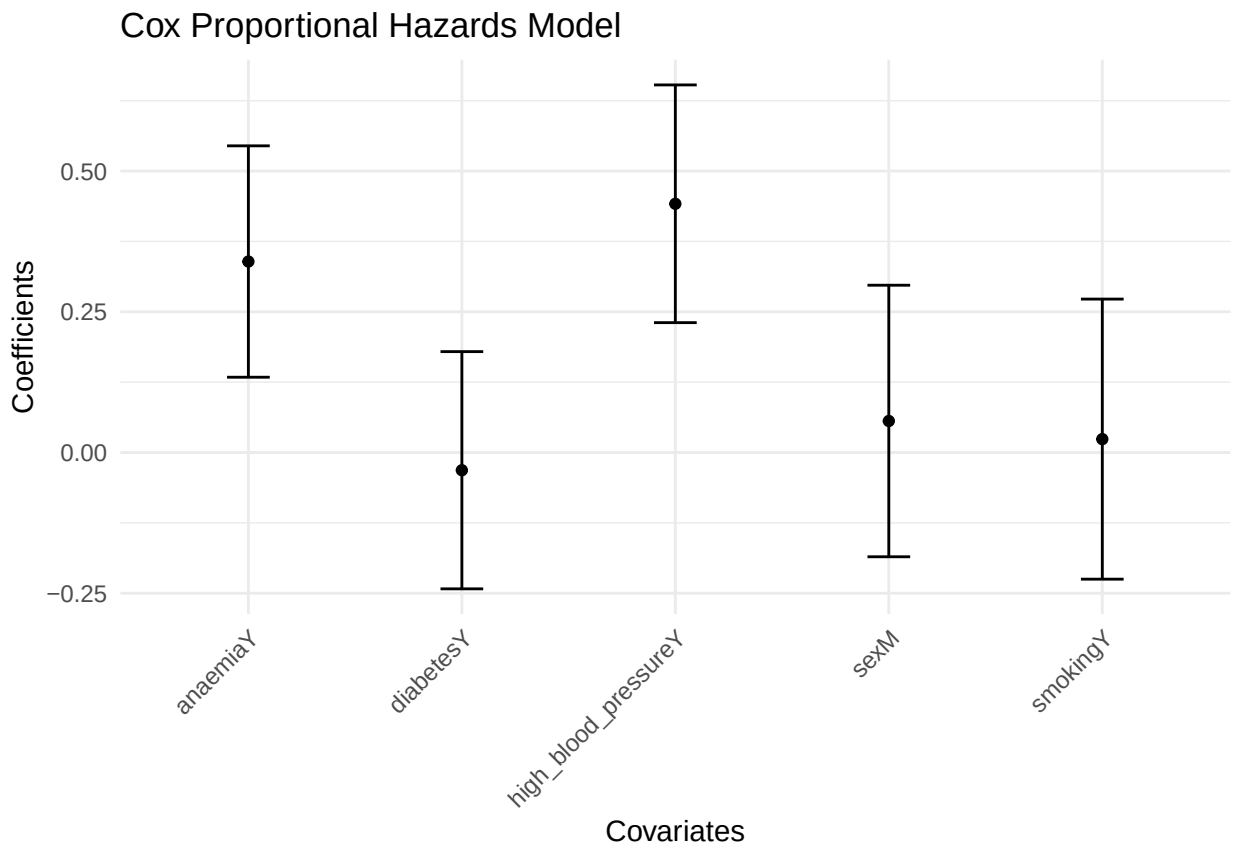
```
## [1] "anaemia"             "diabetes"            "high_blood_pressure"
## [4] "sex"                 "smoking"
```

```r
# Use ggforest to plot the Cox model coefficients
ggforest(cox_model)
```

```
## Warning in .get_data(model, data = data): The ‘data‘ argument is not provided.
## Data will be extracted from model fit.
```




```r
# to visualize the coefficients with error bars:
library(ggplot2)
# Extract coefficients and standard errors from the Cox model
cox_summary <- summary(cox_model)
coefficients <- cox_summary$coefficients
# Create a data frame with the necessary information
cox_coef_df <- data.frame(
  Covariate = rownames(coefficients),
  coef = coefficients[, "coef"],
  se_coef = coefficients[, "se(coef)"]
)
# Plot the coefficients with error bars
ggplot(cox_coef_df, aes(x = Covariate, y = coef)) +
  geom_point() +
  geom_errorbar(aes(ymin = coef - se_coef, ymax = coef + se_coef), width = 0.2) +
  labs(title = "Cox Proportional Hazards Model", x = "Covariates", y = "Coefficients") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Summary:

anaemiaY: The coefficient is positive and the confidence interval does not include zero, indicating that anemia is likely associated with an increased hazard (higher risk) of the event occurring.

high_blood_pressureY : The coefficient is positive and the confidence interval does not include zero, indicating that its likely associated with an increased hazard (higher risk) of the event occurring.

diabetesY, sexM, smokingY: The confidence intervals for these covariates all are below zero, suggesting that their effects are not statistically significant in this model. This means that these covariates may not have a strong impact on the hazard rate in this dataset.

Hence high_blood_pressureY and anemia appears to be a significant risk factor in this Cox proportional hazards model, whereas the other variables do not show a statistically significant effect based on this plot.

The limitations of the study is a small dataset, which restricts us to get more insightful information.

Furthermore, good awareness programs that communicate the benefits of a heart healthy diet and physically active lifestyle could prove to be effective and may help to reduce the mortality rate due to heart failure conditions.