# Survival Analysis of Heart Failure

Yonghan Qiu
Mengzhao Xu
Yuhua Liu

# CONTENTS

01 【Introduction】

02 【Methodology】

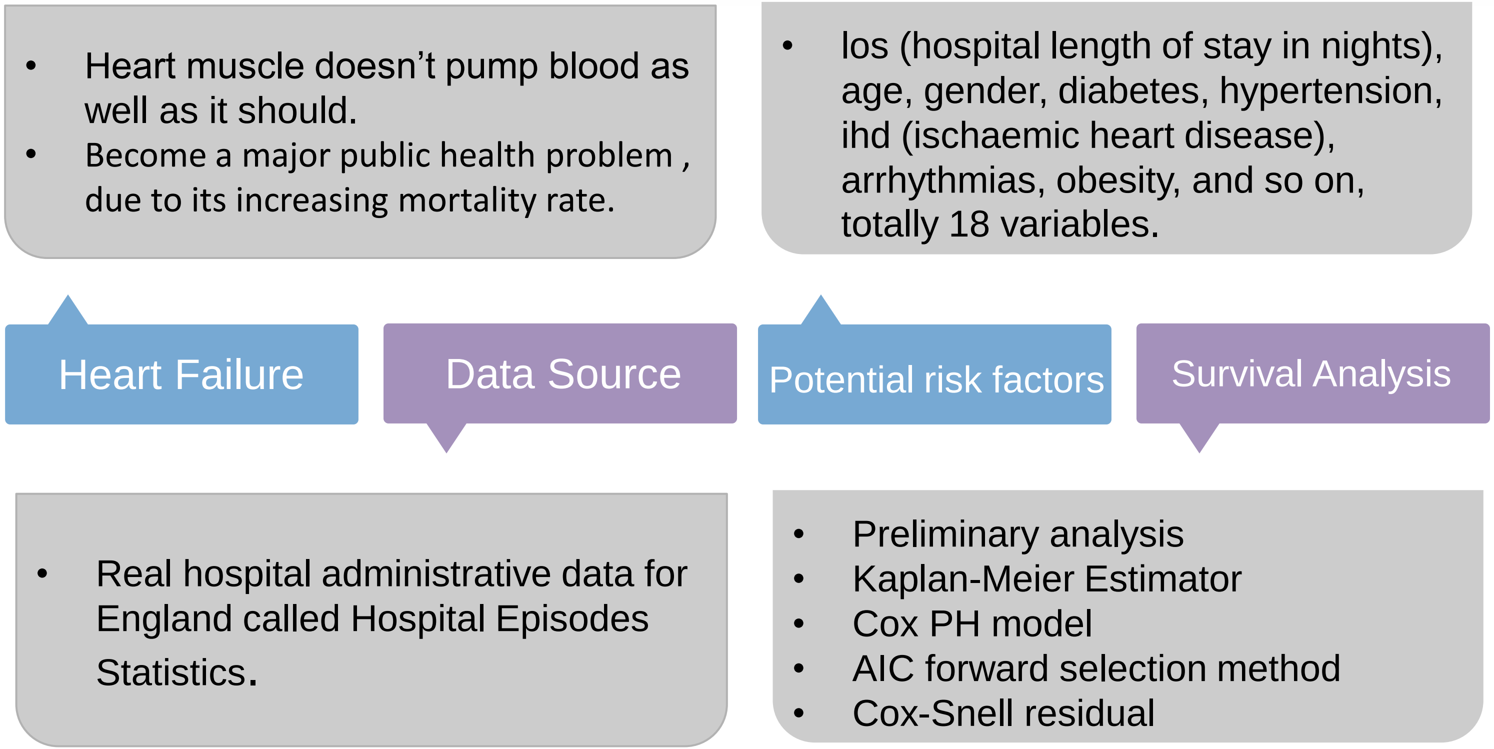
03 【Analysis and Results】

04 【Conclusion】

01

# Introduction

# 1 Introduction



## Heart Failure

- Heart muscle doesn’t pump blood as well as it should.
- Become a major public health problem , due to its increasing mortality rate.

## Data Source

- Real hospital administrative data for England called Hospital Episodes Statistics.

## Potential risk factors

- los (hospital length of stay in nights), age, gender, diabetes, hypertension, ihd (ischaemic heart disease), arrhythmias, obesity, and so on, totally 18 variables.

## Survival Analysis

- Preliminary analysis
- Kaplan-Meier Estimator
- Cox PH model
- AIC forward selection method
- Cox-Snell residual

02

# Methodology

# 2.1 Data Description

- ❖ 1000 heart failure patients
- ❖ 452 female and 548 male
- ❖ 492 dead and 508 alive at the end of the research
- ❖ 31 parameters in total, only 18 parameters interested

- death
- los: hospital length of stay in nights
- age: in years
- gender
- cabg: previous heart bypass
- diabetes: any type
- hypertension
- ihd: ischaemic heart disease
- arrhythmias
- copd: chronic obstructive lung disease
- obesity
- pvd: peripheral vascular disease
- valvular_disease: disease of the heart valves
- pacemaker
- prior_appts_attended: number of outpatient appointments attended in the previous year
- prior_dnas: number of outpatient appointments missed in the previous year
- pci: percutaneous coronary intervention
- fu_time (follow-up time)

# 2.2 Statistic Techniques

- The Cox-proportional hazards model was used to research the relationships between the time to event outcome and a set of explanatory variables and test for the significance of these factors. The model has the form:

$$h(t|Z) = h_0(t)\exp(\beta Z)$$

- To estimate and compare the survival function S(t) for different levels of explanatory variables, we used the Kaplan-Meier (K-M) estimator. The definition is:

$$S(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right), t_1 \leq t$$

# 2.2 Statistic Techniques

- We used forward AIC to select the best Cox Regression model and plotted the Cox-Snell residual to check the overall fit of the model. The definition of the Cox-Snell residual is:

$$r_j = \widehat{H_0}(T_j)\exp(\hat{\beta}^T Z_j) \quad j = 1,2, \dots, n$$

where $Z_j = (Z_{j1}, \dots, Z_{jp})^T$ are all fixed-time covariates. $r_j$ are censored sample from exponential distribution, given the assumed Cox model holds and $\hat{\beta}$, $\widehat{H_0}(t)$ close to the true values β, $H_0(t)$.

- We also plotted $H_r(r_j)$ versus $r_j$. If the Cox model provides a good fit of the data, we expect a straight line through the origin with slope 1.

03

# Analysis and Results

# 3.1 Preliminary Analysis

## ❑ Baseline Characteristics of the Data

| Continuous Variables | | |
|---|---|---|
| Variable | Dead(N=492) | Censored(N=508) |
| | Mean(Standard Deviation) | |
| age | 82.175(8.788) | 75.396(12.119) |
| los | 12.447(14.475) | 9.154(10.180) |
| prior_appts_attended | 5.283(6.271) | 5.785(7.344) |
| prior_dnas | 0.547(1.223) | 0.457(0.985) |

# 3.1 Preliminary Analysis

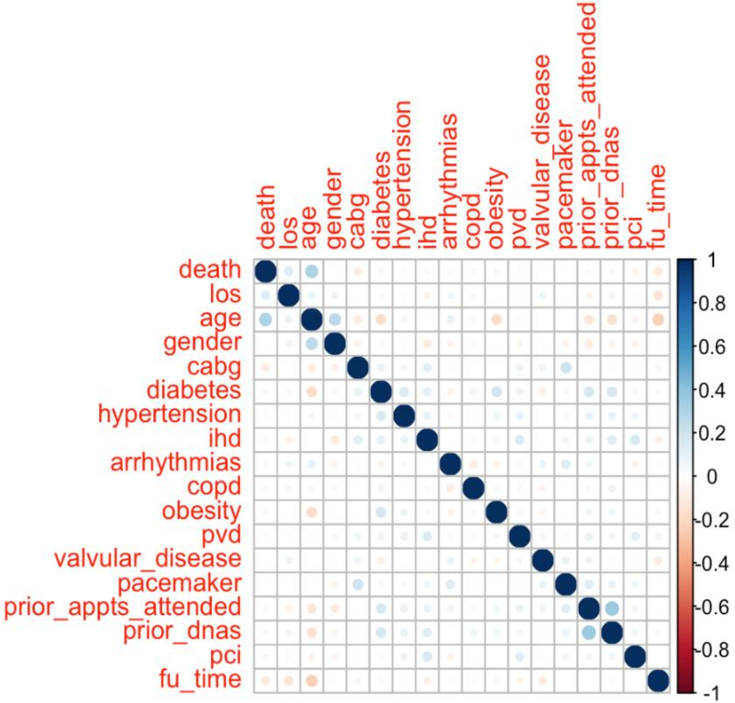
## ❑ Baseline Characteristics of the Data

| Categorical Variables | | | | |
|---|---|---|---|---|
| **Variable** | **Categories** | **Dead(N=492)** | **Censored(N=508)** | **Percentage of Dead (49.2%)** |
| **gender** | Female(1) | 224(45.5%) | 228(44.9%) | 49.6% |
| | Male(0) | 268(54.5%) | 280(55.1%) | 48.9% |
| **cabg** | Yes(1) | 1(0.2%) | 13(2.6%) | 7.1% |
| | No(0) | 491(99.8%) | 495(97.4%) | 49.8% |
| **diabetes** | Yes(1) | 129(26.2%) | 154(30.3%) | 45.6% |
| | No(0) | 363(73.8%) | 354(69.7%) | 50.6% |
| **hypertension** | Yes(1) | 300(61.0%) | 321(63.2%) | 48.3% |
| | No(0) | 192(39.0%) | 187(36.8%) | 50.7% |
| **ihd** | Yes(1) | 253(51.4%) | 242(47.6%) | 51.1% |
| | No(0) | 239(48.6%) | 266(52.4%) | 47.3% |
| **arrhythmias** | Yes(1) | 231(47.0%) | 259(51.0%) | 47.1% |
| | No(0) | 261(53.0%) | 249(49.0%) | 51.2% |
| **copd** | Yes(1) | 127(25.8%) | 115(22.6%) | 52.5% |
| | No(0) | 365(74.2%) | 393(77.4%) | 48.2% |
| **obesity** | Yes(1) | 23(4.7%) | 35(6.9%) | 39.7% |
| | No(0) | 469(95.3%) | 473(93.1%) | 49.8% |
| **pvd** | Yes(1) | 46(9.4%) | 54(10.6%) | 46.0% |
| | No(0) | 446(90.6%) | 454(89.4%) | 49.6% |
| **valvular_disease** | Yes(1) | 116(23.6%) | 128(25.2%) | 47.5% |
| | No(0) | 376(76.4%) | 380(74.8%) | 49.7% |
| **pacemaker** | Yes(1) | 18(3.7%) | 19(3.7%) | 48.6% |
| | No(0) | 474(96.3%) | 489(96.3%) | 49.2% |
| **pci** | Yes(1) | 10(2.0%) | 19(3.7%) | 34.5% |
| | No(0) | 482(98.0%) | 489(96.3%) | 49.6% |

# 3.1 Preliminary Analysis

## ❑ Correlation

There is no high correlation existed

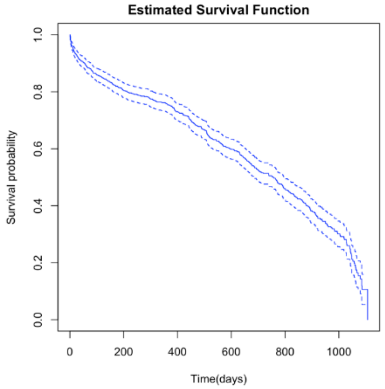


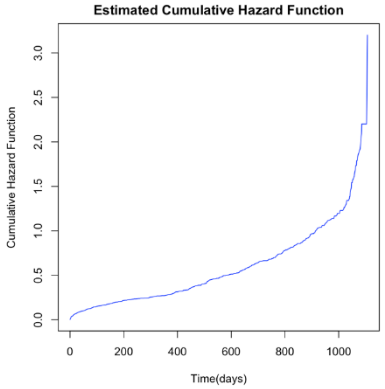
# 3.2 Analysis and Results

## ❑ Survival Function and Cumulative Hazard Function

- We used K-M estimator and N-A estimator to study the survival probability and cumulative hazard rate of death time in the dataset.
- $\hat{S}(748) = 0.5004 > 0.5$ and $\hat{S}(749) = 0.4984 < 0.5$, so the median time $\hat{x}_{0.5} = 749$ days.

| time | n.risk | n.event | surv | std.surv | cumhaz | std.chaz |
|---|---|---|---|---|---|---|
| **0** | 1000 | 3 | 0.9970 | 0.00173 | 0.0030 | 0.00173 |
| **1** | 992 | 9 | 0.9880 | 0.00350 | 0.0121 | 0.00349 |
| **2** | 973 | 7 | 0.9808 | 0.00444 | 0.0193 | 0.00442 |
| **3** | 963 | 5 | 0.9758 | 0.00501 | 0.0245 | 0.00499 |
| | | | **......** | | | |
| **748** | 246 | 0 | 0.5004 | 0.03790 | 0.6911 | 0.03783 |
| **749** | 245 | 1 | 0.4984 | 0.03812 | 0.6951 | 0.03805 |
| | | | **......** | | | |
| **1104** | 2 | 0 | 0.0815 | 0.37498 | 2.4324 | 0.33947 |
| **1107** | 1 | 1 | 0.0000 | Inf | 3.4324 | 1.05605 |





# 3.2 Analysis and Results

## ❑ Cox Model

We use all the variables to fit the Cox model and results are as follows. We can see age, los, gender, ihd and prior_dnas are significant variables and others are non-significant variables.

| Variable | coef | exp(coef) | se(coef) | Z-value | p-value | Variable | coef | exp(coef) | se(coef) | Z-value | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **los** | 0.013 | 1.014 | 0.003 | 4.120 | 3.7e-05 | **copd** | 0.095 | 1.100 | 0.105 | 0.900 | 0.366 |
| **age** | 0.062 | 1.064 | 0.006 | 10.760 | < 2e-16 | **obesity** | 0.108 | 1.114 | 0.224 | 0.480 | 0.629 |
| **gender** | -0.283 | 0.754 | 0.096 | -2.940 | 0.003 | **pvd** | 0.046 | 1.047 | 0.161 | 0.290 | 0.774 |
| **cabg** | -1.839 | 0.159 | 1.007 | -1.830 | 0.068 | **valvular_disease** | 0.191 | 1.211 | 0.109 | 1.760 | 0.079 |
| **diabetes** | -0.012 | 0.988 | 0.113 | -0.100 | 0.918 | **pacemaker** | 0.130 | 1.139 | 0.254 | 0.510 | 0.609 |
| **hypertension** | -0.034 | 0.967 | 0.096 | -0.360 | 0.722 | **prior_appts_attended** | -0.007 | 0.993 | 0.008 | -0.850 | 0.393 |
| **ihd** | 0.245 | 1.278 | 0.096 | 2.540 | 0.011 | **prior_dnas** | 0.134 | 1.144 | 0.039 | 3.420 | 0.001 |
| **arrhythmias** | -0.152 | 0.859 | 0.095 | -1.610 | 0.108 | **pci** | -0.182 | 0.834 | 0.327 | -0.560 | 0.579 |

# 3.2 Analysis and Results

## ❑ Model Selection

- We used forward stepwise selection method for the Cox PH model to find the best model for the heart failure data.

| Start: AIC=5914 | | Step: AIC=5778 | | Step: AIC=5765 | |
|---|---|---|---|---|---|
| **Surv(fu_time, death) ~ 1** | | **Survfu_time, death) ~ age** | | **Surv(fu_time, death) ~ age + los** | |
| Variables | AIC | Variables | AIC | Variables | AIC |
| age | 5778 | los | 5765 | prior_dnas | 5755 |
| los | 5891 | prior_dnas | 5767 | ihd | 5757 |
| …… | | …… | | …… | |

| Step: AIC=5755 | | Step: AIC=5748 | | Step: AIC=5745 | |
|---|---|---|---|---|---|
| **Surv(fu_time, death) ~ age + los + prior_dnas** | | **Surv(fu_time, death) ~ age + los + prior_dnas + gender** | | **Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd** | |
| gender | 5748 | ihd | 5745 | cabg | 5741 |
| ihd | 5750 | cabg | 5745 | arrhythmias | 5744 |
| …… | | …… | | …… | |

| Step: AIC=5741 | | Step: AIC=5740 | | Step: AIC=5739 | |
|---|---|---|---|---|---|
| **Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd + cabg** | | **Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd +cabg + arrhythmias** | | **Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd + cabg + arrhythmias + valvular_disease** | |
| arrhythmias | 5740 | valvular_disease | 5739 | none> | 5739 |
| valvular_disease | 5740 | none> | 5740 | copd | 5740 |
| …… | | …… | | …… | |

# 3.2 Analysis and Results

## ❑ Model Selection

We got 8 risk factors for the final model: **age, los, prior_dnas, gender, ihd, cabg, arrhythmias and valvular_disease.**

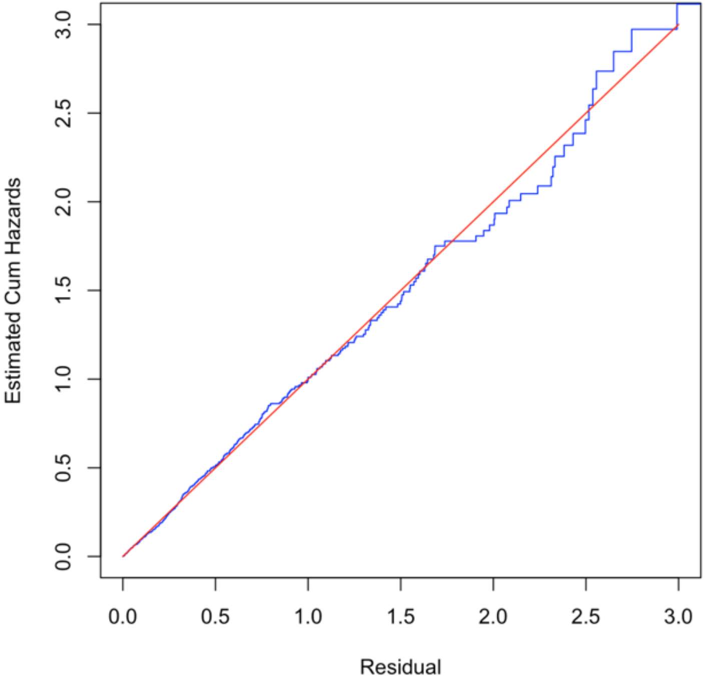
$$\boxed{\begin{array}{l} h(t|\text{age}, \text{los}, \text{prior\_dnas}, \text{gender}, \text{ihd}, \text{cabg}, \text{arrhythmias}, \text{alvular\_disease}) \\ = h_0(t)\exp(0.06195 \cdot \text{age} + 0.01355 \cdot \text{los} + 0.12041 \cdot \text{prior\_dnas} - 0.28157 \cdot \text{gender} \\ +0.23982 \cdot \text{ihd} - 1.81423 \cdot \text{cabg} - 0.16183 \cdot \text{arrhythmias} + 0.18865 \cdot \text{valvular\_disease}) \end{array}}$$

Table 5. Parameter Estimates for the Final Cox Model

| Variables | $\beta$ −coef | exp(coef) | se(coef) | Z-value | P-value |
|---|---|---|---|---|---|
| **age** | 0.062 | 1.064 | 0.006 | 11.040 | < 2e-16 |
| **los** | 0.014 | 1.014 | 0.003 | 4.290 | 1.8e-05 |
| **prior_dnas** | 0.120 | 1.128 | 0.036 | 3.380 | 0.001 |
| **gender** | -0.282 | 0.755 | 0.095 | -2.960 | 0.003 |
| **ihd** | 0.240 | 1.271 | 0.093 | 2.580 | 0.010 |
| **cabg** | -1.814 | 0.163 | 1.004 | -1.810 | 0.071 |
| **arrhythmias** | -0.162 | 0.851 | 0.091 | -1.770 | 0.076 |
| **valvular_disease** | 0.189 | 1.208 | 0.107 | 1.760 | 0.078 |

# 3.2 Analysis and Results

## ❑ Cox-Snell Residual plot

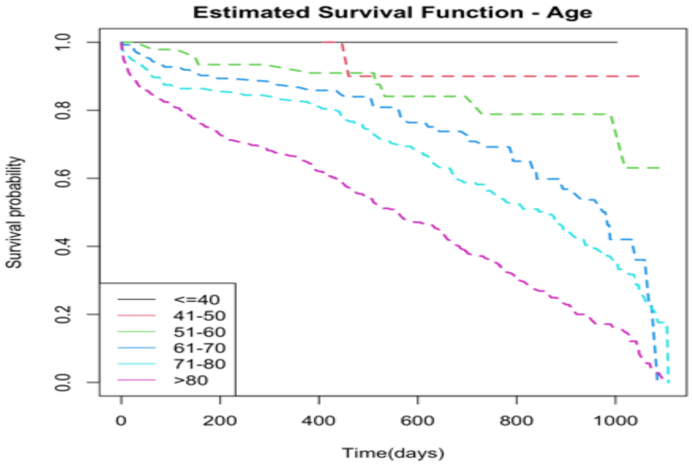


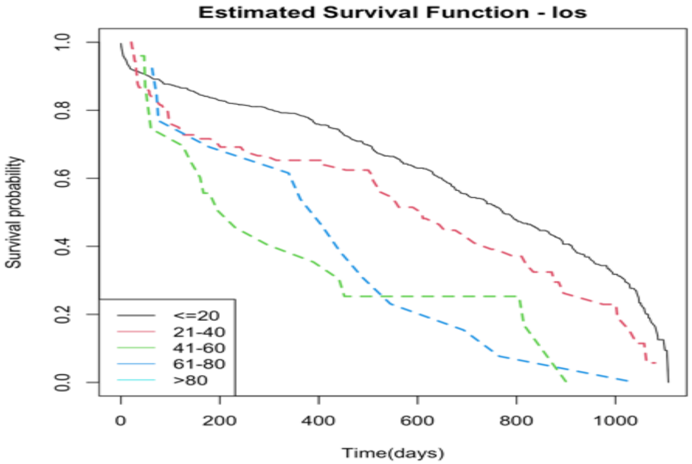
- We conducted a Cox-Snell Residual plot to access the fitness of our model.
- From the plot, we see that the estimated cumulative hazards follow closely to the 45 degree straight line. Therefore, we can conclude that the model is a good fit.

# 3.2 Analysis and Results

## ❑ Highly Correlated Features

Let’s focus on the highly correlated features with p-values <0.05 from the Cox PH model, including age, los, prior_dnas, gender, ihd.

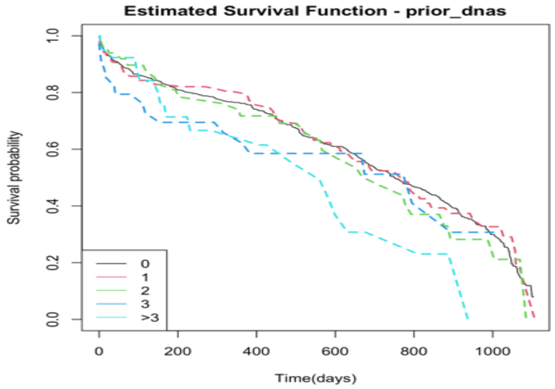


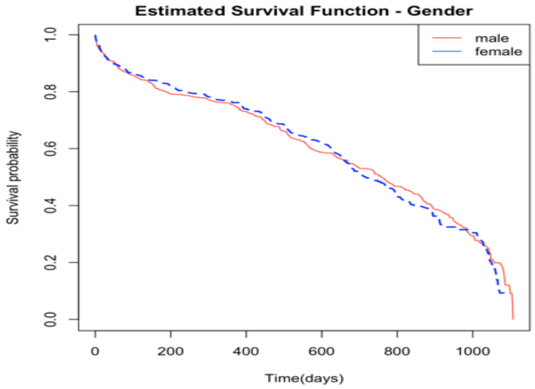


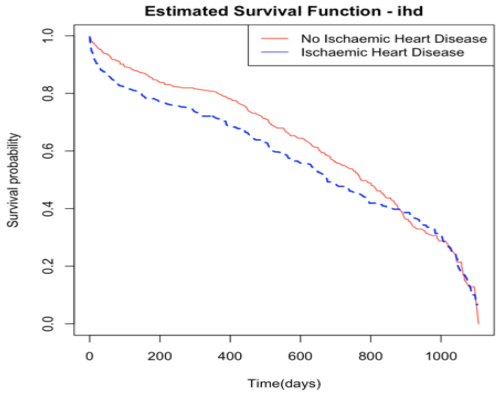
- For age, splitting patients up by age groups shows a large difference between each age group. In particular, patients younger than 40 have a survival probability of 1.
- For los, each groups also have obvious difference, but not as large as age, which is same to Table 5, as exp(coef) of age is 1.064 and exp(coef) of los is 1.014.

# 3.2 Analysis and Results

## ❑ Highly Correlated Features


Estimated Survival Function - prior_dnas


Estimated Survival Function - Gender


Estimated Survival Function - ihd

- For prior_dnas, when the number of appointments missed is less than 3, the survival curves don't have significant different, but if the number is more than 3, the survival probability is significant lower.
- For gender, the differences between the two curves here are not obvious, but the p-value=0.003 tells a different story.
- For ihd, the two curves here are different obviously, and patients who don't have ischaemic heart disease have a higher survival probability.

4

# Conclusion

01

**significant variables:**

los, age, gender, ihd and prior_dnas

02

**outputs:**

- ❑ Patients with high values of los(>61) and high values of prior_dnas(>3) have high death risks.
- ❑ The death rate increases with growing age and los.
- ❑ Patients with ihd (ischaemic heart disease) has significantly high mortality rate.
- ❑ Even no significant differences in the plots were found between gender for death risks, the p-value showed it was significant actually.

# Reference

- Data from https://www.kaggle.com/datasets/jackleenrasmybareh/heart-failure
- Mayo Clinic. Heart failure. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142
- British Heart Foundation( 2020). Heart failure: A blueprint for change.

# THANK YOU!