Heart Failure Survival Study

Mailman School of Public Health at Columbia University

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

Dec 7, 2023

Abstract

Heart failure (HF) results from weakened heart muscles, impairing blood pumping and causing symptoms like breathlessness (Ahmad et al., 2017). Statistics show HF affects 1-2% of adults, especially those over 70, potentially higher due to misdiagnosis (Jones et al., 2019). HF prevalence increased by 25% since 2002 due to aging, improved survival, and risk factors. This study utilizes data from the Institute of Cardiology and Allied Hospital in Faisalabad, Pakistan, which previously investigated the impact of key physiological and clinical factors on the prognosis of heart failure (HF) patients between April and December 2015. Our research employs a comprehensive range of statistical methodologies, including exploratory data analysis, nonparametric techniques, Kaplan-Meier curves, Cox proportional hazards models, and parametric survival models, aiming to ascertain the influence of important predictors such as creatinine levels, age, gender, ejection fraction, blood pressure, anemia, and serum sodium on HF. Our findings aim to refining risk assessment models, thereby strengthening clinical decision-making and optimizing patient care in the future. XXXX (conclusion)

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

Our project aims to assess the influence of key physiological and clinical factors on the outcomes of heart failure patients at the Institute of Cardiology and Allied Hospital, Faisalabad, Pakistan, during April-December 2015. We will examine variables such as creatinine levels, gender, age, ejection fraction, blood pressure, anemia, and serum sodium to determine their impact on patient prognosis. Utilizing a range of analytical techniques including exploratory data analysis, nonparametric methods, Kaplan-Meier curves, Cox proportional hazards modeling, parametric survival models, and validation procedures, our study is designed to identify crucial predictors of patient outcomes and their interrelationships. The insights gained from this analysis are expected to contribute significantly to the development of tailored treatment strategies and improved risk stratification models, thereby enhancing clinical decision-making and patient care for heart failure management.

2 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_cat**: Ejection fraction (Low: $EF \leq 30$, Medium: $30 < EF \leq 45$ and High: $EF > 45$)
- **sodium**: Sodium in mEq/L
- **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 indicator equals 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.

Table 1: Descriptive Statistics Table

	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_cat)				
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)	0.002
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]	

Based on the descriptive table, we can observe that the mean survival time for deletions and events is 158 days and 70.9 days, respectively. Since we have the complete dataset, there is no need to worry about missingness issue. The table also lists the p-values for each variable. Some variables have relatively large p-values. However, we still need to check the distribution of each variable [1] and determine which variables need to be transformed.

[1] Histograms for continuous variables and bar charts for categorical variables

Figure 1.1: Histograms of Continuous Variables

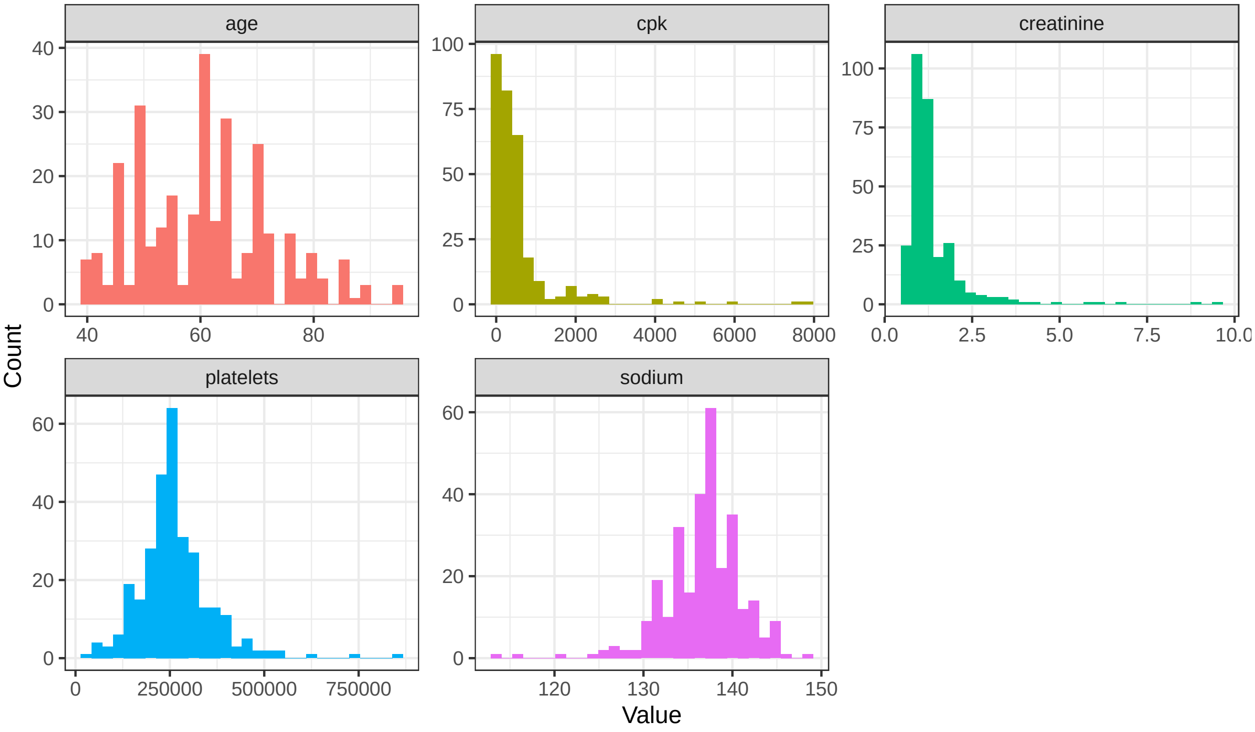
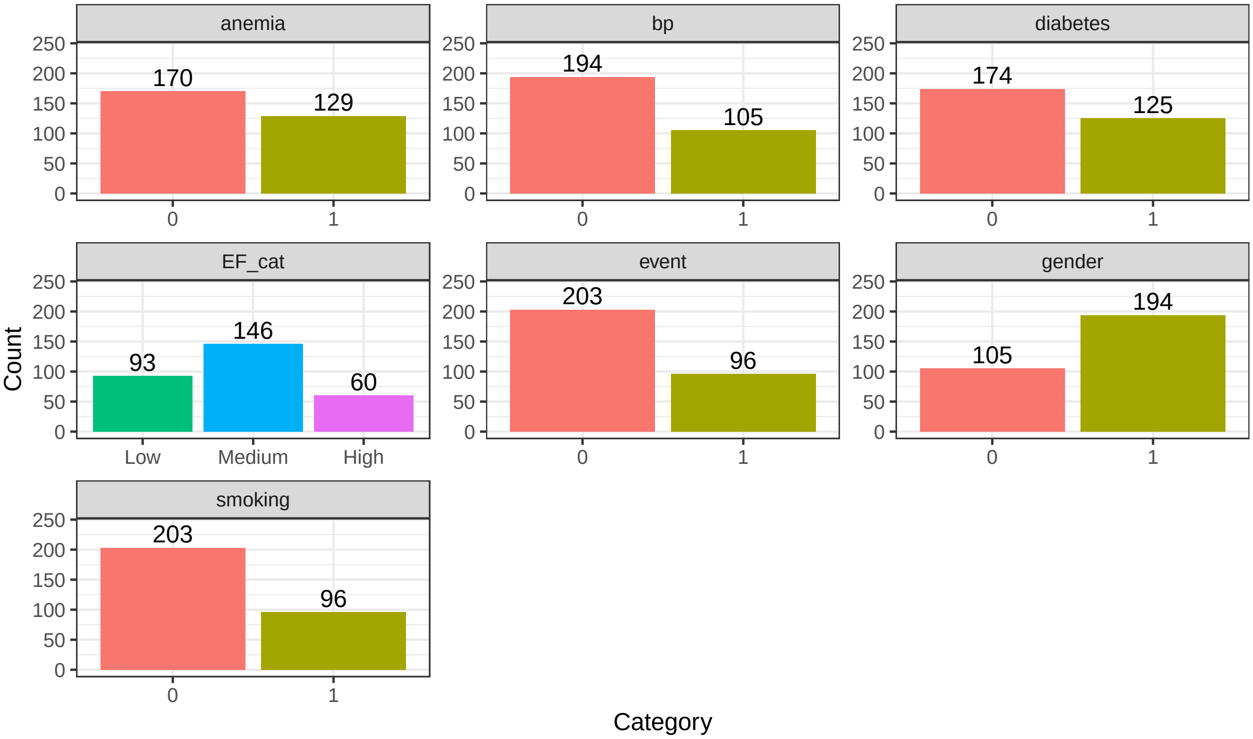


Figure 1.2: Bar Charts of Categorical Variables



After checking the histograms, two continuous variables creatinine phosphokinase (cpk) and serum creatinine are right-skewed. We decide to log-transformed both of them for further model fitting.

3 Methods

3.1 Nonparametric Methods

In our analysis, we applied life table, Kaplan-Meier and Fleming-Harrington Curves to estimate the survival function. All three approaches can handle censored data.

The life table is particularly useful for larger sample sizes and when the data are grouped into intervals. The survival probability at each interval is estimated as:

$$\hat{S}_L(t_i) = \prod_{t_{i-1}<t} (1 - \frac{d_i}{n'_i})$$

where d_i is the number of events in interval[2] $[t_{i-1}, t_i]$, n'_i is the average number at risk in the interval $[t_{i-1}, t_i]$.

The Kaplan-Meier curve allows for varying follow-up times and censored data, making it versatile for smaller samples and individual subject data. It provides a visual representation of the survival experience of the cohort over time. The Kaplan-Meier estimate of survival function can be mathematically expressed as:

$$\hat{S}_K(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{n_i}]$$

where d_i is the number of events at time t_i and n_i is the number at risk at t_i^-, and c_i is the number of censored during the interval $[t_i, t_{i+1}]$

Unlike the Kaplan-Meier estimator, the Fleming-Harrington estimator[3] is designed to weight events differently over time in survival analysis. It focuses on estimating the cumulative hazard function. The estimated survival probability can be computed as:

$$\hat{S}_F(t) = \prod_{t_i \leq t} \exp[-\frac{d_i}{n_i}]$$

where the d_i, n_i and c_i conditions are the same as K-M estimator above, It is true that $\hat{S}_F(t) \geq \hat{S}_K(t)$ because $\exp{[-\frac{d_i}{n_i}]}$ is always great and equal to $1 - \frac{d_i}{n_i}$.

Both K-M and F-H survival curves are presented in the **Result** section for further comparison.

3.2 Hypothesis Testing

The Log-Rank test focuses on comparing the number of observed to expected events across the groups at each time point. The test statistic[4] is calculated as:

$$\frac{L}{\sqrt{\mathrm{Var}(L)}} = \frac{\sum_{i=1}^{k}(d_{0i} - e_{0i})}{\sum_{i=1}^{k}\sqrt{\frac{n_{0i}n_{1i}d_i(n_i-d_i)}{n_i^2(n_i-1)}}} \sim N(0,1)$$

The Gehan's Wilcoxon test[5] gives more weight to events at earlier time points. It achieves a greater sensitivity to differences in survival that manifest at the beginning of the observation period. The test statistic is calculated as:

[2] This is the survival function at the end of interval. However, R often reports the survival function at the beginning of the interval

[3] Also known as Nelson-Aalen estimator

[4] This is the log-rank test statistics with tie

[5] Also known as the Breslow test

$$\frac{L}{\sqrt{\mathrm{Var}(L)}} = \frac{\sum_{i=1}^{k} n_i(d_{0i} - e_{0i})}{\sum_{i=1}^{k} \sqrt{\frac{n_{0i}n_{1i}d_i(n_i-d_i)}{n_i-1}}}$$

The Gehan's Wilcoxon test is actually a special case of weighted log-rank test with weight equals n_i.

3.3 Proportional Hazard Models

We use three propositional hazard models to evaluate the effect of several factors on survival time. It allows us to examine how specified factors influence the rate of the event that we are interested in at a particular point in time. This rate is the hazard rate.

Proportional hazard model is the primary regression model to investigate the effectiveness of treatment X over survival time T, where the i_{th} patient at a time t
is

$$h_i(t) = h_0(t)\exp[\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}]$$

where

- $h_0(t)$ is the baseline hazard function
- $h(t)$ is the hazard function determined by a set of p covariates $(x_1,\ x_2,\ ...,\ x_p)$
- $(\beta_1,\ \beta_2,\ ...,\ \beta_p)$ are the coefficients which measure the impact of covariates.

The **proportional hazard** can be expressed as ratio of two hazard functions at time t in two individuals or groups with covariates X and X', and does not depend on t.

$$\frac{h(t|X=x)}{h(t|X=x')} = e^{\beta(x-x')}$$

There are different ways to formulate the baseline hazard function $h_0(t)$, which lead to different models and estimations.

3.3.1 Cox PH Model

The Cox PH model is a semi-parametric model which does not assume a particular baseline hazard function $\tilde{h}_0(t)$. In contrast to parametric PH models which use the full likelihood, the Cox PH model is formulated by partial likelihood because there is very limited information on β beyond $L_p(\beta)$. The model is given by:

$$h_i(t) = \tilde{h}_0(t)\exp[\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}]$$

3.3.2 Weibull Model

The Weibull model is a popular parametric model wihch assumes a specific functional form for the hazard rate, either increase or decrease over time. Its parameters are intuitively interpretable, with the shape parameter distinctly indicating whether the hazard rate is increasing, decreasing, or constant.

$$h_i(t) = \lambda\gamma t^{\gamma-1}\exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip})$$

where λ is the Scale parameter, and γ is the Shape parameter.

3.3.3 Gompertz Model

The Gompertz model, a parametric approach in survival analysis, assumes an exponential increase or decrease in the hazard function over time. This approach is suitable for studies where hazard rates show distinct exponential trends.

$$h_i(t) = \lambda \exp(\eta t + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip})$$

where λ is the Scale parameter, and η is the rate of increase or decrease in hazard.

3.4 Accelerated Failure Time Model

The Accelerated Failure Time (AFT) model is a parametric model assumes that the effect of covariates is to accelerate or decelerate the life course of a disease by some constant. In the AFT model, the key assumption is that the logarithm of survival time follows a linear relationship with the covariates. Mathematically, it can be expressed as:

$$log(T) = \mu + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

3.5 Model Selection

3.5.1 Survival Tree

The survival tree method is a non-parametric approach used in survival analysis for model selection and identifying significant predictors of time-to-event outcomes. It involves segmenting the data into homogeneous subgroups based on the values of explanatory variables. The method constructs a tree structure where each node represents a subset of the dataset, and each split is based on the value of a predictor variable that best separates the data in terms of survival. Mathematically, this process involves recursively partitioning the data to maximize a criterion like the log-rank statistic. At each node, the split is chosen to maximize the difference in survival between the resulting subgroups. This can be represented as:

$$\text{Max}_{v,s}[\chi^2_{LR}(v, s)]$$

where $\chi^2_{LR}(v, s)$ is the log-rank test statistic computed for a split s on variable v. The process continues until a stopping criterion is met, typically based on the minimum number of observations in a node or a minimum improvement in the survival difference. The result is a tree where the paths from the root to the leaves represent rules for predicting survival, offering a visual and interpretable model of the factors affecting the time to an event like death in this study.

3.5.2 Stepwise Selection

In model selection process, we incorporated bidirectional stepwise selection, alongside using various model assessment criteria such as Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc), and Schwarz Bayesian Criterion (SBC). This comprehensive approach enhanced our ability to identify the most appropriate model for our data.

We chose the final model based on AIC in our survival analysis due to its effective balance between model complexity and fit, particularly valuable in preventing overfitting in complex models. Additionally, given our sufficiently large sample size, AIC provided a more appropriate measure compared to AICc, and its less stringent penalty compared to SBC was better suited for our data's scale and complexity. *Collett (1994)* suggested that the AIC for survival models should be:

$$AIC = -2\partial \log(likeligood) + 2k$$

In this equation, k represents the number of parameters in the model. The term $2k$ serves as a penalty to discourage overfitting by complex models.

For different survival models, the AIC adapts as follows:

$$AIC = -2\log(likeligood) + 2(p + 2 + k)$$

where $k = 0$ for the exponential model, $k = 1$ for the Weibull, log-logistic and log-normal models, and $k = 2$ for the generalized gamma model.

3.6 Model Validation

3.6.1 ROC Curves and AUC

In our analysis, we employed time-dependent Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values to evaluate the discriminative ability of our Cox proportional hazards model over time. Specifically, we focused on two clinically relevant time points: 50 days and 250 days (Ahmad et al., 2017). The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. AUC, a key summary measure of the ROC curve, quantifies the overall ability of the model to discriminate between individuals who will experience the event and those who will not, irrespective of the chosen probability threshold (Heagerty & Zheng, 2005). Higher AUC values indicate better discriminative ability.

3.6.2 C-Index

The concordance index (C-index) was calculated to assess the predictive accuracy of our model. This metric is a measure of the model's ability to correctly rank the survival times of pairs of individuals, considering censored data (Steyerberg & Vergouwe, 2014). The C-index is calculated through pairwise comparisons, where a pair is concordant if the individual predicted to have a shorter survival time indeed experiences the event earlier than the other individual in the pair (Steyerberg & Vergouwe, 2014). A C-index of 0.5 suggests no better predictive accuracy than random chance, while a value of 1 indicates perfect prediction.

3.6.3 Calibration Slope

To evaluate the calibration of our model, we focused on the calibration slope. Calibration reflects the agreement between observed outcomes and predicted probabilities and the calibration slope assesses whether the predicted risks are of the correct magnitude (Steyerberg & Vergouwe, 2014). A slope of 1 indicates perfect calibration, meaning the model's predicted probabilities are accurately scaled. We calculated the calibration slope using logistic regression within a bootstrap framework, which allowed us to robustly assess the scale of the predicted risks relative to the actual event occurrences. The bootstrap approach, involving resampling the dataset 400 times, provided a more comprehensive understanding of the model's calibration under varying sample conditions.

4 Results

4.1 Nonparametric Methods

4.1.1 Life Table

Table 2.1: Heart Failure Life Table (Male)

	tstart	tstop	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-30	0	30	194	1	193.5	23	1.0000000	0.0039621	0.0042125	0.0000000	0.0007755	0.0008766
30-60	30	60	170	4	168.0	12	0.8811370	0.0020979	0.0024691	0.0232651	0.0005862	0.0007123
60-90	60	90	154	25	141.5	9	0.8181986	0.0017347	0.0021898	0.0278070	0.0005626	0.0007295
90-120	90	120	120	20	110.0	5	0.7661577	0.0011608	0.0015504	0.0309802	0.0005094	0.0006932
120-150	120	150	95	18	86.0	2	0.7313323	0.0005669	0.0007843	0.0332571	0.0003970	0.0005546
150-180	150	180	75	2	74.0	5	0.7143246	0.0016088	0.0023310	0.0345899	0.0006991	0.0010418
180-210	180	210	68	26	55.0	3	0.6660594	0.0012110	0.0018692	0.0384014	0.0006834	0.0010787
210-240	210	240	39	16	31.0	2	0.6297289	0.0013543	0.0022222	0.0416431	0.0009305	0.0015705
240-270	240	270	21	16	13.0	1	0.5891013	0.0015105	0.0026667	0.0478504	0.0014564	0.0026645
270-300	270	300	4	4	2.0	0	0.5437858	0.0000000	0.0000000	0.0620201	NaN	NaN
300-Inf	300	Inf	0	0	0.0	0	0.5437858	NA	NA	0.0620201	NA	NA

Table 2.2: Heart Failure Life Table (Female)

	tstart	tstop	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-30	0	30	105	3	103.5	8	1.0000000	0.0025765	0.0026801	0.0000000	0.0008750	0.0009468
30-60	30	60	94	0	94.0	9	0.9227053	0.0029448	0.0033520	0.0262504	0.0009372	0.0011159
60-90	60	90	85	10	80.0	6	0.8343612	0.0020859	0.0025974	0.0367098	0.0008241	0.0010596
90-120	90	120	69	15	61.5	4	0.7717841	0.0016732	0.0022409	0.0419136	0.0008140	0.0011198
120-150	120	150	50	5	47.5	2	0.7215868	0.0010128	0.0014337	0.0460937	0.0007039	0.0010135
150-180	150	180	43	3	41.5	2	0.6912042	0.0011104	0.0016461	0.0489040	0.0007700	0.0011636
180-210	180	210	38	12	32.0	3	0.6578931	0.0020559	0.0032787	0.0519106	0.0011416	0.0018907
210-240	210	240	23	10	18.0	0	0.5962156	0.0000000	0.0000000	0.0579853	NaN	NaN
240-270	240	270	13	11	7.5	0	0.5962156	0.0000000	0.0000000	0.0579853	NaN	NaN
270-300	270	300	2	2	1.0	0	0.5962156	0.0000000	0.0000000	0.0579853	NaN	NaN
300-Inf	300	Inf	0	0	0.0	0	0.5962156	NA	NA	0.0579853	NA	NA

Table 2.1 and Table 2.2 represent the lifetable with a time break of 30 days (one month), stratified by gender. According to the table, we find that the last line (270-300 days) shows a survival probability larger than 0.5 for both genders (0.54 for male and 0.59 for female). This may indicate a high life expectancy and improved health care, where a significant proportion of individuals are expected to live longer than 300 days (10 months). Moreover, males have a relatively shorter survival time than females based on the lifetable. This hypothesis needs future testing in the following modeling fit.

4.1.2 The Kaplan-Meier and Fleming-Harrington Model

Figure 2.1. Kaplan−Meier Survival Curve By Gender

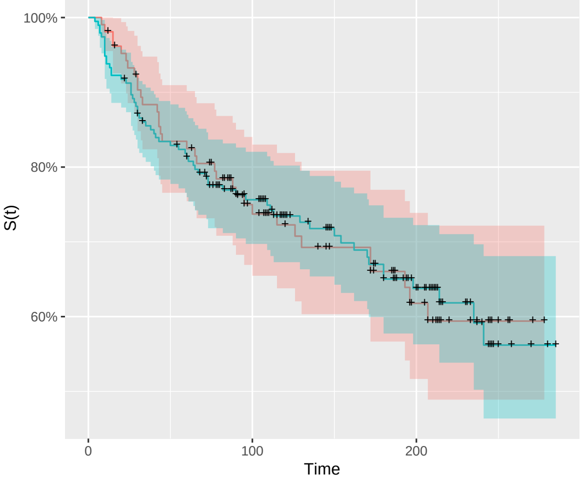
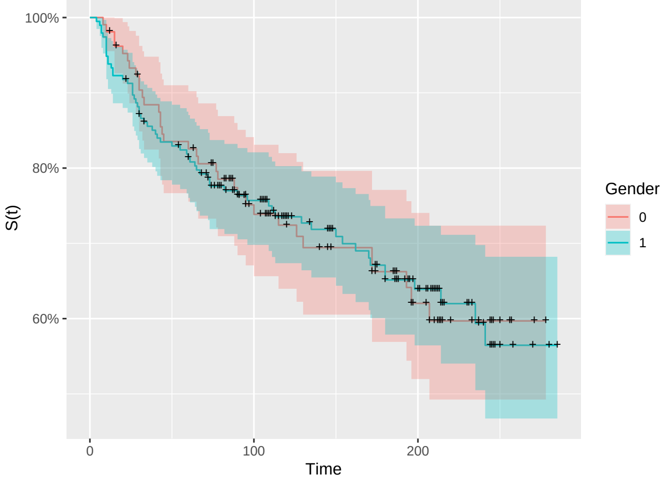


Figure 2.2. Fleming−Harrington Survival Curve By Gender



The Kaplan-Meier and Fleming-Harrington have similar trends and show no significant difference between genders. Given that the p-value is relatively high (p-value = 0.9) for both Kaplan-Meier and Fleming-Harrington estimators, we can further make sure that no significant difference appears in the survival experience between males and females.

4.2 Hypothesis Testing

4.2.1 Log-Rank test and Wilcoxon test

Gender	N	Observed	Expected	Log_Rank	Wilcoxon
0	105	34	34.3	0.00254	0.00397
1	194	62	61.7	0.00141	0.00397

Test	Chi_square	df	p_value
Log Rank	0	1	0.9
Wilcoxon	0	1	0.9

The Log-Rank test and the Wilcoxon test can be used to test for differences in survival experience between genders. Reviewing the results from the table, we find that both the Log-Rank test and the Wilcoxon test have provided a similar result and gave a p-value of 0.9. This indicates that we failed to reject the null hypothesis. Therefore, we conclude that there is no statistically significant difference in survival experiences between males and females.

4.3 Model Selection

$$h(t) = h_0(t)\exp[\beta_1 age + \beta_2 EF_{Medium} + \beta_3 EF_{High} + \beta_4 bp + \beta_5 sodium + \beta_6 \log(creatinine + 1)]$$

4.4 Semi-parametric Models

The final model obtained from Stepwise Selection using AIC contains creatinine, age, ejection fraction, blood pressure status, sodium covariates.

4.5 Parametric Models

(Model checking, fitting the model)

4.6 Model Validation

4.6.1 ROC Curves and AUC

Our time-dependent ROC analysis at 50 days yielded an AUC of approximately 0.738, while at 250 days, the AUC was 0.935. These values indicate that the model's ability to discriminate between those who will experience the event and those who will not improves over time. The ROC curves further visually demonstrate this improvement, with the curve for 250 days being closer to the top left corner, indicating better performance.

4.6.2 C-Index

The average C-index calculated through bootstrapping (n = 400) was 0.746. This suggests that in about 75% of pairwise comparisons, our model correctly ranks the survival times. A C-index of around 0.75 is generally indicative of good predictive ability, especially in clinical settings where accurate risk stratification is crucial for treatment planning.

4.6.3 Calibration Slope

Our calculated mean calibration slope came to approximately 1.1529. This value, slightly above the ideal of 1, is significant in understanding the model's performance. The calibration slope measures the extent to which the model's predicted risks are proportionate to the observed risks. A value of 1 would indicate perfect calibration, meaning the model's predictions are perfectly aligned with the actual observed risks. Our finding of a calibration slope above 1 suggests that our model may be mildly overfitting the data, predicting slightly higher risks than what is observed.

5 Discussion

According to the life table, we find that there is a slight difference in survival probability between genders. However, the hypothesis test and model selection show that there is no difference. Similarities in heart failure presentation, treatment regimen, and sample size may explain the absence of sex-based differences in survival. Specifically, if the severity of heart failure was not related to gender or if patients received appropriate gender-based treatment and the sample size or characteristics of the study limited gender-specific analyses, potential differences could be masked.

The nuanced findings from our analysis provide a comprehensive view of our Cox model's performance. While the model exhibits strong discriminative ability, as indicated by the AUC values and C-index, our calibration assessment, particularly the calibration slope, suggests areas where improvement is needed. Notably, the calibration slope, slightly over the ideal value of 1, implies a mild overestimation in risk predictions. This indicates a complex calibration scenario where the model might be overfitting to some extent.

Such overestimation, although modest, is critical in clinical settings. Accurate risk prediction is vital for informed decision-making and effective patient management. Overestimated risks might lead to more aggressive interventions than necessary, affecting patient care and resource allocation. Conversely, underestimating risks could result in missed opportunities for timely intervention. This highlights the importance of achieving a balance in predictive accuracy, ensuring that the model neither overestimates nor underestimates risks.

The observed improvement in the model's discriminative ability over time, with increasing AUC values from 50 to 250 days, underscores the dynamic nature of risk factors and their evolving impact on patient outcomes. However, the calibration results emphasize the need to focus not just on the model's ability to discriminate but also on the accuracy of its probability predictions.

Future work should, therefore, focus on refining the model's complexity and variable selection. Re-evaluating the model's components and considering alternative modeling approaches might help in aligning the predicted probabilities more closely with actual outcomes. Applying more advanced calibration techniques could also address the observed overfitting, enhancing the model's reliability. Additionally, external validation on an independent dataset is essential to confirm the model's effectiveness and applicability in different clinical contexts. Such efforts will be crucial in enhancing the model's utility and ensuring its robustness in real-world clinical applications, where precise risk assessment directly informs patient care strategies.

6 Conclusions

The non-parametric method showed that more than 50% of both genders (54% for males and 59% for females) can survive for over 300 days. This may indicate a high life expectancy and improved health care. Moreover, the non-parametric method showed no significant difference in survival probability between genders. This result was later consistent with the results of the model selection.

7 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. Collett, D. (1999). Modelling survival data in medical research. Chapman & Hall/CRC.
3. 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

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European Heart Journal, 35(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., & Omar, R. Z. (2015). How to develop a more accurate risk prediction model when there are few events. BMJ, h3868. https://doi.org/10.1136/bmj.h3868

Heagerty, P. J., & Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. Biometrics, 61(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

8 Appendix

8.1 Code

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
library(biostat3)
library(tidyverse)
library(knitr)
library(kableExtra)
library(survival)
library(survminer)
library(ggfortify)
library(ggsurvfit)
library(patchwork)
library(writexl)
library(readxl)
library(table1)
library(rmarkdown)
library(KMsurv)
library(StepReg)
library(ggplot2)
library(timeROC)
library(boot)
library(rms)
library(survivalROC)
library(rpart)
data = read_csv("./data/heart_failure.csv")
dat <- data %>%
  arrange(TIME) %>% janitor::clean_names() %>%
  mutate(ejection_fraction_cat = 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_cat = factor(ejection_fraction_cat,
                                        levels = c("Low", "Medium", "High"))) %>%
  rename(platelets = pletelets,
         anemia = anaemia,
         EF = ejection_fraction,
         EF_cat = ejection_fraction_cat)

# 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"
```

```r
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$EF_cat) = "Ejection Fraction (EF_cat)"
label(dat_table$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)))
}
caption = "Table 1: Descriptive Statistics Table"

table1 = table1(~ time + age + gender + smoking + diabetes + bp + EF_cat +
                  anemia + sodium + creatinine + cpk + platelets | event,
                  data = dat_table,
                extra.col = list(`P-value` = pvalue), caption = caption)
t1kable(table1) %>% kable_styling(font_size = 8, latex_options = "HOLD_position")
# 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 1.1: Histograms of Continuous Variables") +
  theme_bw() +
  theme(legend.position = "none")
cont_hist

# Data contains the categorical vars only
```

```r
cate_data = dat %>%
  select(event, gender, smoking, diabetes, bp, anemia, EF_cat)
# Long format
cate_dat.long = cate_data %>%
  pivot_longer(cols = c(event, gender, smoking, diabetes, bp, anemia, EF_cat))
# 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 = -0.3) +
  facet_wrap(~name, scales = "free") +
  labs(x = "Category", y = "Count", fill = "Category",
       title = "Figure 1.2: Bar Charts of Categorical Variables") +
  theme_bw() +
  theme(legend.position = "none") +
  ylim(0, 240)
cate_barplot
dat_log = dat %>%
  mutate(cpk_log = log(cpk + 1),
         creatinine_log = log(creatinine + 1))
# aft.fit <- survreg(Surv(time, event) ~ gender + smoking + diabetes + bp + anaemia +
#                      age + ef_cat+ sodium + pletelets + logcre + logcpk,
#                    data = model_data, dist = "exponential")
# summary(aft.fit)
nonpara_dat = dat_log %>%
  dplyr::select(-c(EF, cpk, creatinine)) %>%
  relocate(time, event, EF_cat, smoking, everything()) %>%
  mutate(event = as.numeric(event) - 1)
nonpara_male = nonpara_dat %>% filter(gender == 1)
nonpara_female = nonpara_dat %>% filter(gender == 0)
life_table_male <- lifetab2(Surv(time, event) ~ 1, data = nonpara_male,
                            breaks = seq(0, 300, 30))
life_table_female <- lifetab2(Surv(time, event) ~ 1, data = nonpara_female,
                              breaks = seq(0, 300, 30))
life_table_male %>% kable(booktabs = T,
                          caption = "Table 2.1: Heart Failure Life Table (Male)") %>%
  kable_styling(latex_options = c("HOLD_position"), font_size = 6)
life_table_female %>% kable(booktabs = T,
                            caption = "Table 2.2: Heart Failure Life Table (Female)") %>%
  kable_styling(latex_options = c("HOLD_position"), font_size = 6)
km = survfit(Surv(time, event) ~ gender, data = nonpara_dat)
fh <- survfit(Surv(time, event) ~ gender, data = nonpara_dat, type = "fh")
km_plot =
  km %>% autoplot() +
  labs(y = "S(t)",
       x = "Time",
       subtitle = "Figure 2.1. Kaplan-Meier Survival Curve By Gender",
       color = "Gender", fill = 'Gender') + theme(legend.position = "none")
fh_plot =
  fh %>% autoplot() +
  labs(y = "S(t)",
       x = "Time",
       subtitle = "Figure 2.2. Fleming-Harrington Survival Curve By Gender",
       color = "Gender", fill = 'Gender')
```

```
km_plot + fh_plot
results1 <- data.frame(
  Gender = c(0, 1),
  N = c(105, 194),
  Observed = c(34,62),
  Expected = c(34.3, 61.7),
  Log_Rank = c(0.00254, 0.00141),
  Wilcoxon = c(0.00397, 0.00397)
)

results2 <- data.frame(
  Test = c('Log Rank', 'Wilcoxon'),
  Chi_square = c(0, 0),
  df = c(1, 1),
  p_value = c(0.9, 0.9)
)

results_table1 <- kable(results1, align = "c") %>%
  kable_styling(latex_options = c("HOLD_position"))

results_table2 <- kable(results2, align = "c") %>%
  kable_styling(latex_options = c("HOLD_position"))

results_table1
results_table2
# # Calculate predicted risks
# predicted_risks <- predict(step_model, newdata = model_data, type = "risk")
#
# # Time points for ROC analysis
# time_points <- c(50, 250)
#
# # Calculate ROC curves at specified times
# roc_50 <- timeROC(T = model_data$time, delta = model_data$event, marker = predicted_risks, times = 50
# roc_250 <- timeROC(T = model_data$time, delta = model_data$event, marker = predicted_risks, times = 2
#
# # Extract AUC values
# auc_50 <- roc_50$AUC
# auc_250 <- roc_250$AUC
#
# # Print AUC values
# print(paste("AUC at 50 days:", auc_50))
# print(paste("AUC at 250 days:", auc_250))
#
# # Plot ROC curves
# plot(roc_50$FP, roc_50$TP, type = "l", col = "red", xlab = "1 - Specificity", ylab = "Sensitivity", m
# lines(roc_250$FP, roc_250$TP, type = "l", col = "blue")
# legend("bottomright", legend = c("50 days", "250 days"), col = c("red", "blue"), lty = 1)
# abline(0, 1, col = "black", lty = 2)
# Define the function for calculating C-statistic
# boot_c_statistic <- function(original_data, indices) {
#   # Creating a bootstrap sample
#   boot_data <- original_data[indices, ]
#
```

```
#   # Fit the Cox model to the bootstrap sample
#   fit <-   coxph(Surv(time, event) ~ log(creatinine+1) + age + ef_cat + bp +
#                       sodium, data = boot_data)
#
#   # Calculate the concordance statistic using the updated function
#   concordance <- concordance(fit)$concordance
#   return(concordance)
# }
#
# # Perform bootstrapping for C-statistic
# set.seed(123) # for reproducibility
# boot_results_c_stat <- boot(data = model_data, statistic = boot_c_statistic, R = 400)
#
# # Calculate the average C-statistic
# mean_c_stat <- mean(boot_results_c_stat$t)
# print(mean_c_stat)
# # Define the bootstrap function for calibration metrics using logistic regression
# boot_calibration_logistic <- function(original_data, indices) {
#   boot_data <- original_data[indices, ]
#   fit <-  coxph(Surv(time, event) ~ log(creatinine+1) + age + ef_cat + bp +
#                       sodium, data = boot_data)
#
#
#   # Predicted risks for the original dataset
#   predicted_risks <- predict(fit, newdata = original_data, type = "risk")
#
#   # Fit a logistic model for calibration
#   calibration_model_logistic <- glm(event ~ predicted_risks, data = original_data, family = "binomial
#
#   # Calibration slope (coefficient of predicted_risks)
#   calibration_slope_logistic <- coef(calibration_model_logistic)["predicted_risks"]
#
#   return(calibration_slope_logistic)
# }
#
# # Perform bootstrap
# set.seed(123)
# boot_results_logistic <- boot(data = model_data, statistic = boot_calibration_logistic, R = 400)
#
# # Calculate the average calibration slope
# mean_calibration_slope_logistic <- mean(boot_results_logistic$t)
# print(mean_calibration_slope_logistic)
```