Survival Analysis of Clinical, Lifestyle, and Body Predictors in Cardiomyopathy Patients

Jason Withrow

December 2025

Introduction

Cardiomyopathy is a chronic disease of the heart muscle that makes it difficult for the heart to pump adequate amounts of blood to the rest of the body. Understanding the clinical, lifestyle, and body characteristics that influence the survival of patients with the disease is fundamental in the direction of care. This report analyzes a dataset of cardiomyopathy patients, which includes 11 clinical, lifestyle, and body variables, as well as the length of follow-up for each patient.

The main research questions are:

1. Which clinical, lifestyle, and body characteristics on their own show a significant relationship with patient survival (death)?

2. Do the predictors identified as significant in the univariate analyses remain associated with time-to-death when evaluated together in a multivariable survival model, and how does this model compare to one that includes all predictors?

3. Which of the two models provides a better explanation of patient survival

Data Description

The dataset contains clinical data on 299 patients and contains no missing values. The following table gives descriptions of the variables in the dataset.

Variable	**Description**
age	Age of the patient in years
anaemia	1 if the patient has anemia, 0 otherwise
hypertension	1 if the patient has hypertension, 0 otherwise
cpk	Creatinine phosphokinase level (mcg/L)
diabetes	1 if the patient has diabetes, 0 otherwise
ejection_percentage	Percentage of blood pumped by left ventricle
platelets	Platelets in blood (kiloplatelets/mL)
gender	1 if male, 0 if female
creatinine	Creatinine level (mg/dL)
smoking	1 if patient smokes, 0 otherwise
sodium	Sodium level (mEq/L)
follow_up	Length of follow-up (days)
death	1 If the patient died 0 if still alive

Table 1: Data Dictionary

Exploratory Data Analysis

To provide proper context for the forthcoming analysis, it is important to understand some of the summary statistics of the variables. Table 2 and 3 give summary statistics for the continuous variables as well as counts and proportions of the categorical variables

Variable	Mean	SD	Median	Min	Max
age	60.83	11.89	60.00	40.00	95.00
cpk	581.80	970.29	250	23.00	7861.00
ejection_percentage	38.08	11.83	38.00	14.00	80.00
platelets	263358.03	97804.24	262000.00	25100	850000
creatinine	1.39	1.03	1.10	0.50	9.40
sodium	136.63	4.41	137.00	113.00	148.00
follow_up	130.26	77.61	115.00	4.00	285.00

Table 2: Summary statistics for continuous variables

Variable	Num Categories	Counts	Proportions
anaemia	2	0: 170 , 1: 129	0: 56.86% , 1: 43.14%
diabetes	2	0: 174 , 1: 125	0: 58.19% , 1: 41.81%
hypertension	2	0: 194 , 1: 105	0: 64.88% , 1: 35.12%
gender	2	0: 105 , 1: 194	0: 35.12% , 1: 64.88%
smoking	2	0: 203 , 1: 96	0: 67.89% , 1: 37.11%
death	2	0: 203 , 1: 96	0: 67.89% , 1: 37.11%

Table 3: Summary statistics for categorical variables

The continuous variables show moderate variability overall. cpk is strongly right-skewed, with extreme high values pulling the mean far above the median. Platelets and follow-up time also span wide ranges. In contrast, age, ejection percentage, creatinine, and sodium are more tightly distributed.

The categorical variables show relative balanced; none of the classes are extremely unbalanced. The largest difference is with variables `smoking` and `death` where 67.89% of patients belong to the negative class and 37.11% belong to the positive class, but in regards to statistical analysis this is not that noticeable of an imbalance especially for the dependent variable `death`.

Next the distribution of follow-up times across patients is shown in Figure X. The distribution is bimodal, with one group of patients having shorter observation periods and another group having longer periods. This pattern may reflect differences in enrollment, censoring, or patient subgroups, and should be considered when modeling survival outcomes.

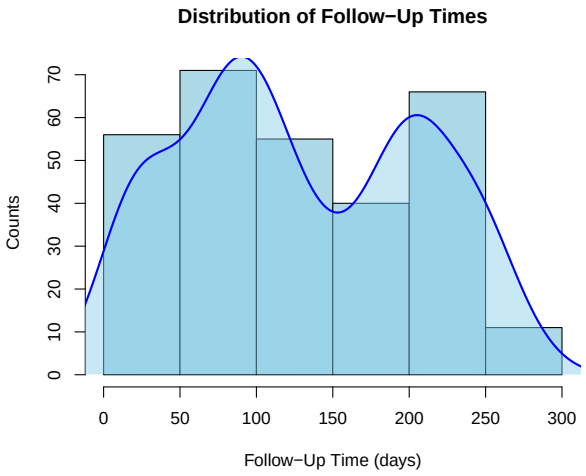
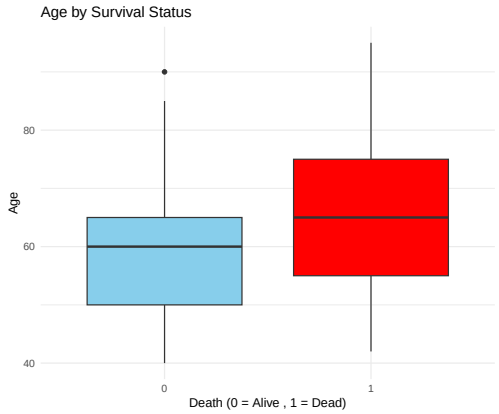
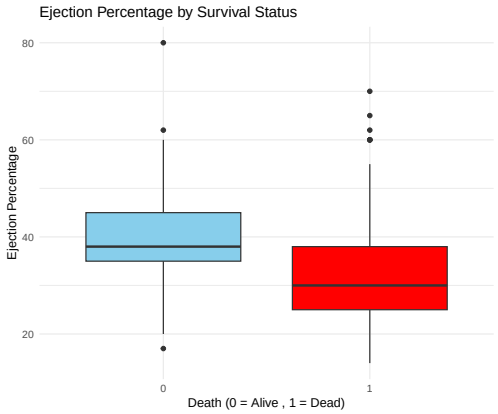


Figure 1: Histogram of Follow Up times (days)

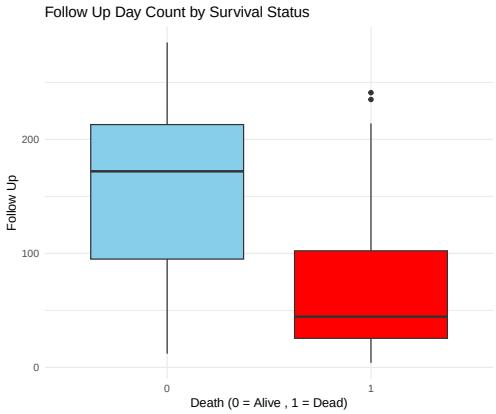
It would be of use to look at certain variables and their relationship with the survival of the patient. Figures 2(a), 2(b) , and 2(c) show box plots of these age, ejection_percentage, and follow_up



(a) Age by death



(b) Ejection Percentage by death

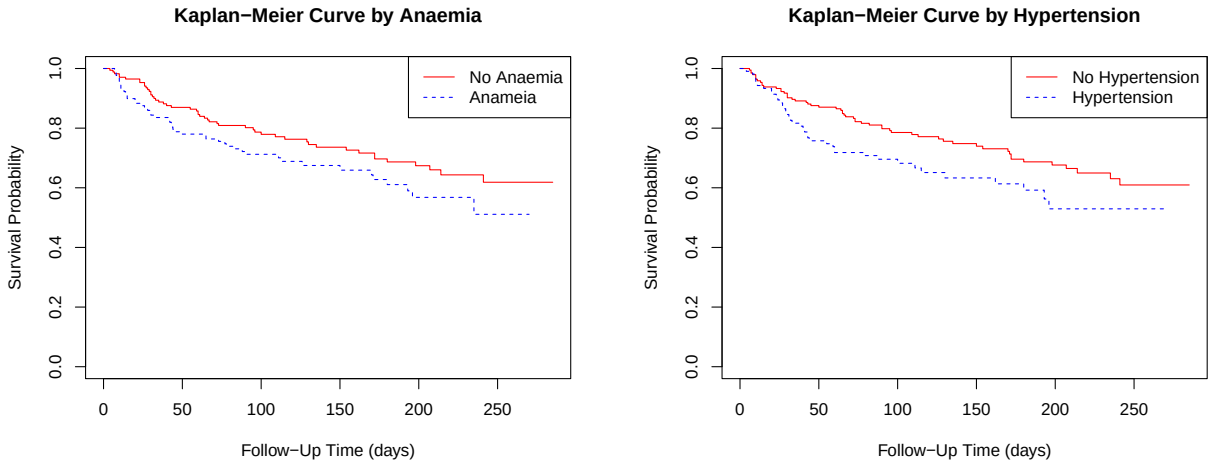


(c) Follow-up by death

Figure 2: Boxplots of selected continuous variables stratified by death status.

These variables had the most noticeable difference in distributions when it came to the death and survival of the patient. The mean age for the patients who died was 65.21, compared to 58.76 for those who survived. The mean ejection percentage for the patients who died was 33.47 and 40.26 who survived. The mean follow up time for patients who lived was 158.34 and 70.89 for patients who survived. These findings indicate that there may be a relationship between each of the variables and patient survival, especially the follow up time. It may be useful to control for follow up time to properly model the data later on

As a preliminary step in the survival analysis, Kaplan–Meier curves were plotted for the categorical patient characteristics to visualize differences in survival over time. Among these variables, anemia and hypertension appeared to show the strongest association with survival as shown in Figures 3(a) and 3(b).



(a) Kaplan–Meier curve: Anemia

(b) Kaplan–Meier curve: Hypertension

Figure 3: Kaplan–Meier curves for anemia (a) and hypertension (b).

The curves do not intersect after the initial time point, indicating that patients without hypertension or anemia tend to have higher survival probabilities when diagnosed with cardiomyopathy.

Next, the correlations between all continuous variables are examined. As shown in Figure 2, none of the variables exhibit very strong linear relationships with each other, suggesting that collinearity is unlikely to be an issue for this dataset.

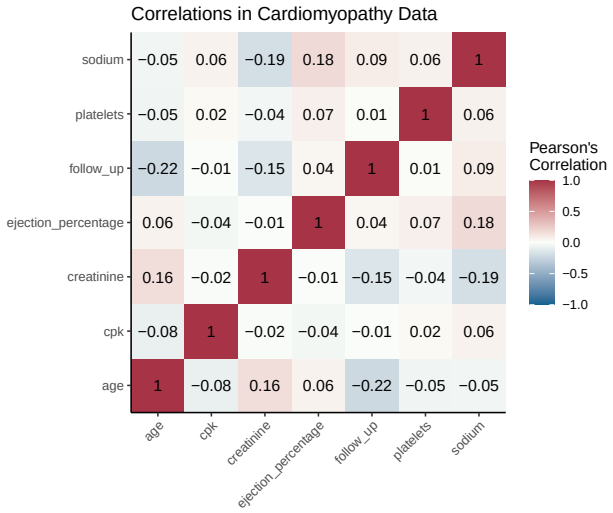


Figure 4: Pearson Correlation Plot

Methods

To answer the questions posed in this analysis several tools were used to model the data and determine significance in the relationships of the predictor variables and the response of whether a patient survived or not.

Simple Logistic Regression

For each of the eleven clinical and lifestyle variables, a separate logistic regression model was fit to assess its univariate association with the response variable `death`. The model equation is as follows:

$$\text{logit}(\hat{p}) = \beta_0 + \beta_1 X,$$

where $\hat{p}$ is the predicted probability of death and X is the predictor of interest. Likelihood Ratio Tests were used to assess significance. Likelihood-ratio tests were then used to assess if the relationship was significant or not. The hypothesis test used followed the following form:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

and was performed at significance level $\alpha = 0.05$

Cox Proportional Hazards Model

To evaluate how the predictors jointly influence the instantaneous risk of death (also known as the hazard), two Cox proportional hazards models were fit. two Cox proportional hazards models were fit. The Cox model is used to quantify the effect of predictors on the hazard (instantaneous risk) of an event occurring over time.

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

where $h(t \mid X)$ represents the hazard function at time t given the covariates X, which can be interpreted as the instantaneous risk of the event occurring at time t, conditional on survival up to that time. The baseline hazard $h_0(t)$ corresponds to the hazard when all predictors are zero, and $\exp(\beta_j)$ represents the hazard ratio associated with a one-unit increase in X_j.where $X_1, ...X_k$ are the predictors included in the model. The first model included all 11 variables and the second model included only the variables found significant in the univariate analysis. Models were assessed using AIC, C-indices, and nested likelihood ratio tests.

Model assumptions were checked using diagnostic plots and tests (Schoenfeld residuals for proportional hazards, assessment of linearity for continuous covariates). All assumptions were reasonably met (see Appendix for code and plots).

Results

Simple Logistic Regression

The resulting table gives a summary of the results of the Likelihood-ratio tests

Variable	Df	Test statistic	P-value	Significant?
age	1	19.356	< 0.001	**Yes**
cpk	1	1.118	0.290	**No**
ejection_percentage	1	23.381	< 0.001	**Yes**
platelets	1	0.738	0.3904	**No**
creatinine	1	28.097	< 0.001	**Yes**
sodium	1	11.327	< 0.001	**Yes**
anaemia	1	1.309	0.253	**No**
diabetes	1	0.001	0.973	**No**
hypertension	1	1.863	0.1723	**No**
gender	1	0.006	0.941	**No**
smoking	1	0.048	0.827	**No**

Table 4: LRT results

Each test was conducted at significance level $\alpha = 0.05$. As shown in Table 4 variables `age`, `ejection_percentage`, `creatinine`, and `sodium` we found to have a significant relationship with `death`.

Cox Proportional Hazards Model

To evaluate how the predictors jointly influence the hazard of death, two Cox proportional hazards models were fitted. The first model included all eleven clinical, lifestyle, and body variables, while the second model included only the variables that showed significant univariate associations with `death` in the simple logistic regression analysis (`age`, `ejection_percentage`, `creatinine`, and `sodium`). Table 5 presents the estimated coefficients, standard errors, z-values, p-values, and hazard ratios for both the full and reduced models with table 6.

Variable	Coefficient ($\hat{\beta}$)	Std. Error	z value	P-value	Hazard Ratio
Full Model (All predictors)					
age	0.0464	0.00932	4.977	< 0.001	1.048
anaemia	0.4601	0.2168	2.122	0.034	1.5843
cpk	0.0002207	0.0000992	2.225	0.026	1.0002
diabetes	0.1399	0.2231	0.627	0.531	1.1501
ejection_percentage	-0.0489	0.01048	-4.672	< 0.001	0.952
hypertension	0.4757	0.2162	2.201	0.028	1.6092
platelets	-0.0000004635	0.000001126	-0.412	0.681	1.0000
creatinine	0.3210	0.07017	4.575	< 0.001	1.379
sodium	-0.0442	0.02327	-1.899	0.0575	0.957
gender	-0.2375	0.2516	-0.944	0.3452	0.789
smoking	0.1289	0.2512	0.513	0.6078	1.138
Reduced Model (Significant predictors only)					
age	0.04437	0.00894	4.963	< 0.001	1.045
ejection_percentage	-0.04545	0.01043	-4.356	< 0.001	0.956
creatinine	0.32902	0.07250	4.538	< 0.001	1.390
sodium	-0.03369	0.02323	-1.450	0.147	0.967

Table 5: Estimated coefficients, standard errors, z-values, p-values, and hazard ratios for the full and reduced Cox proportional hazards models predicting patient hazard.

Model	AIC	C-index	χ^2	df	p-value
Full Model	957.907	0.741			
Reduced Model	958.456	0.719	13.451	7	0.062

Table 6: Models AIC

As shown in Table 5 of the four variables that were found to be significant in the univariate analysis only `age`, `ejection_percentage`, and creatinine were found to be significant predictors in the full and reduced models, while sodium was not significant in either. Certain variables that did not have significant univariate relationships with death were found to be significant in the full model such as `anaemia`, `cpk`, and `hypertension`

When assessing the models we prefer the reduced model that contains only predictors `age`, `ejection_percentage` `creatinine`, and `sodium` over the full model. This model is preferred because it is substantially simpler than the full model while maintaining similar performance. The AICs of the full and reduced models were 957.91 and 958.46, respectively, and the C-indices were 0.741 and 0.719. These differences are small and do not justify choosing the more complex model. In a formal sense a nested likelihood ratio test was performed and it was found that the full model with all predictors did not fit the data significantly better than the reduced model (df = 7, $\chi^2 = 13.451$, p-value = 0.062).

Conclusions

After adjusting for all predictors, age, ejection percentage, and creatinine remained significant independent predictors of survival, emphasizing their impact on patient outcomes.

Variables such as `anaemia`, `cpk`, and `hypertension`, which were not significant in univariate analyses, became significant in the full model, suggesting that their effects are influenced by other variables. These results highlight the importance of evaluating multiple predictors simultaneously to accurately assess risk in cardiomyopathy patients

In the final cox proportional hazard model predicting hazard among cardiomyopathy patients, age, ejection percentage, creatinine, and follow-up time were significant predictors of survival. In the final model, each additional year of age increased the instantaneous risk of death by 4.5% holding all other variables constant , indicating older patients were at higher risk of death. Each one-percent increase in ejection-percentage decreased the instantaneous risk of death by 4.4% holding all other variables constant. Each one unit increase in creatinine increased the instantaneous risk of death by 39%. Sodium was kept in the model despite not being statistically significant, to account for potential clinical relevance and control for confounding. Overall these results indicate that age, cardiac function, and kidney function are primary factors associated with the instantaneous risk of death in this cohort.