Survival Analysis Assignment on Heart Failure Dataset

By

Hazem Abu-Eseifan, Mohammed Bouayoun,

Timmothy Dangeon and Victoria Vigot

Submitted August 31st, 2022

Link to GitHub repository with the code

Contents

1. Introduction

1.1 Abstract

Heart failure – sometimes known as congestive heart failure – occurs when the heart muscle doesn't pump blood as well as it should. When this happens, blood often backs up, flows back and fluid can build up in the lungs, causing shortness of breath. Certain heart conditions, such as narrowed arteries in the heart (coronary artery disease) or high blood pressure, gradually leave the heart too weak or too stiff to fill and pump blood properly, and can lead to heart failure.

Heart failure is a serious condition with high prevalence (about 2% in the adult population in developed countries, and more than 8% in patients older than 75 years). The cost for society is high, reaching up to 2% of total healthcare expenditure in developed countries. Hence the importance of being able to analyse and predict what factors are more likely to shorten the duration of life after heart failure is diagnosed.

1.2 The dataset

This dataset contains the medical records of 299 patients who suffered heart failure. The data has been collected during their follow-up period, and each patient observation has 13 clinical features:[1]

	Feature	Explanation	Measurement	Range
1	Age	Age of the patient	Years	[40,..., 95]
2	Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
3	High blood pressure	If a patient has hypertension	Boolean	0, 1
4	Creatinine phosphokinase (CPK)[2]	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
5	Diabetes	If the patient has diabetes	Boolean	0, 1
6	Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,..., 80]
7	Sex	Woman or man	Binary	0, 1
8	Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
9	Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
10	Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
11	Smoking	If the patient smokes	Boolean	0, 1
12	Time	Follow-up period	Days	[4,...,285]
13	(target) death event	If the patient died during the follow-up period	Boolean	0, 1

1.3 Methodology

- General description of data
- Using Kaplan-Meier for the entire dataset and categorical variables
- Build Cox Proportional Hazard model.
- Variable selection on Cox model to determine the most relevant variables.
- Model diagnostics and tests.

---

1 Source: BMC Medical Informatics and Decision Making.

2 Will just call it CPK in this document for the sake of brevity.

2. Data description

Death event

0 (did not die)	203
1 (died)	96

For the seven numeric variables:

Age

Min.	40.00
1st Qu.	51.00
Median	60.00
Mean	60.83
3rd Qu.	70.00
Max.	95.00

CPK

Min.	23.0
1st Qu.	116.5
Median	250.0
Mean	581.8
3rd Qu.	582.0
Max.	7861.0

Ejection franction

Min.	14.00
1st Qu.	30.00
Median	38.00
Mean	38.08
3rd Qu.	45.00
Max.	80.00

Platlets

Min.	25100
1st Qu.	212500
Median	262000
Mean	263358
3rd Qu.	303500
Max.	850000

Serum creatinine

Min.	0.500
1st Qu.	0.900
Median	1.100
Mean	1.394
3rd Qu.	1.400
Max.	9.400

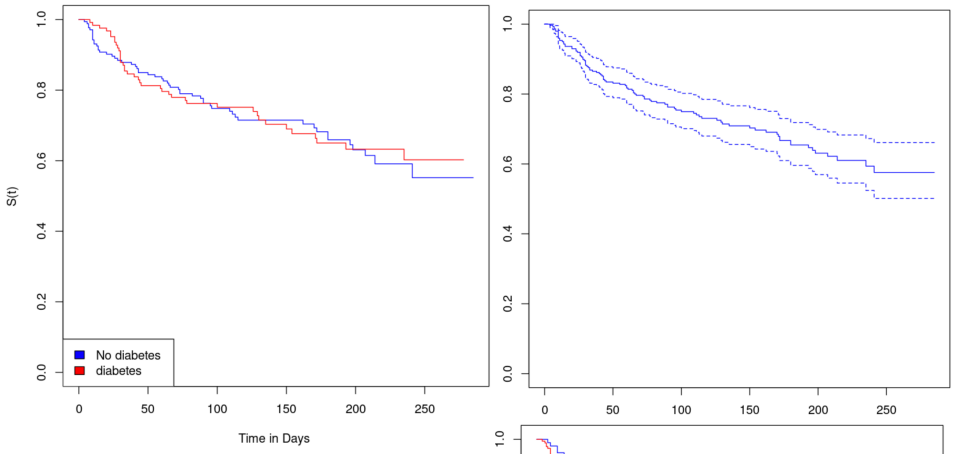
Serum sodium

Min.	113.0
1st Qu.	134.0
Median	137.0
Mean	136.6
3rd Qu.	140.0
Max.	148.0

Time

Min.	4.0
1st Qu.	73.0
Median	115.0
Mean	130.0
3rd Qu.	203.0
Max.	285.0

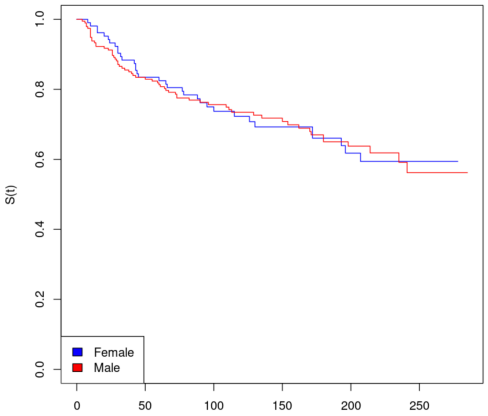
Checking for missing data shows the set is complete, no missing data

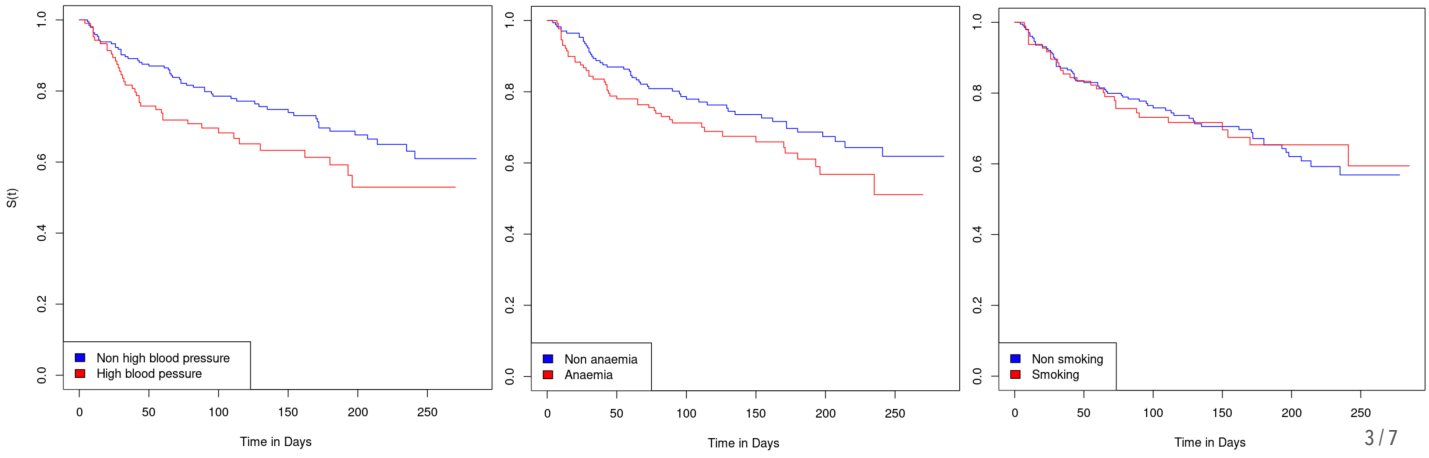


3. Models:

3.1 Kaplan-Meier:

We performed Kaplan-Meier analysis, first with the time and censoring indicator alone. We get the graph to the in up right corner, with a survival rate of 75% at 100 days, and no median survival time since the death rate never falls below 50%. Then we look at the KM functions with categorical variables. From the graphs we can see that anaemia and blood pressure seem to have an impact on survival, whereas the other categorical variables (sex, diabetes and smoking) do not.





The figures seem to indicate that only blood pressure and anaemia have an impact on the outcome. Running a `survdiff` function for each KM function we get the following results:

Sex

	N	Observed	Expected
sex=0	105	34	34.3
sex=1	194	62	61.7
P-value		0.9	

Anaemia

	N	Observed	Expected
df_anaemia=0	170	50	57.9
df_anaemia=1	129	46	38.1
P-value		0.1	

Smoking

	N	Observed	Expected
df_smoking=0	203	66	65.8
df_smoking=1	96	30	30.2
P-value		1	

Diabetes

	N	Observed	Expected
df_diabetes=0	174	56	55
df_diabetes=1	125	40	41
P-value		0.8	

Blood pressure

	N	Observed	Expected
df_blood=0	194	57	66.4
df_blood=1	105	39	29.6
P-value		0.04	

We can see from these figures that indeed anaemia and blood pressure have an impact, however, the difference in the case of anaemia is not statistically significant, with a p-value of 0.1.

3.2 Cox proportional hazard model:

We will now proceed to build a Cox proportional hazard model and do variable selection and model diagnostics. But first we will split the data into a training set and a testing set 80% training and 20% testing. Now we train a cox model using the `coxph` function on all variables (call it cox11). This is the result:

	Term	coef	exp(coef)	p-value	Significance
1	Age	4.145e-02	1.0423	8.53e-05	***
2	CPK	2.018e-04	1.0002	0.0655	
3	Anaemia	3.416e-01	1.4072	0.1577	
4	Diabetes	3.198e-01	1.3768	0.1981	
5	Ejection fraction	-5.013e-02	0.9511	2.21e-05	***
6	High blood pressure	4.596e-01	1.5835	0.0499	*
7	Platelets	-1.035e-07	1.0000	0.9408	
8	Serum creatinine	3.325e-01	1.3945	8.47e-06	***
9	Serum sodium	-6.065e-02	0.9412	0.0374	*
10	Sex	-1.027e-01	0.9024	0.7208	
11	Smoking	2.474e-01	1.2807	0.3747	

It looks clear from the table above that age, ejection fraction and serum creatinine are the most significant variables. We build a model with these three variables and compare it to the full model. So we build a model with these three variables and compare it to the full model using the `anova` function to compare nested models and we get a p-value of 0.931 — high enough, which means we cannot rule out the null hypothesis, namely that all variables outside these three are insignificant. When we removed one of the three variables and compared to the full model we got extremely low p-values, indicating that variables outside the selection are significant.

But we are not stopping here. We will try to perform variable selection using the step AIC method on the full model with 11 variables. The fine result of which was the seven variables to the right.

	AIC
<none>	746.25
Anaemia	746.29
CPK	746.66
High blood pressure	748.11
Serum sodium	748.80
Serum creatinine	757.18
Age	757.94
Ejection fraction	764.89

Now we ran an anova test again to compare the three variable model (call it cox3) to the seven-variable one (cox7) and the p-value came at 0.0235, which means that some additional variables in cox7 are significant. But which ones?

Having narrowed the choice down to between three and seven variables, it is time to compare the different models using the c-statistic technique to see which one preforms better at prediction. To do this trained six models on the training set with 3, 4, 5, 6, and 7 variables, and just for good measure, we threw in the full 11-variable model too. Next we attach the linear predictor (lp) from each model to the testing data set, then we train new cox models on the testing set with the corresponding lp as variable and compare the results. This is what we get:

	Model name	variables	Coefficient	C-statistic
A	cox3	age+ejection fraction+serum creatinine	1.0062	0.681
B	cox4	cox3+anaemia	0.9260	0.685
C	cox5	cox4+high blood pressure	0.9902	0.704
D	**cox6**	**cox5+CPK**	**1.0797**	**0.728**
E	cox7	cox6+serum sodium	0.8931	0.716
F	Cox11	All variables	1.5835	0.712

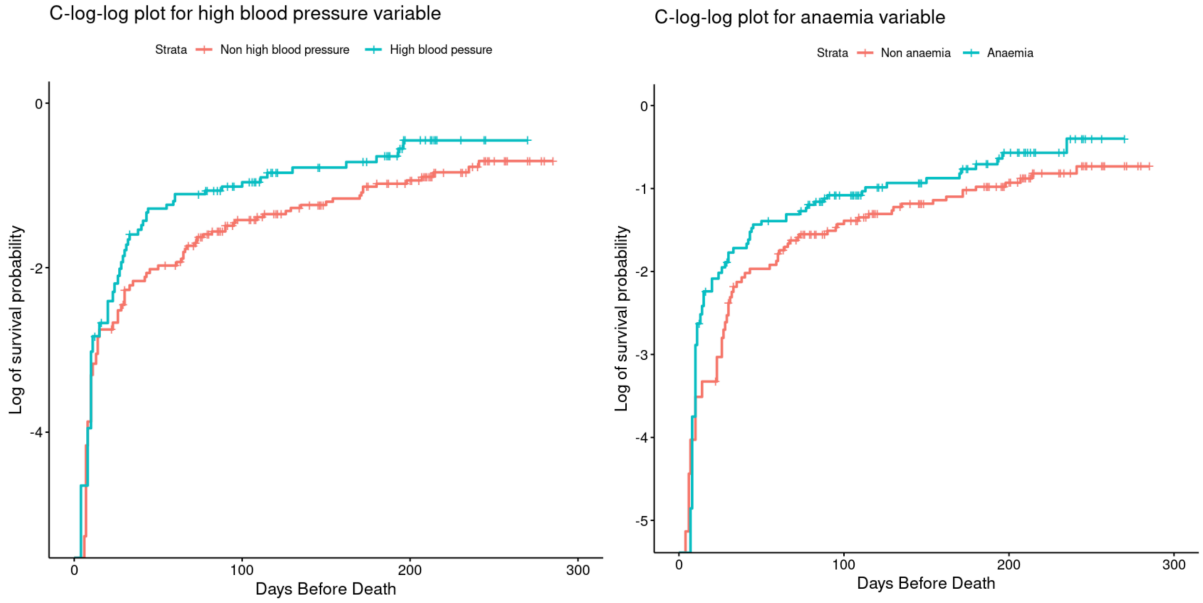
All models have positive coefficients, and we can see that cox6 produces the best results when it comes to prediction; adding the seventh variable recommended by the step AIC method *degrades* the model's predictions. So this is the model we select, with the following variables:

	Term	coef	exp(coef)	p-value
1	Age	4.361e-02	1.0446	8.41e-07
2	Ejection fraction	-5.179e-02	0.9495	2.57e-07
3	Serum creatinine	3.483e-01	1.4167	1.05e-07
4	Anaemia	3.933e-01	1.4818	0.0648
5	High blood pressure	4.668e-01	1.5948	0.0284
6	CPK	1.965e-04	1.0002	0.0462

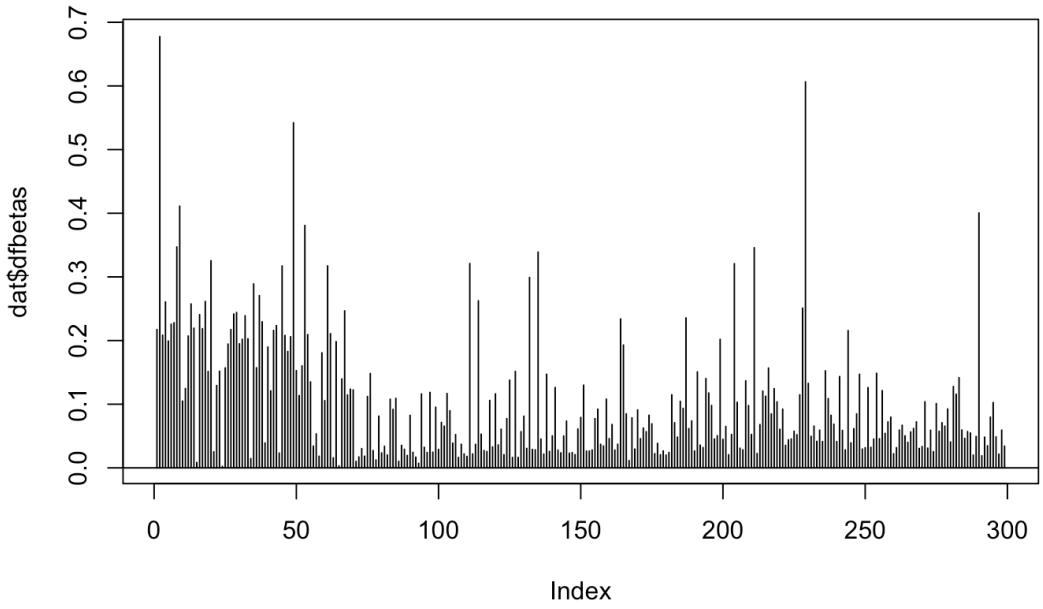
A couple of notes about this result. Although the p-value for anaemia is slightly above the 5% threshold here. However it is not way above 5%, and in some cases, with a different train-test splits, it came below the threshold. So it is a variable that is on the borderline between significance and insignificance. We decided to include it in the model nonetheless because it does enhance the predictive power, albeit slightly. Second, while the coefficient for CPK looks very small, it is in fact meaningful because of the broad range of this variable (23 to 7,861). So for example, everything else being equal, a patient with a CPK of 3,000 has a hazard of death that is 63% higher than a patient with a CPK of 500.

3.3. Model diagnostics:

We will perform two tests to see if the model fits the assumptions and is robust. We can see that for the two categorical variables we have in the final model (anaemia and high blood pressure), the log-log curves for with and without are roughly parallel, which fits with the proportionality assumption of the model.



Next we perform row deletion to test the stability of the model using the `dfbetas` function. We remove one row from the data set, recalculate all the *ß*, then take the normal difference between the new *ß* and the ones calculated with all data points. The idea is to see if removing one data point will have an impact on the outcome of the model. The answer from the graph below is no: our model is pretty stable.

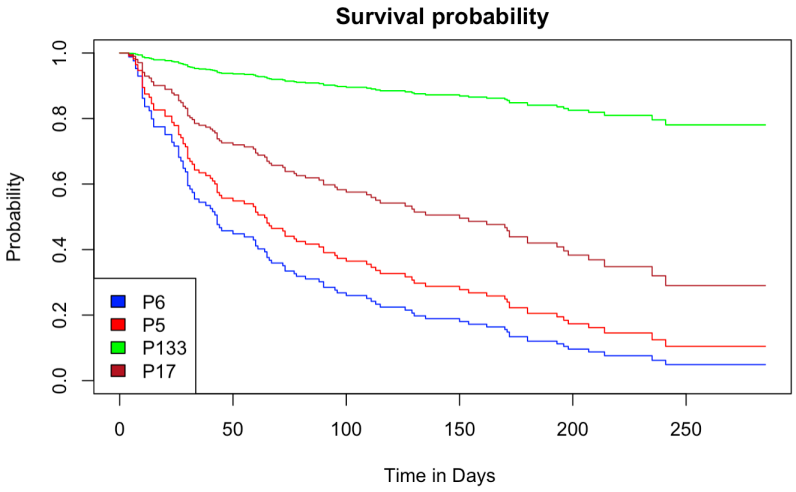


4. Conclusion:

We have a good model with six variables with a strong predictive power, a concordance of 73%. Now we can use our model to make some predictions. Pick four individuals from the dataset at random and compare their survival probability:

No	age	anae	CPK	diabetes	Ejection fracti	high blood pı	platelets	serum creatinine	serum sodium	sex	smoking
P6	90	1	47	0	40	1	204000	2.1	132	1	1
P5	65	1	160	1	20	0	327000	2.7	116	0	0
P133	46	0	719	0	40	1	263358.03	1.18	137	0	0
P17	87	1	149	0	38	0	262000	0.9	140	1	0

And this is the result we get:



	Median survival duration (days)
P6	43
P5	64
P133	NA
P17	150