ISSN 2959-6157

Cox Regression Model with Time-varying Coefficients Applied to Survival Estimation of Heart Failure

Jiahong Zhang[1, *]

[1]School of the Gifted Young, University of science and technology of China, Anhui, 238000, China
*Corresponding author: jiahong_zhang@mail.ustc.edu.cn

Abstract:

This study focuses on survival analysis of heart failure patients. All the patients, who are over 40 years old, have NYHA classes III or IV stages of heart failure. Cox regression model with time-varying coefficients is applied to deal with these data, which is updated in 2015. Kaplan Meier curves are plotted to screen out categorical variables that violates PH assumptions. As for continuous variables violating PH assumption, Schoenfeld residual plot is drawn to realize time stratification. Therefore, for each time period, cox proportional hazard model can be fit to estimate mortality and find out the key factors promoting or inhibiting death event of heart failure. As a result, age, blood pressure, serum creatinine, and ejection fraction are the significant features of hazard of death led by heart failure, which is validated by Wald test, Likelihood ratio test, and Log-rank test. Discrimination ability is also characterized by ROC curves and AUC finally.

Keywords: Heart failure; cox regression; survival analysis; time-varying model.

1. Introduction

As a serious public health concern, heart failure (HF) is affecting millions of individuals worldwide, and is associated with high morbidity, mortality, and healthcare costs. A 2017 survey indicates that roughly 64.34 million individuals globally are affected by heart failure [1]. The occurrence of heart failure among adults in the United States and developed countries in Europe is about 1%~3% [2, 3]. Another survey on basic medical insurance for urban employees in 6 provinces in China shows that the age-standardized prevalence of heart failure among urban residents aged ≥ 25 years is 1.1%, which implies that there are about 12.1 million heart failure patients in China [4]. Therefore, it’s crucial to find out the factors leading to heart failure in order to take precautions. This paper aims to study various potential factors that possibly contribute to death of heart failure and help health care workers prognosis.

The factors that cause incidence and death of heart failure are very complicated. The progression of population ageing and the impact of unhealthy lifestyles have caused an increase in the incidence of various cardiovascular diseases [5]. Scholars have found strong correlation between heart failure and age increasing [6], diabetes mellitus [7], and obesity [8]. Another regional study on heart failure with hypertension have indicated that age, high body index (BMI), lumbar abdominal obesity, dyslipidemia, and genetic history of the disease are major risk factors [9]. However, domestic and foreign studies on heart failure mainly focus on the prophylaxis and treatment of heart failure instead of the prediction of survival of heart failure patients in the next stage and the exploration of the cause of death. Therefore, the purpose of this study is to find out the factors that lead to death of heart failure patients, as well as the positive and negative correlation factors, in order to provide a certain theoretical basis for subsequent treatment options.

In previous studies, Giuseppe and Davide predicted survival of patients by using various machine learning methods [10, 11]. However, consideration of only ejection fraction and serum creatinine may have trouble in comprehensively predicting survival in different follow-up period. Therefore, this paper will use a survival analysis method, cox regression model, to find out the factors that mainly contribute to mortality of heart failure. Different from traditional Cox proportional hazards regression model, this study fit Cox model with time-varying coefficients based on time stratification to find out more concrete expression of coefficient of the features violating the proportional hazard assumption.

2. Methods

2.1 Data Source

The data for this research, collected from Kaggle website, contains medical records of 299 individuals from the Allied Hospital in Faisalabad and the Faisalabad Institute of Cardiology in 2015 who had classes III or IV stages of heart failure, classified by New York Heart Association (NYHA).

2.2 Variable Selection

The dataset analyzed in this study consists of 299 heart failure patients, whose ages range from 40 to 95 years old, among which 105 are female and 194 are male. It contains 13 features (age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum sodium, serum creatinine, sex, smoking, time, and death event), whose descriptions are shown as follows.

Table 1. Description of the 13 Features in the dataset

Variable	Explanation	Measurement	Range
age	Age of the patient	years	40~95
anaemia	Whether or not the patient has anaemia	Boolean	0, 1
creatinine_ phosphokinase	CPK enzyme level in blood	mcg/L	23~7861
diabetes	Whether or not the patient has diabetes	Boolean	0, 1
ejection_fraction	Percentage of blood leaving heart at each heart beat	Percentage	14~80
high_blood_pressure	Whether or not the patient has hypertension	Boolean	0, 1
platelets	Platelets concentration in blood	kiloplatelets/mL	25100~850000
serum_sodium	Sodium level in blood	mEq/L	113~148
serum_creatinine	Creatinine level in blood	mg/dL	0.5~9.4
sex	0: Female 1: Male	Binary	0, 1
smoking	Whether or not the patient smokes	Boolean	0, 1
time	Duration of follow-up period	days	4~285
DEATH_EVENT	Survival status of the patient	Boolean	0, 1

*mcg/L: micrograms per liter.

*mEq/L: milliequivalents per litre

Table 1 shows that each variable has no missing values in the dataset, therefore there's no need to preprocess the data. In the 299 samples, 96 died during the follow-up period.

2.3 Research Protocol

This paper uses Cox regression to research the key factors contributing to the heart failure mortality. 'DEATH_EVENT' is regarded as dependent variable(Y), while the other 12 ones are independent variables(X). Then this article uses Python and R to test the correlation and analyze the relationship between X and Y.

2.4 Model Principle

Cox regression is developed to link the probability of death with several explanatory variables and test the significance of them. Let the hazard of death at moment t be $h(t)$, the values of the variables X_i be x_i. Then Cox model assumes that the probability of death for an individual can be expressed as the following function:

$$h(t) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \qquad (1)$$

In the formula above, β_i is the partial regression coefficient of variable Xi, h_0(t) is the baseline hazard rate. As a semi-parameter model, Cox regression cannot provide functional expression of h0(t), so value of β (and its relationship with time) is what this model focuses on.

As a survival analysis model, Cox model is based on two assumptions: proportional hazard assumption (PH assumption) and log-linear assumption. PH assumption: $h(t)/h_0(t)$ remains constant when time changes, which means that any independent variable has no correlation with time. Log-linear assumption: the logarithm of risk for given individual is linearly related to any continuous variable. Therefore, before applying Cox model, it's necessary to test whether the variables follow these assumptions. If not, the variable should be screened out or particularly processed as time-varying variable.

One way to model continuous time-varying coefficients is to use step functions. The idea of this approach is to divide the analysis time into intervals and use layer Cox regression models for these time intervals. The effect of variables can become stronger or weaker over time, which can be explored through time stratification.

3. Results and Discussion

3.1 Pretreatment of Data

To validate PH assumption, Kaplan-Meier survival curves of the categorical variables are drawn as follows. If there's a clear intersection between the two curves in a figure, the variable violates the assumption, which means that it should not be included in the regression model.

As is shown in Figure 1, curves of 'diabetes', 'sex' and 'smoking' have obvious intersections, indicating their necessity to be screened out. The probable reason why diabetes and smoking should not be considered here is that they are the main contributors of initial heart failure instead of the mortality of heart failure which is what this paper aim on.

However, subjective factors exist in the judgment of the remained variables, so the p-value of each variable left (including the continuous one) is also calculated as follows:

Table 2. Result of proportional hazard test by Python

Variable	test_statistic	p
age	0.03	0.85
creatinine_phosphokinase	0.86	0.35
ejection_fraction	5.65	0.02
platelets	0.18	0.67
serum_sodium	1.83	0.18
serum_creatinine	3.80	0.05
anaemia	0.00	0.98
high_blood_pressure	0.06	0.80

From Table 2, it can be found that the continuous variable 'ejection fraction' failed the non-proportional test: p-value is 0.0174, lower than threshold 0.05, which means that it has to be particularly treated as time-varying variable.

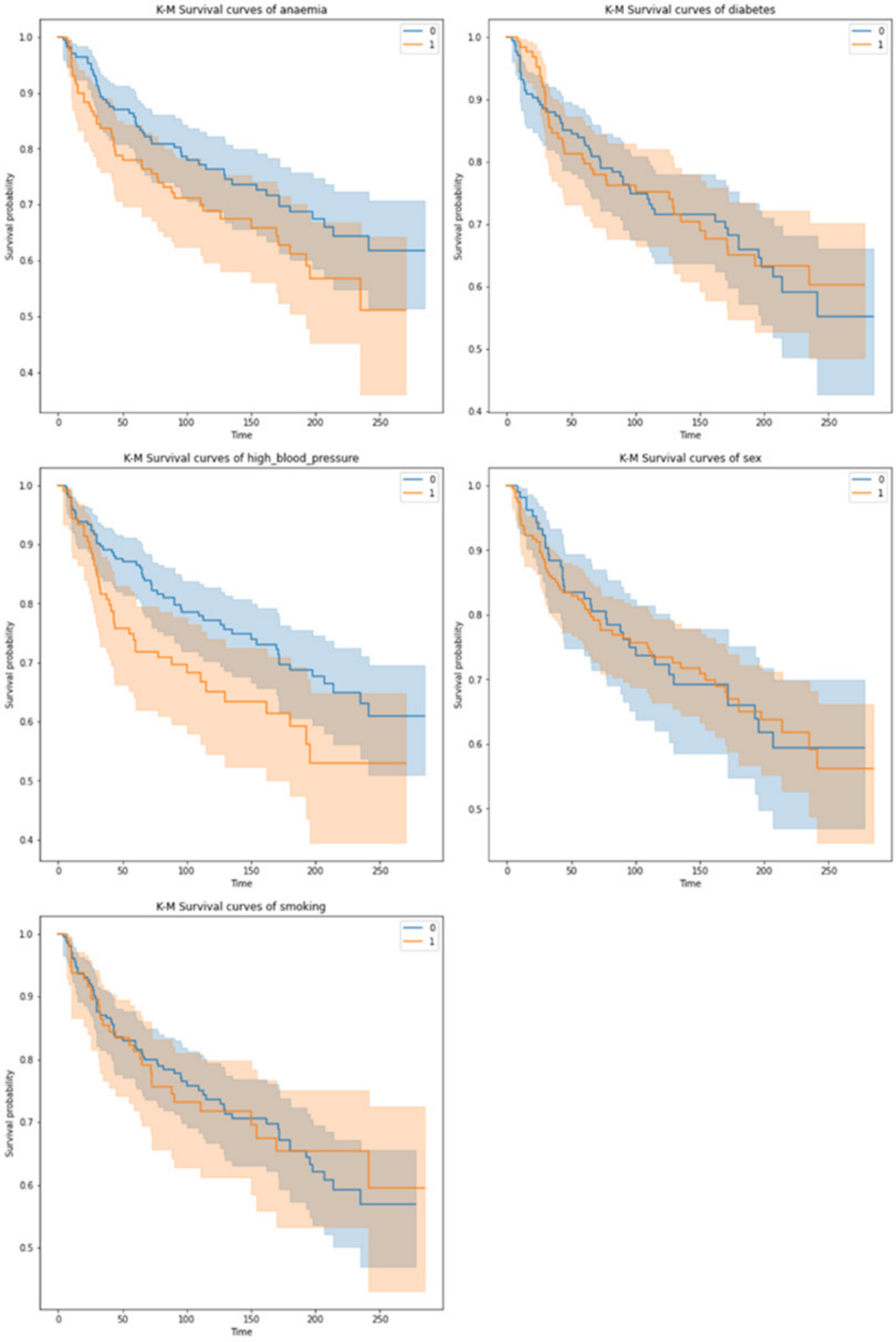

Fig. 1 K-M curves of categorical variables

3.2 Results of Univariate Cox Regression

Before applying multivariate Cox regression, univariate Cox regression was performed on all independent variables that satisfied the hypothesis, and each variable was checked to see if there was a statistically significant correlation with the survival of the sample (table 3).

At the significance level of 0.05, the variables whose p-values are lower than 0.05 in the table are selected, which are considered to be statistically significant in univariate Cox regression model: 'age', 'ejection_fraction', 'serum_creatinine', 'serum_sodium', and 'high_blood_pressure'.

Table 3. Result of univariate cox model by R

Variable	β	Risk ratio	p-value
age	0.042211	1.043115	8.36×10^{-7}
anaemia	0.3374	1.4013	0.0998
creatinine_phosphokinase	0.0001147	1.0001147	0.262
ejection_fraction	-0.04624	0.95481	1.73×10^{-5}
high_blood_pressure	0.4359	1.5464	0.0374
platelets	-7.801e-07	1	0.466
serum_creatinine	0.2897	1.3360	1.19×10^{-7}
serum_sodium	-0.06763	0.93460	0.00052

3.3 Results of Multivariate Cox Regression

Before using time-varying coefficients to fit Cox regression models, relationship between coefficient of 'ejection_fraction' and time is shown in the figure below, from which it can be seen that it does violate the PH assumption: the partial regression coefficient fluctuates over time (Figure 2).

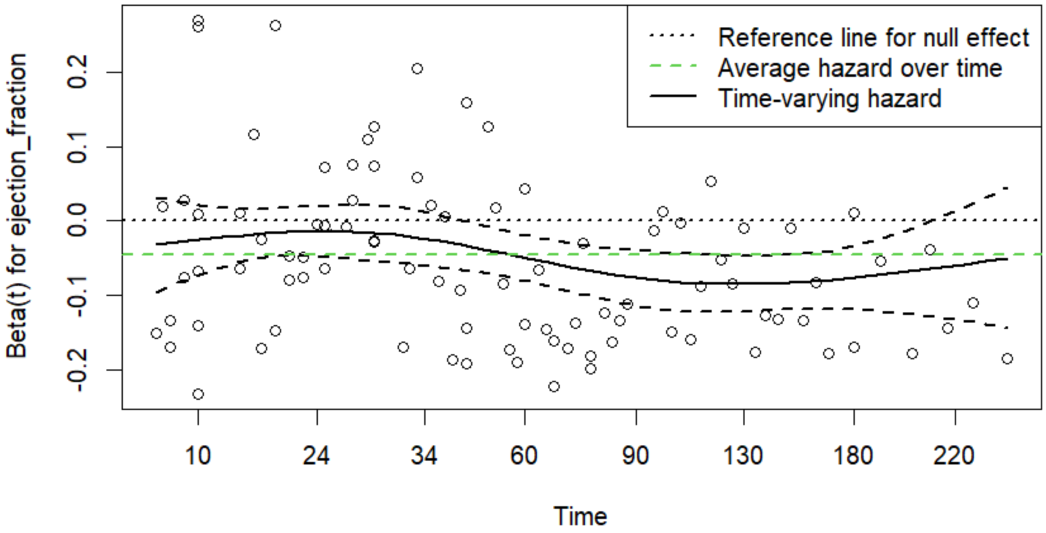

Fig. 2 Schoenfeld residual plot of 'ejection fraction'

Each curve in the figure has its own meaning. The first solid black line represents the curve of coefficient of ejection fraction over time. The second dashed black line is a reference line for the null effect, which is the case of no association between ejection fraction and hazard of death. The third dashed green line shows the average coefficient of ejection fraction based on the fitted Cox model.

Figure 2 shows that β line of variable 'ejection_fraction' has two inflection points when time is approximately equal to 30 and 130. Therefore, the multivariate cox model can be fitted through time stratification. After using 'survSplit' function in R to divide the dataset into three parts bases on time, the model is refitted as follows.

As is shown in Table 4, age is the variable with highest significance. Hazard of death caused by heart failure increases 4.6% as the age increases one year. Serum creatinine is another significant feature with p-value equal to 1.05×10-6, whose one-unit increase leads to 37% increase in hazard of death. Blood pressure is also a significant variable, which has p-value equal to 0.0151. Patients who have high blood pressure have 67% higher mortality than the ones who have no high blood pressure. As for ejection fraction, it's significant in the second and third time period, which correspond to time intervals (30,130) and (130, +∞). When the time is between 30 days and 130 days, one-unit increase in ejection fraction causes 5.9% decrease in probability of death, while it causes 8.5% decrease in mortality when the time exceeds 130 days. According to the results in the table, serum sodium and ejection fraction in the first time period are insignificant variables, since their p-values are all higher than threshold 0.05.

Table 4. Result of the final Cox regression model by R

Variable	e^{β}	p-value
age	1.0461	6.49×10^{-7}
ejection_fraction: tgroup=1	0.9824	0.22744
ejection_fraction: tgroup=2	0.9410	0.00010
ejection_fraction: tgroup=3	0.9153	0.00365
high_blood_pressure	1.6731	0.01521
serum_creatinine	1.3659	1.05×10^{-6}
serum_sodium	0.9642	0.12748

3.4 Model Validation

The overall significance of the final model is displayed in the table 5 below:

Table 5. Significance table of the final Cox regression model by R

Test Method	p-value(95%)
Likelihood ratio test	1×10^{-14}
Wald test	1×10^{-16}
Log-rank test	5×10^{-16}

The p-values of these tests are all much lower than 0.05, and the concordance of this cox model calculated by R is 0.734, indicating that the final cox regression model is statistically significant and has practical significance.

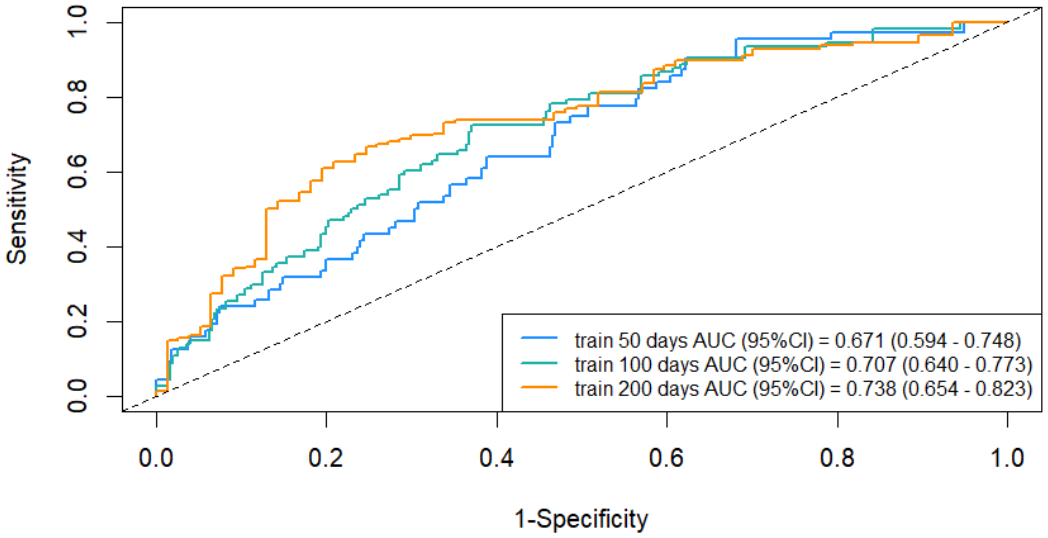

Fig. 3 ROC curves of different follow-up periods

ROC curve is drawn in figure 3 in order to evaluate the performance of the model. Area under curve (AUC) is 0.671, 0.707, and 0.738 respectively when follow-up period is 50 days, 100 days, and 200 days, which indicates the model prediction works relatively well. The main reason why AUC at time of 50 days is lower than expected may be the low significance for mortality of ejection fraction at time lower than 30 days.

4. Conclusion

It can be concluded in this paper that age increasing, hypertension, high level of serum creatinine, and low ejection fraction are the key factors mounting hazard of death for heart failure patients. At the same time, serum sodium may also have negative correlation with mortality of heart failure. As for health care workers, these features should be prioritized in subsequent treatment for heart failure patients. The possible reason for diabetes and smoking not being significant features, which may seems contrary to previous study on heart failure, is that they have impact on morbidity of heart failure instead of its mortality. It can also be concluded from the time-varying coefficient method that higher ejection fraction level can significantly reduce more mortality when time is beyond 130 days, which means that ejection fraction should be more focused on within 130 days after the onset of heart disease.

References

[1] GBD. Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 19902017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet, 2018, 392 (10159): 17891858.

[2] Savarese G, Becher P M, Lund L H, et al. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res, 2023, 118 (17): 3272-3287.

[3] Becher P M, Lund LH, Coats A J S, et al. An update on global epidemiology in heart failure. Eur Heart J, 2022, 43(32): 3005-3007.

[4] Wang H, Cai K, Du Ming, et al. Prevalence and incidence of heart failure among urban patients in China: a national population-based analysis. Circ Heart Fail, 2021, 14(10): e008406.

[5] Conrad N, Judge A, Tran J, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. Lancet, 2018, 391(10120): 572-580.

[6] Gu Dongfeng, Huang Guangyong, Wu Xigui, et al. Epidemiological investigation and prevalence of heart failure in China. Chinese Journal of Cardiovascular Diseases, 2003, 1: 6-9.

[7] Guo Dan. Risk factor prediction analysis of death in elderly patients with severe heart failure. Clinical Practice of Integrated Traditional Chinese and Western Medicine, 2017, 17(09): 130-131.

[8] Gajulapalli R, Kadri A, Gad M, et al. Impact of Obesity in Hospitalized Patientswith Heart Failure: A Nationwide Cohort Study. Southern Medical Journal, 2020, 113(11).

[9] Gao Qiang, Zhang Liangzhi. Current situation and risk factors of patients with hypertension complicated with heart failure in Shenyang. Public Health and Preventive Medicine, 2020, 31(05): 118-121.

[10] Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak, 2020, 20: 16.

[11] Ren Y, et al. Gray's time-varying coefficients model for posttransplant survival of pediatric liver transplant recipients with a diagnosis of cancer. Comput Math Methods Med, 2013, 719389.