Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia

Vol. 18, 2023, page 76 – 81

Survival Analysis and Factors of Heart Failure Disease

Nursaidatul Husna Norsuhaimi, Noraslinda Mohamed Ismail*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: noraslinda@utm.my

Abstract
A risky illness known as congestive heart failure occurs when the heart is unable to efficiently pump blood throughout the body. It is a significant threat to world health and a leading cause of death. Age, gender, tobacco use, high blood pressure, and diabetes are some risk variables affecting heart failure. Knowing what influences survival is essential for enhancing patient outcomes. Data from 299 clinical reports of patients with heart failure were analyzed in a study to find these characteristics and calculate the likelihood of mortality. Researchers examined the effect of several factors on survival using a Cox regression model. This model made it possible to analyze both short and long-term results. The model's determination of the hazard ratio showed the correlation between various parameters and survival times. The study offers useful details on clinical traits and outcomes that can be used to individualized treatment plans that will increase survival rates and enhance patients' quality of life. The model's determination ofthe hazard ratio showed the correlation between various parameters and survival times. The study offers useful details on clinical traits and outcomes that can be used to create individualized treatment plans that will increase survival rates and enhance patients' quality of life.

Keywords: Heart Failure (HF); Cox regression model

1. Introduction

In order to estimate the survival time of patients with heart failure, this study focuses on the application of survival analysis, specifically the non-parametric and semi-parametric methods (Ahmad, Tanvir et al., 2017). The goal of the study is to pinpoint the variables that affect these patients' survival rates and therapeutic outcomes. According to the New York Heart Association categorization, the study covers 299 individuals with heart failure who are in stages III or IV (Chicco and Jurman, 2020). The dataset contains a number of variables, including gender, age, serum creatinine, platelets, creatinine phosphokinase, ejection fraction, high blood pressure, diabetes, anemia, and unhealthy behaviours like smoking.

The time period of time following methods or the period of time patients survive the disease is measured by the researchers using the Kaplan-Meier survival function. This study also use the Cox regression model to examine mortality in relation to various risk factors. The Cox regression model enables simultaneous consideration of the survival time and the investigation of the impact of each risk factor on the defined time of occurrence (Chang, H. -L. and Yeh, T.-H. 2006) and (Ihwah, A 2015). The data are analyzed, and the parameters are estimated, using the statistical programme R Studio.

The study shows the significance of identifying heart failure, comprehending its origins, and finding efficient treatments. The biggest cause of death worldwide and a frequent cause of illness is heart failure. Heart failure can be caused by factors including high blood pressure, diabetes, and obesity, thus it's important to spread awareness of the condition and its symptoms to enable early medical intervention. The study's findings are intended to enhance public health management and our knowledge of heart failure patients' survival rates. It could be possible to lower the prevalence of the condition by determining the main causes of heart failure and raising awareness.

2. Materials and methods

Three models are used in survival analysis: non-parametric, semi-parametric, and parametric. Non-parametric methods, like Kaplan-Meier survival analysis, create the hazard function based on empirical data rather than assuming a certain form for it. Cox regression and other semi-parametric techniques explain the baseline shape and covariate effects rather than making assumptions about the influence of covariates on the hazard function. When time is regarded as an independent variable, these techniques are appropriate for predictive modeling. Making particular assumptions about the hazard function's shape is a necessary step in using parametric approaches, such as the maximum likelihood approach.

Hence the random survival time variable T can also be described by the following functions:

$$S(t) = \Pr(T > t) = 1 - \Pr(T \leq t) = 1 - F(t) = \int_{t}^{\infty} f(x)dx \tag{1}$$

The survival function $S(t)$ represents the probability of surviving beyond a given time t. Theoretically, thesurvival function forms a smooth curve starting at $S(t) = 1$ when $t = 1$ and gradually decreasing to $S(t) = 0$ as t approaches infinity. However, in practical terms, the survival curve is often depicted as a step function due to the limited duration of the study period. The step function ensures that the survival curve eventually reaches 0, indicating no surviving participants beyond a certain time.

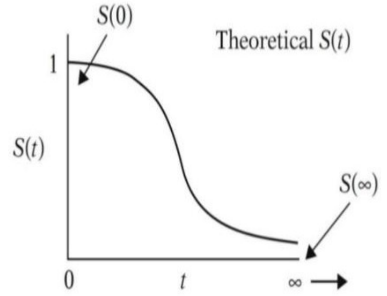

Figure 1: Theoretical Survival Function Graph

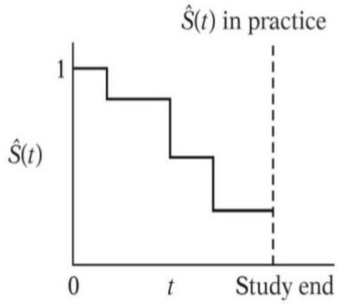

Figure 2: Step Function Graph

Figure 1 shows a theoretical survival function graph, where the probability of survival starts at 1 at the beginning of the study and gradually decreases to 0 as the time (t) increases towards infinity. The graphrepresents a smooth, downward-sloping curve, indicating a non-increasing survival function. On the other hand, Figure 2 illustrates a practical survival function graph, which is obtained when dealing with real data. In this case, the survival curve appears as a step function, reflecting the use of actual data points. The Kaplan-Meier approach is commonly used to calculate survival probabilities for plotting such graphs. Although the study period cannot extend beyond a certain time limit in real life, the interpretation of both graphs remains the same.

3. Cox Regression Model

Cox regression, also known as proportional hazards regression or duration model, was introduced by David Cox in 1972. It allows for examining the impact of different variables on the time it takes for a specific event to occur. The Cox regression model assumes the proportional hazards assumption and determines whether the time to the event increases or decreases based on the predictor variables. It has become the most commonly used regression perspective in survival analysis where predictor variables (Salamzadeh J et al 2003). The Cox regression model incorporates a simple linear regression equation, which is used to explain the dependent variable. If there are multiple explanatory variables, it is referred to as multivariate linear regression; otherwise, simple linear regression is used. The predictor or independent variable (x) is utilised in linear regression to explain the dependent variable (y). If there are multiple explanatory variables, the technique is referred to as multivariate linear regression; otherwise, simple linear regression is used. The model then takes the form of

$$y_i = \beta_0 + \beta_i x_i + \epsilon_i \,; \qquad i = 1,2,3, \dots , n \tag{2}$$

The symbol β represents the coefficients associated with the partial derivatives of the dependent variable with respect to the independent variables in linear regression. The estimation and inference of these coefficients are crucial in statistical analysis. To estimate the parameters β_0 and β_i, the least squares approach is commonly employed. In this study, the formula of hazard function in the form of Cox regression is written as:

$$\begin{aligned} h(t, X) &= h_0(t) exp \sum_{i=1}^{n} \beta_i \, X_i \\ &= h_0(t).(1) = \; h_0(t) \end{aligned} \tag{3}$$

Cox regression is a statistical approach that uses predictor variables, known as covariates, to predict an event variable. It does not require the researcher to specify a baseline hazard rate or calculate absolute risk. The model allows for accurate calculations of hazard ratios, modified survival curves, and closely approximates the results of chosen parametric models. The Cox regression model is particularly suitablewhen there is a strong rationale and when the data assumptions of parametric models are more stringent. It is a semi-parametric model with an undefined baseline hazard function, $h_0(t)$. The covariates can be continuous or categorical, and their modelling in Cox regression depends on whether the time variable isfixed or time dependent. For example, time-fixed covariates may be binary (0 or 1) based on the adjacency of states, while time-dependent covariates could represent changing monthly expenses.

The Hazard Function

The hazard function at time t is defined as the instantaneous rate of failure at time t, can denoted as $h(t)$and the equation is:

$$h(t) = \frac{f(t)}{s(t)} \tag{4}$$

where $h(t)$ is the hazard is an instantaneous potential per unit time for an event to occur, given the individual has survived up to time, t. The instantaneous potential the idea is illustrated by velocity. The hazard function is non-negative and has an upper bound.

R-Software

R is a versatile programming language and environment that is specifically designed for statistical computing and graphics. It provides a comprehensive set of tools and functions for various statistical analyses, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and more. R software's wide range of capabilities makes it suitable for a diverse range of applications in data analysis and research.

The extensibility of R-Software is one of its main advantages. By developing and disseminating their own packages, which may include extra statistical procedures, algorithms, or visualisation strategies, users can quickly increase the usefulness of the system. As a result, researchers and analysts can customise R to their own requirements and take advantage of the community's collective wisdom and contributions.

R is known for having strong graphical functions in addition to its analytical capabilities. It offers a comprehensive selection of plotting tools and libraries, enabling users to produce high-quality charts and graphs that are suitable for publication. R is appropriate for producing figures using mathematical notation because it offers various choices for customising the look and feel of visualisations and permits the incorporation of mathematical symbols and formulas. Therefore, R is an invaluable tool for data analysis, visualisation, and result reporting because it combines statistical processing power with adaptable graphical capabilities. This is because, researchers, statisticians, and data scientists frequently choose it because of how easy it is to use and how well it produces visually pleasing results complicated statistical studies.

The objective of this project is to use R software to analyse a dataset on heart failure in order to better understand the variables influencing heart failure cases and survival rates. Enhancing patient outcomes and creating improved methods of preventing and treating coronary heart disease are the objectives. The procedure of estimating the factors that affect the survival model using the dataset and R programme is described in the paragraph. It illustrates the significance of confounders based on their p-values and the significance of understanding the causes of heart failure and how they affect survival rates. In the end, this information will help improve patient care and heart failure management. To begin the test on the parameters' significance, set up the Model 1 as follows
Full model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} \quad (5)$$

where β_i are unknown parameters as known as regression coefficients. This is due to the parameter β_i that shows the expected changes in response y per unit change in x as all the remaining regressors are set constant. Then we can denote the variables as x_1 is the age of patients, x_2 is serum creatinine, x_3 is the ejection fraction, x_4 is anemia, x_5 is diabetes, x_6 is high blood pressure, x_7 is creatinine phosphokinase, x_8 is platelets, x_9 is serum sodium, x_{10} is sex and lastly x_{11} is smoking.

The paragraph discusses the concept of the null hypothesis in regression models and its significance. When the null hypothesis is accepted, it means that the regressor variable x_i has no significant impact on the dependent variable, allowing its removal from the model without affecting its performance. The decision to accept or reject the null hypothesis is based on p-values associated with regression coefficients. A lower p-value suggests a stronger argument against the null hypothesis and a higher likelihood of a significant correlation between the covariate and survival outcome. In Cox regression, the *coxph* function in R is used to fit the model, and the *summary()* function summarizes the results.
Thus, the best model fit by using R-Software Packages is:

$$Y = \mathbf{0.0484}\,(age) + \mathbf{0.3987}\,(serum\ creatinine) - \mathbf{0.0582}\,(ejection\ fraction) \quad (5)$$

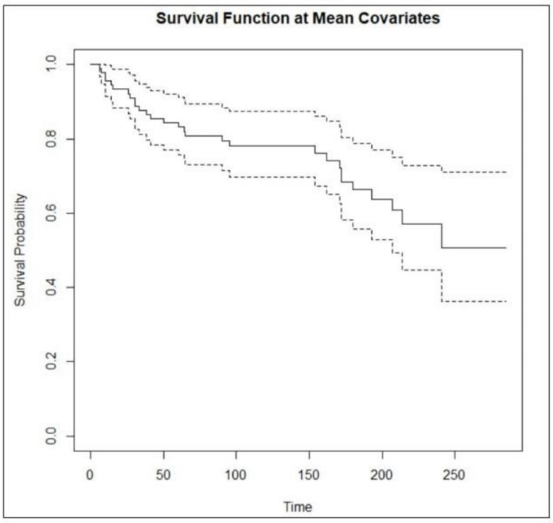

Figure 3 Survival Graph for Lung Cancer's Patients

The graph demonstrates that as the study duration lengthens, patients with heart failure experience shorter survival periods, which suggests a larger risk of encountering heart failure-related events. It is possible to observe how a patient's chance of survival changes with time by looking at the center line on the graph, which shows the survival probability for patients with average covariate values. At 200 days above, the graph shows that about 40% of the patients with average covariate values are estimated to still be alive. This graph provides illuminates trends in survival, the impact of variables, and the clinical relevance of survival outcomes for heart failure patients.

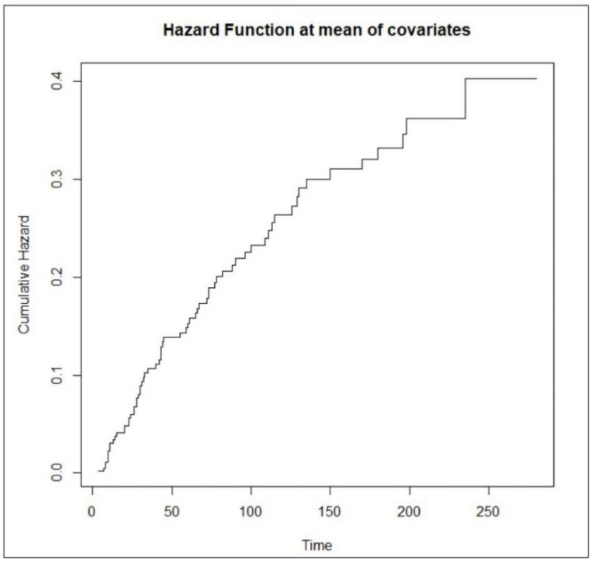

Figure 5 Hazard Graph for Heart Failure Patients

The graph 4 shows the risk of death for heart failure patients over time. It reveals that the risk of dying increases as time passes. By considering the average covariate values, the graph provides insights into how the immediate risk of death varies over time for heart failure patients. It helps us understand the expected survival outcomes for individuals with typical covariate profiles.

Conclusion

In conclusion, the covariates of age, serum creatinine, and ejection fraction are significant factors in heart failure (Ronco et al., 2008). Older age is associated with a higher prevalence of heart failure due to age-related changes in cardiovascular function while elderly patients may have better survival rates, possibly due to adaptation mechanisms and better adherence to treatment. Moreover, elevated serum creatinine levels, indicating impaired renal function, worsen prognosis and increase mortality in heart failure. Hence, to increase patient survival and treatment effectiveness, serum creatinine levels must be closely monitored and managed. Furthermore, heart failure (HF) outcomes are greatly influenced by ejection fraction, a measurement of the heart's pumping effectiveness. Compared to preserved ejection fraction (HFpEF), reduced ejection fraction (HFrEF) has a worse prognosis (Dunlay et al., 2017). Patients with lower ejection fraction have higher survival rates while receiving medical treatment that follows recommended protocols, such as beta-blockers and ACE inhibitors. Due to the limited number of available treatments, controlling HFpEF and mid-range ejection fraction (HFmrEF) presents difficulties (Dunlay et al., 2017).

Reference

[1] Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failurepatients: A case study. *PloS one*, *12*(7), e0181001.

[2] Ahmad, T., Pencina, M. J., Schulte, P. J., et al. (2014). Clinical implications of chronic heart failure phenotypes defined by cluster analysis. Journal of the American College of Cardiology, 64(17),

1765-1774.

[3] Chang, H. L., & Yeh, T. H. (2006). Regional motorcycle age and emissions inspection performance: a Cox regression analysis. *Transportation Research Part D: Transport and Environment*, *11*(5),324-332

[4] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure fromserum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 16. https://doi.org/10.1186/s12911-020-1023-5

[5] Damman, K., Voors, A. A., Navis, G., van Veldhuisen, D. J., & Hillege, H. L. (2012). Current and novelrenal biomarkers in heart failure. Heart failure reviews, 17(2), 241-250.

[6] Dunlay, S. M., Roger, V. L., Redfield, M. M. (2017). Epidemiology of heart failure with preserved ejection fraction. Nature Reviews Cardiology, 14(10), 591-602.

[7] Fang, Y. (n.d.). Machine Learning Survival for Heart Failure [Data set]. Retrieved from https://www.kaggle.com/code/fangya/machine-learning-survival-for-heart-failure/data

[8] Ihwah, A. (2015). The use of Cox regression model to analyze the factors that influence consumer purchase decisions on a product. *Agriculture and Agricultural Science Procedia*, *3*, 78-83

[9] Jones, N. R., Roalfe, A. K., Adoki, I., Hobbs, F. R., & Taylor, C. J. (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.

[10] Ling, H. S., Chung, B. K., Chua, P. F., Gan, K. X., Ho, W. L., Ong, E. Y. L., ... & Fong, A. Y. Y. (2020). Acute decompensated heart failure in a non cardiology tertiary referral centre, Sarawak GeneralHospital (SGH-HF). *BMC cardiovascular disorders*, *20*(1), 1-11.

[11] Malik, A., Brito, D., Vaqar, S., & Chhabra, L. (2017). Congestive heart failure.

[12] Mohammadzadeh, N., Safdari, R., Baraani, A., & Mohammadzadeh, F. (2014). Intelligent data analysis: the best approach for chronic heart failure (CHF) follow up management. *Acta Informatica Medica*, *22*(4), 263.

[13] Rich, M. W. (2017). Heart failure in the elderly. Clinical Geriatric Medicine, 33(4), 507-522. Ronco, C., Haapio, M., House, A. A., Anavekar, N., & Bellomo, R. (2008). Cardiorenal syndrome. Journal of the American College of Cardiology, 52(19), 1527-1539.

[14] Salamzadeh, J., Wong, I. C. K., Hosker, H. S. R., & Chrystyn, H. (2003). A Cox regression analysis of covariates for asthma hospital readmissions. Journal of Asthma, 40(6), 645-652. https://doi.org/10.1081/JAS-120019035

[15] Shlipak, M. G., Katz, R., Kestenbaum, B., Fried, L. F., Siscovick, D., & Sarnak, M. J. (2009). Clinical and subclinical cardiovascular disease and kidney function decline in the elderly. Atherosclerosis, 204(1), 298–303. https://doi.org/10.1016/j.atherosclerosis.2008.08.016

[16] Tafese Ashine Tefera, Geremew Muleta, Kenenisa Tadesse et al. Bayesian Survival Analysis of HeartFailure Patients: A Case Study in Jimma University Medical Center, Jimma, Ethiopia, 05 April 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-339250/v1]

[17] Taylor, C. J., Ryan, R., Nichols, L., Gale, N., Hobbs, F. R., & Marshall, T. (2017). Survival following a diagnosis of heart failure in primary care. Family practice, 34(2), 161-168