Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 36(1), ss. 155-162, Mart 2021
Cukurova University Journal of the Faculty of Engineering, 36(1), pp. 155-162, March 2021

Critical Parameter Selection and Survival Model Development for Heart Failure Patients

Ahmet AYDIN[*1]

[1]*Cukurova University, Faculty of Engineering, Department of Biomedical Engineering, Adana*

Geliş tarihi: 15.02.2021 *Kabul tarihi: 31.03.2021*

Abstract

Cardiovascular diseases are among the diseases that cause the most deaths worldwide. Heart failure is also one of the most common diseases, and mortality rates vary according to the patient's risk level. This distinct difference in mortality revealed the need to develop methods that could predict which patients have a worse prognosis and identify the risk group that would benefit more from intensive medical treatment and/or left ventricular assist devices and heart transplant treatments. In this study, survival models were developed using the dataset of 299 heart failure patients and Cox, Random Survival Forest, and Gradient Boosting Survival. Two different approaches are also used to determine the critical parameters in developing the survival model for heart failure patients. When a model is created using these parameters instead of all parameters in the dataset, higher success has been achieved, and this result is also better than the other studies using the same dataset. In conclusion, a survival model that can predict with high accuracy was developed for heart failure patients using the selected parameter set and Random Survival Forest.

Keywords: Heart failure, Survival analysis, Cox, Random survival forest, Gradient boosting survival

Kalp Yetmezliği Hastalarında Kritik Parametre Seçimi ve Sağkalım Modeli Geliştirilmesi

Öz

Kardiyovasküler hastalıklar dünya çapında en fazla ölüme neden olan hastalıklar arasındadır. Kalp yetmezliği de sık karşılaşılan hastalıklardan biridir ve hastanın taşıdığı risk seviyesine göre ölüm oranları değişiklik göstermektedir. Ölüm oranlarındaki bu belirgin farklılık, hangi hastaların daha kötü prognoza sahip olduğunu tahmin edebilen ve daha yoğun tıbbi tedaviden ve/veya sol ventriküler destek cihazlarından ve kalp nakli tedavilerinden daha fazla yararlanabilecek olan risk grubunu belirleyen yöntemlerin geliştirilmesinin ihtiyaç olduğunu ortaya çıkarmıştır. Çalışma kapsamında kalp yetmezliği bulunan 299 hastanın verileri ve Cox, RSF ve GSB yöntemleri kullanılarak sağkalım modelleri geliştirilmiştir. Ayrıca iki farklı yöntem kullanılarak kalp yetmezliği hastalarının sağkalım modelinin geliştirilmesinde kritik rol oynayan parametreler belirlenmiştir. Veri setindeki tüm parametreler yerine belirlenen bu parametreler kullanılarak bir model oluşturulduğunda daha yüksek başarı elde edilmiştir ve elde edile bu sonuç aynı veri setini kullanan başka çalışmaların sonuçlarında da daha iyidir. Sonuç olarak

[*]Sorumlu yazar (Corresponding author): Ahmet AYDIN, aaydin@*cu.edu.tr*

seçilen parametre seti ve RSF yöntemi kullanılarak kalp yetmezliği hastaları için yüksek doğrulukla tahmin yapabilen bir sağkalım modeli geliştirilmiştir.

Anahtar Kelimeler: Kalp yetmezliği, Sağkalım analizi, Cox, RSF, GBS

1. INTRODUCTION

Heart failure is the condition in which the heart is unable to pump enough blood to the body organs or can do this by increasing filling pressures. There are several causes of heart failure; coronary heart disease, diabetes, hypertension, alcohol or cocaine abuse, chemotherapy, or radiation therapy. Cardiovascular diseases are the top reason for death worldwide [1]. As society's lifespan extends and the life expectancy of heart failure patients increases with modern methods, the frequency of heart failure in society increases [2]. Despite improvements in treatment, heart failure patients' death rates still quite high and continue to be a huge burden on hospitalizations [1]. The life expectancy of patients with heart failure is different from each other, while the one and four-year mortality rate is 5% and 19% in New York Heart Association (NYHA) class 1 patients [3], 15%, and 40% in NYHA class 2-3 [4], and 6-month and 12-month mortality rates in class 4 are 44% and 64% [5]. This distinct difference in mortality has led to the need for risk stratification, which can estimate which patients have a worse prognosis and may benefit more from more intensive medical therapy and/or left ventricular assist devices and heart transplant treatments. Different risk assessment models have been developed to date in the light of population-based data on heart failure [6-9]. However, these risk models were proved to be useful, and there is substantial patient-to-patient variability. This variability can be addressed with novel machine learning-based methods and provide better prediction performance. Also, these methods can be used to determine the critical parameters for the survival modeling.

Ahmad et al. collected the data of 299 patients over the age of 40 with Class III or IV heart failure according to the classification of the NYHA and analyzed which parameters played a critical role in the mortality of heart failure patients [10]. Since the dataset contains censored data, Kaplan&Meier [11] and Cox [12] risk models are used for the analysis and found that age, serum creatin, blood pressure ejection fraction, and anemia play a critical role in the survival model [10]. Chicco et al. used statistical methods on the same dataset and ranked the collected parameters based on their importance [13]. As a result of the study, they concluded that the serum creatinine and ejection fraction parameters would be sufficient to predict the patient's survival with a certain accuracy.

Zahid et al. argued that men's and women's risk parameters would differ from each other due to their lifestyle and physiological differences [14]. They have analyzed the parameter importance for each gender and found that the male and female patients' critical parameters were different from each other. Thereupon, gender-based models were created and compared with the model created with the entire data set. As a result, while the c-index value of the model developed using all male and female patients' data was 0.72, the c-index of the models created with only male and female patients was 0.73 and 0.77, respectively.

Apart from parameter analysis, many studies have been conducted using such heart failure datasets to predict the patients' mortality [15–20]. In these studies, machine learning methods were used, and high accuracy predictions were made. While in some of these studies, the whole dataset was used, in others, parameter analysis was performed, and high estimation results were tried to be obtained with fewer parameters [15,17,20].

It provides valuable information to predict whether the patient will die with artificial intelligence methods. But the main purpose of collecting censored data is to determine the patient's risk level and indicate how long he might live. Besides, there are patients whose follow-up was abandoned in the early period of the study. For example, in the current data set, while there are patients who died on the 4th day of the 285-day follow-up period, there are also those who died on the 241st day; there are patients who were recorded “not die” in mortality prediction studies because they were not followed up on the 12th day, as well as patients who were followed for 285 days and did not die. Therefore, although some patients were not recorded as dead because their follow-up was ended in an early period of the study, a different result would have been obtained if they were followed up throughout the study. Therefore, new survival analysis methods have been developed that use the advantage of censored data and artificial intelligence methods’ learning capabilities. With the help of these methods, a high-accuracy survival model was developed by taking into account the censorship of the data.

This study shows that prediction performance can be increased by using a machine learning-based survival model and finding the parameters that play an important role in the survival analysis of heart failure patients.

2. MATERIALS AND METHODS

2.1. The Dataset

In order to develop a model for survival analysis, it is necessary to follow the patients for a certain period of time and to record when the patient dies. Within the scope of the study, the heart failure clinical dataset shared in the UCI Machine Learning Repository, which has this feature, was used [10,13]. This dataset includes data from 299 patients with Class 3 or 4 heart failure according to NYHA, which was followed up between April and December 2015 for 285 days [21].

Thirteen different parameters were collected from the patients. Some of these parameters are only binary data, while others take continuous values. The details about the collected data, short descriptions, and value ranges are presented in Table 1 [13]. The follow-up period and death event are the main target parameters used to calculate the survival function.

Table 1. Parameters used in the dataset, their meanings, and value ranges [13]

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

mcg/L: micrograms per liter, mL: microliter, mEq/L: milliequivalents per litre

2.2. Survival Analysis Methods

The purpose of survival analysis is to calculate a patient’s survival probability over time. The collected censored data is used to obtain a survival function $S(t)$ for that disease, and this function is used when evaluating the condition of the patients in the future [22]. The survival function will take a value of 1 when the study starts, and as time progresses, the risk of death of the patient will increase, and the survival probability will approach 0 as time passes, Figure 1 [22].

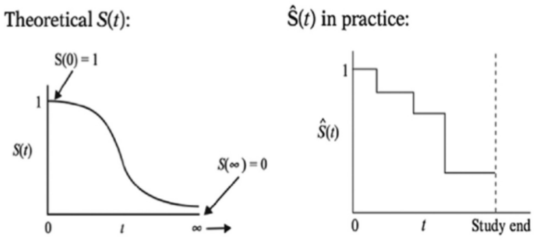

Figure 1. The theoretical (left) and practical (right) survival functions [22]

Survival analysis methods can be divided into three groups in general: non-parametric, semi-parametric, and parametric.

The most widely used non-parametric method is the Kaplan-Meier method [11]. Non-parametric methods are widely preferred as they are computationally simple. Equation 1 is used to obtain $S(t)$ by using the Kaplan-Meier method.

$$S(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} \tag{1}$$

In the given equation, n_i and d_i respectively represent those who are still at risk (patients who still alive) and patients who died at the time t_i.

As can be understood from Equation 1, this method does not consider other factors that may affect survival, such as age, gender, and current health status, and in this case, the prediction success is limited.

In the semi-parametric methods such as Cox Survival Analysis, the model is developed to make predictions considering the patient's current condition [12,23]. Thus, more successful estimates can be made by considering how other parameters affect this risk in addition to the patient's present baseline risk. In this method, the parameters in Equation 2 are found during model training, and the trained model is used to assess subsequent patients' conditions.

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 x_1 + \ldots + \beta_n x_n) \tag{2}$$

In the given equation, λ_0 represents the baseline risk, $x_1 \ldots x_n$ the input paramter values and $\beta_1 \ldots \beta_n$ the trained model paramters.

Although Cox Survival Analysis is widely used in many studies, it lacks in generalizing the dataset. The biggest problem of this method is that the proportional risk between two patients remains the same at all times. In other words, if a patient's condition is two times riskier compared to the other patient when the follow-up is started, the model assumes that he will always be two times risker than the other patient even in the evaluation months later. But this assumption is not always true. For this reason, machine learning methods reveal successful results in this regard, as in many other fields. Machine learning methods generalize the data set better and eliminate this proportional problem.

In this study, machine learning-based Random Survival Forest (RSF) and Gradient Boosting Survival (GBS) methods are used.

RSF is one method that can be used to eliminate the generalization disadvantage of Cox. In this method, firstly, a decision tree divides the data into different groups according to the existing parameters, and then another survival function is modeled for each subgroup [24]. Thus, instead of only a proportionally changing model with the same shape as the reference risk, patients with different characteristics are represented with other models. Therefore more successful predictions are obtained with RSF.

Similarly, the GBS is a decision tree-based method and can also model different survival functions. However, in the RSF method, subgroups are created independently, and the estimation results of the subgroups are averaged. In contrast, in the GSB, subgroups are created consecutively, aiming to increase performance [25].

The data set was first randomly divided into 25% test and 75% training set to evaluate the methods' performance. The model was developed using the training set, and then predictions were obtained on the model with the test data. Harrell's Concordance Index (c-index) was calculated as a performance parameter. [26]. The c-index shows the generated model's ability to provide survival times based on individual risk scores reliably. A c-index value of 1 indicates a perfect model, while 0.5 is equivalent to a random prediction.

3. RESULTS AND DISCUSSION

In this study, survival models were created using Cox, RSF, and GBS, and their performances were compared using the c-index. Then, the most critical parameters were defined and ranked by analyzing each parameter in the dataset. Finally, the best performing method and parameter set are compared with other studies using the same dataset.

Two different methods were used to determine the critical parameters. In the first method, separate Cox models were created using each parameter, and the predictive performance of these models was evaluated by calculating the c-index. Parameters showing higher success than other parameters alone were considered more important. The parameter ranking obtained with this method is given in Table 2.

Table 2. Parameter ranking obtained with Cox survival models

Rank	Parameter
1	Serum creatinine
2	Ejection fraction
3	Serum sodium
4	Age
5	Creatinine phosphokinase
6	High blood pressure
7	Anemia
8	Diabetes
9	Sex
10	Platelets
11	Smoking

Another method used to determine the critical parameters is training a survival model using all parameters with RSF and GSB, then calculating the prediction accuracy decrease after removing one parameter at a time. Thus, a collective parameter selection was made by considering the relationship between parameters.

In this way, the five most critical parameters have been determined as given in Table 3.

Table 3. Selected parameters after ranking with RSF and GSB

Rank	Parameter
1	Age
2	Ejection Fraction
3	Serum Creatinine
4	Platelets
5	High Blood Pressure

The first five parameters in Table 2 and the parameters in Table 3 are different. Therefore, the parameters' importance changes when their relationship is considered. Since the relationship between parameters is important, the parameters obtained in Table 3 will be more suitable to use.

After the critical parameters are defined, in addition to the survival models created using all parameters, other models are created with the selected parameters. In the first group, all five selected parameters are used, and three models are created using Cox, RSF, and GSB, then in the second group, the top three of the parameters are used to create three other models. The obtained c-index results are presented in Table 4.

To optimize the RSF and GSB survival models, different values of the minimum leaf, minimum sample split, learning rate, and the number of the estimators are tried. At first, the optimum values of the minimum leaf, minimum sample split, and learning rate are defined. Then using these defined values as constant, the number of the estimators is changed between 5-150 with five steps.

Table 4. The c-index results of the created models with different parameter sets. The values in the parenthesis show the number of estimators in which the best value is obtained

	All The Parameters	**5 Selected Parameters**	**3 Selected Parameters**
Cox	0.6477 (-)	0.6761 (-)	0.6482 (-)
RSF	0.7532 (35)	0.7798 (55)	0.7628 (30)
GSB	0.7046 (65)	0.7798 (145)	0.7706 (125)

The best prediction is obtained with the selected five parameters using RSF and GSB as 0.7798. However, the number of estimators used at GSB much higher (145) compared to RSF (55). Therefore RSF is performing better at modeling a survival function.

The contribution of the parameter selection on the performance can be seen from the results in Table 4. The best performance is obtained with the selected five parameters and the worst when all the parameters are used. It can be concluded that some parameters have a negative impact on the performance, and choosing too few parameters are not enough to obtain the best performance. Therefore the selected five parameters are the optimum ones.

The obtained results are compared with other studies using the same dataset. Zahid et al. created three different models. While using all patients' data in the first model, the other two models are for men and women. They performed parameter selection for each model. The c-index value is obtained as 0.72 using the model created with all patients' data and 0.73 and 0.77 with the male and female models [14]. A better c-index was obtained in this study as 0.78, using the model created with RSF for all patients.

4. CONCLUSION

This study shows that machine learning-based methods such as Random Survival Forest are more successful in creating a survival model than non-parametric or semi-parametric methods. Only a few parameters can be categorized in classical methods, and a model can be created for that patient group. After the model is trained with the RSF, certain groups are automatically determined according to all parameters in the data set, and separate risk models are automatically created for each subgroup. Thus, unlike other studies, particular models can be obtained for patients of different ages or other physiological conditions, even if their gender is identical.

Besides, when all parameters are used to create a survival model, parameters that negatively affect performance can also be included in the dataset, so a model that predicts with the desired accuracy cannot be obtained. Critical parameters are determined using Random Survival Forest and Gradient Boosting Survival, considering the relationships between parameters, and it was found that Age, Ejection Fraction, Serum Creatinine, Platelets, and High Blood Pressure play an essential role in model creation.

5. ACKNOWLEDGMENTS

I want to thank Cardiology Specialist Süleyman Özbiçer for his medical comments and guidance.

6. REFERENCES

1. Cardiovascular diseases. https://www.who.int/health-topics/ cardiovascular-diseases#tab=tab_1 (accessed Feb. 08, 2021).
2. Ho, K., Pinsky, J., Kannel, W., Levy, D., 1993. The Epidemiology of Heart Failure: The Framingham Study. Journal of the American College of Cardiology. 22(4), 6-42. 6A-13A. 10.1016/0735-1097(93)90455-A.
3. The SOLVD Investigators, 1992. Effect of Enalapril on Mortality and the Development of Heart Failure in Asymptomatic Patients with Reduced Left Ventricular Ejection Fractions, New England Journal of Medicine, 327(10), 685–691, doi:10.1056/NEJM199209033271003.
4. Yusuf, S., 1991. Effect of Enalapril on Survival in Patients with Reduced Left Ventricular Ejection Fractions and Congestive

Heart Failure," New England Journal of Medicine, 325(5), 293–302, doi: 10.1056/NEJM199108013250501.

5. Swedberg, K., Kjekshus, J., 1988. Effects of Enalapril on Mortality in Severe Congestive Heart Failure: Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS), The American Journal of Cardiology, 62(2), 60A-66A, doi: 10.1016/S0002-9149(88)80087-0.

6. Lee, D.S., Austin, P.C., Rouleau, J.L., Liu, P.P., Naimark, D., Tu, J.V., 2003. Predicting Mortality Among Patients Hospitalized for Heart Failure: Derivation and Validation of a Clinical Model, Journal of the American Medical Association, 290(19), 2581–2587. doi: 10.1001/jama.290.19.2581.

7. Aaronson, K.D., Cowger, J., 2012. Heart Failure Prognostic Models Why Bother?, Circulation: Heart Failure, Lippincott Williams & Wilkins Hagerstown, MD, 5(1), 6–9. doi: 10.1161/CIRCHEARTFAILURE.111.965848.

8. Levy, W.C., Mozaffarian, D., Linker, D.T., Sutradhar, S.C., Anker, S.D., Cropp, A.B., Anand, I., Maggioni, A., Burton, P., Sullivan, M.D., Pitt, B., Poole-Wilson, P.A., Mann, D.L., Packer, M., 2006. The Seattle Heart Failure Model: Prediction of Survival in Heart Failure. Circulation, 113(11),1424-1433. doi. org/10.1161/CIRCULATIONAHA.105.584102.

9. Brophy, J.M., Dagenais, G.R., McSherry, F., Williford, W., Yusuf, S., 2004. A Multivariate Model for Predicting Mortality in Patients with Heart Failure and Systolic Dysfunction, the American Journal of Medicine, 116(5), 300-304, doi.org/10.1016/j.amjmed.2003.09. 035.

10. 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, doi: 10.1371/journal.pone. 0181001.

11. Kaplan, E.L., Meier, P., 1958. Non-parametric Estimation from Incomplete Observations, Journal of the American Statistical Association, 53(282), 457–481, doi: 10.1080/01621459. 1958.10501452.

12. Collett, D., 2003. Modelling Survival Data in Medical Research, 2nd ed. Boca Raton, Fla. : Chapman & Hall/CRC, 391.

13. Chicco, D., Jurman, G., 2020. Machine Learning can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone, BMC Medical Informatics and Decision Making, 20(1), 16, doi: 10.1186/s12911-020-1023-5.

14. Zahid, F.M., Ramzan, S., Faisal, S., Hussain, I., 2019. Gender Based Survival Prediction Models for Heart Failure Patients: A Case Study in Pakistan, PLOS ONE, 14(2), doi: 10.1371/journal.pone.0210602.

15. Oladimeji, O.O., Oladimeji, O., 2020. Predicting Survival of Heart Failure Patients Using Classification Algorithms, JITCE (Journal of Information Technology and Computer Engineering), 4(02), 90–94, doi: 10.25077/jitce.4.02.90-94.2020.

16. Rahayu, S., Jaya Purnama, J., Baroqah Pohan, A., Septia Nugraha, F., Nurdiani, S., Hadianti, S., 2020. Prediction of Survival of Heart Failure Patients Using Random Forest, 16(2), 255-260. doi: 10.33480/PILAR.V16I2.1665.

17. Erdas, C.B., Olcer, D., 2020. A Machine Learning-Based Approach to Detect Survival of Heart Failure Patients, 1–4, doi: 10.1109/tiptekno50054.2020.9299320.

18. Le, M.T., Thanh Vo, M., Mai, L., Dao, S.V.T., 2020. Predicting Heart Failure Using Deep Neural Network, in International Conference on Advanced Technologies for Communications, 221–225, doi:10.1109/ ATC50776.2020.9255445.

19. Kucukakcali, Z., Cicek, I.B., Guldogan, E., Colak, C., 2020. Assessment of Associative Classification Approach for Predicting Mortality by Heart Failure, The Journal of Cognitive Systems, 5(2), 41–45, Accessed: Feb. 07, 2021. [Online]. Available: http://dergipark.gov.tr/jcs.

20. Chicco, D., Jurman, G., 2020. Survival Prediction of Patients with Sepsis from Age, Sex, and Septic Episode Number Alone, Scientific Reports, 10(1), 1–12, doi: 10.1038/s41598-020-73558-3.

21. Raphael, C., Briscoe, C., Davies, J., Whinnett, Z.I., Manisty, C., Sutton, R., Mayet, J., Francis, D.P., 2007. Limitations of the New York Heart Association Functional Classification System and Self-reported Walking Distances in Chronic Heart Failure, Heart, 93(4), 476–482, doi:10.1136/ hrt.2006.089656.

22. Deep Learning for Survival Analysis. https://humboldt-wi.github.io/blog/research/information_systems_1920/group2_survivalanalysis/ (accessed Feb. 10, 2021).

23. Cox, D.R., 1972. Regression Models and Life-Tables, Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202, doi: 10.1111/j.2517-6161.1972.tb00899.x.

24. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random Survival Forests, Annals of Applied Statistics, 2(3), 841–860, doi: 10.1214/08-AOAS169.

25. Friedman, J.H., 2001. Greedy Function Approximation: a Gradient Boosting Machine, The Annals of Statistics, 29(5), 1189–1232, Accessed: Feb. 05, 2021. [Online].

26. Uno, H., Cai, T., Pencina, M.J., D’agostino, R.B., Wei, L.J., 2011. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data, Statistics in Medicine, 30(10), 1105-1117. doi: 10.1002/sim.4154.