Proceedings of the 2nd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/53/20240234

Research on Predicting Survival in Heart Failure Patients of Logistic Regression Models

Huixin Man

School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing, 100124, China

21062319@emails.bjut.edu.cn

Abstract. The purpose of this paper is to study a logistic regression model for predicting the survival of patients with heart failure. Heart failure is a serious clinical syndrome that usually develops from multiple heart diseases, and its inadequate pumping function of the heart poses a significant threat to the life and health of the patient and a significant economic burden on the healthcare system. Using medical record data from 299 patients with heart failure in the UCI database, this research utilized logistic regression models to identify the critical factors influencing the survival of heart failure patients and to forecast their survival outcomes. This paper found that age, ejection fraction, serum creatinine concentration and follow-up period are significant factors affecting survival in patients with heart failure. By adopting backward elimination method to optimize the model, the accuracy of prediction is further improved. The optimized logistic regression model yields an area under the Receiver Operating Characteristic (ROC) curve of 0.845, showing high prediction accuracy. The conclusions of this study provide a new perspective for the early diagnosis, risk assessment and personalized treatment of heart failure.

Keywords: Heart failure, survival prediction, logistic regression, backward elimination method.

1. Introduction

Heart failure is a serious clinical syndrome that usually develops from multiple heart diseases, and its inadequate pumping function of the heart poses a significant threat to the life and health of the patient, while imposing a significant economic burden on the healthcare system. The causes of heart failure are diverse and the symptoms are changeable, so it is very important for medical research to use big data and machine learning models to accurately identify the key factors affecting heart failure and predict the survival rate of heart failure patients.

In terms of heart failure diagnosis, Wang et al. found that convolutional neural networks and long short-term memory networks in deep learning models have advantages in their ability to process key biomedical signals such as electrocardiogram and heart rate variability [1]. At present, there are many researches on the clinical symptoms of patients with heart failure in medical field. Among them, Zhao and Liu found that chronic heart failure (CHF) patients have a heavier burden of symptoms, which is related to the number of concurrent diseases, New York Heart Association (NYHA) grade and patient coping style [2]. In addition, there are many treatments for heart failure patients. In patients with chronic heart failure, dapagliflozin combined with sacubitril valsartan is beneficial to the improvement of heart

DOI: 10.54254/2753-8818/53/20240234

function, quality of life and self-care ability of patients, and the overall treatment effectiveness and safety are high [3]. In terms of traditional Chinese medicine, Warming Yang Qushi patch is beneficial to improve the symptoms of pneumonia in patients with viral pneumonia complicated with heart failure, and promote the recovery of heart function in the early stage, with significant efficacy and safety [4].

There are also many studies on the factors that affect heart failure. For example, Shi and Hu's weighted K-Nearest Neighbor (KNN) imputation algorithm based on mutual information has high filling accuracy, and the stochastic forest model based on this method has high prediction accuracy. Shi and Hu's research showed that red blood cell width, crystal infusion, and white blood cell count were the top three most influential features [5]. Heart failure in hemodialysis patients is a complex issue involving multiple risk factors. Through conditional logistic regression analysis, the study identified factors closely related to heart failure, such as age, diabetic nephropathy, hypertensive kidney damage, hemodialysis adequacy, erythropoietin (EPO) dosage, blood pressure, hemoglobin and albumin levels [6]. These findings provide important insights into the assessment and management of heart failure risk in hemodialysis patients. In an alternative approach, Davide Chicco and Giuseppe Jurman employed conventional biostatistical methods for feature ranking. Their analysis demonstrated that serum creatinine and ejection fraction alone were adequate for forecasting the survival rates of heart failure patients based on medical data. They found that predictions made with just these two variables were more precise than those made using the entire set of features from the original dataset [7]. Moreover, Jiang et al. analyzed the status and influencing factors of heart failure (HF) in patients with acute ST-segment elevation myocardial infarction (STEMI) in Bazhong City. A total of 80 STEMI patients were selected from Bazhong Central Hospital from January 1, 2021 to April 30, 2023, and grouped according to whether HF occurred, to explore the risk factors that may cause HF in STEMI patients. Conclusion HF in STEMI patients was associated with high expression of serum PCT, TBiL, hs-CRP, MPV and Hcy [8].

Despite some progress in predictive modeling of heart failure, there are still some challenges. Sarah et al. discussed prognostic uncertainty among those with severe obesity and outlined potential future directions [9]. Furthermore, Sidey-Gibbons and Sidey-Gibbons employed a simple example to elucidate the concepts and applications of machine learning for medical professionals and researchers. The principles they illustrated are readily transferable to other sophisticated tasks, such as natural language processing and image recognition [10]. Therefore, how to effectively integrate predictive models into clinical workflow and make them an auxiliary tool for doctors' decision-making is also a problem that needs to be solved in future research.

In conclusion, the various aspects of heart failure provide new perspectives for early diagnosis, risk assessment and personalized treatment of heart failure. The primary focus of this study is to investigate the determinants of heart failure and to forecast the survival rates of affected patients, utilizing logistic regression as the main analytical tool.

2. Methodology

2.1. Data Source

The data in this paper are derived from the UCI database, a database proposed by the University of California, Irvine, for applying machine learning. This study encompasses a dataset that includes the medical records of 299 heart failure patients, which were compiled over their monitoring period. Each patient's record encompasses 13 distinct clinical attributes.

2.2. Variable Description

The symbolic description of the data set is shown in Table 1 below. Among them, death_event is classified data and dependent variable, and the other variables are independent variables. This dataset has no missing values, so no missing values need to be filled in.

DOI: 10.54254/2753-8818/53/20240234

Table 1. Variable Description.

Variable name	Variable description	Type
age	Patients' age	Integer
anaemia	Reduction in erythrocyte or hemoglobin levels	Binary
cpk	Concentration of creatine phosphokinase in the serum	Integer
diabetes	If the individual is diabetic	Binary
ejec_frac	Ejection fraction of the heart	Integer
bld_pre	If the patient has hypertension	Binary
platelets	If the individual suffers from high blood pressure	Continuous
ser_crt	Serum creatinine concentration	Continuous
ser_sodm	Serum sodium concentration	Integer
sex	Woman or man	Binary
smoke	Whether the patient is a smoker or not	Binary
time	follow-up time	Integer
death_event	If the patient experienced mortality during the follow-up time	Binary

2.3. Method Introduction

Logistic Regression is a statistical technique used for binary classification tasks, where it estimates the likelihood of an outcome by modeling the data with a logistic function. In logistic regression, the predicted target variable is a Boolean variable, usually expressed as 0 and 1. This paper uses logistic regression to predict whether patients survive during the follow-up time.

Logistic regression operates by applying the logistic function to map the outputs of a linear regression to a probability scale ranging from 0 to 1, effectively linking the input features to the predicted probabilities. so as to achieve the prediction of binary events. If the probability is greater than 0.5, the patient is judged to have died during the follow-up period. Otherwise, the patient is considered alive.

Since some independent variables have no significant influence on the dependent variables, this paper further applies the backward elimination method to gradually eliminate the variables that contribute the least to the model until all remaining variables are statistically significant, so as to build a concise and effective regression model. The benefit of this approach is its ability to simplify the model's structure and enhance its interpretability.

3. Results and Discussion

3.1. Feature Visualization

This paper draws thermal maps of correlations between independent and dependent variables in order to intuitively identify variables that are strongly correlated with y, as shown in Figure 1.

By observing the heat map, it can clearly be seen that there is a strong correlation between age, ejec_frac, ser_crt, ser_sodm, and time and death_event. This means that these variables have an important effect on predicting death_event values.

DOI: 10.54254/2753-8818/53/20240234

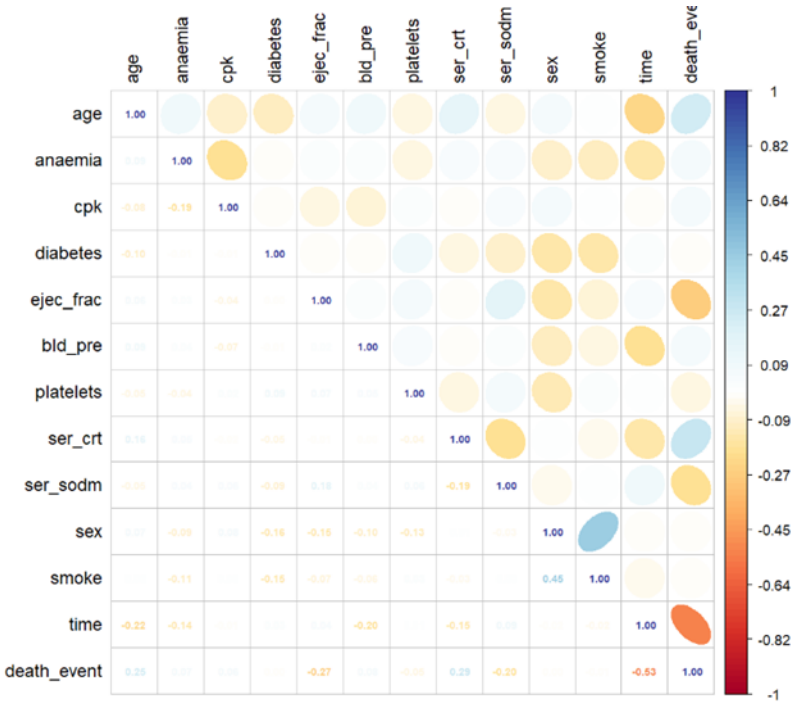

Figure 1. Correlation heat map.

3.2. Logistic Regression Model

Table 2 shows the coefficient estimation and statistical significance of each influencing factor of Logistic regression.

Table 2. Logistic regression result.

	Estimate	Standard Error	Z value	Pr(>|z|)
Intercept	1.018×10^{1}	5.657×10^{0}	1.801	0.07177
age	4.742×10^{-2}	1.580×10^{-2}	3.001	0.00269
anaemia	-7.470×10^{-3}	3.605×10^{-1}	-0.021	0.9835
cpk	2.222×10^{-4}	1.779×10^{-4}	1.249	0.2117
diabetes	1.451×10^{-1}	3.512×10^{-1}	0.413	0.6794
ejec_frac	-7.666×10^{-2}	1.633×10^{-2}	-4.695	0.0000
bld_pre	-1.027×10^{-1}	3.587×10^{-1}	-0.286	0.7747
platelets	-1.200×10^{-6}	1.889×10^{-6}	-0.635	0.5254
ser_crt	6.661×10^{-1}	1.815×10^{-1}	3.670	0.0002
ser_sodm	-6.698×10^{-2}	3.974×10^{-2}	-1.686	0.09186
sex	-5.337×10^{-1}	4.139×10^{-1}	1.289	0.1973
smoke	-1.349×10^{-2}	4.126×10^{-1}	-0.033	0.9739
time	-2.104×10^{-2}	3.014×10^{-3}	-6.981	0.0000

Table 2 indicates that the P-values for age, ejection fraction (ejec_frac), serum creatinine (ser_crt), and time are below the 0.05, signifying that these factors are statistically significant at the 5% level within the model. This suggests that they exert a considerable influence on the predictive outcomes. However, the p values of anemia, cpk, diabetes, platelets, sex and smoke are greater than 0.05, which are not significant in this model, and they have relatively little impact on the predicted results.

DOI: 10.54254/2753-8818/53/20240234

3.3. Backward Elimination Method

To enhance the model's predictive precision and mitigate the issue of overfitting, this paper adopted the backward elimination method to further screen the variables. This is done by gradually eliminating the variables that contribute the least to the model until all remaining variables are statistically significant.

After that, this study allocated the dataset into a training set and a test set with a proportion of 70% to 30%, respectively, for subsequent logistic regression analysis. The findings are presented in Table 3.

Table 3. backward elimination method result.

	Estimate	Standard Error	Z value	Pr(>|z|)
Intercept	-1.3578	0.2463	-5.514	0.0000
age	0.6650	0.2285	2.910	0.0036
ejec_frac	-1.1156	0.2528	-4.413	0.0000
ser_crt	0.8296	0.2289	3.624	0.0003
ser_sodm	-0.4038	0.2567	-1.573	0.1157
time	-1.6086	0.2879	-5.587	0.0000

By regression, Table 3 shows that this paper managed to retain variables that had a significant effect on y value, including age, ejection fraction (ejec_frac), serum creatinine (ser_crt), serum sodium (ser_sodm), and time.

Table 3 shows that the coefficient between age and serum creatinine (ser_crt) is positive, indicating that with the increase of age and serum creatinine content, the probability of death will increase. The coefficient among ejection fraction (ejec_frac), serum sodium (ser_sodm) and time is negative, meaning that an increase in ejection fraction, serum sodium content and a longer follow-up period may reduce the probability of death.

In terms of absolute coefficients, time have the greatest influence on mortality, followed by ejection fraction (ejec_frac) and serum creatinine (ser_crt). This suggests that among these significant factors, time is the most important predictor of survival.

3.4. Model Evaluation

Figure 2 compares the ROC curves of the original logistic regression with those of backward elimination method. By comparing AUC values, it can be seen that the area of the optimized logistic regression model under the ROC curve is 0.845, which is larger than the area of 0.841 before optimization. Therefore, the optimization model is useful.

Figure 2 (b) shows that the ROC curve is significantly away from the base line, and the entire curve is firmly above the critical value. This feature clearly indicates that the Logistic model has a high prediction accuracy. More specifically, the area under the ROC curve (AUC) reached 0.845, which is fairly close to the maximum possible value of AUC, 1.

Given that the AUC is a crucial metric for assessing classifier efficacy, a value closer to 1 indicates superior performance of the classifier. Graphically, the area below the ROC curve occupies a large part of the entire graph, which further confirms the excellent performance of the classifier.

DOI: 10.54254/2753-8818/53/20240234

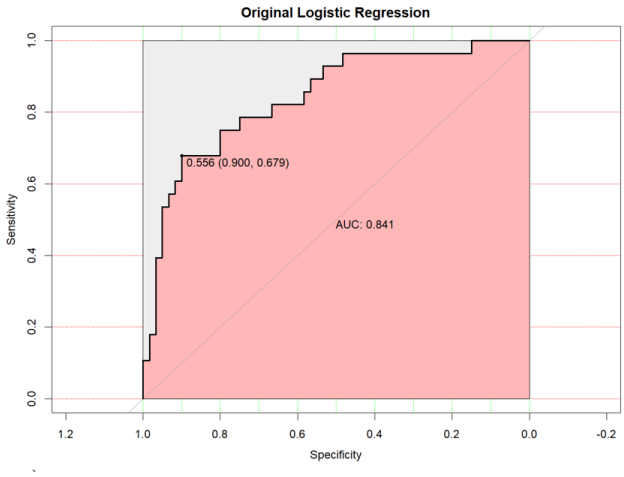

Figure 2. (a)

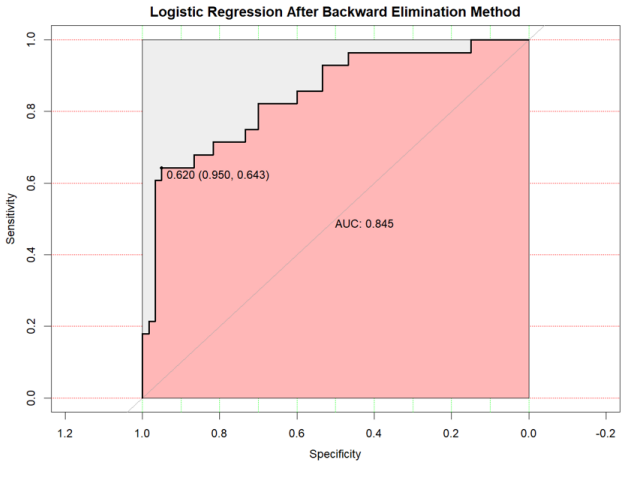

Figure 2. (b)

Figure 2. ROC curve comparison.

3.5. Model Prediction

Based on the optimized logistic model, this paper predicts the probability of patient death under a given predictive value. "0" represents patient survival, "1" represents patient death, and elements predicting a probability of death greater than 0.5 are transformed to 1, drawing the test set confusion matrix, as shown in Figure 3.

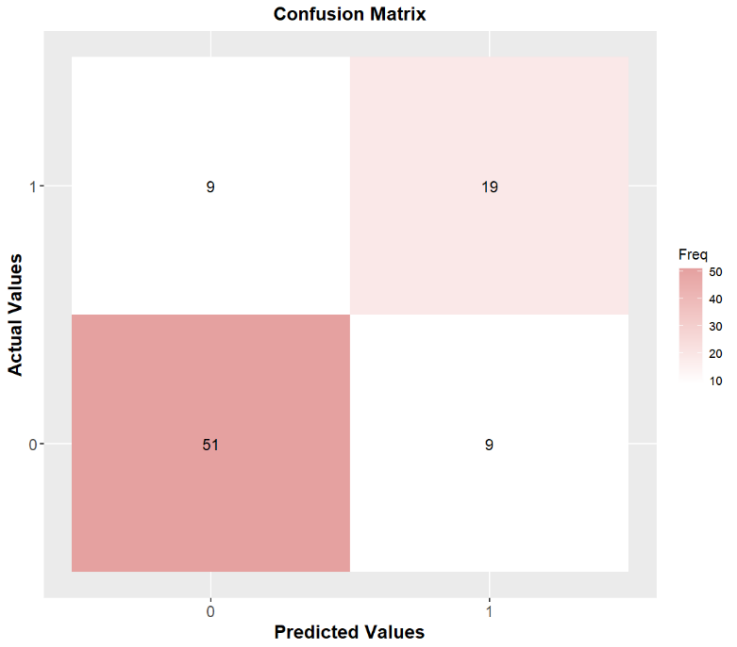

Figure 3. Test Set Confusion Matrix.

By confounding the diagonal elements of the matrix, it can be seen that the model correctly predicts the survival of most patients. Specifically, the model correctly predicted 51 patients to survive and 19 patients to die, for a total of 70 patients, with a correct prediction rate of 79.55%. This indicates that the prediction ability of the model is strong.

DOI: 10.54254/2753-8818/53/20240234

4. Conclusion

In this paper, the survival prediction of patients with heart failure was studied by using logistic regression model. By analyzing medical record data from 299 heart failure patients in the UCI database, this paper found that age, ejection fraction, serum creatinine concentration, and follow-up period were significant factors affecting patient survival. The model was optimized by backward culling method, and the accuracy of prediction was further improved.

The findings suggest that the probability of death increases with age and increased serum creatinine levels. The increase of ejection fraction, serum sodium content, and follow-up time may reduce the probability of death. Among all significant influencing factors, follow-up time had the greatest influence on mortality, followed by ejection fraction and serum creatinine.

The conclusions of this study provide a new perspective for the early diagnosis, risk assessment and personalized treatment of heart failure. However, although the model performs well statistically, how to effectively integrate the predictive model into the clinical workflow in practical clinical applications so that it can be used as a decision aid for physicians is still a question that needs to be addressed by future research.

Furthermore, the research has certain constraints, including a limited sample size, that could potentially impact the model's ability to generalize. Future studies may consider expanding the sample size to further verify the stability and reliability of the model. At the same time, more potential influencing factors and their relationship to heart failure survival can also be explored, with a view to building a more comprehensive and accurate predictive model.

References

[1] Wang, Y., Wei, D., Cao, H., et al. (2024) A review of the application of deep learning in heart failure detection. Computer Science and Exploration, 1-17.

[2] Zhao, Y. and Liu, H. (2019) Study on the status of symptom burden and its correlation with coping style in patients with chronic heart failure. Journal of Psychological Sciences, 19(15), 56-58+88.

[3] Yang, X., Sun, J. and An, X. (2018) Efficacy of Daglipzin combined with sacubactril valsartan in the treatment of patients with chronic heart failure. Journal of Clinical Rational Use of Drugs, 17(24), 17-20.

[4] Fan, M., Chang, L., Liang, L., et al. (2019) Retrospective study on the improvement of cardiac function in patients with viral pneumonia complicated with heart failure. Zhejiang Journal of Traditional Chinese Medicine, 59(08), 681-682.

[5] Shi, C. and Hu, Z. (2024) A mutual information weighted K-nearest neighbor filling algorithm (MIW-KNN): Application of heart failure with Clostridium difficile infection in mortality prediction. Henan Science, 1-12.

[6] Zhao Guozhong, et al. (2008) Logistic regression analysis of risk factors for heart failure in hemodialysis patients. Acta Academiae Medicinae Militaris Tertiae, 28(2), 131-144.

[7] Chicco, D. and 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, 2020,20(1), 16.

[8] Jiang, K., Li, L., Lai, Z., et al. (2019) Risk factors of heart failure in 80 patients with acute ST segment elevation myocardial infarction in Bazhong City. Chinese Journal of Emergency resuscitation and Disaster Medicine, 19(07), 855-857+866.

[9] Margosian, S. et al. (2024) Challenges in Prognostication of an Older Adult with Severe Obesity and End-Stage Heart Failure: A Case Study. Journal of palliative medicine.

[10] Sidey, G.J. and Sidey, G.C. (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol, 19, 64.