SURVIVAL PREDICTION AMONG POPULATIONS OVER 40 YEARS OLD WITH HEART FAILURE IN PAKISTAN BY USING MACHINE LEARNING METHODS

By: Yimin Yuan

B.A. Shaanxi Normal University, Shaanxi, China, 2021

Thesis advisor: Alex Dahlen, Ph.D.

A Master’s thesis submitted in partial fulfillment of the requirements for the degree of Master of Biostatistics

Department of Biostatistics

School of Global Public Health

New York University

May 2024

Survival Analysis and Machine learning prediction for heart failure mortality

Abstract

Introduction: Heart failure (HF), a category within cardiovascular disease, contributes to a considerable yearly global mortality. Patients typically incur approximately $24,000 in out-of-pocket expenses annually for treatment. The situation is more serious in Pakistan since healthcare resources are limited and medical insurance doesn't cover everyone. This study aims to characterize the mortality risk factors (RFs) for mortality and develop predictive models using machine learning techniques to do survival predictions in adults over 40 years old with severe HF symptoms in Pakistan.

Methods: Data for this study came from a publicly available dataset published on Kaggle called "Heart Failure Prediction" collected by Tanvir Ahmad and his colleagues in a case study in 2015. The study included variables such as age, gender, smoking status, anemia, diabetes, and high blood pressure. Cox regression model was used for survival analysis. Logistic regression, decision trees, random forests, and boosting models were fitted to determine the best predictors of mortality. The classification outcome used in this study is the endpoint of day 100 mortality.

Results: Age was the most significant variable, followed by diabetes, high blood pressure, and smoking status. The boosting model is selected to be the best model due to its lowest false negative rate of 0.84 and acceptable accuracy of 0.678 and AUC score of 0.717.

Conclusion: This study built the model for mortality prediction on HF patients, which can aid healthcare providers in prioritizing care effectively. Future research should aim to refine these models and reduce potential predictive errors to enhance their clinical utility.

Word count of abstract: 252 words

Word count of main text: 3802 words

Number of tables: 3

Number of figures: 2

Key word: heart failure, machine learning, survival analysis

Introduction

Cardiovascular disease (CVD) is a leading cause of death all over the world.The 2019 statistics revealed an incidence rate of 918.18 cases per 100,000 individuals and a mortality rate of 357.88 deaths per 100,000 individuals due to CVD.[1] Heart failure, a category within cardiovascular diseases, is also a significant global health concern, contributing notably to the worldwide mortality rate. According to the National Institutes of Health (NIH), HF can happen when the heart cannot pump enough blood and oxygen for the body's needs.[2] According to the European Society of Cardiology (ESC) guidelines, left ventricular ejection fraction (LVEF) is used to classify HF. HF can be categorized as HF with reduced ejection fraction (HFrEF) when LVEF is less than 40%, HF with mid-range ejection fraction (HFmEF) when LVEF is from 40% to 49%, and HF with preserved ejection fraction (EFpEF) when LVEF is larger than 50%.[3] The New York Heart Association (NYHA) has classified HF patients into four classes. Patients in class I have no physical activity limitations, patients in class II feel comfortable at rest but experience fatigue, palpitation, shortness of breath, or chest pain when doing ordinary physical activity, patients in class III feel uncomfortable even doing less than ordinary activity, and patients in class IV can have symptoms of HF at rest.[4]

HF is a serious chronic condition that requires medical treatment. The cost of treating HF is expensive. The median cost for HF in the US is $24,383 for each person per year.[5] If the patient needs left ventricular assist device (LVAD) implantation or heart transplantation, the cost can be even higher. Although most of the costs may be covered by Medicare, the patients may still have to pay significant out-of-pocket costs.[6] However, the situation is more serious in Pakistan. Unlike the developed countries, health insurance in Pakistan does not cover every person,[7] not everyone can afford the expensive cost of treatment. At the same time, hospitalization, recuperation at home, and unemployment due to heart failure can increase the patient's life pressure. So early diagnosis and appropriate treatment can help improve the quality of life and reduce the risk of complications for people with HF.[8,9]

There are many risk factors (RFs) for HF, which can be categorized into several groups: cardiovascular RFs, non-cardiovascular RFs, medications and toxins, familial inherited RFs, previous heart conditions, lifestyle factors, and infections.[10,11] Among the many RFs for HF, high blood pressure, smoking, stable angina, obesity, atrial arrhythmias, unstable angina, cancer, myocardial scar, diabetes mellitus, alcohol, severe anemia, and thyroid disorders were the top 12 most common.[12]

People with high blood pressure are more likely to develop HF. Long-term, high blood pressure will cause cardiac remodeling, which will lead to concentric left vascular hypertrophy (LVH) or predominant volume overload. When pressure overload persists, diastolic dysfunction occurs and results in HFpEF. On the contrary, when volume overload is sustained, the left ventricle dilates, and this can lead to HFrEF.[13] Moreover, the severity of high blood pressure can also affect the probability of getting HF. Research shows that the ambulatory nonresistant high blood pressure group (people whose clinic blood pressure is < or ≥140/90 mmHg and 24-h blood pressure ≥130/80 mmHg and taking ≤2 kinds of drugs) and the ambulatory resistant high blood pressure group (people whose clinic blood pressure ≥140/90 mmHg and 24-h blood pressure ≥130/80 mmHg and taking 3 or more drugs) had 2 times and 3.5 times higher risk of HF than controlled high blood pressure group (people whose clinic blood pressure <140/90 mmHg and 24-h blood pressure <130/80 mmHg).[14]

Evidence shows that people with diabetes are also more likely to develop HF. A retrospective cohort study analyzed the effect of diabetes on HF. Results showed that after the follow-up period of 6 years, the incidence of developing HF was 30.9 per 1000 people in the population with diabetes, and 12.4 per 1000 people in the population without diabetes.[15,16]

Iron deficiency anemia is common in HF. When the human body lacks iron, the oxidative metabolism function will be limited, which will lead to impaired red blood cell function, and finally lead to a lack of enough healthy red blood cells in the body.[17] A cohort study using the Meta-Analysis Global Group in Chronic (MAGGIC) dataset showed that 43% of the HFrEF patients and 42% of the HFpEF patients had anemia. People with anemia have 2 times the risk of dying from HF than those without anemia. [18]

People who smoke frequently may have higher mean left vascular mass index and lower left vascular circumferential strain than those who do not, and this can increase the risk of getting HF.[19] Results of a risk of community atherosclerosis risk study showed that people who smoke currently had two times the risk of getting HF than those who did not.[20] And compared to the patients who never smoked or formerly smoked, patients who currently smoke had the lowest survival probability from HF.[21]

Age and gender can also affect the risk of getting HF. The prevalence of developing HF in females is higher than in males. Most of the HF happened in the population aged 65 years and older. The risk of getting HF can increase as the age increases.[22,23]

In past studies, researchers have conducted survival analyses of patients with HF. The 1month, 1 year, 2 years, 5 years, and 10 years survival rates were 95.7% (95% CI=94.3–96.9), 86.5% (95% CI=85.4–87.6), 72.6% (95% CI=67.0–76.6), 56.7% (95% CI=54.0–59.4) and 34.9% (95% CI=24.0–46.8). The survival rates are significantly affected by the age at diagnosis. The survival rate at 1 year for people younger than 65 years is 10% higher than those who are older than 75 years.[24] The result of a cohort study in the United Kingdom showed that gender can also affect survival rates. Females had higher mortality rates than males both in short-term and long-term outcomes.[25] The type of heart failure is another important factor in survival. The result of a cohort study showed that the 5-year mortality rate for patients with HFrEF is 65%, while for patients with HFpEF is 52%.[26]

With the development of medical treatment, the survival rate of HF has increased during the past two decades. The 1-year survival has increased from 74% in 2000 to 80% in 2016, the 5-year survival has increased from 41% in 2000 to 48% in 2012, and the 10-year survival has increased from 19% to 26%.[25] However, as time passes by, the mortality rate of HF remains high.

In previous studies, researchers usually built models by using the risk factors of HF to predict whether the patients are likely to get HF or not, or they conduct survival analysis to find out the survival rate of the patient. Machine learning models are rarely used to study survival analysis of heart failure, especially for patients with serious symptoms.

The goal of this study is two-fold: first, to characterize the RFs for mortality; and second, to use these factors to create predictive models on mortality in the population with HF in Pakistan. Because of the limitation of the healthcare resources, these models can assist healthcare providers in identifying patients in much more serious conditions, prioritizing care for those who are most vulnerable, and providing critical care to those who need it most. Patients and their families can also benefit from understanding their risk of mortality by using these models. The information provided by the models can help set realistic expectations and motivate individuals to make lifestyle changes and adhere to prescribed treatments to improve their conditions. In this way, the survival rate of the patients can be improved, and the patients can live a better life.

Research question:

1. To what extent the RFs for mortality would affect the survival of adults over 40 years old in Pakistan with HF classed in NYHA III and NYHA IV?
2. Which machine learning model, including logistic regression, decision tree, random forest, and boosting would be the best model for mortality prediction?

Methods

Data Source and Study Design

Data for this study comes from a publicly available dataset published on Kaggle called "Heart Failure Prediction" collected by Tanvir Ahmad et al[27] in a case study in 2015. The goal of this study is to find out the effect of the RFs for mortality like age, ejection fraction, serum creatinine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure, gender, diabetes, and smoking status on patients with heart failure. The case study selected all of the patients with heart failure who were admitted to the Faisalabad Institute of

Cardiology and the Allied Hospital in Faisalabad from April to December 2015 in Pakistan. All of the patients were 40 years old or over, and all fell in NYHA class III defined as marked limitation of physical activity, comfortable at rest, less than ordinary activity causes fatigue, palpitation, shortness of breath, or chest pain, and class IV, defined as symptoms of heart failure at rest, any physical activity causes further discomfort. [27]

Participants and Procedures

All of the patients in this study were 40 years old or over, and all fell in NYHA class III, and class IV. Data was collected from both the patient's blood reports and the physicians' notes. The study was approved by the Institutional Review Board of Government College University, Faisalabad-Pakistan, and the principles of the Helsinki Declaration were followed. In the current study, all of the participants from the parent study will be included, the final sample size is 299.

Measures

Outcomes:

For Cox regression model:

'Time to death': This variable describes the elapsed duration from initiation of follow-up to the endpoint of death or the study's maximum duration of 285 days. It was measured by the physicians from the Institute of Cardiology and Allied Hospital Faisalabad-Pakistan.

For machine learning models:

'Endpoint': This variable describes whether or not the patient survived at day 100. For patients who were alive at day 100 were coded as 0, for patients who had died at day 100 were coded as 1, for patients who dropped out of the study were coded as null. It was also measured by the physicians from the Institute of Cardiology and Allied Hospital Faisalabad-Pakistan.

Exposures:

'Age': This variable describes how old the patients were when they were enrolled in the case study. It is a continuous variable and is taken from the patient's blood report.

'Sex': This variable describes the gender of the patients. It's taken from the patient's blood report. If the patient was male, then coded as 1, otherwise 0.

'Anemia': This variable describes whether the patient had anemia or not. It is accessed by the patient's hematocrit level on the blood report. Patients with hematocrit less than 36 were taken with anemia and were coded as 1, otherwise 0.

'Diabetes': This variable describes whether the patient has diabetes or not. It was assessed by the patient's blood report and was diagnosed by the physicians. Patients with diabetes were coded as 1, otherwise 0.

'High blood pressure': This variable describes whether the patient had high blood pressure or not. It was taken by the physicians' notes and was diagnosed by the physicians. Patients with high blood pressure were coded as 1, otherwise 0.

*'Smoking': Thi*s variable describes the patient's smoking status. It was first reported by the patient to the physician and taken from the physician's notes. If the patient was currently smoking or smoked before, then coded as 1, if the patient had never smoked, then coded as 0.

'Time': This variable describes the follow-up time and is a continuous variable. The follow-up time was 4-285 days, and the average was 130 days. No drop out or attrition during the follow-up.

Statistical Analyses

R version 4.2.3 and STATA version 17.0 were used in the statistical analyses for the current study. Univariate, bivariate, and multivariate analyses were conducted in the study. Statistical significance was determined by the 5% level.

For univariate analysis, mean, median, and standard deviation were calculated for continuous variables (age, time), and frequencies and proportions were calculated for categorical variables (death event, sex, anemia, diabetes, high blood pressure, and smoking).

For bivariate analysis, the association between death events and health conditions and health behaviors was analyzed. The chi-square test was used among categorical variables, and simple logistic regression was done among continuous variables.

For the multivariate analysis, the Cox regression model was utilized for survival analysis, to assess the influence of health conditions and behaviors on mortality rates over time. Logistic regression, decision tree, random forest, and boosting models were fitted for machine learning purposes. Interaction terms of sex and anemia, sex and diabetes, and sex and high blood pressure are added to the dataset, since the impact of these diseases on HF may be affected by the biological difference caused by the difference of sex. Given the limited size of the dataset, the data was split into a 70% training set and a 30% testing set. A larger proportion of the training set is better for enhancing the predictive accuracy of the model. Since only 72 patients had died and 227 patients were alive at day 100, the dataset is unbalanced. To address the imbalance within the dataset, decision tree models were developed using a training set in which observations of deaths during the follow-up period were over-sampled. The ovun.sample() function is applied for this purpose. This approach ensured a balanced distribution of outcomes, enhancing the reliability of the model. A 5-fold cross validation was applied to find the best size for the tree and pruned it. For each model, the true-positive rate (TPR) is measured by $\frac{True\ Positives}{True\ Postives+False\ Negatives}$, the false-positive rate (FPR) is measured by $\frac{False\ Positives}{True\ Negatives+False\ Positives}$, the false-negative rate (FNR) is measured by $\frac{False\ Negatives}{True\ Positives+False\ Negatives}$, and the true-negative rate (TNR) is measured by $\frac{True\ Negatives}{True\ Negatives+False\ Positives}$. Models were evaluated based on the accuracy, AUC scores, FPR, and FNR. The model with the highest accuracy and AUC score and lowest FNR was selected as the best model for future prediction.

Results

Univariate analysis and bivariate analysis

The total sample size was 299, 96 (32%) of the patients died, and 203 (68%) of the patients survived by the end of the follow-up period. The mean age of the patients was 60.83 (SD = 11.89) years, two-thirds (66.88%) of the total sample were male, and most of them (67.89%) had never smoked. For health conditions, 43.14% of the patients had anemia, 41.81% had diabetes, and 35.12% had high blood pressure, which means that one-third of the patients had underlying diseases.

The results of the bivariate analysis showed that only age and follow-up time were significantly associated with dying from HF. For the patients who survived by the end of the follow-up period, they had a mean age of 58.76 (SD = 10.64) years, and the mean follow-up time was 158.34 (SD = 77.61) days; for the patients who died by the end of the follow-up period, the mean age was 65.22 (SD = 13.21) years and the mean follow-up period was 70.89 (SD = 62.38) days (table 1). This conclusion provided evidence that older people are more likely to die from HF, and most of the patients died before day 100.

Multivariate analysis

For the survival analysis, the result of the Cox regression model indicated that age and high blood pressure were significantly associated with dying from HF. For every one-year increase in age, there was a 4.4% increase in the hazard of morality, and patients with high blood pressure would have a 52.1% higher hazard of mortality than those who didn’t.

For the machine learning part, the results of the training error, testing error, accuracy, and AUC of the logistic regression, decision tree, random forest, and boosting models are shown in table 3, and figure 2 shows the result of the ROC-AUC curve.

The result of the logistic regression model showed that age ($p < 0.001$) and diabetes ($p = 0.015$) were significantly related to dying from HF. Each one-year of age increases the odds of dying from HF by 6.28% (95% CI = 1.03-1.10) holding all other variables constant. Patients with diabetes had 4.58 (CI = 1.40-16.64) times the odds of dying from HF than those who didn’t have diabetes. The accuracy of the model was 0.711, and the AUC score was 0.742.

Due to the imbalance of the dataset, the simple decision tree model might develop a bias towards the majority class, which predicts almost all of the patients survived from HF. To address this issue, data was balanced to get a more equal number of patients who survived and had died from HF at day 100. The dataset after being balanced includes 162 patients who survived and 149 patients who had died from HF by the end of day 100. The result of the best tree bifurcates the dataset based on sex, anemia, age, and hypertension, resulting in leaf nodes. The model accuracy was 0.689, and the AUC score was 0.545.

The random forest model was built by using 300 trees, the model had a significant improvement in both training error and testing error. However, although the false-positive rate (FPR) decreased, the false-negative rate (FNR) increased. Since the model is used for mortality prediction, this trend is particularly concerning. For models used for clinical prediction purposes, a higher FPR is more acceptable compared to a higher FNR, since failing to identify the patient at risk is much more severe than identifying the patient at risk. The model accuracy was 0.700, and the AUC score was 0.678. Also, the importance of the variables was tested. The result showed that age and smoking were the two most significant variables, followed by the interaction terms of sex and anemia, sex and diabetes, and sex (figure 1).

The result of the boosting model showed that age has the most significantly highest influence on mortality prediction, followed by the interaction terms sex and high blood pressure, high blood pressure, and diabetes. The FPR slightly increased, however, the FNR was the lowest among all of the four models, which is a better signal. The model accuracy was 0.678, and the AUC score was 0.717.

Discussion

This study examined the impact of the risk factors of mortality on the survival of adults over

40 years old in Pakistan with HF. Although the result of the Cox regression model from the survival analysis, and the logistic regression, decision tree, random forest, and boosting models from the machine learning prediction gave different ranks of the weight of each variable, age was a particularly consistent factor across different analyses. Diabetes and high blood pressure were other two important factors that might affect the survival of the patients.

For the interaction terms, they showed the biological differences influenced by gender further enriching the analysis. The boosting model results identified the importance of the interaction terms sex and high blood pressure and sex and diabetes, which indicated that sex can be a moderator.

Model selection was done based on the model accuracy, AUC score, and FNR. The logistic regression model had the highest accuracy which was 0.711, followed by the random forest 0.700 and the decision tree model 0.689. For AUC scores, the logistic regression model also had the highest AUC score which was 0.742, followed by the boosting model 0.717 and the random forest model 0.678. However, the boosting model has the lowest FNR of 0.84, followed by the decision tree model of 0.88, and the logistic regression model of 0.96. Based on the mortality prediction purposes, the boosting model was chosen as the final prediction model due to its lowest FNR and the acceptable model accuracy and AUC score.

Strengths and Limitations

This study used both survival analysis and machine learning methods to measure the importance of the variables, thereby increasing the credibility of the results. Also, the machine learning models were fitted by using 5-fold cross validation, which decreased the model over-fitting probability. However, there were still several limitations. First, since all the diseases in the parent study were recorded as yes or no, the severity of the diseases is not described, and it is hard to find out the effect of the severity of the diseases on mortality. Also, the sample size of the dataset is only 299, and the number of variables in the current study is only 9 including the interaction terms, this may cause the model accuracy not to be so well and result in a high FPR and FNR. Moreover, selection bias may be existed, because all the patients come from the same area, and this may affect the generalizability of the result. Finally, the variability in admission dates among patients in this study might also affect the result. During the follow-up period, some of the patients had already been in the hospital for a long time, but some were just hospitalized, this might introduce potential biases in survival analysis and may provide

Conclusions

In conclusion, age was a particularly consistent factor across different analyses. Diabetes, high blood pressure, and smoking might also affect the survival of the patients. This is the same as the result of the prior research. The results of the random forest and boosting models identified the importance of the interaction terms, which indicated that sex can be a moderator. The boosting model was the best model among these four models for mortality prediction due to the lowest FNR and the second highest AUC score.

However, since this model is fitted for mortality prediction purposes, there is a pressing need for ongoing research to enhance its model accuracy and decrease both FPR and FNR. The lowest FPR and FNR for this study are 0.03 and 0.84, which is contrary to the objective of the models for clinical prediction purposes. Achieving higher accuracy and lower error rates can provide healthcare providers with more dependable reference values. This improvement will enable the provision of more precise and effective treatment options to patients, ultimately contributing to a reduction in heart failure mortality. Model accuracy evaluation while simultaneously reducing FPR and FNR can be enhanced through several strategic approaches. First of all, increasing the size of the dataset can help the models to have more data to learn from, which can reduce the variability and improve the generalizability of the usage. Second, introducing more RFs for mortality like stable angina, unstable angina, atrial arrhythmias, obesity, alcohol consumption, and thyroid can improve the sensitivity and specificity of the model by focusing on the variables that would have significant direct impacts on HF.[12] Moreover, removing the irrelevant features can help to minimize the noise and focus on the most important features. Last but not least, fitting the advanced model like the bagging model might help to reduce the variance and avoid model overfitting, and finally result in better accuracy and model performance.

Reference

1. Samad Z, Hanif B. Cardiovascular Diseases in Pakistan: Imagining a Postpandemic, Postconflict Future. *Circulation*. Apr 25 2023;147(17):1261-1263. doi:10.1161/CIRCULATIONAHA.122.059122

2. National Heart L, and Blood Institute. What is Heart Failure? Accessed November 2, 2023. https://www.nhlbi.nih.gov/health/heart-failure

3. Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the Diagnosis and Treatment of Acute and Chronic Heart Failure. *Rev Esp Cardiol (Engl Ed)*. Dec 2016;69(12):1167. doi:10.1016/j.rec.2016.11.005

4. Association AH. Classes and Stages of Heart Failure. Accessed December 17, 2023. https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

5. Urbich M, Globe G, Pantiri K, et al. A Systematic Review of Medical Costs Associated with Heart Failure in the USA (2014-2020). *Pharmacoeconomics*. Nov 2020;38(11):1219-1236. doi:10.1007/s40273-020-00952-0

6. Faridi KF, Dayoub EJ, Ross JS, Dhruva SS, Ahmad T, Desai NR. Medicare Coverage and Out-of-Pocket Costs of Quadruple Drug Therapy for Heart Failure. *J Am Coll Cardiol*. Jun 28 2022;79(25):2516-2525. doi:10.1016/j.jacc.2022.04.031

7. Shaikh BT, Ali N. Universal health coverage in Pakistan: is the health system geared up to take on the challenge? *Global Health*. Jan 12 2023;19(1):4. doi:10.1186/s12992-023-00904-1

8. Lesyuk W, Kriza C, Kolominsky-Rabas P. Cost-of-illness studies in heart failure: a systematic review 2004-2016. *BMC Cardiovasc Disord*. May 2 2018;18(1):74. doi:10.1186/s12872-018-0815-3

9. Hessel FP. Overview of the socio-economic consequences of heart failure. *Cardiovasc Diagn Ther*. Feb 2021;11(1):254-262. doi:10.21037/cdt-20-291

10. Bozkurt B, Hershberger RE, Butler J, et al. 2021 ACC/AHA Key Data Elements and Definitions for Heart Failure: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Clinical

Data Standards for Heart Failure). *J Am Coll Cardiol*. Apr 27 2021;77(16):2053-2150. doi:10.1016/j.jacc.2020.11.012

11. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J Am Coll Cardiol*. May 3 2022;79(17):e263-e421. doi:10.1016/j.jacc.2021.12.012

12. Banerjee A, Pasea L, Chung SC, et al. A population-based study of 92 clinically recognized risk factors for heart failure: co-occurrence, prognosis and preventive potential. *Eur J Heart Fail*. Mar 2022;24(3):466-480. doi:10.1002/ejhf.2417

13. Messerli FH, Rimoldi SF, Bangalore S. The Transition From Hypertension to Heart Failure: Contemporary Update. *JACC Heart Fail*. Aug 2017;5(8):543-551. doi:10.1016/j.jchf.2017.04.012

14. Coccina F, Pierdomenico AM, Cuccurullo C, Pizzicannella J, Trubiani O, Pierdomenico SD. Ambulatory Resistant Hypertension and Risk of Heart Failure in the Elderly. *Diagnostics (Basel)*. May 5 2023;13(9)doi:10.3390/diagnostics13091631

15. Michael Lehrke MD NMM. Diabetes Mellitus and Heart Failure. *The American Journal of Cardiology*. 2017;Volume 120(1)(1 July 2017):S37-S47. doi:https://doi.org/10.1016/j.amjcard.2017.05.014

16. Rosano GM, Vitale C, Seferovic P. Heart Failure in Patients with Diabetes Mellitus. *Card Fail Rev*. Apr 2017;3(1):52-55. doi:10.15420/cfr.2016:20:2

17. Singer CE, Vasile CM, Popescu M, et al. Role of Iron Deficiency in Heart Failure-Clinical and Treatment Approach: An Overview. *Diagnostics (Basel)*. Jan 13 2023;13(2)doi:10.3390/diagnostics13020304

18. Berry C, Poppe KK, Gamble GD, et al. Prognostic significance of anaemia in patients with heart failure with preserved and reduced ejection fraction: results from the MAGGIC individual patient data meta-analysis. *QJM*. Jun 2016;109(6):377-382. doi:10.1093/qjmed/hcv087

19. Kamimura D, Cain LR, Mentz RJ, et al. Cigarette Smoking and Incident Heart Failure: Insights From the Jackson Heart Study. *Circulation*. Jun 12 2018;137(24):2572-2582.

doi:10.1161/CIRCULATIONAHA.117.031912

20. Ding N, Shah AM, Blaha MJ, Chang PP, Rosamond WD, Matsushita K. Cigarette Smoking, Cessation, and Risk of Heart Failure With Preserved and Reduced Ejection Fraction. *J Am Coll Cardiol*. Jun 14 2022;79(23):2298-2305. doi:10.1016/j.jacc.2022.03.377

21. Sandesara PB, Samman-Tahhan A, Topel M, Venkatesh S, O'Neal WT. Effect of Cigarette Smoking on Risk for Adverse Events in Patients With Heart Failure and Preserved Ejection Fraction. *Am J Cardiol*. Aug 1 2018;122(3):400-404. doi:10.1016/j.amjcard.2018.04.016

22. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. *Eur J Heart Fail*. Aug 2020;22(8):1342-1356. doi:10.1002/ejhf.1858

23. 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*. Feb 10 2018;391(10120):572-580. doi:10.1016/S0140-6736(17)32520-5

24. Jones NR, Roalfe AK, Adoki I, Hobbs FDR, Taylor CJ. Survival of patients with chronic heart failure in the community: a systematic review and meta-analysis. *Eur J Heart Fail*. Nov 2019;21(11):1306-1325. doi:10.1002/ejhf.1594

25. Taylor CJ, Ordonez-Mena JM, Roalfe AK, et al. Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: population based cohort study. *BMJ*. Feb 13 2019;364:l223. doi:10.1136/bmj.l223

26. Shah KS, Xu H, Matsouaka RA, et al. Heart Failure With Preserved, Borderline, and Reduced Ejection Fraction: 5-Year Outcomes. *J Am Coll Cardiol*. Nov 14 2017;70(20):2476-2486. doi:10.1016/j.jacc.2017.08.074

27. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. *PLoS One*. 2017;12(7):e0181001. doi:10.1371/journal.pone.0181001

Tables and figures

Table 1. Bivariable analyses on key sociodemographic characteristics and health conditions among Heart Failure patients in NYHA Class III over 40 years old, Pakistan, 2015 (n = 299)

	Total(n=299)	Death Event		
		Survived (Censored) (n=203)	Dead (n=96)	
	n(%), Mean ± SD, or Median (IQR)	n(%), Mean ± SD, or Median (IQR)	n(%), Mean ± SD, or Median (IQR)	P-value
Age				
Mean	60.83 (±11.89)	58.76 (±10.64)	65.22 (±13.21)	<0.001[1]
Median	60 (51-70)	60 (50-65)	65 (55-75)	
Sex				
Male	194 (64.88)	132 (65.02)	62 (64.58)	0.941[2]
Female	105 (35.12)	71 (34.98)	34 (35.42)	
Anemia				
Yes	129 (43.14)	83 (40.89)	46 (47.92)	0.252[2]
No	170 (56.86)	120 (59.11)	50 (52.08)	
Diabetes				
Yes	125 (41.81)	85 (41.87)	40 (41.67)	0.973[2]
No	174 (58.19)	118 (58.13)	56 (58.33)	
High blood pressure				
Yes	105 (35.12)	66 (32.51)	39 (40.63)	0.170[2]
No	194 (64.88)	137 (67.49)	57 (59.37)	
Smoking				
Never smoked	203 (67.89)	137 (67.49)	66 (68.75)	0.827[2]
Currently smoking or smoked before	96 (32.11)	66 (32.51)	30 (31.25)	
Follow-up time				
Mean	130.26 (±77.61)	158.34 (±67.74)	70.89 (±62.38)	<0.001[1]
Median	115 (73-205)	172 (95-213)	44.5 (25-104.5)	

1 Pearson's R

2 Chi-square test

Table 2. Result of the Cox regression model

	Hazard Ratio	P-value
Age	1.04 (1.03-1.06)	1.6e-06 ***
Sex	1.01 (0.63-1.61)	0.970
Anemia	1.33 (0.89-1.99)	0.167
Diabetes	1.13 (0.74-1.72)	0.569
High blood pressure	1.52 (1.01-2.30)	0.0469 *
Smoking	1.12 (0.69-1.82)	0.651

Table 3. Results of the model selection

	Training Error	Testing Error	Accuracy	AUC
Logistic regression	0.215	0.289	0.711	0.742
Decision tree	0.249	0.311	0.689	0.545
Random forests	0.148	0.300	0.700	0.768
Boosting	0.158	0.322	0.678	0.717

Figure 1. The importance of the variables in random forest model

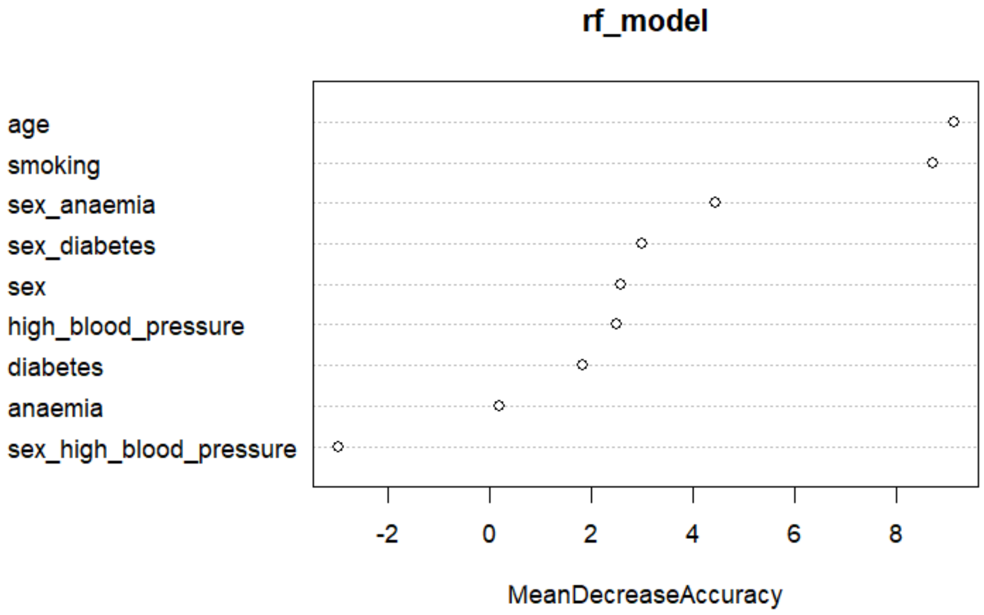
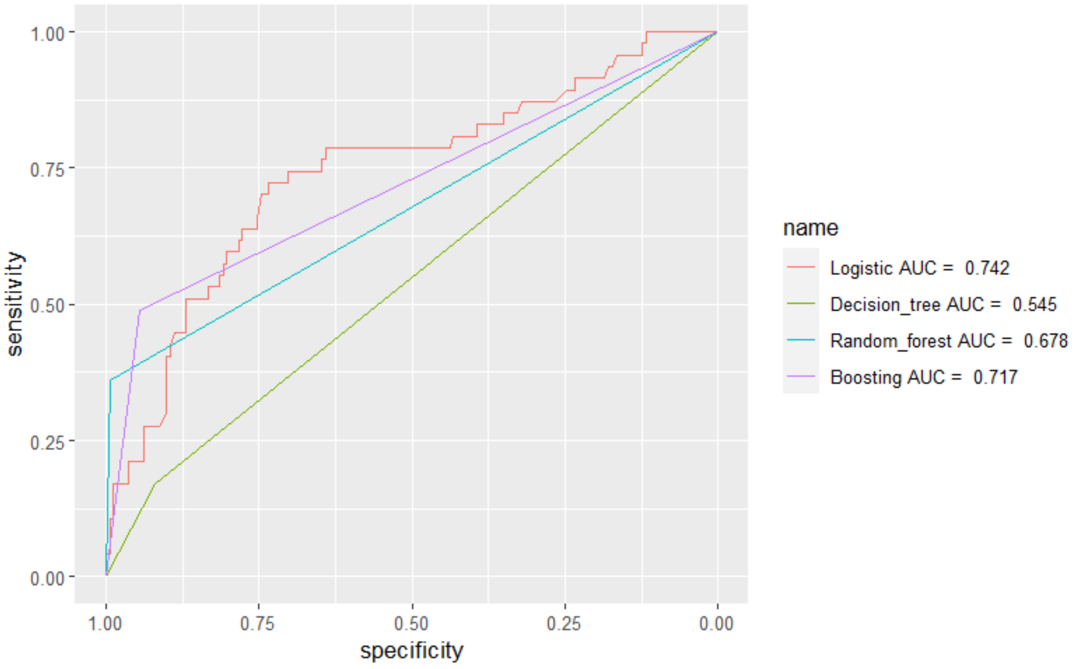


Figure 2. the ROC-AUC curve of different models



Appendix

Appendix 1. Detailed list of variables

Theme	Questions	Responses
Demographics		
Variable Name		
Age	How old is the patient?	#years
Sex	What is the patient's sex?	1 – Male 0 – Female
Time	How long is the follow-up time?	#days
Health Behaviors		
Variable Name		
Smoke	Is the patient currently smoking or had ever smoked?	1 – Currently smoking or had ever smoked 0 – Never smoked
Health conditions		
Variable Name		
Diabetes	Does the patient have diabetes?	1 – Yes 0 – No
Anemia	Does the patient have anemia?	1 – Yes 0 – No
High blood pressure	Does the patient have high blood pressure?	1 – Yes 0 – No
Death event	Does the patient die during the follow-up period?	1 – Yes 0 – No