 doi:10.20944/preprints202105.0677.v1

SUPERVISED ANALYSIS FOR PHENOTYPE IDENTIFICATION: THE CASE OF HEART FAILURE EJECTION FRACTION CLASS

Running head: Artificial Intelligence and heart failure phenotype

Cristina Lopez[1], Jose Luis Holgado[1], Raquel Cortes[1], Inma Sauri[1], Antonio Fernandez[1], Jose M Calderon[1], Julio Nuñez[2], Josep Redon[1,3,4]

[1] Cardiovascular and Renal Research Group INCLIVA Research Institute University of Valencia

[2] Cardiology Hospital Clínico of Valencia

[3] Internal Medicine Hospital Clínico of Valencia

[4] CIBERObn, Carlos III Institute Madrid

Corresponding Author information

Josep Redon

INCLIVA Research Institute

Avda Blasco Ibañez 17

46010 Valencia Spain

e-mail: josep.redon@uv.es

 doi:10.20944/preprints202105.0677.v1

ABSTRACT

Artificial Intelligence are creating a paradigm shift in health care, being phenotyping patients through clustering techniques one of the areas of interest. **Objective**: To develop a predictive model to classify heart failure (HF) patients according to their left ventricular ejection fraction (LVEF), by using available data in Electronic Health Records (EHR). **Subjects and methods**: 2854 subjects more than 25 years old with diagnose of HF and LVEF measured by echocardiography were selected to develop an algorithm to predict patients with reduced EF using supervised analysis. Performance of the algorithm developed were tested in heart failure patients from Primary Care. To select the most influencing variables, LASSO algorithm setting was used and to tackle the issue of one class exceed the other one by a large proportion we used the Synthetic Minority Oversampling Technique (SMOTE). Finally, Random Forest (RF) and XGBoost models were constructed. **Results:** Full XGBoost model obtained the maximized accuracy, a high negative predictive value and the highest positive predictive value. Gender, age, unstable angina, atrial fibrillation and acute myocardial infarct are the variables that most influence FE value. Applied in the EHR data set with a total 25594 patients with an ICD-code of HF and no regular follow-up in Cardiology clinics, 6170 (21.1%) were identified as those pertaining to the reduced EF group. **Conclusion:** The algorithm obtained is able to rescue a number of HF patients with reduced ejection fraction that can be take benefit for a protocol with strong recommendation to succeed. Furthermore, the methodology can be used for studies with data extracted from the Electronic Health Records.

Keywords: heart failure, phenotype, left ventricular ejection fraction, primary care, artificial intelligence, supervised analysis

 doi:10.20944/preprints202105.0677.v1

Introduction

Artificial intelligence (AI), an interdisciplinary science with multiple approaches, is a wide-ranging branch of computer science. Advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector and medicine is not out of this road, being phenotyping patients through clustering techniques one of the areas of interest [1]. The goal of phenotyping patients is to allow identification of patient subgroups with similar presentation, prognosis and response to therapy.

Heart failure (HF) is a major health care problem worldwide for which left ventricular ejection fraction (LVEF) has established clinically useful phenotypes for guiding treatment to reduce associated mortality and morbidity [2,3]. Classically on heart failure has been recognized two LVEF phenotypes, reduced LVEF (HFrEF) and preserved (HFpEF) [4,5], although recently, European Society of Cardiology added a third intermediate LVEF phenotype. Even when ejection fraction (EF) class is an important predictor of treatment response data available in electronic health records (EHR) frequently lack of EF quantitative values, limiting their usefulness in clinical and health service research[6]. An immediate next step is to develop algorithms and strategies to identify their distinct phenotypes in the absence of EF measured by echocardiography.

Looking for a proxy to identify HF phenotypes, AI methods could be useful combining information that usually are collected in the EHRs. Machine learning is an application of artificial intelligence that focuses on how computers learn from data, whose methods and techniques are increasingly being applied in medicine. Disease identification[7] and pathology and image diagnosis[8-9], as well as clinical research, are some of the main applications of machine learning in epidemiology and clinical medicine among others[10]. To explore phenotypes of patients with chronic HF, prior studies have already used hierarchical clusters to classify HFpEF patients[11,12]. These studies are mainly focused on defining phenotypes rather than predict the EF class by using a proxy that takes the decision based on available information in EHR.

In the present study we develop a predictive model that classifies HF patients according to their LVEF by using available features such as age, gender and present diseases. The goal is to overcome the performance of previous studies by using a new approach, the supervised analysis, a subfield of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

machine learning where models can be trained to predict the class of the target variable with earlier knowledge of the output values from prior data[13] .

Subjects and Methods

Data Source and Study Population

Subjects older than 25 years with diagnosis of heart failure, ICD-9 codes (402.X1, 404.X1,404.X3,428 and 398.91) were selected from the EHR system of a Community people older than 25 years in 2012. From this data base we selected a group with available LVEF measured by echocardiography classified as HFrEF and HFpEF according to the EF measurement, LVEF <40% and ≥40%, respectively. A second group of patients with HF diagnosis in the absence of LVEF value in the EHR was collected with or without regular follow-up by Cardiology Departments (Supplementary Figure 1). Variables to be tested were selected from those codified in the ICD-9. The study was approved by the Ethical Committee of the Hospital Clinico in the scope of the BigData@Better Heart, a project founded in the IMI2 program (IMI2-FPP116074-2). Consent form was obtained from the patients with echocardiography study and the data of the second group were documented by a process of pseudo-anonymization, making it impossible to use this information to identify the patients since the only link between the data and the patient is a code not available to the researchers.

Selection of Variables and Analytical Procedure

Variables included demographic information, age and gender and ICD-9 codified diseases. The original dataset was split into two different partitions corresponding to the 80% and 20% of the original dataset. The first, training set, will be used to train the different models developed in the study. The second, test set, will be used to measure the performance of the models developed with the training set. This partition will be performed by using the *caret* package in R. This permits us to split our data by maintaining the proportion of classes in both partitions. The most influential variables were considered by feature selection methods.

Feature Selection

The least absolute shrinkage and selection operator (LASSO) was used. This method creates a regression model where the estimated coefficients β_i for each variable suffer a penalization [14,15] or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

are set to zero. In the following equation, we can see the general formula of the regression model expressed in vector notation:

$$Y = X\beta + \varepsilon, \quad (1)$$

where Y is the end-point vector (our target), X is the vector of the covariates in our model, β is the vector of the coefficients for these covariates and ϵ is a random error.

Estimation of β parameters typically is done in a way such as the sum of squares of the residuals is minimized, this is called the Ordinary Least Squares (*OLS)* approach, and the loss function to minimize is the following one:

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x'_i\hat{\beta})^2, \quad (2)$$

when LASSO is used, the LASSO penalization term is added to this formula, resulting in the following equation:

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x'_i\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|, \quad (3)$$

Here, λ is known as the regularization penalty. Say λ is set to zero, then:

$$L_{OLS}'(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x'_i\hat{\beta})^2 = L_{OLS}(\hat{\beta}), \quad (4)$$

then, minimizing $L_{OLS}'(\hat{\beta})$ means minimizing $L_{OLS}(\hat{\beta})$. Otherwise, if λ is set to 1, equation (3) turns into:

$$L_{OLS}'(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x'_i\hat{\beta})^2 + \sum_{j=1}^{m}|\hat{\beta}_j| = L_{OLS}(\hat{\beta}) + \sum_{j=1}^{m}|\hat{\beta}_j|, \quad (5)$$

and minimizing $L_{OLS}'(\hat{\beta})$ means minimizing $\sum_{j=1}^{m}|\hat{\beta}_j|$, which makes value coefficients much lower than in the case λ=0. To choose an optimum λ value, we will define a set of λ values and for each λ we will estimate $\hat{\beta}$ such that $L_{OLS}'(\hat{\beta})$ is minimum. Then we will have two paired sets of λ and $\hat{\beta}$ values. Those covariates that are not set to zero more often are the final selected.

Also, when features are correlated, the algorithm chooses the variable that provides more information at predicting, reducing the complexity of the model. We perform a LASSO algorithm setting the EF value as the end-point (y_i) and introducing the rest of the covariables in the model (x'_i).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Imbalanced Data Distribution

To tackle the issue of one class exceed the other one by a large proportion we used the Synthetic Minority Oversampling Technique (SMOTE)[16,17] included in *DMwR* package. This algorithm creates new minority class examples by extrapolating between existing ones. Although matching seems to be a convenient procedure before building any classification model, we also performed a predicting model using the original database. In order to avoid the problem of data leakage, the different techniques that we apply over the data, such as SMOTE, feature selection and hyperparameter tuning, should be applied only on the training set but not in the test set.

Model Development

We performed several models based on two different machine learning algorithms, Random Forest (RF) and XGBoost [18], to compare its overall performance. We constructed reduced RF and XGBoost models with 4 possible previous algorithm performances in the dataset: balanced data in combination with feature selection (LASSO) or using all the variables; and unbalanced data in combination with either feature selection or all the variables. Then we defined a set of values for the model hyperparameters to test all possible combinations and find the most convenient tuning. Moreover, to tune the different hyperparameters on the models, instead of splitting training dataset into training and validation sets, we will use the k-fold cross-validation process.

Performance Measurements

In each model several performance measurements were calculated optimizing sensitivity and specificity measurements but also taking care of Negative Predictive Value (NPV), Positive Predictive Value (PPV) and Accuracy. We have made this choice because we aim to isolate the HFrEF class. In this way, we will be sure that those predicted as HFrEF (or Positive) will be truly HFrEF.

When comparing between models with similar sensitivity and specificity values we will select the best by focusing on the other metrics and also the Precision-Recall Curve (PR Curve), since it gives a more informative picture of the model's performance than the ROC Curve when the datasets are highly skewed[19,20].

A diagram showing the full process is shown in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Results

Characteristics of the study population

A total of 2854 subjects with HF diagnoses and LVEF measurement were included. Mean age was 74 years old and 47% were females. Diabetes was present in 53.4% of the participants. The most recorded was hypertension (82.3%). From the complete data set HFrEF was present in 23.4%. Two hundred twenty eight were used to train the models, while fifty hundred seventy were used to test. This partition keeps the proportion of HFrEF and HFpEF registers of the complete data set. From the initial variables contained in the EHR, 13 appears relevant in the models. Age and sex distribution as well as the prevalence of the relevant variables in the study population are shown in Table 1.

Models Developed

We constructed two types of models, reduced and full. For obtaining reduced models we defined a set of different λ values. For each λ value, the algorithm built a single model adding that penalization (λ) to the coefficients. The most relevant variables are those that have appeared in many models, which means that its associated coefficients have been non-zero, Figure 2. The x-axis represents logarithmic lambda values (λ). Gender, age, unstable angina, atrial fibrillation (AF) and acute myocardial infarct (AMI) are the variables that most influence EF value.

To define the most convenient λ value we selected the λ that maximizes the area under the curve (AUC). In Figure 3 the two vertical lines indicate two optimum log(λ) values: the first one from the left corresponds to log (λ_{min}), the value that value maximizes the AUC model's, while the second corresponds to log (λ_{se}). Afterwards, we built a model setting λ at λ_{min}, and the variables whose coefficients are greater or lower than 0 are the predictor variables selected for the final model. The coefficient value associated to each variable: age (-0.02), gender (0.76), atrial fibrillation (-0.18), angina (0.27), hypertension (-0.23), valve disorders (-0.22), diabetes (0.25), anemia (-0.23), EPOC (-0.03), pulmonary hypertension (0.19), Obesity (-0.32), renal dysfunction (0.22) and myocardial infarction (0.14), Figure 4. These values represent the contribution of each covariate to the endpoint, the highest value the highest importance.

After the feature selection process, we construted two new datasets: the first one (Balance 1) results from oversampling the minority class in the original data set until reaching the majority class size. In the second one (Balance 2) the minority class has been also oversampled maintaining a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

reasonable balance between both classes without equalizing their sizes. Original and new datasets sizes and proportions are shown in Table 2.

Table 3 summarizes the performance of the candidate models. All the results were obtained from testing these models in the testing dataset. While NPV is high among all models (ranging between 0.84 and 0.88), PPV presents a higher variance (0.44 as the lowest value vs. 0.75 as the highest). Note that models performed with the original dataset reached a higher accuracy varying from 0.8 to 0.84 as compared to models performed with balanced datasets. C-statistics was around 0,70 and the Precision-Recall Curve (AUCpr) gets its highest values with the original datasets, having its maximum value with the full XGBoost model (0.51).

Models Performance

The model was applied in a large data set of 79502 HF patients and among those 25594 patients treated in Primary Care in the absence of routine Cardiology consultation and without LVEF available, Table 4. Applying the algorithm can identify patients with HFrEF, 19169, among all patients with HF and 6170 among those without regular Cardiology consultation 6170, that can take benefit of a more scalable treatment.

Discussion

Left Ventricle Ejection Fraction phenotypes guide management of HF patients, but frequently is not recorded in the HER from Primary Care. Having alternatives to estimate the class of HF could be helpful not only for research in health care services but also for physicians at the time to choose better treatment. In the present study a machine learning algorithm to predict the phenotype category based on the main characteristics and diseases of the patient has been developed. Full XGBoost model performed offered the better modelling because it maximized the sensitivity, and it reached a high NPV. The present approach could be applied to other clinical conditions.

Few studies have approached to develop methods for predicting left ventricular ejection fraction in patients with heart failure. Some have used Administrative-Claims from Medicare[21,22,23] or specific data base such is the Swedish Heart Failure Registry[24] . In those using Administrative-claims a large number of variables, have been used in the training sample, identified by the ICD-code, while the Swedish Heart Failure used a restricted number but including laboratory parameters and

 doi:10.20944/preprints202105.0677.v1

treatments. The studies differ from the present in that included EF measured patients and uses different methodological approach. Lee et al[23], identified atrial fibrillation, obesity, pulmonary, hypertension and valvular disease to be significantly associated with developing heart failure with HFpEF, while male gender, history of cardiomyopathy, and myocardial infarction was significantly associated with the risk of heart failure with HFrEF. Overall, and despite limitations, routine clinical characteristics could potentially be used to identify different EF subphenotypes in databases.

Previous studies have also developed statistical and unsupervised learning algorithms to classify LVEF phenotypes. In the first one, the analysis included 11073 patients which is much larger than our sample size. What's more, the proportion of HFrEF and HFpEF individuals was well-balanced, leading to an easier distinction between classes. Despite all the above, the overall accuracy of the selected binomial logistic model did not overcome the measures that we obtained with our final model. Our analysis was based on supervised analysis and although machine learning techniques are far from being emergent technologies, its application on LVEF measure prediction is certainly innovative. In this study we mainly used two powerful algorithms: Random Forest and XGBoost. As a brief description of these algorithms, Random Forest is a combination between Decision Tree algorithms and Bagging, where both belong to supervised analysis. Together, they train a model to predict the class of the target variable with earlier knowledge of the output values deduced from prior data[13]. The other leading technique used, XGBoost, combines Boosting and Gradient Boosting algorithms. Boosting corrects sequentially the errors committed by the previous models which have wrongly classified the elements while Gradient Boosting tries to modelized the residuals, that is, transforming the errors into a function to avoid overfitting[13]. We choose these algorithms because these were the most suitable for the dataset and also for the binary class of the target. In addition, they achieved the best results among other models based on alternative machine learning algorithms such as Naive Bayes, Support Vector Machine and Artificial Neural Network. In particular, XGBoost is becoming popular in machine learning competitors and data scientists, as it has been battle tested for production on large-scale problems[20].

Applying the algorithm to the large data of patients with HF allowed to recognize around 24% of patients with HFrEF whom can take benefits of more precise treatment. Future research will include time variables such as time-to-inclusion from diagnoses dates as well as medication and hospital admissions. Furthermore, exploring other balancing techniques such as generating synthetic data based on the individual characteristic distribution could lead to analytical improvement.

 doi:10.20944/preprints202105.0677.v1

There are some limitations in our research that might be mentioned. As we collect the information from the EHR system, it was not a large amount of patient's LVEF measures and particularly, we dealt with an unbalanced dataset as HFrEF represents a minority of the total HF patients. In addition, there is a wide variety of performance measurements that can be used in order to evaluate the models. Depending on the characteristics and the goal of the problem, some metrics will perform better than others. The selection of the optimum λ was based on AUC metric, which is the most intuitive and typically used metric. Finallyu, our final goal was to maximize the PPV value which entails a relative lower value in the sensitivity analysis.

In conclusion, the presented step by step AI approach in the case of HF phenotype, it is a methodology that can help to obtain phenotypes from partially completed data base in different diseases, a common scenario in the EHRs.

 doi:10.20944/preprints202105.0677.v1

FUNDING

This work was supported by the BigData@heart (IMI2-FPP116074-2); BIGMEDILYTICS (ICT-15 - 780495); PI16/01402, CIBERObn Institute of Health Carlos III.

CONFLICTS OF INTEREST

Conflicts of Interest: none declared.

AKNOWLEDGMENT

We acknowledge the contribution of the Health Authorities of the “Conselleria de Salud Universal” Generalitat Valenciana for allowing to use the data.

Bibliography

1. Silkoff, P. E., Moore, W. C. & Sterk, P. J. Three Major Efforts to Phenotype Asthma: Severe Asthma Research Program, Asthma Disease Endotyping for Personalized Therapeutics, and Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome. Clinics in Chest Medicine 40, 13–28 (2019).

2. Redfield, M. M. Heart failure with preserved ejection fraction. New England Journal of Medicine 375, 1868–1877 (2016).

3. McMurray, J. J. V. Clinical practice. Systolic heart failure. N. Engl. J. Med. 362, 228–38 (2010).

4. Yancy, C. W. et al. 2013 ACCF/AHA guideline for the management of heart failure: A report of the american college of cardiology foundation/american heart association task force on practice guidelines. Circulation 128, (2013).

5. Task, A. et al. Guia Insuficiencia Cardiaca 2016. 2016 ESC Guidel. diagnosis Treat. acute chronic Hear. Fail. 2129–2200 (2016). doi:10.1093/eurheartj/ehw128

6. CVD Statistics. Available at: http://www.ehnheart.org/cvd-statistics.html. (Accessed: 30th March 2020)

7. Orange, D. E. et al. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. Arthritis Rheumatol. 70, 690–701 (2018).

8. Wang, S. et al. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). medRxiv 2020.02.14.20023028 (2020). doi:10.1101/2020.02.14.20023028

9. Giger, M. L. Machine Learning in Medical Imaging. J. Am. Coll. Radiol. 15, 512–520 (2018).

10. Deo, R. C. Machine learning in medicine. Circulation 132, 1920–1930 (2015).

11. Ahmad, T. et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J. Am. Coll. Cardiol. 64, 1765–1774 (2014).

12. Shah, S. J. et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 131, 269–279 (2015).

13. Supervised vs. Unsupervised Learning - Towards Data Science. Available at: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d. (Accessed: 30th March 2020)

14. Fonti, V. Feature Selection using LASSO. VU Amsterdam 1–26 (2017). doi:10.1109/ACCESS.2017.2696365

 doi:10.20944/preprints202105.0677.v1

15. (Tutorial) Regularization: Ridge, Lasso and Elastic Net - DataCamp. Available at: https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net. (Accessed: 30th March 2020)

16. Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans. Syst. Man, Cybern. Part ASystems Humans 41, 552–568 (2011).

17. Chawla, N.V., Bowyer, K.W., Hall, L.O., K. W. P. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. J. Artif. Intell. Res. 16, 321–357 (2002).

18. XGBoost Algorithm: Long May She Reign! - Towards Data Science. Available at: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. (Accessed: 30th March 2020)

19. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. in ACM International Conference Proceeding Series 148, 233–240 (2006).

20. XGBoost, a Top Machine Learning Method on Kaggle, Explained. Available at: https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html. (Accessed: 30th March 2020)

21. Bovitz T, Gilbertson DT, Herzog CA. Administrative data and the philosopher's stone: turning heart failure claims data into quantitative assessment of left ventricular ejection fraction. Am J Med 2016; 129: 223–225.

22. Desai RJ, Lin KJ, Patorno E, Barberio J, Lee M, Levin R, Evers T, Wang SV, Schneeweiss S. Development and preliminary validation of a Medicare claims-based model to predict left ventricular ejection fraction class in patients with heart failure. Circ Cardiovasc Qual Outcomes 2018; 11: e004700.

23. Lee MP, Glynn RJ, Schneeweiss S, Lin KJ, Patorno E, Barberio J, Levin R, Evers T, Wang SV, Desai RJ. Risk Factors for Heart Failure with Preserved or Reduced Ejection Fraction Among Medicare Beneficiaries: Application of Competing Risks Analysis and Gradient Boosted Model. Clin Epidemiol. 2020;15:607-616.

24. Uijl A, Lund LH, Vaartjes I, Brugts JJ, Linssen GC, Asselbergs FW, Hoes AW, Dahlström U, Koudstaal S, Savarese G. A registry-based algorithm to predict ejection fraction in patients with heart failure. ESC Heart Fail. 2020;7:2388-2397.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

LEGEND OF FIGURES

Figure 1: Full process diagram.

Figure 2: The most relevant coefficients of the model based on their contribution in the FE class prediction. On the x-axis, Log Lambda refers to the logarithmic of the regularization parameter (λ), which controls the weight that each variable has in the model.

Figure 3: AUC values and its 95% confidence intervals for different logarithmic values of λ represented on x-axis, where λ is the regularization parameter in the model. Vertical lines correspond to $Log(\lambda_{min})$ and $Log(\lambda_{se})$ which are the most optimal Log(λ) values.

Figure 4: Coefficients' values fixing $\lambda = \lambda_{min}$ in the LASSO model. The up and down positions on the y-axis shows the variables which most contribute to the model as its associated coefficients take the highest values. On the middle positions are located the variables which contribute less to the model, as its associated coefficients are close to 0.

Supplementary Figure 1: Flow-chart of patient selection

 doi:10.20944/preprints202105.0677.v1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Table 1. Demographics and chronic diseases of the data sets included in the model

Variable	Training Data Set			Test Data Set		
	HFrEF	HFpEF	Total	HFrEF	HFpEF	Total
	n=535	n=1749	n=2284	n=133	n=437	n=570
Demographics						
Male	386 (72.15)	832 (47.57)	1218 (53.33)	92 (69.17)	203 (46.45)	295 (51.75)
Age, mean (SD)	71.85 (11.14)	75.8 (9.89)	74.88 (10.33)	69.92 (10.62)	75.6 (9.8)	74.27 (10.27)
Comorbidities						
Atrial fibrillation	197 (36.82)	776 (44.37)	973 (42.6)	32 (24.06)	225 (51.49)	257 (45.09)
Anemia	171 (31.96)	745 (42.6)	916 (40.11)	45 (33.83)	199 (45.54)	244 (42.81)
Diabetes	317 (59.25)	911 (52.09)	1228 (53.77)	71 (53.38)	224 (51.26)	295 (51.75)
Hypertension	418 (78.13)	1470 (84.05)	1888 (82.66)	93 (69.92)	368 (84.21)	461 (80.88)
Obesity	49 (9.16)	226 (12.92)	275 (12.04)	11 (8.27)	72 (16.48)	83 (14.56)
Pulmonary HTN	26 (4.86)	71 (4.06)	97 (4.25)	3 (2.26)	22 (5.03)	25 (4.39)
CKD	88 (16.45)	245 (14.01)	333 (14.58)	12 (9.02)	69 (15.79)	81 (14.21)
Valve disorders	66 (12.34)	317 (18.12)	383 (16.77)	9 (6.77)	69 (15.79)	78 (13.68)
Epoc	147 (27.48)	451 (25.79)	598 (26.18)	34 (25.56)	84 (19.22)	118 (20.7)
Myocardial infarction	149 (27.85)	311 (17.78)	460 (20.14)	42 (31.58)	75 (17.16)	117 (20.53)
Angina	239 (44.67)	560 (32.02)	799 (34.98)	61 (45.86)	146 (33.41)	207 (36.32)

Values are number (percentage); rEF reduced ejection fraction, pEF preserved ejection fraction; COPD chronic pulmonary disease
CKD stage 3 glomerular filtration rate <60ml/min/1.73m^2

 doi:10.20944/preprints202105.0677.v1

Table 2. Proportion of HFpEF and HFrEF phenotypes in the original data set and in the two balanced datasets created with *SMOTE*.

N (%)	Original	Balance 1	Balance 2
Total size	2284	2140	3745
HFpEF class	1749 (76.58)	1070 (50)	2140 (42.86)
HFrEF class	535 (23.42)	1070 (50)	1605 (57.14)

HFpEF, heart failure preserved ejection fraction
HFrEF, heart failure reduced ejection fraction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Table 3. Performance measures of the predictive models in the Testing Data Set

		AUC	AUCpr	Accuracy	Sensitivity	Specificity	PPV	NPV	Prediccion BM
XGBoost	**Full models**								
	Original	0.70	0.45	0.80	0.53	0.88	0.57	0.86	24.11
	Smote 50-50	0.69	0.38	0.70	0.69	0.70	0.41	0.88	26.05
	Smote balanced	0.65	0.35	0.72	0.53	0.77	0.41	0.84	21.63
	Reduced models								
	Original	0.70	0.46	0.81	0.49	0.90	0.60	0.85	17.47
	Smote 50-50	0.68	0.38	0.72	0.61	0.76	0.44	0.86	25.05
	Smote balanced	0.66	0.36	0.71	0.53	0.76	0.40	0.84	19.04
RF	**Full models**								
	Original	0.70	0.51	0.83	0.46	0.95	0.72	0.85	4.23
	Smote 50-50	0.69	0.38	0.73	0.65	0.75	0.44	0.88	16.57
	Smote balanced	0.72	0.44	0.77	0.62	0.82	0.51	0.88	15.42
	Reduced models								
	Original	0.70	0.51	0.84	0.46	0.95	0.75	0.85	3.8
	Smote 50-50	0.70	0.38	0.73	0.65	0.75	0.44	0.88	14.38
	Smote balanced	0.72	0.44	0.78	0.62	0.83	0.52	0.88	12.55

AUC area under the curve; AUCpr area under precision recall curve; PPV positive predictive value; NPV negative predictive value, RF Random Forest

 doi:10.20944/preprints202105.0677.v1

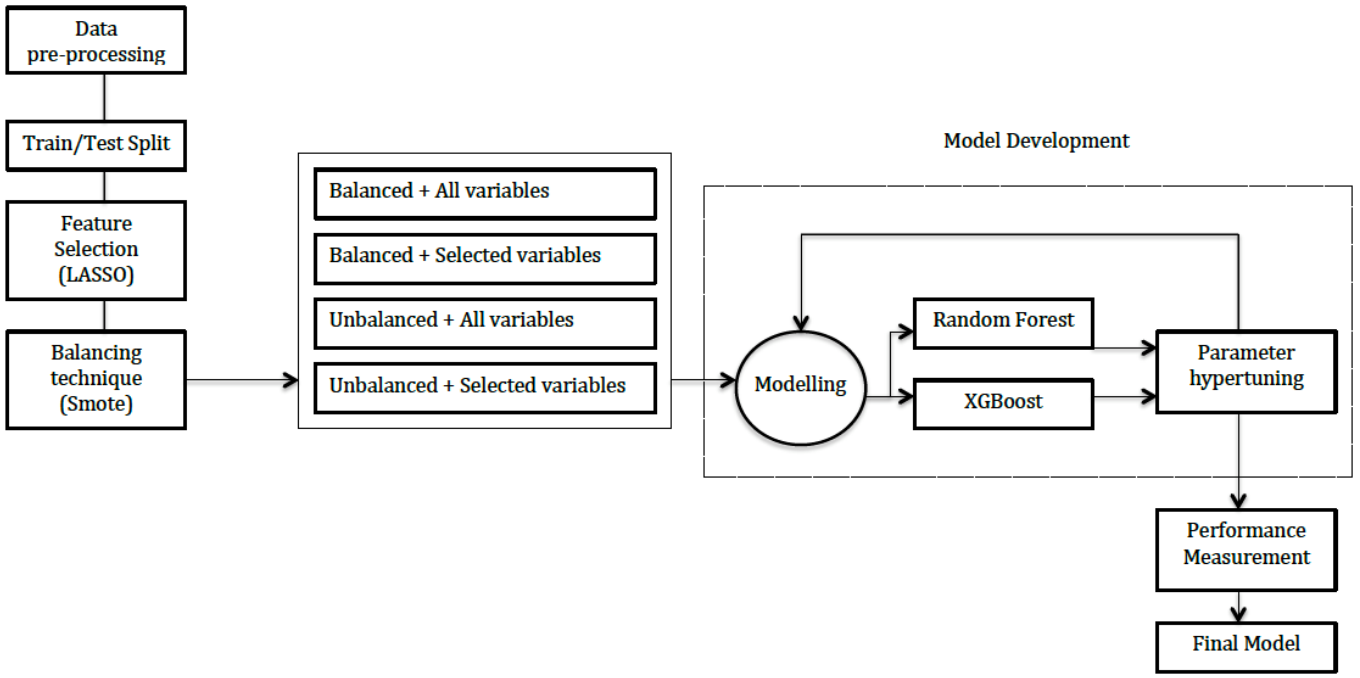
Table 4. Demographics and chronic diseases of the Heart Failure population tested

Variables	All subjects N=79057	Primary Care (N=26376)
Demographics		
Male	36539 (46.22)	10082 (38.22)
Age, mean (SD)	77.75 (11.35)	80.88 (10.36)
Comorbidities		
Atrial fibrillation	31277 (39.56)	6571 (24.91)
Anemia	30132 (38.11)	9197 (34.87)
Diabetes	31607 (39.98)	9998 (37.91)
Hypertension	66181 (83.71)	21048 (79.8)
Obesity	17599 (22.26)	3757 (14.24)
Pulmonary HTN	842 (1.07)	260 (0.99)
CKD	15469 (19.57)	2018 (7.65)
Valve disorders	13061 (16.52)	1016 (3.85)
COPD	20569 (26.02)	6647 (25.2)
Myocardial infarction	13243 (16.75)	2038 (7.73)
Angina	24655 (31.19)	4727 (17.92)

CKD Chronic kidney disease
COPD Chronic obstructive pulmonary disease

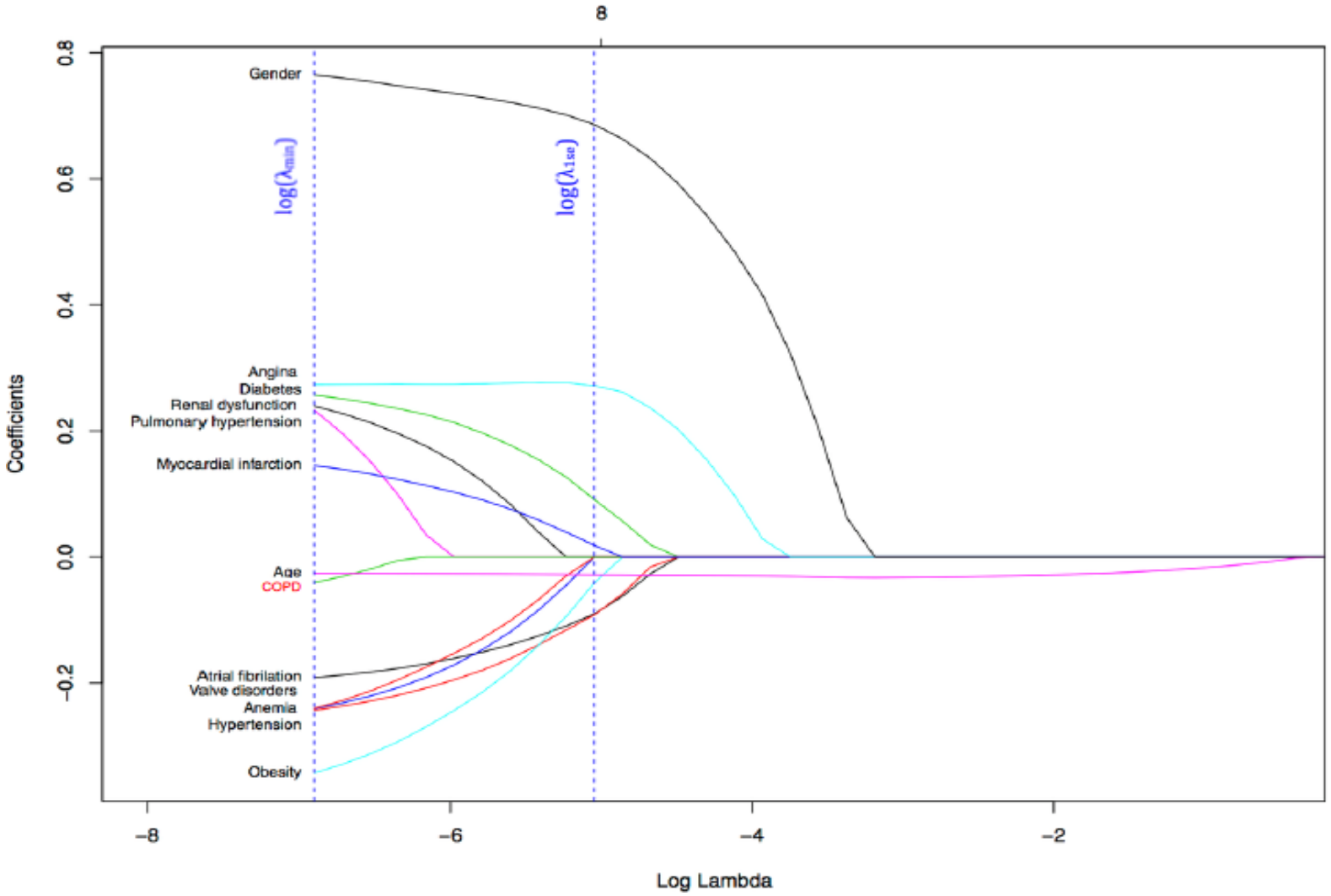
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Figure 1



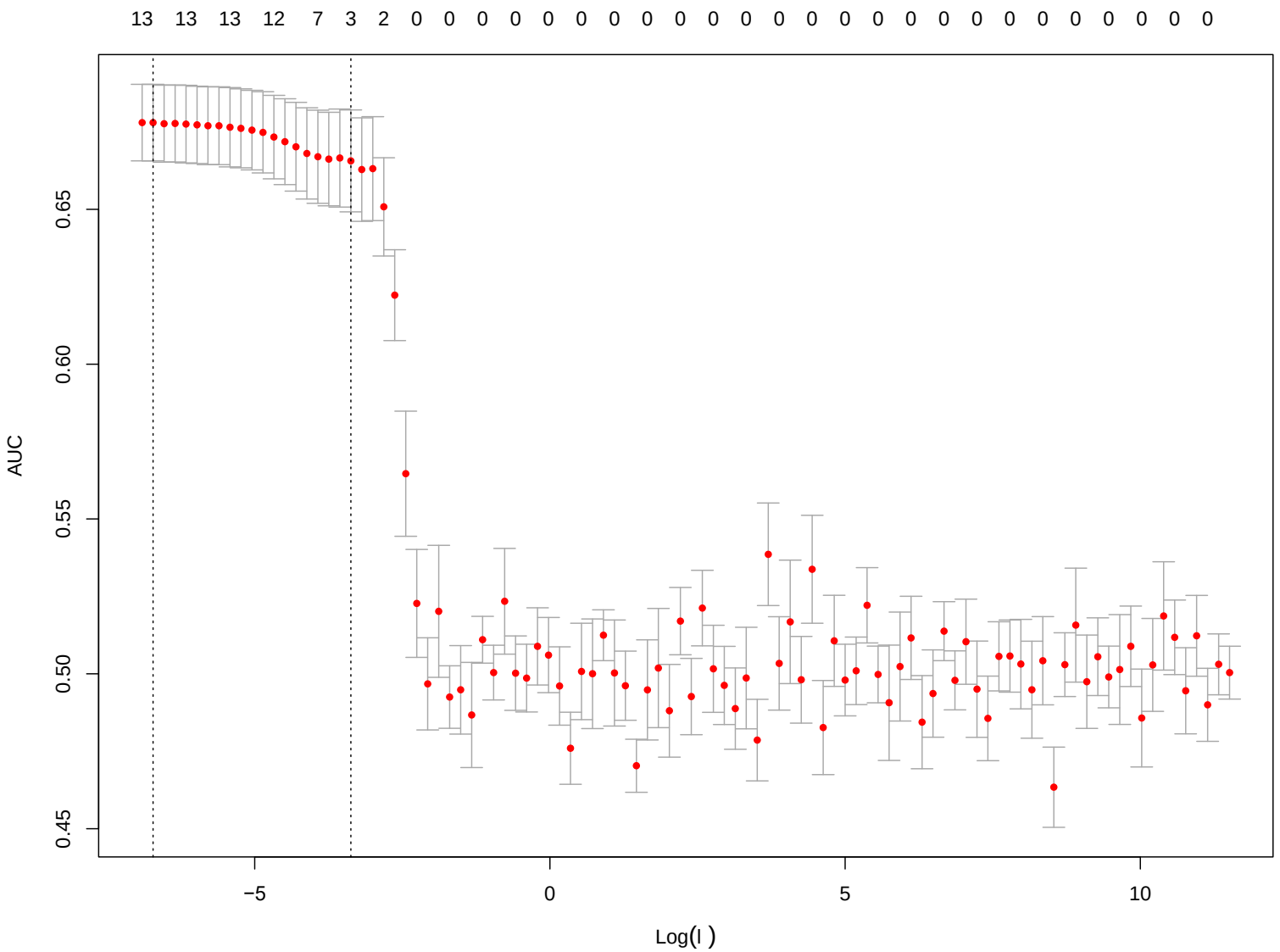
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Figure 2



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Figure 3



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2021 doi:10.20944/preprints202105.0677.v1

Figure 4