A Supervised Learning Method for the Classification of Heart Failure Using Electronic Health Record-Based Phenotypes

Nosheen Reza MD[1], *William Bone BA*[2], *Pankhuri Singhal BS*[2], *Anurag Verma PhD*[2], *Ashwin C. Murthy MD*[1], *Srinivas Denduluri PhD*[1], *Srinath Adusumalli MD MSc*[1], *Marylyn D. Ritchie PhD*[2], *Thomas P. Cappola MD ScM*[1]

[1]*Division of Cardiovascular Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania*

[2]*Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania*

ITMAT
Institute for Translational Medicine and Therapeutics

BACKGROUND

- Current classification schemes fail to capture the broader pathophysiologic heterogeneity in heart failure.
- Electronic health records (EHR) data can be used to phenotype complex diseases.
- We evaluated the performance of a linear support vector machine (SVM) to classify individuals with heart failure (HF) into 2 major HF subtypes using EHR data.

METHODS

- Using the institutional EHR, we identified 7665 patients with HF and extracted 1267 EHR-based features from time of HF diagnosis. Dataset was divided into halves for training and testing.
- We trained an SVM using repeated k-fold cross validation to predict HF subtype, defined as left ventricular ejection fraction < 40% versus > 50%. Model performance was assessed by evaluating accuracy, sensitivity, positive predictive value, negative predictive value, precision, and F1 score. Random forest with bootstrapped resampling was used to identify variable importance.

RESULTS

- The SVM model demonstrated nearly excellent discrimination for the prediction of HF subtype (AUC-ROC 0.79).
- The SVM model predicted 1,896 cases of HFpEF correctly from 2,309 total HFpEF cases and 866 cases of HFrEF correctly from 1,523 total HFrEF cases, resulting in a sensitivity of 82% and positive predictive value of 74%.
- The top five variables contributing to importance for HF subtype classification were systolic blood pressure, high-density lipoprotein cholesterol, diastolic blood pressure, creatinine, total cholesterol.

CONCLUSION

- We demonstrate that a supervised machine learning algorithm using readily clinically available EHR traits can accurately predict the HF subtypes of a broad and heterogeneous population of patients with HF in a large integrated health system.

FIGURES

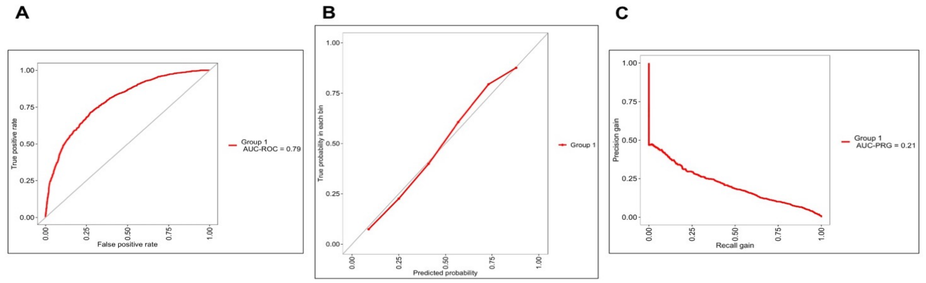

Figure 1. Performance characteristics of linear support vector machine on training dataset. **A.** Receiver Operating Characteristic Curve. **B.** Calibration curve. **C.** Precision gain-recall gain curve.

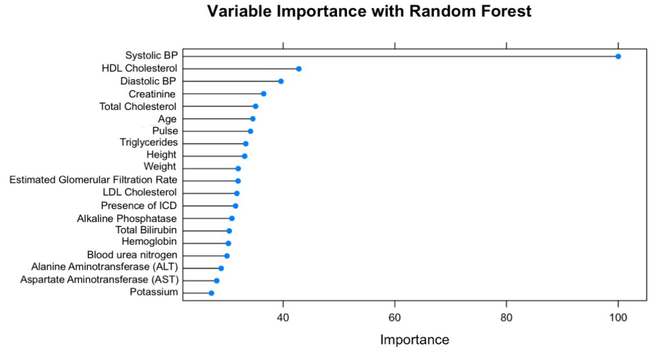

Figure: Plot of variable importance for the prediction of the heart failure subtype in random forest analysis

Abbreviations: BP = blood pressure; HDL = high-density lipoprotein; LDL = low-density lipoprotein; ICD = implantable cardioverter defibrillator

DISCLOSURES

N. Reza (@*noshreza*) is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number KL2TR001879. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Graphics courtesy of Flaticons, Freepik.