ARIC Manuscript Proposal # 3269

PC Reviewed: 11/13/18 **Status: _____** **Priority: ____**
SC Reviewed: _____ **Status: _____** **Priority: ____**

1.a. Full Title: Developing and Validating a Machine Learning Model to Predict Incident Heart Failure in African Americans: An Analysis from the Jackson Heart Study and The Atherosclerosis Risk in Community Cohort

b. Abbreviated Title (Length 26 characters): Machine learning based HF risk prediction

2. Writing Group: Ambarish Pandey, Matthew Segar, Kershaw Patel, Adolfo Correa (JHS investigator & sponsor), Michael Hall (JHS investigator), Carlos Rodriguez (ARIC investigator), Jarett Berry, James De Lemos, Javed Butler (JHS investigator), Justin Grodin, Vijay Nambi (ARIC investigator & sponsor), Christie Ballantyne (ARIC investigator and sponsor)

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _AP_ **[please confirm with your initials electronically or in writing]**

First author: Ambarish Pandey, MD
Address: 5323 Harry Hines Boulevard, Dallas, TX 75390
Phone: 214-645-7541
Fax: 214-645-2480
E-mail: ambarish.pandey@utsouthwestern.edu

Corresponding Author: Ambarish Pandey, MD
Address: 5323 Harry Hines Boulevard, Dallas, TX 75390
Phone: 214-645-7541
Fax: 214-645-2480
E-mail: ambarish.pandey@utsouthwestern.edu

ARIC authors to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).
Name: Salim Virani, MD PhD; Christie Ballantyne, MD
Address: 6565 Fannin Street, Suite B157, Houston, Texas 77030
Phone: 713-798-5800
E-mail: virani@bcm.edu; cmb@bcm.edu

3. Timeline: 01/2019-01/2020

4. Rationale:

Heart failure (HF) affects African-Americans (AA) at substantially higher rates than other ethnic group [1]. The race/ethnic disparities in HF incidence and outcomes are particularly notable among adults < 50 years [2,3]. Prior studies have utilized multivariable-adjusted Cox proportional hazard (CPH) models to evaluate the association between clinical phenotypes and HF incidence. However, CPH has numerous limitations including high variance and poor performance (Breiman), correction for multiple testing and handling of multicollinearity, and linear assumptions between factors [4-6]. Machine learning algorithms, however, can automatically reconstruct relationships between variables and response values from big data and can provide an efficient method of improving the performance of traditional proportional hazard models in identifying critical predictors[7]. Among them, Random Survival Forests (RSF) have shown increased interest for identifying important variables related to outcome without the need for *P* values[8]. Moreover, RSF do not impose a restrictive structure on how the variables should be combined [9]. An increasing number of studies have shown that many of the covariates were excluded from the CPH model analysis due to their violation of the proportional hazard assumption[10]. RSF were recommended as alternative methods for the study as it allows for consideration of more complex exposure/outcome relationships. Therefore, we aim to predict heart failure incidence using RSF in a large population-based cohort of AA using data from the Jackson heart study and then validate the risk prediction model in the non-Jackson cohort of AA in the ARIC study.

5. Main Hypothesis/Study Questions:

Aim 1: We aim to predict heart failure incidence using RSF in a large clinical cohort of AA. We hypothesize that advanced machine learning techniques, like RSF, are superior to the CPH method for determining variable selection and survival probability.

Aim 2: We aim to validate the risk prediction model developed in the Jackson Heart Study in the non-Jackson AA population of the ARIC cohort and compare the calibration and discrimination performance of the machine learning based model with that of well-established HF risk scores such as Health ABC risk score, ARIC clinical HF risk score, and ARIC biomarker HF risk score. We hypothesize that the machine learning based HF risk prediction model will have better discrimination and calibration performance than the other traditional Cox-PH based risk prediction models.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present)

Study design:

Prospective cohort study

Inclusion criteria:

- All participants

Exclusion criteria:

- Participants with existing HF at baseline or un-adjudicated HF events between baseline to 2005
- Patients with missing data on heart failure diagnosis or exacerbation

Primary predictor variables of interest:

- All variables including demographic characteristics (Age, sex, education, annual income), baseline medical history (history of diabetes, hypertension, anti-hypertensive use, smoking, history of MI, CVD, dialysis use, blood pressure medications, statin use), anthropometric measures (body weight, height, waist circumference), physical activity levels, lab data (LDL, HDL, A1c, Serum Creatinine/eGFR, BNP, troponin, CRP), and Echocardiographic parameters.

- These parameters of interest will be harmonized across the two cohorts. We have harmonized the data from JHS and ARIC for pooling in a separate project that we are involved with evaluating the impact of obesity parameters on pooled cohort equation performance in a multi cohort pooled analysis using JHS and ARIC data (PI: Ian Neeland/Rohan Khera at UT Southwestern, Dallas)

Outcomes:

Incident HF events

Data analysis:

- A predictive model will be developed to evaluate the association of multiple demographic, clinical, laboratory, and EKG/echocardiographic predictors and heart failure. The JHS dataset will be divided randomly into 80% training and 20% testing/validation. A random survival forest (RSF) model will be initially trained for predicting heart failure using a binned variable approach. Missing values will be imputed using the missForest R package. A random forest will be generated by creating 1000 trees.
- Variable selection will be calculated by two distinct methods – VIMP and Minimum Depth. Variable importance (VIMP) is calculated by first randomly permuting predictor variable values. VIMP is then defined as the difference between the prediction error of the observed and randomly permuted variables. A large VIMP suggests that misspecification worsens the predictive accuracy in the forest. Conversely, a low VIMP suggests noise is more informative than the observed variable. Therefore, we will ignore variables with negative or near zero VIMP values as they do not indicate that the predictive accuracy of the model is dependent on these variables. Second, minimum depth (MD) is calculated by recording and averaging the distance from the trunk of the tree (root node) across all trees in the forest. It is assumed that variables with high importance or impact on the prediction are nearest the root node. Therefore, lower values (ie, closer to the root of the tree) suggest higher importance in splitting the large group of patients and has a larger impact on the model prediction.
- As discussed previously, we will divide the predictor variables into three distinct hierarchical bins – Demographic/Clinical, Laboratory, and EKG/Echocardiographic. Starting with the Demographic/Clinical variables, we will use all the available associated visit 1 data and change in values between visit 1 and visit 2 to predict heart failure survival outcomes. Among all demographic and clinical variables, we will remove the predictors with a negative VIMP and those below the MD threshold with the remaining variables called Bin 1. We will add laboratory variables and repeat the process to obtain Bin 2, again removing variables with low importance. Finally, the process will be repeated a third time by adding echocardiographic data to Bin 2 to obtain Bin 3.
- The final predictive model will be retrained using the significant variables found in the prior analysis on all JHS data with the hyper-tuned parameters obtained in the testing/validation dataset. The trained model will be tested using data from the Atherosclerosis Risk in Communities (ARIC) study. For this, variables in the the ARIC and JHS study will be harmonized The participants from the Jackson county area that are included in both JHS and ARIC will be excluded from the validation cohort. The RSF model will be compared to the standard Cox proportional hazard model using the variables selected in the prior analysis, AIC Cox model with forward selection, and LASSO-Cox with top 20 RSF variables. Prediction accuracy will be assessed by calculating a Harrell C-index using out-of-bag (OOB) data. A total of 1000 OOB bootstrap samples from the original dataset will be used to compute a prediction model and calculate the C-index.
- The calibration and net-reclassification index will also be compared between the machine learning and Cox models in the ARIC cohort.
- The model will also be compared with the well-established Health ABC risk score [11] and ARIC HF risk score[12] with respect to the calibration, discrimination, and net-reclassification index in predicting HF events in the JHS and ARIC cohorts.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes _x___ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? ____ Yes ____ No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes _x___ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status.
ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

___x___ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

1. Association between lipoprotein(a) levels and cardiovascular outcomes in black and white subjects: the Atherosclerosis Risk in Communities (ARIC) Study.
2. Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density subfractions: the Atherosclerosis Risk in Communities (ARIC) Study.

We are including Dr. Virani and Dr. Ballantyne in our present project

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ___ Yes __x__ No

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number*_Carotid MRI Study)

B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __x__ No.

References:

[1] Husaini BA, Mensah GA, Sawyer D, Cain VA, Samad Z, Hull PC, et al. Race, Sex, and Age Differences in Heart Failure-Related Hospitalizations in a Southern State: Implications for Prevention. Circ Hear Fail 2011;4:161–169. doi:10.1161/CIRCHEARTFAILURE.110.958306.
[2] Yancy CW. Heart Failure in African Americans. Am J Cardiol 2005;96:3–12. doi:10.1016/j.amjcard.2005.07.028.
[3] Bibbins-Domingo K, Pletcher MJ, Lin F, Vittinghoff E, Gardin JM, Arynchyn A, et al. Racial Differences in Incident Heart Failure among Young Adults. N Engl J Med 2009;360:1179–1190. doi:10.1056/NEJMoa0807265.
[4] Breiman L. Heuristics of instability and stabilization in model selection. Ann Stat 1996;24:2350–2383. doi:10.1214/aos/1032181158.
[5] Dietrich S, Floegel A, Troll M, Kühn T, Rathmann W, Peters A, et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol 2016;45:1406–1420. doi:10.1093/ije/dyw145.
[6] Miao F, Cai Y-P, Zhang Y-X, Li Y, Zhang Y-T. Risk Prediction of One-Year Mortality in Patients with Cardiac Arrhythmias Using Random Survival Forest. Comput Math Methods Med 2015;2015:303250. doi:10.1155/2015/303250.
[7] Miao F, Cai Y, Zhang Y, Fan X, Li Y. Predictive Modeling of Hospital Mortality for Patients With Heart Failure by Using an Improved Random Survival Forest. IEEE Access 2018;6:7244–7253. doi:10.1109/ACCESS.2018.2789898.
[8] Rajeswaran J, Blackstone EH. Identifying risk factors: Challenges of separating signal from noise. J Thorac Cardiovasc Surg 2017;153:1136–1138. doi:10.1016/j.jtcvs.2017.01.010.
[9] Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. BMC Res Notes 2017;10:1–18. doi:10.1186/s13104-017-2775-6.
[10] Nasejje JB, Mwambi HG, Achia TNO. Understanding the determinants of under-five child mortality in Uganda including the estimation of unobserved household and community effects using both frequentist and Bayesian survival analysis approaches. BMC Public Health 2015;15:1003. doi:10.1186/s12889-015-2332-y.
[11] Butler J, Kalogeropoulos A, Georgiopoulou V, Belue R, Rodondi N, Garcia M, Bauer DC, Satterfield S, Smith AL, Vaccarino V, Newman AB, Harris TB, Wilson PW, Kritchevsky SB; Health ABC Study. Incident heart failure prediction in the elderly: the health ABC heart failure score. Circ Heart Fail. 2008 Jul;1(2):125-33
[12] Nambi V, Liu X, Chambless LE, de Lemos JA, Virani SS, Agarwal S, Boerwinkle E, Hoogeveen RC, Aguilar D, Astor BC, Srinivas PR, Deswal A, Mosley TH, Coresh J, Folsom AR, Heiss G, Ballantyne CM. Troponin T and N-terminal pro-B-type natriuretic peptide: a biomarker approach to predict heart failure risk--the atherosclerosis risk in communities study. Clin Chem. 2013 Dec;59(12):1802-10

Table 1: General characteristics of the African American participants in JHS and ARIC cohorts

Baseline Characteristics (harmonized for the two cohorts)	JHS cohort	ARIC cohort (excluding JHS participants)

Table 2: The top 20 ranked variables by the importance on the Random survival forest method for incident heart failure in JHS

Rank	Variable

Table 3: Performance C-statistic for Random survival forest model and Cox models for predicting heart failure in the derivation cohort and validation cohort.

Performance measure	Derivation cohort (JHS)	Validation cohort (ARIC)
RSF with top-20 covariates		
standard Cox proportional hazard model using the top 20 RSF variables		
AIC Cox model with forward selection		
LASSO-Cox with top 20 RSF variables		
AIC Cox backward selection model with top 20 RSF variables		

Table 4: Comparison of the O/E and C-statistic for the Random survival forest model vs. the previously published HF risk prediction models in the ARIC cohort

Risk Prediction model	C-statistic	O/E	NRI
RSF model from JHS			
Health ABC model			
ARIC HF risk prediction model (*Agarwal, et al Circ HF*)			
ARIC HF risk prediction model wth biomarkers (*Nambi et al. Clinical Chem*)			
AIC Cox backward selection model with top 20 RSF variables			