ARIC Manuscript Proposal #4273

PC Reviewed: 6/13/23	**Status: _____**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Integration of Clinical and Polygenic Risk Scores, Proteomic and ECG Data to Predict Atrial Fibrillation

b. Abbreviated Title (Length 26 characters): **Data integration and AF risk**

2. Writing Group:

Writing group members: Michael J. Zhang, Yuchen Yao, Wendy Wang, Zhong Zhuang, Ruoyu He, Yuekai Ji, Alex Knutson, Faye L. Norby, Alvaro Alonso, Elsayed Soliman, Weihong Tang, James Pankow, Wei Pan, Lin Yee Chen, others welcome

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

First author: Michael Zhang, MD, PhD
Address: Lillehei Heart Institute and Cardiovascular Division,
University of Minnesota Medical School,
420 Delaware Street SE, MMC 508,
Minneapolis, MN 55455.

Phone: 612-625-9100 Fax:
E-mail: mjzhang@umn.edu

ARIC author 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: **Lin Yee Chen, MD, MS**
Address: Lillehei Heart Institute and Cardiovascular Division,
University of Minnesota Medical School,
420 Delaware Street SE, MMC 508,
Minneapolis, MN 55455.

Phone: 612-625-4401 Fax: 612-624-4937
E-mail: chenx484@umn.edu

3. Timeline: Data analysis: 3 months
Manuscript preparation: 6 month(s)
Anticipated first draft: Fall 2023

4. Rationale:

Atrial fibrillation (AF) is the most common type of sustained cardiac arrhythmia and it is associated with substantial morbidity and mortality (1). Therefore, tools to predict the development of AF has substantial public health benefits. Many clinical risk scores (CRSs), such as the Framingham and CHARGE-AF scores, have been developed to predict risk of AF (2,3); however, their predictive performance have been moderate.

Recent advances in genome-wide association studies (GWASs) have made it possible to construct polygenic risk scores (PRSs) to predict the genetic risk of cardiovascular events, and to combine PRSs with CRSs to improve risk prediction (4-6). In a prior AF risk prediction study, the addition of a PRS to the CHARGE-AF CRS resulted in an increase in C-index of 0.05 (7). However, another study showed that the addition of a PRS to a CRS did not result in C-index increase (8). Protein biomarkers have also been added to CRSs to improve AF risk prediction: addition of NT-pro-BNP and FGF-23 to a CRS improved the C-index by 0.07 (9). Furthermore, electrocardiogram (ECG)-based models have also been added to CRSs to improve AF risk prediction: addition of a convolution neural network-trained ECG model to the CHARGE-AF score resulted in a C-index increase of 0.03 (10).

Despite the large number of studies examining AF risk prediction, few studies have integrated PRS, protein biomarkers and ECG-based models to CRSs and comprehensively assessed their prediction performance. Therefore, the objective of this study is to develop a PRS, protein biomarker model and ECG model to individually and collectively add to the CHARGE-AF CRS to improve AF risk prediction. We will then comprehensively assess all model combinations to identify the best combined model for optimal prediction of AF risk. We will also perform cross-sectional analyses for prevalent AF risk to assess for performance of our models to detect covert AF.

5. Main Hypothesis/Study Questions:

1) Aim 1: Evaluate four individual predictive models for AF: 1) CRS (CHARGE-AF); 2) PRS; 3) Protein biomarker (Somalogic V3 & V5); 4) ECG (pretrained AI).

2) Aim 2: Assess the performance of each combination of predictive model to predict AF risk.

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:

1. Prevalent AF – cross-sectional analysis at visit 3 and visit 5 (separately).
2. Incident AF – prospective observational analysis from visit 3 and visit 5 (separately) until 2019). Will consider follow up restriction to within 5 or 10 years.

Study population:

1) Inclusion criteria:
    a. CHARGE-AF CRS model: ARIC participants attending visit 3 or visit 5 with available covariates (CHARGE-AF clinical variables).
    b. PRS model: ARIC participants attending visit 3 or visit 5 with available genotype data.
    c. Protein biomarker model: ARIC participants attending visit 3 or visit 5 with available Somalogic plasma proteomics data.
    d. ECG model: ARIC participants attending visit 3 or visit 5 with available 12-lead ECG.
    e. Combined model: ARIC participants attending visit 3 or visit 5 with available genotype, plasma proteomics, ECG signals and covariates.
2) Exclusion criteria:
    a. Participants with prevalent HF, race other than Black or White, Black participants in Minneapolis or Washington County field centers.
    b. For incident AF analyses, we will exclude participants with prevalent AF.

Variables:

1) CHARGE-AF Clinical Variables
    a. We will fit regression models for prevalent AF or incident AF by 11 clinical variables in CHARGE-AF for 12,887 participants (visit 3) and 6,538 participants (visit 5). The variables are age, race, height, weight, systolic blood pressure, diastolic blood pressure, current smoking, antihypertensive medication use, diabetes, prevalent heart failure and prevalent myocardial infarction.
2) PRS
    a. We will use a published PRS model with its weights for 1,091,491 SNPs (11) to calculate a PRS for the ARIC data. The weights were calculated using method PRS-CS (12) based on an AF GWAS dataset of over 300,000 Finnish participants in FinnGen Biobank (13) and the 1000 Genomes Project European samples (14) as the reference panel. We will use the genetic data of 10,348 participants (visit 3) and 5,247 participants (visit 5) in ARIC.
3) Protein biomarkers
    a. We will utilize Somalogic-quantified plasma protein levels (approximately 5000 proteins) for 11,471 participants (visit 3) and 5,193 participants (visit 5). To predict AF based on the ARIC proteomics, we will develop regression models with a lasso penalty as well as a residual neural network. Sure Independence Screening (SIS) will be applied before lasso and the neural network to select predictors marginally (15). We then will apply 5-fold cross validation to tune hyper-parameters (with 10 candidate values). For SIS, we will select $\lfloor n/\ log\ n \rfloor$ proteins, where n is the sample size in the training data.
4) ECG

a. We will utilize 12-lead ECG data from 12,730 participants (visit 3) and 5,946 participants (visit 5). A pre-trained convolutional neural network (CNN) model will be applied to predict AF based on the ECG data and the architecture of this CNN model is based on an inception neural network in an incident AF prediction paper (16). The inception network model involves 4 inception blocks and each block concatenates results from 3 CNN layers with different filter numbers.

Outcome: Prevalent AF and Incident AF (until 2019), ascertained by hospitalization discharge codes, study visits and ambulatory patch-based rhythm monitoring.

Statistical analysis:

1) Prevalent and incident AF
   a. We will assess, individually and collectively, the PRS, protein biomarker, and ECG-enhanced AF risk prediction models by assessing discrimination (Harrell's C statistic and its 95% CI) and calibration (Hosmer-Lemeshow $\chi 2$ statistic).

2) Sensitivity analyses
   a. Survival bias –we will consider inverse probability weighting and adjusting for the competing risk of death.
   b. Since renal function can affect plasma proteins, we will explore adjusting for estimated glomerular filtration rate in the protein biomarker model
   c. Prediction model optimism – we will correct for prediction model optimism by performing internal cross-validation (18). Depending on analysis results and data availability, we will consider external cohort validation.

3) Limitations:
   a. Underascertainment of incident AF – we will report AF incidence rates and compare this with other studies examining AF risk to ensure our rates of ascertained AF are clinically relevant.
   b. We are using a PRS derived from European ancestry participants in both Blacks and Whites. We will also assess the performance of the PRS stratified by race to examine the applicability of the PRS to Black participants.

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? ____ Yes __X__ No

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit" ? ____ Yes ____ No
(The 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 current derived consent file ICTDER05 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/mantrack/maintain/search/dtSearch.html

______ Yes ___X____ 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)?

#3398 – Proteomics and the Risk of Incident Atrial Fibrillation in the Elderly – Faye Norby
#3573 – Validation of an artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm in the Atherosclerosis Risk in Communities (ARIC) Study – Peter Noseworthy, Lin Yee Chen
#3905 – Short- and long-term changes in proteomic levels and the risk of incident atrial fibrillation in older adults – Jeffrey Misialek, Pamela Lutsey

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

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number* _2014.18, 2017.14, 2017.27_______)

___ 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 https://sites.cscc.unc.edu/aric/approved-ancillary-studies

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.

References

1. Lip GYaT, Hung-Fat. Management of atrial fibrillation. The Lancet 2007;370:604--618.
2. Alonso AaK, Bouwe P and Aspelund, Thor and Stepas, Katherine A and Pencina, Michael J and Moser, Carlee B and Sinner, Moritz F and Sotoodehnia, Nona and Fontes, Jo{\~a}o D and Janssens, A Cecile J W and Kronmal, Richard A and Magnani, Jare. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. J Am Heart Assoc 2013;2:e000102.
3. Chamberlain AMaA, Sunil K and Folsom, Aaron R and Soliman, Elsayed Z and Chambless, Lloyd E and Crow, Richard and Ambrose, Marietta and Alonso, Alvaro. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). Am J Cardiol 2011;107:85-91.
4. Inoyue M AG, Nelson C, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention 2018;72:1883–93.
5. Khera AV CM, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify participants with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219–24.
6. Phulka JS, Ashraf M, Bajwa BK, Pare G, Laksman Z. Current State and Future of Polygenic Risk Scores in Cardiometabolic Disease: A Scoping Review. Circ Genom Precis Med 2023:e003834.
7. Marston NA, Garfinkel AC, Kamanu FK et al. A polygenic risk score predicts atrial fibrillation in cardiovascular disease. Eur Heart J 2023;44:221-231.
8. Tada H, Shiffman D, Smith JG et al. Twelve-single nucleotide polymorphism genetic risk score identifies participants at increased risk for future atrial fibrillation and stroke. Stroke 2014;45:2856-2862.
9. Lind LaS, Johan and Stenemo, Markus and Hagstrom, Emil and Arnlov, Johan. Discovery of new biomarkers for atrial fibrillation using a custom-made proteomics chip. Heart 2017;103:377-382.
10. Khurshid S, Friedman S, Reeder C et al. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. Circulation 2022;145:122-133.
11. Ripatti NMaJVLaPdBPaEWaJKaAPaS. Systematic comparison of family history and polygenic risk across 24 common diseases. The American Journal of Human Genetics 2022.
12. Ge T, Chen, CY., Ni, Y. et al. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 2019;10.
13. Kurki MIaK, Juha and Palta, Priit and Sipil{\"a}, Timo P. and Kristiansson, Kati and Donner, Kati and Reeve, Mary P. and Laivuori, Hannele and Aavikko, Mervi and Kaunisto, Mari A. and Loukola, Anu and Lahtela, Elisa and et al. FinnGen: Unique genetic insights from combining isolated population and national health register data. medRxiv 2022.
14. Consortium TGP. A global reference for human genetic variation. Nature 2015:68–74.
15. Lv JFaJ. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2008;70:849–911.
16. Raghunath SaP, John M. and Ulloa-Cerna, Alvaro E. and Nemani, Arun and Carbonati, Tanner and Jing, Linyuan and vanMaanen, David P. and McCarty, Bern E. and Hartzel, Dustin N. and et al. Deep Neural Networks can Predict Incident Atrial Fibrillation from the 12-lead Electrocardiogram and may help Prevent Associated Strokes. medRxiv 2020.
17. Armitage P, Colton T. Encyclopedia of biostatistics. 2nd ed. Chichester, England: John Wiley & Sons, 2005.
18. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774-81.