**ARIC Manuscript Proposal #3678**

| | | |
|---|---|---|
| **PC Reviewed: 8/11/20** | **Status: _____** | **Priority: 2** |
| **SC Reviewed: _____** | **Status: _____** | **Priority: ____** |

**1.a. Full Title**: Electrocardiographic Artificial Intelligence Model for Prediction of Heart Failure

**b. Abbreviated Title (Length 26 characters)**: ECG-Based AI HF Model

**2. Writing Group**:

Writing group members: Oguz Akbilgic, Ibrahim Karabayir, Liam Butler, Patricia Chang, Dalane Kitzman, Alvaro Alonso, Lin Y Chen, Elsayed Z Soliman

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _OA_


**First author**: **Oguz Akbilgic, PhD**

Address: Department of Health Informatics and Data Science
Parkinson School of Health Sciences and Public Health
Loyola University Chicago
2160 S 1st Street, CTRE #127, Maywood, IL, 60153
Phone: (708) 216-8971 Fax: (708) 216-8216
E-mail: oakbilgic@luc.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: **Elsayed Z Soliman**, MD, MSc, MS

Address: Epidemiological Cardiology Research Center (EPICARE)
Wake Forest School of Medicine
Medical Center Blvd, Winston-Salem, NC 27157
Phone: (336)716-8632 Fax: (336) 716-0834
Email: esoliman@wakehealth.edu


**3. Timeline**:

Start analyses: within two weeks of receiving data

Submission for publication: Within 3 months of the approval of the proposal.

**4. Rationale**:

There are approximately 6.5 million adults in the United States suffering from Heart failure (HF) [1]. HF contributes up to 12.5% of all cause deaths and its annual cost to healthcare system was estimated as $30.7 billion in 2012 [2]. Early diagnosis and treatment can significantly improve HF prognosis [3], and subsequently help reducing the health and economic burdens of HF.

We previously applied a novel probabilistic symbol pattern recognition approach [4] to identify HF patients using R-R intervals from electrocardiogram (ECG). [5]. Also, in several cohort studies including ARIC, we have shown that there are various ECG markers that are associated with incident HF [6-15]. These findings suggest that applying machine/deep learning approaches on features obtained from ECG can be used in developing automated HF prediction tools for early recognition of patients at risk.

The aim of this proposed study is to develop artificial intelligence (AI) based models to predict the risk for HF by utilizing ARIC raw digital ECG data. The ARIC study with its high-quality digital ECG data and corresponding HF outcomes represents a unique opportunity to answer and address this aim.

**5. Main Hypothesis/Study Questions**: This study aims to:

- To develop AI-based models to predict the risk for HF.
- Compare the prediction performance of our AI-based model to the previously developed ARIC HF Risk Score [16].
- To examine the consistency of these predictive models in sex and race subgroups.

**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).**

**Sample Size**

Inclusion criteria: All ARIC participants with good quality ECG data at baseline as well as information on all relevant risk factors and HF events during study's long-term follow-up will be eligible for inclusion in this analysis.

Exclusion criteria: Participants with race/ethnicity other than black and white, with prevalent HF at baseline visit, missing information on HF during follow-up or with missing/poor-quality ECGs.

**Variables:**

**Outcomes:** Heart failure events anytime during follow up.

**Other Variables**

- Raw digital ECG data available from visits 1, 2, 3, and 4 which will be obtained from the ECG Reading Center (Soliman).
- Key demographic and clinical variables - age, race, gender, body mass index, education, smoking status, hypertension, diabetes mellitus, history of coronary heart disease, HDL cholesterol, LDL cholesterol, total triglycerides, total cholesterol, systolic blood pressure, diastolic blood pressure, fasting blood glucose, use of blood pressure lowering medications, aspirin use, use of lipid lowering medications and creatinine.

**Data analysis:**

**Cohort Characteristics**: Frequency distributions of participant characteristics will be reported.

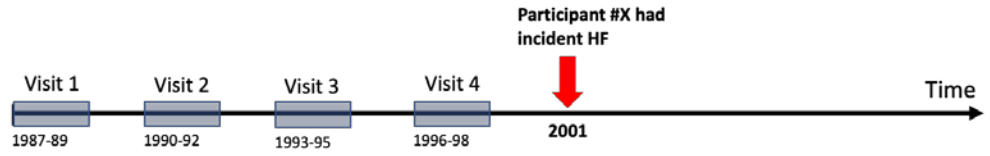
**Outcome Variable**: Outcome variable is incident HF defined as the first occurrence after baseline (visit 1). The outcome variable will be expressed in both binary (Approach 1) and time to event (Approach 2) fashion (Figure 1). Binary outcome will take the value of 1 for participants who experienced HF and 0 for others. The time-to-event variable will represent the time between the incident of HF and the preceding ECG screening time

**Input Variables**: Model inputs will be collected from the visits as described in Figure 1. These will include raw ECG data as well as key demographics and clinical variables. Controls will be matched by age, gender, race, and visit time among participants who did not experience HF.

**Modeling**: We will consider both traditional machine learning via feature engineering and deep learning directly on raw ECG data.

**Feature Engineering**: We will implement signal processing methods on the raw ECG data. These methods will include Sample Entropy, Probabilistic Symbolic Pattern Recognition, Fast and Continuous Fourier Transformation, and Wavelet Transformations. The extracted features will be combined with other clinical data to be used in model building. In Approach 1, we will consider various machine learning and statistical modeling approaches including logistic regression, gradient boosting, random forest, and support vector machines. We will also implement a sensitivity analysis by substituting ECG features from earlier visits to identify how early we can efficiently predict HF. In Approach 2, we will use Cox-Proportional Hazards Regression with Time-Varying Covariates. Therefore, the changes in ECGs over time will be considered in both modeling approach.

**Deep Learning**: This method will be used in Approach 1 only. We will build a convolutional neural network model by feeding the model with raw digital ECG signals to classify the binary HF outcome. We will design a cascaded architecture allowing non-ECG clinical variables into the model once the latest convolutional layer processing raw ECG data is flattened into a dense layer. We will add a normalization layer following this dense layer to handle features in varying units and size.




**Approach 1-Binary Classification**: ECG from the latest visit that is also at least one year prior to the first incident HF will be used as an input in model building. In sensitivity analysis, the earlier ECGs will then be substituted in the final model to identify the effective prediction window.

**Example**: For Participant #X, the ECG from Visit 4 will be used in model building. In sensitivity analysis, the ECGs from Visits 3, 2, and 1 will be substituted in the model to see when the prediction performance becomes poorer.

**Approach 2-Survival Analysis**: ECGs from all visits preceding to the first incident HF will be used as an input in model building with Cox Proportional Hazards Regression with Time-Varying Covariates. Features extracted from these ECGs will represent the time-varying covariates.

**Example**: For Participant #X, we will extract features from the ECG of Visits 1,2,3, and 4. These features will be used as time varying covariates in Survival Analysis.

***Figure 1: Modeling Approaches with an Example of a Participant with HF Event***

**Validation**: We will implement a comprehensive cross-validation strategy. We will first split entire cohort into 80% model building and 20% hold out datasets. Using the 80% model building dataset, we will implement a 10-fold cross validation. We then will create an ensemble of ten models obtained over 10-fold cross-validation as a final model to test on holdout dataset. Alternatively, once the generalization is confirmed on 10-fold cross validation, we will build a new model using entire 80% model building using learned model parameters and test this final model on the hold out dataset. We will evaluate the model performance and compare different models using Area Under the Receiver Operating Characteristic Curve (AUC) statistics.

**Comparison with ARIC HF Risk Score**: The models built in both Approach 1 and 2 will be compared to previously develop ARIC HF Risk score [16] based on AUC statistics.

**Subgroup Analysis**: We will implement a subgroup analysis on the final model for sex and race variables to identify whether the model better or worse represents a certain subgroup.

---

**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/mantrack/maintain/search/dtSearch.html

__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)?**

**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* _____)**

**___ 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 https://www2.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] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," (in eng), *Circulation,* vol. 139, no. 10, pp. e56-e528, Mar 5 2019, doi: 10.1161/cir.0000000000000659.

[2] S. C. Curtin, "Trends in Cancer and Heart Disease Death Rates Among Adults Aged 45-64: United States, 1999-2017," (in eng), *Natl Vital Stat Rep,* vol. 68, no. 5, pp. 1-9, May 2019.

[3] C. f. D. C. a. Prevention. "Heart Failure." https://www.cdc.gov/heartdisease/heart_failure.htm (accessed 06/16/2020.

[4] O. Akbilgic and J. A. Howe, "Symbolic pattern recognition for sequential data," *Sequential Analysis,* vol. 36, no. 4, pp. 528-540, 2017.

[5] R. Mahajan, T. Viangteeravat, and O. Akbilgic, "Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics," *Int J Med Inform,* vol. 108, pp. 55-63, Dec 2017, doi: 10.1016/j.ijmedinf.2017.09.006.

[6] P. M. Rautaharju *et al.*, "Electrocardiographic predictors of incident heart failure in men and women free from manifest cardiovascular disease (from the Atherosclerosis Risk in Communities [ARIC] study)," (in eng), *Am J Cardiol,* vol. 112, no. 6, pp. 843-9, Sep 15 2013, doi: 10.1016/j.amjcard.2013.05.011.

[7] Z. M. Zhang *et al.*, "Different patterns of bundle-branch blocks and the risk of incident heart failure in the Women's Health Initiative (WHI) study," (in eng), *Circ Heart Fail,* vol. 6, no. 4, pp. 655-61, Jul 2013, doi: 10.1161/circheartfailure.113.000217.

[8] Z. M. Zhang, P. M. Rautaharju, R. J. Prineas, L. Loehr, W. Rosamond, and E. Z. Soliman, "Usefulness of electrocardiographic QRS/T angles with versus without bundle branch blocks to predict heart failure (from the Atherosclerosis Risk in Communities Study)," (in eng), *Am J Cardiol,* vol. 114, no. 3, pp. 412-8, Aug 1 2014, doi: 10.1016/j.amjcard.2014.05.011.

[9] Z. M. Zhang, P. M. Rautaharju, R. J. Prineas, L. Loehr, W. Rosamond, and E. Z. Soliman, "Ventricular conduction defects and the risk of incident heart failure in the Atherosclerosis Risk in Communities (ARIC) Study," (in eng), *J Card Fail,* vol. 21, no. 4, pp. 307-12, Apr 2015, doi: 10.1016/j.cardfail.2015.01.001.

[10] M. F. Almahmoud, W. T. O'Neal, W. Qureshi, and E. Z. Soliman, "Electrocardiographic Versus Echocardiographic Left Ventricular Hypertrophy in Prediction of Congestive Heart Failure in the Elderly," (in eng), *Clin Cardiol,* vol. 38, no. 6, pp. 365-70, Jun 2015, doi: 10.1002/clc.22402.

[11] W. T. O'Neal *et al.*, "Electrocardiographic Predictors of Heart Failure With Reduced Versus Preserved Ejection Fraction: The Multi-Ethnic Study of Atherosclerosis," (in eng), *J Am Heart Assoc,* vol. 6, no. 6, May 25 2017, doi: 10.1161/jaha.117.006023.

[12] W. T. O'Neal, P. B. Sandesara, A. Samman-Tahhan, H. M. Kelli, M. Hammadah, and E. Z. Soliman, "Heart rate and the risk of adverse outcomes in patients with heart failure with preserved ejection fraction," (in eng), *Eur J Prev Cardiol,* vol. 24, no. 11, pp. 1212-1219, Jul 2017, doi: 10.1177/2047487317708676.

[13] W. T. Qureshi *et al.*, "Silent Myocardial Infarction and Long-Term Risk of Heart Failure: The ARIC Study," (in eng), *J Am Coll Cardiol,* vol. 71, no. 1, pp. 1-8, Jan 2 2018, doi: 10.1016/j.jacc.2017.10.071.

[14] L. Ilkhanoff *et al.*, "Association of QRS duration with left ventricular structure and function and risk of heart failure in middle-aged and older adults: the Multi-Ethnic Study of Atherosclerosis (MESA)," (in eng), *Eur J Heart Fail,* vol. 14, no. 11, pp. 1285-92, Nov 2012, doi: 10.1093/eurjhf/hfs112.

[15] F. Gunturkun *et al.*, "Deep learning for improved prediction of late-onset cardiomyopathy among childhood cancer survivors: A report from the St. Jude Lifetime Cohort (SJLIFE)," *Journal of Clinical Oncology,* vol. 38, no. 15_suppl, pp. 10545-10545, 2020, doi: 10.1200/JCO.2020.38.15_suppl.10545.

[16] S. K. Agarwal *et al.*, "Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study," (in eng), *Circ Heart Fail,* vol. 5, no. 4, pp. 422-9, Jul 1 2012, doi: 10.1161/circheartfailure.111.964841.