**ARIC Manuscript Proposal # 3077**

PC Reviewed: 11/14/17  Status: _____  Priority: 2

SC Reviewed: _________  Status: _____  Priority: ____

1.a. **Full Title**: Machine Learning for Sudden Cardiac Death (SCD) Prediction: the Atherosclerosis Risk in Communities (ARIC) Study

b. **Abbreviated Title (Length 26 characters)**: SCD Prediction

2. **Writing Group**:
   Writing group members:
   Shannon Wongvibulsin, Kunihiro Matsushita, Elizabeth Selvin, Wayne D. Rosamond, Nona Sotoodehnia, Scott Zeger, others welcome

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

**First author**: Shannon Wongvibulsin  
Address: 615 N. Wolfe Street  
Room E3650  
Baltimore, MD 21205  
Phone: 949-229-1183  
Fax: 410-955-0094  
E-mail: swongvi1@jhmi.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: Kunihiro Matsushita  
Address: Johns Hopkins Bloomberg School of Public Health  
2024 E. Monument St., Suite 2-600 (Rm 2-602), Baltimore, MD 21287  
Phone: 443-287-8766  
Fax: (410) 367-2384  
E-mail: kmatsush@jhsph.edu

3. **Timeline**:
We aim to submit the manuscript to the ARIC publications committee within 12 months of the approval date for this proposal.

4. **Rationale**:
Although sudden cardiac death (SCD) is the leading cause of death in the industrialized world,\(^1\)-\(^4\) effective clinical support tools for identifying individuals at high SCD risk remain lacking. As a result, SCD is a major public health issue which needs to be addressed at both the population and individual levels.
In the United States, each year, there are approximately 400,000 SCDs. These SCDs can be classified broadly into two main categories: 1. healthy individuals with no heart disease, and 2. individuals with heart disease and mild/severe cardiac dysfunction. Approximately 50% of victims do not have a prior diagnosis of heart disease and hence have limited opportunities for prevention. However, even in the group of individuals with heart disease, SCD risk assessment using left ventricular ejection fraction (LVEF) as the sole or primary factor to guide clinical decisions about defibrillator therapy fails to identify high risk individuals who would benefit from implantable cardioverter defibrillator (ICD) implantation. The National Heart, Lung, and Blood Institute (NHLBI) Working Group on SCD Prevention stated: “there is an urgent need to develop effective preventive strategies” of which effective SCD risk assessment can be one important component.

The explosive growth of biomedical and information technologies is an opportunity to substantially improve SCD risk assessment. Methods for statistical prediction and causal inference are increasing our capacity to learn from data. To harness the power of data to benefit patients, effective decision support tools and individualized interventions can be designed, tested, and implemented.

While SCD prediction models exist, current models do not handle the dynamic dependencies among SCDs, heart failure (HF), myocardial infarction (MI), and other time-varying factors. Because dynamic clinical events, such as HF, can substantially impact an individual’s SCD risk (e.g. clinical heart failure results in a fivefold increase in SCD risk), we propose the development of a machine learning random forest model that uses not only baseline predictors but also time-dependent variables for dynamic risk prediction. We will refer to the data used for this analysis as survival, longitudinal, and multivariate (SLAM) data. While there are examples of machine learning methods applied to SLAM data, they have not been widely used in medicine to date. More commonly used methods for SCD risk prediction, Cox proportional hazards model and logistic regression, do not necessarily discover and include non-log-linear relationships or interactions among multiple predictors. Novel machine learning methods that naturally incorporate non-linear/interaction effects and the dependencies among SCD, HF, MI, and survival (or death) could result in individualized SCD prediction. Furthermore, the creation of open-source software for the methods we develop has the potential to impact SCD risk assessment as well as multiple other areas of clinical and translational research.

5. Main Hypothesis/Study Questions:
With Random Forest for Survival, Longitudinal, and Multivariate Data (RF-SLAM) applied to ARIC, we hypothesize that we will be able to estimate an individual’s SCD risk as well as the expected benefit from risk-reduction strategies based upon population-level data. Specifically, we will answer the following question with the methods outlined below: with what accuracy and precision can we predict SCD events from time-varying predictors as well as baseline covariates?

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 with Visit 1 as baseline (1987-1989)
Discrete time survival analysis with time-varying covariates
Joint discrete time survival analysis of SCD and with time-varying covariates

Inclusion criteria
All ARIC patients who participated in Visit 1. Patients with deaths of unknown timing of cardiac symptoms will be excluded from the study. Participants with unknown cardiovascular disease history will also be excluded.

Outcome
We will consider SCD as the main outcome of interest over a median of 21.2 years of follow-up. In our analysis, SCD is defined as death attributable to definite or possible cardiac cause, in which death occurred within 1 hour after the onset of acute symptoms.

Predictors
1. Sociodemographic variables: age, gender, race, center
2. Anthropometric variables: body mass index, height, waist circumference
3. Diabetes mellitus
4. Atrial fibrillation
5. Systolic and diastolic blood pressure
6. Left ventricular hypertrophy by Cornell criteria
7. Biochemical laboratory variables: blood lipids (e.g. LDL, HDL, triglycerides), fasting blood glucose, creatinine, sodium, potassium, magnesium, CRP, white blood cell count, albumin, fibrinogen, NT-proBNP, BUN, CK, CK-MB, cardiac troponin I, myoglobin, hematocrit
8. Medication: antihypertensive, antiarrhythmic, lipid-lowering, or antidiabetic medications
9. Lifestyle: smoking, diet (e.g. total energy intake, saturated fat, caffeine, alcohol), physical activity (intensity and frequency)
10. Electrocardiogram parameters (e.g. Q-wave with or without ST-T changes, ST junction & segment depression, T wave, conduction defect from adjudicated ECG data)
11. Markers of carotid atherosclerosis (e.g. carotid intima-media thickness)
12. Time of hospitalization for cardiovascular events and procedures (i.e. MI, HF, revascularization)

Statistical analysis
We develop Random Forest for Survival, Longitudinal, and Multivariate Data (RF-SLAM) for predicting SCD risk and determining the relative contribution of each risk factor to the individual’s overall SCD risk. To evaluate our method, we will compare our results to those generated by generalized linear models (GLMs) and random forest (RF) survival analysis tools for time-independent variables.

In developing RF-SLAM, we will build upon the open-source RF statistical software available through the Comprehensive R Archive Network (CRAN) where possible. RF is an ensemble learning method based on a collection of decision trees, where the overall RF prediction is the ensemble average or majority vote. RF has been successfully applied in classification,
regression, and survival analysis in many fields including medicine.\textsuperscript{8,16,17} RF exhibits improved prediction performance over single decision tree methods by addressing the issue called “overfitting” whereby the tree is grown to have too many branches, resulting in overfitting the data used to construct the tree and poor generalizability for predictions on new observations. Overfitting is avoided through the introduction of random selection of subjects and of predictor variables during the construction of trees in the random forest. Randomly sampling subjects with replacement (“bootstrapping”) results in approximately $\frac{2}{3}$ of the subjects, called “in-bag”, being selected for fitting a given tree and the remaining $\frac{1}{3}$, “out-of-bag” (OOB) used to obtain an unbiased assessment of prediction error similar to the left-out data in cross-validation. Random sampling of predictor variables at each decision tree node decreases the correlation among the trees in the forest generated by bootstrapping, and thereby improves the precision of the ensemble predictions.\textsuperscript{8,18}

Our approach with RF-SLAM involves partitioning the multiple events and other information for each individual into a set of what we will refer to as “Counting Process Information Units” (CPIUs). Specifically, each CPIU contains the following data for a prespecified bin of time: person indicator, interval indicator, multivariate outcome values (e.g. SCD, HF, and death; 0 meaning event did not occur, 1 meaning event did occur), summary function values of outcome history, predictor values, and the length of the interval. Our proposed method is to use the general RF approach to predict the multivariate outcome indicators within each small unit of time using the predictor values, outcome history values, and interval time variables.

\textit{Outcome distribution specification and splitting criteria:} Because the outcome is multivariate, there are at least three characteristics of its distribution that can be modeled: the joint distribution, the marginal distribution of a selected event, and the conditional distribution of one event given the others. In this project, we will focus on the estimation of the joint distribution and conditional distribution of SCD given the other events. For each approach, we will compare the results for different splitting criteria (i.e. chi-squared test statistic, likelihood ratio test statistic, or weighted Gini index) using the simulation studies described below.

\textit{Control of bootstrapping:} Because we create CPIUs, each individual will have multiple observation intervals rather than only one as in a traditional RF analysis. Rather than bootstrapping CPIUs, we will bootstrap individuals, then assemble together the predictions for each of the CPIUs for an individual to obtain the discrete-time, piecewise constant hazard function for each bootstrap replication. This function, or the corresponding piecewise exponential survival function, will be the basis for visualization of the risk trajectory and post-hoc analyses of how changes in different predictor variables impact an individual’s risk.

\textit{Incorporation of covariates that capture the history of outcomes:} The history of prior outcomes for an individual can serve as important information for the prediction of future events. We will include multiple history variables including time since the previous events of each type (SCD, HF, MI), and total number of prior events of each type.

\textit{Estimation of individual hazard/survival functions:} Our RF algorithm will generate an event risk for each CPIU for each OOB bootstrap replication. We will construct the complete piecewise-constant hazard function for each individual over the study period by assembling the estimated
risks from the terminal nodes to which the CPIUs were classified. The final hazard and survival functions for a person are the ensemble averages over the OOB bootstrap replications.

**Strengths and Limitations**
This study proposes to investigate the use of a novel statistical/machine learning approach for improved prediction of SCD, improved visualization of SCD risk, and quantifying the relative contributions of relevant risk factors. The results of this study will provide valuable information regarding SCD prediction and prevention. Additionally, the open-source software for SCD risk prediction and visualization has the potential to impact SCD risk assessment and prevention as well as multiple other areas of clinical and translational research. A limitation is the low number of SCD cases (544 definite/probable events in 15,792 participants) available in the dataset.

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)?
No previous proposal in ARIC focuses specifically on assessing SCD risk through statistical/machine learning approaches. ARIC Manuscript Proposal #1655 (Sudden Cardiac Death Prediction from Multimodal Data) explores the development of statistical tools (in particular, a decision-tree approach) to analyze multimodal data and classify patients according to their risk of either SCD or overall death. On the other hand, our proposal investigates the use of a novel statistical/machine learning approach, random forest for survival, longitudinal, and multivariate data (RF-SLAM), for predicting SCD, visualizing SCD risk, and quantifying the relative contributions of relevant risk factors. Our proposal aims to further the understanding of
SCD risk and also move beyond prediction and lay the foundation for SCD prevention through the identification of modifiable SCD risk factors.

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
___X___ A. primarily the result of an ancillary study (list number* 2013.07)
 ___  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.cscu.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.cscu.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:


