1.a. **Full Title**: Classification trees for MI identification in community surveillance: Findings from the ARIC study

b. **Abbreviated Title (Length 26 characters)**: SL CHD events

2. **Writing Group**:
   Writing group members:

Matthew Shane Loop, Noah Simon, Eric Whitsel, Brittany Bogle, Gerardo Heiss, and Wayne Rosamond

Others welcome

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

**First author**: Matthew Loop  
Address: 123 West Franklin Street, Suite 450  
CB#8030  
Chapel Hill, NC 27516  

Phone: 919.962.3225  
Fax:  
E-mail: matthew_loop@unc.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: Wayne Rosamond  
Address:

Phone:  
Fax:  
E-mail: wayne_rosamond@unc.edu

3. **Timeline**: Manuscript to be submitted by December 31, 2018
4. **Rationale:**

Surveillance of community events is key to understanding the burden of myocardial infarction (MI). However, classification of hospitalizations for possible MI is resource-intensive, both from human medical record abstraction and expert clinician classification of hospitalizations. The NHLBI ended community surveillance in ARIC in 2014, partially due to the high cost of implementation.

The ARIC study has long-implemented a computer diagnosis of MI, which is given to clinician reviewers during the review process to aid in conducting the review, as well as to reduce the overall number of human reviews by automatically classifying a portion of the hospitalizations. This algorithm is deterministic and was designed based upon subject-matter knowledge. The human reviewers disagree with the computer diagnosis with a non-trivial frequency. If the number of reviews conducted by human reviewers could be reduced, then efficiencies and cost savings in conducting community surveillance could potentially make it more viable to future funding agencies.

Therefore, we propose to evaluate a statistical learning (SL) approach to classify ARIC community surveillance hospitalizations for CHD, compared with the gold standard of expert clinician review.

5. **Main Hypothesis/Study Questions:**

1. Can a classification tree model adequately classify hospitalizations for possible MI compared to a human reviewer?
2. What are the characteristics of events for which the classification tree model fails to agree with the human reviewer?
3. Can a classification tree model adequately predict when a reviewer will disagree with the original ARIC computer diagnosis?

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:** Retrospective analysis of community surveillance

**Inclusion/exclusion:**

**Inclusions:**
- Community surveillance hospitalizations from 1997 – 2014 that had at least one human classification for possible MI (i.e., in-hospital events)
  - We are excluding events from the pre-troponin error, as we are not interested in building a prediction model using types of data that are no longer used to classify an MI.

**Exclusions:**
• Out of hospital deaths

Outcome: Binary version of final MMCC classification of the event (definite/probable MI vs. suspect/no MI; study question 1); whether the MMCC reviewer disagreed with the original ARIC computer MI diagnosis (study question 2)

Other variables of interest

The key variables to be considered to predict final human MI classification will be structured variables on the event summary form (ESF; i.e., not the text of the discharge summary) that the human reviewer uses to classify the event.

Specifically, these elements are:
- Reason for review (hospital diagnosis, death diagnosis)
- Computer MI diagnosis (cardiac pain classification, ECG classification, enzyme classification)
- Special procedures (coronary reperfusion within 24 hours of onset, time from onset to reperfusion attempt, CPR and/or cardioversion)
- History (MI, angina)
- Demographics (age, sex)
- Hospital discharge index(ices)
- Hospital discharge diagnosis(es)
- Underlying cause of death
- Date of event, admission, and discharge/death
- Presence of trauma
- Was enzyme downgraded
- Enzyme levels

Summary of data analysis

Study question 1

Prediction models will be trained on hospitalizations from 1987 – 2013, with hospitalizations in 2014 held out as the test set. Given the potential for non-linearity and non-additivity in human classification of hospitalized events, we will use boosted classification tree models to predict the MI classification. Classification trees in general are known to have higher variance, but lower bias, than other SL methods. The technique of “boosting” reduces the variance of classification trees at the cost of some bias and loss of interpretability, in order to obtain reductions in classification error rates in test data.

Boosted classification trees require two parameters to be chosen a priori: (1) number of trees; and (2) tree depth. We will fix the number of trees at k = 500, but we will use cross validation to evaluate tree depths from 1 to 5 in the training data. After choosing the tree depth that minimizes classification error on the training data, we will re-fit the boosted classification tree model on the
entire training data to create a fitted classification model and training classification error. We will then use that fitted classification model on the test data set (i.e., event year 2014) and estimate the test classification error.

Study question 2

For events in the test dataset from study question 1, we will compare summary statistics for the variables used in the boosted regression trees in events that were classified correctly compared to events that were not classified correctly.

Study question 3

We will build a prediction model in a manner almost identical to study question 1, except that the outcome will be whether the human reviewer disagreed with the original ARIC computer MI diagnosis. This study question will potentially help to clarify the factors that are involved in the clinical judgment of human reviewers that are not considered in the original ARIC computer MI diagnosis.

Methodologic challenges

The key methodological challenge is appropriate cross-validation procedures during model training, given the longitudinal nature/correlation in the data. We often want to be careful about using future data to predict past data. Therefore, we will use standard cross-validation methods for time series that select the first $m$ observations for training and the remaining $n-m$ observations for validation. The group of $m$ observations is progressively increased until most of the observations are used to train the model and the remaining few future observations are used for validation. The classification errors are then averaged across the $k$ iterations of cross-validation to estimate the overall validation error for a particular tree depth value. Dr. Simon and Dr. Loop will pay special attention to the cross-validation procedure, in order to ensure that the training data is not overfit, resulting in poorer test data performance.

Missing data is potentially another methodological challenge. We have excluded fatal events to minimize the amount of missing data, since fatal events sometimes are missing important variables for classification (e.g., enzyme levels). However, during exploratory data analysis we will carefully describe the missing data pattern and evaluate whether some variables should be left out of consideration, or whether an additional category for “missing” should be created for some categorical variables. Missing data procedures for classification tree methods are still an open area of research, and unfortunately sophisticated procedures are not readily available.

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)?

MP2734 – “Machine learning classification algorithms for myocardial infarction in ARIC”

This manuscript proposal overlaps to some extent. However, after discussions between the first author (Brittany Bogle), Wayne Rosamond, Gerardo Heiss, and myself, we all agreed that it would be appropriate for me to go ahead with my own proposal on this topic. The current proposal is significantly narrower in scope, and Dr. Bogle will participate in the current proposal as a co-author.

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 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.