ARIC Manuscript Proposal # 3226

PC Reviewed: 08/14/18  Status: _____  Priority: 2
SC Reviewed: __________  Status: _____  Priority: _____

1.a. Full Title: Clinical Risk Prediction for MCI and Dementia: the Atherosclerosis Risk in Communities (ARIC) Study

b. Abbreviated Title (Length 26 characters): MCI/Dementia Risk Prediction

2. Writing Group:
   Michael E. Griswold, Rajesh Talluri, B. Gwen Windham, Radhikesh Ranadive, Chad Blackshear, Rola Abi Saleh, Myriam Fornage, Eric Boerwinkle, Jeannette Simino, Alden Gross, Melinda Power, Rebecca Gottesman, Ramon Casanova, Dave Knopman, Cliff Jack, Tom Mosley

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

First author: Michael Griswold, PhD
Address: Department of Data Science
         JD Bower School of Population Health
         University of Mississippi Medical Center
         2500 North State St.
         Jackson, MS  39216-4505
         Phone: 601-984-4933
         Email: mgriswold@umc.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: Thomas H. Mosley, Jr.
Address: University of Mississippi Medical Center
         Department of Medicine-WW202
         2500 N. State
         Jackson, MS 39216-4505
         Phone: 601-984-2763  Fax: 601-815-3422
         E-mail: tmosley@umc.edu

3. Timeline:
4. **Rationale:**

The number of people living with dementia worldwide is currently estimated at 47 million, is expected to increase to 75 million by 2030 and is projected to nearly triple by 2050. Early prediction of dementia and mild cognitive impairment (MCI, dementia’s precursor) could provide improved access to preventative care and support systems that help patients maintain additional independent life-years. While early and accurate prognosis of MCI and dementia is challenging, several risk models have been proposed using a variety of data elements including (1) standard clinical markers (e.g. age, sex, race, education, BMI, blood pressure, diabetes, etc.), (2) SocioDemographic\Environmental markers (3) imaging markers (e.g. brain atrophy, cortical thickness, infarct presence, etc.), (4) biomarkers (e.g. CSF concentrations of amyloid-β1-42 (Aβ1-42), tau protein, plasma Abeta, inflammation, etc.), (5) ‘omics markers (e.g. genomic: APOE4, p97/VCP; metabolomic, etc.), (6) historical cognitive testing markers, and (7) combinations of these and other markers over time. Initial studies that train and implement prediction models commonly report area under the curve (AUC) utility statistics of around 0.7 to 0.8, but in the few validation/replication studies conducted, performance is generally worse (AUCs around 0.5 and 0.6) and replication is often lacking altogether.

Standard risk score development and validation protocols have been well described (e.g. equator network TRIPOD guidelines and others), with a commonly adopted approach being (1) construct existing risk scores within a new data source and establish predictive utility overall and within important subgroups, (2) identify new markers not used in the existing risk scores and include them to construct a new risk score, (3) establish and report improved performance of the new risk scores overall and within subgroups (using independent validation data if available or cross-validation/bootstrapping procedures if not). Identification of new markers in this process often occurs in a standard epidemiologic setting with experts *a priori* positing a limited set of additional markers that are then tested for augmented utility through regression/selection models. Newer machine learning and deep learning algorithms can better handle high-dimensional datasets, have potential for enhancing standard epidemiologic risk prediction approaches by expanding the usual *a priori* specified limited additional marker approach to inform potential candidate markers for consideration as risk factors. In addition these approaches commonly employ cross-validation/bootstrapping to set algorithm tuning parameters and use out-of-sample prediction to report validation statistics.

After establishing the predictive utility of a new risk score, estimating its cost-efficacy in implementation is of great practical importance. Phased risk models that operate under a sequential decision framework (3,4) can develop thresholds for interventions and additional measurements in programs specifically designed to reduce disease while justifying the costs of the additional testing. Phased approaches often include an initial screen (Phase 1) using a basic risk prediction based on commonly available measures which are easy to collect (age, gender, education, blood pressure, etc.) and a triage of all
patients from phase 1 into strata of low-risk (on whom no additional action is currently taken) and med/high-risk (who are sent for additional measurements). In Phase 2, medium-cost markers are measured (basic biomarkers, cognitive exams) only on the phase 1 med/high risk group, and the additional information is incorporated into another risk stratification of low, medium and high risk levels. Medium and high risk patients from phase 2 can be sent for additional higher-cost testing (MRI, PET, ‘omics, etc.) and/or offered intervention/support services with patient choice and cost considerations in play. Phased clinical risk frameworks such as this can be tuned to optimize a variety of metrics including sensitivity, specificity, both sensitivity/specificity, costs, and other valuations. Bayesian decision theory has been used for phased risk models and shown to improve clinical decision support systems that facilitate joint clinician-patient decision making.

Our goal is to develop a novel MCI and dementia prediction framework for clinical settings using ARIC data based on three underlying components: (1) improving utilities of existing dementia risk models (CAIDE, BDSI, ABIDE/adaapt) overall and within sex-race-age specific subgroups by using additional a priori -specified risk factor sets; (2) comparing utilities of these updated risk models against machine learning algorithms that dramatically expand the set of potential features to include all available ARIC variables; (3) building clinical decision support tools that empower clinician-patient discussions on best next steps for MCI/Dementia concerns where costs and benefits are incorporated using Bayesian phased risk prediction frameworks and emphasis is placed on modifiable risk factors.

5. Main Hypothesis/Study Questions:

Our overarching goal is to propagate a new clinical decision support system for long-term dementia risk. Central to this goal are the following three aims/questions:

1) Epidemiologic risk prediction (Substantive Epidemiologic Paper):
   a. How do the CAIDE, BDSI and ABIDE risk scores perform in the ARIC study, both overall and within sex-race-age subgroups?
   b. Does the inclusion of additional specifically selected variables improve predictive performance?
      i. Potential set 1: Cardio/Cerebro-vascular risk factors
         1. Waist circumference
         2. Inflammation (CRP)
         3. Atrial fibrillation
         4. Stroke
         5. Heart disease
      ii. Potential set 2: Previous cognition measures
         1. Earlier global cognition scores
            a. V2->V4 for V5 MCI/Dementia outcomes
            b. V2->V5 for V6 MCI/Dementia outcomes
         2. Earlier individual instruments from the ARIC cognitive batteries
iii. Potential set 3: MRI measures
   1. Brain volume
   2. Brain infarct burden
   3. Brain WMH

iv. Potential set 4: genetics
   1. ApoE

c. Does subgroup-specific modeling improve performance?
   i. Sex-race groups
   ii. Normal to MCI, Normal to Dementia, MCI to Dementia

2) Data Science risk prediction (Data Science / Biostatistics Methods Paper):
   a. As generally applied in the area, machine learning models are black boxes that predict dementia. These approaches are less concerned with the etiologic relationships between outcomes and predictors than they are with predictive accuracy. In what situations will data science approaches (e.g. machine learning) that incorporate all available variables improve prediction performance over the standard epidemiologic risk prediction approach taken in 1) above?
      i. We will use approaches that do not require a priori specification of specific variables in the model. We will include all appropriate variables distributed by the ARIC Coordinating Center, including:
         1. over 3,000 variables at visit 2
         2. over 1,700 variables at visit 3
         3. over 2,000 variables at visit 5
   b. Does having more detailed information about the longitudinal progression to dementia (Normal-> MCI-> Dementia) increase the predictive performance of dementia risk models for individuals. Here we will include cognitive features across timepoints.
   c. How do complex machine learning models optimally contribute to better understanding of disease processes such as the ways that individual risk factors may interact to lead to dementia?

3) Decision Theory: (Clinical Algorithm Paper)
   a. Can staged/phased risk prediction approaches, with increasingly sophisticated and costly measures included at each progressive stage, provide better performance while simultaneously balancing cost-effectiveness? Staged decision systems generally start with collecting the lowest cost / lowest burdensome risk factors in Stage 1 (e.g. age, sex, race, education, bmi/waist circumference, etc.) and then stratify the population into groups at low-risk, high-risk and a third group where more information is needed (indeterminate) to better determine next steps. Low and high risk groups generally have some pre-determined clinical action (intervene/treat or do not) while the indeterminate group is pursued in further stages for additional, higher cost / higher burden, more informative data to make a better clinical
recommendations. Staged systems balance the competing goals of minimizing costs while maximizing the chances of correct diagnoses. We will examine sensitivities/specificities and associated costs of predicting Normal/MCI/Dementia outcomes with a staged approach of:

i. Stage 1: Simple Cardio/Cerebro-vascular risk factors generally available at a clinical exam
   1. age, sex, race, education, bmi/waist circumference, previous stroke, etc.

ii. Stage 2: Additional Cardio/Cerebro-vascular risk factors that may need to be additionally ordered at a clinical exam
   1. Inflammation (CRP), Atrial fibrillation (ECG), etc.

iii. Stage 3: Cognitive measures (additional clinical cost & burden)

iv. Stage 4: MRI

v. Stage 5: Genetics (ApoE)

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).

**Design:**
Cross-temporal and Longitudinal/Survival study design with follow-up through visit 6.

**Risk Sets:**
- V2 risk factors: CAIDE, BDSI, ABIDE risk factors + potential sets 1,2,4 from above
- V3 risk factors: CAIDE, BDSI, ABIDE risk factors + potential sets 1-4 from above
- V5 risk factors: CAIDE, BDSI, ABIDE risk factors + potential sets 1-4 from above

**Outcome sets:**
- V2, V3, V5 risk factor associations with:
  - Immediate analyses
    - Time to dementia (basic survival model approach)
    - Time to dementia vs death (competing risk survival model approach)
    - V5 Normal/MCI/Dementia (3-level cross-temporal multinomial approach)
    - V5 Normal/MCI/Dementia/Death (4-level cross-temporal approach)
  - When data become available
    - V6 Normal/MCI/Dementia (3-level cross-temporal multinomial approach)
    - V6 Normal/MCI/Dementia/Death (4-level cross-temporal approach)
    - V5->V6 progression

**Datasets:**
1. ARIC visit 2 derived variables.
2. ARIC visit 3 MRI data (Jackson and Forsyth County).
3. ARIC visit 3 derived variables.
4. ARIC visit 5 derived variables.
5. ARIC visit 5 and visit 6 cognitive status (normal, MCI, dementia)
6. ARIC incident dementia variable
7. ARIC death status variable
8. APOE (WGS if available)

**Analyses:**

**Epidemiologic risk prediction:** Analyses of incident dementia outcomes (time to dementia), will be performed using survival model techniques (e.g. Cox, Weibull, etc.) and Fine & Gray competing risk approaches when appropriate. Dementia status (3-level and 4-level outcomes) and progression will use multinomial and ordinal (where supported) logistic regression models. Attrition effects of death will be explicitly modelled while loss to follow up (LTFU) effects will be examined using inverse proportional weighting (IPW) and joint-modelling approaches. V5 MRI subpopulations will be up-weighted using coordinating center defined v5 stage 3 sampling weights.

**Data Science risk prediction:** Analyses will focus on incorporating all available variables collected at the specified visits (V2: 3046 features curated and available; V3: 1787 features curated and available; V5: 2185 features curated and available). Assuming that all variables have simple relationships with dementia is naïve, and we hope that many of these features being included will have complex relationships with dementia that may not otherwise be detected using logistic regression. Logistic regression techniques will serve as standard anchoring comparison models. Against the anchor we will compare tree-based methods (random forests, xgboost) and deep-learning approaches with tuning parameters determined using cross-validation. Tree-based methods can pick up interactions that could not be modeled using logistic regression. Deep-learning (e.g. feed-forward neural networks, convolutional neural networks, restricted Boltzmann machines) is a recent trend in machine learning that can model highly non-linear data representations.

**Decision Theory:** We will formulate a multi-stage utility-based analysis of decisions for subsequent additional measurements following a prognosis, where the expected utility is maximized over all possible paths (combinations of: predictions -> decisions on new measurements -> updated predictions) in the decision tree. Others have successfully extended these models to incorporate threshold optimization into the decision process as well as including costs, benefits and predictive accuracy into the utility function.

**Comparisons:** We will use standard metrics such as area under the ROC curve (AUC / generalized c-statistic), calibration, sensitivities and specificities for specific thresholds, positive/negative predictive values, positive/negative likelihood ratios and risk reclassification tables and indices (e.g. NRI, IDI), and cost-effectiveness curves for comparisons.

**Validation:** All analyses will use k-fold cross-validation techniques to avoid overfitting and additional external validation will be performed for available risk factor sets on non-ARIC participants from the GENOA study.
Limitations/Challenges
A primary limitation of this study (as with any risk score study) is the reproducibility and
generalizability of the risk models to other populations. Internal validation improves overfitting
issues but does not address basic potential differences across populations with different
characteristics. Including GENOA as an external validation dataset will help, but approaching
additional cohorts to provide further validation datasets may prove useful as well.

References:
(2) Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable
prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD
(3) Clemen, R. T. (1996), Making Hard Decisions: An Introduction to Decision Analysis,
Duxbury Press, 2nd ed.

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? _X__ Yes _____ 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”? _X___ 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)?

- #3118 - Comparison of existing methods for algorithmic classification of dementia status (Power)
- #3054 - Brain Structural MRI Abnormalities Predict Dementia, MCI and Cognitive Decline in an Older Population (Wu)
- #2337 - DNA methylation-derived age predicts changes in brain morphology and cognitive decline (Fornage)
- We are aware that an ancillary study proposal looking at computational techniques for AD prediction has been submitted by one of our co-authors (Ramon Casanova). Ramon has agreed to be a co-author on our proposal here. His ancillary study is complementary but does not fully overlap with our second aim.

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
   _X_  A. primarily the result of an ancillary study (list number* _2008-06_________)
   ___   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.cscce.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.cscce.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.