1.a. **Full Title:** Missing Cognitive Function Imputation Using Multiple Imputation by Chained Equations: ARIC Visit 6 Neural Cognitive Study

b. **Abbreviated Title (Length 26 characters):** Cognition Score Imputation

2. **Writing Group:**
Writing group members:
Aozhou Wu, MHS; and (alphabetically); Karen Bandeen-Roche, PhD; Josef Coresh, MD, PhD; David Couper, PhD; Jennifer Deal, PhD; Michael Griswold, PhD; Alden L. Gross, PhD, MHS; Rebecca F. Gottesman, MD, PhD; Thomas Mosley, PhD; Melinda C Power, ScD; Andreea Rawlings, PhD; A. Richey Sharrett, MD, DrPH; others welcome.

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

**First author:**
Aozhou Wu, MHS
Address: Department of Epidemiology
        Johns Hopkins Bloomberg School of Public Health
        2024 E. Monument Street, B-319
        Phone: (410)736-2558       Fax:
        E-mail: awu32@jhu.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: Josef Coresh, MD, PhD
Address: Department of Epidemiology
         Johns Hopkins Bloomberg School of Public Health
         2024 E. Monument Street, Room 2-635
         Phone: (410)955-0495       Fax:
         E-mail: coresh@jhu.edu
3. Timeline:
Most of the data to be used in this proposal are already available. Analyses and manuscript preparation will be performed with available data in the next few months and completed shortly after the final version of cognition evaluation results (e.g. diagnosed dementia) become available (approximately June 2018).

4. Rationale:
Cohort attrition is a common problem in epidemiological studies, particularly in older populations. Missing data due to cohort attrition is informative when we study attrition-related outcomes. This is particularly problematic in research studying associations between vascular risk factors and dementia or cognitive impairment. Participants with poorer cognitive function are more likely to end up lost to follow-up. The same is true for those with poor cardiovascular profiles.

In the ARIC study, a comprehensive dementia surveillance system has been deployed and implemented. The system contains two parts: one is active dementia ascertainment which was initiated from Visit 5. We actively collected information on participants’ cognitive ability from the participants themselves or their proxies through phone interviews beginning from Visit 5 and continuing during annual follow-up throughout Visit 6. Additionally, a passive ascertainment system identified suspect dementia from hospitalization records and death certificates prior to Visit 6. The surveillance system captured many dementia cases who missed ARIC study visits. It also provided an additional source of cognition information that could be used for imputing cognitive scores for participants who either missed the study visits or did not attempt the cognitive battery at the visits. Since 2015, the dementia surveillance system enhanced the active ascertainment of dementia by annually evaluating participants’ cognitive function through telephone interview using screening tools: Six-Item Screener (SIS) which is derived from mini-mental state exam (MMSE), and AD8 which is derived from CDR. This information will be useful for imputing cognitive scores for those lost to follow-up between Visit 5 and 6, but was not available prior to Visit 5.

Previously, ARIC investigators have used several approaches to address the missing cognitive function when studying cognitive decline, including inverse probability attrition weighting, a Shared-Parameter Model, and multiple imputation. Multiple imputation is a powerful tool for imputing missing data. Compared to single imputation (e.g. mean imputation), multiple imputations provide higher accuracy and variability of imputed values. In a paper published by Dr. Andreea Rawlings, multiple imputation by chained equation (MICE) was used to impute missing cognitive scores prior to Visit 5 and used as an illustration of its effectiveness in the association of diabetes with cognitive decline. Authors concluded that MICE is an effective approach for imputing missing cognition score when there are available informative data. Later ARIC studies have adapted this approach to account for cohort attrition.

In this study, we will apply the same imputation framework as described in Dr. Rawlings’ work but with some modifications to accommodate differences in available predictor data, including the availability of an extended V5 cognitive battery and active dementia surveillance using the SIS and AD8. The detailed description of the prior analysis framework can be found in the published paper and ARIC manuscript proposal #2523. One major change from the previous work is that instead of using the information from telephone interview for cognitive status (TIC) and CDR which were available only at one time point close to Visit 5, we will use the annually measured SIS and AD8 instruments as two of the key imputation variables. Validation studies will be used to evaluate the performance of this new MICE imputation approach. Specifically, we will 1) randomly censor participants’ cognitive scores at Visit 6 assuming the missing mechanism of Missing Completely at Random; 2) censor participants’ cognitive scores based on their attrition probability estimated from covariates assuming a missing mechanism of Missing at
Random; and 3) censor participants’ cognitive scores at Visit 6 in the ARIC-PET subsample who had a cognitive battery administered between Visit 5 and Visit 6 which had not been used in our imputation. Then we will compare the imputed cognitive function score with the observed score in each of these conditions. We will also assess the robustness of the imputation results by conducting a simulation analysis. We expect an improvement in the imputation results for Visit 6 cognitive score with the proposed simulation method compared to the previous work for Visit 5 cognitive score imputation. One reason is that the missing mechanism of Visit 6 scores conditioned on Visit 5 scores is more like Missing at Random, rather than Missing Not at Random which indicates a missing pattern depends on the level of outcome – Visit 6 scores. Because of the short time-interval between Visits 5 and 6, Visit 5 factors score along is predictive for Visit 6 factor score (correlation = 0.88). Since informative attrition is a big concern in cognition study in ARIC. This work could potentially benefit the later studies focusing cognitive decline post Visit 5 and long term cognitive decline from Visit 2 to Visit 6.

We will illustrate the new imputation approach using the association of Visit 4 smoking status with post Visit 5 cognitive decline. Previously, smoking has been shown to be strongly associated with both cognitive impairment \(^7\) and cohort attrition.\(^8\) In an exploratory analysis, we confirmed that smoking is associated with post Visit 5 cohort attrition. This suggested that the smoking-cognitive decline association could be biased if the attrition is not appropriately taken into account. We will apply similar smoking definition as in Dr. Jennifer Deal’s previous work (ARIC MP #2262). Specifically, we will define smoking status primarily as never smoking, former smoker, and current smoker and also use a pack-years measurement to quantify the cumulative smoking dose. We may also explore the potential effect of quitting. As an extension of this project, we may also consider conducting imputation for missing cognitive measurements (both factor scores and DWRT, DSST, and WFT combined Z scores\(^9\)) from Visit 2 to Visit 6 by combining previous Visit 5 imputation approach and the newly developed approach. To achieve a comparability with previous work, we may also consider an addition illustration example – looking at the diabetes status at Visit 2 with cognitive decline trajectory from Visit 2 to Visit 6. For simplicity purpose, this rest parts of this proposal will mainly focus on the illustration example evaluating the Visit 4 smoking status with cognitive decline from Visit 5 to Visit 6 using the imputation approach we proposed.

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

   **Study Aims:**
   
   1) Develop a new imputation model for Visit 6 cognitive scores based on multiple imputation by chained equations method.
   
   2) Validate the performance of the imputation method in terms of the bias in imputed scores, particularly in the subgroup of interest, *i.e.* participant with suspect dementia but missed ARIC Visit 6.
   
   3) Evaluate the association of smoking at Visit 4 with post-Visit 5 cognitive decline with the imputed data.

   **Hypotheses:**
   
   a) The imputed scores by MICE for participants who developed dementia and missed Visit 6 (cognitive battery) will have a similar or lower mean value compared to the observed scores for participants who developed dementia and had factor scores measured at Visit 6.
   
   b) The imputed score by MICE will be unbiased when compared with the observed score in all validation samples.
   
   c) The estimated 5-year decline using MICE will show a larger absolute value when using data with imputation compared to using the original observed data. Comparing to reference exposure category, *i.e.* never smoking, the relative cognitive decline in other exposure groups (current/former smoker) will increase when using the data with imputation.
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 Scope Summary**
Outcome for imputation: missing V6 cognitive factor scores
Illustration exposure: smoking at Visit 4
Illustration outcome: cognitive decline from Visit 5 to Visit 6.
Study population of the illustration example: participants who had both smoking measurement at Visit 4 and factor scores at Visit 5.

**Approach Summary**
We will first impute missing cognitive factor score at Visit 6 using factor score at Visit 5; other variables related to cognition at Visit 6 measured at Visit 6 and between the two visits; and additional relevant covariates. Then, we will examine the association of smoking measured at Visit 4 with cognitive decline from Visit 5 to Visit 6 using the imputed data.

**Study population**
Inclusion Criteria:
- 6,433 ARIC participants who attended ARIC Visit 5 and had cognitive factor score.
Exclusion Criteria:
For the imputation, we will exclude participants:
- Neither white nor African-American and non-white in Washington Co. and Minnesota.
- Missing key covariate information (e.g. education).
- Missing cognitive factor scores at Visit 5 (for simplicity purpose)
- Missing smoking measurement (for simplicity purpose)
For analyzing the association of smoking and cognitive decline, we will further exclude participants:
- Who have dementia at or before Visit 5

**Exposure**
The primary exposure is smoking status at Visit 4 categorized into three groups: never smoking, current smoker, and former smoker, based on self-report smoking status. Cumulative measurements of smoking prior to Visit 5, e.g. pack-years, will be also assessed.

**Outcome**
We will use factor scores at Visit 5 and Visit 6. The factor scores were derived based on a previously published method, and were standardized to the mean and standard deviation of Visit 5 factor scores. The Visit 6 factor score is the outcome we are imputing.

**Imputation Methods**
The imputation method is multiple imputation by chained equations using Stata command “mi impute chained”. The key imputation variables include: MMSE score at Visit 6; surveillance score derived from SIS and AD8; suspect dementia prior to Visit 6 identified by dementia surveillance system; health information collected through annual follow-up interview (e.g. poor health, prevalent CHD, prevalent stroke, hospitalization count, and use of proxy); indicator for death by Visit 6; Visit 5 factor scores; time
between surveillance scores and Visit 5 factor scores; time from Visit 5 to Visit 6; and major dementia risk factors at Visit 4 (measured at the same time as the exposure – smoking).
To evaluate the robustness of our imputation approach, we will apply variation on key imputation steps, e.g. different approaches to utilize SIS and AD8 information; including/excluding deceased participants in the MICE imputation model.

**Modeling Cognitive Decline**
In the primary analysis, we will estimate the 5-year cognitive decline by taking the difference of the factor scores at Visit 5 and Visit 6 (or the imputed score if Visit 6 scores were missing) and then divided by the time difference between Visit 5 and 6 (or the time point for imputation). Because we only have two time points of cognitive function measurements, estimating the decline by taking the difference of cognitive scores measured at two time points should be adequate. We will evaluate the association between smoking and cognitive decline using regression approach with estimated 5-year cognitive decline as the dependent variable and smoking and other covariates as independent variables.
In the sensitivity analysis, we will estimate the cognitive decline by utilizing generalized estimating equation model (GEE) and mixed effect model to characterize the cognitive decline trajectories. GEE and mixed effect model have been previously used to estimate pre-Visit 5 cognitive decline. Model specification and other details can be found in corresponding papers. Briefly, models will be adjusted for age, sex, black race, education level, alcohol consumption, body mass index, hypertension, history of coronary heart disease, history of stroke, and apolipoprotein E ε4 genotype for both intercept and timeslope. The estimated effect of smoking associated 5-year cognitive decline will be compared using the data with vs. without imputation.

**Imputation Validation**
**Validation 1:** We will perform qualitative validation of the distribution of imputed scores by looking for any unreasonable distribution characteristics, e.g. the imputed scores among suspect dementia who missed Visit 6 have a higher mean value than the observed Visit 6 factor scores among suspect dementia cases.

**Validation 2:** We will randomly select 20% participants and set their scores to missing if they had visit 6 factor scores. These participants will be counted as our validation sample. We will then impute their Visit 6 factor scores and compare the imputed scores with the observed score. We should expect no systematic bias between the imputed scores and the observed scores in the overall validation sample and within subgroup defined by key covariates, e.g. exposure category, education level, and dementia status, in the validation sample.

**Validation 3:** We will select 20% participants based on their probability of attrition (estimated from observed covariates at Visit 4 and 5) and set their scores to missing if they had visit 6 factor scores. These participants will be counted as our validation sample for Validation 3. Then a similar validation approach as described in “Validation 2” will be applied here.

**Validation 4:** ARIC PET study, which enrolled about 300 participants, conducted a full cognitive battery between Visit 5 and 6 (call it PET Visit). We will utilize this resource for validation. Similar to aforementioned validation approaches, without using ARIC PET participants’ Visit 6 and PET Visit factor scores, we will impute their cognitive function at the PET Visit. We will then compare the imputed scores with the observed scores. Agreement for the subset of PET participants who missed visit 6 will be of special interest.

**Simulation Study:**
*Cognitive score simulation step:* We will simulate participants cognitive scores at Visit 6 using estimated cognitive decline slope obtained from the observed data (fitted with a mixed effect model), V5 factor scores, dementia status, dementia risk factors, other covariates, and a random component for decline slope.
as well as an error term.

Attrition simulation step: We will censor participants based on their probability of attrition estimated from their covariates to mimic the observed attrition in the ARIC cohort.

Imputation step: The aforementioned imputation approach will be applied to the simulated data. We will compare the estimated decline slope from the imputed data to the true slope we used for the simulation. To assess the robustness of our imputation approach in terms of 1) the potential underestimation of cognitive decline slope in participants who developed dementia but missed the Visit 6 or dead before Visit 6 from the observed data and 2) the degree that missingness is depend on our outcome variable (Visit 6 cognitive score), we will 1) vary the cognitive decline slope for participants who have suspect dementia or dead in the cognitive score simulation step; 2) vary the weight of Visit 6 cognitive scores in estimating attrition probability in attrition simulation step, respectively.

7.a. Will the data be used for non-CVD analysis in this manuscript? __ X __ Yes _____ 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? __X__ 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)?

This proposal uses a similar analysis framework as the previous published work conducted by Dr. Andreea Rawlings (MP #2523: Imputing missing outcome data using multiple imputation by chained equations: simulation and validation in the ARIC study) for Visit 5 cognitive function imputation. The differences are: 1) we are imputing Visit 6 missing cognition function; 2) We are using addition sources of cognition measurement which are different from Dr. Rawling’s approach.

Other related MPs include:
MSP#2115: Sensitivity Analyses with Shared-Parameter Models for studying Cognitive Change in the presence of potentially Informative Dropout – the Atherosclerosis Risk in Communities (ARIC) Neurocognitive Study

Neurocognitive Study
MSP#2160: Diabetes and cognitive change over 20 years: the Atherosclerosis Risk in Communities Study
MSP#2382: Examining the Healthy Cohort Effect: Predictors of Attrition in the Atherosclerosis Risk in Communities (ARIC) Study
MSP#1982: Estimation of cognitive change from repeat measures in observational studies; associations with education: the ARIC NCS
MSP #672: Changes in cognitive test scores in the ARIC cohort over a 6-year period (visit 2 to visit 4) and their correlation with vascular risk factors
MSP #2262: Cigarette smoking in midlife and subsequent 23-year cognitive decline: The Atherosclerosis Risk in Communities Study

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __X__ Yes _____ No  
ARIC NCS (Visit 5 and Visit 6) and ARIC-PET

11.b. If yes, is the proposal
   __X__  A. primarily the result of an ancillary study (list number* __2008.06_and 
2009.29__)  
   ____ 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.

13. Per Data Use Agreement Addendum for the Use of Linked ARIC CMS Data, approved manuscripts using linked ARIC 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.

Reference


