ARIC Manuscript Proposal #2523

PC Reviewed: 4/15/15	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Imputing missing outcome data using multiple imputation by chained equations: simulation and validation in the ARIC study

b. Abbreviated Title (Length 26 characters): MICE simulation, validation

2. Writing Group:

Writing group members: Andreea M. Rawlings, Yingying Sang, Michael Griswold, A. Richey Sharrett, Joe Coresh, Lisa M. Wruck, Jennifer A. Deal, Melinda C. Power, Karen Bandeen-Roche

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

First author: Andreea M. Rawlings
Address: Welch Center
2024 E. Monument St., STE 2-600
Baltimore, MD 21287

Phone: 443 287 4169 Fax: 410 955 0476
E-mail: arawlin2@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: A. Richey Sharrett
Address: Dept. Epidemiology
Johns Hopkins Bloomberg School of Public Health
615 N Wolfe St
Baltimore MD 21205
Phone: 443 287 6178 Fax: 410 955 0863
E-mail: rsharret@jhsph.edu

3. Timeline: Analysis is currently ongoing; plan to submit to ARIC Publications Committee within 3 months.

4. Rationale:

Missing data is a common problem in epidemiologic studies. It is of particular concern in longitudinal studies because participants who drop out over time are likely different from those who attend study visits. Analyses of longitudinal data rely on assumptions about the

mechanisms that gave rise to the missing data, and inappropriate assumptions or analyses may bias estimated parameters, standard errors, or both.

Missingness mechanisms can be classified into three classes using the framework developed by Little and Rubin[1,2]. In this framework, data are missing completely at random (MCAR) when the probability of missing a study visit does not depend on either observed or unobserved data. Data are missing at random (MAR) when the probability of missing, after conditioning on observed data, does not depend on unobserved data. Finally, when the probability of missing depends on unobserved data, even after conditioning on observed data, the data are missing not at random (MNAR). If data are MCAR, then a complete case analysis will yield unbiased associations between risk factors and the outcome. However, an MCAR missingness mechanism is extremely unlikely in practice, and a complete case analyses in the setting of MAR or MNAR may result in biased estimates if the data are not modeled appropriately.

Multiple imputation is a common method of dealing with missing data that depends on the assumption that data are MAR; however, use of imputation for the outcome, or dependent, variable remains limited.[3] This is because if there are sparse data for participants who do not attend study visits (ie there is no between visit follow-up or surveillance available for participants), then imputing the outcome only adds noise[4,5]. However ARIC has available a wealth of data on participants who do not attend visits, data that provide valuable information about cognitive function, namely the score from the informant's clinical dementia rating (CDR), hospital and death certificate dementia codes, and the telephone interview for cognitive status. This information can be used to impute cognitive scores for persons who do not attend study visits. When this auxiliary information exists on persons who do not attend visits, then multiple imputation is more efficient than methods such as inverse probability of attrition weighting, which uses only information from participants who attend study visits[6].

Specific to ARIC and cognitive decline, analyses of the association between risk factors measured at baseline (visit 2) and cognitive change from baseline are biased if participants who do not return for follow-up visits have worse cognitive function, and if the dropout is differential by the risk factors of interest. In this paper we will explore imputation of missing cognitive scores (global Z) using multiple imputation by chained equations (MICE), and will provide practical guidance to researchers regarding validation. We will validate the imputation in both living and deceased participants using existing and simulated data, and test performance of the MICE under several missingness assumptions.

We will use diabetes as the exposure of interest for illustrative purposes. While the main findings between diabetes and 20-year decline have been published[7], this research aims to explore several issues not addressed in the first publication.

First, at the time of our publication on diabetes and cognitive decline, we did not have CDR data available and used IPAW to account for attrition at visits 4 and 5. In IPAW, persons who remain in the study can be weighted so they represent the original cohort.

This assumes that there is no group of participants (given exposure and covariates) that has a zero probability of attending a given visit (positivity assumption). A potential issue that arises is that only approximately 3% of participants suspected of having dementia attended visit 5, and those 3% may have less severe disease. As a result, IPAW may underestimate the true level of cognitive decline.

Second, the imputation of scores for participants who died is debated. In this paper we do not differentiate between participants who are alive at visit 5 and those who have died at visit 5, although the date of imputation of these participants will differ. These results can provide insight into the appropriateness of using MICE in this setting and examples of validation one may conduct in this setting.

Finally, we will attempt to evaluate the missingness mechanism underlying cognitive data. It is by definition impossible to determine if data are MNAR because they are missing based on unobserved characteristics. However, we can attempt to investigate whether the data may be MNAR by using data collected by ARIC on the number of attempts made to contact an individual. Here we will examine whether the probability of the outcome being missing depends on the value of the outcome. Estimation of this model includes participants with observed and unobserved outcome data and will be fit using conditional likelihood and a set of estimating equations as described by Alho et al[8].

5. Main Study Questions:

We will develop a MICE model for imputing cognitive scores at visits 4 and 5, and compare results of estimated 20-year decline before and after the imputation. We will run models similar to those in MSP#2160 to facilitate comparison with IPAW. The goal will also be to validate the imputed values, and to investigate MAR and MNAR assumptions.

We hypothesize that:

- Results of estimated 20-year decline using MICE will show large absolute declines by exposure status and faster rates of decline compared to models not adjust for attrition.
- MICE will show stronger results than using IPAW limited to data observed on all participants if the CDR information differs by exposure and if it differs among persons who did versus those who did not attend visit 5.
- Including imputations for deceased participants (scores imputed 6 months prior to death) will show larger declines compared to including imputations for only living participants.

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).

<u>Motivating example</u>

We will look at the association between diabetes measured at visit 2 and 20-year cognitive decline. We may also include non-traditional markers of glycemia (eg fructosamine, glycated albumin).

Exposure

Diabetes will be defined based on self-reported physician diagnosis, medication use, or a HbA1c $\geq$ 6.5%.

Outcome

Cognitive function was assessed in all participants at visits 2, 4, and 5 using DWRT, DSST, and WFT. We will create a global measure of cognition by averaging Z scores of the three tests and dividing by the standard deviation; the resulting average scores will be standardized to their visit 2 means and standard deviations. Global Z is the outcome we are imputing.

Models

We will utilize generalized linear models fit using GEE and random effects longitudinal models as has been previously done[7,9,10]. Models will be adjusted for age, age squared, sex, race–field center, education, cigarette smoking status, alcohol consumption, body mass index, hypertension, history of coronary heart disease, history of stroke, and apolipoprotein E ε4 genotype.

Exclusions

For simplicity, we will only impute the outcome, so we will exclude persons missing the exposure (diabetes) and covariates used in the model at baseline. In the analyses of diabetes and 20-year change, only 274 persons were excluded (<2% of the visit 2 sample size) because they were missing model covariates or diabetes status (exposure). As a sensitivity analysis, we will rerun models without these exclusions, imputing missing exposure and covariate data.

MICE

MICE will be fit in Stata 13 using the "mi impute chained" command.

MICE Validation

- Method 1: Qualitative validation. This method will provide face validity for the imputation. We will examine whether imputed scores, and distribution of imputed scores, make sense from an "expert knowledge" perspective. For example, we expect persons who have an ICD-9 code for dementia would have lower imputed scores than similar persons who do not have such an ICD-9 code.
- Method 2: Quantitative validation. For this method we will use observed scores of participants who attended visit 5. We will set scores to missing using MCAR and MAR assumptions and see how well MICE performs. For imputation in persons who died, we can use scores from the brain and carotid MRI visits (data that are not used in the imputation) and compare the imputed scores of persons who died shortly after the two visits to scores observed at the visits. We can also use

information of persons who died shortly after completing visit 5 by removing their data from the imputation and then comparing observed and imputed values.

- Method 3: Simulation. We will simulate a complete dataset using ARIC data and MICE. This will allow us to obtain coefficients to use in a simulation model that are realistic. We will use the estimated coefficients to define an LDA model from which to simulate the data. We will simulate the data, including errors and random intercepts and obtain a complete simulated dataset. From this dataset, we can run models for 20-year change to obtain the "true" relationship between diabetes and cognitive change. Next, we will simulate missingness patterns to be similar to what we observe in ARIC. We will then compare the true relationship between diabetes and cognitive decline versus the relationship estimated before and after utilization of MICE.

Examining MNAR assumption

We will examine the MNAR assumption using contact attempt information and notation outlined by White et al[11]. Let r_{ik} be a response indicator for person i on attempt k ($r_{ik}=1$if person i responded on the k^{th} attempt, 0 otherwise, and $r_{ik}=.$ if no attempt was made). Let z_i be a binary indicator for the risk factor of interest (in this case diabetes), $\mathbf{X}_i$ be the vector of covariates, and $y_i$be the global Z score for person i. Using this notation, the probability of response on the k^{th} attempt can be written as:

$$P(r_{ik}=1 | r_{ik}{}^{1}.,z_i,\mathbf{X}_i,y_i)=a_k+bz_i+g\mathbf{X}_i+dy_i$$

The coefficient of interest, the informative missing parameter, is d. This model uses information from participants with and without observed outcome data and is fit using a conditional likelihood and a set of estimating equations[8]. We can formally test whether $d=0$, which would correspond to data that are MAR. A significant d(or large coefficient) means that the probability of missing the outcome depends on the outcome, and indicates that the data may be MNAR.

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 ICTDER03 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)?

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
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

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __x__ Yes ____ No
ARIC NCS
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.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. Accepted

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. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* 2nd ed. Hoboken, NJ: Wiley; 2002.

2. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *J Am Stat Assoc*. 1995;90(431):1112. doi:10.2307/2291350.

3. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12(1):46. doi:10.1186/1471-2288-12-46.

4. Little RJA. Regression With Missing X's: A Review. *J Am Stat Assoc*. 1992;87(420):1227. doi:10.2307/2290664.

5. Von Hippel PT. Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociol Methodol*. 2007;37(1):83-117. doi:10.1111/j.1467-9531.2007.00180.x.

6. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. 2012;68(1):129-137. doi:10.1111/j.1541-0420.2011.01666.x.

7. Rawlings AM, Sharrett a R, Schneider ALC, et al. Diabetes in Midlife and Cognitive Change Over 20 Years: A Cohort Study. *Ann Intern Med*. 2014;161(11):785-793. doi:10.7326/M14-0737.

8. Alho JM. Adjusting for Nonresponse Bias Using Logistic Regression. *Biometrika*. 1990;77(3):617. doi:10.2307/2337000.

9. Gottesman RF, Rawlings AM, Sharrett AR, et al. Impact of Differential Attrition on the Association of Education with Cognitive Change Over 20 Years of Follow-up: The ARIC Neurocognitive Study. *Am J Epidemiol*. 2014;179(8):956-966. doi:10.1093/aje/kwu020.

10. Schneider AL, Sharrett AR, Patel MD, et al. Education and cognitive change over 15 years: the atherosclerosis risk in communities study. *J Am Geriatr Soc*. 2012;60(10):1847-1853. doi:10.1111/j.1532-5415.2012.04164.x.

11. White IR, Kalaitzaki E, Thompson SG. Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an Internet-based alcohol trial. *Stat Med*. 2011;30(27):3192-3207. doi:10.1002/sim.4360.