Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Yishi Jia Date

Adjusting for selective attrition in a longitudinal study assessing the rate of cognitive decline

By

Yishi Jia
Master

Biostatistics and Bioinformatics

John Hanfelt, Ph.D.
Advisor

Robert Lyles, Ph.D.
Reader

Date

Adjusting for selective attrition in a longitudinal study assessing the rate of cognitive decline

By

Yishi Jia
B.S., Southwest University, 2020

Advisor: John Hanfelt, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
Rollins School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master
in Biostatistics and Bioinformatics
2023

Abstract

Adjusting for selective attrition in a longitudinal study assessing the rate of cognitive decline

By Yishi Jia

The thesis aimed to adjust for selective attrition in a longitudinal study that examined the rate of cognitive decline. It used longitudinal data from the Uniform Data Set of the National Alzheimer's Coordinating Center and conducted the analyses on 32,502 participants who were diagnosed as either cognitively normal, mildly cognitively impaired, or demented at the initial visit. The study used two methods, the inverse probability of attrition weighting (IPAW) method and multiple imputations, to investigate the effect of attrition-related selection bias on the estimated association between various factors and cognitive decline. IPAW approach allows for the use of all available data, assuming accurate estimation of attrition probability, and in this specific dataset, it may be preferable to use IPAW rather than imputation. However, the choice of method will depend on the specific characteristics of the dataset and assumptions about attrition.

Adjusting for selective attrition in a longitudinal study assessing the rate of cognitive decline

By

Yishi Jia
B.S., Southwest University, 2020

Advisor: John Hanfelt, Ph.D.

A dissertation submitted to the Faculty of the
Rollins School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master
in Biostatistics and Bioinformatics
2023

Contents

1 Introduction **1**

2 Methods **3**

2.1 Study data and population 3

2.2 Analytic Approach 4

2.2.1 Inverse probability of attrition weighting (IPAW) 4

2.2.2 Generalized estimating equation (GEE) 6

2.2.3 Multiple imputations 6

3 Results **8**

4 Discussion **12**

Bibliography **14**

List of Figures

3.1 Estimated lines of Normal, MCI, Demented from Geeglm model with stabilized weights 10

3.2 Estimated lines of Normal, MCI, Demented from Geeglm model with unstabilized weights 10

3.3 Estimated lines of Normal, MCI, Demented from Geeglm model with multiple imputation 11

List of Tables

2.1 Study participants and demographic table in the first visit 4

3.1 Contrast for the three disease groups (cognitively normal, MCI, and dementia) 8

3.2 Contrast for the three disease groups (cognitively normal, MCI, and dementia) with stable weights 8

3.3 Contrast for the three disease groups (cognitively normal, MCI, and dementia) with unstable weights 9

3.4 Contrast for the three disease groups (cognitively normal, MCI, and dementia) with imputation of missing value 9

Chapter 1

Introduction

The purpose of this project is to adjust for selective attrition in a longitudinal study assessing the rate of cognitive decline. The decline in cognitive function is a common occurrence with aging and is also the hallmark of dementia[1][2]. Mild cognitive impairment (MCI) is the stage between the expected decline in memory and thinking that happens with age and the more serious decline of dementia. The identification of modifiable risk factors for cognitive decline and dementia remains limited. This may, in part, be attributed to the methodological challenges that are unique to longitudinal studies investigating cognitive aging and late-life health outcomes[3]. The results of cognitive tests are strongly influenced by educational background and other cultural factors. Any investigation into the causes of age-related and disease-related cognitive decline must take these factors into consideration[4].

To investigate the effect of attrition-related selection bias on the estimated association between various factors and cognitive decline, we used inverse probability of attrition weighting (IPAW). Prior to computing predicted probabilities of continuation for each observation, we first created models of the probability of continuing in the study—that is, remaining alive and not being lost to follow-up[5]. Then, using these probabilities, we computed analytical weights that were inversely proportional to the probability of surviving the research. In order to ”compensate” for these observations’ underrepresentation in the observed follow-up data, observations with characteristics associated to a lower probability of continuation, such as physical weakness, were given larger weights. The weights were then used in our studies of the relationships between various factors and cognitive decline.

Others have addressed attrition bias by imputing missing values, using the Multiple Imputation method[6]. Multiple imputation is particularly useful when data are available for at least a subset of participants who did not attend all study visits, such as participants with low cognitive performance who are typically less likely to attend follow-up examinations[7].

Chapter 2

Methods

2.1 Study data and population

We conducted our analyses using a longitudinal dataset extracted from the Uniform Data Set (UDS) of the National Alzheimer's Coordinating Center (NACC) as of the December 2019 data freeze. Our cognitive measure of interest was Trailmaking Test B, a measure of executive function. The higher the score on the trail, the worse the impairment. We normed and demographically adjusted this measure. For example, if trailb_adj = 1.5, then the research participant had a score that was 1.5 standard deviations worse than a cognitively normal person of the same age, sex, race, and years of education. We assume attrition has occurred if a participant's most recent non-missing value of Trailmaking Test B was more than 18 months before the data freeze of December 2019.

In our experiment, 1766 participants designated as "impaired not MCI" were deleted, as this is a small group and of little scientific interest to us. Of the remaining 32502 participants, at the initial visit 15707 (48.3%) were classified as normal aging, 8371 (25.8%) were mild cognitive impairment (MCI), and 8424 (25.9%) were in dementia. Among all participants, their age ranged from 18 to 106 and had an average of

70.434 (SD=0.995), 18596 (57.2%) were female, 4432 (13.6%) were black, 1627 (5.0%) had history of alcohol abuse, 8335 (25.6%) were living alone, decades of smoking ranged from 0 (non-smoker) to 8 and had an average of 0.945 (SD=1.482), vascular and metabolomic comorbidity measured quantitatively in the range from 0 to 7 and had an average of 1.382 (SD=1.2), 3111 (9.6%) were underweight according to CDC guidelines, and 7341 (22.6%) were obese according to CDC guidelines.

	Cognitive normal (N=15707)	MCI (N=8371)	Demented (N=8424)	Total (N=32502)	p value
Education, in years					<0.001
Mean(SD)	15.776 (2.970)	15.282 (3.334)	14.794 (3.399)	15.394 (3.207)	
Range	0.000 - 29.000	0.000 - 30.000	0.000 - 30.000	0.000 - 30.000	
Age in decades, centered at 70 years					<0.001
Mean (SD)	0.019 (1.075)	0.286 (0.929)	0.303 (1.000)	0.161 (1.029)	
Range	-5.200 - 3.400	-4.900 - 3.400	-4.900 - 3.100	-5.200 - 3.400	
Sex					<0.001
Male	5413 (34.5%)	4187 (50.0%)	4306 (51.1%)	13906 (42.8%)	
Female	10294 (65.5%)	4184 (50.0%)	4118 (48.9%)	18596 (57.2%)	
Black					<0.001
No	13308 (84.7%)	7139 (85.3%)	7623 (90.5%)	28070 (86.4%)	
Yes	2399 (15.3%)	1232 (14.7%)	801 (9.5%)	4432 (13.6%)	
Attrition					<0.001
No	12701 (80.9%)	5945 (71.0%)	4248 (50.4%)	22894 (70.4%)	
Yes	3006 (19.1%)	2426 (29.0%)	4176 (49.6%)	9608 (29.6%)	

Table 2.1: Study participants and demographic table in the first visit

2.2 Analytic Approach

2.2.1 Inverse probability of attrition weighting (IPAW)

To perform IPAW, we used the 1st order autoregressive (AR_1) covariance structure in the model, which was selected by the Akaike information criterion (AIC) . Weights were applied at the level of observations within individuals, to adjust each person's contribution to our analysis at wave j. Censorship will be denoted (C_{ik}), with (C_{ik}) $= 1$ indicating the i_{th} participant was no longer in study by visit k. Similarly, we denote the baseline time-constant covariates as $(L_{(t)})$, a participant's entire time-varying covariate history up to visit k with a bar as $(\bar{L_{ik}})$, including past measurements of cognitive function, and the baseline covariates. We used the following time-constant predictors of attrition: gender, race, education, site number of each Alzheimer's Disease Center, family history of dementia, smoking status, history of alcohol abuse, living alone status, vascular and metabolomic comorbidity, and BMI. In addition, we used time-varying predictors of attrition, as follows: age, cancer status, depression, global clinical dementia rating. The weight for an individual's contribution to wave j is thus given by:

$$wt_{ij} = \prod_{k=0}^{j} \frac{1}{\hat{Pr}[C_{ik} = 0|C_{i(k-1)} = 0, \bar{L}_{k-1}]} \tag{2.1}$$

These weights are known as nonstabilized weights because, being the inverse of a probability, they have a guaranteed value of 1 for observations that contribute to the analysis. However, for individuals with a low probability of remaining alive and uncensored, these weights can potentially become very large. As a potential remedy, we also computed wave-specific, stabilized IPA weights[5] by multiplying the individual's nonstabilized weight at that wave by the conditional probability of remaining alive and uncensored up to that wave, given a subset of baseline covariates such as race and gender, as V_i (a subset of L_{i0}). The stabilized weights can be obtained by the following formula:

$$stwt_{ij} = \prod_{k=0}^{j} \frac{\hat{Pr}[C_{ik} = 0|C_{i(k-1)=0}, V_i]}{\hat{Pr}[C_{ik} = 0|C_{i(k-1)} = 0, \bar{L}_{k-1}]} \tag{2.2}$$

These probabilities are computed by logistic regression. To obtain the weights, we used the ipw package in R and replaced the extreme values with the 1% and 99% values of our estimation in both unstabilized and stabilized weights.

2.2.2 Generalized estimating equation (GEE)

The generalized linear model (GLM) does not require the specification of the form of the distribution but only of the relationship between the outcome mean and the predictors and between the mean and the variance. The results from applying a glm to our trails data showed that the slope of overall groups is negative, meaning participants would be better than before with visit follow years. This goes against the common sense of cognitive decline.

GEEs represent an extension of the GLM to accommodate correlated data. We focused on estimating and comparing the longitudinal slopes of the trail score in 3 disease groups: the group that is cognitively normal at baseline, the group with MCI at baseline, and the group with dementia at baseline. To fit the GEE, we used unstabilized and stabilized IPA weights and used the geepack package in R.

2.2.3 Multiple imputations

In the multiple imputations, we have retained the first three visits because when visit number reaches the fourth, there are 20,749 (66.0%) missing individuals in this dataset, and when visit number reaches the fifth, there are 23,480 (74.6%) missing individuals. After removing observations with visit number larger than three and individuals who have missing data in the first visit, we incorporated relevant time-constant baseline data (gender, race, education, smoking status, history of alcohol abuse, living alone status, family history of dementia, vascular and metabolomic co-morbidity, BMI) into multiple imputation.

We used multiple imputations to permit multivariate analysis of all participants who had baseline and at least one follow-up cognitive evaluation. We generated ten replications of the original data set, in which 612 missing values (1.95% of data) for 10 covariates considered in the analysis were replaced by values generated according to the Markov Chain Monte Carlo method [8] using the PROC MI SAS procedure. Then we imputed via monotone regression to generate ten replications in 17311 missing values (55.02% of data). Each imputed data set was then analyzed using the generalized estimating equation models described above and the results were pooled to calculate mean estimates and their standard error using the PROC MIANALYZE procedure.

Chapter 3

Results

The results of the study show the estimated intercepts and slopes for the three disease groups (cognitively normal, MCI, and dementia) for each of the four methods: without weights, with stable weights, with unstable weights, and with imputation missing values. The intercepts represent the estimated mean cognitive function score at baseline for each disease group, while the slopes represent the estimated change in cognitive function score per year for each disease group. The higher the scores, the worse the impairment. In all contrasts, the estimated slope is positive, meaning that on average, participants' cognitive function impaired over time. This aligns with the common sense of cognitive decline.

	MCI vs Normal	Demented vs Normal	Demented vs MCI
Intercept	1.083 (SD: 0.020)	2.768 (SD: 0.023)	1.686 (SD: 0.029)
slope	0.102 (SD: 0.007)	0.199 (SD: 0.012)	0.097 (SD: 0.013)

Table 3.1: Contrast for the three disease groups (cognitively normal, MCI, and dementia)

	MCI vs Normal	Demented vs Normal	Demented vs MCI
Intercept	1.083 (SD: 0.060)	2.260 (SD: 0.066)	1.178 (SD: 0.084)
slope	0.129 (SD: 0.009)	0.160 (SD: 0.014)	0.0307 (SD: 0.016)

Table 3.2: Contrast for the three disease groups (cognitively normal, MCI, and dementia) with stable weights

	MCI vs Normal	Demented vs Normal	Demented vs MCI
Intercept	1.107 (SD: 0.135)	2.132 (SD: 0.106)	1.025 (SD: 0.153)
slope	0.126 (SD: 0.010)	0.165 (SD: 0.015)	0.039 (SD: 0.017)

Table 3.3: Contrast for the three disease groups (cognitively normal, MCI, and dementia) with unstable weights

	MCI vs Normal	Demented vs Normal	Demented vs MCI
Intercept	1.053 (SD: 0.020)	2.686 (SD: 2.686)	1.633 (SD: 0.029)
slope	0.083 (SD: 0.010)	0.107 (SD: 0.013)	0.024 (SD: 0.014)

Table 3.4: Contrast for the three disease groups (cognitively normal, MCI, and dementia) with imputation of missing value

Comparing the estimated intercepts and slopes for the different weighting methods and imputation of missing values, the estimates of the intercepts and slopes vary slightly across the different weighting methods and for the imputation of missing values. However, the overall pattern of results remains the same, with positive slopes indicating impairment in cognitive function over time. The estimated intercepts and slopes show that participants with MCI or dementia at baseline had lower cognitive function compared to those who were cognitively normal at baseline. The estimated slopes also suggest that participants with MCI or dementia impaired more over time compared to those who were cognitively normal at baseline.

In terms of the contrasts, the dementia group had an even steeper decline in cognitive function compared to both the MCI and cognitively normal groups, as indicated by the positive slope contrast between the dementia and normal groups, and the dementia and MCI groups. The effect of selective attrition is larger for the demented and MCI participants than for the cognitive normal participants. Among MCI Participants, adjusting for selective attrition results in a steeper change in cognitive impairment over time compared to the unadjusted analysis. Given limited resources, Alzheimer's Disease Centers may have preferred to follow MCI participants over those with dementia.

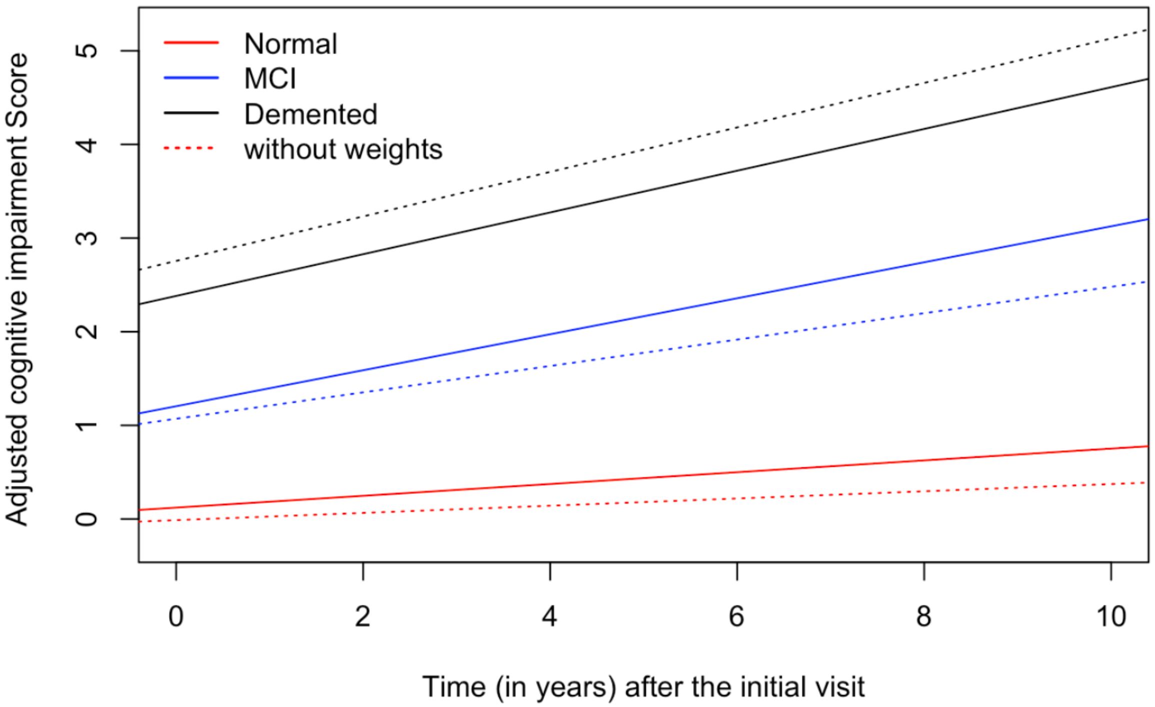

Figure 3.1: Estimated lines of Normal, MCI, Demented from Geeglm model with stabilized weights

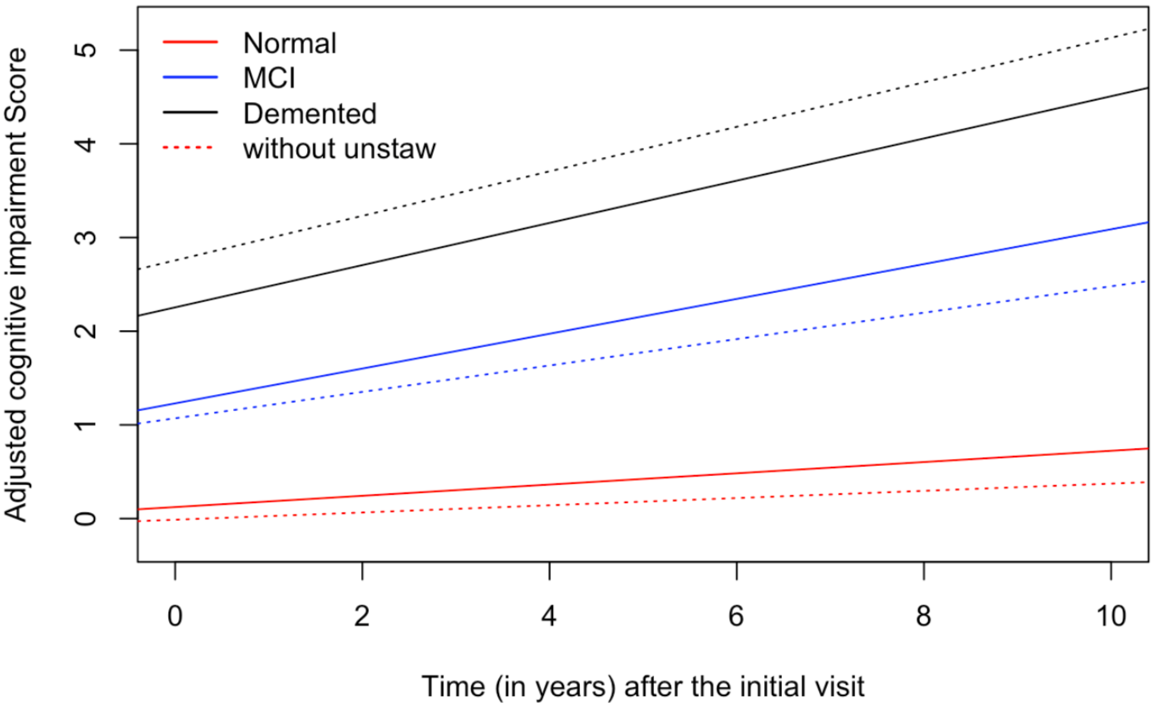

Figure 3.2: Estimated lines of Normal, MCI, Demented from Geeglm model with unstabilized weights

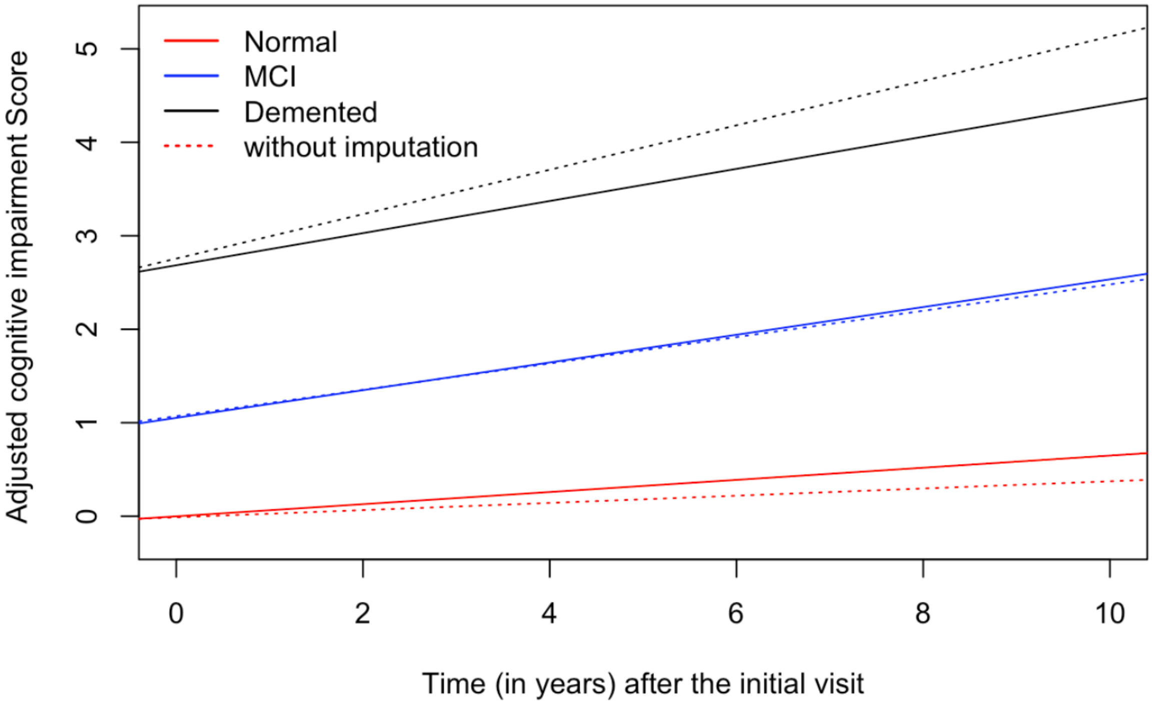

Figure 3.3: Estimated lines of Normal, MCI, Demented from Geeglm model with multiple imputation

Chapter 4

Discussion

For adjusting for selective attrition in a longitudinal study aimed at estimating the rate of cognitive decline and/or assessing the effect of the rate of cognitive decline on survival time, both weights and imputation can be used as methods to adjust for selective bias.

IPAW involves assigning weights to individuals based on their propensity to remain in the study, with those more likely to drop out receiving higher weights. The IPAW are then used in statistical analyses to account for the unequal probabilities of attrition. This approach allows for the use of all available data, but assumes that the probability of attrition can be accurately estimated.

Multiple imputation involves filling in missing data for individuals who have dropped out of the study based on their observed data from earlier time points. This approach can help to reduce bias due to selective attrition, but relies on the assumption that the data are missing at random or missing completely at random. In this analysis, multiple imputations may result in a loss of information, as we only used the first three visits of data, which might not capture the full variation of the missing data.

In this specific dataset, it may be preferable to use IPAW rather than imputation, as the IPAW would allow for the use of almost all available data. However, in other cases where the data are missing at random or missing completely at random, imputation of missing data may be a good method to adjust for selective bias. Ultimately, the choice of method will depend on the specific characteristics of the dataset and the assumptions that can be made about the missing data.

Bibliography

[1] Richard T. Linn, Philip A. Wolf, David L. Bachman, Janice E. Knoefel, Janet L. Cobb, Albert J. Belanger, Edith F. Kaplan, and Ralph B. D'Agostino. The 'Preclinical Phase' of Probable Alzheimer's Disease: A 13-Year Prospective Study of the Framingham Cohort. *Archives of Neurology*, 52(5):485–490, 05 1995.

[2] D. A. Bennett, R. S. Wilson, J. A. Schneider, D. A. Evans, L. A. Beckett, N. T. Aggarwal, L. L. Barnes, J. H. Fox, and J. Bach. Natural history of mild cognitive impairment in older persons. 59(2):198–205, 2002.

[3] Martha L. Daviglus. Preventing alzheimer's disease and cognitive decline. 2010.

[4] R. F. Gottesman, A. M. Rawlings, A. R. Sharrett, M. Albert, A. Alonso, K. Bandeen-Roche, L. H. Coker, J. Coresh, D. J. Couper, M. E. Griswold, G. Heiss, D. S. Knopman, M. D. Patel, A. D. Penman, M. C. Power, O. A. Selnes, A. L. Schneider, L. E. Wagenknecht, B. G. Windham, L. M. Wruck, and T. H. Mosley. Impact of differential attrition on the association of education with cognitive change over 20 years of follow-up: the aric neurocognitive study. 179(8):956–966, 2014.

[5] J. Weuve, E. J. Tchetgen Tchetgen, M. M. Glymour, T. L. Beck, N. T. Aggarwal, R. S. Wilson, D. A. Evans, and C. F. Mendes de Leon. Accounting for bias due to selective attrition: the example of smoking and cognitive decline. 23(1):119–128, 2012.

[6] D. B. Rubin. Multiple imputation after 18+ years. 91(434):473–489, 1996.

[7] A. M. Rawlings, Y. Sang, A. R. Sharrett, J. Coresh, M. Griswold, A. M. Kucharska-Newton, P. Palta, L. M. Wruck, A. L. Gross, J. A. Deal, M. C. Power, and K. J. Bandeen-Roche. Multiple imputation of cognitive performance as a repeatedly measured outcome. 32(1):55–66, 2017.

[8] J. L. Schafer. Analysis of incomplete multivariate data. 1997.