Published in final edited form as:

Stat Med. 2010 September 20; 29(21): 2260–2268. doi:10.1002/sim.4010.

Joint Modeling of Missing Data Due to Non-Participation and Death in Longitudinal Aging Studies

Kumar B. Rajan^{1,*} and Sue E. Leurgans²

¹ Rush Institute for Healthy Aging, Internal Medicine, Rush University Medical Center, Chicago IL

² Rush Alzheimer's Disease Center, Neurological Sciences & Preventive Medicine, Rush University Medical Center, Chicago IL

Summary

Specific age-related hypotheses are tested in population-based longitudinal studies. At specific time intervals, both the outcomes of interest and time-varying covariates are measured. When participants are approached for follow-up, some participants do not provide data. Investigations may show that many have died before the time of follow-up while others refused to participate. Some of these non-participants do not provide data at later follow-ups. Few statistical methods for missing data distinguish between “non-participation” and “death” among study participants. The augmented inverse probability-weighted estimators are most commonly used in marginal structure models when data are missing at random. Treating non-participation and death as the same, however, may lead to biased estimates and invalid inferences. To overcome this limitation, a multiple inverse probability-weighted approach is presented to account for two types of missing data, non-participation and death, when using a marginal mean model. Under certain conditions, the multiple weighted estimators are consistent and asymptotically normal. Simulation studies will be used to study the finite sample efficiency of the multiple weighted estimators. The proposed method will be applied to study the risk factors associated with the cognitive decline among the aging adults, using data from the Chicago Health and Aging Project (CHAP).

1. Introduction

In any longitudinal study, estimation of the mean of an outcome variable as a function of time is compromised due to missing data [1,2,3,4]. Missing data in population-based studies can occur for several reasons, but the most common reasons in studies of older adults are “non-participation” and “death”. Here, “non-participation” refers to the participant's lack of interest in completing the study requirements at any given follow-up interview. The problems with non-participation and death among aging individuals are fairly common, but are inadequately addressed by current methods [5,6,7]. It is standard practice to use available-case analysis by dropping observations with any missing data. If the data are missing completely at random (MCAR), then the available-case analysis will result in unbiased estimations.

When the data is missing at random (MAR), regression models can be used to obtain unbiased parameter estimates, if the missing data model can be specified correctly [8]. Parametric regression methods use maximum likelihood techniques for missing data by using imputation and factorization techniques [9]. Unlike the available-case analysis, methods for missing data require additional specification of a distribution for the variables

*Correspondence to: Kumar B Rajan, Department of Internal Medicine, 1645 W Jackson Blvd, Suite 675, Rush Institute for Healthy Aging, Chicago, IL 60612.

with missing values, and the mechanism that generates the missing values. Many model-based methods employ factorizations of the likelihood of the observed data using Newton-Raphson algorithm and the EM algorithm to obtain maximum likelihood estimates of the parameter under investigation [10,11,12]. These methods arise from defining a probabilistic model for the variables with missing values and making statistical inferences based on the maximum likelihood approach. More elaborate literature on likelihood-based approaches can be found in texts by Allison [13], and Daniels and Hogan [14].

On the other hand, weighted estimating equations are widely used semi-parametric regression methods which account for non-participation among study participants [15,16,17]. Several developments of the weighted estimating equations have been proposed [18,19,20]. The weighted generalized estimating equations (WGEE) do not require full specification of the likelihood. The estimators under the WGEE methods are consistent and asymptotically normal, under plausible conditions. Several authors have studied the difficulty of combining death with non-participation using weighted or competing risk models [21,22,23,24]. Specifically, Dufouil et al. [22] provides a re-weighting approach based on the assumption that mortality rates, following a drop-out from the study, are the same as corresponding rates for subjects in the study. In survival terminology, Dufouil et al. treats drop-out as an “event” and death as a “censoring”. However, we treat drop-out and death as two different events, and observed data as a form of censoring.

The objective of this article is to present a weighted estimating equation that accounts for missing data due to non-participation and death. At any given time of observation, a subject can either provide the necessary data, not respond to the interview, or not provide data due to death. To further clarify, the different conditions of missing data which exist will be explained. Firstly, if the time of the assessment was before death, then the missing data is treated as missing due to non-participation. Secondly, if the time of the assessment was after death, then the response is no longer defined at that given time point, therefore, the data is treated as missing due to death. In this particular situation, the time of the follow-up is also the time of death of the study participant. For example, if the study participants were to be interviewed every three years, and at the time of the follow-up it was found that the study participants had died two years prior to the follow-up, then the time of follow-up for the study participants were one-year and not three years. Clearly, after death, the study participants are no longer followed or assessed. So, data missing due to death is monotone by nature. Based on this idea, a multiple weighted approach is developed by computing weights based on a multinomial probability model to jointly estimate the weights, to account for the non-participation and death among the study participants. The advantage of such an approach is that at any given follow-up time, the individual is in a specific state, which can be used to estimate the appropriate weight necessary to obtain unbiased estimate of the marginal mean from the observed sample.

This article will be organized into five sections. Section 2 provides a motivating example in the Chicago Health and Aging Project (CHAP). In section 3, we review the generalized estimating equations (GEE) approach under different missing data mechanisms, and introduce our multiple weighted generalized estimating equation. In section 4, we study the performance of the multiple weighting approach to available-case GEE and weighted GEE, using simulation models. We will apply the available-case GEE, the weighted GEE, and the multiple weighted GEE approaches to study the risk factors associated with cognitive decline in aging adults in Section 5. Finally, in Section 6, we will discuss the advantages, limitations, improvements, and future research directions for handling non-participation and death in longitudinal studies.

2. Chicago Health and Aging Project

The CHAP study comprised of a cohort of subjects, 65 years and older, who were followed for 10 years or more [25]. To replenish the study cohort, new successive cohorts were added about every four years, beginning in the year 2000. The main scientific question of interest was to study the risk factors associated with lower cognitive function and its change over a period of time. In the CHAP population, the loss of follow-up due to death was a major concern. Deceased participants were substantially different from subjects who provided data, or from living subjects who did not complete an interview. Treating the non-participation and deceased groups of participants to be the same could potentially lead to biased results. On closer examination, the probabilistic model for outcomes were substantially different for those who participated, in comparison to those who did not complete the survey, or for those who did not participate due to death. In order to study the change in the cognitive function among the aging population, we need to account for selection bias due to non-participation and death.

The CHAP study was performed in three adjacent neighborhoods on the south side of Chicago. A census performed from 1994-1996 listed 64,911 residents, of whom 8,509 (13%) were 65 years or older. Of these residents, 6,158 (61% Blacks) were the original cohort of CHAP study participants. Data collection cycles were approximately three years apart. In-home interviews were conducted to collect covariate and outcome data [25]. Cognitive function was evaluated using a battery of four tests and was summarized as a global standardized score. This score combined variables with different ranges and floor-ceiling effects by averaging the four tests together after centering and scaling to the baseline mean and standard deviation [26]. Thus, a participant whose performance matched the average participant at baseline had a composite cognitive score of 0, and a person who performed one SD better than average on every test had a composite cognitive score of +1. The CHAP study enrolled successive age cohorts of community residents as they attained the age of 65. As a result, the CHAP study consists of four age cohorts with varying follow-up times for each cohort. Thus, the CHAP study consisted of 9,871 (67% Blacks) study participants who were followed longitudinally over four cycles of data collection. The average age among study participants at baseline was 73.1 years (SD=7.05 years). The demographic and socio-economic descriptives for the study cohort at baseline are shown in Table I. The design of CHAP is described in more detail by Bienas et al. [25].

Details of the CHAP study follow-ups are shown in Table II. The average time to death among those who provided data at Cycle 1 was 1.88 years (SD=1.17) compared to 1.26 years (SD=0.93) among those who did not provide data. The average time to death among those who provided data at Cycle 2 was 5.02 years (SD=1.20) compared to 5.89 years (SD=1.16) among those who did not provide data. The average time to death among those who provided data at Cycle 3 was 8.04 years (SD=1.24) compared to 9.02 years (SD=1.43) among those who did not provide data. In this cohort, the study participants who had provided data in the previous cycle died earlier than those who did not provide data, with exception to Cycle 1. The theory that participants who did not complete a survey might be closer to death was not supported in this study. From Table II, approximately 25% to 35% of study participants did not complete the survey due to non-participation or death in the follow-up cycles. Excluding these individuals from the analysis will cause severe bias in the coefficient estimates. Also, the number of study participants who were deceased before each cycle was higher than those who did not participate at the follow-up interview. Thus, it can be problematic to assume that these two types of missing data were the same.

The average time to follow-up at Cycle 2 was 3.29 years (SD=.45) among those who completed the survey; the average time to follow-up at Cycle 3 was 6.37 years (SD=0.56);

and the average time to follow-up at Cycle 4 was 9.35 (SD=0.55). Additionally, the average time to follow-up among blacks and whites, along with the descriptive measures of education status, and the level of income, were all fairly similar. The average time to death among blacks and whites, when incorporating the education and income groups for those who completed the survey at the earlier cycle, as well as for those who did not provide data, were also fairly similar. However, the rates of participation and death among the racial and education groups were different. Understandably, the time to death among the older age group (≥ 75 years) was lower than that of the younger age group (65 – 74 years). Moreover, the participation rates of the younger age group was slightly higher than that of the older age group.

The participation rates and mortality rates at each cycle were a function of age, race, education, and sex of the study participant. The probability that a study participant at any given cycle will be followed to the next cycle, can be modeled using baseline demographic and socio-economic variables. In the next section, a multiple weighting approach is proposed, based on the idea that each participant, at any given cycle, will have a probability associated with providing the data at next cycle, not participating in the survey, or dying before the follow-up interview.

3. Multiple Weighting Approach

Let Y_{ij} represent the outcome for participant i at cycle j ; $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m_i$. Let X_{ij} denote the covariate vector, which could be baseline or time-varying and includes a vector of 1 for the intercept. The marginal mean of Y_{ij} given X_{ij} follows the regression model

$$E(Y_{ij}|X_{ij}) = g(\beta^T X_{ij}), \quad (1)$$

where g is a monotone link function and β is a $p \times 1$ vector of unknown parameters. Under MCAR, consistent and asymptotic normal estimators can be obtained by solving the following generalized estimating equation (GEE):

$$U(\hat{\beta}) = n^{-1/2} \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = 0, \quad (2)$$

where Y_i is the vector of responses for subject i , $\mu_i = E(Y_i)$, $D_i = \partial \mu_i / \partial \beta^T$, and V is an $m_i \times m_i$ working covariance matrix of Y_i , with m_i equal to the length of Y_i . Under mild regularity conditions, the roots of the estimating equations provide consistent and asymptotic normal estimates [17]. Our objective is to propose a class of weighted estimating equations that lead to consistent and asymptotically normal estimators of β , provided that the probability of non-participation and death at time j depends only on subject specific data until $j - 1$. The estimator will be unbiased provided the probability model for non-participation and death given previous data is correctly specified or the marginal mean model for the population average is correctly specified [27].

For each participant i at cycle j , we define two indicator variables R_{ij} and S_{ij} to denote non-participation and death. If an outcome was observed for subject i at cycle j , then $R_{ij} = S_{ij} = 0$. If the participant deceased before cycle j , then $S_{ij} = 1$ and $R_{ij} = 0$. Note that S_{ij} is monotone, that is, for a deceased participant with $S_{ij} = 1$, then $S_{i(j+1)} = 1$, for deceased participants. If an outcome was not observed, even though subject i was alive at cycle j , then $R_{ij} = 1$ and $S_{ij} = 0$. We shall assume that R_{ij} is monotone and that the covariate vector X_{ij} is completely

observed. In addition, we shall assume that the random variables R_{ij} and S_{ij} satisfy the following probabilistic model:

$$\begin{aligned}
 P(R_{ij}=S_{ij}=0|R_{i(j-1)}) \\
 &= S_{i(j-1)}=0, X_{ij}, Y_{ij}=P(R_{ij} \\
 &=S_{ij}=0|R_{i(j-1)}) \\
 &= S_{i(j-1)}=0, X_{i(j-1)}, Y_{i(j-1)}=P(R_{ij} \\
 &=0|R_{i(j-1)}) \\
 &= S_{i(j-1)}=0, X_{i(j-1)}, Y_{i(j-1)}) \\
 &\times P(S_{ij} \\
 &=0|R_{i(j-1)}) \\
 &= S_{i(j-1)}=0, X_{i(j-1)}, Y_{i(j-1)})
 \end{aligned} \tag{3}$$

Under this assumption, the probability of participation at any time j is independent of the current outcome and future outcomes are conditioned on the observed covariates. Additionally, the joint probability distribution of (R, S) given the past covariates can be factored into two conditional probability distributions. Each of these distributions needs to be bounded and greater than zero to provide consistent and asymptotically normal estimators. Given assumption (3), we can identify $E(Y_{ij}|X_{ij})$ from the observable random variables in the presence of missing data due to non-participation and death [16]. To review further, equation (3) implies a formula with two random variables [28],

$$E(Y_{ij}|X_{ij}) = \int \dots \int E(Y_{ij}|R_{ij}, S_{ij}, X_{ij}, y_{i(j-1)}) \times \prod_{t=1}^j dF(y_{it}|R_{it}, S_{it}, X_{it}, y_{i(t-1)}), \tag{4}$$

where the right side depends only on the joint distribution of the observed random variables. Thus, the Horvitz-Thompson estimator $E(Y_{ij}|X_{ij})$ is a weighted average of $E(Y_{ij}|R_{ij}, S_{ij}, X_{ij},$

$y_{i(j-1)})$ with specific weights $\prod_{t=1}^j f(y_{it}|R_{it}, S_{it}, X_{it}, y_{i(t-1)})$.

Let ψ_{ij} denote the response probabilities given by $\psi_{ij} = P(R_{ij} = S_{ij} = 0|R_{i(j-1)} = S_{i(j-1)} = 0, X_{i(j-1)}, Y_{i(j-1)})$, defined using a vector of $q \times 1$ vector of unknown parameters α , that is

$$\psi_{ij} = \psi_{ij}(\alpha). \tag{5}$$

Usually $\psi_{ij}(\alpha)$ is chosen to be a multinomial function. Thus, we assume that given $(R_{i(j-1)}, S_{i(j-1)})$, (R_{ij}, S_{ij}) follows a multinomial model on functions parameterized by α . Now, define $\pi_{ij}(\alpha) = \psi_{i1}(\alpha) \times \dots \times \psi_{ij}(\alpha)$, the probability that study participant i completed an assessment at time j . When assumption (3) holds, $\pi_{ij}(\alpha)$ is the conditional probability of observing participant i at time j given past data. Thus, for each subject i , define a diagonal matrix of weighted observations, $\Phi_i(\alpha) = \text{diag}((1 - r_{i1})(1 - s_{i1})/\pi_{i1}(\alpha), (1 - r_{i2})(1 - s_{i2})/\pi_{i2}(\alpha), \dots, (1 - r_{ij})(1 - s_{ij})/\pi_{ij}(\alpha))$.

The estimating equations (2) only involves contributions from available cases. To increase the efficiency of the estimation, we want to use data from individuals with some missing data. This was accomplished by adding a term of zero expectations to the estimating equation. Thus, the generalized estimating equation presented in (2) is modified to

$$U(\widehat{\beta}, \widehat{\alpha}) = n^{-1/2} \sum_{i=1}^n \left\{ D_i^T V_i^{-1} \Phi_i(\widehat{\alpha})(Y_i - \mu_i) - (\Phi_i(\widehat{\alpha}) - 1) \varphi(Y_i; \widehat{\beta}, \widehat{\alpha}) \right\} = 0, \quad (6)$$

where $\Phi_i(\alpha)$ is a $m_i \times m_i$ diagonal matrix of weights for each subject and $\varphi(Y_i; \widehat{\beta}, \widehat{\alpha})$ is the conditional expectation of Y_i given the covariates and observed data. Thus, we are able to adjust for missing data due to the participant non-response and death by weighting the observed responses. The estimator of the mean based on equation (6) is

$$\widehat{\mu} = n^{-1} \sum_{i=1}^n \left\{ \frac{(1 - R_i)(1 - S_i)Y_i}{\pi(X_i, \widehat{\alpha})} - \frac{(1 - R_i)(1 - S_i) - \pi(X_i, \widehat{\alpha})}{\pi(X_i, \widehat{\alpha})} E(Y_i | R_i, S_i, X_i) \right\}, \quad (7)$$

The consistency of $\widehat{\beta}$ follows from the fact that $U(\beta, \alpha)$ has mean 0 under conditions (1), (3), and (4). To show that $E(U_i) = 0$, we use the following identity

$$\Phi_i(\alpha) = I_i + \sum_{j=1}^{m_i} C_{ij}(\alpha) I^j, \quad (8)$$

where I_i is a $m_i \times m_i$ identity matrix and I^j is a partial identity matrix with diagonal elements 1 for indices $\geq j$, or 0 otherwise, and $C_{ij} = ((1 - R_{ij})(1 - S_{ij}) - \psi_{ij}(\alpha)(1 - R_{ij-1})(1 - S_{ij-1}))/\pi_{ij}(\alpha)$. Substituting (7) in (6), the generalized estimating equation is written as

$$U(\beta, \alpha) = n^{-1/2} \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) + n^{-1/2} \sum_{j=1}^{m_i} \sum_{i=1}^n D_i^T V_i^{-1} C_{ij} I^j (Y_i - \mu_i) + o_p, \quad (9)$$

where o_p is a mean 0 term added to the estimating equation in (6). The second term on the right side of equation 7 is a discrete mean zero martingale with respect to some filtration process $F_i(j)$ [28,29]. Since, $E[U(\beta)] = 0$ and the second term is zero, we conclude that $E[U(\beta, \alpha)] = 0$. Also, the consistency of the two terms guarantees the consistency of $U(\beta, \alpha)$ provided the second and third order derivatives exist and are bounded. The limiting distribution of $n^{1/2}(\widehat{\beta} - \beta)$ can be obtained using Taylor series expansions

$$n^{1/2}(\widehat{\alpha} - \alpha) = -E(\partial U_\alpha / \partial \alpha^T)^{-1} U(\alpha_0) + o_p(1), \quad (10)$$

and

$$n^{1/2}(\widehat{\beta} - \beta) = -E(\partial U_{\beta, \alpha} / \partial \beta^T)^{-1} \{U(\beta, \alpha) + E(U_{\beta, \alpha} / \partial \alpha^T)^{-1} n^{1/2}(\widehat{\alpha} - \alpha)\} + o_p(1). \quad (11)$$

The limiting distribution $n^{1/2}(\widehat{\beta} - \beta)$ can be derived by differentiating (9) with respect to α , and (10) with respect to β .

Suppose the probability model (5) is correctly specified but the model (1) might not be correctly specified, then the mean estimator is approximated by

$$\widehat{\mu} = n^{-1} \sum_{i=1}^n \left\{ Y_i + \frac{(1 - R_i)(1 - S_i) - \pi(X_i, \widehat{\alpha})}{\pi(X_i, \widehat{\alpha})} [Y_i - g(X_i, \beta^*)] \right\}, \quad (12)$$

where β^* is the parameter vector under the incorrect model. The consistency of the estimator follows from the fact that the conditional expectation of $((1 - R_i)(1 - S_i) - \pi(X_i, \widehat{\alpha})) / \pi(X_i, \widehat{\alpha}) = 0$. Similarly, the consistency of the mean estimator when the model (1) is incorrectly specified but the model (5) is correctly specified is established by showing that the second term $E(Y_i | R_i, S_i, X_i) - E(Y_i | X_i)$ equals to zero.

4. Simulation Studies

In this section, we compare the performance of the multiple weighted GEE (MWGEE) method to the available-case GEE (GEE) and the weighted GEE (WGEE) method using three different simulation models. All three methods were implemented using an unstructured working correlation matrix. The mean parameters were evaluated with respect to both bias and variability. The observations were simulated from a normal distribution of the form

$$Y_{ij} \sim N(1.39 - 0.29 \times T_j - 0.54 \times X_i - 0.32 \times Z_i),$$

where $j = 0, 1, 2$ denotes the time points, $T = \{0, 3, 6\}$ denotes the time of observation, $X \sim B(1, 0.5)$ was a binary baseline covariate, and $Z \sim N(0, 1)$ was a continuous baseline covariate.

In the first simulation model, the missing indicator was simulated from a multinomial distribution for those with observed responses at the earlier time point, with the probability of providing data to be 0.7, the probability of death to be 0.2, and the probability of non-response to be 0.1. In the second model, the missing indicator was simulated from a multinomial regression model, with the probability of missing observations being dependent on the covariate values. The probability of missing non-participation data and deceased subjects were estimated using the probability model

$$\text{logit}(p_{ij}) = 0.2 + 0.1 \times T_j + 0.2 \times X_i + 0.4 \times Z_i,$$

which were then classified into three groups with about 15% non-participation and 20% death.

In the third simulation study, the missing indicator was simulated from a multinomial distribution, with the probability of missing observations being dependent on the covariate and outcome of interest. The third model was given by the following form

$$\text{logit}(p_{ij}) = 0.2 + 0.1 \times T_j + 0.2 \times X_i + 0.2 \times Z_i + 0.2 \times Y_{ij}.$$

The predicted probabilities from the three simulation models were used to simulate the observed and missing indicator variables. The missing indicators were simulated with an added condition that the data follows monotonicity for non-participation and death indicators.

The first set of simulation studies used a sample size of 200, with two follow-up times. A second set of simulation studies was implemented with the same parameters but with a sample size of 500 and three follow-up times. For each scenario, 1000 replications were conducted.

The weights for participants with missing responses and death were computed using a polychotomous logistic regression model with baseline covariates and time as predictors. The final regression model used the appropriate weighting procedure with an unstructured correlation matrix for the observed participants to estimate the model parameters. The first three columns of Table III shows the simulation results in terms of empirical bias in parameter estimates for a sample size of 200 with two follow-ups. Figure 1 shows the RMSE for the sample size and follow-up time periods. Under the MCAR assumption, the available-case analysis estimates the mean parameter with small empirical biases for all parameters of interest. The weighted GEE and multiple weighted GEE perform fairly well with marginally higher bias and variability. Under the MAR condition, the available-case analysis had a larger bias compared to the weighted GEE and multiple weighted GEE. Also, the multiple weighted GEE had a smaller bias and variability than the traditional weighted GEE. For simulations based on the NMAR assumption, the weighted GEE and multiple GEE methods had a smaller bias than available-case analysis. However, the magnitude of the bias was substantially larger compared to the values of the parameters.

The results for the second set of simulation studies with a sample size of 500 with three follow-up time points are shown in the last three columns of Table III. The RMSE for the three methods with the same configuration are shown in Figure 1. Under the MCAR assumption, the bias and RMSE with available-case analysis was the same as weighted and multiple weighted GEE. Under the MAR condition, the multiple weighted method performed much better than the weighted GEE and available-case analysis. Under the NMAR condition, the weighted GEE and multiple GEE had a smaller bias and RMSE than the available-case analysis.

5. Results of the Chap Study

The CHAP study had a total of 9,871 participants who were followed over a period of 10 years and up to four cycles. A total of 3,268 (33%) of the study participants were deceased before completing all follow-ups. Another 2,538 participants missed 5,811 possible visits due to non-participation, in the four cycles. The mean standardized cognition score at baseline was 0.1356 with a standard deviation of 0.8377. The GEE analysis used an unstructured working correlation matrix with an indicator for race (1=Blacks, 0=Whites), time since baseline (time), gender (1=males, 0=females), age (centered age at baseline in years), and education (centered years of education at baseline).

The marginal population average is the difference in cognitive scores between the two racial groups as it changed over time. A preliminary analysis was performed by an available-case analysis using GEE with an unstructured correlation matrix [17]. To account for the missing responses among the participants, a weighted GEE model was used with a logistic model to estimate the probability of being observed [16]. To account for missing data due to non-response and death, a trichotomous regression model was implemented to estimate the probability of being observed, the probability of missing observations due to non-participation, or the probability of death with age, race, gender, education, and time in the study, as predictor variables. At baseline, all subjects were observed and weights were set to 1. If the subject was still alive at the cycle, the weight for that cycle would be the cumulative probability multiplied by the predicted probability of participating in the survey. If the subject was deceased at the cycle, then the weight would be accumulated and multiplied by

1-predicted probability of non-participation. For deceased subjects, the last observation before death would be the accumulated weight multiplied by the probability of death.

Table IV shows the parameter and SE estimates using the available-case GEE, the weighted GEE, and the multiple weighted GEE for the CHAP study. A bold font indicates a p-value of less than .05. The coefficients for baseline age and baseline education were about the same in the three models. However, the coefficients for Blacks, males, and time in the study showed a higher decline in the cognitive score using the multiple weighted approach compared to the available-case and the weighted GEE approach. Using the multiple weighted GEE model, Blacks had a cognition score lower by 0.4083 compared to non-blacks. The results from the multiple weighted GEE approach shows that Blacks had a slower decline in cognitive score compared to the whites. The change in cognitive function among Blacks was 0.054 standard deviation compared to the non-blacks over 10 years. This association was not observed in the available-case analysis and the weighted GEE approach. The change in cognition score for males and education over time was about 2 to 3 times of that found in the available-case analysis. Also, the decline in cognition score was 0.53, more than half a standard deviation over 10 years, using the multiple weighted GEE approach compared to the weighted GEE approach of 0.39 standard deviation. These differences in the results could be attributed to the fact that the subjects who participated in the study were quite different from those who did not participate, and those who were deceased during the follow-up.

6. Discussion

In longitudinal studies of aging populations, death and missing responses can distort estimates of the parameters under study. The statistical models used to address the scientific question need to adequately account for missing responses and death among participants. This article focuses on one possible way to account for death and missing data by using a multiple weighting approach under missing at random assumption. In previous literature, weighting is frequently used to account for non-participation, and this article extends the idea to simultaneously account for non-participation and death.

The simulation studies presented in this article illustrate that by using a different weighting mechanism for non-participation and death, we can obtain mean estimates with smaller overall bias and variability. The multiple weighting method performs well under different missing data mechanisms. In addition, the multiple weighting method can be viewed as an extension of the weighted GEE method with two random variables rather than one. Under non-MAR assumptions, the multiple weighting method performed fairly well in comparison to the available-case analysis and the weighted GEE approach, but will need further improvement. When applied to the CHAP study, the results from the multiple weighted GEE model were different from the available-case analysis and the weighted GEE method, especially in terms of race, which is one of the primary risk factors under investigation. Also, the cognitive scores changed significantly over time for other demographic groups under investigation. Based on these findings, we recommend the use of the multiple weighting approach when there is a substantial loss of subjects during follow-up due to non-participation and death when the data are missing at random.

The extension of this method to other situations merits further study. The multiple weighted GEE approach was applied and studied using a continuous response. Performance of the multiple weighted GEE approach also needs to be studied for binary and ordinal outcomes. Other types of models can also be important. For continuous outcomes, if interest is focused on subject-specific trajectories, rather than on the marginals means, a random effects model might be the primary approach for modeling. The specification of a working correlation

matrix might also merit study. In reality, it is difficult to know the underlying missing data mechanism. However, in the CHAP study, we found that non-participation and death can be modeled adequately using a MAR assumption. Using baseline and time-varying covariates, the probability of providing data was predicted with fairly good accuracy. About 2.5% of the non-participants provided data at subsequent follow-ups, but they were treated as non-participants in the analysis. In any case, the potential for improving the accommodation of non-participation and death in analysis when studying elderly patients is evident.

Acknowledgments

We wish to thank Dr. Denis A. Evans and Dr. Carlos F. Mendes de Leon for sharing the CHAP study data and providing their suggestions and comments. This research was supported by grants from the National Institute on Aging (AG 11101) and the National Institute of Environmental Health Sciences (ES 10902). We also wish to thank the editor and two reviewers for their thoughtful comments.

Contract/grant sponsor: National Institute for Health Aging; contract/grant number: 98–1846389

References

1. Diehr P, Williamson J, Burke L, Psaty B. The aging and dying processes and the health of older adults. *Journal of Clinical Epidemiology*. 2002; 55:269–278.10.1016/S0895-4356(01)00462-0 [PubMed: 11864798]
2. Brayne C, Spiegelhalter D, Dufoil C, Chi LY, Dening TR, Paykele S, O'Connor DW, Ahmed A, McGee MA, Huppert FA. Estimating the true extent of cognitive decline in the old. *Journal of the American Geriatrics Society*. 1999; 47:1283–88. [PubMed: 10573434]
3. Hoeymans N, Feskens EJ, van den Bos GA, Kromhout D. Age, time, and cohort effects on functional status and self-related health in elderly men. *American Journal of Public Health*. 1997; 87:1620–1625. [PubMed: 9357342]
4. Siegler IC. The terminal drop hypothesis: fact or artifact. *Experimental Aging Research*. 1975; 1:169–185.10.1080/03610737508257957 [PubMed: 802856]
5. Revicki DA, Gold K, Buckamn D, Chan K, Kallich JD, Woolley JM. The aging and dying processes and the health of older adults. *Journal of Clinical Epidemiology*. 2002; 55:269–278.10.1016/S0895-4356(01) [PubMed: 11864798]
6. Diehr P, Johnson LL, Patrick DL, Psaty B. Methods for incorporating death into health related variables in longitudinal studies. *Journal of Clinical Epidemiology*. 2005; 58:1115–1124.10.1016/J.JCLINEPI.2005.05.002 [PubMed: 16223654]
7. Diehr P, Johnson LL. Accounting for Missing Data in End-of-Life Research. *Journal of Palliative Medicine*. 2005; 8:S50–S57.10.1089/jpm.2005.8.5-50 [PubMed: 16499469]
8. Lin DY, Ying Z. Semi-parametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*. 2001; 96:103–118.10.1198/016214501750333018
9. Little RJ, Raghunathan T. On summary measures analysis of the linear mixed effects model for repeated measures when data are not missing completely at random. *Statistics in Medicine*. 1999; 18:2465–2478. [PubMed: 10474153]
10. Little, RJ.; Rubin, D. *Statistical Analysis with missing data*. Wiley; New York: 2002. p. 97-127.
11. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*. 1977; 39:1–38.
12. Ibrahim JG, Chen MH, Lipsitz R. Monte Carlo EM for missing covariates in parametric regression methods. *Biometrics*. 1999; 55:591–596.10.1111/J.0006-341x.1999.00591.x [PubMed: 11318219]
13. Allison, PD. *Missing data*. Sage; California: 2001. p. 12-74.
14. Daniels, MJ.; Hogan, JW. *Missing data in longitudinal studies*. Chapman & Hall; Boca Rotan, Florida: 2008. p. 165-215.
15. Tsiatis, AA. *Semiparametric theory and missing data*. Springer; New York: 2006. p. 53-150.

16. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1999; 90:106–121.
17. Liang Y, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–21.10.1093/biomet/73.1.13
18. Bang H, Robins MT. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005; 61:962–972.10.1111/j.1541-0420.2005.00377.x [PubMed: 16401269]
19. Buzkova P, Lumley T. Semiparametric loglinear regression for longitudinal measurements subject to irregular, biased follow-up. *Journal of Statistical Planning and Inference*. 2008; 138:2450–2461.10.1016/j.jspi.2008.10.013
20. O'Brien LM, Fitzmaurice GM. Analysis of longitudinal multiple source binary data using generalized estimating equations. *Applied Statistics*. 2004; 53:177–193.10.1046/j.0035-9254.2003.05296.x
21. Shardell M, Miller RR. Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine*. 2008; 27:1008–1025.10.1002/sim.2964 [PubMed: 17579923]
22. Dufouil C, Brayne C, Clayton D. Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*. 2004; 23:2215–2226.10.1002/sim.1821 [PubMed: 15236426]
23. Horton NJ, Fitzmaurice GM. Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine*. 2004; 23:2911–2933.10.1002/sim.1879 [PubMed: 15344194]
24. Miller ME, Ten Have TR, Reboussin BA, Lohman KK, Rejeski WJ. A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple-cause nonresponse. *Journal of the American Statistical Association*. 2001; 96:844–857.10.1198/016214501753208555
25. Bienias JL, Beckett LA, Bennett DZ, Wilson RS, Evans DA. Design of the Chicago health and aging project. *Journal of Alzheimer's Disease*. 2003; 5:349–355.
26. Nussbaum, PD. *Handbook of neuropsychology and Aging*. Plenum Press; New York: 1997. p. 351–360.
27. Rosenbaum PR. Model-Based direct adjustment. *Journal of the American Statistical Association*. 1987; 82:387–394.
28. Robins JM. Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods- Applications to control of the healthy worker survivor effect”. *Computers and Mathematics with Application*. 1987; 14:923–945.10.1016/0898-122(87)90238-0
29. Kalbfleisch, JD.; Prentice, RL. *The statistical analysis of failure time data*. New York: John Wiley; 1980.

Figure 1. RMSE for available-case GEE, weighted GEE, and multiple weighted GEE under different missing data mechanisms computed over over 1000 simulated datasets.

Table I

Baseline characteristics among 9,871 participants of the Chicago Health and Aging Project

Characteristic	No. of Observations (%)	Mean Cognitive Score (SD)
Age (yrs.)		
65-74	6868 (70)	0.317 (.689)
75-84	2166 (22)	-0.139 (.931)
≥ 85	837(8)	-0.749 (1.040)
Education (yrs.)		
0-9	1893 (19)	-0.591 (0.885)
10-12	4103 (42)	0.121 (0.746)
13-16	2949 (30)	0.474 (0.632)
≥ 17	859 (9)	0.636 (0.661)
Sex		
Female	6041 (61)	0.138 (0.875)
Male	3830 (39)	0.110 (0.799)
Race		
Black	6635 (67)	-0.002 (0.846)
White	3236 (33)	0.396 (0.780)

Table II

Number of participants (%) followed over the duration of the study

Cycle	Eligible	Completed	Did not participate	Deceased
1	9871	9811 (99)	60 (1)	-
2	8585	6379 (74)	829 (10)	1377 (16)
3	5819	3782 (65)	972 (17)	1065 (18)
4	3839	2336 (61)	677 (18)	826 (21)

Table III

Empirical bias ($\times 100$) of parameter estimates based on 1000 simulated datasets. Two sample size configurations are considered. A sample size of 200 with two follow-up times and a sample size of 500 with three follow-up times

Mechanism	Parameters	Sample size=200				Sample size=500			
		GEE	WGEE	MWGEE	GEE	WGEE	MWGEE	GEE	MWGEE
MCAR	Intercept	.313	.347	.284	.285	.318	.204		
	Time	.074	-.089	.077	.069	-.083	.082		
	X	.021	-.188	-.116	.019	-.185	.107		
	Z	-.100	.116	.132	-.093	-.119	-.116		
MAR	Intercept	.826	.560	.232	.788	.541	.279		
	Time	-.558	-.188	-.105	-.517	-.156	-.085		
	X	-.452	-.347	-.098	-.452	-.334	-.083		
	Z	-.316	-.213	-.107	-.334	-.231	-.076		
NMAR	Intercept	-1.468	-.947	-.755	-1.584	-.858	-.811		
	Time	-5.832	-5.087	-2.687	-5.231	-4.642	-4.478		
	X	.447	.376	.286	.452	.355	.306		
	Z	-.435	-.283	-.292	-.411	-.318	-.291		

Table IV

Parameter Estimates (SE) for the three GEE models using CHAP data

Parameters	Available-case	Weighted	Multiple Weighted
Intercept	0.4245 (0.0137)	0.4158 (0.0146)	0.4610 (0.0170)
Age	-0.0485 (0.0012)	-0.0498 (0.0013)	-0.0493 (0.0013)
Black	-0.3803 (0.0159)	-0.3793 (0.0170)	-0.4083 (0.0194)
Male	-0.0946 (0.0138)	-0.0862 (0.0146)	-0.1267 (0.0169)
Education	0.0736 (0.0023)	0.0739 (0.0025)	0.0741 (0.0025)
Time	-0.0455 (0.0030)	-0.0392 (0.0032)	-0.0530 (0.0014)
Age $\times$ Time	-0.0034 (0.0003)	-0.0025 (0.0003)	-0.0026 (0.0001)
Black $\times$ Time	-0.0005 (0.0034)	-0.0014 (0.0038)	0.0054 (0.0017)
Male $\times$ Time	0.0058 (0.0029)	0.0073 (0.0031)	0.0173 (0.0015)
Education $\times$ Time	-0.0008 (0.0004)	-0.0011 (0.0005)	-0.0021 (0.0002)