

Estimating incidence of cognitive impairment from two-phase sampling with missing values: application of multiple imputation

CHANGYU SHEN^{1,2}

¹*Division of Biostatistics, School of Medicine, Indiana University, 1050 Wishard Boulevard RG R4101,*

Indianapolis, IN 46202, USA

²*Regenstrief Institute for Health Care, 1050 Wishard Boulevard, Indianapolis, IN 46202, USA*

E-mail: chashen@iupui.edu

Phone: 317-274-1641

Fax : 317-274-2678

SUMMARY

Cognitive impairment (CI) is an evolving concept that has been used to describe an intermediate stage between normal aging and dementia. In this paper, we attempt to estimate the incidence rate of CI based on an epidemiological cohort study that adopts a two-phase design. Such study design raises serious issue on how to treat a fairly large amount of missing values that are either MAR (due to the study design) or potentially MNAR (non-response and lost to follow-up). We develop a multiple imputation procedure in the mixture model framework to approach this problem. Sensitivity analysis is carried out to assess the dependence of the estimates on specific model assumptions.

Key words: Cognitive impairment; Incidence rate; MNAR; Multiple imputation.

1. Introduction

Cognitive impairment (CI) generally describes “a cognitive state intermediate between normal and dementia”, clinically suggesting a risk or prodromal state for Alzheimer’s disease (AD) and perhaps other dementias (Ganguli *et al.*, 2004; Luis *et al.*, 2003). Research in this condition has been an active area in the hope to seek effective early diagnosis and intervention of AD and other dementias. From the clinical point of view, CI represents the initial stage of disease progress that is characterized by more severe deterioration of various cognitive functioning than normal aging. In this perspective, subjects with CI provide psychiatrists and neurophysiologists invaluable information that can be used for research in the onset of abnormal neurodegeneration that ultimately results in dementia. CI is an evolving conception that requires epidemiological characterization for its further development. Although there have been an increasing number of epidemiological studies to investigate this intermediate state, a sound epidemiological basis of CI is still not complete. For instance, to our knowledge, no model-based analysis of the prevalence and incidence of CI can be found in the literature, even though several analyses of dementia have been published (Clayton *et al.*, 1998; Gao and Hui, 2000; Gao *et al.*, 2000). In this study, we attempt to estimate the incidence rate of CI using data collected from an African-American cohort in a community-based longitudinal study of dementia for African and native Americans—the Indianapolis-Ibaden Dementia Project.

In longitudinal epidemiological studies where subjects enrolled are followed at a series of time points (or data collection waves) for the examination of characteristics related to the disease or condition of interest, missing values always occur for various reasons. This phenomenon is more pronounced in dementia related cohort studies targeting on the elderly because the study subjects are more susceptible to illness or death. It is well known that the consequence of missing values for

analysis is potential bias in addition to reduced precision. Rubin (1976) defined two general classes of processes that lead to the missingness (missing-data processes), which lay out a theoretical framework to treat the problem of potential bias. Specifically, data are Missing At Random (MAR) when missing-data process does not depend on unobserved values conditional on the observed values. Data are Missing Not At Random (MNAR) when it is not MAR. In other words, data are MNAR when the missing-data process depends on the unobserved values conditional on observed data. It was shown that likelihood based approach ignoring the missing-data process provides valid inference for MAR data if the variables associated with the missing-data process is included in the model; and it can be potentially biased for MNAR data (Little and Rubin, 1987). Therefore, for MNAR data, we need to include the missing-data process in the analysis. The dilemma is that such analysis usually requires unverifiable assumptions such that incorrectly postulated assumptions can also lead to biased inference. For this very reason, a sensitivity analysis is usually required to examine the impact of various assumptions on the result of the analysis.

As described later, the Indianapolis-Ibaden Dementia Project used a two-phase design, which is often applied to studies where a disease is rare and the diagnosis of the disease is expensive. In the first phase, a large random sample from the targeted population is screened and stratified based on the results of the screening. In the second phase, a random subsample is selected within strata to receive formal clinical evaluation to determine the disease status. Such strategy has emerged as a cost-efficient way to obtain population characteristics on a disease, which would otherwise take much more time and resource to obtain. On the other hand, the design itself results in missing values because we do not observe the disease status for participants who are not selected for the diagnosis. In combination with factors leading to unobserved diseases status that is beyond the control of the study designer, missing

values arise as a crucial issue that needs careful treatment. Therefore, it is not a trivial task to estimate the incidence rate of CI from such a study. In particular, we will not be able to observe the disease status for subjects who were (a) not selected for clinical diagnosis; (b) selected but did not respond; (c) lost to follow-up. It is easy to see that (a) is MAR because subjects are randomly selected for diagnosis conditional on the stratum, which is fully observed. On the other hand, (b) and (c) can be potentially MNAR. Although typical likelihood based approaches under the selection model framework (Little, 1995) can be used to deal with MNAR data, it would require self-developed optimization algorithm to obtain the parameter estimates and their variance estimates. In this study, we instead undertake a multiple imputation technique in the framework of mixture models to estimate the incidence rates of CI. As shown in later sections, this procedure can be readily performed using standard software packages (e.g. SAS) and a sensitivity analysis is automatically carried out during the imputation.

This paper is organized as follows. In section 2 we briefly review the underlying rationale of multiple imputation. In Section 3, we provide details on how multiple imputation is used for the estimation of incidence of CI based on data collected from the Indianapolis-Ibaden Dementia Project. We conclude this paper with a discussion in Section 4.

2. Multiple imputation

Multiple imputation (Rubin, 1987) was originally proposed as a general technique to handle missing values in complex surveys and has proven to be valuable in many other settings as well (Rubin, 1996). This procedure repeatedly imputes the blanks in a data set with some value to create multiple “completed” data sets and standard statistical procedures are applied to each one of them. The final estimate is then obtained by combining estimate from each completed data set and variance is calculated to take into account both sampling variation and imputation variation (Rubin, 1987).

Compared with single imputation, multiple imputation successfully adjusts the underestimation of the variability due to uncertainty on missing values. This procedure is essentially set up in a Bayesian framework and has three components:

- i) Modeling task: a model assumption regarding the joint distribution of the outcome and the parameter vector;
- ii) Estimation task: estimation of the posterior distribution of the parameter vector given the observed values;
- iii) Imputation task: draw from the posterior distribution of the parameter vector and then draw from the conditional distribution of the unobserved values given the observed values and the parameter vector just drawn.

Step iii) is repeated to create multiple completed data sets. Although most applications of multiple imputation are for MAR data, it can also be used to handle MNAR data since a sensitivity analysis is automatically embedded within the procedure (Rubin, 1987). The advantage of this approach lies in the fact that it is easy to implement by most statistical software packages (e.g., PROC MI and PROC MIANALYZE in SAS).

Since the data we will deal with include values that are MAR and potentially MNAR, we will illustrate the basic rationale of multiple imputation in its general form, which takes the mixture model framework to impute values that are MNAR. For the sake of argument and simplicity, we will consider univariate scenario. To be specific, suppose the data are composed of n i.i.d. observations such that the i th observation includes a row vector of covariates (\mathbf{X}_i) and the outcome (Y_i), where Y_i is subject to missiness. We will use X to denote the covariate matrix that is fully observed and $\mathbf{Y}=(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ to denote the outcome vector, where \mathbf{Y}_{obs} denote the outcome values that are actually observed

and \mathbf{Y}_{mis} denotes the outcome values that are not observed. In addition, let \mathbf{R} be the missing-data indicator vector such that $R_i=1$ implies that Y_i is observed and $R_i=0$ implies that Y_i is missing. Finally, we will use *obs* to denote the set of observations such that $R_i=1$ for all $i \in \text{obs}$ and *mis* to denote the set of observations such that $R_i=0$ for all $i \in \text{mis}$.

The modeling task formulates distinct models of the conditional distribution of Y_i given X_i for $R_i=1$ and $R_i=0$; and the prior distribution of the parameters involved. We will denote the two models by $p(Y_i | \mathbf{X}_i, \boldsymbol{\theta}_1)$ (for $R_i=1$) and $p(Y_i | \mathbf{X}_i, \boldsymbol{\theta}_0)$ (for $R_i=0$), where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)$ is the parameter vector with a prior distribution $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_0 | \boldsymbol{\theta}_1)$. Such a mixture model structure reflects the fact that the two conditional distributions might be different due to MNAR. It has been used to handle data that are MNAR by a number of authors (Little, 1993, 1994; Little and Wang, 1996; Shen and Weissfeld, 2005). The major issue that arises is that $\boldsymbol{\theta}_0$ is not identifiable based on the observed data. Therefore, any assumption regarding $\boldsymbol{\theta}_0$ for the identification of the model is not testable without external information. A sensitivity analysis is usually used to examine the variation of the results across different assumptions due to the uncertainty of $\boldsymbol{\theta}_0$. In the multiple imputation setting, the prior distribution then reflects our belief on the distribution of $\boldsymbol{\theta}_0$. Since the posterior distribution of $\boldsymbol{\theta}_0$ given $\boldsymbol{\theta}_1$ will be the same as the prior (Rubin, 1987), the estimation task is essentially the calculation of the posterior distribution of $\boldsymbol{\theta}_1$, which is proportional to $p(\boldsymbol{\theta}_1) \prod_{i \in \text{obs}} p(Y_i | \mathbf{X}_i, \boldsymbol{\theta}_1)$. The imputation task will then be performed by the standard procedure.

In summary, the multiple imputation for univariate outcome that is potentially MNAR can be conducted as follows (with covariates):

- (i) Specify models $p(Y_i | \mathbf{X}_i, \boldsymbol{\theta}_1)$ for $R_i=1$ and $p(Y_i | \mathbf{X}_i, \boldsymbol{\theta}_0)$ for $R_i=0$; and a prior distribution

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_0 | \boldsymbol{\theta}_1);$$

(ii) Compute the posterior distribution of θ_1 as

$$p(\theta_1 | X, \mathbf{Y}_{\text{obs}}) = \frac{p(\theta_1) \prod_{i \in \text{obs}} p(Y_i | X_i, \theta_1)}{\int p(\theta_1) \prod_{i \in \text{obs}} p(Y_i | X_i, \theta_1) d\theta_1};$$

(iii) Draw θ_1^* from $p(\theta_1 | X, \mathbf{Y}_{\text{obs}})$;

(iv) Draw θ_0^* from $p(\theta_0 | \theta_1^*)$;

(v) Draw Y_i^* from $p(Y_i | X_i, \theta_0^*)$ for $i \in \text{mis}$;

(vi) Repeat (iii)-(v) m times to create m completed data sets.

Note that when the data are MAR, there will be only one model $p(Y_i | X_i, \theta)$; and the procedures proceed as above with θ_1 replaced by θ and θ_1^* (θ_0^*) replaced by θ^* (step (iv) is omitted).

3. Application to the Indianapolis-Ibaden Dementia Project

3.1. The Indianapolis-Ibaden Dementia Project

The Indianapolis-Ibaden Dementia Project is an on-going longitudinal study of dementia and Alzheimer's disease in the elderly starting 1992 (Hendrie *et al.*, 2001). The study participants are 2212 African Americans living in Indianapolis (U.S.A.) and 2494 native Africans living in Ibaden (Nigeria). All participants were 65 or older at enrollment. A population-based two-phase survey (Pickles *et al.*, 1995) was conducted at each data collection wave for reasons of cost efficiency and high probability of selecting diseased subjects. There was first an in-home screening using the Community Screening Interview for Dementia (CSID) (Hall *et al.*, 2000) that categorizes each subject into 3 performance groups (good, intermediate and poor) based on their screening scores. Then a full clinical assessment was performed for a random subsample of participants from each of the 3 groups with sampling rate 5%, 50% and 100%, respectively. In the clinical assessment phase, subjects are diagnosed as normal, cognitive impaired (CI), or demented. Then subjects diagnosed as normal will proceed to the CSID and

subjects diagnosed as CI will proceed directly to the clinical assessment phase without taking the CSID in the next data collection wave. Subjects diagnosed as dementia were excluded for further follow-up. The study is further complicated by two other features: i) subjects might not respond to the clinical diagnosis even if they were selected; ii) subjects may be lost to follow-up for various reasons between data collection waves. We illustrate the two-phase design in Fig 1, where dashed lines indicate that unobserved diagnosis would occur.

The primary aim of this study is to estimate the incidence rate of CI. However, as explained in the introduction, the disease status is not observed for a large number of subjects for various reasons. To be specific, there are three types of missing values involved: (a) subjects were not selected for the formal clinical diagnosis though their CSID scores are observed; (b) subjects were selected for the formal clinical diagnosis but did not respond; and (c) subjects were lost to follow-up so that neither CSID score nor clinical diagnosis was observed. Item (a) is MAR because the probability of being selected depends on the performance group, which is observed. On the other hand, items (b) and (c) are potentially MNAR since the missingness might depend on the unobserved disease status itself. In Section 3.2, we describe a multiple imputation method to estimate the incidence rate of CI to account for the missing values, using data collected from the African-Americans at Indianapolis.

3.2 Multiple imputation to estimate incidence rate of CI

We will use the baseline and the first follow-up data to estimate the incidence rate of CI. We first provide some notation for the explanation of the imputation procedure. To ease the notation, we suppress the subject index. First, let $\mathbf{Y} = (Y_1, Y_2)$ be the disease status indicator vector such that i) $Y_1 = 1$ if normal at baseline and $Y_1 = 0$ otherwise; ii) $Y_2 = 1$ if CI at the first follow-up and $Y_2 = 0$ otherwise. Note that subjects with $Y_2 = 0$ can be either normal or demented at the first follow-up.

However, the average time interval between baseline and the first follow-up is about 1.74 years, which is relatively short for a normal subject to transition to dementia. Therefore, subjects with $Y_1 = 1$ and $Y_2 = 0$ are predominately normal at the first follow-up. As a matter of fact, among the 51 subjects who were diagnosed as normal and had a diagnosis at the following wave, none of them developed dementia during the time period. Therefore, Y_2 essentially separates CI from normal subjects at the first follow-up for those who were normal at baseline.

We use $\mathbf{M} = (I_1, R_1, D, I_2, R_2)$ to characterize the missing-data pattern of \mathbf{Y} , where $I_k = 1$ implies selection for clinical diagnosis based on CSID performance group at wave k and $I_k = 0$ otherwise ($k=1$ (baseline), 2 (first follow-up)); $R_k = 1$ implies subject would respond if selected for clinical diagnosis at wave k and $R_k = 0$ otherwise; and $D = 1$ if not lost to follow-up and $D = 0$ otherwise. We use $\mathbf{Z} = (Z_1, Z_2)$ to denote the performance group at baseline and first follow-up, where $\mathbf{Z}_k = (Z_{k1}, Z_{k2})$ is a vector of two dummy variables representing “intermediate” and “poor” groups (“good” group is the baseline). Finally, we use \mathbf{X} to denote the covariate vector which includes baseline age (continuous and centered), age², sex (1: female; 0: male) and highest grade finished (continuous and centered).

We excluded 21 subjects who have missing values on education level, which leads to 2191 subjects in our analysis whose \mathbf{X} values and \mathbf{Z}_1 are fully observed. We show the missing-data pattern for the clinical diagnosis at baseline and first follow-up in Table 1, in which each row represents a specific missing-data pattern. Therefore, the largest group includes subjects who are not selected for diagnosis at either wave, though both \mathbf{Z}_1 and \mathbf{Z}_2 were observed for them. The first row (165 subjects) includes subjects who were diagnosed as CI or dementia so that their following diagnosis makes no contribution to the estimation of incidence of CI. Therefore, we do not characterize the

missing-data pattern for their following diagnosis. However, they will still contribute to the estimation of the prevalence of normal subjects at baseline, which is used to estimate the risk set of CI.

The quantity we attempt to estimate is $\Pr[Y_2 = 1 | Y_1 = 1]$. Imputation based procedures offer a straight forward strategy to approach this question since the estimation task is simple after all blanks have been filled. Multiple imputation further allows us to estimate the variance of such estimate, taking into account both sampling variation and uncertainty of missing values. In the mixture model framework, a model for $p(\mathbf{Y} | \mathbf{X}, \mathbf{Z}, \mathbf{M})$ is constructed to drive the imputation as explained in Section 2. In what follows, we describe how to impute missing values of Y_1 and Y_2 sequentially.

3.2.1 Imputation of missing values of Y_1

As seen from Table 1, there are two types of missing values for Y_1 : those caused by non-response ($I_1 = 1$ and $R_1 = 0$) and those caused by non-selection ($I_1 = 0$). To impute missing values of Y_1 , we will make the following assumption:

$$p(Y_1 | \mathbf{X}, \mathbf{Z}, \mathbf{M}) = p(Y_1 | \mathbf{X}, \mathbf{Z}_1, I_1, R_1). \quad (1)$$

In other words, we assume that the distribution of Y_1 does not depend on the performance group and missing-data pattern at first follow-up, conditional on the covariates, performance group and missing-data pattern at baseline. We further assume that

$$\Pr[Y_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1, R_1 = 1] = \text{logit}^{-1}[\alpha_{11} + (\mathbf{X}, \mathbf{Z}_1)\boldsymbol{\theta}_{11}] \quad (2)$$

$$\Pr[Y_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1, R_1 = 0] = \text{logit}^{-1}[\alpha_{10} + (\mathbf{X}, \mathbf{Z}_1)\boldsymbol{\theta}_{10}]. \quad (3)$$

With a non-informative prior, the posterior distribution of $(\alpha_{11}, \boldsymbol{\theta}_{11})$ is approximately normal with mean $(\hat{\alpha}_{11}, \hat{\boldsymbol{\theta}}_{11})$ and variance-covariance matrix \hat{V}_1 , where $(\hat{\alpha}_{11}, \hat{\boldsymbol{\theta}}_{11})$ is the maximum likelihood estimator (MLE) of $(\alpha_{11}, \boldsymbol{\theta}_{11})$ and \hat{V}_1 is the negative inverse of the second derivative of the log-likelihood function evaluated at $(\hat{\alpha}_{11}, \hat{\boldsymbol{\theta}}_{11})$. Since the missing diagnosis

caused by non-response is potentially MNAR, we take the same strategy as described in Section 2 to carry out a sensitivity analysis. Specifically, we assume that $(\alpha_{11}, \boldsymbol{\theta}_{11})$ and $(\alpha_{10}, \boldsymbol{\theta}_{10})$ are *a priori* functionally associated such that $\alpha_{10} = f(\alpha_{11})$ and $\boldsymbol{\theta}_{10} = \boldsymbol{\theta}_{11}$. Hence, the sensitivity analysis is carried out by varying the base probability of being normal (probability evaluated at $(\mathbf{X}, \mathbf{Z}_1) = \mathbf{0}$). Then steps (iii)-(v) in Section 2 are used to impute the missing values of Y_1 due to non-response.

To impute the missing values of Y_1 due to non-selection, note that the distribution of Y_1 does not depend on I_1 conditional on Z_1 . Therefore, the distribution of Y_1 for subjects not selected for diagnosis is simply a mixture of those who were selected and responded and those who were selected but did not respond. Specifically,

$$\begin{aligned} \Pr(Y_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 0) &= \Pr(R_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1) \Pr(Y_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1, R_1 = 1) \\ &+ (1 - \Pr(R_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1)) \Pr(Y_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1, R_1 = 0). \end{aligned} \quad (4)$$

We assume that:

$$\Pr(R_1 = 1 | \mathbf{X}, \mathbf{Z}_1, I_1 = 1) = \text{logit}^{-1}[(1, \mathbf{X}, \mathbf{Z}_1) \boldsymbol{\pi}_1], \quad (5)$$

where $\boldsymbol{\pi}_1$ is *a priori* independent of the α and $\boldsymbol{\theta}$'s. Again, with a non-informative prior, the posterior distribution of $\boldsymbol{\pi}_1$ is approximately normal with the mean and variance similarly computed as that of $(\alpha_{11}, \boldsymbol{\theta}_{11})$. The imputation follows easily after we obtain a draw from the posterior distribution of $\boldsymbol{\beta}_1 = (\alpha_{11}, \boldsymbol{\theta}_{11}, \alpha_{10}, \boldsymbol{\theta}_{10}, \boldsymbol{\pi}_1)$.

3.2.2 Imputation of missing values of Y_2

Obviously, we only need to impute the missing values of Y_2 for subjects with $Y_1 = 1$. Therefore, we need a model for $p(Y_2 | Y_1 = 1, \mathbf{X}, \mathbf{Z}, \mathbf{M})$, for which we make the following assumption:

$$p(Y_2 | Y_1 = 1, \mathbf{X}, \mathbf{Z}, \mathbf{M}) = p(Y_2 | Y_1 = 1, \mathbf{X}, \mathbf{Z}_2, D, I_2, R_2). \quad (6)$$

Hence, we assume that conditional on $Y_1 = 1$ and \mathbf{Z}_2 , the distribution of Y_2 does not depend on their

missing-data pattern at baseline and \mathbf{Z}_1 . Essentially, there are two general types of missing values for Y_2 , one not caused by lost to follow-up and one caused by lost to follow-up. Below we describe the imputation procedures for the two types of missing values.

Missing values not due to lost to follow-up

For missing values not caused by lost to follow-up, the reason is exactly the same as that of Y_1 : non-response and non-selection. Therefore, the procedure in Section 3.2.1 can be used to impute the missing values of Y_2 . Again, we will assume:

$$\Pr[Y_2 = 1 | Y_1 = 1, \mathbf{X}, \mathbf{Z}_2, D = 1, I_2 = 1, R_2 = 1] = \text{logit}^{-1}[\alpha_{21} + (\mathbf{X}, \mathbf{Z}_2)\boldsymbol{\theta}_{21}] \quad (7)$$

$$\Pr[Y_2 = 1 | Y_1 = 1, \mathbf{X}, \mathbf{Z}_2, D = 1, I_2 = 1, R_2 = 0] = \text{logit}^{-1}[\alpha_{20} + (\mathbf{X}, \mathbf{Z}_2)\boldsymbol{\theta}_{20}] \quad (8)$$

$$\Pr[R_2 = 1 | Y_1 = 1, \mathbf{X}, \mathbf{Z}_2, D = 1, I_2 = 1] = \text{logit}^{-1}[(1, \mathbf{X}, \mathbf{Z}_2)\boldsymbol{\pi}_2], \quad (9)$$

where $\boldsymbol{\beta}_2 = (\alpha_{21}, \boldsymbol{\theta}_{21}, \alpha_{20}, \boldsymbol{\theta}_{20}, \boldsymbol{\pi}_2)$ has the same prior distribution as $\boldsymbol{\beta}_1$ and is *a priori* independent of $\boldsymbol{\beta}_1$. One difficulty encountered is the estimation of the posterior distribution of $\boldsymbol{\beta}_2$. To update the distribution of $(\alpha_{21}, \boldsymbol{\theta}_{21})$, we need subjects who were diagnosed as normal at baseline and diagnosed at the first follow-up. As seen in Table 1, only 51 subjects meet this condition (second row). It turns out that 5 developed CI among these people. Similarly, to update the distribution of $\boldsymbol{\pi}_2$, we need subjects who were diagnosed as normal and were selected for diagnosis at the first follow-up. The number of subjects available is 55 with 4 non-responses (Table 1, second and third rows). These numbers are rather small and provide limited information on the parameter vector $(\alpha_{21}, \boldsymbol{\theta}_{21}, \boldsymbol{\pi}_2)$. Since non-response missingness contributes the major part to the missing values of Y_2 (over 80%), the estimation of incidence rate can be very instable if the process of imputation is driven by variable values on these 55 people.

We decide to enrich the set of subjects used to update the distribution of $(\alpha_{21}, \boldsymbol{\theta}_{21}, \boldsymbol{\pi}_2)$. A closer

look at Table 1 suggests that 170 subjects (**group 1**) were respondents at the first follow-up and were not diagnosed at baseline (the 6th and the 10th rows) and 87 subjects (**group 2**) were non-respondents at first follow-up and not diagnosed at baseline (7th and 11th rows). Therefore, group 1 can be used to update the distribution of $(\alpha_{21}, \theta_{21})$ and group 1 and 2 together can be used to update the distribution of π_2 , if we could identify subjects in these two groups who were normal at baseline. Since we did not observe their disease status at baseline, we look at their predicted probability of being normal based on the model in Section 3.2.1. Then we select subjects with high probability of being normal and mark them as $Y_1 = 1$. The set of selected subjects will be called “**additional set**”. Together with the original 55 subjects who were diagnosed as normal at baseline, these subjects serve as the “**enriched set**” used to update the distribution of $(\alpha_{21}, \theta_{21}, \pi_2)$. The details are provided in Section 3.3.

Missing values due to lost to follow-up

For subjects who were lost to follow-up, we did not observe (Z_2, I_2, R_2) , which makes it impossible to impute the missing values of Y_2 using a similar model as (6). The various reasons that lead to the lost to follow-up might have different implications on the value of Y_2 . For instance, subjects who were too busy to be interviewed might be more likely to be cognitive intact as compared with subjects who were too sick to be interviewed. Hence, such missingness is potentially MNAR with heterogeneous missing-data processes. Intuitively, separate models should be constructed for each of the various processes. Nevertheless, sensitivity analysis is still needed due to the non-ignorable nature of the missingness. Since lost to follow-up only contributes less than 20% to the missing values of Y_2 , instead of constructing models to distinguish the various reasons of missingness, we instead use a simple method based on the observed Y_2 values only, which allows direct sensitivity analysis. We believe the range of uncertainty about the missing values in the sensitivity analysis sufficiently covers

what could have been for the unobserved values. Specifically, we divide the age of the cohort into three categories: 65-74, 75-84 and 85 and over (85+). For respondents (at the first follow-up) in the enriched set, we observe for each age group the number of subjects who were CI at the first follow-up (n_{CI}) and the number of subjects who were not CI at the first follow-up (n_{NCI}). The posterior distribution of the incidence rate of CI for these people is then a beta distribution with $p = n_{CI} + 1$ and $q = n_{NCI} + 1$ as the parameters, assuming a flat prior (uniform distribution) of the incidence rate. Then we can use these distributions as our reference distributions to impute the missing values of Y_2 due to lost to follow-up. Potentially, the incidence rate of subjects who were lost to follow-up is different from the incidence rates of the respondents (at first follow-up) in the enriched set. To carry out the sensitivity analysis, we will include for each age group an adjustment term s ($s < q/p$) so that the posterior distribution of the incidence rate for subjects who were lost to follow-up is a beta distribution with $p + ps$ and $q - ps$ as the parameters. Hence, s basically is the percentage increase/decrease of the incidence rate as compared with the reference incidence rates. The imputation procedure for each age group is then composed of drawing incidence rate from the above beta distribution and drawing from a Bernoulli distribution with probability of success being the incidence rate just drawn for each subject in the age group.

3.2.3 Summary

In summary, the multiple imputation procedure for estimation of incidence rate of CI proceeds as follows:

- (i) Impute the missing values of Y_1 as described in Section 3.2.1,
- (ii) Impute the missing values of Y_2 as described in Section 3.2.2 for subjects with $Y_1 = 1$,
- (iii) Repeated (i) and (ii) m times to obtain m completed data sets.

We set $m=10$ in the analysis.

There are two components in the sensitivity analysis to assess the impact of various assumptions regarding the missingness on the results. The first one is the f function in Section 3.2.1, which is used to tune the base level of the prevalence of normal subjects at baseline and base level of incidence rate of CI for non-respondents as compared with respondents. We take a linear form of f (e.g. $\alpha_{10} = k\alpha_{11}$) and consider three scenarios by alternating the values of k in the analysis: (a) at the base level, prevalence of normal subjects among non-respondents is 10% higher than that of the respondents and the incidence rate of CI among non-respondents is 20% less than that of the respondents (non-respondents are healthier than respondents); (b) at the base level, non-respondents and respondents have the same prevalence of normal subjects and incidence rate of CI (non-respondents are equally healthy as respondents); and (c) at the base level, prevalence of normal subjects among non-respondents is 10% less than that of the respondents and the incidence rate of CI among non-respondents is 20% higher than that of the respondents (non-respondents are less healthy than respondents). The second component is the quantity s in Section 3.2.2, which is used to tune the incidence of CI for subjects who were lost to follow-up as compared with the respondents (at the first follow-up) in the enriched set. We consider three values of s and set the same s for each age group: -40% (those who were lost to follow-up has smaller incidence rate), 0 (same incidence rate), 40% (those who were lost to follow-up has greater incidence rate).

3.3 Results

We first fit the logistic models (2) and (5) to the baseline data to obtain the MLEs of the parameters. The results are shown in Table 2 (parameters with p values greater than 0.1 are not included in the models). Therefore, subjects with younger age, higher education level or “good” CSID performance are

more likely to be normal at baseline; and males or subjects with “poor” CSID performance are more likely to respond to the clinical diagnosis at baseline. To create the addition set in Section 3.2.2, we assume the missing values due to non-response at baseline is MAR (f is an identity function) and select all subjects in group 1 and 2 whose predicted probability of being normal is greater than 0.8. This leads to 161 subjects being selected and the enriched set is then composed of 216 subjects. Among these people, 19 and 150 are diagnosed as CI and normal at the first follow-up, respectively, with the rest 47 being non-respondents. Then models (7) and (9) are fitted to the data of these subjects and the results are shown in Table 3. It appears that males, subjects with lower education level or subjects with intermediate CSID performance are more likely to develop CI, though the evidence of sex effect is not as strong as the other two. In addition, subjects with median age range seem to have higher incidence rate than very young or old elders since the coefficient for age^2 is significantly less than 0.

Then the sensitivity analysis is initiated to derive the models of Y for non-respondents, non-selected and those who were lost to follow-up, followed by the multiple imputation procedure described in Section 3.2.3. The final estimates of incidence rates were calculated as the estimated $\Pr[Y_2 = 1 | Y_1 = 1]$ based on multiple completed data sets divided by the mean follow-up time for each age group (65-74: 1.77 years, 75-84: 1.73 years, 85+: 1.65 years, Total: 1.74 years). In Table 4, we show the estimated incidence rate of CI for each age group (65-74, 75-84, and 85+) under different assumptions. Clearly, age group 75-84 has much higher incidence rate than the other two groups, which is due to the quadratic term of age in equation (7). One possible explanation is that people who are normal at age 65-74 are less susceptible to cognitive impairment as compared with age group 75-84; and people who are normal at age 85 or older might be intrinsically superior to the normal population at age 75-84 and therefore have lower incidence rate. Since fewer people in the 85+ group are included

in the study as compared with the other two groups, the standard errors of the estimates for in this group are much greater than the other two.

Most estimates of the incidence rate display some fluctuation under various assumptions, though not substantial. The major difference occurs under age group 85+ when the assumption regarding the incidence rate of those who were lost to follow-up varies, which is due to the limited number of subjects in this group. Note that 40% increase/decrease assumption is rather dramatic and we believe it well covers the true difference in reality.

The extra 161 subjects included in the enriched set were selected based on their predicted probability of being normal (greater than 0.8) at baseline. Hence, it is possible that some of them were actually not normal at baseline. To assess the sensitivity of the results to this assumption, we re-selected 85 subjects based on a critical value of 0.84, leading to an enriched set of size 140. The results are shown in Table 5, which is quite similar to Table 4.

4. Discussion

In this paper we applied multiple imputation approach to estimate the incidence rate of CI using data that are subject to missingness due to study design and factors beyond the control of the investigators. The uniqueness of such data lies in the fact that some data are MAR whereas others are potentially MNAR. Multiple imputation under the mixture modeling framework provides a computationally-efficient and straight forward tool to handle such a problem. All computation is conducted in SAS 9, in which PROC MIANALYZE is used to combine estimates from each imputed data set.

Incidence rate of disease is a fundamental epidemiological quantity that can only be estimated by

large scale longitudinal studies. For rare disease like AD, such studies can be very expensive since a large number of subjects need to be recruited and followed. Although a two-phase design provides a cost-effective alternative, the missing-data problem that arises poses another challenge for valid statistical inference. This is because the non-response and lost to follow-up severely reduce the available diagnosis information, which is already limited due to the missing-values generated by the design. To our knowledge, this is the first attempt to apply multiple imputation to a complex survey for the estimation of incidence rate. It lays out a general strategy that can be potentially used in any epidemiological studies with similar design. In addition, our experience suggests that 10 imputations will have more than 90% efficiency as compared with infinite number of imputations. Since the proposed approach can be easily implemented in many standard software packages, it achieves analytical simplicity at a small price of efficiency loss.

There are two issues we want to give extra explanation. First, as mentioned previously in the paper, subjects who were lost to follow-up represent quite a heterogeneous group, whose missing values are results of various missing-data process. Since these subjects only contribute 20% to the total missing values, we collapse the different missing-data processes and treat them as one pattern in the framework of mixture model for multiple imputation. We believe such a simplification armed with sensitivity analysis is sufficient to assess the contribution of these values to the estimation of CI incidence rate. Second, CI based on current definition is not a stable condition like AD—people can go back to normal from CI (Ganguli, *et al.*, 2004). This means that some incidence cases might already go back to normal at the follow-up investigation. In addition, subjects who were normal at baseline and demented at the first follow-up might already went through the CI stage in the time interval. Nevertheless, they are not counted as incidence cases in the analysis. Therefore, our data and method tend to yield

under-estimates of the incidence rate and estimates presented in Table 4 and 5 should be understood as the lower bound for the incidence rate of CI. Nevertheless, to our knowledge, it provides the first model-based estimate of the incidence of CI.

Our work also opens some improvement possibilities in terms of study design to accommodate potential missing data. To estimate the incidence of CI, we need to enrich the number of subjects whose data are used to generate the imputation models since very few people were diagnosed as normal at baseline and were also diagnosed at the first follow-up. One way to improve the situation is to invite all people who were diagnosed as normal at baseline for a diagnosis at the first follow-up. From Table 1, it can be seen that this would include an extra 102 subjects (4th row in Table 1, suppose all of them will respond). The feasibility and properties of such design need further investigation.

Acknowledgment

This research was supported by NIH grants: R01 AG15813, R01 AG09956 and P30 AG10133. We would like to thank Dr. Sujuan Gao for critical comments on the manuscript.

References

- Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998) Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society Series B*, **60**, 71-87.
- Ganguli, M., Dodge, H. H., Shen, C. and DeKosky, S. T. (2004) Mild cognitive impairment, amnesic type: an epidemiologic study. *Neurology*, **63**, 115-21.

- Gao, S. and Hui, S. L. (2000) Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. *Stat Med*, **19**, 1545-54.
- Gao, S., Hui, S. L., Hall, K. S. and Hendrie, H. C. (2000) Estimating disease prevalence from two-phase surveys with non-response at the second phase. *Stat Med*, **19**, 2101-14.
- Hall, K. S., Gao, S., Emsley, C. L., Ogunniyi, A. O., Morgan, O. and Hendrie, H. C. (2000) Community Screening Interview for Dementia (CSI'D'); Performance in Five Disparate Study Sites. *International Journal of Geriatric Psychiatry*, **15**, 521-531.
- Hendrie, H. C., Ogunniyi, A., Hall, K. S., Baiyewu, O., Unverzagt, F. W., Gureje, O., Gao, S., Evans, R. M., Ogunseyinde, A. O., Adeyinka, A. O., Musick, B. and Hui, S. L. (2001) Incidence of Dementia and Alzheimer Disease in 2 Communities. *Journal of the American Medical Association*, **285**, 739-747.
- Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125-134.
- Little, R. J. A. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471-483.
- Little, R. J. A. (1995) Modeling drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112-1121.
- Little, R. J. A. and Rubin, D. B. (1987) *Statistical analysis with missing data*. New York: Wiley.
- Little, R. J. A. and Wang, Y. (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, **52**, 98-111.
- Luis, C. A., Loewenstein, D. A., Acevedo, A., Barker, W. W. and Duara, R. (2003) Mild cognitive impairment: directions for future research. *Neurology*, **61**, 438-44.

- Pickles, A., Dunn, G. and Vazquez-Barquero, J. L. (1995) Screening for Stratification in Two-Phase ("Two-Stage") Epidemiological Survey. *Statistical Methods in Medical Research*, **4**, 73-89.
- Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.
- Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Rubin, D. B. (1996) Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, **91**, 473-489.
- Shen, C. and Weissfeld, L. (2005) Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. *Biostatistics*, **6**, 333-47.

Fig 1. Schematic representation of the two-phase study (dashed lines indicate where missing values of the clinical diagnosis occur). SNR: subjects who were **selected** for Phase II (clinical diagnosis) but did **not respond**; SR: subjects who were **selected** for Phase II and **responded**; NS: subjects who were **not selected** for Phase II.

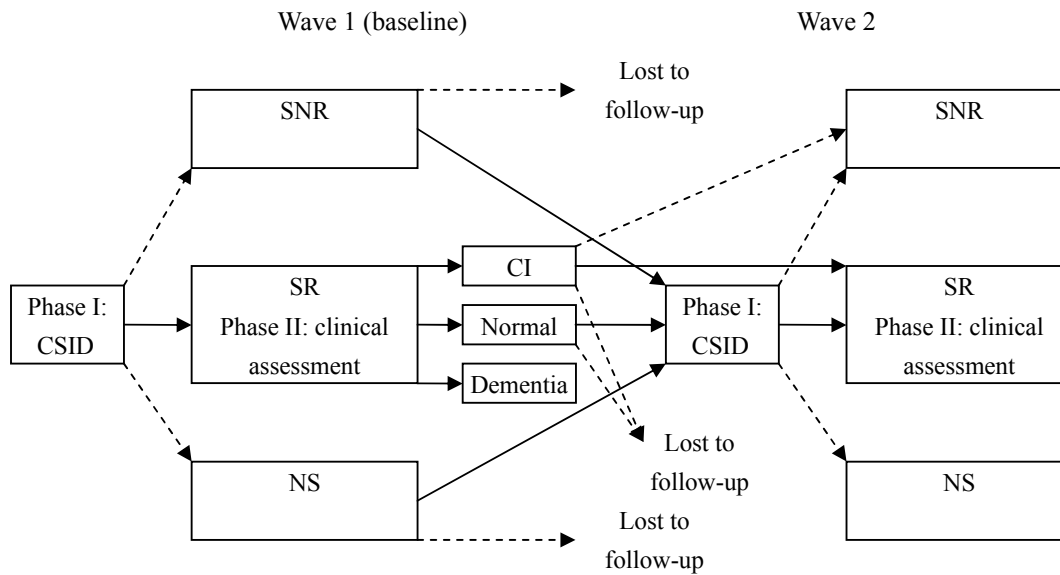


Table 1. Missing-data pattern for clinical diagnosis at baseline and first follow-up (follow-up 1).
 SNR: subjects who were **selected** for Phase II (clinical diagnosis) but did **not respond**; SR: subjects who were **selected** for Phase II and **responded**; NS: subjects who were **not selected** for Phase II.

baseline diagnosis	follow-up 1 diagnosis	I_1	R_1	D	I_2	R_2	# (%)
Diagnosed as CI or dementia at baseline							165 (7.4%)
SR	SR	1	1	1	1	1	51 (2.3%)
	SNR	1	1	1	1	0	4 (0.2%)
	NS	1	1	1	0		102 (4.7%)
	LTF	1	1	0			22 (1.0%)
SNR	SR	1	0	1	1	1	13 (0.6%)
	SNR	1	0	1	1	0	31 (1.4%)
	NS	1	0	1	0		123 (5.6%)
	LTF	1	0	0			81 (3.7%)
NS	SR	0		1	1	1	157 (7.2%)
	SNR	0		1	1	0	56 (2.6%)
	NS	0		1	0		1121 (51.2%)
	LTF	0		0			265 (12.1%)
Total							2191 (100%)

Table 2. Parameter estimates, standard errors (S.E.) and *p* values for model (2) and (5)

parameters	Prevalence of normal subjects among respondents at baseline (model (3.2))			Prevalence of respondents among subjects selected (model (3.5))		
	estimate	S.E.	<i>p</i>	estimate	S.E.	<i>p</i>
intercept	1.53	0.27	<0.0001	0.38	0.16	0.02
age	-0.052	0.017	0.002			NS
age ²			NS*			NS
sex			NS	-0.47	0.18	0.009
grade	0.08	0.038	0.035			NS
intermediate	-0.78	0.36	0.031			NS
poor	-2.02	0.31	<0.0001	0.63	0.17	0.0003

*: Not Significant

Table 3. Parameter estimates, standard errors (S.E.) and *p* values for model (7) and (9) based on the 216 subjects in the enriched set

parameters	Prevalence of normal subjects among respondents at baseline (model (3.7))			Prevalence of respondents among subjects selected (model (3.9))		
	estimate	S.E.	<i>p</i>	estimate	S.E.	<i>p</i>
intercept			NS	1.28	0.16	<0.0001
age			NS			NS
age ²	-0.023	0.011	0.039			NS
sex	-0.90	0.54	0.096			NS
grade	-0.15	0.041	0.0002			NS
intermediate	1.79	0.59	0.0026			NS
poor			NS			NS

Table 4. Incidence of CI (per 1000 person years) and standard errors under various assumptions (216 subjects in enriched set)

incidence rate of lost to follow-up as compared with the incidence rate of respondents in the enriched set	Age group	prevalence: non-respondents 10% more likely to be normal	prevalence: non-respondents equally likely to be normal	prevalence: non-respondents 10% less likely to be normal
		incidence: non-respondents 20% less likely to be CI	incidence: non-respondents equally likely to be CI	incidence: non-respondents 20% more likely to be CI
40% increase	65-74	49 (10)	50 (12)	55 (8)
	75-84	70 (16)	69 (15)	82 (12)
	85+	38 (35)	29 (16)	31 (21)
	Total	55 (10)	55 (11)	62 (7)
equal	65-74	46 (11)	49 (10)	50 (9)
	75-84	66 (12)	66 (13)	72 (12)
	85+	24 (24)	28 (24)	25 (18)
	Total	51 (10)	53 (10)	55 (8)
40% decrease	65-74	44 (9)	44 (13)	47 (8)
	75-84	56 (8)	52 (18)	63 (15)
	85+	20 (19)	14 (10)	16 (14)
	Total	46 (7)	45 (13)	50 (7)

Table 5. Incidence of CI (per 1000 person years) and standard errors under various assumptions (140 subjects in enriched set)

incidence rate of lost to follow-up as compared with the incidence rate of respondents in the enriched set	Age group	prevalence: non-respondents 10% more likely to be normal	prevalence: non-respondents equally likely to be normal	prevalence: non-respondents 10% less likely to be normal
		incidence: non-respondents 20% less likely to be CI	incidence: non-respondents equally likely to be CI	incidence: non-respondents 20% more likely to be CI
40% increase	65-74	52 (9)	49 (13)	57 (16)
	75-84	85 (15)	82 (17)	87 (19)
	85+	31 (15)	31 (19)	29 (23)
	Total	61 (9)	58 (12)	64 (15)
equal	65-74	46 (13)	44 (13)	50 (13)
	75-84	67 (13)	70 (11)	76 (13)
	85+	24 (16)	20 (13)	20 (14)
	Total	51 (12)	50 (10)	56 (11)
40% decrease	65-74	41 (12)	41 (13)	48 (12)
	75-84	62 (13)	63 (17)	67 (12)
	85+	17 (11)	19 (14)	20 (14)
	Total	46 (10)	46 (13)	52 (9)