Journal of Data Science **5**(2007), 503-518

Application of Multiple Imputation to Data from Two-phase Sampling: Estimation of the Incidence Rate of Cognitive Impairment

Changyu Shen
Indiana University and Regenstrief Institute for Health Care

Abstract: Epidemiological cohort study that adopts a two-phase design raises serious issue on how to treat a fairly large amount of missing values that are either Missing At Random (MAR) due to the study design or potentially Missing Not At Random (MNAR) due to non-response and loss to follow-up. Cognitive impairment (CI) is an evolving concept that needs epidemiological characterization for its maturity. In this work, we attempt to estimate the incidence rate CI by accounting for the aforementioned missing-data process. We consider baseline and first follow-up data of 2191 African-Americans enrolled in a prospective epidemiological study of dementia that adopted a two-phase sampling design. We developed a multiple imputation procedure in the mixture model framework that can be easily implemented in SAS. Sensitivity analysis is carried out to assess the dependence of the estimates on specific model assumptions. It is shown that African-Americans in the age of 65-75 have much higher incidence rate of CI than younger or older elderly. In conclusion, multiple imputation provides a practical and general framework for the estimation of epidemiological characteristics in two-phase sampling studies.

Key words: Cognitive impairment, incidence rate, MAR, MNAR, mixture-model, multiple imputation.

1. Introduction

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 (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.

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, Dodge, Shen and DeKosky, 2004; Luis, Loewenstein, Acevedo, Barker and Duara, 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 concept 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, very few analysis of the prevalence and incidence of CI can be found in the literature (Ritchie, Artero and Touchon, 2001), though several analyses of dementia have been published (Clayton, Spiegelhalter, Dunn and Pickles, 1998; Gao and Hui, 2000; Gao, Hui, Hall and Hendrie, 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. 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. 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 a fairly large amount of missing values due to study design and factors beyond

the control of the study designer.

Practical solutions are desirable for the estimation of epidemiological characteristics of diseases (e.g. incidence) that can account for various missing-data processes. With CI as an example, we undertake a multiple imputation technique in the framework of mixture models to estimate the incidence rate. 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. Moreover, it provides a general framework for the estimation of many other epidemiological quantities.

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). Compared with single imputation, it successfully adjusts the underestimation of the variability due to uncertainty on missing values. 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).

Essentially, multiple imputation repeatedly imputes the blanks in a data set with some value to create multiple "completed" data sets. This procedure is set up in a Bayesian framework and composed of three components:

i) Modeling task: a model assumption regarding the distribution of the complete data (e.g. logistic regression model for binary data) and the prior distribution of the model parameters;

ii) Estimation task: estimation of the posterior distribution of the parameters 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. 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). Specifically, if $\hat{Q}_i$ is the point estimate from the ith completed data set with an estimated variance $\hat{U}_i, i = 1, \ldots, m$ then the combined estimate $\bar{Q}$ is computed as

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{2.1}$$

Moreover, define the within-imputation variance $\bar{U}$ and between-imputation variance B as

$$\begin{aligned}\bar{U} &= \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \\ B &= \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2,\end{aligned}$$

then an estimate of the variance of $\bar{Q}$ can be computed as

$$T = \widehat{Var}(\bar{Q}) = \bar{U} + \left(1 + \frac{1}{m}\right)B. \tag{2.2}$$

3. 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, Ogunniyi, Hall, Baiyewu, Unverzagt, Gureje, Gao, Evans, Ogunseyinde, Adeyinka, Musick and Hui, 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, Dunn and Vazquez-Barquero, 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, Gao, Emsley, Ogunniyi, Morgan and Hendrie, 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 Figure 1, where dashed lines indicate that unobserved diagnosis would occur.

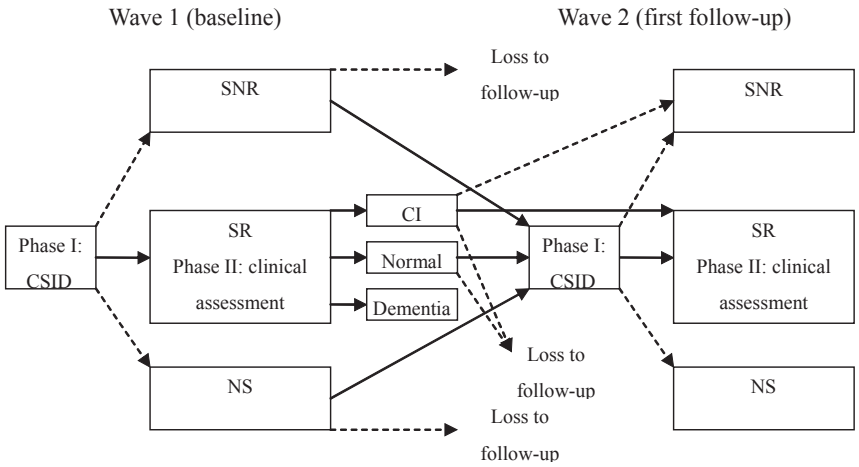

Figure 1: Schematic representation of the two-phase study (dashed lines indicate where missing values of the clinical diagnosis occur). SNR: subjects who were selected for Phase II (clinical diagnosis) but did not respond; SR: subjects who were selected for Phase II and responded; NS: subjects who were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

The primary aim of this paper is to estimate the incidence rate of CI by accounting for the missing diagnosis. 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 4, 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.

4. Estimation of CI Incidence Rate

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. For the sake of simplicity, we use the same notations for baseline and first follow-up. First, let Y be the binary variable that records the diagnosis result. Since we are interested in the estimation of incidence rate, we want to identify normal subjects at baseline and those who have developed CI at first-follow up among the normal subjects. Hence, for baseline, we define $Y = 1$ if

normal and 0 otherwise; for first follow-up, we define $Y = 1$ if CI and 0 otherwise. Next, we denote R as the response indicator for the formal diagnosis (phase II) such that $R = 1$ implies response if selected and 0 otherwise. Finally, we use X to denote the covariate vector which includes baseline age (continuous and centered), age^2, sex (1: female; 0: male) and highest grade finished (continuous and centered), and CSID group (poor, intermediate or good).

We excluded 21 subjects who have missing values on education level, which leads to 2191 subjects in our analysis whose X values 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 (1121 subjects). 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.

Table 1: Missing-data pattern for clinical diagnosis at baseline and first follow-up. 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; LTF: loss to follow-up.

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

To estimate the incidence rate, we need to impute the missing diagnosis values for subjects in the "selected but did not respond" (SNR), "not selected" (NS) and "loss to follow-up" (LTF) categories (Table 1). In doing so, we utilize a mixture-model framework, which essentially divides the whole population into a number of sub-populations based on the missing-data pattern and impute the missing-values for each of them separately. In what follows, we describe how to impute missing values at baseline and first follow-up sequentially in Section 4.1-4.3. We summarize the procedure and provide details of sensitivity analysis in Section 4.4.

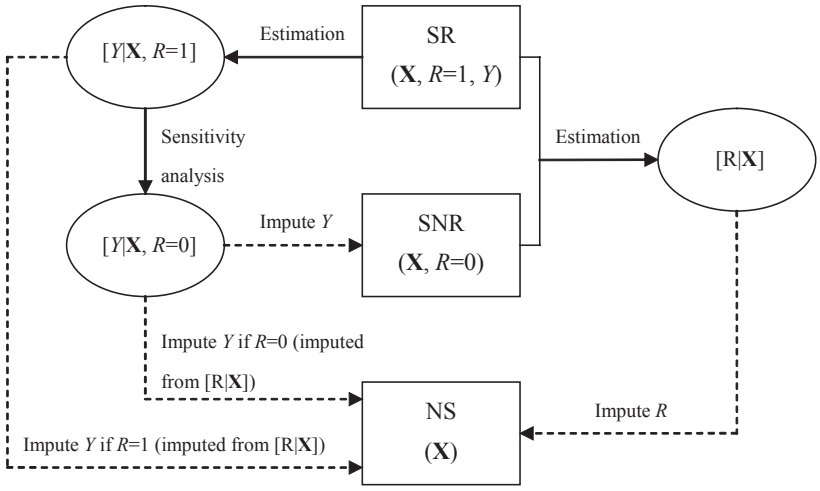

Figure 2: Strategy for the imputation of missing diagnosis for SNR and NS subjects. SNR: subjects who were selected for Phase II (clinical diagnosis) but did not respond; SR: subjects who were selected for Phase II and responded; NS: subjects who were not selected for Phase II.

4.1 Imputation of missing values at baseline

As seen from Table 1, there are two types of missing values for Y: those caused by non-response (SNR) and those caused by non-selection (NS). Since selection is based on the CSID group, which is part of the $\mathbf{X}$ vector, the NS sub-population is simply a mixture of SNR and SR sub-populations for a given $\mathbf{X}$ value. In other words, if we have a model to describe the diagnosis given $\mathbf{X}$ for SR ($[Y|\mathbf{X}, R = 1]$) and SNR ([Y—X, R=0]), and a model to describe the response given $\mathbf{X}$ ($[R|\mathbf{X}]$), we can impute the missing diagnosis for NS based on the model

$$[Y|\mathbf{X}] = [Y|\mathbf{X}, R = 1][R = 1|\mathbf{X}] + [Y|\mathbf{X}, R = 0][R = 0|\mathbf{X}]. \tag{4.1}$$

Here "[ ]" denote the probability distribution. Since we observe the diagnosis for subjects in SR category, $[Y|\mathbf{X}, R = 1]$ can be estimated. Similarly, $[R|\mathbf{X}]$ can be estimated based on data from subjects in SR and SNR categories. The major difficulty is the estimation of $[Y|\mathbf{X}, R = 0]$, which is not possible since we do not observe the Y. Hence, extra assumption is needed. One common strategy is to estimate $[Y|\mathbf{X}, R = 1]$ first and "estimate" $[Y|\mathbf{X}, R = 0]$ by adjusting $[Y|\mathbf{X}, R = 1]$ based on subjective belief and previous experience. This is where the sensitivity analysis is also required for the examination of the difference in results obtained by different model assumptions. We illustrate the overall strategy in Figure 2.

We use logistic regression to model and estimate $[Y|\mathbf{X}, R = 1]$:

$$\text{Logit}([Y = 1]) = \mathbf{X}\boldsymbol{\beta}_1. \tag{4.2}$$

If we assume a flat prior for the parameter vector$\boldsymbol{\beta}_1$, then the posterior distribution approximately follows a normal distribution with mean $\hat{\boldsymbol{\beta}}_1$ and variance-covariance matrix $\hat{\mathbf{v}}_1$, where $\hat{\boldsymbol{\beta}}_1$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}_1$ and $\hat{\mathbf{v}}_1$ is the negative inverse of the second derivative of the log-likelihood function evaluated at $\hat{\boldsymbol{\beta}}_1$. Similarly, we also use a logistic regression to model $[R|\mathbf{X}]$:

$$\text{Logit}([\text{R=1}]) = \mathbf{X}\boldsymbol{\lambda} \tag{4.3}$$

Computation of the posterior distribution of $\boldsymbol{\lambda}$ is the same as $\boldsymbol{\beta}_1$.

For $[Y|\mathbf{X}, R = 0]$, we postulate a logistic model with parameter vector $\boldsymbol{\beta}_2$. Since $\boldsymbol{\beta}_2$ is not estimable, we assume $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_1$, except that the intercepts part of the parameters are different. Hence, we assume that the data in SNR are potentially MNAR and different assumptions regarding the intercept of $\boldsymbol{\beta}_2$ serve as the sensitivity analysis. Since the intercept of the parameter vector controls the base level probability of being normal at baseline, the sensitivity analysis is set up in a way to tune this probability for subjects in SNR. Note that equal intercepts for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ imply MAR.

After we complete the estimation of the three models, $[Y|\mathbf{X}, R = 1]$, $[Y|\mathbf{X}, R = 0]$) and $[R|\mathbf{X}]$, we can impute the Y values for subjects in SNR and NS as shown in Figure 2.

4.2 Imputations of missing values in SNR and NS categories at first follow-up

Obviously, we only need to impute the missing values of Y at the first follow-up for subjects who were normal at baseline (either observed or imputed). For subjects in SNR and NS categories at the first follow-up, the imputation procedure is essentially the same as in the baseline except $\mathbf{X}$ is measured at the first

follow-up. However, this time, the models need to be estimated based on subjects who are normal at baseline. As seen in Table 1, only 51 subjects (second row) are diagnosed as normal at baseline and respond to the diagnosis at the first follow-up (these can be used to estimate $[Y|\mathbf{X}, R = 1]$). Similarly, only 55 subjects (sum of second and third rows) are normal at baseline and selected for diagnosis at the first follow-up (these can be used to estimate $[R|\mathbf{X}]$). These numbers are rather small and provide limited information on the parameter vector. Since non-response missingness contributes the major part to the missing values of diagnosis at the first follow-up (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 estimate the models in Figure 2 for the first follow-up. Specifically, we select subjects who were either SNR or SR at the first follow-up and have a high probability of being normal at baseline based on the estimated models in section 4.1. Together with the original 55 subjects who were diagnosed as normal at baseline, these subjects serve as the "**estimation set**" and are used to estimate $[Y|\mathbf{X}, R = 1]$ and $[R|\mathbf{X}]$.

4.3 Imputation of missing values due to loss to follow-up at the first follow-up

For subjects who were lost to follow-up, we did not observe their Y, R and part of the $\mathbf{X}$ values at the first follow-up, which makes it impossible to impute the missing values by using models in Figure 2. The various reasons that lead to the loss to follow-up might have different implications on the value of . 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 loss to follow-up only contributes less than 20% to the missing values of Y at the first follow-up, instead of constructing models to distinguish the various reasons of missingness, we instead use a simple method based on the observed Y 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 estimation 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 due to loss 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 estimation 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.

4.4 Summary

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

(i) Impute the missing values of Y for baseline as described in Section 4.1,

(ii) Impute the missing values of Y for the first follow-up as described in Section 4.2 and 4.3 for normal subjects at baseline (either observed or imputed)

(iii) Repeated (i) and (ii) m times to obtain m completed data sets.

We set $m = 10$ in the analysis. For each completed data set, the incidence rate is calculated as the proportion of the normal subjects at baseline who develop CI at the first follow-up. The final estimate and associated standard error are calculated based on equations (1) and (2). For more detailed model assumption and construction, see Supplemental Material.

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 relationship between the intercepts of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ in Section 4.1, which is used to tune the base level probability of being normal at baseline and base level probability of being CI at the first follow-up for non-respondents as compared with respondents. We consider three scenarios by alternating the values of the intercept in $\hat{\boldsymbol{\beta}}_2$: (a) at the base level, the probability of being normal for non-respondents is 10% higher than that of the respondents and the probability of being 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 probabilities of being normal

at baseline and CI at first follow-up (non-respondents are equally healthy as respondents); and (c) at the base level, the probability of being normal among non-respondents is 10% less than that of the respondents and the probability of being 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 4.2, 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 estimation 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).

Table 2: Parameter estimates, standard errors (S.E.) and p-values (baseline) for model (4.2) and (4.3)

Parameters	Probability of being normal for respondents (model (4.2))			Probability of being respondents (model (4.3))		
	estimate	S.E.	p	estimate	S.E.	p
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, #: "good" group as the baseline

5. Results

We first fit the logistic models (4) 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 estimation set in Section 4.2, in addition to the 55 subjects who were diagnosed as normal at baseline and were SR or SNR at the first follow-up, we assume the missing values due to non-response at baseline is MAR and select all subjects who are either SNR or SR at the first follow-up and whose predicted probability

of being normal at baseline is greater than 0.8. This leads to 161 subjects being selected and the estimation set is then composed of 216 subjects (161+55 = 216). Then models (4) and (5) 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.

Table 3: Parameter estimates, standard errors (S.E.) and p-values for model (4.2) and (4.3) based on the 216 subjects in the estimation set (first follow-up)

Parameters	Probability of being normal for respondents (model (4.2))			Probability of being respondents (model (4.3))		
	estimate	S.E.	p	estimate	S.E.	p
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

*: Not Significant, #: "good" group as the baseline

Then the sensitivity analysis is initiated to derive the models of Y for non-respondents, non-selected and those lost to follow-up, followed by the multiple imputation. The final incidence rates were calculated as the estimated rates 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 (Table 3). 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 because of younger age; 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.

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

(1)	Age group	(2)	(3)	(4)
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)

(1) Incidence rate of loss to follow-up as compared with the incidence rate of respondents in the estimation set.
(2) **Baseline**: non-respondents 10% more likely to be normal. **First follow-up**: non-respondents 20% less likely to be CI.
(3) **Baseline**: non-respondents equally likely to be normal. **First follow-up**: non-respondents equally likely to be CI.
(4) **Baseline**: non-respondents 10% less likely to be normal. **First follow-up**: non-respondents 20% more likely to be CI.

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 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 estimation 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 threshold of 0.84, leading to an estimation set of size 140. The results are shown in Table 5, which is quite similar to Table 4.

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

(1)	Age group	(2)	(3)	(4)
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)

(1) Incidence rate of loss to follow-up as compared with the incidence rate of respondents in the estimation set.
(2) **Baseline**: non-respondents 10% more likely to be normal. **First follow-up**: non-respondents 20% less likely to be CI.
(3) **Baseline**: non-respondents equally likely to be normal. **First follow-up**: non-respondents equally likely to be CI.
(4) **Baseline**: non-respondents 10% less likely to be normal. **First follow-up**: non-respondents 20% more likely to be CI.

6. Discussion

In this paper we applied multiple imputation approach to estimate the incidence rate of CI using a data set that is 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 loss 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 for the estimation of a number of quantities such as prevalence, incidence, life expectancy and so on. In addition, our data suggest 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 processes. 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. 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 Tables 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. In addition, our approach also provides a convenient tool to estimate the rate of turning back to normal for CI subjects.

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, amnestic type: An epidemiologic study. *Neurology* **63**, 115-121.

Gao, S. and Hui, S. L. (2000). Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. *Stat. Med.* **19**, 1545-1554.

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**, 2101-2114.

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. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Luis, C. A., Loewenstein, D. A., Acevedo, A., Barker, W. W. and Duara, R. (2003). Mild cognitive impairment: Directions for future research. *Neurology* **61**, 438-444.

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.

Ritchie, K., Artero, S. and Touchon, J. (2001). Classification criteira for mild cognitive impairment: A population-based validation study. *Neurology* **56**, 37-42.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*, 473-489.

Received April 1, 2005; accepted February 17, 2006.

Changyu Shen
Division of Biostatistics
School of Medicine, Indiana University
1050 Wishard Boulevard RG R4101
Indianapolis, IN 46202, USA
chashen@iupui.edu

Also: Regenstrief Institute for Health Care
1050 Wishard Boulevard
Indianapolis, IN 46202, USA