DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 769–785, http://dx.doi.org/10.1515/JOS-2016-0039

Nonrespondent Subsample Multiple Imputation in Two-Phase Sampling for Nonresponse

Nanhua Zhang[1], *Henian Chen*[2], *and Michael R. Elliott*[3]

Nonresponse is very common in epidemiologic surveys and clinical trials. Common methods for dealing with missing data (e.g., complete-case analysis, ignorable-likelihood methods, and nonignorable modeling methods) rely on untestable assumptions. Nonresponse two-phase sampling (NTS), which takes a random sample of initial nonrespondents for follow-up data collection, provides a means to reduce nonresponse bias. However, traditional weighting methods to analyze data from NTS do not make full use of auxiliary variables. This article proposes a method called nonrespondent subsample multiple imputation (NSMI), where multiple imputation (Rubin 1987) is performed within the subsample of nonrespondents in Phase I using additional data collected in Phase II. The properties of the proposed methods by simulation are illustrated and the methods applied to a quality of life study. The simulation study shows that the gains from using the NTS scheme can be substantial, even if NTS sampling only collects data from a small proportion of the initial nonrespondents.

Key words: Double sampling; maximum likelihood; missing data; nonignorable missing-data mechanism; quality of life; weighting.

1. Introduction

Nonresponse is very common in population surveys and clinical trials. Complete-case analysis (CC), which discards the incomplete cases, can lead to a substantial loss of information or biased estimation of the key parameters. Since the publication of Rubin's seminal paper on missing data (Rubin 1976), a number of ignorable-likelihood (IL) methods have been developed, including ignorable maximum likelihood, Bayesian inference, and multiple imputation (Dempster et al. 1977; Rubin 1987; Heitjan and Rubin 1991; Little and Zhang 2011). IL methods provide valid inference when missingness does not depend on the underlying missing values after conditioning on available data, a state termed missing at random (MAR) (Rubin 1976; Little and Rubin 2002). When MAR holds, inference can be based on the observed-data likelihood, and thus does not require modeling assumptions about the missingness indicators. When the missingness could depend on the missing values (missing not at random (MNAR) mechanism), nonignorable models (NIM) are developed based on the joint distribution of the variables and the

[1] Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, OH 45229, U.S.A. Email: nanhua.zhang@cchmc.org (corresponding author)
[2] Department of Epidemiology & Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612-3085, U.S.A. Email: hchen1@health.usf.edu
[3] Department of Biostatistics, School of Public Health, University of Michigan and Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48019, U.S.A. Email: mrelliot@umich.edu

missing-data indicators (Heckman 1976; Amemiya 1984; Little 1993, 1994; Nandram and Choi 2002, 2010).

Both IL and NIM methods make use of all available data, but they rely on assumptions about the missing-data mechanism. IL methods are vulnerable to failures of the ignorable missingness assumption; NIM methods are vulnerable to misspecification of the missing-data mechanism and suffer from problems with identifying parameters. The assumptions about the missing-data mechanism are untestable without knowing the underlying values of the missing data. The choice may be aided by learning more about the missing-data mechanism; for example, by recording reasons why particular values are missing. The difficulty in identifying parameters in NIM may be alleviated in some special cases, such as small-area estimation (Nandram and Choi 2002). In cases where the assumptions about the missing-data mechanism cannot be determined, an alternative strategy is to perform a sensitivity analysis to see whether key results are robust to alternative methods and assumptions.

Yet another alternative is to use a study design to relax to some degree the assumptions required under IL and NIM. One such design is two-phase sampling, in which a subsample of nonrespondents to the original survey (Phase I) is selected for further interview attempts (Phase II). This method is called nonresponse two-phase sampling (NTS). It was first proposed by Hansen and Hurwitz (1946) to reduce the nonresponse bias in mail questionnaires by carrying out personal interviews with a fraction of the nonrespondents. Discussions of sample-size selection and estimation of the population mean/total can be found in Hansen and Hurwitz (1946) and Srinath (1971). Some examples of using two-phase sampling to mitigate the effects of nonresponse include the National Comorbidity Survey (Elliott et al. 2000), the 2003 Survey of Small Business Finances (Harter et al. 2007) and the 2011 Canadian National Household Survey (Statistics Canada 2011).

Previous research mainly relies on using case weights developed from the two-phase sample, rather than auxiliary variables, to reduce bias in estimating population means or totals (Hansen and Hurwitz 1946; Srinath 1971; Harter et al. 2007). This article proposes nonrespondent subsample multiple imputation (NSMI), where multiple imputation (Rubin 1987) is performed within the subsample of nonrespondents in Phase I, using additional data collected in Phase II. The rationale of NSMI is that the MAR assumption, which the multiple-imputation method is based on, is valid within the nonrespondent subsample in Phase I, but may be invalid if extended to the whole sample. This is true when the missingness in Phase I is MNAR and the NSMI reduces the nonresponse bias; when the missingness in Phase I is ignorable, the NSMI is still a valid method, although there is some loss of efficiency compared with multiple imputation using all cases.

Section 2 presents a motivating application based on data from a quality of life (QOL) study. In this application, 147 out of the 750 participants did not reply to the initial QOL survey. In Phase II, all 147 nonrespondents were recontacted and 39 provided answers to an abridged version of the QOL instrument. The NSMI method consists of multiple imputation of the missing QOL outcomes within the subsample of nonrespondents in Phase I, that is, using the partial information of the 39 respondents in Phase II to impute the missing QOL data.

Section 3 introduces the framework of NTS and the necessary notation. Section 4 reviews the methods for analyzing data from NTS and proposes NSMI. Section 5 presents

simulations that illustrate the properties of NSMI, while Section 6 applies the method to the motivating data. Section 7 concludes with a discussion.

2. Motivating Problem: A Quality of Life Study

To illustrate the methods, data from 750 participants in a community-based study—the Children in the Community study (CIC) (Cohen et al. 2005)—are considered. The sample was based on a random residence-based cohort, originally drawn from 100 neighborhoods in two upstate New York counties in 1975. Additional information regarding the study is available from Cohen et al. (2005). From 1991 to 1994 (T1), 750 youths (mean age 22 years and SD 2.8 years) were interviewed in their homes by trained interviewers. QOL data were collected as part of the survey. QOL was assessed by the young adult quality of life instrument (YAQOL) (Chen et al. 2004). In 2001–2004 (T2) at mean age 33.0 years (SD = 2.8), the same group of participants was surveyed via the web using the same QOL instrument. Of the 750 subjects assessed for QOL at T1, 603 (80.4%) completed the QOL survey at T2; 147 did not respond to the follow-up survey. For these 147 subjects, an abridged version of the QOL instrument was mailed to their home address. Upon return of the completed surveys, subjects were paid for their participation. Of the 147 eligible subjects, 39 (26.5%) returned their QOL questionnaire. The resources scale used here is taken from the abridged version and identical to that employed at T1.

The goals of the QOL analysis included estimating the mean resources score and determining whether the resources scores are related to major demographic variables—age, gender, race and education. CC analysis suffers from inefficiency and potential bias if the missingness of QOL is MNAR. IL analyses make use of the partial information in the incomplete cases, but assume the missing data are MAR. NSMI is proposed for this problem, which is shown to be valid if the conditions of the Phase II sampling are met, regardless of the missing-data mechanism in Phase I.

3. Continuing Data Collection for Nonresponse

Data with the structure in Table 1 are considered. Let $\{\boldsymbol{y}_i, i = 1, \ldots n\}$ denote n independent observations on a (possibly multivariate) outcome variable $\boldsymbol{Y}$, where $\boldsymbol{Y}$ has missing values. $\boldsymbol{Y}_{obs,1}$ is used to represent the data observed in Phase I, $\boldsymbol{Y}_{obs,2}$ to represent the data missing in Phase I, but observed in Phase II, and $\boldsymbol{Y}_{mis}$ to represent the data missing after Phase II sampling. Let $\boldsymbol{Y}_{obs} = (\boldsymbol{Y}_{obs,1}, \boldsymbol{Y}_{obs,2})$ and $\boldsymbol{Y}_{mis,1} = (\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis})$ represent the observed data after Phase II and the missing data from Phase I, respectively. The vector of

Table 1. Two-phase sampling for nonresponse and general missing-data structure for Section 3.

Pattern	Observation, i	$\boldsymbol{y}_i$	$R_{1,i}$	$S_{2\cdot1,i}$	$R_{2\cdot1,i}$	$R_{2,i}$
1	$i = 1, \ldots, m$	√	1	–	–	1
2	$i = m+1, \ldots, m+r$	x	0	1	1	1
3	$i = m+r+1, \ldots, m+s$	?	0	1	0	0
4	$i = m+s+1, \ldots, n$	?	0	0	0	0

Key: √ denotes observed; ? denotes at least one entry missing; x denotes at least one entry missing in Phase I, but observed in Phase II.

covariates, z_i, is assumed to be fully observed. Interest concerns the parameters $\boldsymbol{\phi}$, which govern the conditional distribution of $\boldsymbol{y}_i$ on $z_i, p(\boldsymbol{y}_i|\boldsymbol{\phi}, z_i)$.

In Phase I, $\boldsymbol{y}_i$s are observed for $i = 1, \ldots, m$, but contain missing values for $i = m + 1, \ldots, n$. The response indicator for Phase I is denoted as $R_{1,i}$, equal to 1 if $\boldsymbol{y}_i$ is observed and 0 otherwise. In Phase II, s subjects were sampled from the nonrespondents in Phase I and r subjects responded. $S_{2\cdot1,i}$ is used to denote whether a subject was sampled among the nonrespondents in Phase I. Let π_i denote the Phase II sampling probability among nonrespondents in Phase I,

$$\pi_i = \Pr(S_{2\cdot1,i} = 1|R_{1,i} = 0; z_i, \boldsymbol{y}_i). \tag{1}$$

After Phase II sampling, data on r additional subjects were collected. $R_{2\cdot1,i}$ is used to denote the Phase II response indicator among the nonrespondents in Phase I. The overall response indicator after completion of Phase II is denoted as $R_{2,i}$. Depending on the context, the second-stage sampling may be a simple random, stratified or other probability sampling scheme. In certain settings such as this example, all nonrespondents may be contacted, with $\pi_i = 1$ for all i, so that $m + s = n$ and the fourth row in Table 1 is empty.

The rows of Table 1 divide the cases into four patterns. Pattern 1 ($i = 1, \ldots, m$) consists of subjects for whom $\boldsymbol{y}_i$ is fully observed after first-phase data collection. Pattern 2 consists of cases that were missing in Phase I, but subsequently observed in Phase II sampling. Pattern 3 consists of cases that were sampled in Phase II, but did not respond, and Pattern 4 were those Phase I nonrespondents were not sampled in Phase II.

4. A Comparison of Methods for Analyzing the Data

4.1. *Ignorable Likelihood Using Multiple Imputation (MI)*

In this subsection, data with the structure in Table 2 are considered. $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ are used to denote the observed and missing component of the data $\boldsymbol{Y}$, respectively. R_i is used to denote the response indicator, equal to 1 if $\boldsymbol{y}_i$ is observed and 0 otherwise. $\boldsymbol{Z}$ denotes the covariates that are fully observed. When the data contain missing values, the full model to describe the data is the joint distribution of $\boldsymbol{Y}_{obs}$, $\boldsymbol{Y}_{mis}$ and $\boldsymbol{R}$ conditional on $\boldsymbol{Z}, P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z})$, where $\boldsymbol{\xi}$ is the parameter associated with the distribution of the response indicator $\boldsymbol{R}$. The observed likelihood can be written as:

$$L(\boldsymbol{\phi}, \boldsymbol{\xi}|\boldsymbol{Y}_{obs}, \boldsymbol{R}; \boldsymbol{Z}) \propto P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) \tag{2}$$

Table 2. General missing-data structure for Subsection 4.1.

Pattern	Observation, i	$\boldsymbol{y}_i$	$\boldsymbol{R}_i$
1	$i = 1, \ldots, m$	$\surd$	1
2	$i = m + 1, \ldots, n$	?	0

Key: $\surd$ denotes observed; ? denotes at least one entry missing.

where

$$P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis}$$
$$= \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\xi}, \boldsymbol{Z}) d\boldsymbol{Y}_{mis}. \quad (3)$$

When the missing-data mechanism is missing completely at random (MCAR) or MAR (Little and Rubin 2002), (3) becomes

$$P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \begin{cases} P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{\xi}; \boldsymbol{Z}) & \text{if MCAR} \\ P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\xi}; \boldsymbol{Z}) & \text{if MAR.} \end{cases} \quad (4)$$

Under the further condition that the parameter spaces of $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are distinct, the likelihood-based inference on ϕ can be conducted based on $P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z})$, ignoring the missing-data mechanism:

$$L(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) \propto P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}). \quad (5)$$

Likelihood-based methods that ignore the missing-data mechanism are called ignorable likelihood (Little and Zhang 2011). Options for IL are maximum-likelihood estimation, Bayesian inference, and multiple imputation. Bayesian inference is based on the posterior distribution of $\boldsymbol{\phi}$ given by:

$$P(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) \propto L(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{\phi}), \quad (6)$$

where $P(\boldsymbol{\phi})$ is the prior distribution of $\boldsymbol{\phi}$.

Another option of IL is multiple imputation, that is, to impute the missing data $\boldsymbol{Y}_{mis}$, and then apply complete-data-based methods to the imputed data to make inference on the parameters $\boldsymbol{\phi}$. Multiple imputation is closely related to Bayesian inference. The imputation of $\boldsymbol{Y}_{mis}$ is based on the posterior predictive distribution of $\boldsymbol{Y}_{mis}$ given $\boldsymbol{Y}_{obs}$, which is the conditional predictive distribution, $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}; \boldsymbol{Z})$, averaged over the posterior distribution of $\boldsymbol{\phi}$, that is,

$$P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) d\boldsymbol{\phi}. \quad (7)$$

In order to generate M sets of imputations given $\boldsymbol{Y}_{obs}$, M values of $\boldsymbol{\phi}$ are independently drawn from the posterior distribution, say $\tilde{\boldsymbol{\phi}}^{(t)} (t = 1, \ldots, M)$. For each $\tilde{\boldsymbol{\phi}}^{(t)}$, one set of imputed values of $\boldsymbol{Y}_{mis}$ is obtained by taking a random draw of $\boldsymbol{Y}_{mis}$ from the corresponding posterior predictive distribution $P\left(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \tilde{\boldsymbol{\phi}}^{(t)}; \boldsymbol{Z}\right)$. Rubin (1987) showed that when the proper imputation method is followed (i.e., an imputation method that accounts for the uncertainty in the model parameters), the resulting inference based on the multiply imputed datasets is valid. The M imputed datasets are then analyzed as if each of them is a complete dataset. The analysis results from M imputed datasets are combined following the multiple-imputation combining rules (Rubin 1987).

Unlike IL methods, CC analysis discards all cases that contain missing values and is based on the following likelihood:

$$L_{\text{cc}}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | \boldsymbol{R} = 1; \boldsymbol{\phi}, \boldsymbol{Z}). \tag{8}$$

The estimation of $\boldsymbol{\phi}$ is obtained through maximizing $L_{\propto}(\boldsymbol{\phi})$; CC analysis is the default method in most statistical packages.

4.2. *Complete-Case and Ignorable-Likelihood Methods*

In this section, data with the structure in Table 1 are considered. The notation is the same as in Section 3. Depending on whether the additional data, $\boldsymbol{Y}_{obs,2}$ are used, there are two versions of complete-case analysis and ignorable-likelihood method (e.g., multiple imputation). The ignorable-likelihood methods that use data in Phase II (IL2) can be written as:

$$\begin{aligned} L_{\text{ign},2}(\boldsymbol{\phi}) &= P(\boldsymbol{Y}_{obs} | \boldsymbol{\phi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\phi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis} \\ &= \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | R_{1,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \times \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_i | R_{1,i} = 0, R_{2,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \\ &\times \int \cdots \int \prod_{i=m+r+1}^{n} p(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis} | R_{2,i} = 0; \boldsymbol{\phi}, \boldsymbol{Z}) d\boldsymbol{y}_{m+r+1,mis} \cdots d\boldsymbol{y}_{n,mis} \end{aligned} \tag{9}$$

where $\boldsymbol{y}_{i,obs}$ consists of the fully observed components of $\boldsymbol{y}_i$.

Rubin's (1976) theory shows that a sufficient condition for valid inference based on (9) is that MAR holds in the Phase II data, that is:

$$P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\xi}, \boldsymbol{Z}) = P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}; \boldsymbol{\xi}, \boldsymbol{Z}). \tag{10}$$

A complete-case analysis using Phase II data (CC2) bases inferences for $\boldsymbol{\phi}$ on the complete observations in Patterns 1 and 2. In a likelihood context, the method bases inference on the conditional likelihood corresponding to the complete cases after Phase II sampling, namely:

$$L_{\text{cc},2}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | R_{1,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \times \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_i | R_{1,i} = 0, R_{2,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}). \tag{11}$$

Note that the first part of (9) is exactly the same as (11), and that the second part explains how (9) uses the partially observed component of the outcome $\boldsymbol{y}_i$ (possibly multivariate) for $i = m+r+1, \ldots, n$. The key assumption under which inference based on $L_{\propto,2}(\boldsymbol{\phi})$ is valid is that the missingness after Phase II is MCAR,

$$P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\xi}, \boldsymbol{Z}) = P(\boldsymbol{R}_2 | \boldsymbol{\xi}, \boldsymbol{Z}). \tag{12}$$

Note that (12) is a special case of (10); when the missingness after Phase II is MCAR, the IL2 method is also valid and more efficient than CC2 because CC2 removes from the analysis all cases that were not observed after Phase II sampling, and fails to use the information in the partially observed data.

The ignorable-likelihood methods that use only the Phase I data (IL1) are based on the following likelihood:

$$\begin{aligned} L_{\text{ign},1}(\boldsymbol{\phi}) = P(\boldsymbol{Y}_{obs,1}|\boldsymbol{\phi},\boldsymbol{Z}) &= \int\int P(\boldsymbol{Y}_{obs,1},\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis}|\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{Y}_{obs,2}d\boldsymbol{Y}_{mis} \\ &= \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i|R_{1,i}=1;\boldsymbol{\phi},\boldsymbol{Z}) \\ &\times \int\cdots\int \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|R_{1,i}=0,R_{2,i}=1;\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{y}_{m+1,mis}\cdots d\boldsymbol{y}_{m+r,mis} \\ &\times \int\ldots\int \prod_{i=m+r+1}^{n} p(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|R_{2,i}=0;\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{y}_{m+r+1,mis}\cdots d\boldsymbol{y}_{n,mis} \end{aligned} \tag{13}$$

They are valid if:

$$P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs,1},\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs,1};\boldsymbol{\xi},\boldsymbol{Z}). \tag{14}$$

Likewise, the CC analysis based only on Phase I uses cases in Pattern 1 (CC1) in Table 1, and is equivalent to (8) when no resampling has occurred. Here, the likelihood can be written as:

$$L_{\text{cc},1}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p\big(\boldsymbol{y}_i|\boldsymbol{R}_1=1;\boldsymbol{\phi},\boldsymbol{Z}\big). \tag{15}$$

The CC1 analysis is valid if the corresponding missing-data mechanism is MCAR:

$$P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = P(\boldsymbol{R}_1|\boldsymbol{\xi},\boldsymbol{Z}). \tag{16}$$

In both cases with or without using data from Phase II, the ignorable-likelihood methods are more efficient than the CC analysis if the corresponding missing-data mechanisms are ignorable (MAR or MCAR). The choice between IL1 and IL2 relies on whether (10) or (14) is a more reasonable assumption, that is, whether the MAR assumption holds among second-wave nonrespondents regardless of the first-wave missingness mechanism (suggesting IL2), or whether MAR holds among first-wave nonrespondents, but missingness is nonignorable at Phase 2 (suggesting IL1).

4.3. Nonrespondent Subsample Multiple Imputation (NSMI)

In this section, data with the structure in Table 1 are also considered. NSMI is proposed, which applies the multiple-imputation method to the cases in Patterns 2, 3, and 4. In the NSMI method, we leave out the subjects in Pattern 1 when the missing values in Pattern 3 and 4 were imputed, and then the imputed datasets from Patterns 2, 3, 4 are combined with data from Pattern 1 for statistical analyses. The method is valid if within the nonrespondents in Phase I (Patterns 2, 3, and 4), the missingness after Phase II sampling is MAR, namely,

$$(\boldsymbol{R}_{2\cdot1}=1|\boldsymbol{R}_1=0,\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = \text{P}(\boldsymbol{R}_{2\cdot1}=1|\boldsymbol{R}_1=0,\boldsymbol{Y}_{obs,2};\boldsymbol{\xi},\boldsymbol{Z}) \tag{17}$$

This missingness mechanism is called nonrespondent subsample missing at random (NS-MAR). Conditioning on $\boldsymbol{R}_1 = 0$, the joint distribution of $\boldsymbol{Y}_{obs,2}$, $\boldsymbol{Y}_{mis}$ and $\boldsymbol{R}_{2\cdot1}$ can be written as:

$$\begin{aligned} &P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) \\ &\quad = P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{\xi}; \boldsymbol{Z}) \end{aligned} \tag{18}$$

The joint distribution of $\boldsymbol{Y}_{obs,2}$ and $\boldsymbol{R}_{2\cdot1}$ conditional on $\boldsymbol{R}_1 = 0$ is obtained by integrating out $\boldsymbol{Y}_{mis}$ (Little and Rubin 2002):

$$P(\boldsymbol{Y}_{obs,2}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis} \tag{19}$$

The key assumption for the NSMI methods is NS-MAR, which ensures that the imputed values are from the predictive distribution of $\boldsymbol{Y}_{mis}$.

It should be noted that the assumption in (17) does not confine the missing-data mechanisms in the whole sample ($\boldsymbol{R}_2$) or the missing-data mechanism in Phase I ($\boldsymbol{R}_1$) to a certain missing-data mechanism, and therefore NSMI may be applied even under the MNAR missingness mechanism in Phase I or Phase I/II combined data as long as Phase II is MCAR or MAR. In contrast, the IL2 assumptions are violated, since under NS-MAR we have:

$$P(\boldsymbol{R}_1 = 1|\cdot) = f_1(\boldsymbol{X}, \boldsymbol{Y}_{obs,1}, \boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}), \mathrm{P}(\boldsymbol{R}_{2\cdot1} = 1|\boldsymbol{R}_1 = 0, \cdot) = f_2(\boldsymbol{X}, \boldsymbol{Y}_{obs,2})$$

where f_1 and f_2 are arbitrary functions, and thus:

$$\begin{aligned} P(\boldsymbol{R}_2 = 1|\cdot) &= P(\boldsymbol{R}_1 = 1|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 1, \cdot) + P(\boldsymbol{R}_1 = 0|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 0, \cdot) \\ &= P(\boldsymbol{R}_1 = 1|\cdot) + P(\boldsymbol{R}_1 = 0|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 0|\cdot) \\ &= f_1 + (1 - f_1) f_2 \end{aligned}$$

Since f_1 involves missing values, the distribution of $\boldsymbol{R}_2$ depends on underlying missing values, and therefore the assumption for IL2 is violated.

5. Simulation Studies

This section illustrates the properties of the NSMI method using simulation studies and compares the performance of NSMI to other methods under different missing-data mechanisms in Phases I and II. For each simulation study, six methods are applied to estimate the mean of the outcome Y and the regression coefficient of Y on scalar covariates Z and X:

1. *BD:* estimates using the data before deletion (BD), that is, the full data generated from simulation before missing values are created, as a benchmark method.
2. *CC1:* complete-case analysis using respondents from Phase I.
3. *CC2:* complete-case analysis using respondents from both Phases I and II.
4. *IL1:* multiple imputation using data from Phase I.

5. *IL2:* multiple imputation using data from both Phases I and II.
6. *NSMI:* multiple imputation in the nonrespondent subsample in Phase I using only additional data from Phase II.

The first three methods (BD, CC1, CC2) were implemented using standard maximum-likelihood estimation procedures in the software package R version 2.15.0 (R Development Core Team 2012). Methods 4–6 were implemented in the R package mice (multiple imputation through chained equations, Van Buuren and Groothuis-Oudshoom 2011); the number of imputed datasets was ten and the default was used for other options.

This article compares the performance of each of the methods using empirical bias, root mean square errors (RMSE), and the coverage probabilities of the 95% confidence intervals.

The first set of simulations generates $(z, x)_i$ from the normal distribution with mean 0, and covariance matrix $\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, for $i = 1,2,\ldots,1{,}000$. $\boldsymbol{Y}$ is related to $\boldsymbol{Z}$ and $\boldsymbol{X}$ by the linear model:

$$y_i = 1 + z_i + x_i + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0,1).$$

The response $\boldsymbol{Y}$ is subject to missingness, while $\boldsymbol{Z}$ and $\boldsymbol{X}$ are fully observed. Two covariates are used to allow the response mechanisms in Phases I and II to depend on different covariates (depending on z in Phase I and on x in Phase II). Let R_i denote the response indicator for y_i in Phase I. Phase I missing values in $\boldsymbol{Y}$ are generated based on the following three missing-data mechanisms:

(I) MCAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-1)$;
(II) MAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-1 + z_i)$;
(III) MNAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-y_i)$;

where $\text{expit}(\cdot)$ is the inverse logit function, $\text{expit}(\cdot) = \exp(\cdot)/[1 + \exp(\cdot)]$. R_i is then generated from a Bernoulli distribution with probability $\Pr(R_i = 0|z_i, x_i, y_i)$. Each missing-data generation scheme results in approximately 27% of the values of $\boldsymbol{Y}$ being missing in Phase I.

Let $R_{2\cdot 1,i}$ denote the response indicator in the subsample of nonrespondents in Phase I. Phase II responses in $\boldsymbol{Y}$ are generated under an MCAR mechanism:

$$\Pr(R_{2\cdot 1,i} = 1|R_i = 0; z_i, x_i, \boldsymbol{Y}_i) = 0.25.$$

The biases, root RMSE, and coverage probabilities of the 95% confidence intervals from each of the six methods are reported in Table 3. Results are based on 1,000 repetitions for each simulated condition.

For the MCAR missing-data mechanism in Phase I, all methods yield approximately unbiased estimates of both the mean of $\boldsymbol{Y}$ and the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. IL2 has the smallest RMSE for the population mean since it makes full use of the data. For the regression parameters, CC2 and IL2 give comparable estimates since the incomplete cases do not contain additional information for the regression of $\boldsymbol{Y}$ on the covariates for this missing-data mechanism (Little and Zhang 2011). The NSMI method has moderately

Table 3. Empirical bias, root mean square error, and 95% CI coverage probabilities under three missing-data mechanisms in Phase I and MCAR in Phase II (1,000 replications).

		MCAR				MAR				MNAR			
		μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z
Bias ($\times 10^4$)	BD	10	−13	1	15	−5	−18	7	−16	16	4	−9	−16
	CC1	15	−14	8	6	−6035	−10	18	−14	7961	2827	−1081	−1082
	CC2	16	−17	9	9	−4038	−14	7	−10	5286	1661	−517	−516
	IL1	8	−15	8	3	2	−11	15	−12	2840	2829	−1082	−1081
	IL2	4	−19	12	10	1	−12	7	−9	1673	1662	−516	−520
	NSMI	−2	−25	17	14	−1	−14	−13	6	26	15	−8	−20
RMSE ($\times 10^4$)	BD	578	311	327	331	602	318	332	326	607	307	341	337
	CC1	673	375	380	397	6071	417	442	429	7984	2854	1161	1161
	CC2	647	357	364	375	4091	373	397	399	5324	1700	651	652
	IL1	614	380	389	400	682	431	450	438	2903	2857	1165	1163
	IL2	606	361	371	381	644	379	404	404	1784	1701	653	660
	NSMI	663	440	454	468	699	477	519	496	681	428	483	487
Coverage	BD	96.5	94.7	95.4	95.7	94.8	93.9	95.0	96.3	94.6	95.8	94.1	94.5
	CC1	95.7	94.4	95.6	93.9	0.0	95.5	94.4	95.9	0.0	0.0	24.7	24.1
	CC2	95.1	94.4	94.9	94.7	0.0	95.1	94.6	95.6	0.0	0.2	73.7	72.5
	IL1	95.8	94.6	95.0	93.1	94.0	95.1	94.2	95.7	0.3	0.0	26.8	28.8
	IL2	96.3	93.7	95.0	93.6	93.7	95.2	93.5	95.3	23.8	0.3	73.5	73.6
	NSMI	96.1	94.8	94.9	95.0	94.0	93.5	93.5	94.9	95.4	96.1	94.7	94.8

larger RMSEs because of increased variability in the imputed values, which uses subjects from Patterns 2, 3, and 4, but not subjects from Pattern 1.

For the MAR missing-data mechanism in Phase I, all methods give approximately unbiased estimates for the regression coefficients, but the CC1 and CC2 methods show significant biases in estimating the mean of $\boldsymbol{Y}$. This is not surprising because the missingness of $\boldsymbol{Y}$ depends on $\boldsymbol{X}$, which is conditioned on in the estimation of the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{X}$, but not in the estimation of the (marginal) mean of $\boldsymbol{Y}$. As in the Phase I MCAR case, the NSMI has a somewhat greater RMSE than the IL methods, because information about the respondents in Phase I was not used in the imputation.

For MNAR missing-data generation in Phase I, the NSMI method is the only method that provides unbiased estimates of the mean of $\boldsymbol{Y}$ and the regression coefficients of $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{X}$. All other methods show significant biases because the MCAR or MAR assumptions are violated.

When the Phase II missingness mechanism is MAR, the results are similar to the results when the Phase II missingness mechanism is MCAR. Please refer to the online supplementary material for related results found at www.dx.doi.org/10.1515/jos-2016-0039.

The second set of simulations uses the same setup as in the MNAR scenario in the previous simulations, but vary the probability of being sampled in Phase II, that is, π is 0.05, 0.15, 0.25 or 0.50. The same six methods are applied on the simulated data. The bias, RMSE, and coverage probabilities are reported in Table 4.

For the simulated MNAR data in Phase I, only NSMI gives approximately unbiased estimates of the mean of $\boldsymbol{Y}$ and the regression coefficients. The precision increases as the sampling proportion in Phase II increases. Note that even randomly sampling five percent of the nonrespondents in Phase I is enough to distinguish the NSMI results from other competing methods. However, if data are collected on a small percentage of nonrespondents in Phase I, the NSMI yields estimates with large variance, and hence increased average lengths of the 95% confidence intervals. Please refer to the online supplementary material for additional simulation studies to examine how the performance of the NSMI method depends on the proportion of missingness.

The performance of different methods when the Phase II missing-data mechanism is MNAR is presented in the online supplementary material and is examined now. When the missing-data mechanism in Phase I is MAR or MCAR, IL1 is the only method that gives approximately unbiased estimates; in this case, both methods utilizing additional data from Phase II (IL2, and NSMI) are biased, because the missingness mechanism in Phase II is MNAR. When both Phases I and II's missingness mechanisms are MNAR, no method gives unbiased estimators for any of the parameters of interest. Please refer to the online supplementary material for additional studies comparing NSMI with alternative methods.

6. Application to Motivating Example

The proposed method will now be applied to the QOL dataset. For illustration purposes, the results for the resources subscale are presented. This is to estimate the mean resources and the regression of resources on gender (male versus female), age (in years),

Table 4. Empirical bias, root mean square error, and 95% CI coverage probabilities under MNAR in Phase I and with different Phase II sampling proportions (MCAR) (1,000 replications).

		$\pi = .05$				$\pi = .15$				$\pi = .25$				$\pi = .50$			
		μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z
Bias ($\times 10^4$)	BD	−32	−11	−5	0	39	21	−18	8	16	4	−9	−16	16	10	−6	4
	CC1	7933	2816	−1083	−1089	7977	2849	−1098	−1055	7961	2827	−1081	−1082	7967	2836	−1078	−1081
	CC2	7336	2522	−929	−933	6289	2066	−705	−665	5286	1661	−517	−516	3150	935	−231	−227
	IL1	2795	2814	−1083	−1085	2866	2850	−1095	−1061	2840	2829	−1082	−1081	2834	2828	−1074	−1073
	IL2	2503	2522	−930	−932	2083	2067	−706	−663	1673	1662	−516	−520	940	935	−232	−230
	NSMI	−65	−45	−21	10	6	−11	−20	22	26	15	−8	−20	16	10	−1	−1
RMSE ($\times 10^4$)	BD	600	323	321	333	604	307	329	333	607	307	341	337	600	325	338	336
	CC1	7958	2845	1158	1169	8001	2874	1171	1135	7984	2854	1161	1161	7991	2866	1154	1159
	CC2	7363	2553	1013	1023	6320	2097	807	776	5324	1700	651	652	3212	1002	440	439
	IL1	2860	2844	1161	1168	2928	2877	1173	1144	2903	2847	1165	1163	2899	2859	1153	1155
	IL2	2576	2555	1017	1025	2167	2099	812	777	1784	1701	653	660	1121	1003	444	445
	NSMI	1171	1058	1059	1145	733	526	592	584	681	428	483	487	624	373	398	393
Coverage	BD	95.1	94.5	95.2	94.2	94.7	95.2	94.9	95.2	94.6	95.8	94.1	94.5	94.7	94.8	94.4	94.2
	CC1	0.0	0.0	23.8	24.0	0.0	0.0	23.6	27.5	0.0	0.0	24.7	24.1	0.0	0.0	26.1	23.0
	CC2	0.0	0.0	36.4	36.7	0.0	0.0	55.3	59.7	0.0	0.2	73.7	72.5	0.1	25.1	89.2	89.8
	IL1	0.6	0.0	27.8	27.8	0.5	0.0	27.8	30.5	0.3	0.0	26.8	28.8	0.2	0.0	28.8	26.0
	IL2	1.6	0.0	39.5	39.7	7.3	0.0	56.8	61.9	23.8	0.3	73.5	73.6	67.3	24.8	89.2	89.9
	NSMI	94.3	94.5	95.5	93.7	95.3	95.9	94.5	93.9	95.4	96.1	94.7	94.8	95.4	94.8	95.0	94.1

race, and education. Race was dichotomized as white versus nonwhite, and education was dichotomized as high school and above versus education below high-school level.

All covariates are fully observed, whereas 147 out of 750 subjects have missing values in resources in Phase I. In Phase II, 39 out of the 147 nonrespondents provided data. Since the Phase II data collection was done within three months of Phase I, it was assumed that resources remained unchanged from Phase I. In implementing the NSMI method, we make the assumption that, among the 147 nonrespondents, the missingness after Phase II is MAR, thus meeting the conditions for NSMI. The validity of other competing methods rests on the mechanism that generates the missing data in Phase I. For instance, if the missing-data mechanism in Phase I is MCAR, then both CC and IL methods provide valid estimates. However, if the missingness in Phase I is MNAR (as suggested by Bonetti et al. 1999 and Fielding et al. 2009), then CC and IL methods will fail to give an unbiased estimation.

For all imputation methods, the fully observed resources measured at the mean age of 22 years are used in the imputation model, but not in the analysis model; this is because the resource scale measured at that age serves as a good predictor for the resources at the mean age of 33, but is not of direct interest in the analysis model (Meng 1994; van Buuren et al. 1999). The results from five methods are shown in Table 5. With respect to the modeling of resources as a function of gender, age, race, and education, NSMI shows a weaker negative association of race with resources compared with the other four methods. In particular, the NSMI method did not reveal a statistically significant association of race with resources, in contrast to the other methods, where whites had significantly greater resources. Age also had a somewhat weaker positive association with resources, although this relationship was not significant in any of the approaches. Those with higher levels of education and females had higher levels of resources, although these relationships were not statistically significant in any of the methods, nor did they differ systematically across the methods.

7. Discussion

Two-phase sampling has been proposed and used in surveys with nonresponse for five decades. However, little research has been done to show the benefit of nonresponse subsampling; traditional methods (i.e., weighting) also fail to make full use of the additional data collected from two-phase sampling. This article proposes an NSMI method to analyze data from NTS. The proposed method yields valid estimates when the missing-data mechanism in the subsample of initial nonrespondents is MAR, regardless of the missing-data mechanism in Phase I. The simulation studies also show that it is beneficial to use the NTS scheme, even when collecting data from only a small proportion of the nonrespondents.

Previous studies suggest that the missing-data mechanism in QOL outcomes was probably not MCAR (Bonetti et al. 1999; Fielding et al. 2009). Therefore, NSMI is considered in this applied example, which utilizes two-phase sampling to obtain data from a subsample of the initial nonrespondents in the Children in the Community study. Using the proposed NSMI method, white race was not found to be significantly associated with

Table 5. Regression analysis of a QOL dataset: resources.

	CC1					CC2					IL1				
	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value
Outcome															
Mean	77.26	0.69	75.9	78.62	<.001	77.20	0.67	75.89	78.51	<.001	77.03	0.72	75.6	78.46	<.001
Regression															
Intercept	62.09	6.07	50.19	73.99	<.001	63.85	5.9	52.28	75.43	<.001	62.32	6.15	50.19	75.46	<.001
Sex (male vs. female)	−2.48	1.38	−5.19	0.23	0.056	−2.70	1.33	−5.30	−0.10	0.043	−2.44	1.31	−5.02	0.14	0.054
Race (white vs. nonwhite)	5.64	2.47	0.80	10.49	0.038	5.35	2.36	0.72	9.98	**0.041**	**5.00**	**2.44**	**0.19**	**9.81**	**0.044**
Education (≥HS* vs.<HS)	1.83	1.45	−1.02	4.68	0.104	1.61	1.40	−1.14	4.36	0.125	1.71	1.45	−1.14	4.56	0.119
Age	0.46	0.26	−0.06	0.97	0.053	0.39	0.25	−0.10	0.89	0.060	0.47	0.28	−0.08	1.02	0.056

	IL2					NSMI				
	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value
Outcome										
Mean	77.09	0.67	75.77	78.42	<.001	77.07	0.71	75.67	78.47	<.001
Regression										
Intercept	64.64	5.73	53.4	75.88	<.001	66.84	5.64	55.77	77.91	<.001
Sex (male vs. female)	−2.27	1.41	−5.07	0.52	0.054	−2.35	1.28	−4.87	0.16	0.053
Race (white vs. nonwhite)	**4.84**	**2.25**	**0.43**	**9.26**	0.042	3.46	2.31	−1.07	7.99	0.067
Education (≥HS* vs.<HS)	2.04	1.33	−0.57	4.66	0.063	1.74	1.32	−0.85	4.34	0.094
Age	0.36	0.25	−0.13	0.84	0.075	0.33	0.24	−0.15	0.80	0.084

*HS: high school. LCL: lower confidence limit. UCL: upper confidence limit.

increased resources, while other alternatives suggested a significant association between race and resources.

By exploring the relationship between missing values and observed data, the NSMI methods use the information of the fully observed variables and improve the efficiency of the estimation. The method of multiple imputation by chained equations provides a valid way of utilizing the information in other variables when imputing the missing values. The NSMI method not only provides a valid estimation of the marginal distribution of the outcome (e.g., mean), but also of the conditional distribution of the outcome on covariates (e.g., regression).

Missing data are ubiquitous and all methods for handling missing data rely on untestable assumptions. NTS provides a valid way to relax these untestable assumptions in part. Ideally, Phase II sampling takes a random sample of Phase I nonrespondents. However, these random subsamples may still be subject to nonresponse. In cases when the sampling yields a missing-data mechanism of MAR for Phase I nonrespondents, the proposed NSMI method is valid, regardless of the first-stage mechanism. In the event that both first- and second-stage missing-data mechanisms are MNAR, neither NSMI nor any multiple-imputation methods that ignore missing-data mechanisms are free of bias. Of course, in practice, assessing MNAR directly is not typically possible—the motivation for the NSMI approach is that, if the Phase I missingness mechanism is strongly MNAR, the Phase II missingness may be less so, because the Phase I nonrespondents may share some common characteristics that make the NS-MAR assumption plausible.

The NTS scheme considered in this article involves collecting data from nonrespondents. This is challenging in practice, but may be achieved by giving an abridged version of the questionnaire, by giving incentives for response, or by using other advanced survey techniques, such as tailoring the questionnaire to the interviewees (Groves and Couper 1998). In the second-stage subsampling within a fixed budget, there is a balance between reducing the nonresponse rate and subsampling more subjects, because by focusing on a moderate number of nonrespondents, it is possible to obtain a high response rate and therefore reduce the nonresponse bias (Elliott et al. 2000). This aspect of the problem is currently under investigation.

Finally, it should be noted that use of the NSMI approach is not fail-safe. As the simulation studies show, if Phase I missingness is MCAR, there is no gain in using the NSMI approach; if Phase I is MCAR or MAR, and Phase II is MNAR, substantial bias can be introduced relative to MAR methods that ignore the Phase II data. While Phase I MAR and MNAR mechanisms cannot be distinguished from observed data, some evidence for Phase I MCAR can be deduced from the observed data. Hence, methods that consider the evidence for MCAR and 'trade off' Phase I versus Phase II imputation may be desirable to enhance robustness under all different mechanisms. In addition, follow-up nonresponse designs that devote more intensive effort to minimizing Phase II MNAR though use of techniques that may not be practical or cost-effective to implement during Phase I data collection (use of targeted incentives, expensive but high response rate data-collection modes such as face-to-face interviews) might be implemented to make NSMI assumptions more plausible. Future research is needed into analytic methods both to improve robustness to NSMI assumption failures and to consider data-collection methods that better meet NSMI assumptions.

8. References

Amemiya, T. 1984. "Tobit Models, a Survey." *Journal of Econometrics* 24: 3–61. Doi: http://dx.doi.org/10.1016/0304-4076(84)90074-5.

Bonetti, M., B.F. Cole, and R.D. Gelber. 1999. "A Method-of-Moments Estimation Procedure for Categorical Quality-of-life Data with Nonignorable Missingness." *Journal of the American Statistical Association* 94: 1025–1034. Doi: http://dx.doi.org/10.1080/01621459.1999.10473855.

Chen, H., P. Cohen, S. Kasen, R. Dufur, E.M. Smailes, and K. Gordon. 2004. "Construction and Validation of a Quality of Life Instrument for Young Adults." *Quality of Life Research* 13: 747–759. Doi: http://dx.doi.org/10.1023/B:QURE.0000021700.42478.ab.

Cohen, P., T.N. Crawford, J.G. Johnson, and S. Kasen. 2005. "The Children in the Community Study of Developmental Course of Personality Disorder." *Journal of Personality Disorder* 19: 466–486.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39: 1–38. Available at: http://www.jstor.org/stable/2984875 (accessed May 2016).

Elliott, M.R., R.J.A. Little, and S. Lewitzky. 2000. "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95: 730–738. Doi: http://dx.doi.org/10.1080/01621459.2000.10474261.

Fielding, S., P.M. Fayers, and C.R. Ramsay. 2009. "Investigating the Missing Data Mechanism in Quality of Life Outcomes: A Comparison of Approaches." *Health and Quality of Life Outcomes* 7: 57. Doi: http://dx.doi.org/10.1186/1477-7525-7-57.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529. Doi: http://dx.doi.org/10.1080/01621459.1946.10501894.

Harter, R.M., T.L. Mach, J.F. Chapline, and J.D. Wolken. 2007. "Determining Subsampling Rates for Nonrespondents." In Proceedings of ICES-III, June 18–21, 2007. 1293–1300. Montreal, Quebec, Canada. Available at: http://50.205.225.65/meetings/ices/2007/proceedings/ICES2007-000197.PDF (accessed May 2016).

Heitjan, D. and D.B. Rubin. 1991. "Ignorability and Coarse Data." *The Annals of Statistics* 81: 2244–2253. Available at: http://www.jstor.org/stable/2241929 (accessed May 2016).

Heckman, J.J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for such Models." *Annals of Economic and Social Measurement* 5: 475–492. Available at: http://www.nber.org/chapters/c10491.pdf (accessed May 2016).

Little, R.J.A. 1993. "Pattern-Mixture Model for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88: 125–134. Doi: http://dx.doi.org/10.1080/01621459.1993.10594302.

Little, R.J.A. 1994. "A Class of Pattern-Mixture Models for Normal Incomplete Data." *Biometrika* 81: 471–483. Doi: http://dx.doi.org/10.1093/biomet/81.3.471.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley.

Little, R.J.A. and N. Zhang. 2011. "Subsample Ignorable Likelihood for Regression Analysis with Missing Data." *Journal of the Royal Statistical Society, Series C* 60: 591–605. Doi: http://dx.doi.org/10.1111/j.1467-9876.2011.00763.x.

Meng, X.-L. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Sciences* 9: 538–573. Available at: http://www.jstor.org/stable/2246252 (accesssed May 2016).

Nandram, B. and J.W. Choi. 2002. "Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability." *Journal of the American Statistical Association* 97: 381–388. Doi: http://dx.doi.org/10.1198/016214502760046934.

Nandram, B. and J.W. Choi. 2010. "A Bayesian Analysis of Body Mass Index Data from Small Domains under Nonignorable Nonresponse and Selection." *Journal of the American Statistical Association* 105: 120–135. Doi: http://dx.doi.org/10.1198/jasa.2009.ap08443.

The R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available at: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (accessed June 2016).

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592. Doi: http://dx.doi.org/10.1093/biomet/63.3.581.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Srinath, K.P. 1971. "Multiphase Sampling in Nonresponse Problems." *Journal of the American Statistical Association* 66: 583–620. Doi: http://dx.doi.org/10.1080/01621459.1971.10482310.

Statistics Canada. 2011. National Household Survey. Available at: http://www12.statcan.gc.ca/nhs-enm/index-eng.cfm (accessed June 1, 2015).

Van Buuren, S., H.C. Boshuizen, and D.L. Knook. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18: 681–694. Available at: http://www.stefvanbuuren.nl/publications/Multiple%20imputation%20-%20Stat%20Med%201999.pdf (accessed May 2016).

Van Buuren, S.‘ and K. Groothuis-Oudshoom. 2011. "Mice: Multivariate Imputation by Chained Equation in R." *Journal of Statistical Software* 45: 1–67. Available at: http://doc.utwente.nl/78938/.

Received February 2015
Revised July 2015
Accepted August 2015

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 1–8, http://dx.doi.org/10.1515/JOS-2016-0039

Online Supplementary Material

Nonrespondent Subsample Multiple Imputation in Two-Phase Sampling for Nonresponse

Nanhua Zhang[1], Henian Chen[2], and Michael R. Elliott[3]

This online supplementary document provides some simulation results to compare the NSMI method with alternatives: (1) when the missingness for Phase II is missing at random; (2) when the missingness for Phase II is missing not at random; (3) when the missingness proportion in Phase I is large (and small).

A SIMULATION STUDY WITH MAR MISSINGNESS MECHANISM IN PHASE II

This studypresents simulations to compare the proposed method with alternative methods when the missingness mechanism in Phase II is MAR. $(z,x)_i$ isgeneratedfrom the normal distribution with mean 0, and covariance matrix $\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, for $i=1,2,...,1000.$ Y is related to Z and X by the linear model:

$$y_i = 1 + z_i + x_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,1).$$

[1] Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, OH 45229, U.S.A. Email: nanhua.zhang@cchmc.org (corresponding author)

[2] Department of Epidemiology & Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612-3085, U.S.A. Email: hchen1@health.usf.edu

[3] Department of Biostatistics, School of Public Health, University of Michigan and Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48019, U.S.A. Email: mrelliot@umich.edu

The response Y is subject to missingness, while Z and X are fully observed. Let R_i denote the response indicator for y_i in Phase I. Phase I missing values in Y are generated based on the following three missing-data mechanisms:

(I) MCAR: $\Pr(R_i = 0 \mid z_i, x_i, y_i) = \text{expit}(-1)$;

(II) MAR: $\Pr(R_i = 0 \mid z_i, x_i, y_i) = \text{expit}(-1 + z_i)$;

(III) MNAR: $\Pr(R_i = 0 \mid z_i, x_i, y_i) = \text{expit}(-y_i)$;

where $\text{expit}(\cdot)$ is the inverse logit function, $\text{expit}(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$. R_i is then generated from a Bernoulli distribution with probability $\Pr(R_i = 0 \mid z_i, x_i, y_i)$. Each missing-data generation scheme results in about 27% of the values of Y being missing in Phase I.

Phase II response indicators in Y are now generated under an MAR mechanism:

$$\Pr(R_{2|1,i} = 1 \mid R_i = 0; z_i, x_i, y_i) = \text{expit}(-1 + x_i),$$

which gives a response rate of around 27%.

The simulation results are summarized in Table S.1, based on 1000 replications. The results are very similar to thosein Section 5. Since the missing-data mechanism in the Phase I nonresponse subsample is MAR, NSMI gives unbiased estimates of all parameters, regardless of the missing-data mechanism in Phase I. The reduction in bias and RMSE is substantial when the missing-data mechanism in Phase I is MNAR. In cases where the missing-data mechanism in phase is MCAR or MAR, there is a slight loss of information in NSMI methods, compared with the other ignorable-likelihood methods.

Table S.1. Empirical Bias, Root Mean Square Error, and 95% CI Coverage Probabilitiesunder Three Missing–Data Mechanisms in Phase I and MAR in Phase II (1,000 Replications)

		MCAR					MAR					MNAR			
		μ	β_0	β_x	β_z		μ	β_0	β_x	β_z		μ	β_0	β_x	β_z
Bias($\times 10^4$)	BD	–16	–11	–3	13		–16	–7	–5	21		–2	0	–14	–4
	CC1	–19	–21	–1	16		–3331	–3	–1	21		7943	2826	–1092	–1087
	CC2	743	–20	1	12		–1059	–2	–5	22		6207	1848	–1023	–525
	IL1	–25	–20	–4	12		–15	–7	1	15		2827	2828	–1092	–1088
	IL2	–25	–20	–3	13		–13	–4	–5	21		1844	1846	–1022	–527
	NSMI	–21	–16	1	11		–6	3	–12	24		33	35	–67	–40
RMSE($\times 10^4$)	BD	596	328	332	345		608	305	314	337		605	328	334	331
	CC1	684	392	371	398		3403	384	387	436		7967	2854	1167	1159
	CC2	992	369	358	382		1264	344	360	400		6238	1886	1101	650
	IL1	635	398	380	406		663	391	396	450		2892	2857	1171	1164
	IL2	622	374	364	388		634	348	368	406		1943	1885	1103	655
	NSMI	681	470	451	447		693	444	460	492		748	549	645	504
Coverage	BD	95.4	94.5	94.8	94.4		95.3	95.9	96.2	95.3		94.1	93.5	94.6	94.9
	CC1	95.7	93.3	95.6	94.9		0.2	95.2	95.9	94.5		0.0	0.0	23.8	23.4
	CC2	81.1	93.5	95.1	94.8		65.3	96.1	95.4	94.5		0.0	0.0	29.4	73.3
	IL1	94.9	93.2	95.3	94.2		93.9	95.8	95.8	93.5		0.3	0.0	27.4	27.5
	IL2	94.9	93.1	94.5	94.5		95.2	95.7	95.4	94.5		15.2	0.0	31.2	73.3
	NSMI	95.2	94.4	95.2	94.9		95.5	96.0	94.6	94.1		94.8	93.6	94.5	95.3

A SIMULATION STUDY WITH MNAR MISSINGNESS MECHANISM IN PHASE II

As in the Phase II MCAR simulations in the main text, missing values in Y in Phase I are generated based on the following three missing-data mechanisms:

(I) MCAR: $\Pr\left(R_i = 0 \mid z_i, x_i, y_i\right) = \text{expit}\left(-1\right)$;

(II) MAR: $\Pr\left(R_i = 0 \mid z_i, x_i, y_i\right) = \text{expit}\left(-1 + z_i\right)$;

(III) MNAR: $\Pr\left(R_i = 0 \mid z_i, x_i, y_i\right) = \text{expit}\left(-y_i\right)$;

For mechanism (I) and (II) in Phase I, Phase II response indicators of Y are generated under an MNAR mechanism:

$$\Pr(R_{2|1,i} = 1 \mid R_i = 0; z_i, x_i, y_i) = \text{expit}\left(-y_i\right);$$

To achieve comparable response rates in Phase II, the response indicator in Phase II for mechanism (III) above was generated based on the following:

$$\Pr(R_{2|1,i} = 1 \mid R_i = 0; z_i, x_i, y_i) = \text{expit}\left(-2 - y_i\right);$$

These schemes are MNAR missing-data mechanisms, which produce response rates between 25 and 35%.

The results are shown in Table S.2. Since the mechanism in Phase I nonrespondents is MNAR, the NSMI method yields biased estimates. When the missing-data mechanism in Phase I is MAR or MCAR, the method IL1 is the only method that gives approximately unbiased estimates; in this case, both methods utilizing additional data from Phase II (IL2, and NSMI) are biased, because the missingness mechanism in Phase II is MNAR.When both Phase I and Phase II missingness mechanisms are MNAR, no methods give unbiased estimators for any of the parameters of interest. The results are generally consistent with theoretical expectations; however, it is possible to get relatively precise estimates for some parameters, such as β_x and β_z in the setup when both Phase I

and Phase II missingness are MNAR. In this case, small values are more likely to be missing in Phase I and also more likely to be recovered in Phase II; thus the counteracting effect leads to good estimates of β_x and β_z. This effect, however, is specific to this simulation design and may not be readily generalized to other settings.

Table S.2. Empirical Bias, Root Mean Square Error, and 95% CI Coverage Probabilitiesunder Three Missing–Data Mechanisms in Phase I and MNAR in Phase II (1,000 Replications)

		MCAR					MAR					MNAR			
		μ	β_0	β_x	β_z		μ	β_0	β_x	β_z		μ	β_0	β_x	β_z
Bias($\times 10^4$)	BD	5	6	17	–5		–27	7	–10	3		41	–2	–14	2
	CC1	24	25	18	0		–3347	6	–16	–2		7980	2824	–1085	–1060
	CC2	–1670	–446	73	60		–3925	–545	101	–319		3828	1031	78	86
	IL1	24	26	21	2		–31	3	–17	–4		2866	2827	–1087	–1063
	IL2	–448	–446	72	60		–577	–543	100	–320		1073	1031	77	87
	NSMI	–1523	–1522	–373	–379		–2123	–2090	–452	–1590		–1919	–1964	559	578
RMSE($\times 10^4$)	BD	604	310	326	336		588	317	329	327		605	314	335	337
	CC1	704	365	384	376		3417	400	397	417		8004	2850	1161	1133
	CC2	1800	563	372	362		3978	665	389	508		3893	1097	395	400
	IL1	644	374	390	385		642	406	406	430		2928	2855	1165	1139
	IL2	776	567	375	366		850	666	392	515		1252	1098	401	402
	NSMI	1665	1589	570	567		2238	2164	673	1696		2059	2029	723	738
Coverage	BD	94.2	95.1	95.6	94.5		95.7	94.4	95.2	94.5		94.4	95.6	95.7	95.0
	CC1	94.6	95.3	95.2	95.5		0.1	94.3	94.8	95.2		0.0	0.0	24.5	26.2
	CC2	27.9	76.4	94.6	95.2		0.0	67.9	94.6	87.9		0.0	20.9	95.0	94.5
	IL1	93.8	94.9	95.8	95.1		95.6	94.6	95.3	94.4		0.3	0.0	26.1	29.2
	IL2	90.0	76.3	94.4	95.1		85.0	70.5	94.6	88.8		60.9	23.3	94.9	94.7
	NSMI	35.6	10.1	86.5	86.4		17.3	5.4	85.6	27.2		27.8	3.9	78.3	77.0

A SIMULATION STUDY WITH VARYING PERCENTAGES OF MISSINGNESS IN PHASE I

This sectionpresents a simulation study to assess how dependent the proposed method is onthe missingness proportions, by varying the missingness proportion in Phases I and II. The outcome and the covariates are simulated the same way as in the previous simulation study. Phase I missing values in *Y*are generated based on two MNAR missing-data mechanisms:

(I) MNAR: $\Pr\left(R_i = 0 \mid z_i, x_i, y_i\right) = \text{expit}\left(-2.3 - y_i\right)$;

(II) MNAR: $\Pr\left(R_i = 0 \mid z_i, x_i, y_i\right) = \text{expit}\left(1 - y_i\right)$.

These two missing-data mechanisms result in 10 and 50% missingness in *Y*, respectively. Phase II response indicators are generated under an MCAR mechanism and give a response rate of about 10% and 50%. These four scenarios examine the dependency of the method to proportion of missingness in both phases.

The simulation results are summarized in Table S.3, based on 1000 replications. Since the missing-data mechanism in the Phase II nonresponse subsample is MCAR, NSMI gives very small empirical biases for all parameters. The coverage probabilities of the 95% confidence intervals are close to the 95% under all four scenarios. When either the nonresponse rate is high or the proportion of nonrespondent subsample collection is high, the NSMI method yields the smallest RMSEs for almost all parameters. However, when both the percentage of missingness in Phase I and the proportion of nonrespondent subsample collection are low, the NSMI method results in less precise estimates with large interval lengths and increased RMSEs.

Table S.3. Empirical Bias, Root Mean Square Error, and 95% CI Coverage Probabilitiesunder MNAR in Phase I (Missingness Proportions 10 and 50%) and MCAR in Phase II (1,000 Replications)

		10% Missingness in Phase I & 10% collection in Phase II				50% Missingness in Phase I & 10% collection in Phase II				10% Missingness in Phase I & 50% collection in Phase II				50% Missingness in Phase I & 50% collection in Phase II			
		μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z
Bias ($\times 10^4$)	BD	–31	–14	–10	–17	4	9	2	–4	29	14	–8	0	29	18	–7	11
	CC1	2637	822	–490	–506	11224	4196	–1318	–1321	2684	844	–486	–486	11271	4201	–1327	–1289
	CC2	2348	721	–428	–441	8304	2689	–621	–629	1283	378	–201	–199	3759	1086	–121	–98
	IL1	805	821	–487	–505	4196	4199	–1321	–1324	858	843	–486	–485	4209	4200	–1330	–1288
	IL2	704	721	–427	–441	2685	2689	–622	–630	394	379	–201	–200	1099	1088	–119	–101
	NSMI	–61	–42	–50	–52	–10	–7	16	4	26	11	2	2	33	21	–6	1
RMSE ($\times 10^4$)	BD	603	318	313	337	594	319	325	346	607	306	341	333	606	318	346	333
	CC1	2700	886	591	615	11245	4223	1392	1405	2747	900	602	600	11291	4228	1408	1365
	CC2	2421	792	539	560	8335	2726	754	774	1414	488	405	396	3820	1147	411	393
	IL1	997	887	592	616	4247	4229	1399	1414	1045	901	604	602	4263	4228	1415	1370
	IL2	918	792	539	564	2764	2728	761	784	718	489	407	398	1271	1150	413	395
	NSMI	2754	2761	3101	2124	797	613	660	661	611	317	365	357	637	378	417	411
Coverage	BD	95.8	95.7	95.9	94.6	95.6	95.4	96.1	94.3	95.1	95.5	94.1	95.2	95.1	95.6	93.8	95.1
	CC1	0.7	29.6	70.8	69.7	0.0	0.0	18.8	21.0	0.3	26.3	72.1	70.8	0.0	0.0	17.2	19.3
	CC2	2.5	41.1	77.0	75.1	0.0	0.0	71.4	68.2	43.6	79.3	90.0	91.6	0.1	15.6	93.3	94.8
	IL1	71.6	30.4	71.8	70.3	0.0	0.0	21.3	25.3	68.5	27.3	71.8	70.7	0.0	0.0	24.5	24.7
	IL2	76.3	41.7	77.6	75.1	2.1	0.0	74.3	68.9	87.9	79.5	89.3	91.6	57.3	17.5	93.9	94.9
	NSMI	94.3	94.9	95.2	94.4	94.6	95.4	94.8	94.0	95.4	96.3	93.8	94.9	95.1	95.5	93.6	94.4