<table>
<thead>
<tr>
<th>Statistica Sinica Preprint No: SS-2017-0016</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Yue Zhang  
Henian Chen and  
Nanhua Zhang |
| **Corresponding Author** | Nanhua Zhang |
| **E-mail** | nanhua.zhang@cchmc.org |

Notice: Accepted version subject to English editing.
BAYESIAN INFERENCE FOR NONRESPONSE TWO-PHASE SAMPLING

Yue Zhang\textsuperscript{1}, Henian Chen\textsuperscript{2} and Nanhua Zhang\textsuperscript{3}

\textsuperscript{1}Shanghai Jiao Tong University, \textsuperscript{2}University of South Florida, \textsuperscript{3}Cincinnati Children’s Hospital Medical Center

Abstract: Nonresponse is an important practical problem in epidemiological surveys and clinical trials. Common methods for dealing with missing data rely on untestable assumptions. In particular, non-ignorable modeling, which derives inference from the likelihood function based on a joint distribution of the variables and the missingness indicators, can be sensitive to misspecification of this distribution and may also have problems with identifying the parameters. Nonresponse two-phase sampling (NTS), which re-contacts and collects data from a subsample of the initial nonrespondents, has been used to reduce nonresponse bias. The additional data collected in phase II provide important information for identifying the parameters in the non-ignorable models. We propose a Bayesian selection model which utilizes the additional data from phase II and develop an efficient Markov chain Monte Carlo algorithm for the posterior computation. We illustrate the proposed model on simulation studies and a Quality of Life (QOL) dataset.

Key words and phrases: Bayesian selection model, Markov chain Monte Carlo, missing not at random, quality of life, two-phase sampling.

1. Introduction

Nonresponse at an appreciable rate exists in various applications such as epidemiological surveys and clinical trials. Commonly used approaches to handling missing data include complete-case (CC) analysis, ignorable likelihood (IL) methods and nonignorable models (NIM). CC, which discards the incomplete cases, can result in substantial loss of information or biased estimation of the key parameters. IL methods are based on the observed likelihood which does not include a model for the missing data indicator, and these models provide valid inference if the missingness does not depend on the missing values. Such missing data mechanism is called missing at random (MAR) (Rubin, 1976; Little and Rubin, 2002). Examples of IL methods include ignorable maximum likelihood (IML) (Dempster, Laird, and Rubin, 1977), Bayesian inferences (Sugden and Smith, 1984),
and multiple imputation (Raghunathan, Lepkowski, Van Hoewyk, and Solenberg-er, 2001; Rubin, 2004). When the data are missing not at random (MNAR), i.e., the missingness could depend on the missing values, non-ignorable models (NIM) are developed based on the joint distribution of the variables and the missing data indicators. The nonignorable models are less common in practice, because of the difficulty in specifying the models for the missing data mechanism, sensitivity to model misspecification, and problems with identifying the parameters (Little and Rubin, 2002; Heckman, 1979; Little, 1993, 1994; Nandram and Choi, 2002, 2010).

All three methods (CC, IL, and NIM) rely on untestable assumptions about the missing data mechanism. Sensitivity analyses have been proposed to systematically examine the effect of perturbations to model assumptions (Little, 1993, 1994; Troxel, Ma, and Heitjan, 2004; Scharfstein, McDermott, Olson, and Wie- gand, 2014; Zhu, Ibrahim, and Tang, 2014). 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 non-respondents to the original survey (phase I) is randomly selected for further interview attempts (phase II). This method is called nonresponse two-phase sampling (NTS). The general missing data structure for NTS is listed in Table 1.

| Pattern | Observation, i | $y_i$ | $R_{1,i}$ | $S_{2|1,i}$ | $R_{2|1,i}$ | $R_{2,i}$ |
|---------|----------------|------|---------|----------|----------|----------|
| 1       | $i = 1, \ldots, m$ | √    | 1       | -        | -        | 1        |
| 2       | $i = m + 1, \ldots, m + r$ | ×    | 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; × denotes at least one entry missing in phase I, but observed in phase II.

Let $\{y_i, i = 1, 2, \ldots, n\}$ denote $n$ independent observations on a outcome variable $Y$, where $Y$ have missing values; if missing, $y_i$ denotes the underlying missing value of the outcome for the $i$-th subject. The response indicator for
phase I is denoted as $R_{1,i}$, equal to 1 if $y_i$ is observed and 0 otherwise. $S_{2|1,i}$ is used to denote whether a subject is sampled among the nonrespondents in phase I. Let $R_{2|1,i}$ denote the phase II response indicator among the nonrespondents in phase I, and $R_{2,i}$ be the overall response indicator after completion of phase II. There are four patterns in Table 1. Pattern 1 consists of subjects from whom $y_i$ 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 not sampled in phase II.

NTS was first proposed by Hansen and Hurwitz (1946) to reduce the nonresponse bias in mail questionnaires by doing personal interviews on a fraction of the nonrespondents. This sampling scheme was referred to as "call-back" and Cochran (1977) studied the effects of call-backs and the optimal sampling fractions among the nonrespondents. Some other examples include the National Comorbidity Survey (Elliott and Little, 2000) and the 2003 Survey of Small Business Finances (Harter, Mach, Wolken, and Chapline, 2007). Different from those approaches relying on using case weights (Hansen and Hurwitz, 1946; Srinath, 1971; Harter, Mach, Wolken, and Chapline, 2007), a method called nonrespondent subsample multiple imputation (NSMI) was proposed by Zhang, Chen, and Elliott (2016) to reduce bias using data from the NTS. NSMI performs multiple imputation within the subsample of nonrespondents in phase I by using additional data collected in phase II. It works well if MAR assumption holds in phase II within the sample of nonrespondents in phase I regardless of the missingness mechanism in phase I. However, this assumption is usually untestable and phase II response mechanism from phase I nonrespondents may be related to outcome values in many practical studies, in which case the NSMI methods yield biased estimates. We propose nonrespondents subsample Bayesian selection model (NSBSM), which makes use of the additional data from phase II when jointly modeling the outcome and the phase I missing data indicator. The rationale of using the additional data in modeling the phase I indicator is that the additional data provide important identifying information for modeling the phase I missingness indicator. For model comparison purpose, we also apply the Bayesian selection model without considering the phase II data (BSM).
The rest of the paper is organized as follows. Section 2 reviews NSMI method, which has been proposed for NTS (Zhang, Chen, and Elliott, 2016). The method works well when phase I missingness is MNAR but the missingness in the phase II among the nonrespondents is MAR. Section 3 introduces the NSBSM method, including the model setup and posterior inference. We illustrate the properties of the NSBSM and compare the performance of different methods in section 4 using simulation studies, while section 5 applies the method to a quality of life (QOL) dataset. Section 6 concludes the paper with discussion.

2. Nonrespondent subsample multiple imputation (NSMI)

Data with the structure in Table 1 are considered. NSMI applies the multiple imputation method to the cases in patterns 2, 3, and 4 of Table 1. Subjects in pattern 1 are excluded when the missing values in pattern 3 and 4 are imputed, and then the imputed datasets from patterns 2, 3, 4 are combined with data from pattern 1 for statistical analyses (Zhang, Chen, and Elliott, 2016). The key assumption of NSMI is that within the nonrespondent in phase I (patterns 2, 3, and 4), the missingness after phase II is MAR. Let $Y_{obs} = (Y_{obs,1}, Y_{obs,2})$ represent observed data in phase I and phase II, and $Y_{mis}$ to represent the data missing after phase II sampling. The assumption for NSMI to be valid can be expressed as follows based on fully observed covariate vector $Z$.

$$\Pr(R_2|1 = 1, R_1 = 0, Y_{obs,2}, Y_{mis}, \gamma; Z) = \Pr(R_2|1 = 1, R_1 = 0, Y_{obs,2}, \gamma; Z),$$

where $\gamma$ is the parameter associated with the distribution of the response indicator $R$. The missingness mechanism is called nonrespondent subsample missing at random (NS-MAR). Note that the assumption does not confine the missing data mechanisms in the whole sample ($R_2$) or the missing data mechanism in phase I ($R_1$) to a certain missing data mechanism, and therefore NSMI may be applied even under the MNAR missingness mechanism in phase I as long as phase II is MCAR or MAR.

However, when the phase II missing data mechanism is MNAR, the NSMI method fails to yield unbiased estimates because the NS-MAR assumption is violated. Also, the NSMI method usually leads to estimates with large variance, due to the increased variability in the imputed values by using subjects from Pat-
tern 2, 3, and 4, but not subjects from Pattern 1. These motivate the proposed method in the next section.

3. Nonrespondents Subsample Bayesian Selection Model (NSBSM)

3.1. Selection Model

In this section, data with structure in Table 1 are also considered. This paper proposes nonrespondents subsample Bayesian selection model (NSBSM), a Bayesian approach based on selection model to address the identifiability issue (Little and Rubin, 2002) by utilizing additional data in phase II. A selection model contains a regression equation for the outcome, and a regression equation for the sample selection mechanism. Suppose the regression equation for the outcome of primary interest is

$$y_i = x_i^T \beta + \eta_i,$$

(3.1)

and the sample selection mechanism is driven by the following latent linear regression equation

$$u_i = z_i^T \gamma + \epsilon_i,$$

(3.2)

where $i = 1, 2, ..., N$, $x_i$ and $z_i$ are the covariates and they may overlap with each other. Assume that the outcome $y_i$ is observed if and only if $u_i > 0$. The missing indicator for sample selection of phase I is

$$R_{1,i} = I(u_i > 0),$$

where $I(\cdot)$ is indicator function.

Heckman (1979) assumed a bivariate normal distribution for $\eta_i$ and $\epsilon_i$ in equations (3.1) and (3.2).

$$\begin{pmatrix} \eta_i \\ \epsilon_i \end{pmatrix} \sim N_2(0, \Sigma),$$

where $0_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$. The second diagonal element of $\Sigma$ is set to 1 for full identification, which is a typical constraint for a binary choice model. To facilitate posterior inference, we factor the bivariate normal distribution
\((\epsilon_i, \eta_i)\) into the product of marginal distribution of \(\epsilon_i\) and conditional distribution of \(\eta_i | \epsilon_i\), and we obtain
\[
y_i = x_i^T \beta + E[\eta_i | \epsilon_i] + \xi_i,
\]
\[
u_i = z_i^T \gamma + \epsilon_i,
\]
where \(\epsilon_i \sim N(0, 1)\), \(\xi_i \sim N(0, \sigma^2)\), \(\eta_i | \epsilon_i = \sigma_{12} \cdot \epsilon_i\), and \(\sigma^2 = \sigma_{11} - \sigma_{12}^2\). By utilizing this re-parameterization, our parameters of interest for the covariance structure are \(\sigma^2\) and \(\sigma_{12}\), for which we assign independent priors. This re-parameterization was first proposed by Koop and Poirier (1997) to address the complication in estimating the two free parameters \((\sigma_{11}, \sigma_{12})\) in the covariance matrix \(\Sigma\); Efficient Gibbs sampling algorithm was also developed by Li (1998) following the re-parameterization. Under the aforementioned bivariate normal assumption, the model (3.1) and (3.2) implies that
\[
Pr(R_{1,i} = 1 | z_i) = \Phi(z_i^T \gamma)
\]
and
\[
E(y_i | R_{1,i} = 1, x_i, z_i) = x_i^T \beta + \sigma_{12} \lambda(z_i^T \gamma),
\] (3.3)
where \(\lambda(\cdot) = \phi(\cdot) / \Phi(\cdot)\) is the inverse Mills ratio (Little and Rubin, 2002).

### 3.2. Model identification

Choosing the appropriate covariates \((x_i^T, z_i^T)\) plays an important role in the selection model. Little and Rubin (2002) points out that the presence of the inverse Mills ratio in equation (3.3) often results in multicollinearity that can lead to profound identification problem for estimating \(\beta\) and \(\sigma_{12}\) in equation (3.3). One possible solution to this problem is to incorporate exclusion restriction assumption, i.e., by including at least one of the elements of \(z_i\) that are not in \(x_i\), which is associated with the selection process but not the outcome. With a valid exclusion restriction assumption, the inverse Mills ratio and the \(x_i\) vector in equation (3.3) will be less correlated, reducing multicollinearity among predictors and therefore facilitating model identification.

However in practice, it can be difficult to identify variables that satisfy the exclusion restriction assumption. With the additional data from phase II, the
model can be identified without assuming exclusion restriction. The inverse Mills ratio is estimated by the non-linear probit model, the correction term $\lambda$ will not be perfectly correlated with $x_i$, even in the absence of exclusion restriction. In addition, the proposed two-phase sampling method provides additional information of nonrespondents of phase I. The model for the initial nonrespondents takes the form (Little and Rubin, 2002)

$$E(y_i|R_{1,i} = 0, x_i, z_i) = x_i^T \beta + \sigma_{12}(-\lambda(-z_i^T \gamma)),$$

where $z_i = x_i$. With the additional data from phase II, equations (3.3) and (3.4) imply that the new correction term $\lambda$ is a vector with both $\lambda(z_i^T \gamma)$ and $-\lambda(-z_i^T \gamma)$ which makes it less linearly correlated with $x_i$, and therefore we can estimate $\beta$ and $\sigma_{12}$ even without exclusion restrictions. We show the benefit of the proposed method in identifying parameters in the simulation study. In the following sections, we use notation $x_i$ as the common covariate in both (3.1) and (3.2). For simplicity, we suppress the conditioning on $x_i$ from all equations in the following sections.

### 3.3. Likelihood function

Let $\theta = (\gamma^T, \beta^T)^T$ denote all the coefficients. It follows from the bivariate normality distribution for $\eta_i$ and $\epsilon_i$ that, when outcome is observed in phase I

$$\Pr\{R_{1,i} = 1|y_i, \theta, \sigma^2, \sigma_{12}\} = 1 - \Phi\left( x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right).$$

On the other hand, when the outcome is missing in phase I, the probability that this occurs is

$$\Pr\{R_{1,i} = 0|y_i, \theta, \sigma^2, \sigma_{12}\} = \Phi\left( x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right).$$

Without incorporating the additional phase II data, the likelihood function of
the traditional Bayesian selection model (BSM) takes the form

\[
L(\theta, \sigma^2, \sigma_{12}, R_{1,i}, R_{2|1,i}) \\
\propto \prod_{i=1}^{m} \phi \left( \frac{y_i - x_i^T \beta}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \left( 1 - \Phi \left( \frac{x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}}} \right) \right) \\
\times \prod_{i=m+1}^{n} \int \phi \left( \frac{y_i - x_i^T \beta}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left( \frac{x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}}} \right) dy_i.
\]

With additional data from phase II, the likelihood is given by

\[
L(\theta, \sigma^2, \sigma_{12}, y, R_{1,i}, R_{2|1,i}) \\
\propto \prod_{i=1}^{m} \phi \left( \frac{y_i - x_i^T \beta}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \left( 1 - \Phi \left( \frac{x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}}} \right) \right) \\
\times \prod_{i=m+1}^{m+r} \phi \left( \frac{y_i - x_i^T \beta}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left( \frac{x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}}} \right) \\
\times \prod_{i=m+r+1}^{n} \int \phi \left( \frac{y_i - x_i^T \beta}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left( \frac{x_i^T \gamma \sqrt{1 + \sigma_{12}^2/\sigma^2} + \frac{\sigma_{12}(y_i - x_i^T \beta)}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}}} \right) dy_i.
\]

In a traditional selection model when no data in phase II are available, identification of the parameter could be a problem. The rationale of the NSBSM methods is that the additional information provides valuable information for identifying the model parameters. In this study, we propose a probit model for \( R_{1,i} \) but not for \( R_{2|1,i} \). Generally, the NSBSM methods are based on a partial likelihood (Cox, 1972) with the component regarding the selection process \( R_{2|1,i} \) discarded from the analysis. The identification problems remains if the selection model also includes a probit model for \( R_{2|1,i} \); we leave this for future research.

3.4. Prior

Assume the following prior distribution

\[
f(\theta, \sigma_{12}, \sigma^2) = f(\theta)f(\sigma_{12})f(\sigma^2),
\]
where

\[
\begin{align*}
  f(\theta) &\sim \mathcal{MN}(\theta_0, \Psi_0^{-1}), \\
  f(\sigma_{12}) &\sim \mathcal{N}(c_0, d_0^{-1}), \\
  f(\sigma^{-2}) &\sim \mathcal{G}\left(\frac{v_0}{2}, \left(\frac{w_0}{2}\right)^{-1}\right).
\end{align*}
\]

Here, $\mathcal{MN}$, $\mathcal{N}$, $\mathcal{G}$ denote multivariate normal, univariate normal, gamma distribution, respectively. $\mathcal{G}\left(\frac{v_0}{2}, \left(\frac{w_0}{2}\right)^{-1}\right)$ denotes a gamma distribution with shape parameter $v_0$ and scale parameter $\left(\frac{w_0}{2}\right)^{-1}$. Prior parameters $\{\theta_0, \Psi_0, c_0, d_0, v_0, w_0\}$ are specified in simulation study and real data application. Throughout the paper, proper priors are used to ensure that the resulting posterior densities have closed form.

### 3.5. Posterior

Let $y_i^*$ denote augmented outcomes. The conditional means and variances of the bivariate normal variables $(u_i, y_i^*)$ have the following forms:

\[
\begin{align*}
  \mu_{y^*|u} &= x_i^T \beta + \sigma_{12}(u_i - x_i^T \gamma), \sigma_{y^*|u} = \sigma^2, \\
  \mu_{u|y^*} &= x_i^T \gamma + \frac{\sigma_{12}}{\sigma^2 + \sigma_{12}^2}(y_i^* - x_i^T \beta), \sigma_{u|y^*}^2 = 1 - \frac{\sigma_{12}^2}{\sigma^2 + \sigma_{12}^2}.
\end{align*}
\]

Gibbs sampling algorithm including data augmentation (imputation) is summarized as following steps. Additional details and expressions for the parameters of the various posteriors are given in the Appendix.

1. If $R_{1,i} = 1$, $y_i^* = y_i$ and $u_i|(y_i^*, \theta, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{u|y^*}^2; 0, \infty)$, where $\mathcal{TN}$ denotes truncated normal distribution.

2. If $R_{1,i} = 0$ and $R_{2|1,i} = 1$, $y_i^* = y_i$ and $u_i|(y_i^*, \theta, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{u|y^*}^2; -\infty, 0)$.

3. If $R_{1,i} = 0$ and $R_{2|1,i} = 0$, $y_i^*|(\theta, \sigma_{12}, \sigma^2) \sim \mathcal{N}(x_i^T \beta, \sigma^2 + \sigma_{12}^2)$ and $u_i|(y_i^*, \theta, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{u|y^*}^2; -\infty, 0)$. 


4. Sample $\theta$ from $\mathcal{MVN}(\tilde{\theta}, \tilde{\Psi}^{-1})$, where

$$
\tilde{\theta} = \tilde{\Psi}^{-1}[Z'(\Sigma^{-1} \otimes I_n)\tilde{y} + \Psi_0 \theta_0],
$$

$$
\tilde{\Psi} = Z'(\Sigma^{-1} \otimes I_n)Z + \Psi_0,
$$

$$
Z = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix},
$$

where $\tilde{y} = \begin{pmatrix} y^* \\ u \end{pmatrix}$, $Z_1 = Z_2 = (1, x)$.

5. Sample $\sigma_{12}$ from $\mathcal{N}(\tilde{c}, \tilde{d}^{-1})$, where $\tilde{c} = \tilde{d}^{-1}[\sigma^{-2}\eta'\epsilon + c_0d_0]$ and $\tilde{d} = \sigma^{-2}\epsilon'\epsilon + d_0$.

6. Sample $\sigma^{-2}$ from $\mathcal{G}(\tilde{v}, (\tilde{w})^{-1})$, where $\tilde{v} = n + v_0$ and $\tilde{w} = (\eta - \epsilon \sigma_{12})'(\eta - \epsilon \sigma_{12}) + w_0$.

7. Return to Step 1 and repeat.

4. Simulation Studies

This section illustrates the properties of the NSBSM method using simulation studies and compares the performance of NSBSM to other methods under different missing data mechanisms in phase II. For each simulation study, six methods are applied to estimate regression coefficient of simple linear regression model based on outcome $Y$ and covariate $X$:

(1) BD: estimates using the data before deletion, i.e., the full data generated from simulation before missing values are created, as a benchmark method.

(2) CC: complete case analysis using respondents from both phase I and phase II, discarding cases where are still missing after phase II.

(3) IL: ignorable likelihood method through multiple imputation using data from both phase I and phase II, assuming ignorable missingness.

(4) NSMI: multiple imputation in the nonrespondent subsample in phase I using only additional data from phase II.

(5) BSM: Bayesian selection model using data from only phase I.
(6) NSBSM: Bayesian selection model using data from both phase I and phase II.

All six methods except method (5) utilize all the observed data in phase I and II. Method (5) uses only information in phase I to estimate parameters in regression model. This paper compares the performance of each of the methods using empirical bias, standard error (SE), root mean square error (RMSE), and the coverage probability of the 95% highest posterior density (HPD) interval.

The outcome is generated by the linear regression model

\[
y_i = 1 + x_i + z_i + x_i \times z_i + \eta_i, \quad \eta_i \overset{i.i.d.}{\sim} N(0, 1),
\]

where \(x_i\) is sampled from standard normal distribution, \(z_i\) is from Bernoulli distribution with probability 0.5, and \(x_i \times z_i\) is the interaction term, for \(i = 1, 2, \ldots, 1,000\). The response is subject to missingness, while \(x_i\) and \(z_i\) are fully observed. Phase I missing values in \(Y\) are generated based on the following MNAR mechanism

\[
\text{Pr}(R_{1,i} = 0 | y_i, x_i, z_i) = \Phi(-3 \times y_i + 2.5).
\]

This missing data generation scheme results in approximately 35% of the values \(Y\) being missing in phase I.

Let \(R_{2|1,i}\) denote the response indicator in the subsample of nonrespondents in phase I. Phase II responses in \(y_i\) are generated under the following three missing data mechanisms:

(1) MAR: \(\text{Pr}(R_{2|1,i} = 0 | y_i, x_i, z_i) = \Phi(-x_i - z_i - x_i \times z_i + c_1)\),

(2) MNAR: \(\text{Pr}(R_{2|1,i} = 0 | y_i, x_i, z_i) = \Phi(-0.3 \times y_i + x_i + z_i + x_i \times z_i + c_2)\),

(3) MNAR: \(\text{Pr}(R_{2|1,i} = 0 | y_i, x_i, z_i) = \expit(-0.3 \times y_i + x_i + z_i + x_i \times z_i + c_3)\),

where \(c_1, c_2\) and \(c_3\) are assigned different values for various selection proportions in phase II. Set appropriate \(c_1, c_2\) and \(c_3\) so that the corresponding selection proportions of phase II responses in \(Y\) are approximately \{40%, 30%, 20%, 10%\}, and \(\expit(.) = \exp(.)/(1 + \exp(.))\). The third scenario is included to evaluate how the NSBSM method performs when data are not simulated from the same model as the NSBSM method. Let \(k\) denote the dimension of coefficients. Prior
Table 2: Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MAR in phase II (200 replicates, \( \Pr(R_{21,i} = 0 | y_i, x_i, z_i) = \Phi(-x_i - z_i - x_i \times z_i + c_1) \))

<table>
<thead>
<tr>
<th>Phase II</th>
<th>Method</th>
<th>( b_0 )</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>( b_1 )</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>( b_2 )</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>( b_3 )</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>40%</strong></td>
<td>BD</td>
<td>14</td>
<td>31</td>
<td>433</td>
<td>0.956</td>
<td>28</td>
<td>32</td>
<td>447</td>
<td>0.94</td>
<td>17</td>
<td>44</td>
<td>623</td>
<td>0.96</td>
<td>-15</td>
<td>49</td>
<td>694</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>2897</td>
<td>32</td>
<td>2932</td>
<td>0</td>
<td>496</td>
<td>32</td>
<td>669</td>
<td>0.785</td>
<td>-1047</td>
<td>48</td>
<td>1243</td>
<td>0.62</td>
<td>-870</td>
<td>51</td>
<td>1125</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>2969</td>
<td>33</td>
<td>2946</td>
<td>0</td>
<td>501</td>
<td>32</td>
<td>674</td>
<td>0.78</td>
<td>-1060</td>
<td>48</td>
<td>1280</td>
<td>0.645</td>
<td>-877</td>
<td>51</td>
<td>1137</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>168</td>
<td>39</td>
<td>546</td>
<td>0.96</td>
<td>170</td>
<td>33</td>
<td>524</td>
<td>0.935</td>
<td>-179</td>
<td>51</td>
<td>808</td>
<td>0.955</td>
<td>-112</td>
<td>52</td>
<td>755</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1213</td>
<td>45</td>
<td>1289</td>
<td>0.6</td>
<td>-498</td>
<td>38</td>
<td>728</td>
<td>0.89</td>
<td>-332</td>
<td>53</td>
<td>824</td>
<td>0.93</td>
<td>-111</td>
<td>60</td>
<td>860</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>235</td>
<td>33</td>
<td>520</td>
<td>0.92</td>
<td>67</td>
<td>32</td>
<td>457</td>
<td>0.955</td>
<td>-47</td>
<td>46</td>
<td>655</td>
<td>0.955</td>
<td>-95</td>
<td>51</td>
<td>719</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td>BD</td>
<td>-24</td>
<td>32</td>
<td>446</td>
<td>0.955</td>
<td>5</td>
<td>32</td>
<td>455</td>
<td>0.95</td>
<td>-9</td>
<td>45</td>
<td>641</td>
<td>0.96</td>
<td>-46</td>
<td>45</td>
<td>643</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>3697</td>
<td>35</td>
<td>3730</td>
<td>0</td>
<td>128</td>
<td>33</td>
<td>487</td>
<td>0.935</td>
<td>-1555</td>
<td>46</td>
<td>1687</td>
<td>0.965</td>
<td>-816</td>
<td>48</td>
<td>1061</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>3700</td>
<td>35</td>
<td>3734</td>
<td>0</td>
<td>129</td>
<td>34</td>
<td>495</td>
<td>0.99</td>
<td>-1561</td>
<td>47</td>
<td>1697</td>
<td>0.975</td>
<td>-818</td>
<td>48</td>
<td>1063</td>
<td>0.715</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>169</td>
<td>47</td>
<td>501</td>
<td>0.935</td>
<td>190</td>
<td>34</td>
<td>567</td>
<td>0.93</td>
<td>-226</td>
<td>47</td>
<td>1260</td>
<td>0.645</td>
<td>-877</td>
<td>51</td>
<td>1137</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1061</td>
<td>40</td>
<td>1201</td>
<td>0.645</td>
<td>-488</td>
<td>39</td>
<td>740</td>
<td>0.88</td>
<td>-341</td>
<td>48</td>
<td>764</td>
<td>0.96</td>
<td>-187</td>
<td>53</td>
<td>766</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>297</td>
<td>35</td>
<td>520</td>
<td>0.96</td>
<td>31</td>
<td>33</td>
<td>462</td>
<td>0.955</td>
<td>-113</td>
<td>46</td>
<td>664</td>
<td>0.98</td>
<td>-147</td>
<td>45</td>
<td>657</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>20%</strong></td>
<td>BD</td>
<td>20</td>
<td>31</td>
<td>490</td>
<td>0.955</td>
<td>-61</td>
<td>32</td>
<td>455</td>
<td>0.95</td>
<td>-11</td>
<td>40</td>
<td>564</td>
<td>0.95</td>
<td>71</td>
<td>43</td>
<td>608</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>4590</td>
<td>31</td>
<td>4615</td>
<td>0</td>
<td>537</td>
<td>37</td>
<td>747</td>
<td>0.78</td>
<td>-2050</td>
<td>43</td>
<td>2139</td>
<td>0.105</td>
<td>-438</td>
<td>53</td>
<td>866</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>4591</td>
<td>32</td>
<td>4615</td>
<td>0</td>
<td>-535</td>
<td>37</td>
<td>749</td>
<td>0.79</td>
<td>-2057</td>
<td>44</td>
<td>2146</td>
<td>0.13</td>
<td>-443</td>
<td>54</td>
<td>873</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>331</td>
<td>63</td>
<td>561</td>
<td>0.955</td>
<td>148</td>
<td>39</td>
<td>625</td>
<td>0.94</td>
<td>242</td>
<td>64</td>
<td>1213</td>
<td>0.94</td>
<td>-156</td>
<td>49</td>
<td>748</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1138</td>
<td>44</td>
<td>1294</td>
<td>0.595</td>
<td>-576</td>
<td>41</td>
<td>814</td>
<td>0.81</td>
<td>-376</td>
<td>47</td>
<td>761</td>
<td>0.945</td>
<td>-37</td>
<td>51</td>
<td>767</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>316</td>
<td>34</td>
<td>707</td>
<td>0.86</td>
<td>-120</td>
<td>36</td>
<td>526</td>
<td>0.945</td>
<td>-200</td>
<td>42</td>
<td>630</td>
<td>0.955</td>
<td>24</td>
<td>47</td>
<td>662</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td>BD</td>
<td>40</td>
<td>32</td>
<td>456</td>
<td>0.95</td>
<td>-36</td>
<td>31</td>
<td>441</td>
<td>0.955</td>
<td>-15</td>
<td>47</td>
<td>660</td>
<td>0.945</td>
<td>-60</td>
<td>47</td>
<td>660</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>5229</td>
<td>38</td>
<td>5357</td>
<td>0</td>
<td>-1141</td>
<td>39</td>
<td>1209</td>
<td>0.835</td>
<td>-2458</td>
<td>53</td>
<td>2570</td>
<td>0.065</td>
<td>-249</td>
<td>55</td>
<td>813</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>5336</td>
<td>39</td>
<td>5365</td>
<td>0</td>
<td>-1149</td>
<td>40</td>
<td>1281</td>
<td>0.415</td>
<td>-2472</td>
<td>54</td>
<td>2591</td>
<td>0.08</td>
<td>-233</td>
<td>57</td>
<td>826</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>309</td>
<td>40</td>
<td>1420</td>
<td>0.95</td>
<td>-258</td>
<td>42</td>
<td>790</td>
<td>0.85</td>
<td>-441</td>
<td>54</td>
<td>885</td>
<td>0.905</td>
<td>-129</td>
<td>60</td>
<td>859</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1183</td>
<td>117</td>
<td>1336</td>
<td>0.565</td>
<td>-141</td>
<td>36</td>
<td>522</td>
<td>0.925</td>
<td>-324</td>
<td>53</td>
<td>816</td>
<td>0.895</td>
<td>-77</td>
<td>51</td>
<td>721</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>679</td>
<td>44</td>
<td>590</td>
<td>0.74</td>
<td>-141</td>
<td>36</td>
<td>522</td>
<td>0.925</td>
<td>-324</td>
<td>53</td>
<td>816</td>
<td>0.895</td>
<td>-77</td>
<td>51</td>
<td>721</td>
<td>0.95</td>
</tr>
</tbody>
</table>

\( b_0 \) is the intercept. \( b_1, b_2, b_3 \) are the coefficients of \( x_i, z_i, x_i \times z_i \), respectively. The true value of these parameters is 1.

Parameters are set as \( \{ \theta = 0, \Psi^{-1} = 100 \cdot I_k, c_0 = 0, d_0 = 1, v_0 = 10, w_0 = 10 \} \) to obtain a balanced variance-covariance matrix \( \Sigma \) and comparable variability for both \( \sigma^2 \) and \( \sigma^2_{12} \) (Li, 1998). Results are based on 200 repetitions for each simulated condition. For each data set the Markov chain Monte Carlo (MCMC) algorithm is run for 12,000 iterations, where the first 2000 draws are discarded as burn-in period. The Gelman-Rubin statistics and trace plot suggest that the Markov chain is mixing well. The biases, RMSEs, and coverage probabilities of the 95% HPD intervals from MAR and MNAR of phase II response are reported. CC and IL yield biased estimates of the regression coefficient since phase I missingness is MNAR; therefore we focus on the results from NSMI, BSM and NSBSM in the following discussion.
Table 3: Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MNAR in phase II (200 replicates, $\Pr(R_{21,i} = 0 | y_i, x_i, z_i) = \Phi(-0.3 \times y_i + x_i + z_i + x_i \times z_i + c_2)$)

<table>
<thead>
<tr>
<th>Phase II Sampling Proportion</th>
<th>Method</th>
<th>$b_0$</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>$b_1$</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>$b_2$</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
<th>$b_3$</th>
<th>SE</th>
<th>RMSE</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>40% BD</td>
<td>18</td>
<td>.33</td>
<td>45</td>
<td>0.94</td>
<td>.01</td>
<td>1</td>
<td>31</td>
<td>44</td>
<td>0.93</td>
<td>2</td>
<td>46</td>
<td>64</td>
<td>0.93</td>
<td>-3</td>
<td>44</td>
<td>62</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>2654</td>
<td>43</td>
<td>2721</td>
<td>0.005</td>
<td>-1875</td>
<td>44</td>
<td>1975</td>
<td>0.14</td>
<td>-1051</td>
<td>56</td>
<td>1312</td>
<td>0.735</td>
<td>185</td>
<td>61</td>
<td>881</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>2661</td>
<td>43</td>
<td>2729</td>
<td>0.01</td>
<td>-1882</td>
<td>45</td>
<td>1989</td>
<td>0.165</td>
<td>-1062</td>
<td>57</td>
<td>1332</td>
<td>0.725</td>
<td>195</td>
<td>64</td>
<td>920</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>-614</td>
<td>45</td>
<td>880</td>
<td>0.765</td>
<td>512</td>
<td>50</td>
<td>873</td>
<td>0.84</td>
<td>335</td>
<td>62</td>
<td>931</td>
<td>0.915</td>
<td>-148</td>
<td>73</td>
<td>1069</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1171</td>
<td>46</td>
<td>1336</td>
<td>0.575</td>
<td>-540</td>
<td>40</td>
<td>777</td>
<td>0.83</td>
<td>-370</td>
<td>52</td>
<td>822</td>
<td>0.93</td>
<td>-108</td>
<td>54</td>
<td>765</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>-187</td>
<td>39</td>
<td>546</td>
<td>0.91</td>
<td>166</td>
<td>39</td>
<td>526</td>
<td>0.955</td>
<td>92</td>
<td>52</td>
<td>745</td>
<td>0.97</td>
<td>-37</td>
<td>54</td>
<td>762</td>
<td>0.97</td>
</tr>
<tr>
<td>30% BD</td>
<td>17</td>
<td>34</td>
<td>475</td>
<td>0.915</td>
<td>.01</td>
<td>23</td>
<td>34</td>
<td>479</td>
<td>0.945</td>
<td>-8</td>
<td>46</td>
<td>655</td>
<td>0.935</td>
<td>-81</td>
<td>47</td>
<td>667</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>3340</td>
<td>36</td>
<td>3378</td>
<td>0</td>
<td>468</td>
<td>36</td>
<td>688</td>
<td>0.795</td>
<td>-1376</td>
<td>49</td>
<td>1540</td>
<td>0.49</td>
<td>-965</td>
<td>49</td>
<td>1184</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>3340</td>
<td>36</td>
<td>3378</td>
<td>0</td>
<td>471</td>
<td>36</td>
<td>695</td>
<td>0.815</td>
<td>-1368</td>
<td>50</td>
<td>1537</td>
<td>0.485</td>
<td>-973</td>
<td>49</td>
<td>1193</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>-578</td>
<td>51</td>
<td>920</td>
<td>0.89</td>
<td>239</td>
<td>38</td>
<td>582</td>
<td>0.905</td>
<td>645</td>
<td>70</td>
<td>1579</td>
<td>0.93</td>
<td>-99</td>
<td>52</td>
<td>737</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1135</td>
<td>42</td>
<td>1278</td>
<td>0.605</td>
<td>-698</td>
<td>40</td>
<td>752</td>
<td>0.87</td>
<td>-383</td>
<td>53</td>
<td>846</td>
<td>0.925</td>
<td>-171</td>
<td>55</td>
<td>798</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>15</td>
<td>36</td>
<td>530</td>
<td>0.945</td>
<td>210</td>
<td>35</td>
<td>534</td>
<td>0.91</td>
<td>16</td>
<td>51</td>
<td>718</td>
<td>0.935</td>
<td>-92</td>
<td>48</td>
<td>707</td>
<td>0.945</td>
</tr>
<tr>
<td>20% BD</td>
<td>10</td>
<td>32</td>
<td>452</td>
<td>0.925</td>
<td>.01</td>
<td>-12</td>
<td>30</td>
<td>419</td>
<td>0.97</td>
<td>39</td>
<td>47</td>
<td>659</td>
<td>0.94</td>
<td>39</td>
<td>42</td>
<td>596</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>4719</td>
<td>47</td>
<td>4766</td>
<td>0</td>
<td>-794</td>
<td>72</td>
<td>1292</td>
<td>0.7</td>
<td>-2015</td>
<td>53</td>
<td>2148</td>
<td>0.2</td>
<td>-449</td>
<td>51</td>
<td>849</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>4730</td>
<td>49</td>
<td>4780</td>
<td>0</td>
<td>-804</td>
<td>74</td>
<td>1313</td>
<td>0.705</td>
<td>-2030</td>
<td>54</td>
<td>2168</td>
<td>0.245</td>
<td>-435</td>
<td>51</td>
<td>845</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>-1077</td>
<td>89</td>
<td>1653</td>
<td>0.87</td>
<td>515</td>
<td>49</td>
<td>861</td>
<td>0.885</td>
<td>880</td>
<td>16</td>
<td>1862</td>
<td>0.91</td>
<td>-136</td>
<td>75</td>
<td>1068</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1161</td>
<td>44</td>
<td>1319</td>
<td>0.615</td>
<td>-577</td>
<td>37</td>
<td>773</td>
<td>0.855</td>
<td>-347</td>
<td>57</td>
<td>882</td>
<td>0.89</td>
<td>-64</td>
<td>53</td>
<td>754</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>201</td>
<td>42</td>
<td>624</td>
<td>0.925</td>
<td>104</td>
<td>39</td>
<td>490</td>
<td>0.96</td>
<td>-32</td>
<td>54</td>
<td>775</td>
<td>0.94</td>
<td>-56</td>
<td>47</td>
<td>671</td>
<td>0.945</td>
</tr>
<tr>
<td>10% BD</td>
<td>16</td>
<td>31</td>
<td>444</td>
<td>0.955</td>
<td>.01</td>
<td>-53</td>
<td>31</td>
<td>447</td>
<td>0.965</td>
<td>13</td>
<td>46</td>
<td>645</td>
<td>0.945</td>
<td>8</td>
<td>46</td>
<td>650</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>4940</td>
<td>36</td>
<td>4974</td>
<td>0</td>
<td>-744</td>
<td>37</td>
<td>905</td>
<td>0.685</td>
<td>-2286</td>
<td>49</td>
<td>2390</td>
<td>0.09</td>
<td>-299</td>
<td>52</td>
<td>795</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>1054</td>
<td>37</td>
<td>4982</td>
<td>0</td>
<td>-755</td>
<td>38</td>
<td>924</td>
<td>0.68</td>
<td>-2208</td>
<td>52</td>
<td>2414</td>
<td>0.11</td>
<td>-281</td>
<td>53</td>
<td>802</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>NSMI</td>
<td>-1011</td>
<td>116</td>
<td>1925</td>
<td>0.91</td>
<td>454</td>
<td>44</td>
<td>772</td>
<td>0.905</td>
<td>900</td>
<td>2244</td>
<td>1522</td>
<td>0.92</td>
<td>-56</td>
<td>61</td>
<td>868</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>BSM</td>
<td>1160</td>
<td>43</td>
<td>1311</td>
<td>0.59</td>
<td>-699</td>
<td>39</td>
<td>818</td>
<td>0.83</td>
<td>-407</td>
<td>48</td>
<td>793</td>
<td>0.945</td>
<td>16</td>
<td>53</td>
<td>781</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>NSBSM</td>
<td>313</td>
<td>38</td>
<td>616</td>
<td>0.935</td>
<td>61</td>
<td>35</td>
<td>491</td>
<td>0.975</td>
<td>-163</td>
<td>49</td>
<td>715</td>
<td>0.95</td>
<td>-12</td>
<td>47</td>
<td>698</td>
<td>0.955</td>
</tr>
</tbody>
</table>

$b_0$ is the intercept. $b_1$, $b_2$, $b_3$ are the coefficients of $x_i$, $z_i$, $x_i \times z_i$, respectively. The true value of these parameters is 1.
Table 4: Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MNAR in phase II (200 replicates, logit(Pr(R21,i = 0|yi, xi, zi)) = −0.3 × yi + xi + zi + xi × zi + c3))

<table>
<thead>
<tr>
<th>Sampling Proportion</th>
<th>Method</th>
<th>b0</th>
<th>SE (×10^5)</th>
<th>RMSSE (×10^4)</th>
<th>CR</th>
<th>b1</th>
<th>SE (×10^5)</th>
<th>RMSSE (×10^4)</th>
<th>CR</th>
<th>b2</th>
<th>SE (×10^5)</th>
<th>RMSSE (×10^4)</th>
<th>CR</th>
<th>b3</th>
<th>SE (×10^5)</th>
<th>RMSSE (×10^4)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>40%</td>
<td>BD -24</td>
<td>34</td>
<td>479.093</td>
<td>33</td>
<td>407.955</td>
<td>35</td>
<td>43</td>
<td>613.960</td>
<td>-31</td>
<td>46</td>
<td>646.940</td>
<td>-96</td>
<td>54</td>
<td>843.955</td>
<td>-56</td>
<td>56</td>
<td>802.955</td>
</tr>
<tr>
<td>CC 2133 39 2231 0.045</td>
<td>2035</td>
<td>39</td>
<td>1175.555</td>
<td>-803</td>
<td>54</td>
<td>1195.0.8</td>
<td>-96</td>
<td>54</td>
<td>843.955</td>
<td>-56</td>
<td>56</td>
<td>802.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL 2125 40 2199 0.045</td>
<td>-1028</td>
<td>41</td>
<td>1179.500</td>
<td>-867</td>
<td>55</td>
<td>1117.0.84</td>
<td>-91</td>
<td>62</td>
<td>881.945</td>
<td>-84</td>
<td>64</td>
<td>913.945</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSMI -666 41 846 0.79</td>
<td>500</td>
<td>43</td>
<td>784.460</td>
<td>195</td>
<td>56</td>
<td>820.935</td>
<td>-84</td>
<td>64</td>
<td>913.945</td>
<td>-84</td>
<td>64</td>
<td>913.945</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSM 1116 40 1250 0.065</td>
<td>-537</td>
<td>40</td>
<td>797.450</td>
<td>-366</td>
<td>49</td>
<td>777.960</td>
<td>-91</td>
<td>56</td>
<td>802.955</td>
<td>-91</td>
<td>56</td>
<td>802.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSBSM -323 37 621 0.925</td>
<td>258</td>
<td>38</td>
<td>594.920</td>
<td>108</td>
<td>49</td>
<td>702.960</td>
<td>-56</td>
<td>55</td>
<td>772.955</td>
<td>-56</td>
<td>55</td>
<td>772.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td>BD 46</td>
<td>34</td>
<td>480.892</td>
<td>-2</td>
<td>32</td>
<td>454.95</td>
<td>-82</td>
<td>44</td>
<td>628.945</td>
<td>11</td>
<td>46</td>
<td>647.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC 1079 41 3314 0</td>
<td>-1325</td>
<td>44</td>
<td>1647.385</td>
<td>-1230</td>
<td>53</td>
<td>1444.655</td>
<td>-65</td>
<td>61</td>
<td>861.955</td>
<td>-65</td>
<td>61</td>
<td>861.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL 3063 42 3124 0</td>
<td>-1514</td>
<td>45</td>
<td>1644.325</td>
<td>-1223</td>
<td>55</td>
<td>1451.67</td>
<td>-71</td>
<td>62</td>
<td>879.965</td>
<td>-71</td>
<td>62</td>
<td>879.965</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSMI -671 45 927 0.79</td>
<td>498</td>
<td>51</td>
<td>872.845</td>
<td>193</td>
<td>63</td>
<td>903.935</td>
<td>-127</td>
<td>77</td>
<td>1008.995</td>
<td>-127</td>
<td>77</td>
<td>1008.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSM 1146 42 1291 0.655</td>
<td>-534</td>
<td>41</td>
<td>798.965</td>
<td>-426</td>
<td>54</td>
<td>872.930</td>
<td>-70</td>
<td>56</td>
<td>796.955</td>
<td>-70</td>
<td>56</td>
<td>796.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSBSM -185 40 594.944</td>
<td>164</td>
<td>40</td>
<td>586.94</td>
<td>18</td>
<td>53</td>
<td>727.95</td>
<td>18</td>
<td>53</td>
<td>780.955</td>
<td>18</td>
<td>53</td>
<td>780.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>BD 18</td>
<td>30</td>
<td>427.966</td>
<td>48</td>
<td>52</td>
<td>455.965</td>
<td>-10</td>
<td>42</td>
<td>600.955</td>
<td>-31</td>
<td>50</td>
<td>701.945</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC 3674 39 3913 0</td>
<td>-1858</td>
<td>37</td>
<td>1910.12</td>
<td>-1466</td>
<td>53</td>
<td>1664.55</td>
<td>-218</td>
<td>65</td>
<td>941.925</td>
<td>-218</td>
<td>65</td>
<td>941.925</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL 3805 39 3923 0</td>
<td>-1853</td>
<td>37</td>
<td>1922.14</td>
<td>-1478</td>
<td>52</td>
<td>1653.55</td>
<td>-205</td>
<td>66</td>
<td>932.995</td>
<td>-205</td>
<td>66</td>
<td>932.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSMI -924 50 1164 0.695</td>
<td>717</td>
<td>39</td>
<td>1049.84</td>
<td>490</td>
<td>69</td>
<td>1087.91</td>
<td>-402</td>
<td>93</td>
<td>1366.915</td>
<td>-402</td>
<td>93</td>
<td>1366.915</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSM 1116 45 1281 0.61</td>
<td>-491</td>
<td>41</td>
<td>752.875</td>
<td>-441</td>
<td>53</td>
<td>849.910</td>
<td>-67</td>
<td>60</td>
<td>854.995</td>
<td>-67</td>
<td>60</td>
<td>854.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSBSM -152 40 579.94</td>
<td>177</td>
<td>36</td>
<td>584.935</td>
<td>59</td>
<td>59</td>
<td>714.955</td>
<td>-53</td>
<td>60</td>
<td>847.995</td>
<td>-53</td>
<td>60</td>
<td>847.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>BD 26</td>
<td>30</td>
<td>424.966</td>
<td>-13</td>
<td>32</td>
<td>446.960</td>
<td>-28</td>
<td>40</td>
<td>566.975</td>
<td>-31</td>
<td>40</td>
<td>569.975</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC 5414 37 5459 0</td>
<td>-2556</td>
<td>41</td>
<td>2624.02</td>
<td>-2059</td>
<td>59</td>
<td>2178.195</td>
<td>-210</td>
<td>64</td>
<td>930.919</td>
<td>-210</td>
<td>64</td>
<td>930.919</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL 5451 38 5471 0</td>
<td>-2565</td>
<td>43</td>
<td>2617.03</td>
<td>-2072</td>
<td>52</td>
<td>2198.26</td>
<td>-199</td>
<td>67</td>
<td>961.992</td>
<td>-199</td>
<td>67</td>
<td>961.992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSMI -926 66 5160 0.85</td>
<td>585</td>
<td>76</td>
<td>1241.885</td>
<td>376</td>
<td>103</td>
<td>1494.92</td>
<td>-205</td>
<td>129</td>
<td>1830.995</td>
<td>-205</td>
<td>129</td>
<td>1830.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSM 1396 42 1337 0.55</td>
<td>-513</td>
<td>39</td>
<td>782.835</td>
<td>-425</td>
<td>49</td>
<td>821.910</td>
<td>-42</td>
<td>54</td>
<td>793.995</td>
<td>-42</td>
<td>54</td>
<td>793.995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSBSM 306 40 641.995</td>
<td>-134</td>
<td>40</td>
<td>580.96</td>
<td>-142</td>
<td>50</td>
<td>715.950</td>
<td>36</td>
<td>58</td>
<td>813.935</td>
<td>36</td>
<td>58</td>
<td>813.935</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b0 is the intercept, b1, b2, b3 are the coefficients of x, z, x × z, respectively. The true value of these parameters is 1.
For the data simulated based on MAR mechanism in phase II (Table 2), both NSMI and NSBSM yield approximately unbiased estimates of the regression coefficients. The NSMI method has moderately larger SEs and RMSEs because of increased variability in the imputed values, which uses subjects from Pattern 2, 3 and 4, but not subjects from Pattern 1 (Zhang, Chen, and Elliott, 2016). The BSM method shows significant bias in estimating the regression coefficients. This is not surprising because of the identification problem with the traditional selection model. By varying the proportion of being sampled in Phase II, we see the precision increases as the sampling proportion in Phase II increases. Note that even sampling 10% of the nonrespondents in phase I is enough to distinguish the NSBSM and NSMI method from other competing methods.

When the Phase II missing data mechanism is MNAR (Table 3), the NSBSM is the only method that provides unbiased estimates of the regression. All other methods show significant biases because the MAR assumptions are violated. We again see that precision increases as sampling proportion in Phase II increases, and the improvement is substantial even if we only collect data from 10% of the nonrespondents in phase I.

When the Phase II missing data mechanism is simulated based on the logit model (Table 4) rather than a probit model, the NSBSM still outperforms competing methods. This implies that the NSBSM method is robust to slight violation of the model assumptions.

5. Application to QOL Dataset

We applied the proposed method to a quality of life (QOL) dataset from a community-based study — the Children in the Community study (CIC) (Cohen, Crawford, Johnson, and Kasen, 2005). A brief description of this dataset can be found in Chen and Cohen (2006). The 750 participants sample was originally drawn from 100 neighborhoods in two upstate New York counties in 1975 (Cohen, Crawford, Johnson, and Kasen, 2005). From 1991 to 1994 (T1), these 750 youths (mean age of 22.0 years and SD of 2.8 years) were interviewed in the home by trained interviewers. QOL was assessed using the young adult quality of life instrument (YAQOL) (Chen, Cohen, Kasen, Gordan, Dufur, and Smailes, 2004). In 2001-2004 (T2) at mean age of 32.0 years (SD=2.8 years), the same
Table 5: Posterior means, lower and upper 95% HPD intervals of QOL analysis

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>IL</th>
<th>NSMI</th>
<th>BSM</th>
<th>NSBSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.195 (4.024, 4.365)</td>
<td>4.193 (4.014, 4.373)</td>
<td>4.167 (3.935, 4.399)</td>
<td>4.193 (4.010, 4.396)</td>
<td>4.194 (4.006, 4.387)</td>
</tr>
<tr>
<td>Sex (male vs. female)</td>
<td>0.009 (-0.030, 0.047)</td>
<td>0.009 (-0.030, 0.046)</td>
<td>0.003 (-0.046, 0.055)</td>
<td>0.015 (-0.030, 0.061)</td>
<td>0.009 (-0.034, 0.053)</td>
</tr>
<tr>
<td>Race (White vs. non-White)</td>
<td>0.091 (0.021, 0.150)</td>
<td>0.090 (0.021, 0.150)</td>
<td>0.086 (0.021, 0.168)</td>
<td>0.094 (0.021, 0.172)</td>
<td>0.091 (0.016, 0.169)</td>
</tr>
<tr>
<td>Education (≥HS vs. &lt;HS)</td>
<td>0.089 (0.049, 0.130)</td>
<td>0.089 (0.048, 0.130)</td>
<td>0.081 (0.021, 0.141)</td>
<td>0.092 (0.044, 0.137)</td>
<td>0.089 (0.043, 0.133)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.003 (-0.010, 0.005)</td>
<td>-0.003 (-0.010, 0.005)</td>
<td>-0.003 (-0.009, 0.005)</td>
<td>-0.003 (-0.011, 0.006)</td>
<td>-0.003 (-0.011, 0.006)</td>
</tr>
</tbody>
</table>

A group of participants was surveyed using the same YAQOL instrument. Of the 750 participants assessed for QOL at T1, 603 (80.4%) completed the survey at T2 while 147 did not respond to the survey at T2 (phase I). For those 147 subjects who did not respond to the survey, an abridged version of the YAQOL instrument was mailed to their home address (phase II). Subjects were paid for their participation upon return of the completed surveys. Of the 147 eligible subjects, 39 (26.5%) returned their YAQOL instrument. Since phase II data collection was completed within three months of phase I, it was assumed that the YAQOL outcomes remained unchanged from phase I. We focus our analysis on the resources subscale of the YAQOL instrument.

The goal of the QOL analysis is to determine whether the resources subscale at T2 is related to major demographic variables — age, gender, race and education. We regressed the resources on gender (male versus female), age (in years), race (White vs. non-White), and education (High school or above vs. less than high school). We applied the CC analysis, IL method using multiple imputation, the NSMI method, the BSM method, and NSBSM method to the dataset. The NSMI method is valid if, among the 147 nonrespondents, the missingness after phase II is MAR, meeting the assumptions for NSMI. When phase I missingness is MNAR, the NSBSM method will be more efficient than the BSM because the additional data collected from phase II provide valuable information for modeling the missing data mechanism and improve the identification of the model parameters. When both phase I and phase II missingness are MNAR, the NSBSM method is the only method that provides valid estimation of the regression model for the resources subscale.

To correct for the skewness of the outcome and improve the posterior esti-
mation, the resources subscales was log-transformed. The results from all five methods are shown in Table 5. The results are quite consistent across five different methods, although the effect of sex and age are a little weaker for the NSMI method. All methods showed a significant effect of age with White race having significantly more resources than non-White participants. Those with at least high-school education also had significantly more resources at an mean age of 33 compared to those with less than high school education.

6. Discussion

Two-phase sampling has been proposed and used in surveys to adjust for nonresponse bias for more than five decades. The traditional methods (i.e., weighting) fail to make full use of the additional data collected in the second phase. Also, little research has demonstrated the utility of NTS sampling and answered the question what proportion of the nonrespondents should be surveyed in the second phase. Zhang, Chen, and Elliott (2016) provided an efficient NSMI method which yields valid estimates when the missing data mechanism in the subsample of nonrespondents is MAR, regardless of the missing data mechanism in phase I. In this paper, we proposed the NSBSM method which improved over the NSMI methods by yielding valid inference even when the missing data mechanism in phase II is MNAR. Our simulation studies also showed that it is beneficial even by collecting data from a small proportion of the nonrespondents.

Prior literature in missing data has primarily focused on preventing and minimizing nonresponse in the data collection stage (Groves and Couper, 2012), and developing methods to handle missing data after the data collection (Little and Rubin, 2002). Many of the methods rely on assumptions that are untestable, which motivated the research in sensitivity analyses (Troxel, Ma, and Heitjan, 2004; Scharfstein, McDermott, Olson, and Wiegand, 2014). Two-phase sampling provides an alternative way to remedy nonresponse. The utilities of NTS sampling are two-fold: (1) it minimizes nonresponse by collecting additional data from nonrespondents; (2) the additional data from the initial nonrespondents provide valuable information regarding the missing data mechanism and therefore improve the modeling of the missingness. The proposed NSBSM methods and the NSMI methods are effective means to make use of the additional data
from phase II sampling.

In theory, the NSBSM method could be extended to a full selection model specification by including a selection model for the missingness indicator in phase II, conditional on missingness in phase I. Such models are subject to the same identification issues with the traditional selection models. Other approaches, such as the multiply robust estimators (Han, 2014), the instrumental variable approach (Wang, Shao, and Kim, 2014), and the pattern mixture model developed for repeated attempt design (Daniels, Jackson, Feng, and White, 2015), may be alternative methods to a full selection model specification. The proposed NSBSM could be extended by modeling the covariance structure following Barnard, McCulloch, and Meng (2000) to incorporate prior information in the posterior inference. The parametric assumptions of the NSBSM model could be extended to a $t$-model (Marchenko and Genton, 2012) for the error distributions or be relaxed through nonparametric approach studied in Chib, Greenberg, and Jeliazkov (2009). For data arising from complex survey setting, the weights can be incorporated in the pseudo-likelihood to reflect the design feature (Chambers, Steel, Wang, and Welsh, 2012). These extensions will be the subject of future work.

Acknowledgements

We thank the Associate Editor and the three referees for their thoughtful and constructive comments which greatly improved the paper.

Appendix: MCMC Sampling Algorithm

We provide some additional details about MCMC sampling algorithm regarding to the augmented likelihood

$$L(\theta, \sigma^2, \sigma_{12} | \tilde{y}) \propto |\Sigma \otimes I_n|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\tilde{y} - Z\theta)^T(\Sigma \otimes I_n)^{-1}(\tilde{y} - Z\theta) \right\},$$

where $\tilde{y} = \begin{pmatrix} y^* \\ u \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma^2 + \sigma_{12}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$, and $I_n$ is identity matrix with dimension $n$.

Gibbs sampling on $\theta$
The full conditional distribution of $\theta$ is of the form
\[
p(\theta|\bar{y}, \sigma^2, \sigma_{12}) \propto |\Sigma \otimes I_n|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\bar{y} - Z\theta)^T(\Sigma \otimes I_n)^{-1}(\bar{y} - Z\theta) \right\} \\
\times |\Psi^{-1}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\theta - \theta_0)^T \Psi(\theta - \theta_0) \right\} \\
\propto \exp \left\{ -\frac{1}{2}(\bar{y} - Z\theta)^T \Sigma^{-1} \otimes I_n (\bar{y} - Z\theta) - \frac{1}{2}(\theta - \theta_0)^T \Psi(\theta - \theta_0) \right\} \\
\propto \exp \left\{ -\frac{1}{2}(-\bar{y}^T \Sigma^{-1} \otimes I_n Z(\theta - \theta^T \Sigma^{-1} \otimes I_n \bar{y}) + \theta^T \Psi(\theta - \theta^T \Psi(\theta - \theta_0))) \right\} \\
\propto \exp \left\{ -\frac{1}{2}(\theta - \tilde{\theta})^T \Psi(\theta - \tilde{\theta}) \right\},
\]
where $\tilde{\theta} = \tilde{\Psi}^{-1}(Z'\Sigma^{-1} \otimes I_n)\bar{y} + \Psi_0\theta_0$ and $\tilde{\Psi} = Z'(\Sigma^{-1} \otimes I_n)Z + \Psi_0$.

**Gibbs sampling on $\sigma_{12}$**

The full conditional distribution of $\sigma_{12}$ is given by
\[
p(\sigma_{12}|\cdot) \propto |\Sigma \otimes I_n|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\bar{y} - Z\theta)^T(\Sigma \otimes I_n)^{-1}(\bar{y} - Z\theta) \right\} \\
\times \exp \left\{ -\frac{1}{2}d_0(\sigma_{12} - c_0)^2 \right\}
\]

Since $\bar{y} - Z\theta = \begin{pmatrix} \eta \\ \epsilon \end{pmatrix}$ and $\Sigma^{-1} = \begin{pmatrix} 1/\sigma^2 & -\sigma_{12}/\sigma^2 \\ -\sigma_{12}/\sigma^2 & 1 + \sigma_{12}^2/\sigma^2 \end{pmatrix}$,
\[
p(\sigma_{12}|\cdot) \propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} \eta \\ \epsilon \end{pmatrix}^T \begin{pmatrix} 1/\sigma^2 & -\sigma_{12}/\sigma^2 \\ -\sigma_{12}/\sigma^2 & 1 + \sigma_{12}^2/\sigma^2 \end{pmatrix} \otimes I_n \begin{pmatrix} \eta \\ \epsilon \end{pmatrix} \right\} \\
\times \exp \left\{ -\frac{1}{2}d_0(\sigma_{12} - c_0)^2 \right\}
\]
\[
\propto \exp \left\{ -\frac{1}{2} (\epsilon^T \epsilon \sigma_{12}^2/\sigma^2 - 2\eta^T \sigma_{12}/\sigma^2 + d_0(\sigma_{12} - c_0)^2) \right\}
\]
\[
\propto \exp \left\{ -\frac{1}{2}d(\sigma_{12} - \bar{c})^2 \right\},
\]
where $\bar{c} = \bar{d}^{-1}[\sigma^{-2}\eta'\epsilon + c_0d_0]$ and $\bar{d} = \sigma^{-2}\epsilon'\epsilon + d_0$.

**Gibbs sampling on $\sigma^{-2}$**

...
The full conditional distribution of \( \sigma_{12} \) takes the form

\[
p(\sigma^{-2}|\cdot) \propto |\Sigma \otimes I_n|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\hat{y} - Z\theta)^T (\Sigma \otimes I_n)^{-1}(\hat{y} - Z\theta) \right\} \\
\times (\sigma^{-2})^{\frac{m+n}{2}-1} \exp \left\{ -\frac{w_0}{2\sigma^{-2}} \right\} \\
\propto (\sigma^{-2})^{\frac{m+n}{2}} \exp \left\{ -\frac{1}{2}\sigma^{-2}(\eta^T \eta - 2\eta^T \epsilon \sigma_{12} + \epsilon^T \epsilon \sigma_{12}^2) \right\} \\
\times (\sigma^{-2})^{\frac{m+n}{2}-1} \exp \left\{ -\frac{w_0}{2\sigma^{-2}} \right\} \\
\propto (\sigma^{-2})^{\frac{m+n}{2}+n} \exp \left\{ -\frac{w_0 + (\eta - \epsilon \sigma_{12})^T (\eta - \epsilon \sigma_{12})}{2\sigma^{-2}} \right\}.
\]

References


REFERENCES


Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, PR China

E-mail: yue.zhang@sjtu.edu.cn

Department of Epidemiology and Biostatistics, University of South Florida, Tampa, FL 33612, USA

E-mail: hchen1@health.usf.edu

Division of Biostatistics and Epidemiology, Cincinnati Childrens Hospital Medical Center, Cincinnati, OH 45229, USA Department of Pediatrics, University of Cincinnati College of
REFERENCES

Medicine, Cincinnati, OH 45229, USA
E-mail: Nanhua.Zhang@cchmc.org