Published in final edited form as:

Stat Med. 2015 May 20; 34(11): 1925–1939. doi:10.1002/sim.6456.

Accounting for Informatively Missing Data in Logistic Regression by Means of Reassessment Sampling

Ji Lin¹ and Robert H. Lyles²

¹ Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285

² Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322

Abstract

We explore the “reassessment” design in a logistic regression setting, where a second wave of sampling is applied to recover a portion of the missing data on a binary exposure and/or outcome variable. We construct a joint likelihood function based on the original model of interest and a model for the missing data mechanism, with emphasis on non-ignorable missingness. The estimation is carried out by numerical maximization of the joint likelihood function with close approximation of the accompanying Hessian matrix, using sharable programs that take advantage of general optimization routines in standard software. We show how likelihood ratio tests can be used for model selection, and how they facilitate direct hypothesis testing for whether missingness is at random. Examples and simulations are presented to demonstrate the performance of the proposed method.

Keywords

Binary data; logistic regression; maximum likelihood; non-ignorable missingness

1. Introduction

Missing data are commonly encountered in epidemiologic and clinical research, e.g., due to survey non-response, study subjects failing to report for evaluations, respondents refusing to answer certain items on a questionnaire, and loss of data. Commonly used approaches for dealing with missing data are based on the missing at random (MAR) assumption [1], which posits that missingness of a variable depends only on observed information. Imputation (e.g. multiple imputation (MI)) or likelihood-based methods that rely on MAR will often provide estimates with reduced bias relative to the usual complete-case (CC) analysis in regression settings. However, these approaches can produce badly biased regression coefficient estimates when the MAR assumption is violated, i.e., when data are not missing at random (NMAR).

Some authors have advocated sensitivity analyses in epidemiologic or clinical settings in order to assess the impact of departures from the MAR assumption [2]. Such techniques generally entail examining changes in estimated parameters of interest over a range of

subjective specifications of the missing data mechanism. Although it is sometimes unavoidable, this inherent subjectivity often prevents a fully satisfying solution in practice.

An alternative to sensitivity analysis and/or to reliance on the MAR assumption when it is questionable is the notion of obtaining direct information about the missing data process based on a supplemental sample. This concept relates to the idea of two-stage sampling, which has been advocated in the context of case-control studies [3-5]. However, our interest lies in a random (rather than targeted) subsample of those whose data were originally missing, perhaps more directly akin to the notion of internal validation sampling in the measurement error literature [6]. Glynn *et al.* [7] were early advocates of this notion in the missing data context, while Lyles and Allen [8] proposed the term “reassessment” in the setting of case-control studies with an NMAR binary exposure variable. In a reassessment sample, one applies a more cost- and/or labor-intensive data collection method to recoup missing information on a subsample of those for whom it was not obtained originally. For example, reassessment might be made by examining medical records, follow-up phone interviews or mailings with incentives, or conducting physical examinations when data cannot be obtained initially by means of a health-related questionnaire [9, 10]. As in the case of internal validation sampling, reassessment will not always be possible. When feasible, however, it shares many of the same important advantages. In particular, it allows identification of potentially complex missing data mechanisms, while offering cost benefits relative to the alternative of applying the intensive data collection procedure to all subjects.

Lyles and Allen [11] provided analytic expressions for the odds ratio and relative risk for cross-sectional studies of association between two binary variables (X and Y), under general models for their joint missing data mechanism. They proposed reassessment sampling and subsequent maximum likelihood (ML) analysis for cases in which the potentially informative missing data mechanism precludes valid estimates of association based only on the data originally observed. In this paper, we generalize these developments to permit valid estimation of more epidemiologically relevant covariate-adjusted odds ratios when both X and Y are subject to potentially informative missingness.

In Section 2, we introduce a logistic regression model relating Y to X while allowing for additional covariates (C), which can be categorical, continuous, or a mixture of both. A secondary logistic model then permits missingness of exposure (X) information to depend on the value of X itself, the disease status (Y), and the covariates. A similarly general model is posed for the missingness of disease information (Y), and missingness of Y is permitted to be conditionally dependent upon (i.e., associated with) that of X . Under such complex but realistic circumstances, the MAR assumption becomes far too simplified and is non-testable based only on the data originally observed. However, information from a reassessment subsample can be incorporated to permit identification of the underlying missing data mechanism. Valid estimation of the adjusted odds ratios of interest is achieved by numerical joint maximization of the full likelihood for the subjects comprising the original and reassessment samples. Section 3 summarizes simulation studies to evaluate the proposed approach versus alternatives that include CC analysis and multiple imputation assuming MAR. In Section 4, we present a real data example in which we examine the covariate-

adjusted association between serum folate levels and body mass index among non-pregnant women participants in the National Health and Nutrition Examination Survey (NHANES).

2. Methods

2.1. Outcome or Exposure Missing with Reassessment

We assume that the following logistic regression model is of primary interest:

$$\text{logit} [Pr(Y=1|x, \mathbf{c})] = \beta_0 + \beta_1 x + \beta_2' \mathbf{c} \quad (1)$$

where Y is a binary outcome, X is a binary exposure variable, and the vector $\mathbf{C}$ represents an arbitrary set of covariates. Initially, we consider the case where X is missing (and potentially NMAR) for some non-negligible proportion of subjects. We assume that a subsample of those with X missing are selected into a reassessment sample, and that the missing data on X is recovered for each subject in the subsample (although the method remains applicable if X is MAR for some subjects in the reassessment sample).

Define M_i and R_i , respectively, as indicators for whether X is missing for subject i and for whether subject i is included in the reassessment sample. Let p_{Ri} be the conditional probability of selection into that sample for those with X missing, the reassessment mechanism is as follows:

$$Pr(R_i=1|y_i, x_i, \mathbf{c}_i, m_i) = \begin{cases} 0 & \text{if } m_i=0, \text{ i.e., the exposure is observed} \\ p_{Ri} & \text{if } m_i=1, \text{ i.e., the exposure is missing} \end{cases} \quad (2)$$

We allow the conditional probability of reassessment to depend on the outcome Y and the covariates $\mathbf{C}$, but given these we assume it does not depend on the true value of X ; that is, we assume $p_{Ri} = Pr(R_i=1|y_i, x_i, \mathbf{c}_i, m_i=1) = Pr(R_i=1|y_i, \mathbf{c}_i, m_i=1)$. We term this “reassessment at random” (RAR). This assumption is more relaxed than the MAR assumption on the first round of sampling and is often quite reasonable [12]. The scenario in which reassessment sampling is completely at random, i.e., $p_{Ri} = Pr(R_i=1)$, is encompassed as a special case.

To construct the joint likelihood function for the main study and reassessment sample data, we consider the 3 possible types of observations. First, a subject with X originally observed contributes the term

$$\begin{aligned} p(y_i, x_i, m_i=0, R_i=0|\mathbf{c}_i) &= Pr(R_i=0|y_i, x_i, \mathbf{c}_i, m_i=0) \times Pr(m_i=0|y_i, x_i, \mathbf{c}_i) \times p(y_i|x_i, \mathbf{c}_i) p(x_i|\mathbf{c}_i) \\ &= 1 \times Pr(m_i=0|y_i, x_i, \mathbf{c}_i) \times p(y_i|x_i, \mathbf{c}_i) p(x_i|\mathbf{c}_i) \end{aligned}$$

A subject with X missing originally but reassessed and found to have $X_i = x_i$ contributes the term

$$\begin{aligned} p(y_i, x_i, m_i=1, R_i=1|\mathbf{c}_i) &= Pr(R_i=1|y_i, x_i, \mathbf{c}_i, m_i=1) \times Pr(m_i=1|y_i, x_i, \mathbf{c}_i) \times p(y_i|x_i, \mathbf{c}_i) p(x_i|\mathbf{c}_i) \\ &= p_{Ri} \times Pr(m_i=1|y_i, x_i, \mathbf{c}_i) \times p(y_i|x_i, \mathbf{c}_i) p(x_i|\mathbf{c}_i) \end{aligned}$$

Finally, a subject with X missing originally but not reassessed contributes the term

$$\begin{aligned} p(y_i, m_i=1, R_i=0|\mathbf{c}_i) &= \Pr(R_i=0|y_i, \mathbf{c}_i, m_i=1) \times \Pr(Y_i=y_i, m_i=1|\mathbf{c}_i) \\ &= (1 - p_{Ri}) \times \sum_{x=0}^1 [\Pr(m_i=1|y_i, X_i=x, \mathbf{c}_i) \times p(y_i|x, \mathbf{c}_i) p(x|\mathbf{c}_i)] \end{aligned}$$

To estimate the conditional probability of missingness, we assume the following logistic model with consideration of up to 3 way interactions:

$$\text{logit} [\Pr(m=1|y, x, \mathbf{c}, \phi)] = \phi_0 + \phi_1 y + \phi_2 x + \phi_3' \mathbf{c} + \phi_{12} yx + \phi_{23}' x\mathbf{c} + \phi_{13}' y\mathbf{c} + \phi_{123}' yx\mathbf{c} \quad (3)$$

Note that the conditional probability terms $\Pr(Y = y | x, \mathbf{c})$ follow from the main model of interest (1). For the conditional probability of exposure (X) given covariates ($\mathbf{C}$), we propose the following logistic sub-model:

$$\text{logit} [\Pr(X=1|\mathbf{C}=\mathbf{c})] = \theta_0 + \theta_1' \mathbf{c} \quad (4)$$

With the above specifications and after ordering subjects with respect to the three observation types, the overall log-likelihood can be written as

$$\begin{aligned} l(\phi, \beta, \theta) &= \sum_{i=1}^n l(\phi, \beta, \theta; y_i, x_i, m_i, R_i | \mathbf{c}_i) \\ &= \sum_{i=1}^{n_{cc}} \log [p(y_i, x_i, m_i=0, R_i=0 | \mathbf{c}_i)] + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \log [p(y_i, x_i, m_i=1, R_i=1 | \mathbf{c}_i)] \\ &\quad + \sum_{i=n_{cc}+n_R+1}^{n_{cc}+n_R+n_{\text{missing}}} \log [p(y_i, m_i=1, R_i=0 | \mathbf{c}_i)] \\ &= \sum_{i=1}^{n_{cc}} \{ \log [\Pr(m_i=0|y_i, x_i, \mathbf{c}_i; \phi)] + \log [p(y_i|x_i, \mathbf{c}_i; \beta)] + \log [p(x_i|\mathbf{c}_i; \theta)] \} \\ &\quad + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \{ \log(p_{Ri}) + \log [\Pr(m_i=1|y_i, x_i, \mathbf{c}_i; \phi)] + \log [p(y_i|x_i, \mathbf{c}_i; \beta)] + \log [p(x_i|\mathbf{c}_i; \theta)] \} \\ &\quad + \sum_{i=n_{cc}+n_R+1}^{n_{cc}+n_R+n_{\text{missing}}} \left\{ \log(1 - p_{Ri}) + \log \left[\sum_{x=0}^1 \Pr(m_i=1|y_i, X_i=x, \mathbf{c}_i; \phi) p(y_i|X_i=x, \mathbf{c}_i; \beta) p(X_i=x|\mathbf{c}_i; \theta) \right] \right\} \end{aligned} \quad (5)$$

where n_{cc} is the number of subjects with exposure observed, n_R is the number of subjects with exposure originally missing but recovered by reassessment, n_{missing} is the number of subjects with exposure originally missing and not selected for reassessment, and $n_{cc} + n_R + n_{\text{missing}} = n$. Note that the terms containing p_{Ri} have factored out in (5) given the RAR assumption. Following an analogous argument to that behind “ignorable” missingness [1], if reassessment is at random and the parameters τ involved in modeling $p_{Ri} = \Pr(R_i=1 | y_i, \mathbf{c}_i, \tau)$

are distinct from the set of parameters (ϕ, β, θ) , then the terms $\sum_{i=n_{cc}+1}^{n_{cc}+n_R} \log(p_{Ri})$ and

$\sum_{i=n_{cc}+n_R+1}^n \log(1 - p_{Ri})$ can be omitted in the maximization of the log-likelihood with regard to (ϕ, β, θ) . Therefore, we can bypass modeling the reassessment mechanism under RAR assumptions. Up to a constant, the log-likelihood simplifies to

$$\begin{aligned}
l(\phi, \beta, \theta) &= \sum_{i=1}^n l(\gamma; y_i, x_i, m_i, R_i | \mathbf{c}_i) \\
&= \sum_{i=1}^{n_{cc}} \{ \log [Pr(m_i=0|y_i, x_i, \mathbf{c}_i; \phi)] + \log [p(y_i|x_i, \mathbf{c}_i; \beta)] + \log [p(x_i|\mathbf{c}_i; \theta)] \} \\
&+ \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \{ \log [Pr(m_i=1|y_i, x_i, \mathbf{c}_i; \phi)] + \log [p(y_i|x_i, \mathbf{c}_i; \beta)] + \log [p(x_i|\mathbf{c}_i; \theta)] \} \\
&+ \sum_{i=n_{cc}+n_R+1}^n \left\{ \log \left[\sum_{x=0}^1 Pr(m_i=1|y_i, X_i=x, \mathbf{c}_i; \phi) p(y_i|X_i=x, \mathbf{c}_i; \beta) p(X_i=x|\mathbf{c}_i; \theta) \right] \right\}
\end{aligned} \tag{6}$$

Now, suppose it is disease status Y rather than exposure X that is subject to missing values originally and recovered by reassessment. The full log-likelihood can be derived analogously and is identical to (6) except that the last term is replaced by

$$\sum_{i=n_{cc}+n_R+1}^n \left\{ \log \left[\sum_{y=0}^1 Pr(m_i=1|Y_i=y, x_i, \mathbf{c}_i; \phi) p(Y_i=y|x_i, \mathbf{c}_i; \beta) p(x_i|\mathbf{c}_i; \theta) \right] \right\},$$

where m_i is the missingness indicator for disease status, R_i is the reassessment indicator, $p_{Ri} = \Pr(R_i=1|x_i, \mathbf{c}_i, \tau)$ is the reassessment rate assumed independent of Y given X and $\mathbf{C}$, and γ is the vector of (ϕ, β, θ) .

Outcome and Exposure Both Missing with Reassessment

Suppose both disease status (Y) and exposure (X) are subject to missing values, and reassessment is conducted on either or both of them. Here we allow the missing data mechanism for each variable to be dependent on Y , X , and the covariates ($\mathbf{C}$); we also allow the missingness of Y and X to be dependent, conditional on these variables. The reassessment subjects are assumed to be randomly chosen within subjects with missing values, independently of the values of Y and X . Let m_{D_i} be the missingness indicator for disease status, and the missingness m_{E_i} the missingness indicator for exposure. Let R_{D_i} be the reassessment indicator for disease status, and R_{E_i} the reassessment indicator for exposure. With reassessment applied at random with respect to both variables (see previous section), the likelihood contribution for subject i can be factored in two ways as follows:

$$\begin{aligned}
p(y_i, x_i, m_{D_i}, m_{E_i}, R_{D_i}, R_{E_i} | \mathbf{c}_i) &= p(R_{D_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}, R_{E_i}) p(R_{E_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}) \\
&\times p(m_{D_i} | y_i, x_i, \mathbf{c}_i, m_{E_i}) p(m_{E_i} | y_i, x_i, \mathbf{c}_i) \\
&\times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i) \\
&= p(R_{D_i} | \mathbf{c}_i, m_{D_i}) p(R_{E_i} | \mathbf{c}_i, m_{E_i}) \\
&\times p(m_{D_i} | y_i, x_i, \mathbf{c}_i, m_{E_i}) p(m_{E_i} | y_i, x_i, \mathbf{c}_i) \\
&\times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i)
\end{aligned} \tag{7}$$

Or equivalently,

$$\begin{aligned}
p(y_i, x_i, m_{D_i}, m_{E_i}, R_{D_i}, R_{E_i} | \mathbf{c}_i) &= p(R_{D_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}, R_{E_i}) p(R_{E_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}) \\
&\times p(m_{E_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}) p(m_{D_i} | y_i, x_i, \mathbf{c}_i) \\
&\times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i) \\
&= p(R_{D_i} | \mathbf{c}_i, m_{D_i}) p(R_{E_i} | \mathbf{c}_i, m_{E_i}) \\
&\times p(m_{E_i} | y_i, x_i, \mathbf{c}_i, m_{D_i}) p(m_{D_i} | y_i, x_i, \mathbf{c}_i) \\
&\times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i)
\end{aligned} \tag{8}$$

The second equality in both (7) and (8) reflects particular reassessment at random assumptions in this setting. To be specific, we assume that the conditional probability of being reassessed for Y or X is not dependent on the underlying value of Y or X given the other covariates $\mathbf{C}$. The two factorizations of the joint density represent different mechanisms for how the missingness of exposure and disease interact. In the case of all categorical covariates ($\mathbf{C}$), the two factorizations could be made equivalent by invoking a saturated model for the joint distribution of the two missingness indicators. When there are continuous covariates, one can choose either factorization and seek to approximate the saturated model via higher order terms and model selection. We illustrate this process for a real data example in Section 4.

Adopting the first factorization of the joint density in (7), we denote the conditional probabilities of missingness as follows:

$$Pm_{m_E}^D = Pr(m_D=1|y, x, \mathbf{c}, m_E=1), \quad Pm_{m_E}^D = Pr(m_D=1|y, x, \mathbf{c}, m_E=0), \quad \text{and} \quad Pm^E = Pr(m_E=1|y, x, \mathbf{c}).$$

The reassessment mechanism is described by the following conditional probabilities:

$$\begin{aligned}
Pr(R_{D_i}=1 | \mathbf{c}_i, m_{D_i}) &= \begin{cases} 0 & \text{if } m_{D_i}=0, \text{ i.e., the disease is observed} \\ p_{Ri}^D & \text{if } m_{D_i}=1, \text{ i.e., the disease is missing} \end{cases} \\
Pr(R_{E_i}=1 | \mathbf{c}_i, m_{E_i}) &= \begin{cases} 0 & \text{if } m_{E_i}=0, \text{ i.e., the exposure is observed} \\ p_{Ri}^E & \text{if } m_{E_i}=1, \text{ i.e., the exposure is missing} \end{cases}
\end{aligned} \tag{9}$$

Based on these general specifications, we can categorize the subjects based on different missingness and reassessment patterns. The likelihood contribution for a subject falling into each of nine possible observation categories is summarized in Table 1; upon enumerating the values of Y and X, these correspond to the 25 observation types described by Lyles and Allen [9] for the no covariate case. For instance, a subject with both Y and X observed contributes the term

$$\begin{aligned}
p(y_i, x_i, m_{D_i}=0, m_{E_i}=0, R_{D_i}=0, R_{E_i}=0 | \mathbf{c}_i) &= Pr(R_{D_i}=0 | \mathbf{c}_i, m_{D_i}=0) Pr(R_{E_i}=0 | \mathbf{c}_i, m_{E_i}=0) \\
&\times Pr(m_{D_i}=0 | y_i, x_i, \mathbf{c}_i, m_{E_i}=0) Pr(m_{E_i}=0 | y_i, x_i, \mathbf{c}_i) \\
&\times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i) \\
&= 1 \times 1 \times \left(1 - Pm_{\frac{D}{m_E}}\right) (1 - Pm^E) \times p(y_i | x_i, \mathbf{c}_i) p(x_i | \mathbf{c}_i)
\end{aligned}$$

The log-likelihood can be readily constructed by looking up the specific likelihood contribution of each subject in Table 1.

The conditional probabilities $\Pr(Y=1 | x, \mathbf{c}, \beta)$ and $\Pr(X=1 | \mathbf{c}, \theta)$ follow directly from models (1) and (4). As in section 2.1, if reassessment is at random and the parameters τ^D and τ^E involved in modeling $p_R^D = \Pr(R_D = 1 | \mathbf{c}, \tau^D)$ and $p_R^E = \Pr(R_E = 1 | \mathbf{c}, \tau^E)$ are distinct from the set of parameters (φ, β, θ) , then the terms containing p_R^D and p_R^E can be omitted in the maximization of the log-likelihood with respect to (φ, β, θ) . Thus, one can easily bypass modeling the reassessment mechanism by invoking “at random” selection into the reassessment sample.

By extension of model (3) to the case of both Y and X subject to missingness, the two conditional probabilities defining the missing data mechanism can be modeled by a pair of logistic regression models:

$$\begin{cases} \text{logit} [Pr(m_D=1 | y, x, \mathbf{c}, m_E)] = \phi_0^D + \phi_1^D y + \phi_2^D x + \phi_3^{D'} \mathbf{c} + \phi_4^D m_E + (\text{additional terms}) \\ \text{logit} [Pr(m_E=1 | y, x, \mathbf{c})] = \phi_0^E + \phi_1^E y + \phi_2^E x + \phi_3^{E'} \mathbf{c} + (\text{additional terms}) \end{cases} \quad (10)$$

Alternatively, the missingness model for the factorization in (10) can be also defined as:

$$\begin{cases} \text{logit} [Pr(m_E=1 | y, x, \mathbf{c}, m_D)] = \phi_0^E + \phi_1^E y + \phi_2^E x + \phi_3^{E'} \mathbf{c} + \phi_4^E m_D + (\text{additional terms}) \\ \text{logit} [Pr(m_D=1 | y, x, \mathbf{c})] = \phi_0^D + \phi_1^D y + \phi_2^D x + \phi_3^{D'} \mathbf{c} + (\text{additional terms}) \end{cases} \quad (11)$$

where the “additional terms” refer to interaction terms involving some or all of the predictors and higher order terms involving continuous covariates $\mathbf{C}$ when needed. Model selection in this context is discussed in Section 2.4.

2.3. Estimation

We have found maximization of the joint log-likelihood functions in Sections 2.1 and 2.2 quite feasible using built-in Quasi-Newton optimization routines available in SAS IML [13]. For standard errors to accompany the MLEs of the regression coefficients of interest, one can also utilize a very close approximation to the Hessian matrix of the maximized log-likelihood based on available numerical routines. Such available computational tools help to enhance the accessibility of the proposed methods for practical use, and readily adaptable programs are available from the first author.

2.4. Model Selection and Testing the MAR Assumption

In practice, models for missingness that contain only main effects might be inadequate to describe a potentially informative missing data mechanism, especially if missingness is dependent on interactions between certain variables. The form of the missingness model can have considerable impact on parameter estimation and inference for the main model. This is highly problematic in the absence of reassessment data, since one must generally rely on untestable assumptions about an informatively missing data mechanism, perhaps with accompanying sensitivity analysis [2], which leaves statistical inference vulnerable to bias. Nevertheless, although formal tests to assess missing completely at random (MCAR) versus MAR assumptions have been proposed [14], in most cases it is not possible to formally address the question of whether a MAR or an NMAR mechanism is at play. As demonstrated here, however, the reassessment approach (when feasible) potentially offers an opportunity to achieve this desirable objective. We can conveniently use likelihood ratio (LR) tests and/or the Akaike information criterion (AIC) to evaluate the goodness-of-fit of different missingness models [e.g., via models (3), (10), or (11)]. Under the null hypothesis that the missing data are MAR, the LR test boils down to assessing whether a subset of the coefficients in the missingness model are equal to 0. The approximate reference distribution will thus be chi-square, with degrees of freedom equal to the number of coefficients in the subset. We illustrate this process by means of simulations (Section 3) and an example (Section 4).

3. Simulations

Simulation studies were conducted to assess the performance of the proposed reassessment-based methods. A full dataset was first generated according to model (1), in which all variables were observed for all subjects. The analysis result based on this dataset set a benchmark for the comparisons of the different missing data methods. Missing values were then induced by generating missingness indicator variable(s) by model (3), when only X is subject to missingness, or by model (10) and (11), when both Y and X are subject to missingness, to form the obtained dataset. The subset of it, which includes only the subjects with all variables observed, forms the complete case dataset, on which the CC analysis is performed. The reassessment dataset was formed by restoring the missing values by the reassessment process described in Section 2. This dataset is built of the subjects for whom the variable values were either not missing in the obtained dataset, or restored via reassessment. The multiple imputation (MI) method was selected as a presentation of the traditional method of handling missing data under the MAR assumption. The MI method could have been performed on the obtained dataset as usually. However, the MI method in this exercise was performed on the reassessment dataset, to offset the advantage of the additional information utilized by the proposed method via reassessment. This step allows for a more fair comparison of MI with the proposed approach in terms of statistical efficiency. The proposed and the traditional method were compared by the estimates of regression coefficients and corresponding 95% confidence interval coverage rates, as relative to the benchmark. Performance of the LR test for evaluating the MAR hypothesis was also assessed. Specifically, data were generated under MAR to examine the type I error of the proposed LR test for H_0 : MAR against H_a : NMAR (see Section 3.1) and then data

were generated under NMAR to examine the power of the LR test (Section 3.2). A set of simulations was also conducted to verify the consistency of the two factorizations (10) and (11) of the missingness model when both disease status and exposure are subject to missing values (Section 3.3).

3.1. Comparison of Methods with MAR X

With only the binary exposure (X) subject to missing values, simulation studies were conducted to assess point estimates and associated standard errors. A random covariate C was first generated as $N(0,1)$. Then X was generated from a logistic model as in (4). Disease status (Y) was then generated according to model (1). Missing values were produced by a missingness indicator following model (3) with $\phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$ as under the MAR setting. The reassessment indicators were generated by a logistic model with Y and C as predictors, to produce a RAR (but not completely random) reassessment strategy.

Table 2 summarizes results from this simulation study, and provides the true model (1) regression coefficients; the table footnotes provide the true coefficients corresponding to models (3) and (4) along with the reassessment model. A total of 1000 simulations were performed, each with sample size 1000. The parameter values in the simulation were chosen to result in an overall missingness rate of 20.7% under these MAR conditions. Within the subset of subjects with missing X, approximately 36.7% were reassessed. As the data were generated under MAR, CC analysis and the other MAR-based methods all produced minimal bias in the coefficient estimate for X while CC analysis produced a biased estimate for the coefficient of C. To explain the latter result, note that when X is MAR the missingness indicator can depend on C. However, CC analysis discards entire records for which X is missing, despite the availability of C. As a result, the induced missingness of C can in fact be effectively NMAR because X was originally missing in a C-dependent way. Note from Table 2 that the proposed joint modeling method assuming MAR managed to gain efficiency compared to CC analysis due to the additional information from reassessment, and performs comparably to MI. The joint modeling method generalized to accommodate potentially NMAR conditions lost a small amount of efficiency, given that the true missingness mechanism was MAR.

The difference between the likelihood functions at the maximum were used to construct the LR test of the MAR hypothesis for each simulated dataset, i.e., $H_0 : \phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$. For the aforementioned simulation scenario, the rejection rate was 5.0%, i.e., under the MAR conditions the type I error achieved the desired nominal level.

3.2. Comparison of Methods with NMAR X

Missing values were produced by a missingness indicator following model (3), where up to three-way interactions involving X were considered,

$$\text{logit} [Pr(m=1|y, x, c, \phi)] = \phi_0 + \phi_1 y + \phi_2 x + \phi_3 c + \phi_{12} yx + \phi_{23} xc + \phi_{13} yc + \phi_{123} yxc.$$

where ϕ_2 , ϕ_{12} , ϕ_{23} , and ϕ_{123} were not all equal to 0. The interaction terms involving both Y and X were to induce bias into each of the regression coefficient estimates with CC analysis.

The reassessment indicators were generated by a logistic model with Y and C as predictors, to produce a RAR (but not completely random) reassessment strategy.

Table 3 summarizes results from this simulation study. The parameter values in the simulation were chosen so that the overall missingness rate for X was 23.4%, and 32.5% of subjects with missing X were reassessed. The CC analysis produced biased estimates for the regression coefficients related to both X and C due to the NMAR missingness mechanism operating in model (3). It is interesting to notice that with the interaction terms in the missingness model, bias was introduced to the coefficient not only for X, but also for C as well. A logistic regression imputation model as formulated in model (3), but eliminating the terms containing X, i.e., setting $\varphi_2 = \varphi_{12} = \varphi_{23} = \varphi_{123} = 0$, as under the MAR assumption, was implemented for MI by invoking the Logistic Regression Imputation [15, 16] using the MONOTONE statement in the standard MI procedure in SAS/STAT [17]. This reduced the bias and improved the efficiency compared to CC, although the bias in the point estimates is still noticeable due to the false assumption of MAR. When the proposed method was applied under the MAR assumption by forcing the coefficients in (3) related to X to be zero, i.e. $\varphi_2 = \varphi_{12} = \varphi_{23} = \varphi_{123} = 0$, the results were very close to those from MI under the MAR assumption. This was expected, given that in this case the proposed method effectively reduced to the common likelihood method under MAR.

Finally, with a general model for the missing data mechanism as specified above which allows three-way interactions involving X, we found that the bias of the point estimates diminished to minimal, mean standard errors approached the empirical standard deviation of the parameter estimates, and CI coverage rates approached the designated 95% value. Although not presented in Table 3 to avoid redundancy, results based on adding a reassessment model in the likelihood were identical to those without the reassessment model. This confirmed via the simulation that under the “reassessment at random” assumption, we can omit modeling the reassessment model as in (6).

The difference between the likelihood functions at the maximum under the MAR and NMAR assumptions were used to construct and evaluate the likelihood ratio test at level 0.05 for hypothesis testing on the MAR assumption, i.e., $H_0 : \varphi_2 = \varphi_{12} = \varphi_{23} = \varphi_{123} = 0$. Under the conditions described above, the rejection rate was 26.0%, i.e., given the magnitude of NMAR operating in this simulation, the power is around 26% to reject H_0 : MAR, at significance level 0.05. Examination of the mean standard errors and p-values associated with Wald tests for the terms in model (3) agreed qualitatively with the low power (results not shown); i.e., the initial simulation conditions represent an NMAR missingness mechanism that deviates only mildly from MAR. This likely explains the fact that the parameter estimates in Table 3 did not deviate markedly from the truth for models incorrectly invoking the MAR assumption, with the exception of the CC analysis.

A second simulation was performed in which the values of the coefficients in the missingness model (3) were chosen to produce a greater magnitude of deviation from MAR. As expected, the CC analysis and other methods assuming MAR produce more bias (see Table 4 as compared to Table 3), while the proposed reassessment-based ML method yielded virtually unbiased coefficient estimates in model (1). The differences between the

likelihood functions at the maximum for each simulated dataset were again used to construct and evaluate the LR test for the MAR hypothesis at significance level 0.05. Based on 1000 simulations under these greater NMAR conditions, the rejection rate was around 67%.

3.3. Both Outcome and Exposure NMAR with Reassessment

We conducted additional simulation studies to examine the performance of the proposed methods when both outcome and exposure are subject to missing values and reassessment as designed. In this case, it is interesting to examine whether the two factorizations of the joint missingness model [eqns. (10) and (11)] yield equivalent results.

Table 5 summarizes 1000 simulation runs, with a total sample size of 1000 for each generated dataset. The missingness indicator was generated by model (10), and produced an overall missingness rate of 21.6% for Y and 17.7% for X . The reassessment rate was 11.6% for Y and 7.9% for X . The proposed method based on both missing data mechanisms (10) and (11) were applied.

In Table 5, we see that when the underlying missing data mechanism is (10), the joint modeling via (10) and (11) yield very similar results. This supports the notion that the investigator can choose either parameterization of the missingness model to construct the likelihood. Note that validity of the estimates for the coefficients corresponding to both X and C is restored by means of the reassessment-based analysis. In contrast the CC displays bias, particularly in the estimate of the X coefficient.

4. Example

Higher pre-pregnancy body mass index (BMI) is associated with increased risk of neural tube defects (NTDs) and possibly other negative birth outcomes in the offspring. The mechanism for this association remains uncertain. Lower maternal folate level has been implicated in the etiology of NTDs in general. Better understanding of the association of the BMI and folate distribution and metabolism has important public health implications [18]. This example examines the association of BMI with folate level in adult women using data from a cross-sectional survey of the U.S. population (National Health and Nutrition Examination Survey (NHANES), 1999–2008), after the 1998 U.S. folate fortification program for cereal products. A specific goal is to address the hypothesis that higher BMI is associated with lower serum levels of folate after controlling for age and race/ethnicity.

Of the 51,623 participants in NHANES 1999–2008, 11,834 were non-pregnant women aged 20 and above, with the serum folate level available. Age and race/ethnicity information were complete. Subjects' BMI was obtained by questionnaire and also by body examination. In this example, we consider the BMI obtained by questionnaire as the first wave of sampling, and a portion of the missing BMI values are then recovered by pulling the record in the body examination as the reassessment measure. For illustration purposes, the measured value was treated as an accurate replacement for the questionnaire value without any measurement error or misclassification. This results in an overall missing rate of 10.4% in the first wave of sampling (1,226 subjects), and among these with missing values, 37.7% were reassessed by body examination (462 subjects).

The association of BMI with serum folate was assessed by logistic regression [model (1)] in which dichotomized serum folate was the dependent variable of interest and BMI was the independent variable of interest. The analysis was controlled for the effect of age and race/ethnicity. BMI was divided into two categories: less than 30 kg/m² and equal to and above 30 kg/m². Serum folate level was dichotomized at the 75th percentile (19.7 ng/mL) of the target population.

The NHANES 1999-2008 used a stratified multistage probability sampling design to survey U.S. household civilian populations. The NHANES requires the use of weights and specific design elements to make the samples representative of the U.S. population and to adjust estimated standard errors. However, given the focus on missingness and reassessment of BMI information and to streamline demonstration of the proposed methods, we omit the sampling weights in this example.

We performed complete case (CC) analysis, multiple imputation (MI) under the MAR assumption, and the proposed ML method based on the reassessment data for BMI. The MI method was implemented by invoking the MCMC statement in the standard MI procedure in SAS/STAT [15-17], and the combined set of BMI data was utilized from the questionnaire and recovered by pulling the body examination records. The proposed approach was implemented as introduced in Section 3.2, with the missing data mechanism modeled in three ways. First the missing data mechanism was defined as MAR by setting the regression coefficients related to X to zero in model (3):

$$\text{logit} [Pr (m=1|y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_3' \mathbf{c} \quad (12)$$

Secondly, the missing data mechanism was allowed to be dependent on BMI (i.e., NMAR), but only the interaction between serum folate and BMI was accounted for:

$$\text{logit} [Pr (m=1|y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \phi_3' \mathbf{c} + \phi_{12} yx \quad (13)$$

Lastly, the missing data mechanism was allowed to be dependent on BMI (i.e., NMAR), and up to three way interactions involving BMI were allowed:

$$\text{logit} [Pr (m=1|y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \phi_3' \mathbf{c} + \phi_{12} yx + \phi_{23}' x\mathbf{c} + \phi_{123}' yx\mathbf{c} \quad (14)$$

The values of the likelihood function at the MLE were recorded for each model, permitting separate LR tests formulated to address the hypothesis that the missing data mechanism is MAR, and to assess the necessary complexity of the NMAR missing data mechanism. The parameter estimates of the main model are summarized in Table 6.

The LR test for model (13) versus (12) yields $\chi^2 = 29.26$ with degrees of freedom equal to 2 ($p < 0.0001$), whilst the likelihood ratio test for (14) versus (12) yields $\chi^2 = 134.29$ with 7 degrees of freedom ($p < 0.0001$). Both LR tests are highly significant, indicating that the missingness of BMI is not at random. The estimates of the regression coefficients in the more general missingness model (14) are summarized in Table 7. We conclude that the

missing data mechanism in this study appears to be dependent on the missing values (NMAR). However, it is interesting to note that the joint modeling does not yield a meaningfully different estimated odds ratio for BMI until the relatively comprehensive missingness model with up to three-way interactions is considered. This point illustrates the importance of adequate reassessment data to allow the analyst to arrive at a sufficiently rich model for the missing data mechanism. In Table 7, further LR testing was used to verify that model (14) also provided significantly better fit than a simpler three-way interaction model that omitted the non-significant higher-order terms ($\chi^2 = 48.48$ vs. 3 degrees of freedom, $p < 0.0001$).

As an exercise, we artificially induced more missing values on the BMI questionnaire results by invoking the model (14). We used the parameter values as summarized in Table 7 but adjusted the value of the intercept, so that the overall missingness rate is only inflated by a constant and the correlation structure with the related effects remains unchanged. Results from this exercise (not shown) demonstrated more deviation of the parameter estimates compared to the results in Table 6 for models relying on MAR (e.g. CC and MI), indicating more significant impact of the underlying NMAR mechanism by the higher missingness rate.

5. Discussion

The impact of missing data on statistical analyses can be particularly extreme when the mechanism is dramatically NMAR. As a case in point, a missingness rate for dichotomized BMI as low as 10.4% in our example produced dramatic deviations in the corresponding estimated coefficient when invoking a missingness model involving interactions indicated by the data.

To make it possible to identify such mechanisms, we have explored the “reassessment” design in the logistic regression setting with NMAR binary data, where a second wave of sampling is implemented to recover a small to moderate portion of the missing data in the original sample. We construct the joint likelihood based on the original model of interest and a model for the missing data mechanism, where the latter allows non-ignorable missingness. The estimation is carried out by numerical maximization of the joint likelihood, and standard errors are estimated via close approximation of the Hessian matrix. We demonstrate that when the reassessment is at random (herein termed “RAR”), the model for the reassessment mechanism can be omitted in the likelihood specification.

Statistical inference is highly dependent on assumptions about the missing data mechanism. Common approaches tend to take the MAR assumption without question, or at best may be supplemented by sensitivity analysis. While the latter approach will sometimes be the only alternative to explore the potential for NMAR mechanisms, it lacks the potential for definitive conclusions or for valid hypothesis testing about the validity of the MAR assumption. A substantial advantage of the reassessment strategy, when feasible, is that it makes it possible to perform formal hypothesis tests about the missing data mechanism. In particular, it is possible to test NMAR against MAR via a likelihood ratio test based on the data. Simulation studies in Sections 3.1 and 3.2 (and others not presented) demonstrate that

this LR test provides valid empirical type I error at the designated level, at least in moderate to large samples. The power of the proposed test is dependent on the magnitude of deviation from MAR, sample size, and the missingness and reassessment rates.

The second crucial advantage of the reassessment design is the fact that it facilitates the consistent estimation of log odds ratios in model (1), despite a potentially complex NMAR mechanism. The proposed joint likelihood function is formulated based on a “selection model” [1]; therefore, the careful choice of the missingness model is important. As demonstrated in Sections 3 and 4, the analyst should employ careful model fitting for missingness [model (3)] in order to reap the full benefits of the reassessment approach. In addition to validating the LR test for the MAR assumption, our simulation studies demonstrate the restoration of valid point estimation using the reassessment-based joint likelihood function. In contrast, alternative approaches (e.g., CC analysis or MI based on the MAR assumption) are shown to produce biased estimates under NMAR conditions.

In this paper, we treated observed data on Y and X at both the initial and reassessment sampling stages as known without error. However, misclassification could be involved in either stage, bringing additional complexity to the problem. In future work, we may seek to extend the proposed methods to incorporate an additional validation component in order to deal with possible measurement error or misclassification in addition to informative missingness. Such an extension could build off prior work along similar lines, in which no additional covariates were assumed present in the missingness models or in the primary model relating Y to X [19].

In the current report, explicit treatment was provided for the setting of a binary outcome and a binary exposure, with either or both missing and subject to reassessment. However, it will be feasible and potentially valuable to extend the proposed method to a general setting with quantitative variables and/or multiple independent variables subject to missing values. From a study design perspective, the model for p_R (or p_R^D and p_R^E in the case of both disease status and exposure missing) could be chosen in a targeted fashion with efficiency of estimation in mind. Similar design considerations have been proposed in the measurement error literature [20, 21] and might be adapted to inform the reassessment sampling approach for dealing with missing data. Future explorations could also place additional focus on a quantitative description of the power of the LR test for the MAR assumption, as well as power/sample size calculations to hone in on the appropriate size and design of the reassessment sample. The allowance for RAR permits selection into the reassessment sample to depend on fully observed variates, and the balance between gains in practical accessibility and efficiency based on tuning the reassessment process can be further explored.

As pointed out previously [11], the type of reassessment discussed here is focused on studies in which missing data occur naturally, in contrast to two-stage designs proposed in previous literature [3-5]. Therefore, it is only the distribution of reassessment given missingness that is under the control of the investigator. We acknowledge that nonresponse at the reassessment stage may be an issue [7], depending on how the second wave of sampling is conducted. If the response to reassessment is MCAR, which may likely be the case if

reassessment is done via access to health or work records, then one can simply ignore that missingness at the reassessment stage. If missingness at the reassessment stage is assumed to be MAR, one could model that missingness without difficulty and without requiring sensitivity analysis or additional reassessment effort. However, if one expects that non-response to the available reassessment strategy would be extensive or highly informative, then the reassessment approach could not be recommended. Nevertheless, large-scale epidemiological studies may benefit from incorporation of a reassessment strategy at the design stage, so that formal assessment and compensation for NMAR missingness can be accounted for at the stage of analysis.

Acknowledgements

This research was supported in part by grants from the National Institute of Nursing Research (1RC4NR012527-01), the National Institute of Environmental Health Sciences (5R01ES012458-07), and the National Center for Advancing Translational Sciences (UL1TR000454). We thank Dr. W. Dana Flanders and Dr. Candice Johnson for access to the motivating data and for helpful discussions.

References

1. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Second Edition.. Wiley; New York: 2002.
2. Vach W, Blettner M. Logistic regression with incompletely observed categorical covariates—investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*. 1995; 14(12):1315–29. [PubMed: 7569490]
3. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika*. 1988; 75(1): 11–20.
4. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*. 1991; 10(5):739–47. [PubMed: 2068427]
5. Zhao LP, Lipsitz SR. Designs and analysis of two-stage studies. *Statistics in Medicine*. 1992; 11(6): 769–82. [PubMed: 1594816]
6. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu C, M. Measurement Error in Nonlinear Models: A Modern Perspective. 2nd Edition. Chapman and Hall CRC Press; Boca Raton , Florida: 2006.
7. Glynn RJ, Laird NM, Rubin DB. Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*. 1993; 88(423):984–993.
8. Lyles RH, Allen AS. Estimating crude or common odds ratios in case-control studies with informatively missing exposure data. *American journal of Epidemiology*. 2002; 155(3):274–81. [PubMed: 11821253]
9. Hansen MH, Hurwitz WN. The problem of non-response in sample surveys. *Journal of the American Statistical Association*. 1946; 41(236):517–529. [PubMed: 20279350]
10. Crawford SL, Johnson WG, Laird NM. Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey. *Case Studies in Bayesian Statistics*. 1993:78–117.
11. Lyles RH, Allen AS. Missing data in the 2×2 table: patterns and likelihood-based analysis for cross-sectional studies with supplemental sampling. *Statistics in Medicine*. 2003; 22(4):517–534. [PubMed: 12590411]
12. Bartholomew DJ. A method of allowing for not-at-home bias in sample surveys. *Applied statistics*. 1961:52–59.
13. SAS Institute, Inc. SAS/IML 9.2 User's Guide. 2008
14. Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. 1988; 83.404:1198–1202.

15. Allison, PD. Missing Data. Sage; Thousand Oaks, CA: 2002.
16. Allison PD. Imputation of categorical variables with PROC MI. SAS Focus Session, SUGI; 2005; 30
17. SAS Institute, Inc. SAS/STAT 9.2 User's Guide. The MI Procedure. 2008:3738–3831. Chapter 54.
18. Mojtabai R. Body mass index and serum folate in childbearing age women. *European Journal of Epidemiology*. 2004; 19(11):1029–36. [PubMed: 15648596]
19. Lyles RH, Allen AS, Flanders WD, Kupper LL, Christensen DL. Inference for case-control studies when exposure status is both informatively missing and misclassified. *Statistics in Medicine*. 2006; 25:4065–4080. [PubMed: 16463349]
20. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics*. 1991:851–869. [PubMed: 1789885]
21. Lyles RH, Williamson JM, Lin HM, Heilig CM. Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics*. 2005; 61(1):287–294. [PubMed: 15737105]

Table 1

Likelihood contribution of a subject in each of the nine categories

Observation Categories	Likelihood Contribution
Y and X both observed	$\left(1 - Pm \frac{D}{m_E}\right) \times (1 - Pm^E) \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c})$
Y observed X missing, reassessed	$p_R^E \times \left(1 - Pm \frac{D}{m_E}\right) \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c})$
Y observed X missing not reassessed	$\left(1 - p_R^E\right) \times \frac{1}{x=0} \left[\left(1 - Pm \frac{D}{m_E}\right) \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c}) \right]$
Y missing, reassessed X observed	$p_R^D \times Pm \frac{D}{m_E} \times (1 - Pm^E) \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c})$
Y missing not reassessed X observed	$\left(1 - p_R^D\right) \times \frac{1}{y=0} \left[Pm \frac{D}{m_E} \times (1 - Pm^E) \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c}) \right]$
Y, X both missing both reassessed	$p_R^D \times p_R^E \times Pm \frac{D}{m_E} \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c})$
Y, X both missing Y reassessed, X not	$p_R^D \times \left(1 - p_R^E\right) \times \frac{1}{x=0} \left[Pm \frac{D}{m_E} \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c}) \right]$
Y, X both missing X reassessed, Y not	$\left(1 - p_R^D\right) \times p_R^E \times \frac{1}{y=0} \left[Pm \frac{D}{m_E} \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c}) \right]$
Y, X both missing both not reassessed	$\left(1 - p_R^D\right) \times \left(1 - p_R^E\right) \times \frac{1}{y=0} \frac{1}{x=0} \left[Pm \frac{D}{m_E} \times Pm^E \times p(y \mid x, \mathbf{c})p(x \mid \mathbf{c}) \right]$

Table 2

Comparison of methods under MAR

	Intercept 0	X 1	C -0.5
Full Data	-0.002 (0.105) [0.102] {95.2% }	1.011 (0.161) [0.161] {95.1% }	-0.505 (0.045) [0.045] {95.1% }
CC Analysis	-0.219 (0.121) [0.117] {53.5% }	1.009 (0.185) [0.183] {94.4% }	-0.580 (0.054) [0.053] {69.7% }
MI	-0.003 (0.110) [0.107] {95.1% }	1.011 (0.171) [0.174] {95.2% }	-0.505 (0.046) [0.046] {95.0% }
Joint Modeling Under MAR	-0.003 (0.110) [0.107] {94.8% }	1.011 (0.170) [0.173] {95.5% }	-0.505 (0.046) [0.046] {94.9% }
Joint Modeling Under NMAR, without modeling R	0.000 (0.115) [0.112] {94.8% }	1.004 (0.183) [0.188] {95.2% }	-0.503 (0.047) [0.047] {94.6% }

Numbers in each cell reflect mean (standard deviation) based on 1000 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

The true values to generate the missingness indicator were taken as $(\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_{12}, \varphi_{23}, \varphi_{13}, \varphi_{123}) = (-2, 1, 0, 0.2, 0, 0, 0.15, 0)$, as in model (3).

The true values to generate X were taken as $(\theta_0, \theta_1) = (0, 0.5)$, as in model (4).

Table 3Comparison of methods with X mildly not missing-at-random (NMAR)

	Intercept 0	X 1	C -0.5
Full Data	-0.002 (0.103) [0.102] {95.0%}	1.010 (0.166) [0.161] {94.6%}	-0.504 (0.046) [0.045] {95.3%}
CC Analysis	-0.213 (0.121) [0.117] {54.6%}	1.204 (0.194) [0.189] {80.8%}	-0.571 (0.054) [0.054] {76.1%}
MI	-0.051 (0.109) [0.107] {91.5%}	1.143 (0.182) [0.178] {87.3%}	-0.520 (0.048) [0.047] {93.0%}
Joint Modeling Under MAR	-0.051 (0.109) [0.106] {91.4%}	1.143 (0.182) [0.177] {86.7%}	-0.520 (0.048) [0.047] {92.8%}
Joint Modeling Under NMAR, without modeling R	-0.003 (0.116) [0.113] {95.4%}	1.013 (0.204) [0.195] {94.1%}	-0.505 (0.048) [0.047] {94.0%}

Numbers in each cell reflect mean (standard deviation) based on 1000 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

The true values to generate the missingness indicator were taken as $(\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_{12}, \varphi_{23}, \varphi_{13}, \varphi_{123}) = (-2, 1, 1, 0.2, -1, -0.1, 0.15, 0.05)$, as in model (3).

The true values to generate X were taken as $(\theta_0, \theta_1) = (0, 0.5)$, as in model (4).

Table 4

Comparison of methods with X markedly not missing-at-random (NMAR)

	Intercept 0	X 1	C -0.5
Full Data	0.001 (0.105) [0.102] {94.3%}	1.003 (0.170) [0.161] {94.1%}	-0.502 (0.046) [0.045] {94.9%}
CC Analysis	-0.083 (0.111) [0.108] {86.7%}	1.354 (0.190) [0.182] {51.6%}	-0.489 (0.050) [0.050] {94.5%}
MI	-0.081 (0.105) [0.103] {86.2%}	1.239 (0.183) [0.175] {71.2%}	-0.528 (0.047) [0.046] {91.3%}
Joint Modeling Under MAR	-0.081 (0.105) [0.103] {86.3%}	1.238 (0.182) [0.174] {71.2%}	-0.528 (0.047) [0.046] {91.4%}
Joint Modeling Under NMAR, without modeling R	0.005 (0.113) [0.109] {91.3%}	0.998 (0.197) [0.187] {91.8%}	-0.500 (0.047) [0.046] {93.7%}

Numbers in each cell reflect mean (standard deviation) based on 1000 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

The true values to generate the missingness indicator were taken as $(\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_{12}, \varphi_{23}, \varphi_{13}, \varphi_{123}) = (-2, -1, 1.5, 0.2, -1.5, -0.1, 0.15, 0.05)$, as in model (3).

The true values to generate X were taken as $(\theta_0, \theta_1) = (0, 0.5)$, as in model (4).

Table 5

Comparison of methods when both outcome and exposure are NMAR and reassessed

	Intercept 0	X 1	C -0.5
Full Data	-0.004 (0.092) [0.094] {95.7%}	1.010 (0.144) [0.144] {94.3%}	-0.503 (0.074) [0.073] {95.3%}
CC Analysis	-0.212 (0.113) [0.113] {54.2%}	1.360 (0.181) [0.178] {47.8%}	-0.512 (0.091) [0.090] {94.5%}
Joint Modeling [model (10)]	-0.015 (0.172) [0.177] {94.7%}	1.037 (0.297) [0.302] {93.2%}	-0.501 (0.088) [0.088] {95.0%}
Joint Modeling [model (11)]	-0.011 (0.174) [0.178] {94.6%}	1.028 (0.295) [0.304] {93.0%}	-0.499 (0.088) [0.088] {95.0%}

Numbers in each cell reflect mean (standard deviation) based on 1000 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

The true values to generate the missingness indicators were taken as $(\phi_0^D, \phi_1^D, \phi_2^D, \phi_3^D, \phi_4^D, \phi_{12}^D) = (-2, 0.8, 0.5, 0.2, 1.5, -1)$ and $(\phi_0^E, \phi_1^E, \phi_2^E, \phi_3^E, \phi_{12}^E) = (-2, 0.5, 0.8, 0.2, -1)$, as in (10), with the interactions between Y and X as the additional term.

The true values to generate X were taken as $(\theta_0, \theta_1) = (0, 0.5)$, as in model (4).

Table 6

Parameter Estimates by Different Methods for Serum Folate/BMI Example

	BMI	Age	Race
CC Analysis	−0.404 (0.052) [0.67]	0.033 (0.001) [1.03]	0.701 (0.050) [2.02]
MI	−0.392 (0.050) [0.68]	0.026 (0.001) [1.03]	0.643 (0.047) [1.90]
Joint Modeling MAR [model (12)]	−0.394 (0.079) [0.67]	0.025 (0.001) [1.03]	0.652 (0.047) [1.92]
Joint Modeling NMAR [model (13)]	−0.395 (0.052) [0.67]	0.025 (0.001) [1.03]	0.635 (0.047) [1.89]
Joint Modeling NMAR [model (14)]	−0.549 (0.053) [0.577]	0.026 (0.001) [1.03]	0.636 (0.047) [1.89]

Dichotomized serum folate level was considered as the dependent variable in a logistic regression, with BMI, age and race/ethnicity as the independent variables. The overall missing rate is 10.4%, and reassessment rate 37.7%. The values in each cell represent the estimated logarithm odds ratio, standard error in (), and odds ratio in [].

Table 7

Parameter Estimates for the Missingness Model [Model (14)] in Serum Folate/BMI Example

Effect	Estimate	Std. Error	Chi-Square	P-value
Intercept	−3.953	0.207	365.06	<0.0001
Serum Folate (SF)	−0.860	0.173	24.70	<0.0001
BMI	−0.850	0.388	4.785	0.0287
Age	0.035	0.004	82.16	<0.0001
Race	−1.000	0.143	48.81	<0.0001
BMI*SF	1.105	0.853	1.68	0.195
BMI*Age	0.017	0.007	5.90	0.0152
BMI*Race	−0.920	0.504	3.333	0.0679
BMI*SF*Age	−0.008	0.013	0.357	0.550
BMI*SF*Race	−1.624	0.333	23.80	<0.0001
BMI*Age*Race	0.035	0.007	23.95	<0.0001