Published in final edited form as:

Stat Med. 2010 March 15; 29(6): 671–679. doi:10.1002/sim.3821.

Correction of Bias from Non-random Missing Longitudinal Data Using Auxiliary Information

Cuiling Wang^{1,*†} and Charles B. Hall¹

¹Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461

SUMMARY

Missing data are common in longitudinal studies due to drop-out, loss to follow-up, and death. Likelihood-based mixed effects models for longitudinal data give valid estimates when the data are ignorably missing; that is, the parameters for the missing data process are distinct from those of the main model for the outcome, and the data are missing at random (MAR). These assumptions, however, are not testable without further information. In some studies, there is additional information available in the form of an auxiliary variable known to be correlated with the missing outcome of interest. Availability of such auxiliary information provides us with an opportunity to test the MAR assumption. If the MAR assumption is violated, such information can be utilized to reduce or eliminate bias when the missing data process depends on the unobserved outcome through the auxiliary information. We compare two methods of utilizing the auxiliary information: joint modeling of the outcome of interest and the auxiliary variable, and multiple imputation (MI). Simulation studies are performed to examine the two methods. The likelihood-based joint modeling approach is consistent and most efficient when correctly specified. However, mis-specification of the joint distribution can lead to biased results. MI is slightly less efficient than a correct joint modeling approach but more robust to model mis-specification when all the variables affecting the missing data mechanism and the missing outcome are included in the imputation model. An example is presented from a dementia screening study.

Keywords

auxiliary variable MAR (A-MAR); joint modeling; linear mixed effects model; missing data; MNAR; multiple imputation (MI)

1. INTRODUCTION

Longitudinal studies are widely used in epidemiological research to study the pattern of change of certain outcomes. Missing data, often due to drop-out, is a common challenge in analyzing longitudinal data. If the missing data process is unrelated to the value of the unobserved outcome conditioning on the observed outcome and covariates (i.e. the data are missing at random (MAR) [1]), and the parameters governing the missing data process and the model for the outcome are disjoint, then the missing data are ignorable [1–2] and

*Correspondence to: Cuiling Wang, 1300 Morris Park Ave, Bronx, NY 10463, U.S.A.

†cuwang@aecom.yu.edu

inference on the model parameters can be based on the likelihood of the observed data. The MAR missing data mechanism is the underlying assumption for many statistical methods used in the analysis of longitudinal data, such as the likelihood-based mixed effects model. However, when the missing data mechanism is missing not at random (MNAR), i.e., missing data depends on the unobserved outcome, the missing data is non-ignorable. In many studies this situation cannot be ruled out and serious biases in the parameter estimates may result if the missing data process is ignored.

Without additional information, the MAR assumption is unverifiable. If it is possible that the missing data mechanism is MNAR, the missing data process has to be modeled. It is usually difficult to justify a specific choice for the missingness model and the parameters may not even be identifiable. Hence sensitivity analysis is often performed under MNAR, in which several plausible missingness models are assumed or the parameters governing the dependence of the missing data on the unobserved outcome are varied over a plausible range of values. This allows one to examine the degree to which inferences for the main model are influenced across the different scenarios for missing data.

Most studies collect information beyond the specific outcome and covariates used in a model. If there are auxiliary variables associated with the outcome which are available when the outcome is missing, we can test the MAR assumption using the auxiliary variable. If there is evidence that the missing data mechanism is not MAR, it is possible to utilize the auxiliary variable to make inferences without fitting a non-ignorable missingness model. For example, in a dementia screening study in Bronx, NY [3], one of the research aims is to study the decline in memory measured by Free and Cued Selective Reminding Test (FCSRT) [4] at one year of follow-up compared to baseline. Fifty-nine (25%) of the 238 subjects studied did not have a follow-up measure of FCSRT. However, the primary care physicians' assessment of memory in the clinical dementia rating system (CDR) [5] was available for all the subjects at both baseline and follow-up. Because both measure memory to some degree, FCSRT and CDR are correlated. We can thus check the MAR assumption by examining the dependence of the missing data process on CDR, conditional on the observed outcome and the covariate. If there is evidence the missing data mechanism is MNAR, the CDR information can be utilized to eliminate or reduce bias.

In terms of using auxiliary data to eliminate or reduce bias due to non-ignorable missing data process, Ibrahim et al [6] and Daniels and Hogan [7] proposed a joint modeling approach for cross-sectional and longitudinal studies, respectively. Alternatively, adding auxiliary variables to a multiple imputation procedure to correct bias under non-ignorable missingness is a natural extension of the MI approach under MAR and has been proposed in the literature [8–9], and most recently by Collins et al [10].

To our knowledge, however, neither the relative performance nor the sensitivity to model mis-specification of the joint modeling and MI approaches has been previously examined. In this paper, we compare the two approaches and examine their sensitivity to model misspecifications through simulation studies, and apply the methods to the dementia screening study. Section 2 provides additional background to the problem, and describes how the auxiliary variable can be used to test the MAR assumption. In Section 3 the joint

modeling and MI approaches for utilizing the auxiliary variable are described in detail. Simulation studies to examine the performance of the two methods under different missing data mechanisms, and the effect of model mis-specification on the parameter estimation, are presented in Section 4. In Section 5 the methods are applied to data from a dementia screening study. We conclude the paper with a discussion in Section 6.

2. Using auxiliary information to test the MAR assumption

2.1 Background and notation

Suppose there are m independent subjects with n planned visits in a longitudinal study. Due to attrition, only $n_i, 1 \leq n_i \leq n$, measurements are observed for subject i . Denote Y_{ij} as the j^{th} outcome measurement for subject i , and $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$. For a continuous outcome, the following linear mixed effects model [11] is often used:

$$Y_i = W_i \beta + D_i b_i + \varepsilon_i, \quad (1)$$

where W_i is the design matrix for the fixed effects which consists of the vector of covariates, X_i functions of time, and/or their interactions, b_i is a vector of random effects, D_i is the design matrix for the random effects (usually a subset of W_i) and ε_i is the random error. The parameter vector β for the fixed effects is of primary interest here. For a categorical outcome, the generalized linear mixed effects model [12], in which the expectation of the outcome is proportional to the linear combination of the fixed and random effects, modelled through a link function, can be used instead.

We decompose the outcome into two parts according to observation status, $Y = (Y^o, Y^m)$, with Y^o and Y^m denoting the observed and missing part, respectively. Let R denote the vector of the observation indicator of Y . According to Rubin [1], and Little and Rubin [2], three missing data mechanisms can be defined :

- Missing completely at random (MCAR) if $R \perp (Y^o, Y^m) \mid X$
- Missing at random (MAR) if $R \perp Y^m \mid Y^o, X$
- Missing not at random (MNAR) if R depends on Y_m conditional on Y_o and X .

The likelihood for the observed data (Y^o, R) is

$$f(y^o, r | x) = \int f(y^o, y^m | x) f(r | y^o, y^m, x) dy^m.$$

When the missing mechanism is MAR, that is,

$$P(R=1 | Y^o, Y^m, X) = P(R=1 | Y^o, X), \quad (2)$$

the part for R can be factored out of the integral so that $f(y^o, r | x) = f(y^o | x) f(r | y^o, x)$. Furthermore, if the parameters governing the missing data process, i.e., the distribution of R , and the parameters for the outcome Y are disjoint, the portion of the likelihood corresponding to the missing data process $f(r | y^o, x)$ can be ignored. Thus, inference for the parameters in the model for Y can be based only on $f(y^o | x)$ and the missing data is deemed

to be ignorable; otherwise the missing data process is non-ignorable or informative. To avoid further complication, we assume that the condition of disjoint parameters holds throughout this paper.

Denote the auxiliary information as Z , where Z and Y are correlated. For simplicity, we assume Z is fully observed. Suppose that the missing data mechanism is MNAR, i.e., $P(R=1|Y^o, Y^m, X)$ depends on Y^m . Assume further that conditional on Y^o and X , R depends on Y^m only through Z . Then

$$P(R=1|Y^o, Y^m, X, Z) = P(R=1|Y^o, X, Z). \quad (3)$$

The missing data assumption (3) is called auxiliary variable MAR (A-MAR) by Daniels and Hogan [7].

2.2 Testing the MNAR assumption using an auxiliary variable

If the A-MAR assumption holds, then it can be shown that R depends on the unobserved outcome Y^m , conditional on Y^o and X , if and only if it depends on the auxiliary variable Z . Thus, the auxiliary variable Z can be used to test the assumption of MAR. For example, suppose $R_{i1}=1$ and the missing data pattern is monotone, i.e., $R_{i1} \geq R_{i2} \geq \dots \geq R_{in}$. Denote $H_{it} = (X_i, Y_{i1}, Y_{i2}, \dots, Y_{it}, Z_{i1}, Z_{i2}, \dots, Z_{it})$ as the history for subject i up to t , $1 \leq t \leq n$. Assume that the probability of observing a particular Y_{ij} may depend on the past and the current values of the outcome and auxiliary variable, but not the future. Then (3) implies that for $1 < t \leq n$, $\lambda_{it} \triangleq P(R_{it}=1|R_{it-1}=1, X_i, Y_i, Z_i) = P(R_{it}=1|R_{it-1}=1, X_i, H_{it-1}, Z_{it})$.

The unconditional observation probability $\pi_{it} = P(R_{it}=1|X_i, H_{it-1}, Z_{it})$ can then be expressed as $\prod_{j=1}^t \lambda_{ij}$. Assume a parametric model, $\lambda_{it}(\alpha)$, for λ_{it} (e.g., a logistic model), where $\lambda_{it}(\alpha)$ is a known function indexed by unknown parameter vector α . The estimator for α , $\hat{\alpha}$, can be obtained by maximizing the partial likelihood

$$L(\alpha) = \prod_i L_i(\alpha) = \prod_i \prod_t \left[\lambda_{it}(\alpha)^{R_{it}} \{1 - \lambda_{it}(\alpha)\}^{1-R_{it}} \right]^{R_{it}-1}.$$

Equivalently, $\hat{\alpha}$ is the solution to the following estimating equation:

$$U(\alpha) = \sum_i U_i(\alpha) = \sum_i \frac{\partial L_i(\alpha)}{\partial \alpha} = 0.$$

The estimator $\hat{\alpha}$ is consistent and asymptotically normally distributed with variance $E\{U(\alpha)U(\alpha)^T\}^{-1}$. The MAR missing data mechanism is equivalent to a situation where all elements of the parameter vector α corresponding to the Z_{ij} are equal to zero, $1 \leq j \leq t$, $1 \leq t \leq n$. Hence, testing the null hypothesis that the missing data mechanism is MAR versus the alternative hypothesis that it is MNAR is equivalent to testing the null hypothesis that all the coefficients for any measurements of the auxiliary variable are zero under the A-MAR assumption (3).

For example, in the case of two planned visits in a longitudinal study, only one model needs to be fit for $\lambda_{i2}(\alpha) = \pi_{i2}$, the probability of observing the outcome Y_{i2} at the second visit, such as the following logistic model:

$$\text{logit} \{P(R_{i2}=1|Y_{i1}, Z_{i1}, Z_{i2})\} = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \alpha_3 Y_{i1} = B^T \alpha,$$

where $B_i = (1, Z_{i1}, Z_{i2}, Y_{i1})^T$ is the design matrix in the model for R_{i2} , $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$. Here $\lambda_{i2}(\alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \alpha_3 Y_{i1})$, where $\text{logit}^{-1}(x) = \exp(x) / \{1 + \exp(x)\}$. The parameter vector α is the solution of the estimating equation

$$\sum_{i=1}^n B_i \{R_{i2} - \lambda_{i2}(\alpha)\} = 0.$$

Testing the MAR assumption is equivalent to testing the null hypothesis that $\alpha_1 = \alpha_2 = 0$. with appropriate modeling to utilize the auxiliary variable, we can eliminate the bias if the data is missing A-MAR, and reduce the bias if it is not. In the next section we show methods for utilizing an auxiliary variable to eliminate or reduce bias under a non-ignorably missing data mechanism.

3. Methods utilizing auxiliary information to reduce bias under non-ignorably missing data mechanism

3.1 Joint modeling of the outcome of interest and the auxiliary variable

Denote $Y^* = (Y, Z)$, and the observed and missing part of Y^* as $Y^{*o} = (Y^o, Z)$ and $Y^{*m} = Y^m$, respectively. Then the A-MAR assumption (3) is equivalent to

$$P(R=1|Y^{*o}, Y^{*m}, X) = P(R=1|Y^{*o}, X),$$

which is MAR if Y^* , the combination of the outcome of interest and the auxiliary variable, is now the outcome. This implies that if we jointly model Y and Z as a multivariate longitudinal outcome, we can obtain valid results under A-MAR using conventional statistical software. Ibrahim et al [6] proposed this method to reduce the bias under non-ignorable missingness for a cross-sectional study. Daniels and Hogan [7] also proposed this method for longitudinal studies.

Several approaches to jointly modeling multivariate longitudinal data are available. A recent review can be found in Fitzmaurice et al [13]. We describe three approaches that are suitable for our purposes: multivariate marginal models, shared random effects models and random effects models. These methods can all be easily applied using standard statistical software, e.g., sample programs are provided in the website <http://www.biostat.harvard.edu/~fitzmaur/lda/> for the book of Fitzmaurice et al [13].

3.1.1 Multivariate marginal models—Multivariate marginal models directly specify the joint density of Y and Z conditional on the covariate X , $f(y, z|x)$. This means that not only the

marginal means of Y and Z , but also the marginal association among the repeated measurements within each variable, as well as the assumptions on the association between elements of Y and Z , are specified as part of the model. This method is straightforward to apply when Y and Z can be assumed to be multivariate normal. For example, in SAS 9.1, the MIXED procedure can fit this model assuming a Kronecker product covariance structure. If the inter-variable correlation matrix is Σ and the intra-variable variance-covariance matrix at the same time point is V , then the variance-covariance of Y and Z is $V \otimes \Sigma$, where $\otimes$ denotes the Kronecker product.

For example, in a longitudinal study with two visits, the outcome of interest, $Y = (Y_1, Y_2)$, and the auxiliary variable, $Z = (Z_1, Z_2)$ are both assumed to be continuous and $(Y, Z) = (Y_1, Y_2, Z_1, Z_2)$ is multivariate normal, with $\text{Var}(Y_j) = \sigma_Y^2$, $\text{Var}(Z_j) = \sigma_Z^2$, $\rho = \text{corr}(Z_1, Z_2) = \text{corr}(Y_1, Y_2)$, $r = \text{corr}(Y_j, Z_j)$, $j=1, 2$. That is, the variance-covariance matrix of (Y, Z) is

$$V \otimes \Sigma = \begin{pmatrix} \sigma_Y^2 & r\sigma_Y\sigma_Z \\ r\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} \sigma_Y^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & r\sigma_Y\sigma_Z \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ r\sigma_Y\sigma_Z \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & \sigma_Z^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \end{pmatrix}.$$

This method, however, is difficult to apply when Y and Z are of different types (e.g., continuous and discrete); in this case, one of the approaches below may be considered.

3.1.2 Shared random effects models—The shared random effects models assume that Y and Z are associated through common random effects. Specifically, denote b as a vector of random effects, and assume Y and Z are independent from each other conditional on b . The joint distribution of Y and Z can then be expressed as

$$f(y, z|x) = \int f(y, z|b, x) f(b|x) db = \int f(y|b, x) f(z|b, x) f(b|x) db$$

where $f(b|x)$ denotes the density function of the random effects. The random effect b serves as the common underlying characteristic that governs both the outcome Y and Z .

A major advantage of this approach is that Y and Z need not to be of the same type. For example, if Y is continuous and Z is binary, a linear mixed effects model for Y and a logistic random effects model for Z can be fit, with both models sharing the same random effects as specified below:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_i + \varepsilon_{ij}, \\ \text{logit} \{P(Z_{ij}=1)\} &= \tau_0 + \tau_1 t_{ij} + \lambda b_i, \end{aligned} \quad (4)$$

where b_i is the random effect distributed as $N(0, \sigma_b^2)$, and ε_{ij} is a normally distributed random error with variance σ_ε^2 and is independent from b_i . The SAS procedure NLMIXED can be used to fit this model. A disadvantage of this approach is the strong assumption regarding the association between Y and Z that is implied by the common random effect governing the two processes.

3.1.3 Random effects models—The random effects models assume that Y and Z are associated through separate but correlated random effects.

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + u_{1i} + \varepsilon_{ij}, \\ \text{logit} \{P(Z_{ij}=1)\} &= \tau_0 + \tau_1 t_{ij} + u_{2i}, \end{aligned} \quad (5)$$

where u_{1i} and u_{2i} are the random effects with variance σ_1^2 and σ_2^2 , respectively, and correlation coefficient $\rho(u_1, u_2)$.

The random-effects model is an extension of the shared random effects model. The latter is a special case of the former when $\rho(u_1, u_2)=1$. The random-effects model has all of the advantages of the shared random effects model but the assumption about the association between Y and Z is less restrictive. It can be fit using the SAS procedure NLMIXED. However, the computational burden is also increased with the inclusion of more random effects.

All of the above approaches require the introduction of assumptions about the joint distribution of Y and Z . If this joint distribution is correctly specified, the joint modeling approach should yield not only consistent estimates under A-MAR, but also efficient estimates because it is likelihood based. However, if the joint distribution is mis-specified, significant bias could result. We will examine the extent of such potential biases in Section 4.

3.2 Multiple imputation

The A-MAR assumption in (3) implies that the distribution of the missing outcome is independent of the missing data pattern conditional on the observed outcome, the auxiliary variable and the covariate, i.e., $f(y^m | y^o, x, z, r) = f(y^m | y^o, x, z)$; hence it can be consistently estimated, or imputed, from the non-missing data. For a monotone missing data pattern, the missing outcome can be sequentially imputed such as that proposed by Paik [14] and described by Kenward and Carpenter in chapter 21 of [13]. Specifically, denote the complete data for subject i as $y_i^c = (y_{i1}^c, \dots, y_{in}^c)^T$. Since only n^i measurements are observed, the j th element of y_i^c is missing for $j > n_i$. We sequentially impute the missing data as follows: for all subjects missing the second measurement, the missing data is imputed by random draw from the conditional distribution $y_{i2}^c | y_{i1}^c, z, x$ which can be estimated from the regression of y_{i2}^c on y_{i1}^c, z and x among those with y_{i2}^c observed (linear regression for continuous outcome and logistic regression for binary or ordinal outcome). The imputed observations, $\tilde{y}_{i2}^*$, are filled in for the missing values so a complete second measurement of the outcome is y_{i2}^{c*} which equals y_{i2} if $r_{i2} = 1$, and equals $\tilde{y}_{i2}^*$ if $r_{i2} = 0$. This is used as if it is the observed measurement to impute the third measurements that are missing using the imputation model for the distribution $y_{i3}^c | y_{i1}^c, y_{i2}^{c*}, z, x$. This process is repeated up to the last measurement which is imputed from the distribution $y_{in}^c | y_{i1}^c, y_{i2}^{c*}, \dots, y_{in-1}^{c*}, z, x$. Thus a complete set of imputations is generated.

The above process is repeated M times to obtain M imputed data sets. For each imputed data set, standard methods such as the linear mixed effects model (1) are used to model the

outcome of interest y . Denote the parameter estimate as $\hat{\beta}_j$, with variance V_j , $j = 1, \dots, M$. The results from the M analyses are then combined to yield the MI estimator $\bar{\beta} = \sum_{j=1}^M \hat{\beta}_j$, with variance $\bar{V} + B_m$, where $\bar{V} = \sum_{j=1}^M V_j$ and $B_m = \sum_{j=1}^M \left(\hat{\beta}_j - \bar{\beta} \right)^T \left(\hat{\beta}_j - \bar{\beta} \right) / (M - 1)$, the within- and between-imputation variability. Other approaches, e.g., predicted mean matching, propensity score, can also be used to impute the data.

The idea of adding auxiliary variables in multiple imputation to correct bias, even though the auxiliary variables are not included in the main model of interest, has been proposed by others [8–9] and most recently by Collins et al [10]. As emphasized in the literature [15–16], the imputation model should include variables known to be predictive of the missingness and related to the missing variables, and should contain all structure, e.g., interaction and power of a variable, that are included in the main model [13, 17]. If the missing data mechanism is MAR, then the imputation model does not need to include the auxiliary variable. But under A-MAR, the auxiliary variable, in addition to the observed outcome and the covariates, must be included in the imputation model. Assuming that all the required variables and their structures are included in the imputation model, it has been suggested that the resulting inferences are fairly robust to the specific choice of the imputation distribution [e.g. 16, 18].

3.3 Other methods of utilizing auxiliary information

Our focus thus far has been on a linear mixed effects model for the main outcome. Another popular method for longitudinal data analysis is the generalized estimating equations (GEE) approach [19]. This approach is valid when the data are missing completely at random (MCAR). Robins et al [20] showed that the inverse probability weighting (IPW) method will give consistent estimates when data are missing at random (MAR). If the auxiliary variables are added to the model for the probability of observation in the IPW approach, then the assumption on missing data is reduced to A-MAR.

3.4 Extension to more complicated missing data case

We have assumed that the auxiliary variable Z is fully observed and the missing data pattern for the outcome Y is monotone. These assumptions can be relaxed to allow more complicated cases such as when the auxiliary variable is not fully observed and/or the missing data pattern is not monotone. Denote Z^o as the observed part of Z , according to Daniels and Hogan [7], the missing data mechanism is A-MAR if $P(R = 1 | Y^o, Y^m, X, Z^o) = P(R = 1 | Y^o, X, Z^o)$. The joint modeling method can easily handle missing auxiliary variable and/or non-monotone missing pattern. For the multiple imputation approach, the Markov Chain Monte Carlo (MCMC) [16] method and the univariate conditional method [21] are available for general missing data patterns, though the extensions to unbalanced longitudinal data (uncommon measurement times among subjects) are less well developed.

4. Simulation Studies

4.1 Performance of joint modeling and MI estimates under correct specification

A simulation study was conducted by generating 1000 replicates of a longitudinal data set with two scheduled visits and a sample size of $n = 500$. The outcome of interest was a normally distributed continuous variable. Two types of auxiliary variables were considered: continuous, with a normal distribution; and categorical, with a Bernoulli distribution. In the continuous auxiliary variable case, the outcome of interest $Y = Y_1, Y_2$ and the auxiliary variable $Z = Z_1, Z_2$ were generated from multivariate normal distributions, with

$Var(Y_j) = \sigma_Y^2$, $Var(Z_j) = \sigma_Z^2$, $\rho = \text{corr}(Z_1, Z_2) = \text{corr}(Y_1, Y_2)$, $r = \text{corr}(Y_j, Z_j)$, $j = 1, 2$. The covariance structure of the multivariate longitudinal outcome is a Kronecker product in this case, as in the example in section 3.1.1. The model for the mean of Y is $E(Y_{ij}) = \beta_0 + \beta_1 t_{ij}$, where t_{ij} is 0 if $j = 1$ and 1 if $j = 2$. The parameter β_1 measures the expected decline on Y at visit 2 compared to visit 1 and is the parameter of interest here. A similar model was assumed for the mean of Z is $E(Z_{ij}) = \gamma_0 + \gamma_1 t_{ij}$. The parameter was estimated using marginal multivariate models. We set $(\beta_0, \beta_1) = (0, -3)$, $(\gamma_0, \gamma_1) = (0, -2)$, $\sigma_Y^2 = 9$, $\sigma_Z^2 = 4$, $\rho = 0.5$ and $r = 0.8$. In the binary auxiliary variable case, the shared random effects model (4) was used to generate Y and Z . We $(\beta_0, \beta_1) = (0, -3)$, $(\tau_0, \tau_1) = (0, -0.2)$, $\sigma_b^2 = 4$, $\sigma_\varepsilon^2 = 6$, and $\lambda = 1$. In both cases, all variables were completely observed except that Y_2 , the second measurement of Y , could be missing. The observation indicator R for Y_2 was generated from the following logistic model:

$$\text{logit} \{P(R=1|Y, Z)\} = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Y_1 + \alpha_4 Y_2. \quad (6)$$

We set $(\alpha_0, \alpha_1, \alpha_3) = (1, 0, 0.3)$ for continuous Z and $(0.2, 0, 0.3)$ for binary Z case. Different values for (α_2, α_4) were considered: the MAR case corresponds to $(\alpha_2, \alpha_4) = (0, 0)$; the A-MAR case corresponds to $\alpha_4 = 0$ but with a nonzero value for α_2 . We set α_2 to 0.5 and 0.1 for continuous Z , and 0.3 and 0.8 for binary Z . Finally, for the MNAR case which corresponds to a nonzero value for α_4 , we set $(\alpha_2, \alpha_4) = (0.5, 0.1)$ for continuous Z (0.8, 0.1) binary Z case in which the observation probability depends on the unobserved Y_2 . The regression method is used in the MI approach using SAS PROC MI. Simulation bias (Bias), standard error (STD) and the percentage of times the 95% confidence intervals covered the true parameter for β_1 (Coverage) are 1 presented in Table 1 for the different approaches.

The results show that both the joint modeling and multiple imputation methods that utilize the auxiliary information correct the bias from non-random missing longitudinal data under A-MAR, while the naive mixed effects model for Y yields biased estimates. With the other parameters fixed, the value of α_2 in the observation indicator model measures the extent to which the MAR assumption is violated. The more α_2 deviates from 0, the larger the violation. The bias in the 2 estimate from the naive model for Y increases with the magnitude of the violation of the MAR assumption. Under MAR, estimates from the linear mixed effects model for Y give consistent estimates as well and it is not necessary to utilize auxiliary variables in this circumstance. However, the two approaches utilizing the auxiliary information showed improved efficiency of the parameter estimate. Under MNAR, all

methods resulted in biased estimates. However, utilizing the auxiliary information reduced the bias.

4.2 Effects of model mis-specification

The above results assume that the joint distribution of Y and Z in the joint modelling approach and the imputation model for the multiple imputation approach are correctly specified. To examine the effect of model mis-specification on the estimate of the parameter of interest, we performed another simulation study where the auxiliary variable Z is generated from a right-skewed log-normal distribution but is modelled as a normal random variable in the joint modeling approach, and linearly associated with the mean of the outcome in the imputation model in MI approach. Specifically, Y and Z are generated as follows:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \varepsilon_{ij},$$

$$\log(Z_{ij}) = \tau_0 + \tau_1 t_{ij} + b_i + e_{ij},$$

where b_i is the random effect distributed as $N(0, \sigma_b^2)$, and ε_{ij} and e_{ij} are independent normally distributed random errors, also independent from b_i , with variance σ_ε^2 and σ_e^2 , respectively. Here $\beta_0 = 0$, $\beta_1 = -3$, $\tau_0 = 0$, $\tau_1 = -0.1$, $\sigma_b^2 = 1$, $\sigma_\varepsilon^2 = 1$ and $\sigma_e^2 = 0.5$. The observation indicator for Y_2 is generated from logistic model (6) with $(\alpha_0, \alpha_1, \alpha_3) = (0.4, 0, 0.3)$. The value of (α_2, α_4) is $(0, 0)$, $(0.3, 0)$ and $(0.2, 0.3)$ for the case of MAR, A-MAR and MNAR, respectively. In the mis-specified joint modeling approach, the distribution of Y and Z are assumed multivariate normal. In the mis-specified imputation model of the MI approach, Y_2 is assumed linearly related with Z .

Table 2 shows that the mis-specified joint modeling and MI approaches do not eliminate bias under A-MAR, and do not reduce the bias under MNAR either, as the correct methods should do, though the biases under MAR are negligible. In fact the bias from mis-specified joint modeling and MI approaches can be close to or even larger than the bias from modeling Y only. This suggests that it is important that the correct model specification should be used in the joint modeling and MI approaches when utilizing the auxiliary information, otherwise the benefit from utilizing the auxiliary information would be lost. The bias from the MI approach is slightly smaller than that from the joint modeling approach, suggesting that mis-specifying the joint distribution of Y and Z is more severe than mis-specifying the mean structure of the missing Y in relation to Z . In reality, such mis-specification problem can be avoided by checking the distribution of the auxiliary variable and applying normalizing transformation as appropriate.

In another simulation study, we generated multivariate normal Y and Z but mis-specified the covariance structure in the joint modeling approach, and found that it also caused bias in the parameter estimates (results not shown).

The simulation studies show that the joint modeling and MI approaches should be applied with caution. The joint distribution of Y and Z in the joint modeling approach and the imputation model in the MI approach should be carefully examined before a specific model is assumed.

5. Example

In a dementia screening study in a primary care geriatrics practice [3], the decline in memory as measured by the Free and Cued Selective Reminding Test (FCSRT) [4] between the follow-up visit and baseline is of interest. We used a subset of the data in which the primary care physicians' assessment of memory in the clinical dementia rating system (CDR) [5] was available at both baseline and follow-up ($n=238$). Observed baseline FCSRT ranged from 0 to 44 (mean=27.6, std=8.47). The follow-up FCSRT was missing for 59 (25%) subjects. The original CDR assessment of memory impairment is graded on a scale of 0–3, with 0 = no impairment (66.4% at baseline and 62.6% at follow-up); 0.5 = memory impairment (27.3% at baseline and 25.6% at follow-up); 1 = mild dementia (6.3% at baseline and 8.8% at follow-up); 2 = moderate dementia (0% at baseline and 2.9% at follow-up); and 3 = severe dementia (none in this data set). Because of the low frequency of CDR values greater than 1 in this population, we defined three-categories of CDR which were modeled with the following two indicator variables : $CDR_{half} = 1$ for $CDR = 0.5$ and 0 otherwise; and $CDR_{1P} = 1$ for $CDR \geq 1$ and 0 otherwise. The physicians' CDR memory impairment rating was highly associated with FCSRT performance. At baseline, mean (STD) of FCSRT among the groups with $CDR = 0$, $CDR = 0.5$ and $CDR \geq 1$ were 30.15 (6.54), 23.78 (8.99) and 16.80 (9.97), respectively, with significant differences ($p < 0.0001$) among the CDR categories. Hence the CDR memory rating is a potential auxiliary variable for FCSRT.

We then fit a logistic model for probability of missing the follow-up FCSRT as a function of baseline FCSRT and CDR memory rating at baseline and follow-up. The results, shown in Table 3, suggest that subjects with impaired baseline CDR memory rating are more likely to have missing follow-up FCSRT compared to those with unimpaired CDR memory rating at baseline ($p=0.016$). The likelihood ratio test for assessing whether the CDR memory rating can be omitted from the logistic model shows that CDR memory rating is significantly associated with the missing data process after adjusting for baseline FCSRT (Chi-square=9.666, degree of freedom=4, $p=0.046$). This suggests that the missing data process might be A-MAR rather than MAR.

Next, we estimated the decline in FCSRT using a linear mixed effects model for FCSRT only and the two methods that utilize the auxiliary information, CDR. In the joint modeling approach, the multinomially distributed CDR memory rating and the multivariate normally distributed FCSRT were jointly modeled using correlated random effects as described below.

$$FCSRT_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + o_{ij},$$

$$\log \left\{ \frac{P(\text{CDR}_{ij}=0.5)}{P(\text{CDR}_{ij}=0)} \right\} = \alpha_{01} + \alpha_{11}t_{ij} + b_{1i},$$

$$\log \left\{ \frac{P(\text{CDR}_{ij} \geq 1)}{P(\text{CDR}_{ij}=0)} \right\} = \alpha_{02} + \alpha_{12}t_{ij} + b_{1i},$$

where $i = 1, \dots, n$, $j = 1, 2$, are the subject and time index, respectively; t_{ij} is 0 if $j = 1$ and 1 if $j = 2$; (b_{0i}, b_{1i}) are the subject specific random effects distributed as bivariate normal with mean $(0, 0)$, marginal variance (σ_0^2, σ_1^2) and correlation coefficient ρ ; δ_{ij} is the normally distributed error term for FCSRT which is independent of the random effects. The parameter of interest β_1 represents the decline of FCSRT at follow-up compared to baseline. SAS 9.1 (SAS Institute Inc., Cary, N.C) procedure NLMIXED was used to fit this model.

In the multiple imputation approach, a linear regression model for the observed follow-up FCSRT was fit using baseline FCSRT, and both baseline and follow-up CDR memory rating. New parameters were randomly drawn from the posterior distribution of the parameters using a non-informative prior. The missing follow-up FCSRT was imputed using these new parameters and the baseline FCSRT and CDR memory rating at baseline and follow-up. This process was repeated 5 times. Each of the 5 imputed data sets was then used as a complete data to calculate the FCSRT decline using regular linear mixed effects model. The 5 sets of this parameter estimates were averaged to yield the point MI estimate. The standard errors of each parameter estimate and the variation among the 5 estimates were combined to calculate the variance of the MI estimate. SAS 9.1 procedures MI and MIANALYZE were used to obtain the MI estimate.

The results are shown in Table 4. Because subjects with poorer CDR memory rating tended to miss their follow-up visits for FCSRT, the linear mixed effects model which did not take account the CDR information underestimated the magnitude of FCSRT decline compared to the joint modeling or multiple imputation approach. The model we assumed for the joint modeling of FCSRT and CDR resulted in a slope estimate corresponding to a 4% greater decline in FCSRT than the naïve model. The multiple imputation estimates which are based on fewer assumptions corresponded to a 14% greater decline in FCSRT than the naïve model, suggesting that more flexible alternative joint models might need to be considered. The multiple imputation method makes assumptions only on the mean structure of the FCSRT and thus we believe its estimate of FCSRT decline may be closer to the true value if A-MAR holds.

6. DISCUSSION

Informative loss to follow-up is a major problem in longitudinal studies. Despite this fact, the use of models that assume MAR or MCAR remains common. We have shown that auxiliary information can be valuable for testing the MAR assumption for the main model of interest and for eliminating or reducing the bias when the missing process for the main

model is MNAR. In a dementia screening example with a single follow-up visit, we have shown that the estimate of cognitive decline may indeed be biased by informative loss to follow-up, and that the methods proposed here may mitigate that bias. The methods shown here should be applicable to studies with many waves of follow-up and to non-monotone patterns of missingness; future work should evaluate their effectiveness under such circumstances.

The auxiliary variable can be used to test the MAR assumption; if the data suggest that the missing data mechanism is MNAR, bias can be reduced if we take the auxiliary variables into account when analyzing the data even though these variables are not of primary interest. Joint modeling and multiple imputation methods can be applied to utilize the auxiliary information. However, additional modeling assumptions are introduced when utilizing the auxiliary information, and therefore, this information must be used judiciously. In practice, the distribution of the outcome of interest and the auxiliary variable should be carefully examined; we recommend careful checking of the joint distribution and flexible specification of the covariance structure in the joint modeling approach, and careful specification of the imputation model in the MI approach.

Others have advocated collection of auxiliary information that might be related to missing values (e.g., [22]); our work provides a framework for the productive use of the auxiliary information, either to provide a more informative sensitivity analysis, or possibly correct the bias from informative loss to follow-up. In many situations, it will be possible to collect auxiliary information in a cost-effective manner in such a way that the auxiliary information will be available when the outcome of interest is censored. We have demonstrated the practicality of using existing medical records from primary care providers for study participants enrolled in a longitudinal study; other possibilities of information auxiliary to that collected in an in-house research encounter might include information collected via telephone contact, or from informants. All longitudinal studies should take seriously the potential for bias from MNAR data and investigators should seriously consider collecting auxiliary information to evaluate and possibly correct for this bias.

Acknowledgments

The authors are grateful to Dr. Ellen Grober for providing the data and to Dr. Mimi Kim and referees for helpful comments. This research was supported by National Institute of Aging grants P01-AG03949 (PI: Richard Lipton) and R01-AG025119 (PI: Joe Verghese). Dr. Charles B. Hall was also supported by R01-AG017854 (PI: Ellen Grober).

REFERENCES

1. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
2. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd edition. John Wiley; New York: 2002.
3. Grober E, Hall C, Lipton RB, Teresi JA. Primary Care Screen for Early Dementia. *Journal of the American Geriatrics Society*. 2008; 56:206–213. DOI: 10.1111/j.1532-5415.2007.01553.x. [PubMed: 18179486]
4. Grober E, Buschke H. Genuine memory deficits in dementia. *Developmental Neuropsychology*. 1987; 3:13–36.

5. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*. 1993; 43:2412–2414. [PubMed: 8232972]
6. Ibrahim JG, Lipsitz SR, Horton N. Using auxiliary data for parameter estimation with nonignorably missing outcomes. *Applied Statistics*. 2001; 50:361–373.
7. Daniels, MJ.; Hogan, JW. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall; New York: 2008.
8. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 1994; 9:538–573. DOI: 10.1214/ss/1177010269.
9. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996; 91:473–489.
10. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedure. *Psychological Methods*. 2001; 6:330–351. [PubMed: 11778676]
11. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
12. Agresti A, Booth JG, Hobart JP, Caffo B. Random effects modeling of categorical response data. *Sociological Methodology*. 2000; 30:27–80. DOI: 10.1111/0081-1750.t01-1-00075.
13. Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. *Longitudinal data analysis*. Chapman & Hall; Boca Raton, FL: 2009.
14. Paik MC. The generalized estimating equation approach when data are not missing completely at random. *Journal of American Statistical Association*. 1997; 92:1320–1329.
15. Rubin, DB. *Multiple imputation for nonresponse in surveys*. Wiley; New York: 1987.
16. Schafer, JL. *Analysis of incomplete multivariate data*. Chapman & Hall; New York: 1997.
17. Fay RE. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 1992:227–232.
18. Liu M, Taylor JMG, Belin TR. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*. 2000; 56:1157–1163. [PubMed: 11213759]
19. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
20. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90:106–121.
21. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*. 2007; 16:219–242. DOI: 10.1177/0962280206074463. [PubMed: 17621469]
22. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*. 1995; 90:1112–1121.

Table 1Simulation results comparing joint modeling and multiple imputation approaches with modeling *Y* only

Missing data cases	Method	Bias	STD	Coverage
Continuous auxiliary variable <i>Z</i>				
MAR	Model for <i>Y</i> only	0.0003	0.1573	0.942
	Joint modeling of <i>Y</i> and <i>Z</i>	0.0009	0.1307	0.945
	Multiple imputation	0.0002	0.1410	0.956
A-MAR (weaker dependence on <i>Z</i>)	Model for <i>Y</i> only	0.1154	0.1589	0.887
	Joint modeling of <i>Y</i> and <i>Z</i>	0.0008	0.1329	0.949
	Multiple imputation	0.0019	0.1422	0.956
A-MAR (stronger dependence on <i>Z</i>)	Model for <i>Y</i> only	0.7404	0.1916	0.027
	Joint modeling of <i>Y</i> and <i>Z</i>	0.0001	0.1388	0.948
	Multiple imputation	0.0019	0.1661	0.948
MNAR	Model for <i>Y</i> only	1.0370	0.2057	0.000
	Joint modeling of <i>Y</i> and <i>Z</i>	0.0826	0.1446	0.907
	Multiple imputation	0.1237	0.1779	0.883
Binary auxiliary variable <i>Z</i>				
MAR	Model for <i>Y</i> only	−0.0038	0.2019	0.947
	Joint modeling of <i>Y</i> and <i>Z</i>	−0.0082	0.1940	0.953
	Multiple imputation	−0.0051	0.2116	0.944
A-MAR (weaker dependence on <i>Z</i>)	Model for <i>Y</i> only	0.0462	0.1988	0.938
	Joint modeling of <i>Y</i> and <i>Z</i>	−0.0049	0.1893	0.957
	Multiple imputation	−0.0021	0.2047	0.950
A-MAR (stronger dependence on <i>Z</i>)	Model for <i>Y</i> only	0.1101	0.1938	0.896
	Joint modeling of <i>Y</i> and <i>Z</i>	−0.0033	0.1839	0.952
	Multiple imputation	0.0017	0.2011	0.945
MNAR	Model for <i>Y</i> only	0.4531	0.2079	0.367
	Joint modeling of <i>Y</i> and <i>Z</i>	0.2603	0.1938	0.744
	Multiple imputation	0.3159	0.2214	0.704

Table 2

Simulation results on effects of mis-specification where the log-normally distributed auxiliary variable is modelled as normal random variable in the joint modeling and MI approaches

Missing data cases	Method	Bias	STD	Coverage
MAR	Model for Y only	-0.0010	0.0804	0.936
	Correct joint modeling of Y and Z	-0.0007	0.0748	0.940
	Multiple imputation	0.0010	0.0772	0.945
	mis-specified joint modelling of Y and Z	0.0293	0.0826	0.934
	mis-specified multiple imputation	0.0053	0.0792	0.945
A-MAR	Model for Y only	0.0593	0.0787	0.877
	Correct joint modeling of Y and Z	-0.0007	0.0758	0.951
	Multiple imputation	-0.0016	0.0794	0.946
	mis-specified joint modelling of Y and Z	0.0798	0.0841	0.857
	mis-specified multiple imputation	0.0511	0.0811	0.908
MNAR	Model for Y only	0.3029	0.0942	0.113
	Correct joint modeling of Y and Z	0.1884	0.0890	0.433
	Multiple imputation	0.2089	0.0982	0.513
	mis-specified joint modelling of Y and Z	0.3281	0.1006	0.082
	mis-specified multiple imputation	0.3018	0.1011	0.241

Table 3

Estimates from the logistic model for missing follow-up FCSRT

Effects	Estimates	Standard Error	p-value
Baseline FCSRT	−0.005	0.023	0.811
CDRhalf (baseline)	0.936	0.389	0.016
CDR1p (baseline)	0.056	0.726	0.938
CDRhalf (follow-up)	−0.001	0.417	0.998
CDR1p (follow-up)	0.659	0.584	0.259

Table 4

Estimates of FCSRT decline using different methods

Method	Estimate	Standard Error	p-value
Regular linear mixed effects model for FCSRT	-2.283	0.432	<0.0001
Joint modeling of FCSRT and CDR	-2.384	0.429	<0.0001
Multiple Imputation	-2.600	0.612	0.0014