BIOMETRICS 56, 1157–1163
December 2000

# Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies

**Minzhi Liu,[1] Jeremy M. G. Taylor,[2,*] and Thomas R. Belin[3]**

[1]Clinical Biostatistics, Merck and Co., Inc., Rahway, New Jersey 07065, U.S.A.
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[3]Department of Biostatistics, UCLA School of Public Health, Los Angeles, California 90095, U.S.A.
**email:* jmgt@umich.edu

SUMMARY. This paper outlines a multiple imputation method for handling missing data in designed longitudinal studies. A random coefficients model is developed to accommodate incomplete multivariate continuous longitudinal data. Multivariate repeated measures are jointly modeled; specifically, an i.i.d. normal model is assumed for time-independent variables and a hierarchical random coefficients model is assumed for time-dependent variables in a regression model conditional on the time-independent variables and time, with heterogeneous error variances across variables and time points. Gibbs sampling is used to draw model parameters and for imputations of missing observations. An application to data from a study of startle reactions illustrates the model. A simulation study compares the multiple imputation procedure to the weighting approach of Robins, Rotnitzky, and Zhao (1995, *Journal of the American Statistical Association* **90,** 106–121) that can be used to address similar data structures.

KEY WORDS: Gibbs sampling; Missing data; Multiple imputation; Multivariate longitudinal data.

## 1. Background

In designed longitudinal studies, missing data often occur because subjects miss visits during the study, because some variables may not be measured at particular visits, or because subjects drop out. The aim in such studies is frequently to relate fixed covariates to the longitudinally measured response variables, to relate the response variables to each other, or to estimate the mean response at a certain time. The absence of complete data is a serious impediment to pursuing such aims.

Methods for handling incomplete data can be classified based on the nature of assumptions made about a data model and about the missing-data mechanism. Methods can also be classified according to whether the statistical solution to the problem involves reweighting of observations or imputation of missing values. Maximum-likelihood strategies have been developed to handle certain types of missing data in longitudinal studies (Laird and Ware, 1982; Jennrich and Schluchter, 1986) but not for the case of both missing response variables and missing covariates.

Weighting approaches are commonly used in sampling methodology when there is unit nonresponse, i.e., when no outcome data is available on an individual. The contribution of such an individual to the analysis can be reflected by attaching greater weight to an individual with observed outcomes who has similar covariate data to the nonrespondent. Robins et al. (1995) have developed weighting approaches to analyze incomplete longitudinal data to avoid having to rely on a multivariate model for the data. Their generalized estimating equations approach uses a model for the mechanism giving rise to nonresponse to determine weights.

In contrast, imputation approaches involve filling in missing items with plausible values given the observed data, where plausible values are obtained from either an explicit parametric model or an implicit model, as with hot-deck imputation (Rubin, 1987), or in a partially parametric way (Schenker and Taylor, 1996). To account properly for uncertainty due to values being missing, it is standard to produce multiple imputations and to obtain inferences about quantities of interest by combining estimates of within-imputation and between-imputation variability (Rubin, 1987). Imputation approaches have considerable flexibility because they can be used to address missing data at either the individual (unit) level or at the measurement (item) level. General purpose algorithms for imputation of missing values have been developed for cross-sectional data (Schafer, 1997).

Multiple imputation is usually thought of as a model-based approach since imputed values are often produced from a model for the data. There are situations where a model might be used to produce imputations but where a less restrictive or an uncongenial model (Meng, 1994; Little and Yau, 1996) may be used to analyze the completed data. Despite the popularity of both weighting and multiple imputation schemes, little empirical work has been done on comparing their properties.

The current paper develops a method for analyzing incomplete multivariate longitudinal data assuming an ignorable nonresponse mechanism. Gibbs sampling is used to fit the model. The sequence of variables in the sampler include both parameters and missing observations. Thus, the procedure can be used not only to draw inferences about model parameters but also to produce multiply imputed data sets for further analysis. We thus aim to accommodate both inference under a random coefficients model for longitudinal data and the distinct alternative of using such a model only for purposes of filling in missing values, thereby permitting flexibility at the later analysis stage. An experiment on children measuring startle reactions to a series of auditory stimuli provides a motivating example for the model. Two response variables were measured, one concerned with blinking and one related to heart rate. A random effects model in which the intercept and slope of both response variables were considered as random is a reasonable choice for these data. Unfortunately, there was some missing data, particularly toward the end of each child's sequence of responses.

Section 2 describes the multivariate random coefficients model. Section 3 gives the conditional distributions of parameters and missing data needed to implement Gibbs sampling. Section 4 describes an application to the startle reaction data. Section 5 summarizes the results of two simulation studies. Section 6 contains some discussion.

## 2. Multivariate Model for Longitudinal Data

### 2.1 *Framework*

We assume an ignorable missing-data mechanism (Rubin, 1976) so that the posterior predictive distribution of the missing data given the observed data can be used to multiply impute missing values without specifying a model for the process giving rise to the missing data.

The multivariate hierarchical model we use is an extension of the popular two-stage hierarchical model for a single longitudinal outcome variable (Laird and Ware, 1982). The model is similar to that described in Schluchter et al. (1990) and Zucker, Zerbe, and Wu (1995), although we use the Gibbs sampling estimation method in contrast to their use of either the method of moments or maximum likelihood.

### 2.2 *Model Details and Notation*

The variables are partitioned into time-dependent and time-independent variables. Furthermore, there is a defined set of time points at which every time-dependent variable could be measured. Let $Y_{ijk}$ be the observation, possibly missing, for the $i$th person of $j$th time-dependent variable at sampling time $t_k$, $i = 1, \dots, n$, $j = 1, \dots, J$, and $k = 1, \dots, K$. Let $X_{iq}$ be the observation of the $q$th time-independent variables for the $i$th person, $q = 1, 2, \dots, Q$.

For the time-independent variables, let $X_i = (X_{i1} X_{i2} \cdots X_{iQ})'$, let $X_i^{(1)} = (X_{i1} X_{i2} \cdots X_{iQ_1})'$ be variables with potentially missing values, and let $X_i^{(2)} = (X_{iQ_1+1} X_{iQ_1+2} \cdots X_{iQ})'$ be variables that are always observed for all $i$. $X_i^{(2)}$ may contain some categorical variables. We only need to assume a distribution for $X_i^{(1)}$ conditional on $X_i^{(2)}$, which is multivariate normal, i.e.,

$$X_i^{(1)} \mid X_i^{(2)} \sim \mathrm{MVN}\left(\mu_f + \beta_f' X_i^{(2)}, \Sigma_f\right),$$

where the subscript $f$ is used for the portion of the model where the variables are fixed in time. The parameters in this model are not the main interest but are introduced to complete the specification since they are required for the estimation procedure.

Each time-dependent variable is assumed to be a linear growth curve,

$$Y_{ijk} = \eta_{0ij} + \eta_{1ij} t_k + e_{ijk},$$

with individual random intercepts and slopes as follows:

$$\begin{aligned} \eta_{0ij} &= \beta_{0j}^0 + \beta_{0j}^1 X_{i1} + \cdots + \beta_{0j}^Q X_{iQ} + \alpha_{0ij} \\ \eta_{1ij} &= \beta_{1j}^0 + \beta_{1j}^1 X_{i1} + \cdots + \beta_{1j}^Q X_{iQ} + \alpha_{1ij}. \end{aligned} \tag{1}$$

Let $\alpha_{ij} = (\alpha_{0ij} \alpha_{1ij})'$ and $\alpha_i = (\alpha_{i1}' \alpha_{i2}' \cdots \alpha_{iJ}')'$. We assume $\alpha_i \sim \mathrm{MVN}(0, \Sigma_\alpha)$ and $e_{ijk} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_{jk}^2 I)$. In many applications, $\sigma_{jk}^2$ might be assumed to depend only on $j$.

We use $\mathrm{Y}_{ij}$ to denote the vector of repeated measurements for the $j$th variable of the $i$th subject. Then

$$Y_{ij} = B_i \beta_j + w \alpha_{ij} + e_{ij},$$

where $\beta_j^q = (\beta_{0j}^q \beta_{1j}^q)'$, $\beta_j = (\beta_j^{0\prime} \beta_j^{1\prime} \cdots \beta_j^{Q\prime})'$, $B_i = (1 X_{i1} X_{i2} \cdots X_{iQ}) \otimes w$, $w = (w_1 w_2 \cdots w_K)'$, and $w_k = (1 t_k)'$. Concatenating the $\mathrm{Y}_{ij}$'s into one vector $\mathrm{Y}_i$ gives

$$Y_i = (I_J \otimes B_i)\beta + (I_J \otimes w)\alpha_i + e_i = U_i \beta + Z \alpha_i + e_i, \tag{2}$$

where $U_i$ and $Z$ are block diagonal, $e_i \sim \mathrm{MVN}(0,\ \Sigma_e)$, and $\Sigma_e$ is diagonal with elements $\sigma_{jk}^2$.

We note that $Z$ and $w$ do not depend on $i$; this is a consequence of assuming the same set of potential time points for each person. The models and theory in this paper easily extend to other designs in which there is a different set of potential time points for each person.

## 3. Model Fitting Using Gibbs Sampling

Gibbs sampling in which the parameters and missing values are drawn iteratively from appropriate conditional distributions is used to obtain the joint posterior distribution of parameters and missing values given observed data. The Gibbs sampler is also used to produce the multiple imputes.

### 3.1 *Specification of Prior Distributions*

Let $\gamma = (\Sigma_f, \mu_f, \beta_f, \Sigma_\alpha,\ \alpha_i, \beta, \sigma_{jk}^2)$ denote the parameters. We use convenient priors so that it is easy to simulate from conditional distributions. Specifically, we use flat priors for $\mu_f$, $\beta_f$, and $\beta$ and conjugate priors for other parameters. Conjugate priors for the covariance matrices $\Sigma_f$ and $\Sigma_\alpha$ are inverse Wishart distributions $\Sigma_f^{-1} \sim \mathrm{W}(c, D)$ and $\Sigma_\alpha^{-1} \sim \mathrm{W}(m, \Lambda)$, respectively. A conjugate prior for the scalar parameter $\sigma_{jk}^2$ is an inverse gamma $\mathrm{IG}(a_{jk}/2, b_{jk}/2)$. Values for the hyperparameters c, D, m, Λ, $a_{jk}$, and $b_{jk}$ can be based on prior knowledge or chosen to give diffuse priors. We follow Schafer (1995, 1997) in our choice of $m = 2J + 1$, $\Lambda = I_{2J}/m$, $c = Q_1 + 1$, and $D = I_{Q_1}/c$, where $I_d$ is the $d \times d$ identity matrix and, we choose large values for $a_{jk}$ and $b_{jk}$. In practice, it is advisable to rerun the analysis with different priors to ensure that the results are not sensitive to the choice.

### 3.2 *Full Conditional Distributions*

Let $[\xi \mid \cdot]$ denote the full conditional distribution of a single component $\xi$ of $\gamma$ given all other components and the complete data. The full conditional distributions are

$$[\Sigma_f^{-1} \mid \cdot]$$

$$\sim \mathrm{W}\left(n+c-1, R^{-1}\right)$$

$$[\mu_f \mid \cdot]$$

$$\sim \mathrm{MVN}\left(\sum_{i=1}^{n}\left(X_i^{(1)} - \mathbf{1}\beta_f' X_i^{(2)}\right)/n, \Sigma_f/n\right),$$

$$[\beta_f \mid \cdot]$$

$$\sim \mathrm{MVN}\left(Q^{-1}\sum_{i=1}^{n} X_i^{(2)}\mathbf{1}'\Sigma_f^{-1}\left(X_i^{(1)} - \mu_f\right), Q^{-1}\right),$$

$$[\Sigma_\alpha^{-1} \mid \cdot]$$

$$\sim \mathrm{W}\left(n+m, \left(\sum_{i=1}^{n}\alpha_i\alpha_i' + \Lambda^{-1}\right)^{-1}\right),$$

$$[\alpha_i \mid \cdot]$$

$$\sim \mathrm{MVN}\Bigg(\left(Z'\Sigma_e^{-1}Z + \Sigma_\alpha^{-1}\right)^{-1}\left(Z'\Sigma_e^{-1}(Y_i - U_i\beta\right),$$

$$\left(Z'\Sigma_e^{-1}Z + \Sigma_\alpha^{-1}\right)^{-1}\Bigg),$$

$$[\beta \mid \cdot]$$

$$\sim \mathrm{MVN}\left(\left(\sum_{i=1}^{n}U_i'\Sigma_e^{-1}U_i\right)^{-1}\right.$$

$$\times \sum_{i=1}^{n}U_i'\Sigma_e^{-1}(Y_i - Z\alpha_i),$$

$$\left.\times\left(\sum_{i=1}^{n}U_i'\Sigma_e^{-1}U_i\right)^{-1}\right),$$

$$[\sigma_{jk}^2 \mid \cdot]$$

$$\sim \mathrm{IG}\left(\frac{n + a_{jk}}{2}, \frac{RSS_{jk} + b_{jk}}{2}\right),$$

where

$$R = \sum_{i=1}^{n}\left(X_i^{(1)} - X_{\cdot}^{(1)}\right)\left(X_i^{(1)} - X_{\cdot}^{(1)}\right)' + D^{-1},$$

$$Q = \sum_{i=1}^{n} X_i^{(2)}\mathbf{1}'\Sigma_f^{-1}\mathbf{1}X_i^{(2)\prime},$$

and

$$RSS_{jk} = \sum_{i=1}^{n}[(Y_{ij} - B_i\beta_j - w\alpha_{ij})]_k^2$$

for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$ and where the subscript $k$ denotes the $k$th element of the given vector.

In the model where we assume $\sigma_{jk}^2 = \sigma_j^2$, the required conditional distribution for $\sigma_j^2$ is

$$[\sigma_j^2 \mid \cdot] \sim \mathrm{IG}\left(\frac{Kn + a_j}{2}, \frac{\sum_{k=1}^{K} RSS_{jk} + b_j}{2}\right).$$

### 3.3 *Using the Model to Produce Multiply Imputed Data Sets*

As noted earlier, it may be desirable to use a multivariate model for the data only to fill in missing values, permitting more flexibility in subsequent analyses. For our model, all of the required conditional distributions to impute missing values given parameters and observed data are Gaussian.

For missing time-dependent variables, if the $l$th component of $Y_i$ is missing, then $[Y_{i(l)} \mid X_i, \beta, \alpha_i, \Sigma_e]$ is a univariate normal distribution obtained as the $l$th component of $(Y_i \mid X_i, \beta, \alpha_i, \Sigma_e) \overset{\text{ind.}}{\sim} \mathrm{N}(U_i\beta + Z\alpha_i, \Sigma_e)$ from equation (2). For time-independent variables for person $i$, we have

$$\left[X_i^{(1)} \mid X_i^{(2)}, Y_i, \gamma\right] \sim \mathrm{N}\left(H^{-1}G_i, H^{-1}\right), \qquad (3)$$

where $H = V^{(1)\prime}\Sigma_e^{-1}V^{(1)} + \Sigma_f^{-1}$ and $G_i = V^{(1)\prime}\Sigma_e^{-1}(Y_i - Z\beta^0 - V^{(2)}X_i^{(2)} - Z\alpha_i) + \Sigma_f^{-1}(\mu_f + \beta_f'X_i^{(2)})$. The terms $V^{(1)}, V^{(2)}$, and $\beta^0$ are defined by

$$\begin{aligned} Y_i &= U_i\beta + Z\alpha_i + e_i \\ &= Z\beta^0 + V^{(1)}X_i^{(1)} + V^{(2)}X_i^{(2)} + Z\alpha_i + e_i, \end{aligned}$$

where $\beta^0 = (\beta_1^0\beta_2^0\cdots\beta_J^0)'$ and

$$V = \begin{pmatrix} w\beta_1^1 & \cdots & w\beta_1^Q \\ w\beta_2^1 & \cdots & w\beta_2^Q \\ \vdots & \vdots & \vdots \\ w\beta_J^1 & \cdots & w\beta_J^Q \end{pmatrix}.$$

Then to obtain the conditional distribution of the missing $X_i^{(1)}$'s, we use equation (3) and the standard result about conditional distributions from multivariate normals.

After the Gibbs sampler converges, the samples of missing values will converge to the predictive distribution of $[Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}]$, where $Y_{\mathrm{obs}}$ denotes observed and $Y_{\mathrm{mis}}$ denotes missing measurements. From this sequence, we can select $M$ draws of missing values, with long lag times between iterations of the Gibbs sampler to avoid autocorrelation, to obtain $M$ complete data sets for further analysis.

## 4. Application to Startle Response Data

The multivariate modeling can be used to draw inferences directly if the target quantities of interest are model parameters. Here we illustrate such an approach in analyzing data from an experiment on startle response (Ornitz et al., 1996). The study collected data on individual differences in the blink responses to 40 sequential repetitive acoustic stimuli (trials) on 40 school-age boys. Besides the startle response, heart rate and various central nervous system measures were recorded. A major goal of the study was to determine to what extent individual variation in startle response can be explained by association with other measures. Some data were missing from the study because some trials were not usable due to fluctuations in background electromyography or subjects' spontaneous blinks had occurred just before the stimulus. The average number of rejected trials per subject was 3.35, giving 8.4% missingness in the data set. In a previous analysis of these data, the average of four consecutive trials was taken to construct one block, and block averages were analyzed to minimize the impact of missing data. Here we study a bivariate response consisting of the log transformed startle response (*LOGAMP*) and the heart rate just prior to the stimuli (*HR*).

**Table 1**

*Gibbs sampling results for $\beta$, $\sigma^2$, and covariance parameters $\Sigma_\alpha$ for LOGAMP and HR of the startle data*

| Parameter | Posterior mean | Standard deviation | Two-sided $P$-value |
|---|---|---|---|
| $\beta_{01}$ | 4.94 | 0.15 | 0.000 |
| $\beta_{11}$ | −0.12 | 0.015 | 0.000 |
| $\beta_{02}$ | 74.81 | 1.83 | 0.000 |
| $\beta_{12}$ | 0.40 | 0.15 | 0.006 |
| $\sigma_1^2$ | 0.13 | 0.013 | 0.000 |
| $\sigma_2^2$ | 23.51 | 2.22 | 0.000 |
| $\Sigma_{\alpha,11}$ | 0.78 | 0.21 | 0.000 |
| $\Sigma_{\alpha,22}$ | 0.0048 | 0.0017 | 0.004 |
| $\Sigma_{\alpha,33}$ | 122.79 | 32.23 | 0.000 |
| $\Sigma_{\alpha,24}$ | −0.013 | 0.012 | 0.289 |
| $\Sigma_{\alpha,44}$ | 0.30 | 0.15 | 0.042 |

We use Gibbs sampling to fit the random coefficient regression model,

$$LOGAMP_{ik} = \beta_{01} + \beta_{11}k + \alpha_{0i1} + \alpha_{1i1}k + \varepsilon_{i1k}$$
$$HR_{ik} = \beta_{02} + \beta_{12}k + \alpha_{0i2} + \alpha_{1i2}k + \varepsilon_{i2k},$$

with random effects $\alpha_i=(\alpha_{0i1}\alpha_{1i1}\alpha_{0i2}\alpha_{1i2})' \sim \mathrm{N}(0,\Sigma_\alpha)$. Of particular substantive interest is the covariance component estimate ($\Sigma_{\alpha,24}$) for the association between the slopes of the startle response and the heart rate. Ten multiple Gibbs sampler sequences of 2000 iterations each were run, and convergence was checked by monitoring the potential scale reduction factor of Gelman and Rubin (1992).

Table 1 shows the posterior means and standard deviations calculated from the last 1000 iterations of the first Gibbs sampler sequence. The results indicate that there are significant variations among individual sizes and rates of habituation of startle responses ($LOGAMP$) and those of $HR$ (represented by $\Sigma_{\alpha,11}, \Sigma_{\alpha,22}, \Sigma_{\alpha,33}, \Sigma_{\alpha,44}$). But there are not significant associations between the slope of startle response and that of heart rate (represented by $\Sigma_{\alpha,24}$).

To accommodate inferences about target quantities that are not explicit parameters of the multivariate model, a straightforward strategy would be to analyze data in a multiple imputation framework by taking values from the last of each of the parallel Gibbs sampling sequences.

## 5. Simulation Studies

Two studies were designed to evaluate multiple imputation inference based on the Gibbs sampling strategy. We used five imputations widely separated from a single converged Gibbs sampler chain.

### 5.1 *Comparison with Complete-Case Analysis*

The first study compared our multivariate modeling framework with complete-case analysis in a setting with five repeated measurements on two time-dependent variables where missingness of the time-dependent variable depended strongly on a binary time-independent covariate. Not surprisingly, for the cross-sectional mean of one of the time-dependent variables at the last time point, complete-case analysis was severely biased with very poor coverage (only 0.5% for a nominal 95% interval), while multiple imputation inference had minimal bias, 95.5% coverage, and efficiency comparable to estimation based on the originally generated data before deletion of missing values. For a comparison of the difference in outcomes between the levels of the binary variable driving the missing-data mechanism, complete-case analysis showed little bias and 93% coverage, but the multiple imputation approach, which had similar bias and coverage properties, was 40% more efficient. The reader is referred to Liu, Taylor, and Belin (1995) for a more complete description and additional results.

### 5.2 *Comparison with Weighted Estimating Equations*

Robins et al. (1995) describe a method for the analysis of longitudinal data containing missing values using semiparametric regression models for missing repeated outcomes. The parameters of the regression models are estimated from a class of estimating equations that do not require full specification of the likelihood. The method requires the estimation of weights that are derived from a model for the missing-data mechanism. In a simulation study, the authors compared a number of estimation methods across seven nonresponse models with different combinations of covariates. Included in their study were some scenarios that were deliberately chosen to illustrate limitations of the method. Their simulation showed that correctly specified nonresponse models give results with little bias and good coverage properties, while deliberately misspecified nonresponse models can lead to some bias, somewhat worse coverage, and potentially tremendous variability in estimates of target quantities when the proposed weighted estimating equation approach is used.

We assess our Gibbs-sampling/multiple-imputation method using the same simulation study design. Two repeated-measures outcome variables with missing values, $Y_{it}$ (called CD4) and $V_{it}$ (called WBC), were generated for $t = 0, 1, 2, 3$ and $i = 1, \ldots, 500$ according to the following model:

$$Y_{it} = 200 - 40t + d_{0i}(6 - t) + \varepsilon_{0it}$$
$$V_{it}^{1.33} = 3,000 - 100t + d_{1i}(10 - t) + \varepsilon_{1it}.$$

The random effects ($d_{0i}$,$d_{1i}$) were bivariate normal with mean zero, squared correlation coefficient $\rho^2$ either .81 or .36, and variances $(4.5^2, 100^2)$. We note that this model has a nonstandard covariance structure since it requires only two random effects. The measurement errors were generated independently as follows: $\varepsilon_{1it} \sim \mathrm{N}(0, 200^2)$ for all t, $\varepsilon_{0i0} \sim \mathrm{N}(0, 40^2)$, $\varepsilon_{0i1} \sim \mathrm{N}(0, 35^2)$, $\varepsilon_{0i2} \sim \mathrm{N}(0, 25^2)$, $\varepsilon_{0i3} \sim \mathrm{N}(0, 10^2)$. The missing data were generated based on the response probability $\lambda_{it}$, which depended only on the population tercile of $V_{i(t-1)}$, with the conditional probability $\lambda_{it}$ of remaining on study at $t$ being .9, .75, and .5 for the highest, middle, and lowest terciles of $V_{i(t-1)}$. The framework further assumes that, once individuals leave the study, they remain off the study from then on.

The results for $\beta_{03}$, the average of $Y$'s at $t = 3$, from the Robins et al. (1995) paper are reproduced here in Table 2. For later comparison with our own method, we have reexpressed the Monte Carlo variance as relative to the sample average case. The true value of $\beta_{03}$ is 80.0. We can see that, when the last tercile WBC was included in the nonresponse models, the weighted estimating equation approach showed little bias, provided good coverage, and the variance of the target

**Table 2**

*Results from Robins et al.* (1995) *simulation study at* $t = 3$ *for* $\beta_{03} = \mathrm{E}[Y_{i3}] = 80.0$

| | | Monte Carlo average | | Monte Carlo relative variance | | 95% Actual coverage rate | |
|---|---|---|---|---|---|---|---|
| Method | Nonresponse model | $\rho^2(.81)$ | (.36) | $\rho^2(.81)$ | (.36) | $\rho^2(.81)$ | (.36) |
| Sample average | | 86.4 | 84.2 | 1.00 | 1.00 | 0.0 | 4.0 |
| Weighted | Linear CD4 | 84.0 | 82.5 | 1.25 | 1.08 | 3.0 | 44.0 |
| | Last tercile WBC | 80.0 | 80.0 | 1.50 | 1.50 | 94.5 | 96.0 |
| | Linear CD4, last tercile WBC | 80.1 | 80.0 | 1.42 | 1.50 | 94.0 | 95.5 |
| | Linear CD4, last WBC | 74.8 | 76.3 | 30.25 | 19.08 | 90.0 | 91.0 |
| | Linear CD4, last WBC, last tercile WBC | 80.1 | 80.1 | 1.33 | 1.42 | 96.0 | 95.5 |

quantity was moderate. When the last tercile of WBC was not included in the nonresponse model, there was some downward bias in the estimate of the target quantity and somewhat less than nominal coverage with a dramatic increase in the variance of estimates of the target quantity in one case.

Table 3 presents results applying our multiple imputation approach to datasets generated using the same model as Robins et al. (1995). We fit a model in which the variables $Y$ and $V^{1.33}$ have linear mean structure, random intercepts, random slopes, and heterogeneous error variances over time. We note that this model has the same mean structure as the model used to simulate the data but a different covariance structure. This model was used to create the multiple imputes for the missing $Y$'s at time 3, and the results for standard multiple imputation inference are shown in Table 3, labeled as the correct mean model. Values of the hyperparameters were chosen to correspond to proper but diffuse priors; specifically, for $j = 0, 1$ and $t = 0, 1, 2, 3$, $a_{jt} = 12$ and $b_{jt}$ is 10 times the value of $\sigma^2_{jt}$ used to generate the data.

The results from Table 3 indicate that the multiple-imputation approach produces little bias and good coverage. Furthermore, the multivariate modeling method appears more efficient than the weighted estimating equation method: the relative variances of the estimates were 1.03 ($\rho^2 = 0.81$) and 1.14 ($\rho^2 = 0.36$) for the multiple-imputation method and between 1.3 and 1.5 using appropriately specified weighted estimating equations.

Other simulations to investigate the effect of misspecification of mean and error structures for the imputation model were conducted. Results are shown in the last two rows of Table 3. Both models are based on a standard intercept and slope random effects model applied to the untransformed data $Y$ and $V$. In the model labeled as the linear model, heterogeneous $\sigma_{jt}$, we assume heterogeneous error variance (i.e., eight different values of $\sigma^2_{jt}$); in the model labeled as the linear model, homogeneous $\sigma_j$, we assume homogeneous error variance (i.e., two different values of $\sigma^2_j$). We see that incorrectly specifying the model does lead to some bias but much smaller bias compared with the sample average. The coverage rates are generally around 95%, although the specific example of 87% would be considered too low. The efficiency for the heterogeneous model is comparable to the correct model scenario. The efficiency in the homogeneous model case is much worse due to inappropriately large values of $\sigma_j$ being used to impute the missing values at time 3. The structure of both these models is such that it would be possible to ascertain from the observed data that the model is not adequate. This emphasizes that it is important to use a model that gives a good description of the observations. A complex question is whether one could always find such a model in order to give sufficiently accurate inference using the multiple imputation approach. In simulations concerned with robustness to the measurement error distribution, we found (results not shown) that mis-

**Table 3**

*Results for comparing with Robins et al.* (1995) *simulation study for* $\beta_{03} = \mathrm{E}[Y_{i3}] = 80.0$; *multiple imputation using model* (2)

| | Monte Carlo average | | Monte Carlo relative variance | | 95% coverage | |
|---|---|---|---|---|---|---|
| Method | $\rho^2(.81)$ | (.36) | $\rho^2(.81)$ | (.36) | $\rho^2(.81)$ | (.36) |
| All data | 80.03 | 80.01 | 0.40 | 0.41 | 96.5 | 96.0 |
| Sample average | 86.28 | 84.11 | 1.00 | 1.00 | 0.0 | 5.0 |
| Five imputations | | | | | | |
| Correct mean model | 79.83 | 80.23 | 1.03 | 1.14 | 94.5 | 93.0 |
| Linear model, heterogeneous $\sigma_{jt}$ | 80.44 | 80.25 | 1.09 | 1.13 | 87.0 | 93.0 |
| Linear model, homogeneous $\sigma_j$ | 79.93 | 80.23 | 2.28 | 2.73 | 92.0 | 94.0 |

specifying error distributions as normal when the underlying distribution is a Student's $t$ gives some loss of efficiency for parameter estimation but does not result in substantially worse coverage.

## 6. Discussion

We have alluded to three approaches to handling missing data: (i) a formal Bayesian analysis with a full probability model for the data, (ii) a full probability model but only used to fill in missing data with subsequent analysis of the observed and imputed data, and (iii) weighting methods. The issues associated with choosing an approach for handling missing data in multivariate settings are complex. Robins et al. (1995) note that, by modeling nonresponse probabilities, one is able to avoid parametric assumptions about multivariate data. Schafer (1997, pp. 143–144) notes that multivariate models might need to be restricted because there might not be enough information in the observed data to provide stable estimates of all conceivably relevant parameters. Concerns also arise regarding robustness since there may be substantial reliance on modeling assumptions when the percentage of missing data is high. Schafer (1997, pp. 211–212) describes research showing multiple imputation is robust to modest departures from normal assumptions. Of the three approaches, we would expect the second approach to be more robust than the first because it has less reliance on the model (Schafer, 1997, p. 144), although it is possibly less efficient.

The simplicity of the multiple imputation approach relies on the assumption of ignorable missing data as well as an appropriate model for the data. Because nonignorable effects are apt to be dependent on the context, it is hard to develop a general-purpose approach to nonignorable missingness. One recommendation in the literature is to base imputations on a model that includes as many relevant variables as possible (Rubin, 1996), even though the ultimate analysis may be based on a smaller set of variables. It is possible for missingness to be nonignorable when conditioning on only a few covariates, but additional covariates may account for the source of most of the nonignorability (e.g., David et al., 1986).

Another limitation of the present approach is its focus on continuously scaled outcomes. Generalizing the procedure to mixed categorical and continuous longitudinal variables would be challenging because of the lack of convenient multivariate models for such a mix of outcomes.

Our simulation results indicate that a model-based imputation strategy is a feasible and attractive methodology. The model-based approach produced more efficient estimates than the approach described in Robins et al. (1995) even though the covariance structure of the model-based approach was misspecified. A further difference is that separate estimating equations are needed for separate estimates, whereas Gibbs sampling can be implemented without regard to the quantity that will be estimated and can be used for a variety of estimates provided one has a good probability model for the observations.

Alternative weighted estimating equations similar to that described in Robins et al. (1995) have been proposed, e.g., the locally efficient augmented inverse probability estimators for monotone missingness (Robins, Rotnitzky, and Zhao, 1994). Theoretical work (Robins and Ritov, 1997; Lipsitz, Ibrahim, and Zhao, 1999; Scharfstein, Rotnitzky, and Robins, 1999) suggests that these estimators are more efficient than the Robins et al. (1995) estimator and are consistent if either the model for the observations or the model for the missingness is correctly specified. Further work comparing the properties of model-based and efficient weighted estimating equation approaches in moderate sized samples would be of interest.

Some experience in data sets with large numbers of available covariates suggests that the dimension of $\beta$ can get quite large, even with a modest number of time-independent variables (Belin et al., 1997). This happens because each additional time-fixed covariate adds multiple components to $\beta$ in expression (1). One possible extension is to allow different sets of covariates to predict intercepts and slopes for different time-dependent variables. There are other possible extensions to make it applicable to a wider range of data, e.g., to allow the categorical time-independent variables to be missing or the measurement error terms to be correlated or have nonnormal distributions. Another issue worthy of investigation is misspecification of random effects.

Not surprisingly, we see that certain violations of modeling assumptions have only minor effects on bias and coverage of the multiple imputation method, while other violations of assumptions have more serious effects (e.g., assuming homogeneous errors in the linear model results in a substantial loss of efficiency, although the bias remains minor and the coverage remains good in this case). The current paper, with its empirical comparison, contributes some insights into the relative strengths of imputation modeling versus the use of weighted estimating equations, a challenging question that is apt to remain of continuing interest to applied statisticians.

## Acknowledgement

This work was supported by NIH grants AI07370, AI29196, MH58107, and MH57082.

## Résumé

Cet article décrit une méthode d'imputation multiple pour le traitement des données manquantes dans les études longitudinales planifiées. Les observations manquantes sont prises en compte à l'aide d'un modèle à effets aléatoires pour données multivariées longitudinales continues qui traite conjointement les mesures répétées multivariées. Plus précisément, on suppose que les variables indépendantes du temps sont distribuées indépendamment et identiquement selon des lois normales et que les variables dépendant du temps suivent un modèle de régression hiérarchique à coefficients aléatoires, conditionnellement au temps et aux variables indépendantes du temps, avec des variances résiduelles qui fluctuent selon les variables et le temps. Le tirage des paramètres du modèle et des imputations remplaçant les valeurs manquantes sont effectuées à l'aide de l'échantillonneur de Gibbs. Nous appliquons ce modèle aux données d'une étude sur les réactions d'alarme. Nous comparons la procédure d'imputation multiple à l'approche pondérée de Robins Rotnitzky et Zhao également applicable à des observations de ce type.

## References

Belin, T. R., Piacentini, J. C., Rotheram-Borus, M. J., and Song, J. (1997). Handling missing items in a multivariate study of a suicide prevention program. *Proceedings of the American Statistical Association, Biometrics Section*, 60–65.

David, M., Little, R. J. A., Samuhel, M. E., and Trieste, R. K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association* **81,** 29–41.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **4,** 457–511.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics* **42,** 805–820.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94,** 1147–1160.

Little, R. J. A. and Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* **52,** 1324–1333.

Liu, M., Taylor, J. M. G., and Belin, T. R. (1995). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Proceedings of the American Statistical Association, Biometrics Section,* 142–147.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9,** 538–557.

Ornitz, E. M., Russell, A. T., Yuan, H., and Liu, M. (1996). Autonomic, electroencephalographic, and myogenic activity accompanying startle and its habituation during mid-childhood. *Psychophysiology* **33,** 507–513.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16,** 285–319.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90,** 106–121.

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

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

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

Schafer, J. L. (1995). Model-based imputation of census shortform items. In *Proceedings of the 1995 Annual Conference, Bureau of the Census,* 267–299. Washington, D.C.: U.S. Department of Commerce.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94,** 1096–1146.

Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* **22,** 425–446.

Schluchter, M. D. and the Modification in Diet in Renal Disease Study Group. (1990). Estimating correlation between alternative measures of disease progression in a longitudinal study. *Statistics in Medicine* **9,** 1175–1188.

Zucker, D. M., Zerbe, G. O., and Wu, M. C. (1995). Inference for the association between coefficients in a multivariate growth curve model. *Biometrics* **51,** 413–424.

*Received July* 1998. *Revised January* 2000.
*Accepted February* 2000.