# The Generalized Estimating Equation Approach When Data Are Not Missing Completely at Random

Myunghee Cho PAIK

We propose two methods for handling missing data in generalized estimating equation (GEE) analyses: mean imputation and multiple imputation. Each provides valid GEE estimates when data are missing at random. Missing outcomes are imputed sequentially starting from the outcome nearest in time to the observed outcome. The estimators from the two kinds of imputation are compared with the weighting method of Robins et al. We show that multiple imputation with an infinite number of replications is asymptotically equivalent to mean imputation. The methods are applied to a stroke study in which neurological outcomes are measured over time after stroke but some outcomes are missing due to death or loss to follow up.

KEY WORDS: Generalized estimating equation; Importance sampling; Missing at random; Missing data; Multiple imputation; Nonignorable missingness.

## 1. INTRODUCTION

In longitudinal studies, outcomes that are repeatedly measured over time may be correlated and some may be missing. Liang and Zeger (1986) proposed the generalized estimating equation (GEE) approach to longitudinal data, which makes assumptions only on the mean and variance–covariance structure for a vector outcome. The GEE approach may not be valid however when data are not missing completely at random (MCAR) (Rubin 1976); that is, when the missingness probability depends on the responses (Laird, 1988; Liang and Zeger 1986). Robins, Rotnitzky, and Zhao (1995) (RRZ) proposed a weighting method for rendering GEE analyses correct under missing-at-random (MAR) missing mechanisms. Xie and Paik (1997) considered imputation for GEE when covariates are MAR and outcomes are MCAR. In this article I consider imputation for non-MCAR outcomes in GEE analyses.

A motivating example is a longitudinal study of neurological changes after stroke (Tatemichi et al. 1992). In this study 241 stroke patients were examined at six time points; 7–10 days after stroke, at a 3-month follow-up visit, and then annually for 4 years after the stroke. The complete outcome for each subject consists of six indicators of normal cognitive function from six visits. Outcomes are partially missing for some patients because of death or loss at follow-up. Because patients who have normal cognitive functions were more likely to drop out, the GEE estimates for this study may be biased (Laird 1988; Liang and Zeger 1986).

## 2. STRUCTURE OF PROPOSED METHODS

Let $y_{it}$, and $\mathbf{x}_{it} = (1, \mathbf{x}_i, t)^T, i = 1, \ldots, K$, be the outcome and a vector of independent variables for individual $i$ at visit $t, t = 1, \ldots, N$, where $N$ denotes the number of scheduled visits. I denote complete outcomes and independent variables by $\mathbf{Y}_i^c = (y_{i1}, \ldots, y_{in_i}, \ldots, y_{iN})^T$ and $\mathbf{X}_i^c = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}, \ldots, \mathbf{x}_{iN})^T$. My primary interest is $E(\mathbf{Y}_i^c | \mathbf{X}_i^c) = \boldsymbol{\mu}_i^c = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iN})^T$. We assume that $X_i^c$ is completely observed for all $i$. Let $r_{it}$ take value 1 if $y_{it}$ is observed, and 0 otherwise, and assume a monotone missing pattern (e.g., Rubin 1987, p. 170), where $r_{i1} \geq r_{i2} \geq \cdots r_{iN-1} \geq r_{iN}$. In this study, $r_{i1} = 1$ for all $i$; that is, all patients are completely observed at baseline. Under the monotone missing pattern, the observed data can be expressed as $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots y_{in_i})^T, \mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots \mathbf{x}_{in_i})^T$. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots \mu_{in_i})^T$, where $n_i$ indicates the number of visits of the $i$th subject. Subject $i$ belongs to missing pattern $k$ if $n_i = k$. Let $\mathbf{D}_{it}$ be the history up to $t$; that is, $\mathbf{D}_{it} = (\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{it}), t = 1, \ldots, n_i$.

The GEE approach (Liang and Zeger 1986) allows regression modeling of longitudinal data specifying only the mean and variance of the outcome variables. The mean for the complete data is assumed to satisfy $\mu_{it} = h^{-1}(\eta_{it}) = h^{-1}(\mathbf{x}_{it}\boldsymbol{\beta})$, where $h$ is a link function and the estimating equation for $\boldsymbol{\beta}$ is

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_i \frac{\partial \mu_i^T}{\partial \boldsymbol{\beta}} V_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1}(\mathbf{Y}_i - \mu_i) = 0,$$

where $V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i(\mu_i)^{1/2} \boldsymbol{\Omega}_i(\boldsymbol{\rho}) A_i(\boldsymbol{\mu}_i)^{1/2}, A_i = \mathrm{diag}\{\mathrm{var}(y_{it})\}, \boldsymbol{\alpha} = (\phi, \boldsymbol{\rho})$, and $\boldsymbol{\Omega}_i(\boldsymbol{\rho})$ is a "working" correlation matrix of $\mathbf{Y}_i$. One can choose $\boldsymbol{\Omega}_i(\boldsymbol{\rho})$ as the identity matrix or the equicorrelation matrix, and the specified correlation structure need not be the correct correlation structure of $\mathbf{Y}_i$. The GEE estimator, say $\hat{\boldsymbol{\beta}}_{\mathrm{LZ}}$, is the root of $\mathbf{U}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) = 0$, where $\hat{\boldsymbol{\alpha}}$ is any $\sqrt{K}$-consistent estimator. Liang and Zeger (1986) showed that $\hat{\boldsymbol{\beta}}_{\mathrm{LZ}}$ is consistent and asymptotically normal and its variance can be consistently estimated by a sandwich-type estimator.

The estimating equation can be reexpressed using complete data as

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_i \frac{\partial \boldsymbol{\mu}_i^{cT}}{\partial \boldsymbol{\beta}} V_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \mathbf{R}_i (\mathbf{Y}_i^c - \boldsymbol{\mu}_i^c) = 0,$$

where $\mathbf{R}_i = \mathrm{diag}\{r_{it}\}$ with order $N$, and $\mathbf{V}_i^c = \mathrm{var}(\mathbf{Y}_i^c | \mathbf{X}_i^c)$. If data are missing completely at random, then

Myunghee Cho Paik is Associate Professor, Division of Biostatistics, School of Public Health, Columbia University, New York, NY 10032. This research was supported in part by National Institutes of Health grant R01-NS26179. The author is grateful to David Desmond for providing the neurology dataset and to Bruce Levin, Wei-Yann Tsai, Melissa Begg, Dan Heitjan, the editor, an associate editor, and referees for helpful comments.


Journal of the American Statistical Association
December 1997, Vol. 92, No. 440, Applications and Case Studies

$E(y_{it}|\mathbf{x}_{it}, r_{it} = 1) = E(y_{it}|\mathbf{x}_{it}, r_{it} = 0) = \mu_{it}$, and inference conditional on $r$ is valid, because the estimating function has mean 0. When the probability of missingness depends on the outcome, $E\{r_{it}(y_{it} - \mu_{it})\} \neq 0$, and the root of $\mathbf{U}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) = 0$ is a biased estimate of $\boldsymbol{\beta}$.

RRZ proposed a weighted estimating equation for obtaining unbiased GEE estimates under MAR:

$$U_{\mathrm{W}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma) = \sum \frac{\partial \boldsymbol{\mu}_i^{cT}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \boldsymbol{\Delta}_i (\mathbf{Y}_i^c - \boldsymbol{\mu}_i^c) = 0, \tag{1}$$

where the $t$th diagonal element of $\boldsymbol{\Delta}_i$ is $r_{it}/\Pr\{(\prod_{j=1}^{t} r_{ij}) = 1|\mathbf{D}_{it-1}, \gamma\}$. RRZ showed that if $\boldsymbol{\Delta}_i$ is estimated consistently, then the root $\hat{\boldsymbol{\beta}}_{\mathrm{W}}$ is consistent and asymptotically normal under MAR and monotone missing patterns.

My approach is to instead replace the missing $y$'s with imputed $y$'s to obtain an unbiased imputed estimating function. To remove the bias introduced by missing data, missing $y_{it}$ is replaced with $\tilde{y}_{it}$, an estimate of $E(y_{it}|\mathbf{D}_{in_i}, r_{it} = 0)$. The resulting imputed estimating equation is

$$\begin{aligned} \tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum \tilde{\mathbf{u}}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \sum \frac{\partial \boldsymbol{\mu}_i^{cT}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\tilde{\mathbf{Y}}_i^c - \boldsymbol{\mu}_i^c), \end{aligned} \tag{2}$$

where $\tilde{\mathbf{Y}}_i^c = (y_{i1}, y_{i2}, \ldots, y_{in_i}, \tilde{y}_{in_i+1}, \ldots, \tilde{y}_{iN})^T$. If $(\tilde{y}_{in_i+1}, \ldots, \tilde{y}_{iN})^T$ is consistent, then the $E\{\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})\}$ is 0 because

$$u_{it} = r_{it}(y_{it} - \mu_{it}) + (1 - r_{it})\{E(y_{it}|\mathbf{D}_{in_i}, r_{it} = 0) - \mu_{it}\}$$

has expectation 0, where the expectation is taken over the joint distribution of $(y_{it}, r_{it}, y_{it-1}, r_{it-1}, \ldots y_{i1}, r_{i1})$. The problem then reduces to estimating $E(y_{it}|\mathbf{D}_{in_i}, r_{it} = 0)$, which depends on the missingness probabilities. I consider the following three mechanisms for $t = 2, \ldots, N$:

M1. $\Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = \Pr(r_{it} = 1|y_{i1}, \mathbf{x}_i, r_{it-1} = 1)$

M2. $\Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = \Pr(r_{it} = 1|y_{i1}, \ldots, y_{it-1}, \mathbf{x}_i, r_{it-1} = 1)$

M3. $\Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = \Pr(r_{it} = 1|y_{i1}, \ldots, y_{it}, \mathbf{x}_i, r_{it-1} = 1)$.

M1 and M2 are MAR (Rubin 1976), and M3 is nonignorable (Laird 1988; Little and Rubin 1987).

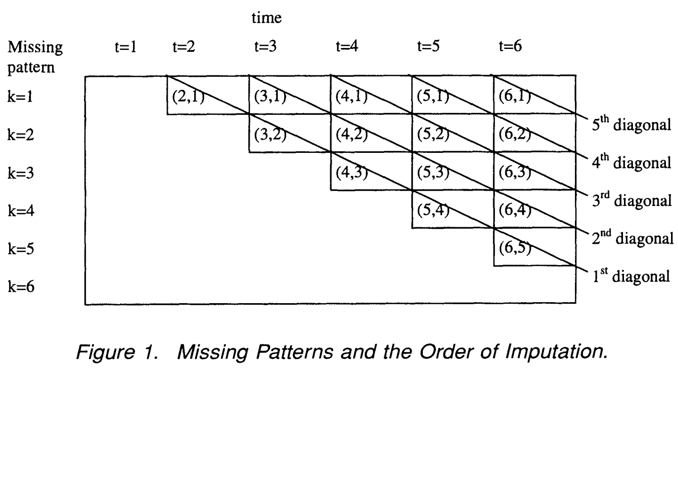


*Figure 1. Missing Patterns and the Order of Imputation.*

## 3. MEAN IMPUTATION

If $\mathbf{D}_{in_i}$ is categorical, then $E(y_{it}|\mathbf{D}_{n_i})$ can be estimated by the sample mean of observed $\mathbf{y}_{it}$'s that have the same history as $\mathbf{D}_{in_i}$. This sample mean has two drawbacks, however. First, it need not be consistent for $E(\mathbf{y}_{it}|\mathbf{D}_{n_i})$. This problem is attacked in Section 3.1 by considering sequential imputation. Second, there may be little or no data with exactly the same history. Actually, this is always true if some of elements in $\mathbf{D}_{n_i}$ are continuous. This problem is attacked in Section 3.2 by modeling $E(y_{it}|\mathbf{D}_{n_i})$.

### 3.1 Sequential Imputation

To justify imputation, the conditions under which missing outcomes have the same distribution as observed outcomes must be found. Under mechanisms M1 and M2, application of Bayes's theorem gives

$$\begin{aligned} &E(y_{it}|\mathbf{D}_{it-1}, r_{it-1} = 1, r_{it} = 0) \\ &\qquad = E(y_{it}|\mathbf{D}_{it-1}, r_{it-1} = 1, r_{it} = 1). \end{aligned} \tag{3}$$

This implies that $E(y_{it}|\mathbf{D}_{it-1}, r_{it-1} = 1, r_{it} = 0)$ can be consistently estimated by the sample mean of observed $y_{jt}$'s having the same history as $\mathbf{D}_{it-1}$; that is,

$$\begin{aligned} \tilde{y}_{it} &= \tilde{E}(y_{it}|\mathbf{D}_{it-1}, r_{it} = 1, r_{it-1} = 1) \\ &= \frac{\sum_{j=1}^{K} y_{jt} r_{jt} I(\mathbf{D}_{jt-1} = \mathbf{D}_{it-1})}{\sum_{j=1}^{K} r_{jt} I(\mathbf{D}_{jt-1} = \mathbf{D}_{it-1})}. \end{aligned} \tag{4}$$

For example, (3) implies that the expected value of the first diagonal cells in Figure 1 can be consistently estimated by observed sample means as shown in (4). However, the second through fifth diagonal cells can be consistently estimated by observed sample means only under M1, not under M2. Using Bayes's theorem, $E(y_{it}|\mathbf{D}_{it-2}, r_{it-2} = 1, r_{it-1} = 0)$, the expected value of a second diagonal cell, is equal to $E(y_{it}|\mathbf{D}_{it-2}, r_{it-1} = 1)$ but not equal to $E(y_{it}|\mathbf{D}_{it-2}, r_{it-1} = 1, r_{it} = 1)$, which is the term that can be estimated by a sample mean. For example, the expected value of cell (3, 1) in Figure 1, $E(y_{i3}|\mathbf{D}_{i1}, r_{i1} = 1, r_{i2} = 0)$, is not equal to $E(y_{i3}|\mathbf{D}_{i1}, r_{i2} = 1, r_{i3} = 1)$ (the expected value of observed outcome at visit 3), but $E(y_{i3}|\mathbf{D}_{i1}, r_{i2} = 1)$ (the expected value of observed outcome at visit 3 *and* cell (3, 2) in Fig. 1). $E(y_{i3}|\mathbf{D}_{i1}, r_{i2} = 1)$ cannot be estimated by the sample mean, say, $[\sum_{j=1}^{K} y_{j3} r_{j2} I(\mathbf{D}_{j1} = \mathbf{D}_{i1})]/[\sum_{j=1}^{K} r_{j2} I(\mathbf{D}_{j1} = \mathbf{D}_{i1})]$, because this quantity involves $y_{i3}$'s in cell (3, 2) that are unobserved. However, unobserved values in cell (3, 2) can be estimated using (4). Then a consistent estimate of the expected value of cell (3, 1) in Figure 1 can be obtained by the mean of observed and imputed data for cell (3, 2); that is,

$$\begin{aligned} &\left\{\sum r_{j2} I(\mathbf{D}_{j1} = \mathbf{D}_{i1})\right\}^{-1} \left\{\sum y_{j3} r_{j3} I(\mathbf{D}_{j1} = \mathbf{D}_{i1}) \right. \\ &\qquad \left. + \sum \tilde{y}_{j3} r_{j2} (1 - r_{j3}) I(\mathbf{D}_{j1} = \mathbf{D}_{i1}) \right\}. \end{aligned} \tag{5}$$

Therefore, we suggest the following sequential imputation: first, fill in the first diagonal cells using observed data; second, fill in the second diagonal cells using observed data *and* the imputed values for the first diagonals. This procedure is repeated along the diagonals to the right. For the fifth diagonal cell, observed data and imputed data from the first through fourth diagonal cells are used.

Note that the mean of the observed and imputed data in expression (5) consistently estimates $E(y_{i3}|\mathbf{D}_{i1}, r_{i2} = 1)$, which equals $E(y_{i3}|\mathbf{D}_{i1}, r_{i1} = 1, r_{i2} = 0)$. That is, sequential estimation is based on the equality

$$E(y_{it}|\mathbf{D}_{it-j}, r_{t-j} = 1, r_{it-j+1} = 0) = E(y_{it}|\mathbf{D}_{it-j}, r_{it-j+1} = 1). \quad (6)$$

This sequential imputation is valid under M1 and M2. Under M1, however, filling in each cell with observed sample means also yields valid results. This sequential imputation results in imputing a weighted mean. For example, after replacing $\tilde{y}_{j3}$ in (5) with $\tilde{y}_{j3}$ in (4), expression (5) can be written as a weighted mean of observed $y_{j3}$'s with history $\mathbf{D}_{i1}$, where the weight is an estimated density $d\hat{F}(y_2|D_1)$ computed from the completely observed data up to $t = 2$. If this density is estimated from the data completely observed up to $t = 3$ instead of $t = 2$, then the imputed value is a simple average of observed $y_{i3}$'s, which yields biased estimates under (M2).

Appendix A shows that after substituting the sample mean into (2) sequentially, $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be rearranged as a sum of independent units, and $\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ equals the weighted estimating function, $\mathbf{U}_{\mathrm{W}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \hat{\gamma})$, defined by RRZ.

### 3.2 Modeling Missing Data

Sometimes there are few records with the same history, perhaps because the history includes continuous variables. If so, then calculating the sample mean using (4) or (5) becomes infeasible. We then model the conditional mean, $E(y_{it}|\mathbf{D}_{in_i}, r_{it} = 0)$, and impute a predicted value from the model. A sample mean is also a predicted value from saturated model with categorical variables. When fitting a saturated imputation model is not practical, the model can be trimmed by deleting less crucial variables, or a mildly informative prior can be imposed (see, e.g., Schafer 1994). I take the first approach and assume that the conditional mean can be consistently estimated by estimating a finite number of unknown parameters.

Let $\mu_{it}^* = E(y_{it}|\mathbf{D}_{in_i}), t = n_i + 1, \ldots, N$ and let $\mathbf{X}_{in_i}^*$ be a known function of $\mathbf{D}_{in_i}$. Note that the rank of $\mathbf{X}_{in_i}^*$ depends on the missing pattern of subject $i$, and each cell in Figure 1 must be modeled separately. Assume that the conditional mean at $t$ for subject $i$ with missing pattern $k$ is $\mu_{it}^* = h_{tk}^{-1}(\mathbf{X}_{ik}^* \gamma_{tk})$, where $h_{tk}(\cdot)$ is a link function and $\gamma_{tk}$ is a vector of $q_{tk}$ unknown parameters. Different cells are allowed to have different link functions. For missing cell $(t, k)$, let $\hat{\gamma}_{tk}$ be the root of

$$\sum_{i=1}^{K} \mathbf{w}_{ijk} = \sum_{i=1}^{K} \frac{\partial \mu_{it}^*}{\partial \gamma_{tk}} \operatorname{var}(y_{it}|\mathbf{D}_{ik})^{-1} \times I(n_i > k)(y_{it}^* - \mu_{it}^*) = 0, \quad (7)$$

where $y_{it}^* = y_{it}$ if $r_{it} = 1$ and $y_{it}^* = \tilde{y}_{it} = \tilde{E}(y_{it}|\mathbf{D}_{in_i})$ if $r_{it} = 0$. The indicator function $I(n_i > k)$ is needed because imputation is sequential, as described in Section 3.1. Fitted values from higher indices of missing patterns are used as outcome in the model fitting. However, fitted values are not used as covariates in the models, so mean imputation does not introduce bias.

Imputing fitted values in the estimating function leads to Equation (2), whose root, $\tilde{\beta}$, is the mean imputation estimator.

*Theorem 1.* If every $\tilde{y}_{it}$ obtained by solving Equation (7) converges to $E(y_{it}|\mathbf{D}_{in_i})$, then $K^{1/2}(\tilde{\beta} - \beta_0)$ is normally distributed as $K \to \infty$, with mean 0 and variance $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{\Gamma} = E\{[\partial\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})]/\partial\boldsymbol{\beta}^T\}$. Moreover, $\Lambda$ can be consistently estimated by

$$\sum \left\{ \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})^{-1}(\tilde{\mathbf{Y}}_i^c - \tilde{\mu}_i^c) + \boldsymbol{\Xi}(\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \hat{\gamma})\mathbf{v}_i(\hat{\gamma}) \right\}^{\otimes 2}, \quad (8)$$

where

$$\boldsymbol{\Xi}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lim_{K \to \infty} K^{-1} \sum_{i}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} (\mathbf{I} - \mathbf{R}_i)\mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\gamma})$$

and $\boldsymbol{\xi}_i(\boldsymbol{\gamma})$ and $\mathbf{v}_i(\boldsymbol{\gamma})$ are defined in Appendix B, which also provides a sketch of the proof.

For nonignorable missingness, condition (6) does not hold; thus the conditional mean of missing data cannot be consistently estimated without additional assumption. Under M3, we assume that the two quantities in (6) satisfy

$$h_{jt}\{E(y_{it}|\mathbf{D}_{it-j}, r_{t-j} = 1, r_{it-j+1} = 0)\} = \delta_{jt} + h_{jt}\{E(y_{it}|\mathbf{D}_{it-j}, r_{it-j+1} = 1)\}. \quad (9)$$

The imputed value under M3, say $\check{\mathbf{y}}_{it}$, is then obtained by transformation; for example, $\check{y}_{it} = \operatorname{logit}^{-1}\{\delta_{jt} + \operatorname{logit}(\tilde{y}_{it})\}$, because $\tilde{y}_{it}$, an estimate under MAR, is an estimate of $E(y_{it}|\mathbf{D}_{it-j}, r_{it-j+1} = 1)$ under M3.

## 4. MULTIPLE IMPUTATION

In mean imputation, missing $y_{it}$ are replaced with sample means or fitted values that depend on observed and imputed values. The observed or fitted values used to calculate imputed values are called *donors* (Rubin 1977). In multiple imputation, $y_{it}$ is replaced with a random sample from the donors.

This gives multiple completed datasets, each one of which can be analyzed as if it were the complete data. The inferences from the multiple imputed datasets are then combined into a single inference. Repeated analyses enable assessment of the variability due to imputation. Rubin and Schenker (1986) suggested various bootstrap-like sampling schemes for generating a set of plausible values without specifying the full distribution of data while reflecting the sampling error of the observed data. One such scheme is the approximate Bayesian bootstrap (ABB), in which one first samples with replacement from the nonmissing observations, then draws imputes from this bootstrap sample.

*Table 1. Estimated Logistic Model for the Probability of Observing MMSE at Each Time Point*

| | Year 2 | | Year 3 | | Year 4 | | Year 5 | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE | Estimate | SE |
| Intercept | 3.48 | 2.21 | 1.65 | 2.84 | 1.60 | 1.85 | −.880 | 2.51 |
| Age | −.030 | .025 | −.003 | .034 | −.017 | .022 | −.014 | .031 |
| Education | −.001 | .010 | .003 | .011 | −.002 | .009 | .015 | .013 |
| Barthel | .111 | .047 | .063 | .062 | .046 | .040 | .060 | .055 |
| MMSE | −.853 | .489 | −.335 | .649 | −.310 | .412 | −.793 | .551 |

Analysis of multiple imputation for GEE is analogous to my analysis of mean imputation for GEE. First, going back to the case in which all elements in history are categorical, under MAR, the first diagonal cells in Figure 1 are filled in with randomly drawn values from a bootstrap sample of the observed outcomes having the same history. For the second diagonal cells, imputes are randomly drawn from a bootstrap sample of the observed outcomes *and* imputed values from the first diagonal cells. This procedure is repeated up to the fifth diagonal cell. The ABB procedure for GEE is as follows: (a) randomly sample missing data from a bootstrap sample of donors; (b) after imputation, solve (2) as if the imputed data had been observed, and estimate $\hat{\boldsymbol{\beta}}_k$ for the $k$th replication. Steps (a) and (b) are repeated $m$ times.

The final estimate and its estimated variance are $\bar{\boldsymbol{\beta}}_m = m^{-1}\sum_{k=1}^{m}\hat{\boldsymbol{\beta}}_k$ and $\widehat{\text{var}}(\bar{\boldsymbol{\beta}}_m) = \bar{\mathbf{V}}_m + (1+m^{-1})\mathbf{B}_m$, where $\bar{\mathbf{V}}_m = m^{-1}\sum\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_k)$ is within-imputation variance and $\mathbf{B}_m$ is between-imputation variance. Theorem 2 in Appendix B shows that the estimators $\tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}_m$ obtained from the two imputation methods are equivalent as the number of replications for the multiple imputation method and the sample size approach infinity.

For nonignorable missingness, I suggest the sampling–importance resampling method (Rubin 1987). The first stage involves sampling with replacement with equal probability from donors. The sample is then resampled, with probability proportional to

$$\exp(\delta_{tk})^{y_{it}} = \left[\frac{E(y_{it}|\mathbf{D}_{ik}, r_{it}=0)\{1-E(y_{it}|\mathbf{D}_{ik}, r_{it}=1)\}}{E(y_{it}|\mathbf{D}_{ik}, r_{it}=1)\{1-E(y_{it}|\mathbf{D}_{ik}, r_{it}=0)\}}\right]^{y_{it}},$$

where $\delta_{tk}$ is defined in (9). Therefore, given history $\mathbf{D}_{ik}$, a missing outcome is more (less) likely to be 1 (0) than the observed values with odds ratio $\exp(\delta_{tk})$.

When data are imputed by fitting models for $E(y_{it}|\mathbf{D}_{in_i})$, the model is sequentially fit for the conditional mean of cell $(j,k)$, $\mu_{it}^*(\mathbf{X}_{ik}^*) = h_{tk}^{-1}(\mathbf{X}_{ik}^*\gamma_{tk})$, just as in mean imputation, then $q_{tk}$ independent standard normal random variables, $\varepsilon_{tk}$, are generated and $\gamma_{tk}^* = \hat{\gamma}_{tk} + \widehat{\text{var}}(\hat{\gamma}_{tk})^{1/2}\varepsilon_{tk}$ is calculated. Given $\gamma_{tk}^*$, a binary random variable is generated and replaced with $\mathbf{y}_{it}$.

## 5. ANALYSIS OF THE STROKE STUDY

Tatemichi et al. (1992) studied cognitive function among 241 elderly stroke patients 7–10 days after the stroke, 3 months after the stroke, and then yearly for 4 years. At each visit, the mini–mental state examination (MMSE), which measures cognitive function, was administered. The MMSE consists of 30 questions. A score of 24 or more correct answers is interpreted as normal cognitive function. As is common practice, MMSE is dichotomized as less than 24 or at least 24.

Ninety two patients were recruited during the first year, and 104 and 45 patients were recruited during the second and third years. During the 4-year follow-up period, 69 patients died: 26 during the first year of follow-up and 21, 10, and 12 in subsequent years. The numbers of observed patients were 241 (7–10 days), 237 (3 months), 183 (year 1), 158 (year 2), 98 (year 3), and 35 (year 4). All 241 patients received an initial evaluation.

Let $\mathbf{y}_{it} = 1$ if the $i$th individual's MMSE score is above 2 at time $t$, 0 and otherwise. It is assumed that if a patient is alive at time $t$, then

$$\text{logit}\{E(y_{it})\} = \beta_0 + \sum_{k=1}^{5}\beta_k I(t=k+1) + \beta_6\text{age}_i + \beta_7\text{edu}_i + \beta_8\text{str}_i,$$

where $\text{age}_i$ denotes age of the $i$th patient at the time of entry, $\text{edu}_i$ denotes years of education, and $\text{str}_i$ is a binary stroke severity indicator. Let $T_i$ be the number of scheduled visits that were scheduled before death. If patient did not die during the study, then $T_i = 6$. The complete outcome is defined as $Y_i = (y_{i1}, \ldots y_{iT_i})$.

In this study, 226 of 241 patients had a monotone missing pattern, and the rest had a nonmonotone missing pattern. For these, the data were monotonized by ignoring 41 observed outcomes from 23 patients after a first missed visit, as suggested by RRZ. The ignored data were treated as missing. This monotonization does not affect the inference if

$$\begin{aligned}&E(y_t|y_{t-k}, r_{t-k}=1, r_{t-k+1}=0, \ldots, r_{t-1}=0, r_t=1)\\&\quad = E(y_t|y_{t-k}, r_{t-k}=1, r_{t-k+1}=0, \ldots, r_{t-1}=0, r_t=0).\end{aligned}$$

I examined the probability of observation at time $t$ given that patients are alive at $t$, and observed at time $t-1$, by fitting logistic models using age, years of education, Barthel index, and indicator of having a normal cognitive function at time $t-1$ as covariates, and $r_{it}$ as outcome. Table 1 shows the results at 1–4 years after stroke. At 3 months after stroke, there were only three losses to follow-up, and the coefficient estimates did not converge. Patients who had better cognitive functions at previous time points were more likely to drop out at each visit. Other variables did not show any trend. Using the previous outcomes $y_{i1}, y_{i2}\ldots y_{it-1}$, as covariates did not improve the goodness of fit nor change the coefficient estimates.

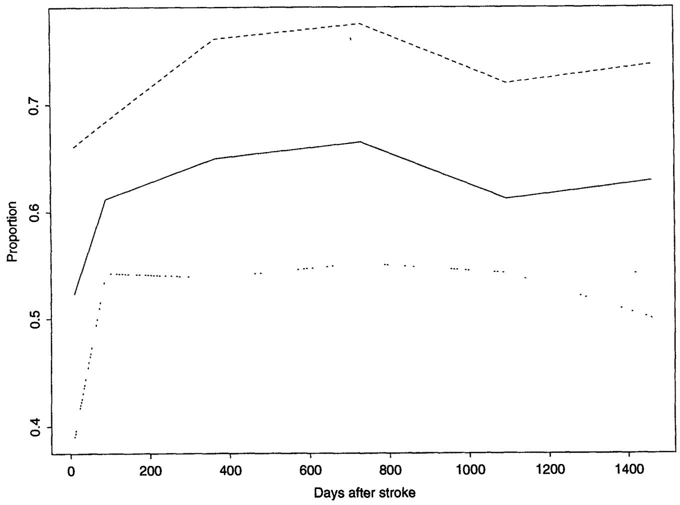


*Figure 2. Observed Proportion of Normal Neurological Function. · · ·, less than high school; – – –, more than high school; ——, all.*

Using the observed data only, with no imputation, the proportions of patients with normal cognitive function were .52, .61, .65, .66, .61, and .63 at 7–10 days, 3 months, and 1, 2, 3, and 4 years. Figure 2 shows the proportions over time for all subjects and stratified by education.

For missing outcomes of cell $(t,k)$, I fit

$$\operatorname{logit} E(y_{it}|\mathbf{D}_{ik}) = \gamma_{0tk} + \gamma_{1tk}\mathrm{age}_i + \gamma_{2tk}\mathrm{edu}_i + \gamma_{3tk}\mathrm{str}_i + \gamma_{4tk}y_{ik} \quad (10)$$

sequentially as described in Section 3 assuming MAR. Replacing $\gamma_{4tk}y_{ik}$ in (10) with $\sum_{l=1}^{k}\gamma_{(4+l)tk}y_{il}$ had little effect on the final estimates for $\boldsymbol\beta$. To assess sensitivity to the assumed model, I transformed $\tilde{y}_{it}$, the predicted value from (10), using $\operatorname{logit}(\tilde{y}_{it}) = \delta + \operatorname{logit}(\tilde{y}_{it})$; imputed $\tilde{y}_{it}$; and then estimated $\boldsymbol\beta$ by $\tilde{\boldsymbol\beta}(\delta)$ and $\bar{\boldsymbol\beta}_5(\delta)$, where $\delta = -\log 2, \delta = \log 2$. Note that $\tilde{\boldsymbol\beta}(0)$ and $\bar{\boldsymbol\beta}(0)$ are estimates under MAR. By setting $\delta = -\log 2$, I assume that missing outcomes are more likely to be 0 with odds ratio 2, even after conditioning on the covariates and previous outcome.

Table 2 shows estimates obtained from the fitted models. First, $\hat{\boldsymbol\beta}_{\mathrm{LZ}}$, estimates based on only the observed data and estimates after the monotonization are similar. The next rows show estimates $\hat{\boldsymbol\beta}_{\mathrm{W}}$, where the weights were calculated from the fitted models shown in Table 1. The following summarizes my findings:

1. All estimates are similar for all visits except for those at year 3 and year 4. The younger and better educated the patient and the less severe the stroke, the more likely the patient is to have normal cognitive function. These results did not change under the assumed nonignorable missing mechanisms. The odds ratio for normal cognitive function at 3 months after stroke relative to 7–10 days after stroke is about 1.4. It increases at years 1 and 2.
2. Imputation and weighting yielded similar point estimates, but the imputation estimates often had smaller standard errors.
3. Using observed data only, at years 3 and 4 the odds ratios of having normal cognitive function relative to 7–10 days are around 1.1. The same odds ratios are 1.2–1.3 under MAR ($\delta = 0$), .9–1 under nonignorable missingness ($\delta = -\log 2$), and about 1.6 ($\delta = \log 2$). Figure 3 shows the estimated time effects for $\hat{\boldsymbol\beta}_{\mathrm{LZ}}, \tilde{\boldsymbol\beta}(0), \tilde{\boldsymbol\beta}(-\log 2)$, and $\tilde{\boldsymbol\beta}(\log 2)$.

Estimates based on the observed data only are similar for age, stroke severity, and education, but estimates for the time effect may be slightly understated, because more patients with normal cognitive functions tended to drop out of the study.

## 6. SIMULATION RESULTS

I conducted simulation studies for bivariate binary data and for multivariate binary data under condition similar to

*Table 2. Estimated Coefficients and Standard Error for the Probability of Having Normal Cognitive Function, $\operatorname{logit}\mu_{it} = \beta_0 + \beta_1 age_i + \beta_2 str_i + \beta_3 edu_i + \Sigma_{j=1}^{5}\beta_{4+j}year_j$, Where $year_j$ Indicates Whether t is j Year(s) After Stroke*

| Estimate | Intercept | Age | Str. | Edu. | 3 months | Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|---|---|---|---|---|---|
| Obs | 3.72 | −.0693 | −1.23 | .174 | .450 | .572 | .540 | .265 | .0976 |
| | (1.09) | (.0155) | (.281) | (.0297) | (.147) | (.182) | (.195) | (.234) | (.365) |
| Mono | 3.89 | −.0715 | −1.24 | .173 | .451 | .581 | .521 | .177 | .114 |
| | (1.11) | (.0158) | (.281) | (.03) | (.148) | (.186) | (.209) | (.249) | (.412) |
| Robins | 4.68 | −.0811 | −1.13 | .160 | .448 | .639 | .607 | .361 | .677 |
| | (1.35) | (.0193) | (.362) | (.0311) | (.158) | (.228) | (.248) | (.267) | (.528) |
| Mean0 | 4.22 | −.0765 | −1.28 | .177 | .456 | .644 | .608 | .334 | .273 |
| | (1.39) | (.0197) | (.304) | (.0283) | (.153) | (.199) | (.249) | (.275) | (.313) |
| Mult0 | 4.52 | −.0786 | −1.24 | .160 | .448 | .590 | .651 | .226 | .261 |
| | (1.03) | (.0149) | (.256) | (.0289) | (.147) | (.194) | (.247) | (.241) | (.280) |
| Mean1 | 4.32 | −.0783 | −1.28 | .180 | .451 | .561 | .426 | −.0334 | −.148 |
| | (1.41) | (.0196) | (.631) | (.0459) | (.162) | (.199) | (.241) | (.298) | (.409) |
| Mult1 | 3.93 | −.0730 | −1.39 | .184 | .485 | .606 | .430 | −.0576 | −.0773 |
| | (1.01) | (.0142) | (.272) | (.030) | (.152) | (.179) | (.194) | (.207) | (.247) |
| Mean2 | 4.00 | −.0726 | −1.23 | .169 | .455 | .711 | .761 | .663 | .686 |
| | (1.48) | (.0205) | (.317) | (.0283) | (.151) | (.199) | (.258) | (.301) | (.419) |
| Mult2 | 4.22 | −.0762 | −1.24 | .174 | .451 | .677 | .813 | .561 | .661 |
| | (1.11) | (.0157) | (.266) | (.0273) | (.146) | (.176) | (.193) | (.238) | (.239) |

NOTE: Imputation model under MAR logit $\tilde{y}_{it} = \tilde{\gamma}_0 + \tilde{\gamma}_1 age_i + \tilde{\gamma}_2 str_i + \tilde{\gamma}_3 edu_i + \tilde{\gamma}_4 y_{it-1}$, and under nonignorable missingness was logit $\tilde{y}_{it} = \delta +$ logit $\tilde{y}_{it}$; str, binary stroke severity indicator; and edu, years of education. The following estimates are shown: obs, observed data only; mono, monotonization by deletion; Robins, Robins et al.'s estimates under MAR; mul0 multiple imputation estimate under MAR ($\delta = 0$); mul1 ($\delta = -log\ 2$); mul2 ($\delta = log\ 2$); mean0 mean imputation estimate under MAR ($\delta = 0$); mean1 ($\delta = -log\ 2$); mean2 ($\delta = log\ 2$).

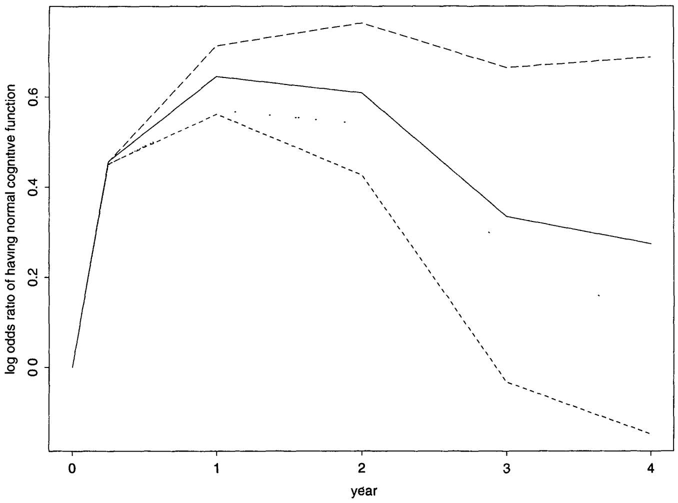


*Figure 3. Time Trends in Log Odds Ratio Under Different Imputation Models. - - -, miss more if normal; · · ·, observed; ——, MAR; — — —, miss less if normal.*

those of the stroke study. I generated bivariate binary data $(t = 1, 2)$ with marginal mean

$$E(y_{it}) = \mu_{it} = \exp\{\beta_0 + \beta_1 \mathbf{x}_i + \beta_2(t-1) + \beta_3 \mathbf{x}_i(t-1)\}$$

and $\Pr(y_{i1} = 1, y_{i2} = 1|\mathbf{x}_i) = p_{11i} = \rho\sqrt{\mu_{i1}(1-\mu_{i1})}\sqrt{\mu_{i2}(1-\mu_{i2})} + \mu_{i1}\mu_{i2}, p_{10i} = \mu_{i1} - p_{11i}, p_{01i} = \mu_{i2} - p_{11i}$, and $p_{00i} = 1 - \mu_{i1} - p_{01i}$, where $\rho = .25, .5, .75$. A monotone missing pattern was assumed, so that $r_{i1} = 1$ for all $i$. Two models of missingness were also considered:

- m1: logit $\Pr(r_{i2} = 1) = y_{i1}$ (MAR),
- m2: $\Pr(r_{i2} = 1) = \Phi(y_{i2} + y_{i1})$ (nonignorable missingness).

For all estimates, the working correlation matrix is $\boldsymbol{\Omega}_i = I_i$. There is one binary indicator $x_i$ with $E(x_i = 1) = .5$. Five estimates are compared under m1: $\hat{\boldsymbol{\beta}}_F$, the GEE estimate using all the data; $\hat{\boldsymbol{\beta}}_{\text{LZ}}$, the GEE estimate using observed data alone; $\tilde{\boldsymbol{\beta}}$, the mean imputation estimate; $\bar{\boldsymbol{\beta}}_5$, the multiple imputation estimate; and $\hat{\boldsymbol{\beta}}_{\text{W}}$, the weighted GEE estimate of Robins et al. (1995). Simulation results are based on 200 replications. The mean and simulation variance of the estimates of $\beta_2$, and the 95% coverage probability of the true parameter; are reported. Table 3 shows results under m1 when true values of $(\beta_0, \beta_1, \beta_2, \beta_3)$ are (a) (0, 0, 0, 0), (b) (0, 0, .5, 0), (c) (0, 1, 0, 0), and (d) (0, 1, .5, 0). Both models for weight and imputation were correctly specified. For the weight, estimated $\Pr(r_i = 1|y_{i1})^{-1}$ was used for all four sets of parameter values. For the imputation model, $\tilde{E}(y_{i2}|y_{i1})$ was used for (a) and (b), and $\tilde{E}(y_{i2}|y_{i1}, x_i)$ for (c) and (d). Note that the same stratification was used for weighting and imputation method in (a) and (b), and the finer stratification was used in (c) and (d) for imputa-

*Table 3. Generalized Estimating Equation Estimates for Bivariate Binary Data Where corr($y_{i1}$, $y_{i2}$) = $\rho$, Working Correlation Matrix is the Identity Matrix, and Logit $\mu_{it} = \beta_0 + \beta_1 x_i + \beta_2(t-1) + \beta_3 x_i(t-1)$*

| | $\rho$ = .25 | | | $\rho$ = .5 | | | $\rho$ = .75 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Var | 95% | Estimate | Var | 95% | Estimate | Var | 95% |
| | | | | $\boldsymbol{\beta}$ = (0, 0, 0, 0) | | | | | |
| $\hat{\beta}_F$ | −.184 | .0725 | .940 | .0130 | .0435 | .930 | .0100 | .0206 | .960 |
| $\hat{\beta}_{LZ}$ | .0825 | .0996 | .940 | .1860 | .0668 | .900 | .2770 | .0483 | .760 |
| $\tilde{\beta}$ | −.0114 | .0805* | .975 | −.0014 | .0438* | .990 | .0081 | .0243* | .995 |
| $\bar{\beta}_5$ | −.0061 | .0808 | .975 | .0027 | .0490 | .960 | .0045 | .0257 | .960 |
| $\hat{\beta}_W$ | −.0100 | .0976* | .925 | −.0027 | .0637* | .975 | .0038 | .0423* | .990 |
| | | | | $\boldsymbol{\beta}$ = (0, 0, .5, 0) | | | | | |
| $\hat{\beta}_F$ | .488 | .0673 | .935 | .475 | .0418 | .930 | .514 | .0198 | .960 |
| $\hat{\beta}_{LZ}$ | .583 | .0945 | .960 | .691 | .0665 | .925 | .819 | .0558 | .765 |
| $\tilde{\beta}$ | .492 | .0655* | .990 | .491 | .0428* | .995 | .515 | .0228* | .990 |
| $\bar{\beta}_5$ | .494 | .0736 | .970 | .492 | .0472 | .970 | .512 | .0238 | .965 |
| $\hat{\beta}_W$ | .489 | .0967* | .935 | .492 | .0699* | .975 | .517 | .0511* | .980 |
| | | | | $\boldsymbol{\beta}$ = (0, 1, 0, 0) | | | | | |
| $\hat{\beta}_F$ | −.0115 | .0642 | .945 | −.0063 | .0409 | .930 | −.0103 | .0327 | .970 |
| $\hat{\beta}_{LZ}$ | .1800 | .1000 | .925 | .1910 | .0670 | .890 | .2510 | .0844 | .860 |
| $\tilde{\beta}$ | −.0109 | .0960 | .975 | .0019 | .0591 | .925 | −.0142 | .0566* | .980 |
| $\bar{\beta}_5$ | −.0154 | .1040 | .970 | −.0070 | .0613 | .935 | −.0148 | .0629 | .955 |
| $\hat{\beta}_W$ | −.0083 | .0994 | .955 | −.0001 | .0634 | .965 | −.0199 | .0748* | .980 |
| | | | | $\boldsymbol{\beta}$ = (0, 1, .5, 0) | | | | | |
| $\hat{\beta}_F$ | .493 | .0557 | .940 | .501 | .0411 | .965 | .505 | .0203 | .945 |
| $\hat{\beta}_{LZ}$ | .595 | .0890 | .950 | .695 | .0759 | .880 | .828 | .0560 | .725 |
| $\tilde{\beta}$ | .489 | .0888 | .935 | .501 | .0719* | .940 | .515 | .0434* | .940 |
| $\bar{\beta}_5$ | .502 | .0998 | .930 | .498 | .0748 | .930 | .521 | .0520 | .920 |
| $\hat{\beta}_W$ | .491 | .0870 | .950 | .501 | .0773* | .970 | .513 | .0533* | .990 |

NOTE: Based on 200 simulation, mean simulation variance and coverage probability for four estimates of $\beta_2$ are given: $\hat{\beta}_F$, estimates based on full data; $\hat{\beta}_{LZ}$, GEE estimates by Liang and Zeger (1986); $\tilde{\beta}$, mean imputation estimates; $\bar{\beta}_5$ multiple imputation estimates; $\hat{\beta}_W$, weighting estimates by Robins et al. (1995). The missing mechanism is logit $\Pr(r_{i2} = 1) = y_{i1}$.
* Mean squared errors between $\tilde{\beta}$ and $\hat{\beta}_W$ are significantly different using sign test at $\alpha$ = .05.

*Table 4. Generalized Estimating Equation Estimates for Bivariate Binary Data Where corr($y_{i1}$, $y_2$) = $\rho$, Working Correlation Matrix is the Identity Matrix, and Logit $\mu_{it} = \beta_0 + \beta_1 x_i + \beta_2(t-1) + \beta_3 x_i(t-1)$, Where $x_i$ is Binary With $E(x_i) = .5$ or Continuous With N(0, 1)*

| | $\rho$ = .25 | | | $\rho$ = .5 | | | $\rho$ = .75 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Var | 95% | Estimate | Var | 95% | Estimate | Var | 95% |
| | | | | $x_i \sim$ Bernoulli(.5) | | | | | |
| $\hat{\beta}_F$ | −.0183 | .0975 | .970 | .0109 | .0563 | .970 | .0125 | .0385 | .975 |
| $\hat{\beta}_{LZ}$ | .3250 | .1320 | .855 | .4560 | .0888 | .735 | .5070 | .0884 | .500 |
| $\tilde{\beta}$ | .2480 | .1360 | .895 | .2580 | .0826 | .945 | .1540 | .0669 | .960 |
| $\bar{\beta}_5(0)$ | .2420 | .1420 | .890 | .2630 | .0937 | .905 | .1570 | .0705 | .900 |
| $\bar{\beta}_5(.25)$ | .1800 | .1310 | .915 | .2010 | .0789 | .925 | .1160 | .0640 | .865 |
| $\bar{\beta}_5(.5)$ | .1110 | .1290 | .910 | .1360 | .0763 | .950 | .0831 | .0592 | .905 |
| $\bar{\beta}_5(.75)$ | .0422 | .1210 | .940 | .0776 | .0737 | .965 | .0479 | .0558 | .895 |
| $\bar{\beta}_5(1)$ | −.0088 | .1140 | .950 | .0253 | .0669 | .965 | .0222 | .0525 | .910 |
| | | | | $x_i \sim$ N(0, 1) | | | | | |
| $\hat{\beta}_F$ | .0149 | .0375 | .940 | −.0110 | .0227 | .965 | −.0209 | .0142 | .935 |
| $\hat{\beta}_{LZ}$ | .3580 | .0536 | .620 | .4000 | .0350 | .485 | .4790 | .0236 | .120 |
| $\tilde{\beta}$ | .2860 | .0552 | .805 | .2110 | .0323 | .855 | .1210 | .0209 | .920 |
| $\bar{\beta}_5(0)$ | .2880 | .0569 | .765 | .2080 | .0330 | .835 | .1280 | .0208 | .890 |
| $\bar{\beta}_5(.25)$ | .1940 | .0637 | .795 | .1390 | .0358 | .840 | .0739 | .0202 | .910 |
| $\bar{\beta}_5(.5)$ | .0976 | .0630 | .850 | .0659 | .0356 | .920 | .0244 | .0197 | .935 |
| $\bar{\beta}_5(.75)$ | .0094 | .0608 | .900 | −.0075 | .0358 | .940 | −.0262 | .0204 | .940 |
| $\bar{\beta}_5(1)$ | −.0680 | .0576 | .910 | −.0770 | .0351 | .940 | −.0749 | .0210 | .905 |

NOTE: Based on 200 simulation, mean, simulation variance and coverage probability for eight estimates of $\beta_2$ are given: $\hat{\beta}_F$, estimates based on full data; $\hat{\beta}_{LZ}$, GEE estimates by Liang and Zeger (1986); $\tilde{\beta}$, mean imputation estimates; $\bar{\beta}_5(\delta)$ multiple imputation estimates, where $\delta$ = 0, .25, .5, .75, 1. The missing mechanism is $\Pr(r_{i2} = 1) = \Phi(y_{i1} + y_{i2})$.

tion than weighting. Although point estimates for weighting and imputation methods were nearly unbiased, the simulated variance of $\tilde{\beta}$ was significantly smaller ($\alpha = .05$, sign test) than that of $\hat{\beta}_W$ for nine out of 12 scenarios, and the efficiency increased more than 50% in certain case as the correlation increased. This is not unexpected. The higher the correlation, the more precisely missing data can be predicted in imputation, whereas such an advantage is not there in weighting. For all parameter values, $\hat{\beta}_{LZ}$ showed substantial bias that increased as the correlation between binary outcomes increased. This is because the higher correlation corresponds to the stronger dependence of missing-data indicators on outcomes. Simulation variances of the multiple imputation estimates were about 10% larger than the variances of the mean imputation estimates. This increase in variance agrees with the asymptotic relative efficiency calculation. Because about 50% of the data are missing and the number of multiple imputations is 5, the approximate relative efficiency of var($\bar{\beta}_5$) to var($\bar{\beta}_\infty$) [or, equivalently, var($\tilde{\beta}$)] is $50/5 = 10\%$. Confidence intervals based on the multiple-imputation estimates were calculated using the $t$ distribution with degrees of freedom given by Rubin (1987, p. 91). The approximation to the $t$ distribution seems satisfactory, because coverage probabilities were within the 95% confidence interval (.92, .98). Confidence intervals for mean imputation estimates tend to be conservative when the coefficient for $x$ is 0.

Table 4 summarizes simulation results under m2 for $x$ is binary and for $x$ randomly sampled from the standard normal distribution. In addition to $\hat{\beta}_F, \hat{\beta}_{LZ}, \tilde{\beta}$, and $\bar{\beta}_5$, four more estimates—$\bar{\beta}_5(\delta_k), \delta_1 = .25, \delta_2 = .5, \delta_3 = .75$, and $\delta_4 = 1$—were computed via multiple imputation. Under MAR, $\tilde{y}_{i2}$ was obtained by fitting logit $E(y_{i2}|x_i, y_{i1}) = \gamma_0 + \gamma_1 x_i + \gamma_2 y_{i1}$, and for $\bar{\beta}_5(\delta_k)$ imputes obtained by changing intercept by $\delta_k$ in the MAR imputation model were filled in. Note that the imputation models are misspecified. For both binary and continuous $x$, the bias of estimated time effect ($\beta_2$) using multiple imputation decreased as $\delta$ approached 1. The bias in the estimate of time and $x$ interaction ($\beta_3$), not reported, was relatively small under m2. When the missingness mechanism depends on both $x_i$ and $y_{i2}$, such as $\Phi(x_i + y_{i2})$, the estimates of $\beta_3$ showed substantial biases.

In Table 5, the multivariate binary distribution by Zhao and Prentice (1990) was used to generate responses. The marginal mean was logit $E(y_{it}|\mathbf{x}_{it}) = -.5 + \mathbf{x}_i + \sum_{j=2}^{6} I(t = j)\beta_j (i = 1, \ldots, 241, t = 1, \ldots, 6)$, where $\mathbf{x}_i$ is binary with $E(x_i) = .5$, and the time coefficients $\beta_2$–$\beta_6$ are the same as mean imputation estimates under MAR from the stroke study (as in Table 2). All pairwise correlations are .25. The missing mechanism is $\Pr(r_{it} = 1|r_{it-1} = 1) = 1 + (\log 2) y_{it-1}$. The correct missing mechanism was specified for the weighted method. For imputation, the imputation model was misspecified as logit $E(y_{it}|\mathbf{D}_{in_i}) = \mathbf{D}_{in_i}\gamma$: the correct model includes all two-way interaction terms between $\mathbf{x}_i$ and $(y_{i1}, y_{i2}, \ldots, y_{in_i})$. Table 5 shows that imputation estimates are nearly unbiased, even though the mean model is misspecified. Furthermore, the standard errors of mean imputation estimates of coefficients for $x$ and visit 6 were significantly smaller ($\alpha = .05$, two-tailed sign test) than those of the weighted estimates.

## 7. DISCUSSION AND CONCLUSION

Historically, treatments of missing data in the survey sampling literature have followed one of two paradigms (Little 1986). *Weighting methods* discard incomplete data and weigh observed data inversely to the probability of observation, and *imputation methods* impute missing values.

*Table 5. Simulation Results Based on 200 Replications Under the Setting Similar to the Stroke Example*

| | Intercept | x | Time 2 | Time 3 | Time 4 | Time 5 | Time 6 |
|---|---|---|---|---|---|---|---|
| | | | | *Bias* | | | |
| $\hat{\beta}_F$ | −.00132 | .00172 | .00156 | −.00149 | −.00413 | −.0193 | −.0123 |
| $\hat{\beta}_{LZ}$ | −.00839 | .0157 | .0387 | .0779 | .128 | .115 | .171 |
| $\tilde{\beta}$ | −.0113 | .0217 | .00186 | .00171 | .0108 | −.0316 | −.0168 |
| $\tilde{\beta}_m$ | −.00662 | .0122 | .00224 | .000546 | .00114 | −.0316 | −.0135 |
| $\hat{\beta}_W$ | −.00849 | .0159 | .00167 | .00165 | .0126 | −.0296 | −.014 |
| | | | | *Simulation variance* | | | |
| $\hat{\beta}_F$ | .0282 | .0269 | .0278 | .0264 | .0338 | .0242 | .0273 |
| $\hat{\beta}_{LZ}$ | .0328 | .0387 | .0324 | .0384 | .0581 | .0532 | .0670 |
| $\tilde{\beta}$ | .0329 | .0418* | .0326 | .0382 | .0590 | .0533 | .0515* |
| $\tilde{\beta}_m$ | .0332 | .0424 | .0337 | .0389 | .0602 | .0520 | .0537 |
| $\hat{\beta}_W$ | .0348 | .0463* | .0327 | .0383 | .0595 | .0555 | .0583* |
| | | | | *95% Coverage probability* | | | |
| $\hat{\beta}_F$ | .915 | .950 | .935 | .960 | .930 | .955 | .950 |
| $\hat{\beta}_{LZ}$ | .925 | .945 | .960 | .955 | .895 | .930 | .915 |
| $\tilde{\beta}$ | .910 | .935 | .965 | .965 | .935 | .980 | .985 |
| $\tilde{\beta}_m$ | .910 | .910 | .955 | .955 | .895 | .955 | .930 |
| $\hat{\beta}_W$ | .925 | .965 | .950 | .965 | .940 | .960 | .970 |
| | | | | *Power* | | | |
| $\hat{\beta}_F$ | .880 | 1.00 | .795 | .970 | .945 | .470 | .350 |
| $\hat{\beta}_{LZ}$ | .850 | 1.00 | .805 | .955 | .935 | .455 | .410 |
| $\tilde{\beta}$ | .860 | 1.00 | .685 | .880 | .760 | .225 | .080 |
| $\tilde{\beta}_m$ | .845 | 1.00 | .675 | .875 | .795 | .295 | .260 |
| $\hat{\beta}_W$ | .835 | 1.00 | .690 | .910 | .760 | .270 | .150 |

NOTE: Multivariate binary data are generated according to Zhao and Prentice (1990) with all pairwise correlation among $y$ as .25, logit $\mu_{it} = -5 + x_i + \sum_{j=2}^{6} I(t = j)\beta_j$, where $\beta_2$ through $\beta_6$ are time coefficient estimates of $\tilde{\beta}$ in Table 2. $\hat{\beta}_F$, estimates based on full data; $\hat{\beta}_{LZ}$, GEE estimates by Liang and Zeger (1986), $\tilde{\beta}$, mean imputation estimates; $\tilde{\beta}_m$ multiple imputation estimates; $\hat{\beta}_W$, weighting estimates by Robins et al. (1995). The missing mechanism is logit $\Pr(r_{i2} = 1) = 1 - \log 2y_{i1}$.

* Mean squared errors between $\tilde{\beta}$ and $\hat{\beta}_W$ are significantly different using sign test at $\alpha = .05$.

RRZ can be viewed as the weighting approach to the GEE, whereas my method can be seen as the imputation approach. The two methods yield different estimates in general. The validity of weighting depends on correct specification of the missingness mechanism, and the validity of imputation depends on correct specification of the conditional mean of the missing outcome given observed data. An attractive feature of weighting over imputation is that the weight, the inverse of the probability of observation, is not a function of mean parameters in GEE, whereas in imputation the conditional mean given previous outcomes is related to the marginal mean. As for the imputation method, it is not always obvious that a multivariate distribution exists that satisfies the assumptions on conditional and marginal means. This question of existence cannot be answered within the context of GEE, because it does not make full distributional assumptions. For that reason, I did not assume any joint likelihood function that dictates the imputation model, but instead directly modeled the conditional mean. Nevertheless, the existence of a multivariate distribution that embeds the imputation and marginal model would justify a set of assumptions made in the imputation method. Such multivariate distributions exist, however. One example is the partial exponential family (Zhao and Prentice 1990, 1992). Moreover, simulation results suggest that misspecification of the imputation yields fairly robust estimates. Standard errors of imputation estimates obtained under misspecified imputation model were significantly smaller for some of the coefficients than those of the weighting estimates in our simulation setting. A referee pointed out that the estimating equation of RRZ belongs to the class of estimating equations discussed by Robins and Rotnitzky (1995), and thus it could be adapted to attain semiparametric information bound in that class.

As shown in Appendix A, the estimators via mean imputation and RRZ are identical, when all elements in $\mathbf{D}_{it}$ are categorical variables and models for both the probability of missingness and the conditional mean of missing data are saturated. Similar results were given by Little (1986) in the analysis of variance (ANOVA) setting.

In conclusion, I have described imputation estimators for GEE when data are not MCAR. The main idea is to correct the bias of the estimating function by imputation. The imputation methods can provide consistent estimators under MAR if the model for the missing data is correctly specified. I considered multiple imputation estimators that are asymptotically equivalent to the mean imputation estimators as the number of replications approaches infinity. The multiple imputation method has advantages over the mean imputation method in that coefficients and their standard errors can be estimated more easily. Careful modeling of missing data is important to minimize bias caused by misspecification of imputation models.

## APPENDIX A: RELATIONSHIP BETWEEN IMPUTATION AND WEIGHTING

I rearrange $\tilde{U}(\beta, \alpha)$ by considering the contribution of each observation to the first through fifth diagonal cells defined in Figure 1. The contribution of $y_{jt}$ as a donor to the first diagonal cell $(t, t-1)$ is

$$\frac{\sum_i (1 - r_{it}) r_{it-1} I(\mathbf{D}_{it-1} = \mathbf{D}_{jt-1})}{\sum_i r_{it} I(\mathbf{D}_{it-1} = \mathbf{D}_{jt-1})}. \tag{A.1}$$

The numerator of (A.1) is the number of times that $y_{jt}$ serves as a donor, and the denominator is the number of fellow donors. Denoting

$$\lambda_{it} = \Pr(r_{it} | \mathbf{D}_{it-1}, r_{it-1} = 1),$$

(A.1) is $(1-\hat{\lambda}_{it})/\hat{\lambda}_{it}$. The $y_{jt}$ can contribute to the second diagonal cell $(t, t-2)$ in two ways, as observed data or as the imputed value of the first diagonal cell:

$$\frac{1-\hat{\lambda}_{it-1}}{\hat{\lambda}_{it-1}} + \frac{(1-\hat{\lambda}_{it})}{\hat{\lambda}_{it}} \frac{(1-\hat{\lambda}_{it-1})}{\hat{\lambda}_{it-1}} = \frac{1-\hat{\lambda}_{it-1}}{\hat{\lambda}_{it-1}\hat{\lambda}_{it}}.$$

Summing up the contribution of $\mathbf{y}_{it}$ as a donor to all diagonal cells,

$$\frac{1-\hat{\lambda}_{it}}{\hat{\lambda}_{it}} + \frac{1-\hat{\lambda}_{it-1}}{\hat{\lambda}_{it-1}\hat{\lambda}_{it}} + \frac{1-\hat{\lambda}_{it-2}}{\hat{\lambda}_{it-2}\hat{\lambda}_{it-1}\hat{\lambda}_{it}} + \cdots + \frac{1-\hat{\lambda}_{i1}}{\prod_{k=1}^{t}\hat{\lambda}_{ik}} = \frac{1-\prod_{k=1}^{t}\hat{\lambda}_{ik}}{\prod_{k=1}^{t}\hat{\lambda}_{ik}}.$$

Therefore,

$$\begin{aligned}\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_i^K \left\{ \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) + \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right\} \\ &= \sum_i^K \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{I}_i + \mathbf{W}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i),\end{aligned}$$

where the $t$th diagonal element of $\mathbf{W}_i$ is $(1 - \prod_{i=1}^{t} \hat{\lambda}_i)/(\prod_{i=1}^{t} \hat{\lambda}_i)$. Now the $t$th diagonal element of $\mathbf{I}_i + \mathbf{W}_i$ is $(\prod_{i=1}^{t} \hat{\lambda}_i)^{-1}$, and, therefore, $\mathbf{U}_w(\boldsymbol{\beta}, \boldsymbol{\alpha}, \hat{\gamma}) = \tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

## APPENDIX B: PROOF OF THEOREM 1

Replacing missing $y_{it}$ with $\tilde{y}_{it}$ yields the imputed estimating function

$$\begin{aligned}&\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \sum \frac{\partial \boldsymbol{\mu}_i^{cT}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\tilde{\mathbf{Y}}_i^c - \boldsymbol{\mu}_i^c) \\ &= \sum \left\{ \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right. \\ &\qquad \left. + \frac{\partial \boldsymbol{\mu}_i^{cT}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{I} - \mathbf{R}_i)(\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_i) \right\} \\ &\quad + \sum \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{I} - \mathbf{R}_i)(\tilde{\mathbf{Y}}_i^c - \boldsymbol{\mu}_i^*),\end{aligned} \tag{B.1}$$

where $\boldsymbol{\mu}_i^* = (\mathbf{0}_{1\times n_i}, \mu_{in_i+1}^*, \ldots, \mu_{iN}^*)$, a vector of conditional mean for missing outcome. Let $\tilde{\mathbf{Y}}_i^c = \mathbf{a}_i(\hat{\gamma})$, where $\gamma^T = (\gamma_{21}^T, \gamma_{31}^T, \ldots, \gamma_{32}^T, \ldots \gamma_{N(N-1)}^T)$ with order $\sum_{t=2}^{N} \sum_{k=1}^{t-1} q_{tk} = q_{..}$. Note that $(\tilde{\mathbf{Y}}_i^c - \boldsymbol{\mu}_i^*)$ can be approximated by $\mathbf{a}_i(\hat{\gamma}) - \mathbf{a}_i(\gamma) \approx \{[\partial \mathbf{a}_i(\gamma)]/\partial \gamma^T\}(\hat{\gamma} - \gamma)$. Note that the elements of $\hat{\gamma}$ corresponding to the second diagonal or higher are conditioned on the estimates of lower diagonal cells. For instance, $\hat{\gamma}_{42}$ is a function of $\hat{\gamma}_{43}$, say $\hat{\gamma}_{42}(\hat{\gamma}_{43})$. Therefore,

$$\begin{aligned}&\hat{\gamma}_{42}(\hat{\gamma}_{43}) - \gamma_{42} \\ &\quad = \{\hat{\gamma}_{42}(\hat{\gamma}_{43}) - \gamma_{42}(\hat{\gamma}_{43})\} + \{\gamma_{42}(\hat{\gamma}_{43}) - \gamma_{42}\} \\ &\quad = \psi_{42} \sum_{i=1}^{K} \mathbf{w}_{i42} + \frac{\partial \gamma_{42}(\gamma_{43})}{\partial \gamma_{43}^T} \psi_{43} \sum_{i=1}^{K} \mathbf{w}_{i43} \\ &\quad = \sum_{i=1}^{K} \left\{ \psi_{42} \mathbf{w}_{i42} + \frac{\partial \gamma_{42}(\gamma_{43})}{\partial \gamma_{43^T}} \psi_{43} \mathbf{w}_{i43} \right\},\end{aligned} \tag{B.2}$$

where $\mathbf{w}_{ijk}$ is defined in (7) and $\boldsymbol{\Psi}_{jk} = E[\sum(\partial \mathbf{w}_{ijk}/\partial \gamma_{jk}^T)]$. The first part of (B.2), $\psi_{42}\mathbf{w}_{i42}$, represents a contribution as observed data; the second part, $\{[\partial \gamma_{42}(\gamma_{43})]/\partial \gamma_{43}^T\}\ \psi_{43}\mathbf{w}_{i43}$, represents an indirect contribution via $\hat{\gamma}_{43}$. All other elements can also be expressed as a sum of independent terms. Let $(\hat{\gamma} - \gamma) = \sum_{i=1}^{K} \mathbf{v}_i$. Subject $i$, who belongs to the missing pattern 1, contributes only as observed data; therefore, $\mathbf{v}_i = \boldsymbol{\Psi}(\gamma)\mathbf{w}_i$, where $\boldsymbol{\Psi}(\gamma) = E\sum\{[\partial \mathbf{w}_i(\gamma)]/\partial \gamma^T\}$, a $q_{..} \times q_{..}$ block diagonal matrix, and $\mathbf{w}_i = (\mathbf{w}_{i21}, \mathbf{w}_{i31}, \ldots \mathbf{w}_{i65})^T$, a vector of order $q_{..}$.

Then the third term of (B.1) can be reexpressed as

$$\sum_i \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{I} - \mathbf{R}_i) \boldsymbol{\xi}_i(\gamma) \sum_j \mathbf{v}_j(\gamma) + O_p(1),$$

where $\boldsymbol{\xi}_i(\gamma) = E\{[\partial \mathbf{a}_i(\gamma)]/\partial \gamma^T\}$, a $N \times q_{..}$ matrix. Note that if $n_i$ measurements are observed, then the first $n_i$ row(s) of $\boldsymbol{\xi}$ are 0, and if $n_i = N$, then $\boldsymbol{\xi} = 0$. Also, blocks of the upper right submatrices of $\boldsymbol{\xi}$ are 0, because $\tilde{y}_{it}$ does not involve estimates of $\gamma$ for lower diagonal cells. By exchanging the order of summation, the third term of (B.1) can be rewritten as

$$\sum_j \left\{ \sum_i \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} (\mathbf{I} - \mathbf{R}_i) \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \boldsymbol{\xi}_i(\gamma) \right\} \mathbf{v}_j(\gamma).$$

Let

$$\boldsymbol{\Xi}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \lim_{K\to\infty} K^{-1} \sum_i^K \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} (\mathbf{I} - \mathbf{R}_i) \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \boldsymbol{\xi}_i(\gamma).$$

Then (B.1) can be reexpressed as

$$\begin{aligned}&\sum \left\{ \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right. \\ &\qquad \left. + \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^c(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{I} - \mathbf{R}_i)(\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_i^c) = \boldsymbol{\Xi}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) \mathbf{v}_i(\gamma) \right\} \\ &\quad + O_p(K^{-1/2}).\end{aligned} \tag{B.3}$$

Because (B.3) is a sum of independent random vectors with mean 0 and finite variance, $\tilde{\mathbf{U}}(\boldsymbol{\beta})$ is multivariate normal with mean 0, and the variance can be consistently estimated by (8).

*Theorem 2.* $K^{1/2}(\bar{\boldsymbol{\beta}}_m - \tilde{\boldsymbol{\beta}}) = O_p(m^{-1/2}) + O_p(K^{-1/2})$.

*Proof.* Let $\mathbf{H}_{jik}$ be a $n_i \times n_i$ diagonal matrix with diagonal elements as indicators whether the $i$th subject is used as a donor for the $j$th subject at the $k$th imputation at each time point.

Then the estimating function for $\hat{\boldsymbol{\beta}}_k$ via ABB is

$$\mathbf{L}_k = K^{-1/2} \sum_i \left\{ \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i) + \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \sum_j \mathbf{H}_{jik} \mathbf{W}_{ji}^*(\mathbf{Y}_i - \boldsymbol{\mu}_i) \right\}$$

$$= K^{-1/2} \sum_i \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \left( \mathbf{I}_i + \sum_j \mathbf{H}_{jik} \mathbf{W}_{ji}^* \right) \times (\mathbf{Y}_i - \boldsymbol{\mu}_i),$$

where $\mathbf{W}_{ji}^*$, an eligibility matrix, is an $n_i \times n_i$ diagonal matrix whose $t$th diagonal element indicates whether $y_{it}$ can serve as a donor for the $j$th subject. The $t$th diagonal element of $\mathbf{W}_{ji}^*$ is $(1 - r_{jt})I(t > n_j, \mathbf{D}_{il} = \mathbf{D}_{jl})$. Let $\mathbf{Q} = [E(\partial \mathbf{L}_k / \partial \boldsymbol{\beta}^T)]^{-1}$, where the expectation is taken over $\mathbf{Y}$ and sampling indicators $\mathbf{H}$. Then

$$K^{1/2}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0) = \mathbf{Q}\mathbf{L}_k + O_p(K^{-1/2}).$$

The $t$th element of $E(\mathbf{H}_{jik}\mathbf{W}_{ji}^* | \mathbf{Y}, \mathbf{X}, \mathbf{R})$ is the probability of selecting $y_{it}$ to replace missing $y_{jt}$, and $E(\sum_j \mathbf{H}_{jik}\mathbf{W}_{ji}^* | \mathbf{Y}, \mathbf{X}, \mathbf{R})$ is the expected number of contribution. By sorting out contributions to the first through fifth diagonal cells as in Appendix A, it is clear that $E(\sum_j \mathbf{H}_{jik}\mathbf{W}_{ji}^* | \mathbf{Y}, \mathbf{X}, \mathbf{R}) = \mathbf{W}_i$, where $\mathbf{W}_i$ is defined in Appendix A. It immediately follows that $E(\mathbf{L}_k | \mathbf{Y}, \mathbf{X}, \mathbf{R})$, say $\mathbf{L}$, is equal to $K^{-1/2}\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and $\mathbf{Q} = K\boldsymbol{\Gamma}^{-1}$. Then

$$\begin{aligned}
&K^{1/2}(\bar{\boldsymbol{\beta}}_m - \tilde{\boldsymbol{\beta}}) \\
&= \left\{ m^{-1} \sum \mathbf{Q}\mathbf{L}_i - K^{1/2}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right\} + O_p(K^{-1/2}) \\
&= [\mathbf{Q}\mathbf{L} + m^{-1} \sum \{\mathbf{Q}(\mathbf{L}_i - \mathbf{L})\} - K^{1/2}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha})] \\
&\quad + O_p(K^{-1/2}) \\
&= \mathbf{Q} \left\{ m^{-1} \sum (\mathbf{L}_i - \mathbf{L}) \right\} + O_p(K^{-1/2}) \\
&= O_p(m^{-1/2}) + O_p(K^{-1/2}).
\end{aligned}$$

*[Received June 1995. Revised May 1997.]*

## REFERENCES


Laird, N. (1988), "Missing Data in Longitudinal Studies," *Statistics in Medicine*, 7, 305–315.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Little, R. A. (1986), "Survey Nonresponse Adjustments For Estimates of Means," *International Statistical Review*, 54, 139–158.

Little, R., and Rubin, D. (1987), *Statistical Analysis With Missing Data*. Wiley.

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

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.

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1987), "Multiple Imputation for Survey Nonresponse," New York: Wiley.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

Schafer, J. (1994), Comment on "Multiple-Imputation Inferences With Uncongenial Sources of Input," by X.-L. Meng, *Statistical Science*, 9, 560–561.

Tatemichi, T. K., Paik, M., Bagiella, E., Desmond, D. W., Pirro, M., and Hanzawa, L. K. (1992), "Dementia After Stroke is a Predictor of Long-Term Survival," *Stroke*, 25, 1915–1919.

Xie, F., and Paik, M. C. (in press), "Generalized Estimating Equation Model for Binary Outcomes With Missing Covariates," *Biometrics*, 53.

Zhao, L. P., and Prentice R. L. (1990), "Correlated Binary Regression Using a Quadratic Exponential Model," *Biometrika*, 77, 642–648.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992), "Multivariate Mean Parameter Estimation by Using a Partly Exponential Model," *Journal of Royal Statistical Society*, Ser. B, 54, 805–811.