NIH Public Access
Author Manuscript
J R Stat Soc Ser A Stat Soc. Author manuscript; available in PMC 2010 June 21.

Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2009 January ; 172(1): 3–20. doi:10.1111/j.1467-985X.2008.00564.x.

Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to AIDS data

STUART R. LIPSITZ,
Harvard Medical School, Boston, U.S.A.

GARRETT M. FITZMAURICE,
Harvard Medical School, Boston, U.S.A.

JOSEPH G. IBRAHIM,
University of North Carolina, Chapel Hill, U.S.A.

DEBAJYOTI SINHA,
The Florida State University, Talahassee, FL, U.S.A.

MICHAEL PARZEN, and
Emory University, Atlanta, GA, U.S.A.

STEVEN LIPSHULTZ
University of Miami School of Medicine, Miami, FLA, U.S.A

SUMMARY

In a large, prospective longitudinal study designed to monitor cardiac abnormalities in children born to HIV-infected women, instead of a single outcome variable, there are multiple binary outcomes (e.g., abnormal heart rate, abnormal blood pressure, abnormal heart wall thickness) considered as joint measures of heart function over time. In the presence of missing responses at some time points, longitudinal marginal models for these multiple outcomes can be estimated using generalized estimating equations (GEE) (Liang and Zeger, 1986), and consistent estimates can be obtained under the assumption of a missing completely at random (MCAR) mechanism. When the missing data mechanism is missing at random (MAR), that is the probability of missing a particular outcome at a time-point depends on observed values of that outcome and the remaining outcomes at other time points, we propose joint estimation of the marginal models using a single modified GEE based on an EM-type algorithm. The proposed method is motivated by the longitudinal study of cardiac abnormalities in children born to HIV-infected women and analyses of these data are presented to illustrate the application of the method. Further, in an asymptotic study of bias, we show that under an MAR mechanism in which missingness depends on all observed outcome variables, our joint estimation via the modified GEE produces almost unbiased estimates, provided the correlation model has been correctly specified, whereas estimates from standard GEE can lead to substantial bias.

Keywords

EM-type algorithm; generalized estimating equations; missing at random; missing completely at random

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

1 Introduction

Longitudinal data are frequently collected in social science studies as well as In health studies such as AIDS, cardiovascular, and cancer clinical trials. Although most statistical methods focus on a single outcome of interest at each time point, in many longitudinal studies, multiple outcomes are measured at each time point. For example, in longitudinal studies of cardiac function, many binary measures of heart function are collected at each time point, and focusing on just a single outcome over time, say abnormal blood pressure, may provide an incomplete picture of cardiac function. This is particularly true for the Pediatric Pulmonary and Cardiac Complications (P^2C^2) of Vertically Transmitted HIV Infection Study (Lipshultz et al., 1998), which was a large, prospective longitudinal study designed to monitor heart disease and the progression of cardiac abnormalities in children born to HIV-infected women. Previous results (Lipshultz et al., 1998; Lipshultz et al., 2000; Lipshultz et al., 2002) from the P^2C^2 study have shown that subclinical cardiac abnormalities develop early in children born to HIV-infected women, and that they are frequent, persistent, and often progressive. Cardiac abnormalities include cardiomyopathy (decreased left ventricular (LV) contractility) and reduced pumping ability of the heart (low LV fractional shortening). In the P^2C^2 study, cardiovascular function was measured approximately every year, including at birth, for up to six years, in a birth cohort of 393 infants born to women infected with HIV-1; this yielded up to 7 measurements on each child. The 393 children in this study (Lipshultz et al., 1998) were enrolled between May 1990 and April 1993. To better understand longitudinal change in heart function, multiple dichotomous measures of heart function (low LV fractional shortening, decreased LV contractility, abnormal heart rate, and abnormal blood pressure) must be jointly modeled over time.

Thus, the P^2C^2 data can be considered to be multivariate in two aspects: more than one outcome variable at any time-point, and multiple time points. In this paper, we focus on marginal regression models for multivariate longitudinal binary data, where the marginal probability of an abnormal outcome over time is related to a set of covariates. Here, we are primarily interested in estimating the marginal regression parameters for each outcome. We treat the association among multiple measures, and across time, as a nuisance characteristic of the data, but propose a parsimonious model for the association structure. In the marginal models, we assume that each outcome has a different set of marginal regression parameter. For example, the marginal regression parameters for abnormal blood pressure and abnormal heart rate over time are distinct.

Although most studies are designed to collect complete data on all participants, missing data very commonly arise and must be properly accounted for in the analysis. For example, in the P^2C^2 study, each patient was supposed to have an echocardiogram every year for the first 6 years of life, including at birth. However, a feature of this study which complicates the analysis is missing outcome data; for example, only 1 (0.25 %) of the 393 patients have outcomes measured at all 7 occasions. All four outcomes (fractional shortening, contractility, heart rate, and blood pressure) were either measured or not measured at each point in time; thus, we do not have missingness within time points, either the whole set of outcomes is observed at any time point or is missing. Table 1 gives the frequency distribution of the number of echocardiograms per individual, and Table 2 shows the number of subjects seen at each of the 7 possible occasions. As we see from Table 1, only 21 % of the subjects were seen more than three times. In Table 2, we see that 262 of the 393 children (66.7 %) had baseline measurements; after birth, the percentage of children with measurements of the outcomes slowly drops until only 7 (1.8 %) of the 393 subjects have the measures at 6 years of age. Most of the missing data are due to patients who “drop-out”, i.e., once the patient misses a scheduled visit, no more measurements of the outcome variables are obtained thereafter. However, there are 29 (3.9 %) patients who missed at least one measurement occasion, but returned at a later measurement

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

occasion. Furthermore, as we will see in Section 4, patients with HIV, abnormal blood pressure, abnormal heart rate, and abnormal fractional shortening are more likely to be seen at later measurement occasions. This implies that missingness cannot be assumed to be a completely random process. None of the 393 children died, so that we do not need to jointly model survival time along with the repeated measures data, as might be the case if death was related to the values of the repeated measures.

To estimate the regression parameters of marginal models, Liang and Zeger (1986) proposed the "standard" generalized estimating equations (GEE) to obtain consistent parameter estimates. This approach does not require the complete specification of the joint distribution of the repeated responses, but only the first two moments. When some individuals' response vectors are only partially observed, the standard GEE approach circumvents the problem of missing data by simply basing inferences on the observed responses, with correlations estimated using "all-available-pairs". This approach yields consistent marginal regression parameter estimates provided that the responses are *missing completely at random* (MCAR) (Rubin, 1976; Laird, 1988). In particular, when the outcome data are MCAR, missingness depends only on the covariates (that are included in the model), and the standard GEE provides consistent regression parameter estimates. However, when missingness is related to the observed data (covariates and observed responses), but conditionally independent of the missing responses given the observed data, the missing data are said to be *missing at random* (MAR) (Rubin, 1976; Laird, 1988) and standard GEE can yield biased regression parameter estimates. In this paper we consider a modification of GEE that yields regression parameter estimates with considerably less bias than the standard GEE when data are MAR and the "working correlation" structure is the true correlation structure. The proposed modification uses the EM-type algorithm proposed by Lipsitz et al (2000) for estimation of the correlation parameters. Lipsitz et al., (2000) showed that with MAR missing data, their modified GEE was practically unbiased for the marginal model for a single outcome repeatedly measured over time, whereas the "standard" generalized estimating equations can be heavily biased. Although the association structure is usually treated as a nuisance in the GEE approach, the association structure must be correctly specified in the modified GEE in order to minimize the bias in estimating the marginal regression parameters.

Assuming the longitudinal model for each outcome variable has a separate set of marginal regression parameters, one can estimate the marginal models using separate generalized estimating equations for each outcome. These estimates will be consistent under MCAR. However, just as in the case of a single outcome measured repeatedly over time, when data are missing at random (MAR), these estimates are potentially biased. The modified GEE of Lipsitz et al. (2000) applied separately to each outcome will, in general, reduce the bias of the standard GEE, but bias will remain if missingness depends on all of the observed data (e.g., if missingness depends on observed outcomes other than the one being estimated via the separate GEE's). When data are MAR, it is likely that missingness can depend on all of the observed outcome data, and estimating the marginal models for each outcome separately, even using the modified GEE, can still produce biased results. In this paper, we propose joint estimation of the marginal models for all outcomes using a single modified GEE; in this modified GEE, one must also specify the association parameters among the different outcomes (e.g., between heart rate and blood pressure). We compare the proposed method with the standard GEE approach of Liang and Zeger (1986).

In Section 4, using the binary measures of cardiac abnormalities in children born to HIV-infected women, we show that that discernably different regression parameter estimates are obtained when using the various GEE approaches. In this example, we also describe a logistic regression procedure for exploring whether the data are MCAR versus MAR. Using the results of this logistic regression procedure, we discuss whether MAR is a plausible assumption for

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the data on cardiac abnormalities in children born to HIV-infected. These analyses illustrate the potential for bias under different assumptions about missingness. To make more general recommendations to the applied investigator, as well as to complement the results of these data analyses, in Section 5 we conduct an asymptotic study of bias of the different GEE approaches. In this study of asymptotic bias, we show that if the missing data are MAR, and missingness depends on all observed outcomes, then joint estimation via the modified GEE produces almost unbiased estimates, assuming the correlation model has been correctly specified; in contrast, the standard GEE can yield highly biased estimates.

An alternative to GEE is estimation of the parameters via a full likelihood approach. To formulate a full likelihood under MAR, one must specify a joint model for the binary outcomes within each time point as well as across time points. Unfortunately, the full likelihood approach has many nuisance parameters, and it can be conceptually difficult to model higher-order associations in a flexible and interpretable manner that is consistent with the model for the marginal expectations (e.g., Bahadur, 1961). Full likelihood approaches are complicated algebraically since, given a marginal model for the vector of repeated outcomes, the multinomial probabilities cannot, in general, be expressed in closed-form as a function of the model parameters. Finally, maximum likelihood estimation can be computationally prohibitive, especially when the number of outcomes at each time and the number of times is large, since the number of multinomial probabilities grows exponentially with the number of repeated measures. For instance, in our example, there are 7 measurement occasions and 4 outcomes, meaning a full likelihood under MAR requires the specification of a multinomial distribution with $2^{7\cdot 4} = 268, 435, 456$ joint probabilities. As a result, ML estimation is feasible for only a relatively small number of repeated measures (say, less than 5). Also, unlike GEE, to obtain asymptotically unbiased estimates, the full joint distribution of the data must be correctly specified.

Thus, one of the chief attractions of our modified GEE approach over maximum likelihood is that it significantly eases the numerical complexities of the full likelihood approach by only requiring specification and estimation of pairwise association parameters. Further, it alleviates the need to specify and estimate many nuisance parameters that are needed in a full likelihood approach. In addition, with MAR missing data, approximately asymptotically unbiased estimators of the regression parameters can be obtained provided that the first two moments are correctly specified when using the modified GEE. Thus, when using the proposed modified GEE approach with MAR missing data, the key requirement is that the marginal model and the model for bivariate associations for all outcomes must be correctly specified and estimated simultaneously.

2 Notation and Distributional Assumptions

Suppose K binary random variables are collected at pre-specified time points t, $t = 1, ..., T$, as in the P^2C^2 study in which echocardiograms were to be taken every year from birth until six years of age. Let Y_{ikt} be the k^{th} binary random variable ($k = 1, ..., K$) collected on subject i ($i = 1, ..., n$) at time t. Then, for the k^{th} outcome variable from the i^{th} individual measured at T times, we can form a ($T \times 1$) response vector, $\mathbf{Y}_{ik} = [Y_{ik1}, ..., Y_{ikT}]'$, (for $k = 1, ..., K$). In principal, each of the K outcome variables can have its own set of covariates. However, for simplicity, we assume each outcome has the same set of covariates and these covariates are fully observed. We denote the $J \times 1$ covariate vector for subject i as $\mathbf{x}_i$. The main interest here is in the marginal model for each binary outcome Y_{ikt}, which we assume follows a logistic regression. The marginal distribution of Y_{ikt} is Bernoulli with success probability,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$p_{ikt}=p_{ikt}(\beta_k)=E(Y_{ikt}|\mathbf{x}_i,\beta_k)=pr(Y_{it}=1|\mathbf{x}_i,\beta_k)=\frac{\exp\left(\mathbf{x}_i'\beta_k\right)}{1+\exp\left(\mathbf{x}_i'\beta_k\right)}. \quad (1)$$

Even though we assume the covariate vector $\mathbf{x}_i$ is the same for all outcome variables, we assume the regression parameter vector $\boldsymbol{\beta}_k$ is distinct across the K outcomes. For outcome k, the p_{ikt}'s can be grouped together to form a $(T \times 1)$ vector $\mathbf{p}_{ik}$ containing the marginal probabilities of success over time, $\mathbf{p}_{ik} = E[\mathbf{Y}_{ik}|\mathbf{x}_i, \boldsymbol{\beta}_k] = [p_{ik1}, ..., p_{ikT}]'$. Further, the K vectors $\{\mathbf{Y}_{ik}\}$ and $\{\mathbf{p}_{ik}\}$ can be grouped into overall $TK \times 1$ vectors $\mathbf{Y}_i'=\left[\mathbf{Y}_{i1}',\ldots,\mathbf{Y}_{iK}'\right]'$ and $\mathbf{p}_i'=\left[\mathbf{p}_{i1}',\ldots,\mathbf{p}_{iK}'\right]'$. Note that we are primarily interested in making inference about $\boldsymbol{\beta}_k$. In this paper we are interested in the case where individuals are not observed at all T times; however, we assume that no covariates are missing.

The association between a pair of binary outcomes is typically measured in terms of marginal odds ratios (Plackett, 1965) or marginal correlations (Bahadur, 1961). For ease of exposition, here, we discuss marginal correlations, which are a function of the unknown parameter vector. We propose an autoregressive type correlation structure that is an extension of the correlation model proposed by (Galecki, 1994). In general, for outcomes (j, k) and times (s, t), the correlation model is

$$\rho_{i,js,kt}=Corr\left(Y_{ijs},Y_{ikt}|\mathbf{x}_i\right)=\alpha_{jk}^{|t-s|}\alpha_{2jk}^{I(j\neq k)}, \quad (2)$$

where $-1 < \alpha_{jk} < 1$, $-1 < \alpha_{2jk} < 1$, and $I(\cdot)$ is an indicator variable. In particular, for the same outcome variable $(k = j)$ at two different points in time $s \neq t$, the model is first-order autoregressive,

$$\rho_{i,ks,kt}=Corr(Y_{iks},Y_{ikt}|\mathbf{x}_i)=\alpha_{kk}^{|t-s|}.$$

For different outcomes $(j \neq k)$ at the same point in time $(s = t)$, the model is

$$\rho_{i,jt,kt}=Corr\left(Y_{ijt},Y_{ikt}|\mathbf{x}_i\right)=\alpha_{2jk}, \quad (3)$$

and for different outcomes $(j \neq k)$ at different points in time $(s \neq t)$, the model is

$$\rho_{i,js,kt}=Corr\left(Y_{ijs},Y_{ikt}|\mathbf{x}_i\right)=\alpha_{jk}^{|t-s|}\alpha_{2jk}. \quad (4)$$

Note that as $|t - s| \to 0, \alpha_{jk}^{|t-s|} \to 1$ and $\rho_{i,js,kt} \to \alpha_{jk}$, so that (4) agrees with (3). For the correlation structure given by (2), there are $K(K + 1)/2$ α_{kk}'s and $K(K - 1)/2$ α_{jk}'s.

Note, in general, the joint distribution of Y_{ijs} and Y_{ikt} is bivariate binary (Bahadur, 1961),

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$f\left(y_{ijs},y_{ikt}|\mathbf{x}_i,\beta,\alpha\right)=p_{ijs}^{y_{ijs}}\left(1-p_{ijs}\right)^{(1-y_{ijs})}p_{ikt}^{y_{ikt}}(1-p_{ikt})^{(1-y_{ikt})}\left\{1+\rho_{i,js,kt}\frac{\left(Y_{ijs}-p_{ijs}\right)(Y_{ikt}-p_{ikt})}{\sqrt{p_{ijs}\left(1-p_{ijs}\right)p_{ikt}\left(1-p_{ikt}\right)}}\right\}. \quad (5)$$

From (5), the joint probability that $Y_{ijs} = 1$ and $Y_{ikt} = 1$ equals

$$p_{i,js,kt}=pr\left(Y_{ijs}=1,Y_{ikt}=1|\mathbf{x}_i,\beta,\alpha\right)=p_{ijs}p_{ikt}+\rho_{i,js,kt}\sqrt{p_{ijs}\left(1-p_{ijs}\right)p_{ikt}\left(1-p_{ikt}\right)}; \quad (6)$$

this result is used in the GEE approach discussed in the next section.

3 Generalized estimating equations

Lipsitz et al., (2000) showed that, for a single outcome variable measured repeatedly over time, a modified GEE which uses an EM-type algorithm was practically unbiased for the marginal model when the missing data are MAR, whereas the standard GEE of Liang and Zeger (1986) could be highly biased. In datasets such as ours with multivariate longitudinal data, i.e., multiple outcomes measured over time, one would typically estimate the regression parameters $\boldsymbol{\beta}_k$ by applying separate generalized estimating equations to each outcome. The modified GEE of Lipsitz et al. (2000) applied separately to each outcome will reduce the bias of the standard GEE, but will still lead to bias if missingness depends on all of the observed data (for example, if separate GEEs are used, there will be bias in the estimated marginal model for abnormal blood pressure if missingness depends on the previous value of fractional shortening). When data are MAR, it is likely that missingness can depend on all of the observed outcome data, and estimating the marginal models for each outcome separately, even using the modified GEE, can still produce biased results. In this section, we describe joint estimation of the marginal models for all outcomes using a single modified GEE with the $TK \times 1$ outcome vector $\mathbf{Y}_i$ containing all outcomes over times.

When there are no missing data, the generalized estimating equations (GEE) for $\boldsymbol{\beta}$ are given by

$$\mathbf{u}_1\left(\widehat{\beta}\right)=\sum_{i=1}^{N}\mathbf{u}_{1i}\left(\widehat{\beta}\right)=\sum_{i=1}^{N}\widehat{\mathbf{D}}_i'\widehat{\mathbf{V}}_i^{-1}\left[\boldsymbol{Y}_i-\mathbf{p}_i\left(\widehat{\beta}\right)\right]=0, \quad (7)$$

where $\mathbf{D}_i = \partial\mathbf{p}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\mathbf{V}_i = \mathbf{V}_i(\alpha,\boldsymbol{\beta})$ is the $TK \times TK$ "working" or approximate covariance matrix of $\mathbf{Y}_i$ (Liang and Zeger, 1986); $\boldsymbol{\beta}$ is the JK vector of regression parameters. Since Y_{ikt} is binary, the corresponding diagonal elements of $\mathbf{V}_i$ is $\mathrm{Var}(Y_{ikt}) = p_{ikt}(1 - p_{ikt})$, which is specified entirely by the marginal distributions (i.e., by $\boldsymbol{\beta}$). A general off-diagonal element of $\mathbf{V}_i$ is $\mathrm{Cov}(Y_{ijs},Y_{ikt}) = p_{i,js,kt} - p_{ijs}p_{ikt}$, where $p_{i,js,kt}$ is specified in equation (6).

When there are missing outcome data, we can write $\boldsymbol{Y}_i'=\left(\boldsymbol{Y}_{m,i}',\boldsymbol{Y}_{o,i}'\right)$, where $\mathbf{Y}_{o,i}$ is a $(C_i \times 1)$ vector containing the *observed* components of $\mathbf{Y}_i$, and $\mathbf{Y}_{m,i}$ is a $[(TK - C_i) \times 1]$ vector containing the *missing* components of $\mathbf{Y}_i$. If the missing data are MAR, a consistent estimate of $\boldsymbol{\beta}$ can be obtained by setting the conditional expectation of $\mathbf{u}_1(\boldsymbol{\beta})$ in (7), denoted $\mathbf{u}_1^*(\beta)$, to $\mathbf{0}$ and solving for $\widehat{\beta}$. Here, the conditional expectation is taken with respect to the conditional distribution of the missing data given the observed data. In particular,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathbf{u}_1^*(\beta) = \sum_{i=1}^{N} E\left[\mathbf{u}_{1i}(\beta) | \mathbf{Y}_{o,i}, \mathbf{x}_i\right] = \sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1}\left[E(\mathbf{Y}_i | \mathbf{Y}_{o,i}, \mathbf{x}_i) - \mathbf{p}_i\right]. \quad (8)$$

In (8), we have conditioned of the observed data (a partition of the full vector $\mathbf{Y}_i$). This conditional expectation $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$ is a function of the observed data $\mathbf{Y}_{o,i}$. If we then take the expection of $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$ with respect to $\mathbf{Y}_{o,i}$, we get $E_{yo,i}\{E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)\} = E(\mathbf{Y}_i|\mathbf{x}_i) = \mathbf{p}_i$. It then follows that

$$E\left[\mathbf{u}_1^*(\beta)\right] = \sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1}\left[E_{y_{o,i}}\{E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)\} - \mathbf{p}_i\right] = 0.$$

Heuristically, using method of moment ideas, since $E\left[\mathbf{u}_1^*(\beta)\right] = 0$, and we are solving $\mathbf{u}_1^*\left(\widehat{\beta}\right) = 0$ for $\widehat{\beta}$, $\widehat{\beta}$ is consistent.

Note, however, that the computation of the conditional expectation of $\mathbf{Y}_i$ given $(\mathbf{Y}_{o,i}, \mathbf{x}_i)$ requires the full specification of the distribution of $\mathbf{Y}_i$. With a vector of TK binary responses, there are 2^{TK} possible response sequences, and $\mathbf{Y}_i$ has a multinomial distribution with 2^{TK} joint cell probabilities. If we specify all 2^{TK} joint cell probabilities to calculate the conditional expectation of $\mathbf{Y}_i$ given $(\mathbf{Y}_{o,i}, \mathbf{x}_i)$, we might as well use maximum likelihood since the full likelihood will be specified; further, (8) with $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$ correctly specified would be identical to the part of the maximum likelihood score vector for estimating $\boldsymbol{\beta}$. The primary appeal of GEE lies in avoiding the full specification of this joint distribution of $\mathbf{Y}_i$. In particular, as opposed to maximum likelihood, our proposed GEE only requires specification of the first two moments. Therefore, we consider an approximation for $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$, based on the multivariate normal distribution, that avoids the full specification of the joint distribution of $\mathbf{Y}_i$. Thus, our motivation for using the multivariate normal approximation is to find as simple approximation as possible to the first two moments of the joint multinomial distribution of the data; in other settings, we have found this approximation works very well (Lipsitz et al., 2000). In particular, we pro- pose replacing $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$ in (8) by the corresponding expression for this conditional expectation when $\mathbf{Y}_i$ is assumed to have a multivariate normal distribution,

$$E\left[\begin{array}{c} \boldsymbol{Y}_{m,i} \\ \boldsymbol{Y}_{o,i} \end{array} \middle| \mathbf{Y}_{o,i}, \mathbf{x}_i\right] = \left[\begin{array}{c} \mathbf{p}_{m,i} + \mathbf{V}_{mo,i}\mathbf{V}_{o,i}^{-1}\left[\boldsymbol{Y}_{o,i} - \mathbf{p}_{o,i}\right] \\ \boldsymbol{Y}_{o,i} \end{array}\right] \quad (9)$$

where

$$\mathbf{p}_i' = E\left(\boldsymbol{Y}_{m,i}', \boldsymbol{Y}_{o,i}'\right) = \left(\mathbf{p}_{m,i}', \mathbf{p}_{o,i}'\right)$$

and

$$\mathbf{V}_i = Var\left[\begin{array}{c} \boldsymbol{Y}_{m,i} \\ \boldsymbol{Y}_{o,i} \end{array} \middle| \mathbf{x}_i\right] = \left[\begin{array}{cc} \mathbf{V}_{m,i} & \mathbf{V}_{mo,i} \\ \mathbf{V}_{mo,i}' & \mathbf{V}_{o,i} \end{array}\right],$$

i.e., $\mathbf{p}_{o,i}$ and $\mathbf{V}_{o,i}$ are the elements of $\mathbf{p}_i$ and $\mathbf{V}_i$ corresponding to the observed data $\mathbf{Y}_{o,i}$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Estimating $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$ based on the multivariate normal distribution, it can be shown that the estimating equations for $\boldsymbol{\beta}$ in (8) reduce to

$$\mathbf{u}_1^*(\widehat{\beta}) \approx \sum_{i=1}^{N} \widehat{\mathbf{D}}_{o,i}' \widehat{\mathbf{V}}_{o,i}^{-1} \left[\boldsymbol{Y}_{o,i} - \mathbf{p}_{o,i}(\widehat{\beta}) \right] = 0, \tag{10}$$

where $\mathbf{D}_{o,i} = \partial \mathbf{p}_{o,i}(\beta)/\partial \boldsymbol{\beta}$. Although not derived using the multivariate normal approximation for $E(\mathbf{Y}_i|\mathbf{Y}_{o,i}, \mathbf{x}_i)$, the "standard" generalized estimating equations for $\boldsymbol{\beta}$, as originally proposed by Liang and Zeger (1986) and Prentice (1988), are identical to (10). The difference between the "standard" generalized estimating equations of Liang and Zeger and our approach is in the estimating equations for $\boldsymbol{\alpha}$ (and thus $\mathbf{V}_{o,i}$). Different estimates of $\mathbf{V}_{o,i}$ produce different solutions to (10), so that, even though the form of the estimating equations for $\boldsymbol{\beta}$ are identical, the estimates of $\boldsymbol{\beta}$ will be different when $\mathbf{V}_{o,i}$ is estimated by different approaches. In our experience, we have found, when data are MAR, as long as $\mathbf{V}_i$ is specified correctly, and consistently estimated, we expect the estimate of $\boldsymbol{\beta}$ to have little bias. With MAR missing data, for any GEE, the solution $\widehat{\beta}$ is asymptotically normal with mean $\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ may not necessarily equal the true $\boldsymbol{\beta}$. Further, the covariance matrix is given by the so-called sandwich estimator proposed by Huber (1967),White (1982) and Royall (1986). In particular, the asymptotic covariance matrix of $\widehat{\beta}$ can be consistently estimated with

$$\left[\sum_{i=1}^{N} \widehat{\mathbf{D}}_{o,i}' \widehat{\mathbf{V}}_{o,i}^{-1} \widehat{\mathbf{D}}_{o,i} \right]^{-1} \left[\sum_{i=1}^{N} \widehat{\mathbf{D}}_{o,i}' \widehat{\mathbf{V}}_{o,i}^{-1} (\boldsymbol{Y}_{o,i} - \widehat{\mathbf{p}}_{o,i})(\boldsymbol{Y}_{o,i} - \widehat{\mathbf{p}}_{o,i})' \widehat{\mathbf{V}}_{o,i}^{-1} \widehat{\mathbf{D}}_{o,i} \right] \left[\sum_{i=1}^{N} \widehat{\mathbf{D}}_{o,i}' \widehat{\mathbf{V}}_{o,i}^{-1} \widehat{\mathbf{D}}_{o,i} \right]^{-1}. \tag{11}$$

In the usual case where $\mathbf{V}_i$, and specifically, is unknown, we must parameterize and estimate $\rho_{i,js,kt} = Corr(Y_{iks}, Y_{ikt}|\mathbf{x}_i)$. Prentice (1988) suggests a second set of estimating equations for $\boldsymbol{\alpha}$ by first forming the cross-products $Y_{ijs}Y_{ikt}$, which have expected value $p_{i,js,kt} = E(Y_{ijs}Y_{ikt}|\mathbf{x}_i, \boldsymbol{\alpha})$ (the joint probabilities in (6)). The estimating equations for α are then based on linear combinations of $TK(TK-1)/2$ pairs $[Y_{ijs}Y_{ikt} - p_{i,js,kt}]$, which have mean 0 when no data are missing. With missing data, we propose replacing $Y_{ijs}Y_{ikt}$ in the estimating equations for $\boldsymbol{\alpha}$ with $E[Y_{ijs}Y_{ikt}|\mathbf{Y}_{o,i}, \mathbf{x}_i]$, the conditional expectation of $Y_{ijs}Y_{ikt}$ given the observed data $\mathbf{Y}_{o,i}$, where this conditional expectation is again calculated as if the complete data $\mathbf{Y}_i$ is multivariate normal. In particular, the conditional expectation $Y_{ijs}Y_{ikt}$ given $\mathbf{Y}_{o,i}$ under multivariate normality is an off-diagonal element of

$$E\left[\boldsymbol{Y}_i \boldsymbol{Y}_i' | \boldsymbol{Y}_{o,i}, \mathbf{x}_i \right] = \begin{pmatrix} \mathbf{V}_{m,i} - \mathbf{V}_{mo,i}\mathbf{V}_{o,i}^{-1}\mathbf{V}_{mo,i}' & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \mathbf{p}_{m,i} + \mathbf{V}_{mo,i}\mathbf{V}_{o,i}^{-1}[\boldsymbol{Y}_{o,i} - \mathbf{p}_{o,i}] \\ \boldsymbol{Y}_{o,i} \end{pmatrix} \begin{pmatrix} \mathbf{p}_{m,i} + \mathbf{V}_{mo,i}\mathbf{V}_{o,i}^{-1}[\boldsymbol{Y}_{o,i} - \mathbf{p}_{o,i}] \\ \boldsymbol{Y}_{o,i} \end{pmatrix}'.$$

In contrast, Liang and Zeger's standard GEE approach is based on an "all-available-pairs" estimator. To estimate $Corr(Y_{is}, Y_{it})$, the "all-available-pairs" method uses all subjects who are observed at times s and t; thus a subject contributes all pairs of times at which she/he is observed. Since the number of subjects observed at the different pairs of times can be different, the sample size used to estimate the different pairwise correlation coefficients can also be different. It is well-known (Little and Rubin, 2002) that this method can lead to an estimate of $\mathbf{V}_{o,i}$ that is not positive definite, and very biased when data are MAR. In contrast, our proposed estimator of $\boldsymbol{\alpha}$ based on the multivariate normal conditional expectation $Y_{ijs}Y_{ikt}$ given $\mathbf{Y}_{o,i}$, lead to an estimate of $\boldsymbol{\alpha}$ and thus $\mathbf{V}_{o,i}$ that will be positive definite and have minimal bias (Fitzmaurice et al., 2001). In order to get approximately unbiased estimates using GEE, the mutltivariate normal approximation must be used for both the first, $E[Y_{ijs}|\mathbf{Y}_{o,i}, \mathbf{x}_i]$ and the second, E

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$[Y_{ijs}Y_{ikt}|\mathbf{Y}_{o,i}, \mathbf{x}_i]$, moments. In summary, the primary difference between the proposed "modified" GEE and the "standard" GEE is in the method of estimating $\mathbf{V}_{o,i}$. As will be demonstrated later, this will have a huge impact on the resulting bias in the estimate of $\boldsymbol{\beta}$.

In summary, the proposed modified GEE estimate is the solution to (10) with $\boldsymbol{\alpha}$ replaced by the solution to our second set of estimating equations after replacing $Y_{ijs}Y_{ikt}$ with its approximate conditional expectation given $\mathbf{Y}_{o,i}$ under multivariate normality. The modified GEE yields consistent estimates of $\boldsymbol{\beta}$ when data are MCAR, and based on the results of Lipsitz et al. (2000) for a single outcome measured repeatedly over time, we expect minimal bias under MAR when the covariance structure ($\mathbf{V}_i$) is correctly specified. Note, though, this requires that all correlations, both for each outcome over time, and across different outcomes (at a given time or at different times), be correctly specified. Also, the modified GEE applied separately to each outcome can be thought of as a special case of our proposed modified GEE, with "working correlation" of independence across different outcomes. If the data are MAR, but not MCAR, then $E(\mathbf{Y}_{o,i}|\mathbf{x}_i,\boldsymbol{\beta})$ may not equal $\mathbf{p}_{o,i}$, but the weighted linear combination, $\sum_{i=1}^{N}\mathbf{D}'_{o,i}\mathbf{V}_{o,i}^{-1}(\boldsymbol{Y}_{o,i}-\mathbf{p}_{o,i})$, with $\mathbf{V}_{o,i}$ estimated using the multivariate normal approximation for $E\{Y_{ijs}Y_{ikt}|\mathbf{Y}_{o,i}, \mathbf{x}_i\}$, may nonetheless have mean close to $\mathbf{0}$. When this is the case, then the modified GEE will be approximately unbiased. The unbiasness of any GEE approach requires unbiased estimates of $\mathbf{V}_i$, so that correct linear combinations of the residuals $[\mathbf{Y}_{o,i}-\mathbf{p}_{o,i}]$ are taken in these estimating equations; our experience has found that if the estimate of $\mathbf{V}_i$ is poor and highly biased, such as under MAR with an "all-available-pairs" approach, then the resulting estimate of $\boldsymbol{\beta}$ can be highly biased. We explore this conjecture in a study of asymptotic bias in Section 5, where the bias of the estimate from our joint modified GEE is compared to the bias of the estimate using standard GEE. Further, since they will be used often in practice, we also explore the bias of the GEE from separate estimation for each outcome, using both the modified approach and the standard approach.

4 Application: Analysis of cardiac abnormalities in children born to HIV-infected women

Data from cross-sectional and short-term longitudinal studies (Lipshultz et al., 1998) have shown that children infected with HIV-1 have an increased risk of cardiovascular abnormalities. We aimed to investigate this hypothesis using data from the P^2C^2 study described in the Introduction. The P^2C^2 study is also used to illustrate the application of the proposed methodology. In the P^2C^2 study, a birth cohort of 393 infants born to women infected with HIV-1 were to have cardiovascular function measured approximately every year from birth to age 6; giving up to 7 measurements on each child. Of these 393 infants, 74 (18.8%) were HIV positive, and 319 (81.2%) were HIV negative. The main scientific interest is in determining if HIV-1 infected children have worse heart function over time. To truly understand the change in heart function over time, four dichotomous measures of heart function are jointly modelled over time. These four measures of abnormal heart function are: abnormal LV fractional shortening (1=yes, 0=no); decreased LV contractility (1=yes, 0=no); abnormal heart rate (1=yes, 0=no); and abnormal blood pressure (1=yes, 0=no). The main covariate of interest is the effect of HIV infection; other possible covariates that could be confounders are mother's smoking status during pregnancy (1=yes, 0=no); gestational age (in weeks) and birthweight standardized for age (1=abnormal, 0=normal). A child of a mother who smokes is expected to have worse heart function. Children with younger gestational age and lower birthweight (standardized for gestational age) may also be at risk for cardiac problems.

Thus, to examine the effect of HIV-1 effect in these children born to HIV-infected women, we considered the following marginal logistic regression model,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\log\left[\frac{p_{ikt}}{1-p_{ikt}}\right]=\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 \mathrm{HIV}_i+\beta_4 \mathrm{smoke}_i+\beta_5 \mathrm{age}_i+\beta_6 \mathrm{wt}_i,$$

for $t = 0, 1, ..., 6$, where HIV_i equals 1 if the i^{th} child is born with HIV-1 and equals 0 if otherwise; smoke_i equals 1 if the mother smoked during pregnancy, and 0 otherwise; age_i is the gestational age (in weeks); and wt_i equals 1 if the child's birthweight for gestational age was abnormal, and 0 otherwise. To account for the association among the binary outcomes, the autoregressive correlation structure given in (2) was used. As seen in Table 1, a feature of this study which complicates the analysis is that there is a lot of missing data, with only 1 out of the 393 children having outcomes measured at all 7 occasions. To explore how missing data affects various estimation techniques, we compare the proposed joint modified GEE estimates of $\boldsymbol{\beta}$ to those obtained using three alternative approaches using an AR1 correlation structure: 1) the standard GEE approach using "all-available-pairs", separately for each outcome; 2) the modified GEE approach, separately for each outcome; 3) joint standard GEE using "all-available-pairs". In effect, approaches 1) and 2) assume the correlations between different outcome variables, at the same or different times, is 0, i.e., $Corr(Y_{ijt}, Y_{ikt}|\mathbf{x}_i) = 0$ for $j \neq k$.

Before describing the results of the different GEE approaches, it is of interest to explore the missing data mechanism that might be generating the missing data. A somewhat informal way to assess if the data are missing completely at random is to formulate a logistic regression model for missingness at each time point, given the outcome data at the previous time point was observed. In particular, we define the indicator random variable R_{it} which equals 1 if the outcomes $\{Y_{ijt}\}$ are observed at time t and 0 if (Y_{ijt}) is unobserved, for $t = 1, ..., 6$. Then the conditional probability of interest is

$$\pi_{it}=\mathrm{pr}\left(R_{it}=1|R_{i,t-1}=1, Y_{i1,t-1}, \ldots, Y_{i4,t-1}, x_i, \gamma\right), \tag{12}$$

for $t = 1, ..., 6$. Note that this probability is estimable since the values of $(Y_{i1,t-1}, ..., Y_{i4,t-1})$ are observed when $R_{i,t-1} = 1$. We fit a logistic regression model to π_{it} with a linear time effect (quadratic was not significant), covariate effects corresponding to the four outcomes, $(Y_{i1,t-1}, ..., Y_{i4,t-1})$, and interactions between time and $(Y_{i1,t-1}, ..., Y_{i4,t-1})$, between x_i and $(Y_{i1,t-1}, ..., Y_{i4,t-1})$; and between the elements of $(Y_{i1,t-1}, ..., Y_{i4,t-1})$. Under MCAR, all effects of the outcomes, and interactions with the outcomes will be 0, e.g.,

$$\mathrm{pr}\left(R_{it}=1|R_{i,t-1}=1, Y_{i1,t-1}, \ldots, Y_{i4,t-1}, x_i, \gamma\right)=\mathrm{pr}\left(R_{it}=1|R_{i,t-1}=1, x_i, \gamma\right)$$

and any GEE approach will be approximately unbiased There were 904 observations over time that contributed to this logistic regression; since these 904 were repeated measures from the 393 children, we fit a GEE under independence to estimate the logistic regression parameters for π_{it}. Out of these 904 times when $(Y_{i1,t-1}, ..., Y_{i4,t-1})$ was observed, $(Y_{i1t}, ..., Y_{i4t})$ was observed at the next visit ($R_{it} = 1$) 448 times (49.6%). The results are given in Table 3. We kept all interactions in the model that were significant at .10. We see that older patients with abnormal blood pressure at the previous visit are more likely to be seen at the current visit ($p < .05$); patients with abnormal gestational age and abnormal heart rate at the previous visit are more likely to be seen at the current visit ($p < .10$); and patients with both abnormal blood pressure and fractional shortening at the previous visit are more likely to be seen at the current visit ($p < .10$). The latter two effects are only marginally significant, but could still be a factor as to whether the standard GEE will be approximately unbiased. Thus, it does appear that the sicker patients are more likely to be seen; a GEE approach that minimizes the bias is warranted.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4 gives the estimates of $\boldsymbol{\beta}$ obtained using the four different approaches; joint modified GEE, joint standard GEE, separate modified GEE, and separate standard GEE. In general, assuming that the proposed joint modified GEE is correct, we see that the estimated relative differences (calculated as 1 minus the ratio of a given estimate to the proposed GEE estimate) are large for some effects. In particular, the standard GEE and the modified GEE (separately for each outcome), as well as joint standard GEE gave different estimates than the newly proposed estimate for the low birthweight effect for most outcomes, and for the mother smoked and gestational age effects on contractility. Although the joint standard GEE tends to have smaller relative difference than standard GEE (separately for each outcome), it is not uniformly smaller than the modified GEE (separately for each outcome). We do note here, though, that if one chooses a .05 level of significance as a cutoff, all three approaches give the same conclusions as to which effects are significant. Further the estimated standard errors are very similar using all approaches; we are mainly concerned with bias in this paper, but a simulation comparing finite sample mean square error is a topic for future exploration. Overall, the results based on the newly proposed joint modified GEE suggest that the covariates appear to only significantly affect the heart rate outcome. Children with HIV have $exp(0.9753) \approx 2.7$ times the odds of having an abnormal heart rate than children without HIV; further, children whose mother smoked during pregnancy have $exp(0.4453) \approx 1.6$ times the odds of having an abnormal heart rate than children whose mother did not smoke.

Finally, without knowledge of the true model generating the data, we can only remark that these different approaches can yield discernibly different regression parameter estimates, but we cannot assess which method produces the most or least bias. To address the latter issue, we conducted an asymptotic study of bias that compared these methods for handling missing data.

5 Asymptotic Study of Bias of $\boldsymbol{\beta}$

In the asymptotic study of bias that follows, we assume that the models for the means, $E(\mathbf{Y}_i|\mathbf{x}_i,\boldsymbol{\beta})$, are correctly specified. Thus, bias will result only from the fact that the missing data are not MCAR.

For simplicity, we consider the case of two binary outcomes at three time points, resulting in 6 correlated binary outcomes per subject. We assume a simple two group configuration, e.g. active treatment versus placebo. Subjects are assumed to belong to either group with equal probability. To specify the true underlying joint distribution of the binary responses, we choose the model for correlated binary data first described by Bahadur (1961), and later by Cox (1972). With 2 binary outcomes at each of 3 times, there are 6 binary outcomes, and the joint distribution of an individual's responses is multinomial with 2^6 probabilities. Thus, in Bahadur's correlated binary model, the joint distribution of an individual's responses at the three times is multinomial,

$$\begin{aligned}&\text{pr}\{Y_{i11}=y_{11}, Y_{i12}=y_{12}, Y_{i13}=y_{13}, Y_{i21}=y_{21}, Y_{i22}=y_{22}, Y_{i23}=y_{23}, |x_i, \beta, \alpha\}\\&=\left\{\prod_{t=1}^{3}\prod_{k=1}^{2} p_{ikt}^{y_{ikt}}(1-p_{ikt})^{(1-y_{ikt})}\right\}\left\{1+\sum_{s<t}\sum_{k=1}^{2}\rho_{i,ks,kt}z_{iks}z_{ikt}+\sum_{s\neq t}\sum_{j\neq k}\rho_{i,js,kt}z_{ijs}z_{ikt}\right\},\end{aligned} \tag{13}$$

where

$$z_{ikt}=\frac{y_{ikt}-p_{ikt}}{\{p_{ikt}(1-p_{ikt})\}^{1/2}}$$

and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\log\left[\frac{p_{ikt}}{1-p_{ikt}}\right]=\beta_{0k}+\beta_{Gk}x_i+\beta_{\tau k}(t-1)+\beta_{G\tau k}x_i(t-1);$$

for $t = 1, 2, 3$ and $k = 1, 2$. Here, x_i is a dichotomous covariate indicating group membership for the i^{th} individual.

For the study of asymptotic bias, the parameters of the true model are as follows. The marginal regression parameters for the outcomes $k = 1, 2$ are

$$(\beta_{01},\beta_{G1},\beta_{\tau 1},\beta_{G\tau 1})=(-.25,.5,.2,-.5) \quad \text{and} \quad (\beta_{02},\beta_{G2},\beta_{\tau 2},\beta_{G\tau 1})=(.25,-.5,-.2,.5).$$

A variety of different correlation structures were examined and the same overall pattern of results were obtained. For simplicity, the results from a true exchangeable correlation structure are presented here, in which $\rho_{i,js,kt}$ = for $\alpha \in \{0.1, 0.25\}$ for all $js \neq kt$. Because of constraints on the joint distribution in (13), the maximum possible value of α is approximately .25; however, while not particularly large, this value still illustrates the substantial bias that can occur using various GEE approaches.

The true drop-out mechanism is assumed to depend on (Y_{ijt}, Y_{ikt}) at the previous times and on the group membership, with subjects dropping out at times 2 or 3. We define the indicator random variable R_{it} which equals 1 if (Y_{ijt}, Y_{ikt}) is observed and 0 if (Y_{ijt}, Y_{ikt}) is unobserved, for $t = 2, 3$, and we define the dropout probability to equal

$$\begin{aligned}\text{logit}\{\text{pr}(R_{it}=0|R_{i1}=\ldots=R_{i,t-1}=1,y_{i11},\ldots,y_{i1t},y_{i21},\ldots,y_{i2t},x_i,\gamma)\}&\\ =\gamma_0+\gamma_G x_i+\gamma_{y_1}y_{i1,t-1}&\\ +\gamma_{y_2}y_{i2,t-1}&\\ +\gamma_{G_{y_1}}x_i y_{i1,t-1}&\\ +\gamma_{G_{y_{2-}}}x_i y_{i2,t-1},&\end{aligned} \tag{14}$$

($t = 2, 3$). In (14), the probability of being missing (or observed) at time t, given that the subject is observed at the previous occasions ($R_{i1} = ... = R_{i,t-1} = 1$), depends on the previous responses and on group membership. Note, if $\gamma_{y1} = \gamma_{y2} = \gamma_{Gy1} = {}_{Gy2} = 0$, then the data are MCAR.

Next, we consider the derivation of the asymptotic bias of $\widehat{\beta}$. First, suppose that the missing data are MCAR, then $\widehat{\beta}$ from any of the GEE methods described in Section 3 is consistent, i.e., $\widehat{\beta} \xrightarrow{P} \beta$. However, if the data are MAR, then $\widehat{\beta} \xrightarrow{P} \beta^*$, where $\boldsymbol{\beta}^*$ is not necessarily equal to $\boldsymbol{\beta}$. The goal is to assess ($\boldsymbol{\beta}^* - \boldsymbol{\beta}$), the asymptotic bias of $\widehat{\beta}$. Following Rotnitzky & Wypij (1994), the asymptotic bias of $\widehat{\beta}$ can be ascertained by solving the expected value of an estimating equation $\mathbf{u}(\boldsymbol{\beta})$

$$E[\mathbf{u}(\beta^*)]=0 \tag{15}$$

for $\boldsymbol{\beta}^*$, where the expectation is taken with respect to the discrete distribution of ($\mathbf{Y}_i$, x_i,R_{i2},R_{i3}). Basically, the expectation in (15) is a weighted sum, where the weights are the probability of the given realization of ($\mathbf{Y}_i$,x_i,R_{i2},R_{i3}). Since there are 2^6 possible values of $\mathbf{Y}_i$,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and two possible values for each of x_i, R_{i2}, and R_{i3}, then the multinomial distribution for ($\mathbf{Y}_i$, x_i, R_{i2}, R_{i3}) will have $J = 2^9$ probabilities. In particular, $E[\mathbf{u}(\boldsymbol{\beta}^*)]$ in (15) equals

$$\sum_{(y,x,r_2,r_3)} pr\left(\mathbf{Y}_i=y, x_i=x, R_{i2}=r_2, R_{i3}=r_3\right) \mathbf{u}\left(\beta^*; \mathbf{Y}_i=y, x_i=x, R_{i2}=r_2, R_{i3}=r_3\right),$$

where $\mathbf{u}(\boldsymbol{\beta}; \mathbf{Y}_i, x_i, R_{i2}, R_{i3})$ denotes $\mathbf{u}(\boldsymbol{\beta})$ as a function of ($\mathbf{Y}_i$, x_i, R_{i2}, R_{i3}), and the sum is over all $J = 2^9$ patterns of ($\mathbf{Y}_i$, x_i, R_{i2}, R_{i3}). We can solve for $\boldsymbol{\beta}^*$ using any GEE program, where the 'data' consist of $J = 2^9$ 'observations', each with weights $pr[\mathbf{Y}_i = y, x_i = x, R_{i2} = r_2, R_{i3} = r_3]$.

Our main concern is with the bias in estimating $\boldsymbol{\beta}$ when missing data follow a MAR drop-out process. We consider the following approaches, all of which give asymptotically unbiased estimates under MCAR: 1)IND=estimation under the naive assumption of independence, i.e., $\rho_{i,js,kt} = 0$ for all j, k, s, t; 2) sep-standard-GEE=GEE using "all-available-pairs" separately for each outcome; 3) sep-mod-GEE=the modified GEE approach, separately for each outcome; 4) joint-standard-GEE=GEE with joint estimation using "all-available-pairs"; 5) joint-mod-GEE=our proposed modified GEE with joint estimation of the correlation for all outcomes. Since all estimates are unbiased under an MCAR dropout mechanism, any possible bias results only from the fact that the missing data are not MCAR.

Table 5 gives the asymptotic bias of the various GEE approaches for different values of (γ_0, γG, γ_{y1}, γ_{y2}, γG_{y1}, γG_{y2}), corresponding to drop-out rates of approximately 15%, 30%, 50%. We specified three sets of γ's. In the first set, missingness depends on both outcomes at the prior time, and all of the GEE approaches use the correct exchangeable correlation model. In the second set, missingness depends on only the first outcome variable (Y_{i1t}) at the prior time ($\gamma_{y2} = \gamma G_{y2} = 0$), and all of the GEE approaches use the correct exchangeable correlation model. In this case, we might expect bias in the estimates of the parameters for outcome 2 for the GEE approaches with separate estimation for outcomes 1 and 2 over time, since this is akin to non-ignorable missingess for outcome 2 (dropout depends on $Y_{i1,t-1}$, which is not in the estimation procedure for Y_{i2t}). In the third set, missingness depends on both outcomes at the prior time, and all of the GEE approaches use the wrong correlation model (AR1 instead of the true exchangeable model).. This will provide insight into how the various GEE approaches perform when the mean is correctly specified but the correlation model is incorrect.

Examining the results in Table 5, we see that the estimates under the naive assumption of independence have the largest bias; this approach should not be used when there is dropout. The sep-standard-GEE has the next largest bias. Even with only 15 % dropout, the sep-standard-GEE can have as much as 15 % relative bias. With 30 % dropout, the relative bias of the sep-standard-GEE can be as high as 33 %, and with 50 % dropout, the relative bias can be as high as 47 %. The relative bias of the sep-standard-GEE seems to be similar regardless of whether dropout depends on first outcome or both of the previous outcomes. Using sep-mod-GEE reduces the relative bias of the sep-standard-GEE for all configurations. Finally, joint-standard-GEE tends to have a similar magnitude of bias as sep-mod-GEE. In general, when dropout depends on both outcomes at prior times, using sep-mod-GEE for the outcomes or joint-standard-GEE reduces the relative bias of the sep-standard-GEE by approximately 33 %; however, the relative bias of sep-mod-GEE can still be substantial (as high as 23 %), as can the bias of joint-standard-GEE (as high as 21 %). When dropout depends only on the first outcome at the prior time, sep-mod-GEE is unbiased for the regression parameters of this first outcome; this is to be expected since this is the exact case considered by Lipsitz et al. (2000). However, in this case, the relative bias of sep-mod-GEE for the outcomes can be high for the regression parameters of the second outcome, as high as 44 %; further, for the second outcome, the sep-standard-GEE and sep-mod-GEE perform very similarly. The joint-standard-GEE

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

tends to have smaller bias for the second outcome than the sep-standard-GEE, but has very high bias for the first outcome (as high as 36 %). Our proposed approach (joint-mod-GEE) is asymptotically unbiased in all configurations when dropout depends on both outcomes at prior times, and the correlation model is correctly specified as exchangeable.

Finally, from Table 5, we see that when dropout depends on both outcomes at prior times, and the correlation structure is incorrectly specified as AR1, the bias tends to be 2% larger in absolute value for any GEE approach when compared to correctly specifying the correlation. In particular, the joint-mod-GEE tends to have approximately 2% bias; thus, joint-mod-GEE, at least for this configuration, appears to be robust to mis-specification of the correlation model.

When the four GEE approaches are considered, the results in Table 5 indicate that the sep-standard-GEE can have quite appreciable bias. Although it reduces the bias over sep-standard-GEE, sep-mod-GEE can still have substantial bias. Although tending to reduce the bias over sep-standard-GEE, the joint-standard-GEE still has appreciable bias. Not surprisingly, the magnitude of the bias in estimating increases with increasing drop-out rates and increasing correlation. It is worth emphasizing that the overall pattern of results reported in Table 5 have been replicated in many other configurations that were considered but not reported here; because of the complexity of specifying a joint distribution for multivariate longitudinal binary data, all of these configurations had two binary outcomes at each of 3 times points.

Next, the findings from this asymptotic study of bias can be put in the context of the results from the example in Section 4. In the example, the missingness mechanism (Table 3) appears to depend on all outcomes except Contractility. In this case, as in the second set of asymptotic calculations in which missigness does not depend on all outcome variables at the previous time, we see (Table 4) the largest relative differences in the GEE estimates (versus joint modified GEE) for the parameters of Contractility. When using the GEE approaches with separate estimation for the outcomes over time, the missingness mechanism for Contractility can be considered non-ignorable missingess (in Table 3, dropout depends the other outcomes, which are not in the estimation procedure for Contractility), and can lead to considerable bias. For the other three outcomes, as in the first set of asymptotic calculations in which missigness depends on all outcome variables at the previous time, there can still be substantial relative differences in the GEE parameter estimates (versus joint modified GEE). Finally, in general, as in the asymptotic study, sep-standard-GEE tends to produce the largest bias, with sep-mod-GEE having the next largest bias. The joint-standard-GEE tends to have the least bias, although as in the second set of asymptotic studies for $\boldsymbol{\beta}_{G\tau,1}$, we see that joint-standard-GEE has a larger relative difference than sep-standard-GEE for the Gestational Age effect for the Fractional Shortening outcome.

6 Discussion

In this paper we consider mutlivariate binary data measured longitudinally. We have shown that joint estimation with all outcomes using a modified GEE for handling missing at random response data yields regression parameter estimates with less bias than the standard GEE or a modified GEE separately for each outcome, as well as joint estimation with standard GEE. The proposed modified GEE uses an EM-type algorithm, where the EM-type algorithm is based on the multivariate normal distribution. Use of a multivariate normal distribution in the EM-type algorithm avoids having to completely specify the full joint distribution of the vector of mutlivariate longitudinal binary responses.

The results of the asymptotic study suggest that joint estimation using the modified GEE, with a correctly specified model for the correlation, has negligible bias. Note that if the “working” correlation is misspecified, some bias can arise using this approach. We found that the joint

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

modified GEE had minimal (2%) bias when the true correlation was exchangeable, and we estimated an AR1 correlation. At one extreme end of a misspecified correlation model, using the modified GEE separately for each outcome can be considered a special case of our proposed method in which the working correlation between different outcome variables is set to 0. Therefore, the proposed modified GEE approach must incorporate careful modelling of the correlations, which can be considered a potentially unattractive feature of the approach. However, most alternative approaches, including maximum likelihood and multiple imputation, also require correct specification of the correlations with MAR missing data. The only approach with MAR missing data that does not require correct specification of the correlations is weighted estimating equations (WEE), which requires specification and estimation of the missing data mechanism. The downside to WEE in this setting are two-fold. First, it is less suitable for non-monotone missing data patterns such as ours, and second, the estimation of the missing data mechanism can involve many more additional nuisance parameters than the joint modified GEE.

The configurations used in the asymptotic study of bias were somewhat simpler than the scenario actually encountered in the example. Despite this, the pattern of results from the asymptotic study suggest what methods are more suitable for the data from this example. Because of the broad range of possible data configurations and underlying probability distributions generating the data, it is difficult to draw definitive conclusions from the asymptotic studies. Nonetheless, in terms of bias, in the asymptotic study reported here, the joint estimation using the modified GEE appears to perform discernibly better than the standard GEE (joint or separate estimation) and modified GEE separately across outcomes. In this paper, we are mainly concerned with bias. In the example, the estimated standard errors are very similar using all approaches; however, in simulations for univariate longitudinal data, Lipsitz (2000) found that the modified GEE estimate in some cases could have substantially smaller variances than the standard (all available pairs) GEE estimate. We would expect this relationship to hold for joint GEE estimation, but this is a topic for future exploration.

Since the proposed method is computationally feasible, we can recommend that it replace the standard GEE in cases where there are missing data. Thus, when estimating the marginal regression parameters for multivariate, longitudinal binary data, to protect against missingness that could depend on any or all of the outcomes, we suggest the use of our proposed method to jointly estimate the regression parameters. Even in settings where there is interest only in the marginal regression model for a single outcome variable over time, the proposed method has the potential to protect against biases that might arise when missingness depends on other outcome variables. Finally, although not explored here, the approach can be easily extended to handle the case when there are partially observed data at each time point.

Acknowledgments

We are grateful for the support from the United States National Institutes of Health grants GM 29745, HL 69800, AI 60373, CA 74015, CA 70101, CA69222 and CA 68484.

References

Bahadur, RR. A representation of the joint distribution of responses to n dichotomous items.. In: Solomon, H., editor. Studies in Item Analysis and Prediction. Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press; 1961. p. 158-68.

Cox DR. The analysis of multivariate binary data. Applied Statistics 1972;21:113–20.

Crowder MJ. Gaussian estimation for correlated binomial data. J. Roy. Statist. Soc. B 1985;47:229–237.

Finkelstein DM, Williams PL, Molenberghs G, Feinberg J, Powderly W, Kahn J, Dolins R, Cotton D. Patterns of opportunistic infections in patients with HIV infection. J. Acq. Immune Def. Syndr. Hum. Retrovir 1996;12:38–45.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Fitzmaurice GM, Lipsitz SR, Molenberghs G. Bias in Estimating Association Parameters for Longitudinal Binary Responses with Drop-outs. Biometrics 2001;57:15–21. [PubMed: 11252590]

Galecki AT. General Class of Covariance Structures for Two or More Repeated Factors in Longitudinal Data Analysis. Communications in Statistics - Theory and Methods 1994;23:3105–3119.

Huber, PJ. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press; Berkeley: 1967. The behavior of maximum likelihood estimates under nonstandard conditions.; p. 221-233.

Laird NM. Missing data in longitudinal studies. Statistics in Medicine 1988;7:305–315. [PubMed: 3353609]

Lee, H.; Laird, NM.; Johnston, G. Combining GEE and REML for estimation of generalized linear models with incomplete multivariate data.. 1998. Unpublished Manuscript

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

Lipshultz SE, Easley KA, Orav EJ, Kaplan S, Starc TJ, Bricker JT, Lai WW, Moodie DS, McIntosh K, Schluchter MD, Colan SD. Left ventricular structure and function in children infected with human immunodeficiency virus: the prospective P2C2 HIV Multicenter Study. Pediatric Pulmonary and Cardiac Complications of Vertically Transmitted HIV Infection (P2C2 HIV) Study Group. Circulation 1998;97:1246–1256. [PubMed: 9570194]

Lipshultz SE, Easley KA, Orav EJ, Kaplan S, Starc TJ, Bricker JT, Lai WW, Moodie DS, Sopko G, Colan SD. Cardiac dysfunction and mortality in HIV- infected children: The Prospective P2C2 HIV Multicenter Study. Pediatric Pulmonary and Cardiac Complications of Vertically Transmitted HIV Infection (P2C2 HIV) Study Group. Circulation 2000;102:1542–1548. [PubMed: 11182983]

Lipshultz SE, Easley KA, Orav EJ, Kaplan S, Starc TJ, Bricker JT, Lai WW, Moodie DS, Sopko G, Schluchter MD, Colan SD. Cardiovascular status of infants and children of women infected with HIV-1 (P(2)C(2) HIV): a cohort study. Lancet 2002;360:368–73. [PubMed: 12241776]

Lipsitz SR, Molenberghs G, Fitzmaurice GM, Ibrahim JG. GEE with Gaussian estimation of the correlations when data are incomplete. Biometrics 2000;56:528–536. [PubMed: 10877313]

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Second Edition. Wiley; New York: 2002.

Mardia KV. Some contributions to the contingency– type bivariate distributions. Biometrika 1967;54:235–249. [PubMed: 6049540]

Plackett RL. A class of bivariate distribution. Journal of the American Statistical Association 1965;60:516–522.

Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033–1048. [PubMed: 3233244]

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Amer. Statist. Assoc 1995;90:106–121.

Rotnitzky A, Wypij D. A note on the bias of estimators with missing data. Biometrics 1994;50:1163–1170. [PubMed: 7786998]

Royall RM. Model robust confidence intervals using maximum likelihood estimators. International Statistical Review 1986;54:221–26.

Rubin DB. Inference and missing data. Biometrika 1976;63:581–592.

SAS Institute Inc.. SAS/STAT Software: Changes and Enhancements through Release 6.12. SAS Institute Inc.; Cary, NC: 1996.

White H. Maximum likelihood estimation under mis-specified models. Econometrica 1982;50:1–26.

Whittle P. Gaussian estimation in stationary time series. Bull. Internat. Statist. Inst 1961;39:1–26.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Frequncy distribution of the number of echocardiograms for children in the (P^2C^2) longitudinal study

Number of echocardiograms	Number of Subjects	Percent
1	148	37.66
2	105	26.72
3	56	14.25
4	45	11.45
5	30	7.63
6	8	2.04
7	1	0.25
Total	393	100.00

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Number of subject seen at each occasion

Age at visit (years)	Number of Subjects	Percent (out of n = 393)
birth	262	66.67
1	260	66.16
2	149	37.91
3	116	29.52
4	80	20.36
5	37	9.41
6	7	1.78

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Parameter Estimates for missingness model pr($R_{it} = 1 | R_{i,t-1} = 1, Y_{i1,t-1}, Y_{i2,t-1}, Y_{i3,t-1}, Y_{i4,t-1}, x_i, \gamma$)

Effect	$\hat{\beta}$	SE	*Z*-statistic	P-value
Intercept	0.280	1.072	0.26	0.794
Age	-0.278	0.051	-5.42	0.000
HIV	0.602	0.165	3.66	0.000
Gest. Age	0.011	0.028	0.40	0.688
Mom Smoked	-0.087	0.152	-0.57	0.567
Low Birth Wt	-0.016	0.163	-0.10	0.920
$BP_{i,t-1}$	-1.769	0.584	-3.03	0.003
$HR_{i,t-1}$	-5.983	3.273	-1.83	0.068
$FS_{i,t-1}$	-0.246	0.175	-1.40	0.160
$Cont_{i,t-1}$	-0.163	0.182	-0.90	0.369
Age*$BP_{i,t-1}$	0.649	0.205	3.17	0.002
Gest. Age*$HR_{i,t-1}$	0.155	0.086	1.81	0.070
$BP_{i,t-1}$*$FS_{i,t-1}$	0.873	0.505	1.73	0.084

BP=Blood pressure, HR=Heart Rate,

FS =Fractional Shortening, Cont=Contractility

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 4

Regression parameter (β) estimates for the (P^2C^2) longitudinal study of cardiac abnormalities in children born to HIV-infected women, using AR(1) correlation model

		Separate standard GEE		Separate modified GEE		Joint Standard GEE		Joint Modified GEE
Outcome Variable	**Covariate**	**Estimate (SE)**	**Relative Difference (%)[a]**	**Estimate (SE)**	**Relative Difference(%)[a]**	**Estimate (SE)**	**Relative Difference(%)[a]**	**Estimate (SE)**
Blood Pressure	Intercept	1.086 (2.259)	-5.6	1.145 (2.257)	-0.5	1.204 (2.251)	4.7	1.150 (2.258)
	time	-0.416 (0.270)	9.9	-0.417 (0.272)	10.1	-0.381 (0.272)	0.6	-0.379 (0.271)
	time^2	0.020 (0.057)	28.8	0.021 (0.057)	32.7	0.016 (0.057)	4.5	0.016 (0.057)
	hiv	0.171 (0.290)	19.1	0.170 (0.290)	18.5	0.145 (0.289)	0.9	0.143 (0.289)
	Mom Smoked	0.095 (0.249)	-4.7	0.094 (0.249)	-5.7	0.096 (0.248)	-3.7	0.100 (0.247)
	Gest. Age	-0.081 (0.060)	-2.8	-0.082 (0.059)	-0.8	-0.084 (0.059)	1.7	-0.083 (0.060)
	Low Birth Wt	-0.128 (0.279)	1.3	-0.126 (0.279)	0.0	-0.132 (0.278)	4.7	-0.126 (0.277)
Contractility	Intercept	0.189 (1.542)	-131.5	0.148 (1.563)	-124.7	-0.338 (1.528)	-43.7	-0.600 (1.552)
	time	-0.345 (0.164)	7.3	-0.342 (0.166)	6.5	-0.313 (0.165)	-2.5	-0.321 (0.167)
	time^2	0.050 (0.037)	0.8	0.053 (0.037)	5.2	0.045 (0.037)	-9.2	0.050 (0.038)
	hiv	0.062 (0.262)	4.2	0.048 (0.263)	-18.4	0.086 (0.254)	44.8	0.059 (0.257)
	Mom Smoked	0.073 (0.205)	-22.4	0.062 (0.206)	-34.5	0.093 (0.199)	-1.3	0.094 (0.199)
	Gest. Age	-0.041 (0.039)	86.2	-0.040 (0.039)	82.1	-0.028 (0.038)	29.4	-0.022 (0.039)
	Low Birth Wt	0.134 (0.218)	-37.6	0.151 (0.220)	-29.7	0.188 (0.209)	-12.2	0.214 (0.209)
Fractional Shortening	Intercept	2.541 (1.583)	39.8	1.842 (1.596)	1.3	2.377 (1.578)	30.8	1.818 (1.610)
	time	-1.296 (0.153)	-2.3	-1.292 (0.159)	-2.6	-1.314 (0.154)	-0.9	-1.327 (0.162)
	time^2	0.186 (0.030)	-0.4	0.183 (0.032)	-2.1	0.188 (0.031)	0.5	0.187 (0.033)
	hiv	0.305 (0.219)	-2.7	0.343 (0.219)	9.5	0.282 (0.221)	-9.9	0.314 (0.223)
	Mom Smoked	-0.269 (0.181)	-8.2	-0.267 (0.181)	-8.7	-0.301 (0.183)	3.0	-0.292 (0.184)
	Gest. Age	-0.057 (0.040)	51.7	-0.039 (0.040)	4.3	-0.052 (0.040)	38.4	-0.038 (0.041)
	Low Birth Wt	0.073 (0.205)	-44.9	0.126 (0.202)	-4.3	0.092 (0.206)	-30.0	0.132 (0.204)
Heart Rate	Intercept	-1.112 (1.840)	1.4	-1.110 (1.840)	1.2	-1.248 (1.811)	13.7	-1.097 (1.791)
	time	-0.749 (0.377)	7.6	-0.751 (0.377)	7.9	-0.690 (0.393)	-0.8	-0.696 (0.414)
	time^2	0.059 (0.118)	53.8	0.060 (0.118)	54.8	0.041 (0.128)	7.0	0.039 (0.137)
	hiv	0.972 (0.258)	-0.3	0.973 (0.258)	-0.2	0.962 (0.256)	-1.4	0.975 (0.256)
	Mom Smoked	0.449 (0.250)	0.9	0.449 (0.250)	0.9	0.447 (0.247)	0.3	0.445 (0.248)
	Gest. Age	-0.030 (0.047)	-1.0	-0.030 (0.047)	-1.0	-0.026 (0.046)	-12.4	-0.030 (0.046)

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Outcome Variable	Covariate	Separate standard GEE		Separate modified GEE		Joint Standard GEE		Joint Modified GEE
		Estimate (SE)	Relative Difference (%)[a]	Estimate (SE)	Relative Difference(%)[a]	Estimate (SE)	Relative Difference(%)[a]	Estimate (SE)
	Low Birth Wt	0.181 (0.270)	16.2	0.181 (0.270)	15.9	0.172 (0.269)	10.2	0.156 (0.269)

[a]Relative to Joint Modified GEE.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 5

The value $(\beta^*_{G\tau,1}, \beta^*_{G\tau,2})$ to which the estimate $(\widehat{\beta}^*_{G\tau,1}, \widehat{\beta}^*_{G\tau,2})$ converges. The true marginal logistic model has parameters $(\beta_{G\tau,1}, \beta_{G\tau,2}) = (-0.5, 0.5)$, and correlation $\rho_{i,js,kt} = \alpha$ for all $js \neq kt$ (exchangeable)

% Drop-out	Dropout model $(\gamma_0, \gamma_G, \gamma_{y1}, \gamma_{y2}, \gamma_{G_{y1}}, \gamma_{G_{y2}})$	APPROACH[a]	α = 0.10		α = 0. 25	
			$\beta_{G\tau,1}$ (RB %)[b]	$\beta_{G\tau,2}$ (RB %)[b]	$\beta_{G\tau,1}$ (RB %)[b]	$\beta_{G\tau,2}$ (RB %)[b]
Missingness depends on both outcomes at prior time, exchangeable correlation estimated						
15	(2.0,1,-1,-1,1,1)	IND	-0.447 (-10.6)	0.554 (10.8)	-0.367 (-26.6)	0.635 (27.0)
		sep-standard	-0.457 (-8.6)	0.542 (8.4)	-0.422 (-15.6)	0.577 (l5.4)
		sep-mod	-0.477 (-4.6)	0.524 (4.8)	-0.454 (-9.2)	0.550 (l0.0)
		Joint-standard	-0.470 (-6.0)	0.528 (5.6)	-0.458 (-8.4)	0.540 (8.0)
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
30	(1.0,1,-1,-1,1,1)	IND	-0.410 (-18.0)	0.591 (18.2)	-0.278 (-44.4)	0.727 (45.4)
		sep-standard	-0.417 (-11.6)	0.552 (l0.4)	-0.356 (-28.8)	0.641 (28.2)
		sep-mod	-0.460 (-8.0)	0.541 (8.2)	-0.420 (-16.0)	0.582 (l6.4)
		Joint-standard	-0.437 (-12.6)	0.560 (l2.0)	-0.416 (-16.8)	0.583 (l6.6)
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
50	(-0.2,1,-1,-1,1,1)	IND	-0.367 (-26.6)	0.634 (26.8)	-0.183 (-63.4)	0.825 (65.0)
		sep-standard	-0.382 (-23.6)	0.614 (22.8)	-0.299 (-40.2)	0.693 (38.6)
		sep-mod	-0.440 (-12.0)	0.557 (ll.4)	-0.383 (-23.4)	0.609 (21.8)
		Joint-standard	-0.418 (-16.4)	0.581 (l6.2)	-0.399 (-20.2)	0.603 (20.6)
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
Missingness depends on first outcome at prior time, exchangeable correlation estimated						
15	(2.0,1,-2, 0,2,0)	IND	-0.447 (-10.6)	0.554 (10.8)	-0.368 (-26.4)	0.634 (26.8)
		sep-standard	-0.441 (-11.8)	0.547 (9.4)	-0.429 (-14.2)	0.599 (l9.8)
		sep-mod	-0.500 (0.0)	0.546 (9.2)	-0.499 (-0.2)	0.593 (l8.6)
		Joint-standard	-0.442 (-11.6)	0.521 (4.2)	-0.436 (-12.8)	0.531 (6.2)
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
30	(1.0,1,-2, 0,2,0)	IND	-0.413 (-17.4)	0.588 (17.6)	-0.286 (-42.8)	0.719 (43.8)
		sep-standard	-0.375 (-25.0)	0.579 (l5.8)	-0.354 (-29.2)	0.666 (33.2)
		sep-mod	-0.500 (0.0)	0.577 (l5.4)	-0.498 (-0.4)	0.654 (30.8)
		Joint-standard	-0.377 (-24.6)	0.539 (7.8)	-0.366 (-26.8)	0.561 (l2.2)

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

% Drop-out	Dropout model $(\gamma_0, \gamma_G, \gamma_{y1}, \gamma_{y2}, \gamma_{G_{y1}}, \gamma_{G_{y2}})$	APPROACH[a]	α = 0.10		α = 0. 25	
			$\beta_{G\tau,1}$ (RB %)[b]	$\beta_{G\tau,2}$ (RB %)[b]	$\beta_{G\tau,1}$ (RB %)[b]	$\beta_{G\tau,2}$ (RB %)[b]
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
50	(-0.5,1,-2, 0,2,0)	IND	-0.378 (-24.4)	0.623 (24.6)	-0.202 (-59.6)	0.805 (61.0)
		sep-standard	-0.330 (-34.0)	0.610 (22.0)	-0.307 (-38.6)	0.734 (46.8)
		sep-mod	-0.499 (-0.2)	0.608 (21.6)	-0.497 (-0.6)	0.720 (44.0)
		Joint-standard	-0.333 (-33.4)	0.545 (9.0)	-0.321 (-35.8)	0.567 (l3.4)
		Joint-mod	-0.500 (0.0)	0.500 (0.0)	-0.500 (0.0)	0.500 (0.0)
Missingness depends on both outcomes at prior time, AR1 correlation estimated						
15	(2.0,1,-1,-1,1,1)	sep-standard	-0.453 (-9.4)	0.547 (9.4)	-0.405 (-19.0)	0.594 (18.8)
		sep-mod	-0.466 (-6.8)	0.535 (7.0)	-0.435 (-13.0)	0.568 (l3.6)
		Joint-standard	-0.460 (-8.0)	0.539 (7.8)	-0.436 (-12.8)	0.563 (l2.6)
		Joint-mod	-0.490 (-2.0)	0.509 (1.8)	-0.486 (-2.8)	0.514 (2.8)
30	(1.0,1,-1,-1,1,1)	sep-standard	-0.414 (-17.2)	0.585 (17.0)	-0.334 (-33.2)	0.665 (33.0)
		sep-mod	-0.445 (-11.0)	0.556 (ll.2)	-0.395 (-21.0)	0.608 (21.6)
		Joint-standard	-0.420 (-16.0)	0.579 (l5.8)	-0.380 (-24.0)	0.620 (24.0)
		Joint-mod	-0.487 (-2.6)	0.514 (2.8)	-0.483 (-3.4)	0.519 (3.8)
50	(-0.2,1,-1,-1,1,1)	sep-standard	-0.375 (-25.0)	0.623 (24.6)	-0.274 (-45.2)	0.721 (44.2)
		sep-mod	-0.426 (-14.8)	0.572 (l4.4)	-0.364 (-27.2)	0.632 (26.4)
		Joint-standard	-0.385 (-23.0)	0.614 (22.8)	-0.352 (-29.6)	0.651 (30.2)
		Joint-mod	-0.489 (-2.2)	0.513 (2.6)	-0.494 (-1.2)	0.511 (2.2)

[a] IND=naive assumption of independence; sep-standard=GEE using all available pairs, separately for each outcome; sep-mod=modified GEE, separately for each outcome; Joint-standard=GEE using all available pairs, and joint estimation with all outcomes. Joint-mod=modified GEE, and joint estimation with all outcomes.

[b] RB %=Relative Bias percent