A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data

Reference (Published version):
Beunckens, Caroline; Sotto, Cristina & Molenberghs, Geert(2008) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 52(3). p. 1533-1548

DOI: 10.1016/j.csda.2007.04.020
Handle: http://hdl.handle.net/1942/9691

A Simulation Study Comparing Weighted Estimating Equations with Multiple Imputation Based Estimating Equations for Longitudinal Binary Data

Caroline Beunckens [a,*], Cristina Sotto [a,b], Geert Molenberghs [a]

[a] *Center for Statistics, Hasselt University, Agoralaan 1, Building D, 3590 Diepenbeek, Belgium*

[b] *School of Statistics, University of the Philippines, Diliman, Quezon City, Philippines*

Abstract

Missingness frequently complicates the analysis of longitudinal data. A popular solution for dealing with incomplete longitudinal data is the use of likelihood-based methods, when, for example, linear, generalized linear, or non-linear mixed models are considered, due to their validity under the assumption of *missing at random* (MAR). Semi-parametric methods such as generalized estimating equations (GEE) offer another attractive approach but require the assumption of *missing completely at random* (MCAR). Weighted GEE (WGEE) has been proposed as an elegant way to ensure validity under MAR. Alternatively, multiple imputation (MI) can be used to pre-process incomplete data, after which GEE is applied (MI-GEE). Focusing on incomplete binary repeated measures, both methods are compared using the so-called asymptotic, as well as small-sample, simulations, in a variety of correctly specified as well as incorrectly specified models. In spite of the asymptotic unbiasedness of WGEE, results provide striking evidence that MI-GEE is both less biased and more accurate in the small to moderate sample sizes which typically arise in clinical trials.

Key words: missing at random, weighted GEE, multiple imputation GEE, asymptotic bias

* Corresponding author. Tel. +3211268257. Fax: +3211268299.
Email address: caroline.beunckens@uhasselt.be (Caroline Beunckens).

1 Introduction

Longitudinal binary, or in general non-Gaussian, data are common in biomedical research and beyond. A typical study, for instance, would consist of repeatedly observing the presence or absence of some characteristic, taken in relation to covariates of interest. Data arising from such investigations, however, are often prone to incompleteness, or missingness. In the context of longitudinal studies, missingness predominantly occurs in the form of dropout, in which subjects fail to complete the study for one reason or another. The focus of this paper will be on this type of missingness. In what follows, we will discuss methodology that applies to all non-Gaussian settings, but illustrations and simulations will be confined to the prevalent binary case.

The nature of the dropout mechanism affects both the analysis and interpretation of the remaining data. Since one can almost never be certain about the cause of dropout, certain assumptions have to be made. Therefore, when referring to the missingness process, we will use the terminology introduced by Rubin (1976) and Little and Rubin (1987). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data, and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). Note that specific names for these mechanisms for the case of longitudinal data were cornered by Diggle and Kenward (1994). Moreover, Little (1995) further splits the MCAR case in situations where missingness is independent of both outcomes and covariates on the one hand, and cases where missingness is covariate-dependent only. For reasons of simplicity and generality, we prefer to retain the generic MCAR–MAR–MNAR terminology. Full details can be found in Molenberghs and Kenward (2007). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable*, while an MNAR process is non-ignorable. This is not the case for frequentist inference, where the stronger condition of MCAR is required to ensure ignorability (Rubin, 1976). Indeed, frequentist methods, such as standard generalized estimating equations, for which dropout does not need to be modelled, are only valid under the restrictive MCAR assumption. *Weighted generalized estimating equations* (WGEE) and *multiple imputation based generalized estimating equations* (MI-GEE) are two possible alternatives that make it possible to model the data under the MAR missingness mechanism. However, in both methods, dropout needs to be addressed, either by means of a dropout model for WGEE or by an imputation model for MI-GEE, meaning the missing-data mechanism is then not ignorable.

A general taxonomy of models for longitudinal non-Gaussian data consists of three families: marginal, random-effects, and conditional models. Within these model families, a broad set of methods are available, although the marginal and random-effects models are most often used in longitudinal non-Gaussian settings. Such random-effects models, known as generalized linear mixed models, are typically estimated through maximum likelihood, or variations to this theme, implying that ignorability under MAR can be invoked. This is not the case for non-likelihood marginal models, such as the semi-parametric method of generalized estimating equations (Liang and Zeger, 1986), henceforth GEE, which is a second prevalent modelling approach in this area. Such models give valid inferences under the restrictive assumption of MCAR. To be able to analyze the longitudinal non-Gaussian profiles under the weaker MAR assumption, Robins *et al* (1995) extended generalized estimating equations by using inverse probability weights, resulting in weighted estimating equations, or WGEE. An alternative approach is multiple imputation, developed by Rubin (1987). A detailed account is given in Schafer (2003). Missing values are imputed several times, and the resulting complete datasets are analyzed using a standard method, such as GEE. Afterwards, the obtained inferences are combined into a single one (MI-GEE). Regarding the missingness process, standard multiple imputation requires MAR to hold, even though extensions exist. Pros and cons of inverse probability weighting methods with respect to multiple imputation have been the subject of some debate (the discussion of Scharfstein, Rotnitzky, and Robins, 1999; Clayton *et al*, 1998; Carpenter, Kenward, and Vansteelandt, 2006).

In this paper, the focus will be on the comparison between the two GEE versions for incomplete data mentioned above: WGEE and MI-GEE. Comparisons will be made by means of a simulation study, including both small-sample simulations, as well as so-called asymptotic simulations (Rotnitzky and Wypij, 1994). The behavior of both methods in terms of mean squared error (MSE), variance and bias of the estimators will be studied, under correctly specified and misspecified models. In this way, robustness of both methods under misspecification of either the dropout model, the imputation model, or the measurement model, can be explored.

The outline of this paper is as follows. In Section 2, an overview of methods for analyzing incomplete longitudinal non-Gaussian data is given, with main attention on WGEE and multiple imputation together with GEE as analysis method. A description of the asymptotic and small-sample simulation design, as well as the results of the simulation study, is provided in Section 3. We conclude with a discussion in Section 4.

2 Methods for Incomplete Non-Gaussian Longitudinal Data

Whereas the linear mixed model is seen as a unifying parametric framework for Gaussian repeated measures (Verbeke and Molenberghs, 2000), there are a variety of methods in common use in the non-Gaussian setting.

In line with Fahrmeir and Tutz (2001), Diggle *et al* (2002), and Molenberghs and Verbeke (2005), we distinguish between three model families. In a *marginal model*, marginal distributions are used to describe the outcome vector, given a set of predictor variables. The correlation among the components of the outcome vector can then be captured either by adopting a fully parametric model specification or by means of working assumptions, such as in GEE (Liang and Zeger, 1986).

Alternatively, in *conditional models*, any response within the sequence of repeated measures is modelled conditional upon (subsets of) the other outcomes. This could be the set of all past measurements or a subset thereof, as is the case in so-called transition models. A well-known member of this family is the log-linear model (Agresti, 2002).

Finally, in a *subject-specific* or *random-effects model*, the responses are assumed independent, given a collection of subject-specific parameters.

Although we will focus on the marginal model family to which GEE belongs, we will also explore the effect of generating the data from a conditional model. We will describe both of these model families in more detail in turn.

2.1 Data Setting and Notational Conventions

Assume that for subject $i = 1, \ldots, N$ a sequence of responses Y_{ij} is designed to be measured at occasions $j = 1, \ldots, J$. The outcomes are grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$. Define now a dropout indicator D_i for the occasion at which dropout occurs and make the convention that $D_i = J+1$ for a complete sequence. Further, split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components, respectively. For this definition to be meaningful, a balanced design is necessary, in the sense that the measurement occasions are common to all subjects.

In general, one would want to analyze the joint distribution of the measurement and dropout process, $f(\boldsymbol{y}_i, \boldsymbol{d}_i)$ say. Different routes can be taken. A common approach is through the use of a selection model (Rubin, 1976; Little and Rubin, 1987), in which the joint distribution is factorized as the marginal distribution of the measurement process, $f(\boldsymbol{y}_i)$, on the one hand, and the conditional distribution of the dropout process given the measurement process,

$f(\boldsymbol{d}_i|\boldsymbol{y}_i)$, on the other. This factorization could also be reversed, resulting in a pattern-mixture model (Little, 1993, 1994). Finally, one could also use a shared-parameter model (Wu and Carrol, 1988; Wu and Bailey, 1989), in which the measurement and dropout process are assumed to be independent, given a certain set of shared parameters. In this paper, the focus will be on the selection model approach.

2.2 Conditional Models

In a conditional model the parameters describe a feature (e.g., probability, odds, logit) of (a set of) outcomes, given values for the other outcomes (Cox, 1972). The best known example is undoubtedly the log-linear model.

In a transition model, a measurement Y_{ij} in a longitudinal sequence is described as a function of previous outcomes, or history $\boldsymbol{h}_{ij} = (Y_{i1}, \ldots, Y_{i,j-1})$ (Diggle *et al*, 2002, p. 190). One can write a regression model for the outcome Y_{ij} in terms of $\boldsymbol{h}_{ij}$, or alternatively, the error term ε_{ij} can be written in terms of previous error terms. In the case of linear models for Gaussian outcomes, one formulation can be translated easily into another and specific choices give rise to well-known marginal covariance structures such as, e.g., AR(1). The order of a transition model is the number of previous measurements that is still considered to influence the current one. A model is called stationary if the functional form of the dependence does not vary over time.

A particular version of a transition model is a stationary first-order autoregressive model for binary longitudinal outcomes, which follows a logistic-regression type model:

$$\text{logit}[P(Y_{ij} = 1|\boldsymbol{x}_{ij}, Y_{i,j-1} = y_{i,j-1}, \boldsymbol{\beta}, \alpha)] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha y_{i,j-1}. \quad (1)$$

Extension to the second or higher orders is obvious.

2.3 Marginal Models

In marginal or population-averaged models, the parameters characterize the marginal expectation of a subset of the outcomes, without conditioning on other outcomes.

Bahadur (1961) proposed a marginal model for binary outcomes, accounting for the association via marginal correlations. Define the marginal probability

$\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$, and define standardized deviations

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}, \tag{2}$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$, $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3})$,..., and $\rho_{i12\ldots J} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{iJ})$. Then, the general Bahadur model can be represented by the expression

$$f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i), \tag{3}$$

where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2}e_{ij_1}e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3}e_{ij_1}e_{ij_2}e_{ij_3} + \ldots + \rho_{i12\ldots J}e_{i1}e_{i2}\ldots e_{iJ}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. One viewpoint is to consider the factor $c(\boldsymbol{y}_i)$ as a model for overdispersion.

Besides the Bahadur model, a broad set of marginal models have been proposed by Dale (1986), Plackett (1965), Lang and Agresti (1994), Molenberghs and Lesaffre (1994), and Molenberghs and Lesaffre (1999). Even though a variety of flexible full-likelihood models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high-dimensional vectors of correlated data arise, as alluded to in the context of the Bahadur model.

As a consequence, alternative methods have been in demand. Liang and Zeger (1986) proposed so-called *generalized estimating equations* (GEE), useful to circumvent the computational complexity of full likelihood, and which can be considered whenever interest is restricted to the mean parameters. It requires only the correct specification of the univariate marginal distributions, provided one is willing to adopt so-called *working assumptions* about the association structure of the vector of repeated measurements.

Let us introduce more formally the classical form of GEE (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005). The score equations for a non-Gaussian outcome are

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}(A_i^{1/2}C_iA_i^{1/2})^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \;=\; \mathbf{0}, \tag{4}$$

where $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$, $\boldsymbol{\beta}$ is the vector of regression parameters, A_i is a diagonal matrix with the marginal variances, and C_i is the marginal correlation matrix for the repeated measures. Although $A_i = A_i(\boldsymbol{\beta})$ follows directly from the marginal mean model, $\boldsymbol{\beta}$ commonly contains no information about C_i. Therefore, the correlation matrix C_i typically is written in terms of a vector $\boldsymbol{\alpha}$ of unknown parameters, $C_i = C_i(\boldsymbol{\alpha})$. Liang and Zeger (1986) dealt with this set of nuisance parameters $\boldsymbol{\alpha}$ by allowing for specification of an incorrect structure or so-called working correlation matrix. Some of the more popular choices for the working correlations include independence ($\text{Corr}(Y_{ij}, Y_{ik}) = 0,\ j \neq k$), exchangeability ($\text{Corr}(Y_{ij}, Y_{ik}) = \alpha,\ j \neq k$), AR(1) ($\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t,\ t = 0, 1, \ldots, J - j$), and unstructured ($\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk},\ j \neq k$).

Assuming that the marginal mean $\boldsymbol{\mu}_i$ has been correctly specified as $h(\boldsymbol{\mu}_i) = X_i\boldsymbol{\beta}$, they showed that, under mild regularity conditions, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from solving (4) is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and with covariance matrix

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{5}$$

where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right) \quad \text{and} \quad I_1 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\boldsymbol{y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right),$$

with $V_i = A_i^{1/2} C_i A_i^{1/2}$. When the working correlation structure differs strongly from the true underlying structure, there is no price to pay in terms of the consistency of the asymptotic normality of $\widehat{\boldsymbol{\beta}}$, but such a poor choice may result in loss of efficiency. With incomplete data that are MAR or MNAR, an erroneously specified working correlation matrix may additionally lead to bias (Molenberghs and Kenward, 2007).

The GEE moments that are specified coincide with those of the Bahadur model, so that the former can be seen as a non-likelihood version of the latter. In summary, GEE for binary data can be seen as a moment-based version of the Bahadur model. Alternatively, it may be helpful to view it as a "correlation-corrected version of logistic regression."

2.4 Non-Gaussian Longitudinal Data and MAR

While full likelihood methods are appealing because of their flexible ignorability properties, their use for non-Gaussian outcomes can be problematic due to prohibitive computational requirements. Therefore, GEE is an attractive

alternative within the marginal model family. Since GEE is motivated by frequentist considerations, the missing-data mechanism needs to be MCAR for it to be ignorable. This motivated so-called *weighted generalized estimating equations* (WGEE). An alternative mode of analysis, proposed by Schafer (2003), consists of multiply imputing the missing outcomes using a full-parametric model, e.g., of a random-effects or conditional type, followed by analysis of the so-completed sets of data using a conventional marginal (e.g., GEE) or conditional model (e.g., a transition model), and finally performing multiple-imputation inference on the so-analyzed sets of data.

2.4.1 Weighted Generalized Estimating Equations

As Liang and Zeger (1986) pointed out, GEE-based inferences are valid only under MCAR, due to the fact that they are based on frequentist considerations. An important exception, mentioned by these authors, is the situation where the working correlation structure happens to be correct, since then the estimates and model-based standard errors are valid under the weaker MAR. In general, the working correlation structure will not be correctly specified, and hence, Robins *et al* (1995) proposed a class of weighted estimating equations to allow for MAR.

The idea of weighted estimating equations (WGEE) is to weight each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he dropped out. Thus, anyone staying in the study is considered representative of himself as well as of a number of similar subjects that did drop out from the study. The incorporation of these weights, reduces possible bias in the regression parameter estimates. Such a weight can be expressed as

$$\begin{aligned}\nu_{ij} \equiv P[D_i = j] = \prod_{k=2}^{j-1} (1 - P[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1]) \times \\ P[R_{ij} = 0 | R_{i2} = \ldots = R_{i,j-1} = 1]^{I\{j \leq J\}},\end{aligned}$$

where $j = 2, 3, \ldots, J+1$.

Recall that we partitioned $\boldsymbol{Y}_i$ into the unobserved components $\boldsymbol{Y}_i^m$ and the observed components $\boldsymbol{Y}_i^o$. Similarly, we can make the exact same partition of $\boldsymbol{\mu}_i$ into $\boldsymbol{\mu}_i^m$ and $\boldsymbol{\mu}_i^o$. In the weighted GEE approach, the score equations to be solved are:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{d=2}^{J+1} \frac{I(D_i = d)}{\nu_{id}} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}(d)(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)(\boldsymbol{y}(d) - \boldsymbol{\mu}_i(d)) \;=\; \mathbf{0},$$

where $\boldsymbol{y}_i(d)$ and $\boldsymbol{\mu}_i(d)$ are the first $d-1$ elements of $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$ respectively. We

define $\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^t}(d)$ and $(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)$ analogously, in line with the definitions of Robins *et al* (1995).

2.4.2 MI-GEE and MI-Transition

Multiple imputation (MI) was formally introduced by Rubin (1978). The key idea of the procedure is to first replace each missing value with a set of M plausible values drawn from the conditional distribution of the unobserved values, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute. In this way, M imputed datasets are generated (imputation stage), which are then analyzed using standard complete data methods (analysis stage). Finally, the results from the M analyses have to be combined into a single inference (pooling stage) by means of the method laid out in Rubin (1978). In its basic form, multiple imputation requires the missingness mechanism to be MAR, even though versions under MNAR have been proposed (Rubin, 1987; Molenberghs, Kenward, and Lesaffre, 1997).

In line with the notation in Section 2.1, suppose the parameter vector of the distribution of $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ is denoted by $\boldsymbol{\theta}$. Multiple imputation uses the observed data $\boldsymbol{Y}^o$ to estimate the conditional distribution of $\boldsymbol{Y}^m$ given $\boldsymbol{Y}^o$. The missing data are sampled several times from this conditional distribution and augmented to the observed data. The resulting completed data are then used to estimate $\boldsymbol{\theta}$. If the distribution of $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ were known, with parameter vector $\boldsymbol{\theta}$, then $\boldsymbol{Y}_i^m$ could be imputed by drawing a value of $\boldsymbol{Y}_i^m$ from the conditional distribution $f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{\theta})$. The objective of the imputation phase is to sample from this true predictive distribution. However, $\boldsymbol{\theta}$ in the imputation model is unknown, and therefore needs to be estimated from the data first, say $\widehat{\boldsymbol{\theta}}$, after which $f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \widehat{\boldsymbol{\theta}})$ is used to impute the missing data. Precisely, this implies one first generates draws from the distribution of $\widehat{\boldsymbol{\theta}}$, thereby taking sampling uncertainty into account. Generally, the vector $\boldsymbol{\theta}$ in the imputation model differs from the parameter vector $\boldsymbol{\beta}$ that governs the analysis model. Alternatively, a Bayesian approach, in which uncertainty about $\boldsymbol{\theta}$ is incorporated by means of some prior distribution for $\boldsymbol{\theta}$, can also be taken. In the context of multiple imputation, a random $\boldsymbol{\theta}^*$ is first drawn from this prior distribution, which is then put into the distribution of $\boldsymbol{Y}_i$, and then a random $\boldsymbol{Y}_i^m$ is selected from $f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{\theta}^*)$. The estimate of $\boldsymbol{\beta}$ and its estimated variance are calculated using the completed data and a, potentially different, analysis model, $(\boldsymbol{Y}^o, \boldsymbol{Y}^{m*})$: $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{Y}^o, \boldsymbol{Y}^{m*})$, and the *within* imputation variance is $\boldsymbol{U} = \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})$. These steps are repeated a number of M times, producing $\widehat{\boldsymbol{\beta}}^m$ and $\boldsymbol{U}^m$, for $m = 1, \ldots, M$.

In the last phase of multiple imputation, the results of the analyses for the M imputed datasets are pooled into a single inference. The combined point esti-

mate for the parameter of interest $\boldsymbol{\beta}$ from the multiple imputation is simply the average of the M complete-data point estimates (Schafer, 1999). That is, the estimate and its estimated variance are given by:

$$\overline{\widehat{\boldsymbol{\beta}}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\beta}}^m \quad \text{and} \quad \boldsymbol{V} = \boldsymbol{W} + \left(\frac{M+1}{M}\right)\boldsymbol{B},$$

where

$$\boldsymbol{W} = \frac{\sum_{m=1}^{M}\boldsymbol{U}^m}{M} \quad \text{and} \quad \boldsymbol{B} = \frac{\sum_{m=1}^{M}(\widehat{\boldsymbol{\beta}}^m - \overline{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}^m - \overline{\widehat{\boldsymbol{\beta}}})'}{M-1},$$

with $\boldsymbol{W}$ denoting the average *within* imputation variance and $\boldsymbol{B}$ the *between* imputation variance (Rubin, 1987).

Since in WGEE all subjects are given weights, calculated using the hypothesized dropout model, any misspecification of this dropout model will affect all subjects, and thus the results. On the other hand, one can also consider MI together with GEE or with a transition model (in what follows, we refer to these as MI-GEE and MI-Transition, respectively). In essence, this method comes down to first using the predictive distribution of the unobserved outcomes given the observed ones and perhaps covariates. After this step, the missing-data mechanism can be further ignored, provided the missing-data mechanism is MAR. In these MI cases, a misspecification made in the imputation step will only effect the unobserved (i.e., imputed) but not the observed part of the data. Meng's (1994) results show that, as long as the imputation model is not grossly misspecified, this approach will perform well. Considering all this, one might be inclined to expect the MI-GEE or MI-Transition to be more robust against model misspecification than WGEE. In the next section, we will use a simulation study to investigate this idea.

3 A Simulation Study

In the previous section, two approaches have been proposed to overcome the bias occurring in GEE under MAR. WGEE is unbiased for a correctly specified dropout and mean structure of the measurement model. MI-GEE requires compatibility between the imputation and estimation model to be correctly specified. Therefore, it is of interest to quantify the bias and precision under various types of misspecification. To this end, an asymptotic simulation study, as well as small-sample simulations, were conducted on various underlying data-generating models. Whereas asymptotic simulations give a nice paradigm to explore the situation of "large" samples, small-sample simulations give insight into the behavior of the methods in real-life settings.

In the simulation study, we distinguish between two stages: (1) the data-generating stage and (2) the analysis stage. In the first stage, a data-generating

model is defined. Under the selection model framework, this generating model consists of a measurement model on the one hand, and a dropout model, given the measurement model on the other. In the analysis stage, a distinction should be made among three types of models: a measurement model, a dropout model and an imputation model. For the WGEE approach, only a marginal measurement model and a dropout model need to be specified. In contrast, the analysis stage for MI-GEE would entail the specification of an imputation model, rather than a dropout model, as well as a marginal measurement model. Finally, for MI-Transition, a conditional rather than marginal measurement model is needed, as well as an imputation model. For the MI-GEE and MI-Transition approaches, the predictors of dropout are included in the imputation model.

To assess the distinctive and relative merits of the methods of interest, we consider their performance, first in the case without any misspecification, then under various misspecifications. Since our interest lies in comparing WGEE and MI-GEE as methods for dealing with missing data in a binary longitudinal setting, the misspecification can be made either in the dropout model, in the imputation model, or in the measurement model. Misspecification in the missingness mechanism, however, e.g., using MCAR for an underlying MAR mechanism, is not further explored, as this is not the main focus here and has already been investigated extensively (Jansen *et al*, 2006).

In this section, we first define the various generating models employed for the simulations. A description of the design of the simulation study follows, after which the results of the simulation, under each of the various scenarios, are presented.

3.1 Data-Generating Models

For the simulation study, we generated an outcome at 3 time points using three different measurement models: first, three-dimensional binary outcomes were generated from a Bahadur model as well as from a second-order autoregressive, AR(2), transition model; further, a three-dimensional continuous outcome (that was later dichotomized) was generated from a trivariate Gaussian distribution. Whereas the choice of the first two is obvious, since our focus lies on binary repeated measures, the third case depicts real-life settings for which a continuous outcome is available, but the scientific question is based on a dichotomized version of it. For all three cases, the measurement model incorporated a binary treatment indicator, such as a treatment *versus* placebo classification. In addition, for the dropout model, an MAR mechanism was considered. Assuming that dropout can occur only after the first time point, there are three possible dropout patterns: (1) dropout at the second time point, (2) dropout at the third time point, or (3) no dropout. The combination of

the various measurement models and the dropout model gives rise to three data-generating models, which will hereinafter be denoted as GM I (Bahadur measurement model and MAR dropout model), GM II (AR(2) measurement model and MAR dropout model) and GM III (Gaussian measurement model and MAR dropout model). Let us define these three data-generating mechanisms in turn.

Note that we restrict the simulation setting to short sequences, since the higher-order Bahadur models would become prohibitive to generate from. Nevertheless, both the WGEE as well as the MI-GEE methods and then especially also the MI-Transition models can be used, in fact are very appealing, for longer sequences of repeated measures. When sequences become very long, the transition model is appealing owing to its computational convenience.

Denote by t_j the time point at which measurement j is taken and by x_i the treatment indicator. GM I is based on a Bahadur model, which follows general formulation (3), with

$$\text{logit}(\pi_{ij}) = \text{logit}[P(Y_{ij} = 1|x_i, t_j)] = \beta_0 + \beta_x\, x_i + \beta_t\, t_j + \beta_{xt}\, x_i\, t_j, \qquad (6)$$

where we choose $\beta_0 = -0.25$, $\beta_x = 0.5$, $\beta_t = 0.2$ and $\beta_{xt} = -0.8$, with two- and three-way correlation coefficients equal to $\rho_{ij_1j_2} = 0.2$ and $\rho_{ij_1j_2j_3} = 0$, respectively. The latter define an exchangeable correlation structure. The missingness process for GM I is assumed to be MAR, and the probability of dropout at time point j given x_i and the measurement at the previous time point, is modelled by a logistic regression of the form

$$\text{logit}[P(D_i = j|x_i, y_{i,j-1}, D_i \geq j)] = \psi_0 + \psi_x\, x_i + \psi_{prev}\, y_{i,j-1},$$

where $j = 2, 3, 4$, $\psi_0 = -0.5$, $\psi_x = -0.6$ and $\psi_{prev} = -3.5$. Combining this dropout model with the measurement model yields, for GM I, 68% completers, 15% with the last outcome missing (7% for $x = 0$ and 8% for $x = 1$), and 18% with only the first outcome observed (10% for $x = 0$ and 8% for $x = 1$).

The same dropout model is used to generate the missingness for GM II, but now combined with the AR(2) transition model. Such a model can be described as follows:

$$\begin{aligned} P(x_i) &= \mu_x, \\ \text{logit}[P(Y_{i1} = 1|x_i)] &= \alpha_0 + \alpha_x\, x_i, \\ \text{logit}[P(Y_{i2} = 1|x_i, y_{i1})] &= \phi_0 + \phi_x\, x_i + \phi_1 y_{i1} \\ \text{and} \quad \text{logit}[P(Y_{i3} = 1|x_i, y_{i1}, y_{i2})] &= \gamma_0 + \gamma_x\, x_i + \gamma_1\, y_{i1} + \gamma_2\, y_{i2}, \end{aligned}$$

where $\mu_x = 0.5$, $\alpha_0 = -0.2$, $\alpha_x = 0.3$, $\phi_0 = -0.1$, $\phi_x = 0.5$, $\phi_1 = 0.7$, $\gamma_0 = -0.25$, $\gamma_x = 0.35$, $\gamma_1 = 0.4$ and $\gamma_2 = 0.6$. On this generation model, the

missingness proportions are 73% for completers, 11% with the last outcome missing (7% for $x = 0$ and 4% for $x = 1$), and 17% with only the first outcome observed (11% for $x = 0$ and 6% for $x = 1$).

Since the methods of interest, WGEE and MI-GEE, involve marginal models, so as to allow comparison, the above conditional model needs to be further marginalized to obtain so-called marginalized "true" parameters, which then approximately describe a marginal logistic function. This marginalization assumes that the corresponding underlying marginal model is of the form given in (6). Inasmuch as the underlying measurement model is in fact conditional, rather than marginal, there is no way to verify whether this assumed underlying marginal model is "true". This marginalization was done by computing the marginal probabilities from the underlying conditional AR(2) transition model probabilities, i.e., for a given outcome vector and treatment level, $(y_{i1}, y_{i2}, y_{i3}, x_i)$,

$$P(y_{i1}, y_{i2}, y_{i3}, x_i) = P(y_{i3}|x_i, y_{i1}, y_{i2})P(y_{i2}|x_i, y_{i1})P(y_{i1}|x_i)P(x_i). \tag{7}$$

On a hypothetical dataset consisting of all 16 possible combinations of the form $(y_{i1}, y_{i2}, y_{i3}, x_i)$, with corresponding probability weights $P(y_{i1}, y_{i2}, y_{i3}, x_i)$, we fitted a GEE model of the form (6). The resulting marginalized "true" parameters of GM II are $\beta_0 = -0.3658$, $\beta_x = 0.2673$, $\beta_t = 0.2265$ and $\beta_{xt} = 0.0790$.

Finally, for GM III, we assume a Gaussian outcome, W_{ij}, at three time points, where:

$$\mu_{ij} = E(W_{ij}|x_i, t_j) = \eta_0 + \eta_x\, x_i + \eta_t\, t_j + \eta_{xt}\, x_i\, t_j,$$

for $i = 0, 1$ and $j = 1, 2, 3$, with $\eta_0 = 3.5$, $\eta_x = 0$, $\eta_t = 1.75$ and $\eta_{xt} = 0.5$. That is,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{bmatrix} = \begin{bmatrix} (\mu_{01}, \mu_{02}, \mu_{03})' \\ (\mu_{11}, \mu_{12}, \mu_{13})' \end{bmatrix} = \begin{bmatrix} (5.75, 8.00, 10.25)' \\ (5.25, 7.00, 8.75)' \end{bmatrix}.$$

Moreover, we assume the following unstructured covariance structure:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.80 & 0.35 \\ 0.80 & 1 & 0.50 \\ 0.35 & 0.50 & 1 \end{pmatrix}.$$

The missingness process for this GM is given by:

$$\text{logit}[P(D_i = j|x_i, w_{i,j-1}, D_i \geq j)] = \delta_0 + \delta_x\, x_i + \delta_{prev}\, w_{i,j-1},$$

where $j = 2, 3, \ldots, J+1$, $\delta_0 = -0.15$, $\delta_x = 0.8$ and $\delta_{prev} = -0.35$. Combining this dropout model with the measurement model yields, on average, over all the 500 samples generated from GM III, 76% completers, 7% with the last outcome missing (3% for $x = 0$ and 4% for $x = 1$), and 17% with only the first outcome observed (7% for $x = 0$ and 10% for $x = 1$).

The binary outcome Y_{ij} was then obtained from the continuous outcome W_{ij} by defining a cut-off value of 6.5, i.e., $Y_{ij} = 1$, if $W_{ij} \geq 6.5$, and 0, otherwise. Although the generated outcomes are continuous in nature, the focus here is on the analysis of the binary version Y_{ij}. For this reason, we need to obtain "true" parameters corresponding to this dichotomized response by fitting a GEE model of the form (6). Note however that this model is not necessarily the unknown underlying marginal model for the binary outcomes. The resulting parameters are $\beta_0 = -3.0373$, $\beta_x = 0.0095$, $\beta_t = 1.7812$ and $\beta_{xt} = 0.4828$.

Our choice for linear time evolutions, at the scale of the linear predictor and within each of the treatment arms, allows us to distinguish between misspecification effects on cross-sectional parameters (β_0 and β_x), longitudinal parameters (β_t), and parameters combining aspects of both (β_{xt}). In practice, for example in a clinical trial, it might be advisable to allow for an unstructured, saturated treatment-by-time model, reducing the risk of model misspecification and in line with recommendations made by Molenberghs *et al* (2004) and several references listed therein.

3.2 Design of the Simulation Study

We now proceed to describe the details of our simulation study. Given that the sequence of outcomes and the missing data process for GM I and GM II are discrete, quantification of bias under specific assumptions about the nonresponse process can be done via an algorithm first proposed by Rotnitzky and Wypij (1994). This so-called asymptotic simulation method entails first creating a hypothetical dataset consisting of all possible outcome sequences for each level of the covariate(s). In addition, for each of these, there are J possible missingness patterns. The probability mass with which each of these outcome sequences occurs can be computed based on the assumed data-generating model (measurement and dropout models).

For our case, we consider a binary outcome at 3 time points, denoted by $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})'$, and a single covariate x_i with 2 levels, i.e., a binary treatment indicator. This gives rise to $2^3 = 8$ possible sequences at each level of the covariate, yielding a total of 16 possibilities. From the assumed measurement model, the probability masses for each of these 16 sequences can be computed, $P(\boldsymbol{y}_i, x_i)$ say. Now, for each such case, there are 3 possible dropout patterns – dropout at second time point, dropout at the third time point, and no

dropout – yielding a total of 48 possibilities. The probabilities $P(\boldsymbol{y}_i, x_i)$ are thus further split among the 3 missingness patterns according to the dropout probabilities. Specifically, denoting by $P(D_i = 2|D_i \geq 2), P(D_i = 3|D_i \geq 3)$ and $P(D_i = 4|D_i \geq 4)$ the probabilities of dropout at time points 2, 3, and 4, respectively, we obtain:

$$P(\boldsymbol{y}_i, x_i, D_i = 4|D_i \geq 4) = P(\boldsymbol{y}_i, x_i) \prod_{j=2}^{4} [1 - P(D_i = j|D_i \geq j)],$$

$$P(\boldsymbol{y}_i, x_i, D_i = 3|D_i \geq 3) = P(\boldsymbol{y}_i, x_i) \prod_{j=2}^{3} [1 - P(D_i = j|D_i \geq j)] P(D_i = 4|D_i \geq 4)$$

and

$$P(\boldsymbol{y}_i, x_i, D_i = 2|D_i \geq 2) = P(\boldsymbol{y}_i, x_i)[1 - P(D_i = 2|D_i \geq 2)] \prod_{j=3}^{4} P(D_i = j|D_i \geq j).$$

The estimating equations are then applied to this hypothetical dataset with the application of the resulting probability weighting. The solutions obtained are the limiting (i.e., asymptotic) solutions, which can then be compared with the known parameters of the simulation model, so as to conveniently derive the asymptotic bias of the estimators.

For the small-sample simulations, we assume a sample of size $N = 100$ subjects, equally divided between the two treatment groups. Such a choice is directly applicable to practitioners, since many biopharmaceutical trials employ about 50 to 100 patients per treatment arm. Based on the underlying probabilities from GM I or GM II, 50 observations were generated randomly for each treatment group. $S = 500$ such samples were then generated. Similarly, for GM III, we generated $S = 500$ samples with $n_0 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $n_1 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. While asymptotic simulations were conducted only for GM I and GM II, small-sample simulations were done for all three generation models. When using a GEE approach for analysis, the same working correlation structure as assumed during data generation is employed.

3.3 Results of the Simulation Study

For the ensuing discussion, in assessing and comparing WGEE and imputation-based GEE, various properties are quantified. First, we define bias as the difference between the estimate and the true value of the parameter, i.e., $\mathrm{Bias}(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. For the asymptotic version, probability weights, computed from the underlying GM, are applied in solving the estimating equations (Rotnitzky and Wypij, 1994). The resulting estimates are the limiting solutions, which can then be used to compute the asymptotic bias (Bias_∞), while the resulting variances are the asymptotic variances (Var_∞) of the parameter estimators.

For the small-sample simulations, the average ($\overline{\mathrm{Est}}$) of the estimators over all

$S = 500$ samples, its true variance for a sample of size N (Var_N), its estimated variance for a sample of size N ($\widehat{\mathrm{Var}}_N$) and MSE are computed as:

$$\overline{\mathrm{Est}} \equiv \overline{\widehat{\boldsymbol{\beta}}} = \sum_{i=1}^{S} \frac{\widehat{\boldsymbol{\beta}}_i}{S},$$

with

$$\mathrm{Var}_N \equiv \mathrm{Var}_N(\overline{\mathrm{Est}}) = \frac{\mathrm{Var}_\infty}{N}, \qquad \widehat{\mathrm{Var}}_N \equiv \widehat{\mathrm{Var}}_N(\overline{\mathrm{Est}}) = \sum_{i=1}^{S} \frac{(\widehat{\boldsymbol{\beta}}_i - \overline{\widehat{\boldsymbol{\beta}}})^2}{S-1}$$

and

$$\mathrm{MSE} \equiv \mathrm{MSE}(\overline{\mathrm{Est}}) = \mathrm{Bias}_N^2(\overline{\mathrm{Est}}) + \widehat{\mathrm{Var}}_N(\overline{\mathrm{Est}}).$$

3.3.1 Everything Correctly Specified

We first investigate the individual merits of each method when every one of its aspects is correctly specified. Recall that GM I is based on a Bahadur measurement model and a logistic model for dropout that is reflective of an MAR mechanism, i.e., depending on the previous measurement as well as the treatment indicator. An appropriate analysis model would consist of a measurement model and a dropout model that match those of this GM. Since GEE methods are moment-based versions of the Bahadur model (Section 2.3), a GEE-based version, with the same structure as that of the underlying measurement model would be suitable. To address the MAR nature of the missingness, the GEE-based approach is supplemented with a weighting scheme, obtained from a model of the same form as that of the underlying dropout model, resulting now in WGEE. Thus, WGEE was fitted for GM I, using weights taken from fitting a logistic dropout model with the treatment indicator and the previous measurement as predictors. It should be noted that under WGEE the imputation model is not relevant since the missingness is addressed, not by imputation, but rather, by means of the dropout model. The results for both the asymptotic and small-sample simulations are shown in Table 1.

It can be observed that the asymptotic unbiasedness of the WGEE estimators under a correctly specified mean structure is demonstrated by our asymptotic simulation. The same cannot be said, however, for the small-sample simulation, under which a substantial amount of bias is observed. Moreover, the estimated variances of the parameter estimators are considerably larger than the true variances, demonstrating the inefficiency of WGEE for small samples. These observations are indicative that, for a sample of size $N = 100$, the consistency of the WGEE estimators does not seem to be achieved, at least not for this particular generating model.

For GM II, which uses an AR(2) transition model for the mean structure and a

Table 1
Asymptotic and small-sample simulation results for WGEE, with everything correctly specified, under GM I. Asymptotic results include asymptotic bias (Bias_∞) and asymptotic variance (Var_∞), while small-sample simulation results include the average ($\overline{\text{Est}}$), bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
β_0	-1.87E-06	0.44095	-0.6457	-0.3956	1.0779	0.0044	1.2345
β_x	1.99E-07	1.10959	0.6225	0.1225	2.1108	0.0111	2.1258
β_t	2.02E-07	0.11942	0.3018	0.1018	0.2388	0.0012	0.2491
β_{xt}	-1.66E-07	0.27815	-0.9355	-0.1356	0.4441	0.0028	0.4625

conditional logistic model for dropout, we considered fitting an AR(2) transition model, which is consistent with the underlying measurement model, after multiple imputation (MI-Transition). The multiple imputations were carried out with the SAS procedure MI, which employs a conditional logistic imputation model for binary outcomes, a model in line with the underlying measurement model of GM II and fully parametric, admitting valid inferences under MAR (Schafer, 2003). Thus, our analysis model, both the imputation as well as the measurement models, are correctly specified in the sense that they are compatible with the underlying measurement model. Note also that a dropout model need not be defined for this mode of analysis, since imputations, rather than dropout weights, are used to cope with the missingness. For the asymptotic simulation, $M = 500$ datasets were imputed, while for the small-sample simulations, since efficient results can be obtained even under a small number of imputations (Rubin, 1987), we chose a more practically relevant value of $M = 5$. Table 2 gives the results for both types of simulations.

The first panel shows asymptotically unbiased parameter estimates, since, for this outcome, data for all subjects are assumed available and are thus not imputed. The small-sample simulations for this outcome, on the other hand, show slightly biased estimates as can be expected whenever taking finite samples. For the second and third panels, some bias is observed, asymptotically and for small samples, but the amounts are generally of small magnitudes. Some degree of difference can also be observed between the estimated and true variances, pointing to a slight inefficiency of the estimators. This might be attributed to the fact that, when applying multiple imputation, small-sample behaviour stems from both the actual sample size, N, as well as from the number of imputations, M. Thus, in cases where the former is large while the latter is relatively small, it should not come as a surprise that the estimated variance is relatively large.

Table 2
Asymptotic and small-sample simulation results for MI-Transition, with everything correctly specified, under GM II. Asymptotic results include asymptotic bias (Bias_∞) and asymptotic variance (Var_∞), while small-sample simulation results include the average ($\overline{\text{Est}}$), bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
α_0	-0.0000	8.0803	-0.2313	-0.0313	0.0925	0.0808	0.0935
α_x	-0.0000	16.1003	0.3369	0.0369	0.1791	0.1610	0.1805
ϕ_0	-0.0096	12.0926	-0.0683	0.0317	0.2046	0.1209	0.2056
ϕ_x	-0.0666	18.0194	0.5041	0.0041	0.2635	0.1802	0.2635
ϕ_1	0.0343	18.1493	0.7241	0.0241	0.2692	0.1815	0.2698
γ_0	0.0236	17.4438	-0.1702	0.0798	0.3472	0.1744	0.3535
γ_x	-0.0568	18.5632	0.3590	0.0090	0.3023	0.1856	0.3024
γ_1	-0.0594	19.7766	0.5029	-0.0971	0.2354	0.1978	0.2448
γ_2	0.0072	18.9333	0.4382	0.0382	0.3249	0.1893	0.3264

Finally, we consider GM III, which is based on a Gaussian measurement model and a logistic dropout model. The analysis model used for this GM was MI-GEE, which requires an imputation model and a measurement model, but not a dropout model. Multiple imputations of the missing Gaussian outcomes were first obtained using a Gaussian imputation model, thereby ensuring a correctly specified imputation model, that is, one that uses the underlying measurement process to generate the imputations for the missing observations. The Gaussian outcome was then dichotomized based on the previously defined cutoff value, after which standard GEE, using a probit link, was applied to the dichotomized outcome of the imputed datasets. Since the underlying distribution for the outcomes is not discrete, only small-sample simulations are possible. Although initially $S = 500$ samples were generated, after dichotomization of the Gaussian outcome, there were 51 samples for which convergence was not attained. Inspection of these samples showed that the treatment-by-time interaction could not be estimated because, at one time point, all dichotomized outcomes belonged to only one treatment group.

Table 3 gives the results of the simulation only for the $S' = 449$ convergent samples. The "true" parameter values used to compute the bias were obtained by fitting the same measurement model using the complete (binary) data from the $S' = 449$ samples. Consistent with the theory on MI, we obtained only very small bias for the estimates, which might be expected to decrease even

Table 3
Small-sample simulation results for MI-GEE, with everything correctly specified, under GM III. Results include the average ($\overline{\text{Est}}$), bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$) and mean squared error (MSE), of the parameter estimators, for $N = 100$.

	Small-Sample			
Parameter	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	MSE
β_0	-3.0358	0.0015	0.1978	0.1978
β_x	0.0151	0.0056	0.3968	0.3968
β_t	1.7808	-0.0004	0.0601	0.0601
β_{xt}	0.4767	-0.0061	0.1480	0.1481

further under larger samples.

3.3.2 Dropout and Measurement Models Correct, Imputation Model Incorrect

We now consider a comparison between WGEE and MI-GEE, both having a correctly specified measurement model, but the latter using an incorrectly specified imputation model and the former specifying the dropout model correctly. For the two methods, the measurement model used is consistent with the underlying Bahadur measurement model of GM I. We know from the discussion in the previous section that fitting WGEE for GM I, using the same mean structure as that of the underlying measurement model and with weights obtained from a logistic dropout model with the treatment indicator and the previous measurement as predictors, ensures every aspect is correctly specified. For MI-GEE, imputations are done using a conditional logistic imputation model for binary outcomes – a model that is *not* consistent with the marginal nature of the underlying Bahadur measurement model and is, therefore, incorrectly specified. Thus, the said comparison, of WGEE with correctly specified dropout and measurement models against MI-GEE with correctly specified measurement model but incorrectly specified imputation model, is possible under GM I. The results are given in Table 4.

As was already noted in the previous section, WGEE does not yield unbiased and consistent estimators for the particular sample size used, whereas the bias is considerably smaller for MI-GEE. The latter also leads to more precise estimators than those obtained for WGEE, as evidenced by smaller differences between the estimated and true variances for MI-GEE, despite the fact that the WGEE analysis model used was entirely correctly specified. Moreover, comparison of the MSEs indicate more efficient estimators for MI-GEE. All of these observations suggest a certain amount of robustness of MI-GEE when misspecifying the imputation model.

Table 4
Small-sample simulation results for WGEE, with correctly specified dropout, and MI-GEE, with incorrectly specified imputation model, under GM I. Results include the bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.

	WGEE				MI-GEE			
Parm	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
β_0	-0.3956	1.0779	0.0044	1.2345	-0.0169	0.2332	0.1896	0.2335
β_x	0.1225	2.1108	0.0111	2.1258	0.0195	0.4835	0.3938	0.4839
β_t	0.1018	0.2388	0.0012	0.2491	0.0088	0.0548	0.0414	0.0548
β_{xt}	-0.1356	0.4441	0.0028	0.4625	-0.0058	0.1172	0.0885	0.1172

3.3.3 Imputation and Measurement Models Correct, Dropout Model Incorrect

Whereas in the previous section the relative performances of WGEE with correctly specified dropout and MI-GEE with incorrectly specified imputation model were compared, in this section we proceed to look at the reverse. That is, we consider a comparison of WGEE with incorrectly specified dropout model against MI-GEE with correctly specified imputation model. In both cases, the measurement model corresponds to the assumed underlying measurement model for the dichotomized version of the continuous response. For this assessment, we apply the methods under GM III. We have seen, in Section 3.3.1, that for GM III, imputing the missing observations using a Gaussian imputation model and subsequently fitting standard GEE to dichotomized outcomes of the completed sets of data, results in MI-GEE with everything correctly specified. To enable comparison with WGEE using an incorrectly specified dropout model, we obtain weights from a logistic dropout model with the treatment indicator and the binary version of the previous measurement as predictors. The latter is a clear misspecification in the dropout model, since the underlying dropout model uses the continuous form of the previous measurement as predictor. The results of this comparison are given in Table 5. Only small-sample simulations are possible since the underlying GM does not consist of a discrete set of outcomes.

Bias is much smaller for MI-GEE, which can be expected as this is a correctly specified analysis model. With respect to the estimated precision ($\widehat{\text{Var}}_N$) for a sample of size $N = 100$, the estimators obtained from MI-GEE are superior to those from WGEE. Overall, the MI-GEE estimators are more efficient, with MSEs for the WGEE estimators about 1.5 times those of MI-GEE. These results seem to highlight the sensitivity of WGEE to misspecifications in the dropout model, in contrast to MI-GEE, which was noted to be somewhat robust to misspecifications in the imputation model.

Table 5
Small-sample simulation results for WGEE, with incorrectly specified dropout, and MI-GEE, with correctly specified imputation model, under GM III. Results include the bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.

	WGEE			MI-GEE		
Parm	Bias_N	$\widehat{\text{Var}}_N$	MSE	Bias_N	$\widehat{\text{Var}}_N$	MSE
β_0	-0.1855	0.3113	0.3457	0.0015	0.1978	0.1978
β_x	-0.1380	0.5644	0.5834	0.0056	0.3968	0.3968
β_t	0.3100	0.1376	0.2336	-0.0004	0.0601	0.0601
β_{xt}	0.0367	0.2312	0.2325	-0.0061	0.1480	0.1481

3.3.4 Imputation and Dropout Models Correct, Measurement Model Incorrect

We finally proceed to looking at a comparison between WGEE and MI-GEE when the measurement model is specified incorrectly. For this setting, we consider GM II. We first present the results of the asymptotic and small-sample simulations for the marginalized version of MI-Transition, with which WGEE and MI-GEE are subsequently compared. Recall that the resulting parameter estimates, from the correctly specified MI-Transition model fitted in Section 3.3.1 (Table 2), define three sets of conditional probabilities, from which marginal probabilities can be derived as in (7). These estimated probabilities were then used as weights in fitting a GEE model of the form (6) on a dataset consisting of all possible combinations of outcome sequences and treatment level, yielding the marginalized version of MI-Transition. The resulting parameter estimates were subsequently compared with the "marginal" parameters in (6); the results are shown in Table 6. Asymptotic bias for the parameter estimates is generally small, while its small-sample counterpart is larger. Estimated and true variances for a sample of size $N = 100$ differ substantially, indicating some degree of inefficiency under this sample size.

Assuming now that these "marginal" parameters define some underlying marmarginal model for GM II, we fit both WGEE and MI-GEE, with a correctly specified dropout model and a correctly specified imputation model, respectively. For WGEE, weights are obtained from a dropout model consistent with the underlying dropout model of GM II, while for MI-GEE, imputations are generated from a conditional AR(2) transition model, which is in line with the underlying measurement model of GM II. In this way, both the dropout and imputation models are correctly specified. However, the fitted measurement models for both WGEE and MI-GEE are clearly misspecified, in the sense that the outcomes are modelled marginally (i.e., GEE), rather than conditionally (i.e., AR(2)). The results of this comparison (Table 7) indicate less biased estimates for MI-GEE and marginalized MI-Transition. In addition,

Table 6
Asymptotic and small-sample simulation results for marginalized MI-Transition, with everything correctly specified, under marginalized GM II. Asymptotic results include asymptotic bias (Bias_∞) and asymptotic variance (Var_∞), while small-sample simulation results include the average ($\overline{\text{Est}}$), bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
β_0	-0.0045	1.11659	-0.4253	-0.0595	1.1230	0.0112	1.1265
β_x	0.0285	2.39565	0.3134	0.0461	2.4702	0.0240	2.4723
β_t	-0.0022	0.20405	0.2644	0.0379	0.2060	0.0020	0.2074
β_{xt}	-0.0363	0.43727	0.0648	-0.0142	0.4524	0.0044	0.4526

Table 7
Small-sample simulation results for WGEE, with correctly specified dropout, and MI-GEE, with correctly specified imputation model, under GM II. Results include the bias (Bias_N), estimated variance ($\widehat{\text{Var}}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.

	WGEE				MI-GEE			
Parm	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
β_0	-0.4223	1.1310	0.0047	1.3098	-0.0562	0.2508	0.1901	0.2539
β_x	-0.1451	2.9804	0.0104	3.0014	0.0530	0.4927	0.3841	0.4955
β_t	0.1241	0.2149	0.0014	0.2303	0.0343	0.0608	0.0414	0.0620
β_{xt}	0.0792	0.5877	0.0030	0.5940	-0.0233	0.1184	0.0847	0.1190

MI-GEE outperforms both WGEE and marginalized MI-Transition in terms of precision and efficiency.

4 Concluding Remarks

When the analysis of incomplete binary longitudinal data is envisaged, several routes are available. Apart from likelihood-based methods, such as the generalized linear mixed-effects model (Molenberghs and Verbeke, 2005), non-likelihood methods are attractive, especially when a so-called marginal model is of interest. Since standard generalized estimating equations (Liang and Zeger, 1986) are unbiased only under MCAR, a variety of modifications and alternatives to GEE have been proposed. Undoubtedly the most popular route is through weighted estimating equations, proposed by Robins *et al* (1995),

and a number of later extensions. Also of attraction is a combination of GEE and multiple imputation (Rubin, 1987) methods, i.e., MI-GEE. Once multiple imputation is considered an option, it has the merit of allowing for a variety of imputation techniques, whereafter several analysis methods can be considered. Two such routes considered in this paper are MI-GEE and MI-Transition.

In this paper we have provided quantitative evidence, based on asymptotic, as well as small-sample, simulations, that can be usefully applied in the decision-making process. We have considered WGEE, MI-GEE, and MI-Transition under a variety of scenarios. While simulations are necessarily limited, we believe both methods have been put to the test in a fair fashion. Although asymptotically WGEE exhibits the desirable properties that it theoretically is known to possess, these are barely reproduced for small samples, even when every aspect of the analysis is correctly specified. Moreover, the observed sensitivity of WGEE to misspecification in either the dropout or measurement model renders these asymptotic properties meaningless. MI-GEE and MI-Transition, on the other hand, demonstrate a certain degree of robustness to misspecification in either the imputation or measurement model, this, despite a further marginalization for the MI-Transition case. Furthermore, WGEE's applicability to the case where also covariates are missing is less straightforward, while application of MI is relatively easy. Moreover, one can do MI under MAR with intermittent missing data. Although the results of this study provide insight about the methods under consideration, and thus are useful in the decision-making process, whenever inference is critical, it is always wise to try a couple of different methods, by way of sensitivity analysis.

In view of previous work on the merits of inverse probability weighting methods *versus* multiple imputation (the discussion of Scharfstein, Rotnitzky, and Robins, 1999; Clayton *et al*, 1998; Carpenter, Kenward, and Vansteelandt, 2006), we now compare our findings with theirs. Clayton *et al* (1998) investigated the use of inverse probability weighting (IPW) and multiple imputation, among others, in the context of longitudinal binary data in a multi-phase sampling setting. They found that, while IPW was inefficient for such a 2×2-phase design, MI showed remarkable efficiency. Moreover, this, along with possible extension to data arising from other designs, indicates the substantial strengths of MI. Carpenter, Kenward, and Vansteelandt (2006), on the other hand, used simulations to study a so-called doubly-robust IPW estimator, introduced by Scharfstein, Rotnitzky, and Robins (1999), in comparison with standard IPW, maximum likelihood, and MI. The doubly-robust IPW estimator is a modified version of the usual IPW, proposed to improve the efficiency of IPW estimators. IPW estimators were again found to be inefficient and sensitive to the choice of the weight model, but the doubly-robust version proves to be as efficient as MI and robust to misspecification. Although applied to continuous Gaussian data, they expect the results to generalize to the discrete case. Whereas Clayton *et al* (1998) used actual data and Carpenter,

Kenward, and Vansteelandt (2006) used simulations of a small-sample nature, we complement a small-sample simulation study with asymptotic simulations. Through our simulations, we reinforce the strength of MI over IPW, specifically in application to GEE. WGEE can be viewed as a type of IPW scheme that uses as weights the inverse of the probability of dropout (taken from some dropout model), while MI-GEE uses imputations for the missing data. WGEE was found to be inefficient for small-samples, in line with the findings of these two papers regarding the inefficiency of such IPW schemes. However, this (lack of) efficiency might well be addressed by adopting the doubly-robust IPW version in obtaining the WGEE solutions.

Misspecifications are common in practice and it is seldom the case that one would have an entirely correctly specified analysis model. This, along with the fact that the nice properties of WGEE are not attained for modest sample sizes, which is common in typical clinical trials, discourages its recommendation. On the other hand, although theoretically MI-GEE does not provide consistent results when there is a misspecification, overall, it still yields more precise estimates than WGEE.

Thus, we provided evidence for the important fact that MI-GEE is less biased and more precise in small and moderate samples, in spite of the asymptotic unbiasedness of WGEE. As a consequence, in practice, MI-GEE would be the preferred method for analysis over WGEE. Moreover, although the focus of this paper is on missingness in the response, in real-life settings, missingness in covariates is often encountered. In such cases, the choice for MI-GEE is even more convincing, since the use of WGEE would be ruled out. Finally, with MI, the imputation model is not restricted to the use of covariates that will be conditioned upon in the measurement model. Other covariates that are available, without necessarily being of interest in the measurement model, can be incorporated in the imputation model, thereby yielding presumably better imputations as well as wider applicability.

Importantly, it ought to be clear that in the case of the conditionally specified model, a so-called direct likelihood approach, exploiting ignorability results, is a very viable alternative and may well be the user's preferred one. However, we wanted to focus on a comparison between inverse probability weighting methods and multiple imputation. Hence, not to overly clutter the simulation setting, we have left direct likelihood out of the picture. Additionally, direct likelihood would *not* apply to the marginal model settings, given the prohibitive nature of fitting such models as the Bahadur in other than the simplest settings.

As a final remark, recall that asymptotic simulations were done to obtain the asymptotic bias and asymptotic variance. These have theoretical use only, and may provide guidance as to what happens in large to very large samples.

Supplementing them with small-sample simulations is therefore an attractive route. Needless to say the method is of no use with conventional data analysis.

Acknowledgments

The authors gratefully acknowledge support from *Fonds Wetenschappelijk Onderzoek-Vlaanderen* Research Project G.0002.98 "Sensitivity Analysis for Incomplete and Coarse Data" and from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Agresti, A., 2002. Categorical Data Analysis (2nd ed.). New York: John Wiley & Sons.

Bahadur, R.R., 1961. A representation of the joint distribution of responses to n dichotomous items. In: Studies in Item Analysis and Prediction, H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

Carpenter, J.R., Kenward, M.G., Vansteelandt, S., 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J. Roy. Statist. Soc. Ser. A 3, 571–584.

Clayton, D., Spiegelhalter, D., Dunn, G., Pickles, A., 1998. Analysis of longitudinal binary data from multi-phase sampling (with discussion). J. Roy. Statist. Soc. Ser. B 60, 71–87.

Cox, D.R., 1972. The analysis of multivariate binary data. Appl. Statist. 21, 113–120.

Dale, J.R., 1986. Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics 42, 909–917.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y., Zeger, S.L., 2002. Analysis of Longitudinal Data (2nd ed.). Oxford Science Publications. Oxford: Clarendon Press.

Diggle, P.J., Kenward, M.G., 1994. Informative drop-out in longitudinal data analysis (with discussion). Applied Statistics 43, 49–93.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Heidelberg: Springer-Verlag.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C., 2006. Analyzing incomplete binary longitudinal clinical trial data. Statist. Sci. 21, 52–69.

Kenward, M.G., Molenberghs, G., 1998. Likelihood based frequentist inference when data are missing at random. Statist. Sci. 13, 236–247.

Lang, J.B., Agresti, A., 1994. Simultaneously modelling joint and marginal distributions of multivariate categorical responses. J. Amer. Statist. Assoc. 89, 625–632.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Little, R.J.A., 1993. Pattern-mixture models for multivariate incomplete data. J. Amer. Statist. Assoc. 88, 125–134.

Little, R.J.A., 1994. A class of pattern-mixture models for normal incomplete data. Biometrika 81, 471–483.

Little, R.J.A., 1995. Modeling the drop-out mechanism in repeated measures studies. Journal of the American Statistical Association 90, 1112–1121.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. New York: Wiley.

Meng, X.L., 1994. Multiple-imputation inferences with uncongenial sources of input (with discussion). Statist. Sci. 10, 538–573.

Molenberghs, G., Kenward, M.G., 2007. Handling Incomplete Data From Clinical Studies. New York: John Wiley.

Molenberghs, G., Kenward, M.G., Lesaffre, E., 1997. The analysis of longitudinal ordinal data with non-random dropout. Biometrika 84, 33–44.

Molenberghs, G., Lesaffre, E., 1994. Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. J. Amer. Statist. Assoc. 89, 633–644.

Molenberghs, G., Lesaffre, E., 1999. Marginal modelling of multivariate categorical data. Statist. Med. 18, 2237–2255.

Molenberghs, G., Thijs, H., Jansen, I, Beunckens, C., Kenward, M.G., Mallinkrodt, C., Carroll, R.J., 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics 5, 445–464.

Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. New York: Springer-Verlag.

Plackett, R.L., 1965. A class of bivariate distributions. J. Amer. Statist. Assoc. 60, 516–522.

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

Rotnitzky, A., Wypij, D., 1994. A note on the bias of estimators with missing data. Biometrics 50, 1163–1170.

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

Rubin, D.B., 1978. Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In: Imputation and Editing of Faulty or Missing Survey Data. Washington, DC: U.S. Department of Commerce, pp. 1–23.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Schafer, J.L., 1999. Multiple imputation: a primer. Statist. Methods Med. Res. 8, 3–15.

Schafer, J., 2003. Multiple imputation in multivariate problems when the imputation and analysis models differ. Statist. Neerl. 57, 19–35.

Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with comments). J. Amer. Statist. Assoc. 94, 1096–1146.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

Wu, M.C., Bailey, K.R., 1989. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. Biometrics 45, 939–955.

Wu, M.C., Carrol, R.J.. 1988. Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. Biometrics 44, 175–188.