A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to non-ignorable non-monotone missingness

Peer-reviewed author version

Troxel, Andrea B.; Lipsitz, Stuart R.; Fitzmaurice, Garrett M.; IBRAHIM, Joseph; Sinha, Debajyoti & MOLENBERGHS, Geert (2010) A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to non-ignorable non-monotone missingness. In: STATISTICS IN MEDICINE, 29(14). p. 1511-1521.

DOI: 10.1002/sim.3867
Handle: http://hdl.handle.net/1942/11048

A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to nonignorable non-monotone missingness

Andrea B. Troxel[1,*], Stuart R. Lipsitz[2], Garrett M. Fitzmaurice[2], Joseph G. Ibrahim[3], Debajyoti Sinha[4], and Geert Molenberghs[5]

[1] *University of Pennsylvania School of Medicine, Philadelphia, PA, U.S.A.*
[2] *Harvard Medical School, Boston, MA, U.S.A.*
[3] *University of North Carolina, Chapel Hill, NC, U.S.A.*
[4] *Florida State University, Tallahassee, FL, U.S.A.*
[5] *Hasselt University, Diepenbeek, Belgium*

SUMMARY

For longitudinal binary data with non-monotone non-ignorably missing outcomes over time, a full likelihood approach is complicated algebraically, and with many follow-up times, maximum likelihood estimation can be computationally prohibitive. As alternatives, two pseudo-likelihood approaches have been proposed that use minimal parametric assumptions. One formulation requires specification of the marginal distributions of the outcome and missing data mechanism at each time point, but uses an "independence working assumption," i.e., an assumption that observations are independent over time. Another method avoids having to estimate the missing data mechanism by formulating a "protective estimator." In simulations, these two estimators can be very inefficient, both for estimating time trends in the first case and for estimating both time-varying and time-stationary effects in the second. In this paper, we propose use of the optimal weighted combination of these two estimators, and in simulations we show that the optimal weighted combination can be much more efficient than either estimator alone. Finally, the proposed method is used to analyze data from two longitudinal clinical trials of HIV-infected patients.

1. INTRODUCTION

Longitudinal studies in which each subject is to be observed at a fixed number of time points have become very popular in social science and medical applications. For example, longitudinal data are often collected in AIDS, cardiovascular, and cancer clinical trials and observational studies. We focus on the case where the response variable over time is binary (e.g., success

*Correspondence to: Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 423 Guardian Drive, 632 Blockley Hall, Philadelphia, PA 19104, U.S.A.; atroxel@mail.med.upenn.edu

or failure) and are interested in modeling the marginal means or success probabilities; this setting has been well-described [1, 2, 3, 4, 5]. Such modeling is often complicated by the fact that in longitudinal studies, the outcome is not always observed at all assessment times. In addition, this missing data is often non-ignorable [6], since the probability that an outcome is missing at a given time can depend on the potentially missing value of the outcome at that time. The missing outcome data must be properly accounted for in the analysis, and numerous approaches have been proposed [7, 8, 9, 10, 11]. In clinical trials, an individual's response is often missing at one follow-up time but observed at the next follow-up time, resulting in a large class of distinct missingness patterns, often called "non-monotone" missingness. We will, however, assume that all subjects have the outcome measured at the first time point; e.g., to be part of the study, the subject must be seen at baseline.

An example of a data set with this structure comes from two longitudinal clinical trials of HIV-infected patients sponsored by the AIDS Clinical Trials Group (ACTG): ACTG 116A [12] and 116B/117 [13]. The two studies were randomized phase III double-blind trials, designed to compare two treatments, zidovudine (AZT) and didanosine (ddI); they differed with respect to the length of prior treatment with AZT and have been used in several combined analyses [14]. The response of interest is normal CD4 cell count ($>$ 200 cells per cubic millimeter) versus abnormal CD4 cell count ($\leq$ 200) measured at baseline (week 0) and every week for up to 5 weeks from baseline. The cutoff of 200 was initially chosen because of its strong predictive value for development of opportunistic infections and has been adopted as a standard threshold of clinical importance [15]. Previously, we analyzed these data for HIV patients with and without AIDS [16]; here we consider only the 431 patients with AIDS at baseline. The main question of scientific interest is the effect of treatment on changes in CD4 cell count sufficiency over time.

As with most longitudinal studies, missing outcome data over time complicate the analysis. For example, fewer than 50% of the patients (202/431= 46.9%) have outcomes measured at all 6 occasions.

Table I shows the number of subjects seen at each of the six possible occasions. In Table I, we see that 383 of the 431 patients (88.9%) had a measurement at week 1; the percentage of patients seen slowly drops until 285 (66.1%) of the 431 patients are seen at week 5. A majority of the missing data is due to patients who drop out, i.e., once the patient misses a scheduled visit, no further measurements of the response variable are obtained. However, there are 109 (25.3%) patients who missed at least one measurement, but returned for a later measurement. In this setting, it is quite plausible that patients with high or normal CD4 counts are more likely to miss the scheduled study visits. If this is true, then missingness depends on the unobserved outcome of interest and is nonignorable. Indeed, some have argued that the only plausible non-monotone missing at random mechanisms are those that derive from randomized monotone missingness processes [17, 18]. In the longitudinal data setting, these processes require that missingness at an assessment depends on the prior assessment if and only if the prior assessment is observed; such processes are thus highly implausible here.

To formulate a full likelihood for non-ignorable non-monotone missing outcomes over time, one must specify a joint distribution for the T repeated binary outcomes of interest, of dimension 2^T, and a model for the missingness mechanism. To estimate the parameters, a full likelihood approach has many nuisance parameters and is complicated algebraically; furthermore, estimation can be computationally prohibitive, especially when the number of times is large. As alternatives to a full likelihood procedure, two pseudo-likelihood [19, 20] procedures have been proposed by Troxel et al. [21] and Fitzmaurice et al. [22] under minimal

parametric assumptions.

First, Troxel et al. [21] proposed a pseudo-likelihood that is formed by an "independence working assumption," i.e., assuming for the purpose of estimation that the longitudinal binary measurements are independent over time and ultimately applying a robust "sandwich" variance estimate [23] to achieve proper inference. Specifically, their pseudo-likelihood first assumes a marginal logistic regression model for the outcome at each time point; it also assumes that the missingness probability at a given time depends only on the possibly missing response at that time and the covariates (the covariates are assumed to be fully observed). The chief attraction of this pseudo-likelihood approach is that it substantially eases the numerical complexities of the full likelihood approach by reducing high-dimensional sums to sums of a single dimension. Further, it alleviates the need to specify and estimate many nuisance parameters that are needed in a full likelihood approach. In addition, asymptotically unbiased estimators of the regression parameters and missingness parameters can be obtained. However, by assuming independence of repeated measures across measurement occasions, the method can be highly inefficient for estimating the regression parameters. For example, results from Table 1 of Troxel et al. [21] indicate that their pseudo-likelihood method can be very inefficient compared to the MLE, and in particular in estimating time trends.

Alternatively, Fitzmaurice et al. [22] proposed a pseudo-likelihood based on the idea of formulating a "protective estimator" [24] without having to estimate the missing data mechanism. Specifically, they assume that the baseline response is fully observed and that the probability that a response is missing at any future occasion is conditionally independent of the baseline response given the response at that occasion. This assumption ensures that the conditional distributions of the outcome at time 1 given the outcome at any future time

are fully identifiable. These conditional distributions are functions only of the parameters of primary scientific interest (the regression parameters), and not the parameters of the missing data mechanism. Their pseudo-likelihood is based on the conditional distributions of the baseline response, given the response at each future occasion, for estimation of the regression parameters. The pseudo-likelihood requires only specification of the bivariate distribution of the outcome at time 1 and any future time, and is thus computationally much more feasible than maximum likelihood. The resulting parameter estimates are asymptotically unbiased when the identifying assumption holds. However, the results of their simulation study showed that, with high correlation, the "protective estimator" can be highly efficient for estimating time trends, but inefficient for estimating the effects of time-stationary covariates.

Since the estimate from Troxel et al.'s [21] approach is very inefficient for estimating time trends, and the estimate from Fitzmaurice et al.'s [22] approach is very inefficient for estimating time-stationary effects, this suggests formulating a new estimator of the marginal regression parameters that is a combination of these two estimators. In this paper, we propose forming a new estimator that is the asymptotic minimum variance linear combination of the two estimators [25, 26]. The new estimator is basically a weighted least squares estimate, where the weight matrix is the inverse of the estimated asymptotic covariance matrix of the vector formed from concatenating the Troxel et al. and Fitzmaurice et al. estimates. This estimated asymptotic covariance matrix is obtained using the "sandwich" variance estimator of White [23].

The remainder of the paper is organized as follows. In Section 2, we describe the underlying data models and introduce the necessary notation. In Section 3, we review the pseudo-likelihoods of Troxel et al. and Fitzmaurice et al., and our proposed weighted combination

of the two. Section 4 illustrates the methods with the AIDS example. In Section 5, we present results from our simulation study, showing that our proposed estimator produces much more efficient estimates of the time trends than the Troxel et al. estimator, and much more efficient estimates of the time-stationary effects than the Fitzmaurice et al. estimator.

2. UNDERLYING DATA MODEL

We assume that n independent subjects are to be observed at a fixed set of T occasions, $t = 1, \ldots, T$. For the i^{th} individual ($i = 1, \ldots\ldots, n$), we can form a $T \times 1$ vector, $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{iT}]'$, where the binary random variable Y_{it} equals 1 if the i^{th} individual has response 1 (e.g., "success") at time t, and 0 otherwise. Each individual also has a $J \times 1$ covariate vector $\mathbf{x}_{it}$; we assume that all covariates are fully observed. The main interest here is in the marginal model for each binary outcome Y_{it}, which we assume follows a logistic regression. The marginal distribution of Y_{it} is Bernoulli with success probability

$$p_{it} = p_{it}(\beta) = E(Y_{it}|\mathbf{x}_{it}, \beta) = pr(Y_{it} = 1|\mathbf{x}_{it}, \beta) = \frac{\exp(\mathbf{x}'_{it}\beta)}{1 + \exp(\mathbf{x}'_{it}\beta)}. \quad (1)$$

In a marginal model, the goal is to make inferences about the marginal regression parameters β, whereas the within-subject association among the repeated responses is regarded as a nuisance characteristic of the data. Although the association model is not even specified in the pseudo-likelihood of Troxel et al., the pairwise associations between the outcome at time 1 and the follow-up times must be correctly specified in the protective pseudo-likelihood of Fitzmaurice et al. to obtain consistent estimates. Thus, we briefly discuss the association model here.

The association between a pair of binary outcomes is typically measured in terms of marginal odds ratios [28] or marginal correlations [29]. For ease of exposition, as well as to be compatible with the original protective pseudo-likelihood of Fitzmaurice et al., here we discuss marginal

correlations. In general, we propose a generalized autoregressive(1)-type correlation structure. For two different points in time $s \neq t$, the generalized autoregressive(1) model states that

$$\rho_{ist} = Corr(Y_{is}, Y_{it}|\mathbf{x}_i) = \rho^{|t-s|^\theta},$$

where $-1 < \rho < 1$ and $-\infty < \theta < \infty$. Note that if $\theta = 0$, this correlation reduces to an exhangeable correlation, $\rho_{ist} = \rho$, and if $\theta = 1$, this correlation reduces to the usual autoregressive(1). Depending on the context, sometimes θ will be estimated, and sometimes it will be specified as 0 or 1. Notation-wise, we generally let α represent the parameter vector of the correlation model.

In many longitudinal studies, individuals are not observed at all T occasions on account of some stochastic missing data mechanism. Here, we assume that all subjects are observed at baseline ($t = 1$). However, subjects can be missing at any follow-up time. It is convenient then to introduce $(T-1)$ random variables, R_{it}, $(t = 2, \ldots, T)$, that equal 1 if Y_{it} is observed and 0 if Y_{it} is missing. As we discuss briefly in the following section, under the protective assumption used in the pseudo-likelihood of Fitzmaurice et al., a model for R_{it} does not even need to be specified. However, when using the pseudo-likelihood of Troxel, et al., the marginal model for R_{it} given Y_{it} and $\mathbf{x}_{it}$ does need to be correctly specified. Since R_{it} is binary, the marginal model for R_{it} involves specifiying the probability of being observed ($R_{it} = 1$). This probability is assumed to follow a logistic regression,

$$\pi_{it} = \pi_{it}(Y_{it}, \mathbf{x}_{it}, \gamma) = pr(R_{it} = 1|y_{it}, \mathbf{x}_{it}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 y_{it} + \gamma_2' \mathbf{x}_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{it} + \gamma_2' \mathbf{x}_{it})}\,. \qquad (2)$$

In this marginal model, if $\gamma_1 \neq 0$, then the missing data mechanism is non-ignorable, since the probability of being missing depends on possibly unobserved data Y_{it}. In the next section, we briefly discuss the pseudo-likelihoods of Troxel et al. and Fitzmaurice et al., and we describe our proposed estimator.

3. ESTIMATORS

3.1. Troxel et al. Pseudo-Likelihood under Working Assumption of Independence

In this section we review the pseudo-likelihood approach proposed by Troxel et al. [21] that uses an "independence working assumption," i.e., assumes that observations are independent over time. The resulting pseudo-likelihood is a product of simple marginal terms and can be used to estimate the marginal regression parameters β and the marginal missingness parameters γ, but not the association parameters α. To describe this pseudo-likelihood, we let $f(y_{it}, r_{it}|\mathbf{x}_{it}, \beta, \gamma)$ denote the marginal distribution of (Y_{it}, R_{it}) at time t. We can write this distribution as

$$f(y_{it}, r_{it}|\mathbf{x}_{it}, \beta, \gamma) = f(y_{it}|\mathbf{x}_{it}, \beta) f(r_{it}|y_{it}, \mathbf{x}_{it}, \gamma),$$

where $f(y_{it}|\mathbf{x}_{it}, \beta)$ is Bernoulli with success probability given in (1), and $f(r_{it}|y_{it}, \mathbf{x}_{it}, \gamma)$ is Bernoulli with probability of being observed as given in (2). If we consider only the data at time t, then our observed data likelihood would be

$$f(y_{it}, r_{it}|\mathbf{x}_{it}, \beta, \gamma)$$

if Y_{it} were observed, and would be

$$\sum_{y_{it}=0}^{1} f(y_{it}, r_{it}|\mathbf{x}_{it}, \beta, \gamma)$$

if Y_{it} were missing.

The pseudo-likelihood [21], then, which treats the observations at different times as

independent, is

$$\begin{aligned}
\mathcal{L}_{ind}(\beta,\gamma) &= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it},r_{it}|\mathbf{x}_{it},\beta,\gamma)]^{r_{it}}\left[\sum_{y_{it}=0}^{1} f(y_{it},r_{it}|\mathbf{x}_{it},\beta,\gamma)\right]^{(1-r_{it})} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it}|\mathbf{x}_{it},\beta)f(r_{it}|y_{it},\mathbf{x}_{it},\gamma)]^{r_{it}}\left[\sum_{y_{it}=0}^{1} f(y_{it}|\mathbf{x}_{it},\beta)f(r_{it}|y_{it},\mathbf{x}_{it},\gamma)\right]^{(1-r_{it})} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it}|\mathbf{x}_{it},\beta)\pi_{it}]^{r_{it}}\left[\sum_{y_{it}=0}^{1} f(y_{it}|\mathbf{x}_{it},\beta)(1-\pi_{it})\right]^{(1-r_{it})}.
\end{aligned}$$

This pseudo-likelihood is simply a product of terms at each measurement occasion: when an observation is present, the Bernoulli probability function $f(y_{it}|\mathbf{x}_{it},\beta)$ is multiplied by the probability of being observed (π_{it}), and when the observation is missing, the product of $f(y_{it}|\mathbf{x}_{it},\beta)$ and the missingness probability $(1-\pi_{it})$ is summed over the range of the possible values of Y_{it}. Note that these marginal distributions are not a function of the association parameter α.

The maximum pseudo-likelihood estimate of Troxel et al. [21] under independence maximizes the log pseudo-likelihood, which can be obtained by setting the first derivative of the log pseudo-likelihood, i.e., the pseudo-score vector,

$$S_{ind}(\beta,\gamma) = \frac{\partial}{\partial(\beta,\gamma)}\log\mathcal{L}_{ind}(\beta,\gamma) = \sum_{i=1}^{n} S_{i,ind}(\beta,\gamma),$$

equal to $\mathbf{0}$ and solving for $(\hat{\beta}_{ind},\hat{\gamma}_{ind})$. Using method of moments ideas, the pseudo-likelihood estimator $(\hat{\beta}_{ind},\hat{\gamma}_{ind})$ can be shown to be consistent and asymptotically normal if the marginal bivariate distribution $f(y_{it},r_{it}|\mathbf{x}_{it},\beta,\gamma)$ is correctly specified. The maximum pseudo-likelihood estimate can be obtained using a Newton-Raphson algorithm, or the same EM-algorithm [30] that would be used if the (Y_{it},R_{it}) pairs were truly independent. Finally, we note that the negative second derivative of the log pseudo-likelihood will not provide a consistent estimator of the asymptotic variance; instead, the so-called "robust" or "sandwich" variance estimator

can be used [23].

3.2. Fitzmaurice et al. Protective Estimator

The pseudo-likelihood of Fitzmaurice et al. [22] under the protective assumption is a product of $(T-1)$ simple conditional distributions and the marginal distribution of the outcomes at time 1. Recall that we assume Y_{i1} is observed for all subjects in the dataset, as in the AIDS study. The marginal distribution at time 1 for all subjects is the product of Bernoulli distributions over the n subjects, denoted by

$$\prod_{i=1}^{n} f(y_{i1}|\mathbf{x}_{i1},\beta) = \prod_{i=1}^{n} p_{i1}^{y_{i1}}(1-p_{i1})^{(1-y_{i1})} \; . \tag{3}$$

Note that since no data are missing at time 1, one could obtain a consistent, albeit inefficient, estimate of β (excluding time effects or interactions with time) from (3).

Next, consider the conditional probability of the outcome at time 1 given the outcome at time t $(t>1)$ (and that Y_{it} is observed),

$$f(y_{i1}|y_{it},\mathbf{x}_{it},R_{it}=1,\beta,\alpha,\gamma) = \frac{\text{pr}(R_{it}=1|y_{i1},y_{it},x_{it},\gamma)f(y_{i1},y_{it}|x_{it},\beta,\alpha)}{\sum_{y_{i1}}\text{pr}(R_{it}=1|y_{i1},y_{it},x_{it},\gamma)f(y_{i1},y_{it}|x_{it},\beta,\alpha)}.$$

Now suppose the conditional probability $\text{pr}(R_{it}=1|y_{i1},y_{it},x_{it},\gamma)$ does not depend on y_{i1}, i.e.,

$$\text{pr}(R_{it}=1|y_{i1},y_{it},\mathbf{x}_{it},\gamma) = \text{pr}(R_{it}=1|y_{it},\mathbf{x}_{it},\gamma). \tag{4}$$

This implies that the probability of being missing at a time-point can be predicted by all (or some combination) of the data at that time. Under (4),

$$\begin{aligned} f(y_{i1}|y_{it},\mathbf{x}_{it},r_{it}=1,\beta,\alpha,\gamma) &= \frac{\text{pr}(r_{it}=1|y_{it},x_{it},\gamma)f(y_{i1},y_{it}|x_{it},\beta,\alpha)}{\text{pr}(r_{it}=1|y_{it},x_{it},\gamma)\sum_{y_{i1}}f(y_{i1},y_{it}|x_{it},\beta,\alpha)} \\ &= \frac{f(y_{i1},y_{it}|x_{it},\beta,\alpha)}{\sum_{y_{i1}}f(y_{i1},y_{it}|x_{it},\beta,\alpha)} \\ &= f(y_{i1}|y_{it},x_{it},\beta,\alpha). \end{aligned} \tag{5}$$

This implies that the conditional distribution of Y_{i1} given $(y_{it}, \mathbf{x}_{it})$ for those observed at time t ($R_{it} = 1$), equals the population conditional distribution, $f(y_{i1}|y_{it}, \mathbf{x}_{it}, \beta, \alpha)$, which is a function of the parameters of interest β (as well as α). Since y_{it} will be observed for all subjects with $R_{it} = 1$, one can get a consistent estimate of (β, α) by maximizing the following pseudo-likelihood [22],

$$\mathcal{L}_{prot}(\beta, \alpha|\mathbf{Y}) = \prod_{i=1}^{n} \left(f(y_{i1}|\mathbf{x}_{it}, \beta, \alpha) \prod_{t=2}^{T} [f(y_{i1}|y_{it}, \mathbf{x}_{it}, \beta, \alpha)]^{r_{it}} \right),$$

which includes all subjects at time 1 and $f(y_{i1}|y_{it}, \mathbf{x}_{it}, \beta, \alpha)$ when y_{it} is observed. The marginal parameters are identified using arguments analogous to those presented by Brown for the normal case [24]; the means at the baseline assessment are clearly identified by the complete data at time 1, and subsequent means and correlation parameters are identified using combinations of the marginal distribution at time 1 and the conditional distributions involving the later assessments.

The maximum pseudo-likelihood can again be obtained by setting the pseudo-score vector,

$$S_{prot}(\beta, \alpha) = \frac{\partial}{\partial(\beta, \alpha)} \log \mathcal{L}_{prot}(\beta, \alpha) = \sum_{i=1}^{n} S_{i,prot}(\beta, \alpha),$$

equal to $\mathbf{0}$ and solving for $(\hat{\beta}_{prot}, \hat{\alpha}_{prot})$. Using method of moments ideas, the pseudo-likelihood estimator $(\hat{\beta}_{prot}, \hat{\alpha}_{prot})$ can be shown to be consistent and asymptotically normal if $f(y_{i1}|y_{it}, \mathbf{x}_{it}, \beta, \alpha)$ is correctly specified and (5) holds. The maximum pseudo-likelihood estimate can again be obtained using a Newton-Raphson algorithm. Again, the so-called "sandwich" variance estimator [23] must be used to obtain a consistent estimate of the variance.

3.3. Comparison of Pseudo-likelihood Approaches

The two approaches described above require different but in each case non-trivial assumptions related to the missing data mechanism. The independence approach of Troxel et al. requires

correct specification of a missingness model in which the missingness probabilities at a given time may depend only on outcomes at that time. The protective approach of Fitzmaurice et al. requires that the missingness probabilities at a given time may depend on outcomes at that time but *must not* depend on outcomes observed at baseline, an assumption that obviates the need to specify the missing data model directly. While the protective assumption in (4) is not the most general non-ignorable missing data mechanism, it is still non-ignorable due to dependence of R_{it} on the outcomes at time t. This assumption is often quite reasonable, since for many nonignorable missing data mechanisms, missingness depends primarily on the unobserved data at time t, Y_{it}, and possibly the covariates $\mathbf{x}_{it}$, but, conditional on Y_{it} (and $\mathbf{x}_{it}$), missingness is independent of Y_{i1}. However, even though the missing data mechanism does not have to be estimated, this protective assumption is, in a sense, stronger than the missingness assumptions made in the pseudo-likelihood proposed by Troxel et al. The Troxel et al. approach does not make any assumptions about the missingness probabilities at one time given data at that time *and another time*, but only makes assumptions about the missingness probability at one time given data at that time. On the other hand, the pseudo-likelihood proposed by Troxel et al. does require correct specification of the model for the missingness probability given in (2), which is not required by the Fitzmaurice et al. approach.

There are numerous scenarios in which both models would appear to be reasonable, for example, repeated assessments of highly correlated indicators of symptom occurrence such as tingling of the hands and feet in cancer patients receiving chemotherapy. One can plausibly hypothesize that the *current* occurrence of the symptom almost entirely determines the patient's ability to attend the clinic and thus have the symptom measured; one can equally plausibly be confident of modeling correctly (or nearly correctly) the predictors of missingness,

including the symptom value itself but also numerous other known complicating factors such as patient age, presence of family members to assist, treatment with anthracycline-based chemotherapy, etc.

Of greater interest are scenarios in which one set of assumptions holds but not the other. Consider, for example, a setting in which the likelihood of missingness depends on both the current assessment and the baseline value. This is plausible in the setting of quality of life where difficulty coping at baseline is often indicative of later missingness, but difficulty coping at later assessments also increases the likelihood of being missing; in addition, repeated assessments of quality of life tend to be highly variable and poorly correlated. In this setting, the protective assumption is violated, and the low correlation means that the protective estimator will not be robust to the violation; while the missingness model using the independence approach will also be misspecified by not including the baseline values, it will still correctly capture the nonignorability and thus suffer minimal bias. On the other hand, there are many scenarios in which the protective assumption is satisfied, but the missingness model in the independence approach is so badly misspecified that the subsequent estimates will be biased. In the coping example above, we might specify a model in which those with difficulty coping are more likely to be missing. In reality, however, it may be that both those with very poor coping and very high levels of coping may be equally likely to be missing, the former because they can't manage their disease and the latter because they see no need for follow-up care. Subjects with missing values will be a mixture of these two populations; the monotone model for missingness that links difficulty coping with higher rates of missingness will fail to capture the higher rates of missingness among a subset of those who are coping well, resulting in biased estimates.

3.4. Proposed Weighted Least Squares Estimator

As shown in the previous subsections, under the protective assumption given in (4), and assuming the missingness probability given in (2) is correctly modeled, both the estimate $\hat{\beta}_{ind}$ from Troxel et al.'s pseudo-likelihood, and the estimate $\hat{\beta}_{prot}$ from Fitzmaurice et al.'s pseudo-likelihood will be consistent. Compared to the assumptions required for consistency of the full likelihood, namely correct specification of the full joint distribution of the outcome Y_{it} and the missingness indicators R_{it}, these are still weak assumptions. Both pseudo-likelihood approaches are particularly attractive in this setting since the full likelihood can be far more complicated algebraically. In addition, ML estimation is computationally very demanding for $T > 4$, due to the additional nuisance parameters induced by the specification of the full joint distribution mentioned above. Note, however, because the association among the repeated measures is not used at all in the pseudo-likelihood of Troxel et al., their estimate can be very inefficient for estimating time trends. Further, when using Fitzmaurice et al.'s pseudo-likelihood, if the repeated measures were truly independent, then their pseudo-likelihood would simply reduce to the likelihood at time 1 (since the estimated correlations would be close to 0, and the pseudo-likelihood would mainly be a function of data at time 1). In this case, $\hat{\beta}_{prot}$ would be inefficient for estimating both time-stationary effects, since it only uses data at time 1, and time trends; in fact, there may be very little information in the pseudo-likelihood for estimating time trends in this case. In the simulations given in Section 5, with high correlations, we have found $\hat{\beta}_{ind}$ to be inefficient for estimating time trends and $\hat{\beta}_{prot}$ to be inefficient for estimating time-stationary effects. This suggests formulating a new estimator of the marginal regression parameters that is the optimal combination of $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$.

First, note that under the protective assumption, the joint asymptotic distribution of

$(\hat{\beta}_{ind}, \hat{\gamma}_{ind}, \hat{\beta}_{prot}, \hat{\alpha}_{prot})$ is multivariate normal with mean vector $(\beta, \gamma, \beta, \alpha)$ and variance-covariance matrix

$$\left[\begin{array}{cc} \mathcal{I}_{ind} & 0 \\ 0 & \mathcal{I}_{prot} \end{array} \right]^{-1} \sum_{i=1}^{n} E \left(\left[\begin{array}{c} S_{i,ind}(\beta,\gamma) \\ S_{i,prot}(\beta,\alpha) \end{array} \right] \left[\begin{array}{c} S_{i,ind}(\beta,\gamma) \\ S_{i,prot}(\beta,\alpha) \end{array} \right]' \right) \left[\begin{array}{cc} \mathcal{I}_{ind} & 0 \\ 0 & \mathcal{I}_{prot} \end{array} \right]^{-1} ,$$

where

$$\mathcal{I}_{ind} = E \left[\frac{\partial}{\partial(\beta,\gamma)} S_{ind}(\beta,\gamma) \right] ,$$

and

$$\mathcal{I}_{prot} = E \left[\frac{\partial}{\partial(\beta,\alpha)} S_{prot}(\beta,\alpha) \right] .$$

The variance estimate is obtained by replacing (β, γ) in $S_{i,ind}(\beta, \gamma)$ and $\mathcal{I}_{ind}$ with $(\hat{\beta}_{ind}, \hat{\gamma}_{ind})$ and (β, α) in $S_{i,prot}(\beta, \alpha)$ and $\mathcal{I}_{prot}$ with $(\hat{\beta}_{prot}, \hat{\alpha}_{prot})$. We denote the submatrix of this estimated variance-covariance matrix corresponding to $V_\beta = Var(\hat{\beta}_{ind}, \hat{\beta}_{prot})$ by $\hat{V}_\beta$.

Under the protective assumption,

$$E \left[\begin{array}{c} \hat{\beta}_{ind} \\ \hat{\beta}_{prot} \end{array} \right] = \left[\begin{array}{c} \mathbf{I}_J \\ \mathbf{I}_J \end{array} \right] \beta = \mathbf{Z}\beta ,$$

where $\mathbf{I}_J$ is a $(J \times J)$ identity matrix, and J is the dimension of $\mathbf{x}_i$.

We propose forming a new estimator that is the asymptotic minimum variance linear combination of $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$ [25, 26]. The new estimator is basically a weighted least squares estimate, where the weight matrix is the inverse of $\hat{V}_\beta$, the estimate of $V_\beta = Var(\hat{\beta}_{ind}, \hat{\beta}_{prot})$. In particular, our proposed estimate is

$$\hat{\beta}_{wls} = [\mathbf{Z}'\hat{V}_\beta^{-1}\mathbf{Z}]^{-1}[\mathbf{Z}'\hat{V}_\beta^{-1}(\hat{\beta}'_{ind}, \hat{\beta}'_{prot})'] ,$$

which has asymptotic variance estimated by

$$\widehat{Var}(\hat{\beta}_{wls}) = [\mathbf{Z}'\hat{V}_\beta^{-1}\mathbf{Z}]^{-1} .$$

The estimate $\hat{\beta}_{wls}$ has the minimum asymptotic variance of any linear combination of $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$, including both $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$. Thus, with large n, $\hat{\beta}_{wls}$ will have smaller variance than both $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$. The decrease in variance of $\hat{\beta}_{wls}$ with respect to $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$ will depend on the configuration of the data, which we explore in simulations in Section 5.

4. APPLICATION: Response of CD4 Lymphocytes to Treatment with AZT or ddI

We present an analysis of the CD4 count data from the AIDS clinical trials described in the Introduction. The parameters are estimated using the protective pseudo-likelihood, the non-ignorable pseudo-likelihood under independence, WLS, and generalized estimating equations (GEE) [2] under ignorable assumptions, described below. The two AIDS clinical trials are randomised phase III double-blind trials, designed to compare two therapeutic treatments: zidovudine (AZT) and didanosine (ddI); the dataset contains records on $n = 431$ patients diagnosed with AIDS or AIDS-related complex. The response of interest at time (week) $t = 0, 1, \ldots, 5$ is the patient's CD4 count sufficiency, with $Y_{it} = 1$ if the CD4 count exceeds 200 and 0 otherwise. As discussed in the Introduction and given in Table I, CD4 count data are missing for 11% to 44% of patients at the five follow-up occasions; moreover, the missing data patterns are non-monotone.

To describe the treatment effect, we form the following indicator variable

$$\text{AZT}_i = \begin{cases} 1 \text{ if the } i^{th} \text{ subject is randomized to AZT} \\ 0 \text{ if the } i^{th} \text{ subject is randomized to ddI} \end{cases}.$$

Because of the stratified randomization, to control for baseline age we define the indicator variable

$$\text{age}_i = \begin{cases} 1 \text{ if the } i^{th} \text{ subject has baseline age} \geq 35 \\ 0 \text{ otherwise} \end{cases}.$$

We model the logit of $p_{it} = pr(Y_{it} = 1|x_{it})$, the probability that CD4 count > 200 at a given

time, as a function of treatment, time and baseline age,

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 \text{age}_i + \beta_3 t + \beta_4 t * \text{AZT}_i,$$

for $t = 0, 1, ..., 5$. Note the exclusion of a main effect of treatment (AZT_i). The main effect of AZT corresponds to the baseline ($t = 0$) treatment effect, and, because of randomization, there is no treatment effect at baseline, i.e., the main effect of AZT equals 0. For the ignorable GEE approach, we used the glimmix macro in SAS, which uses a linearization approach and allows incorporation of a random effect to accommodate the assumption of MAR data [27]. We assume a compound symmetry correlation structure for simplicity; results were robust to various other choices.

Recall that the protective pseudo-likelihood requires specification of the correlations, ρ_{1t}. We estimated the parameters under both AR(1) and exchangeable correlations; the results were so similar that for simplicity, we present results under an exchangeable assumption. Further, recall that for the non-ignorable pseudo-likelihood under independence, we must model the probability of being observed at each time point. It was conjectured that CD4 count is nonignorably missing since sicker patients may not come in for a further GP visit, e.g., sicker patients may have been hospitalized. We considered the following missing data mechanism:

$$\begin{aligned} \text{logit}(\pi_{it}) &= \text{logit}[\text{pr}(R_{it} = 1 | y_{it}, x_{it}, \gamma)] \\ &= \gamma_0 + \gamma_1 y_{it} + \gamma_2 \text{AZT}_i \\ &\quad + \gamma_3 \text{age}_i + \gamma_4 t + \gamma_{12} y_{it} \text{AZT}_i + \gamma_{14} y_{it} t \,, \end{aligned} \tag{6}$$

for $t > 0$. Using the pseudo-likelihood approach in (6), both the $y_{it}\text{AZT}_i$ and $y_{it}t$ interactions are significant at the 0.1% level. In general, the non-ignorable models suggest that subjects

with normal CD4 counts and on AZT are less likely to be seen over time.

Table II displays estimates and standard errors for the parameters β for all models and methods. Note that the WLS estimator of the k^{th} element of β is not just a weighted combination of the k^{th} elements of the independence and protective estimators, $\hat{\beta}_{ind,k}$ and $\hat{\beta}_{prot,k}$, but rather a weighted combination of the full vectors $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$, since the weight matrix $\hat{V}_{\beta}^{-1}$ is not diagonal; thus the WLS estimates do not always fall between the protective and pseudo-likelihood under independence estimates. From Table II, we see that the estimates from the WLS, protective approach, and pseudo-likelihood approach under independence are all similar, but the WLS has the smaller standard errors than these other two non-ignorable approaches. For example, for the time-stationary age effect, the estimated relative efficiency (ratio of estimated variances) is 44% for protective versus WLS and 87% for the pseudo-likelihood under independence versus WLS. For the AZT*TIME interaction, the estimated relative efficiency (ratio of estimated variances) is 46% for protective versus WLS and 15% for the pseudo-likelihood under independence versus WLS. The estimated exchangeable correlation is 0.54, indicating high correlation among the repeated responses, and we show in the simulation section that very substantial efficiency gains over the protective and pseudo-likelihood under independence approaches can be made using the WLS estimator when the correlation is high. This highlights the efficiency that can be gained using the WLS approach. However, this is just one example. To examine the finite sample properties of these approaches, we conducted a simulation study in the next section.

From Table II, we see that, among the non-ignorable approaches, the estimates are similar, except for the AGE effect using the protective estimate, which is over 50% smaller (and, as discussed above, also over 50% more variable). Comparing GEE to the non-ignorable

approaches, we see that the GEE estimate of the time by treatment interaction is much smaller than the estimate using the non-ignorable approaches. This also highlights how different assumptions about the missing data mechanism can produce discernibly different, and possibly conflicting, estimates of effects.

5. SIMULATION STUDY

We compared the WLS estimator, the protective estimator, the pseudo-likelihood estimator under independence, the ML estimator using the correct non-ignorable missingness mechanism, and GEE under an ignorable missing data mechanism. To ensure feasibility of the simulation study, we restricted the number of occasions to $T = 3$ and considered a simple two-group study design configuration (e.g., evenly randomized between active treatment and placebo).

Let $x_i = 0, 1$ indicate treatment group membership. The binary outcomes, denoted by (Y_{i1}, Y_{i2}, Y_{i3}), are assumed to follow a Bahadur model [29], with joint probabilities

$$\text{pr}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}|x_{it}, \beta, \alpha) =$$
$$\left\{\prod_{t=1}^{3} p_{it}^{y_{it}}(1-p_{it})^{(1-y_{it})}\right\}\left\{1 + \rho_{12}z_{i1}z_{i2} + \rho_{13}z_{i1}z_{i3} + \rho_{23}z_{i2}z_{i3} + \rho_{123}z_{i1}z_{i2}z_{i3}\right\},$$

where

$$\begin{aligned}
Z_{it} &= \frac{Y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}}\,, \\
\rho_{st} &= \text{Corr}(Y_{is}, Y_{it}) = \frac{E[(Y_{is} - p_{is})(Y_{it} - p_{it})|x_i]}{\sqrt{p_{is}(1-p_{is})p_{it}(1-p_{it})}}, \\
\rho_{123} &= \frac{E[(Y_{i1} - p_{i1})(Y_{i2} - p_{i2})(Y_{i3} - p_{i3})|x_i]}{\sqrt{p_{i1}(1-p_{i1})p_{i2}(1-p_{i2})p_{i3}(1-p_{i3})}}, \\
\text{logit}(p_{it}) &= \beta_0 + \beta_x x_i + \beta_\tau(t-1)\,,
\end{aligned}$$

for $t = 1, 2, 3$. We group $\alpha = [\rho_{12}, \rho_{13}, \rho_{23}, \rho_{123}]'$. For the simulation study, we choose $\beta_0 = -0.25$, $\beta_x = 0.5$, and $\beta_t = 0.20$. A variety of different correlation structures were

examined and the same overall pattern of results was obtained. For reasons of parsimony, we present the results from an exchangeable correlation with $\rho_{ist} = \rho$.

We performed simulations with the following true non-ignorable missingness mechanism,

$$\text{logit}(\pi_{it}) = \text{logit}[\text{pr}(R_{it} = 1|y_{it}, x_{it}, \gamma)] = \gamma_0 + \gamma_1 x_i + \gamma_2(t-1) + \gamma_3 y_{it}, \tag{7}$$

for $t > 1$, and we let the missingness indicators be independent at the three occasions. For the simulation study, the true model parameters in (7) are $\gamma_0 = -0.5$, $\gamma_1 = 1.0$, $\gamma_2 = 0.2$, and $\gamma_3 = 1.0$. Here, missingness at a given occasion depends upon group membership, time, and the possibly missing outcome at that occasion. In this mechanism, non-monotone missingness can occur in that an outcome can be missing at time s ($R_{is} = 0$), but observed at a future time t ($R_{it} = 1$ for $t > s$). Given the true γ parameters, the percentage missing is approximately 34% at time 2 and and 30% at time 3. The full distribution $f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_{it}, \gamma)$ is

$$\text{pr}[R_{i1} = r_{i1}, R_{i2} = r_{i2}, R_{i3} = r_{i3}|y_{i1}, y_{i2}, y_{i3}, x_{it}, \gamma] = \prod_{t=2}^{3} \pi_{it}^{r_{it}}(1-\pi_{it})^{(1-r_{it})}.$$

In the simulations reported in Table III, all of the non-ignorable methods are approximately unbiased, whereas GEE is clearly biased. The main interest of this simulation is to explore the efficiency gains of WLS over the protective and pseudo-likelihood estimator under independence. We provide both the average of the estimated variance and the empirical simulation variance; in general they match closely, except for the protective estimator when the correlation is low and the variance is poorly estimated. We see that the WLS estimator displays considerable gains in efficiency over the protective estimator for both β_τ and β_x for all correlations and sample sizes. In fact, the protective estimator is never more than 65% efficient compared to the WLS estimator; it is usually considerably less efficient. When the correlation is weak (e.g., $\rho = 0.1$), the pseudo-likelihood under independence is nearly as efficient as the MLE

(and thus also the WLS), since in this case, this estimator is close to the MLE. The variance of the WLS estimator is always smaller than the pseudo-likelihood under independence, although it performs less well when the correlation is low. In general, for the group effect, the pseudo-likelihood under independence is at least 90% as efficient as WLS. However, for the time effect, the pseudo-likelihood under independence can be substantially less efficient when the correlation is moderate to high. For example, when $n = 450$ and $\rho = .25$, for the time effect, the pseudo-likelihood under independence is only 64% as efficient as WLS. Further, when $n = 450$ and $\rho = .4$, for the time effect, the pseudo-likelihood under independence is only 48% as efficient as WLS. Comparing WLS to maximum likelihood, we see that WLS has at least 90% efficiency for any configuration, except for the group effect when $n = 300$ and $\rho = .4$, in which case it is 85% efficient.

6. DISCUSSION

We have proposed a weighted least squares estimator (WLS) which is an optimal combination of $\hat{\beta}_{ind}$ and $\hat{\beta}_{prot}$. The WLS estimator is appropriate for the estimation of marginal models for longitudinal binary data with non-monotone, nonignorably missing outcomes. Unlike the full likelihood, WLS requires specification of the bivariate distribution of the data at time 1 given all future times on the same subject (for the protective estimator) and the marginal missing model at each time point. Further, compared to maximum likelihood, which requires the full likelihood to be correctly specified in order to obtain consistent estimates, the pseudo-likelihood estimates are consistent as long as the protective assumption holds and the marginal missingness model is correctly specified. We have discussed above some of the scenarios in which both the protective assumption should hold and the missingness model can be specified with

a fair degree of confidence. In such cases, the use of the WLS estimator has the benefit of added efficiency compared to both of its components. In some scenarios, however, the analyst may be unwilling to require additional assumptions in order to achieve this efficiency. As many other authors have noted, sensitivity analyses are a crucial component of any analysis involving potentially nonignorable missing data [6, 31, 32, 33]. Comparisons of results obtained using models such as those described here, that allow for nonignorable missing data, with models making assumptions of MAR data are extremely instructive; in addition, comparisons of results from models making different assumptions about the mechanisms of nonignorability, as in the various approaches discussed here, can help elucidate the missing data mechanism.

Because of the broad range of possible missing data configurations and underlying probability distributions generating the data, it is difficult to draw definitive conclusions from simulation studies, and we can make only general suggestions. Based on our simulation studies, however, we have shown that one can take two relatively inefficient estimators (the protective and pseudo-likelihood under independence), and create a highly efficient estimator in the WLS estimator.

ACKNOWLEDGEMENTS

The authors are grateful for constructive comments from two reviewers, and for the support provided by the following grants from the US National Institutes of Health: AI 60373, GM 29745, CA 74015, CA 70101, MH 054693, and CA 68484. Andrea Troxel gratefully acknowledges support from the Columbia University Institute for Scholars at Reid Hall, Paris. Geert Molenberghs gratefully acknowledges financial support from the Belgian Science Policy IAP research network #P6/03.

REFERENCES

1. Cox DR. The analysis of multivariate binary data. *Applied Statistics* 1972; **21**: 113-20.
2. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
3. Le Cessie , Van Houwelingen JC. Logistic regression for correlated binary data. *Applied Statistics* 1994; **43**: 95-108.
4. Meester SG, MacKay J. A parametric model for cluster correlated categorical data. *Biometrics* 1994; **50**: 954-963.

5. Molenberghs G, Lesaffre E. Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* 1994; **89**: 633-644.
6. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Wiley & Sons: New York, 1987.
7. Baker SG. Marginal regression for repeated binary data with outcomes subject to nonignorable nonresponse. *Biometrics* 1995; **51**: 1042-1052.
8. Baker SG, Laird NM. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* 1988; **83**: 62-69.
9. Diggle P, Kenward MG. Informative drop-out in longitudinal analysis (with discussion). *Applied Statistics* 1994; **43**: 49–93.
10. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 1999; **55**: 688-698.
11. Ibrahim JG, Chen MH, Lipsitz, SR. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 2001; **88**: 551-564.
12. Kahn JO, Lagakos SW, Richman DD, AIDS Clinical Trials Group. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New England Journal of Medicine* 1992; **327**: 581-7.
13. Gallant JE, Moore RD, Richman DD, Keruly J, Chaisson RE, Zidovudine Epidemiology Study Group. Incidence and natural history of cytomegalovirus disease in patients with advanced human immunodeficiency virus disease treated with zidovudine. *Journal of Infectious Diseases* 1992; **166**: 1223-7.
14. Finkelstien DM, Williams PL, Molenberghs G, Feinberg J, Powderly WG, Kahn J, Dolin R, Cotton D. Patterns of opportunistic infections in patients with HIV infection. *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology* 1996; **12**, 38-45.
15. Phair J, Munoz A, Detels R, Kaslow R, Rinaldo C, Saah A, the Multicenter AIDS Cohort Study Group. The risk of *Pneumocystis carinii* pneumonia among men infection with human immunodeficiency virus type 1. *New England Journal of Medicine* 1990; **332**: 161-5.
16. Fitzmaurice G, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B* 1996; **57**: 691-704.
17. Gill R, Robins JM. Sequential models for coarsening and missingness. In *Proceedings of the First Seattle Symposium on Bisotatistics: Survival Analysis*, Lin DY, Fleming TR (eds). Springer-Verlag: New York, 1997; 295-305
18. Robins JM, Gill R. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* 1997; **16**: 39-56.
19. Gong G, Samaniego F. Pseudo maximum likelihood estimation: theory and applications. *Annals of Statistics* 1981; **9**: 861–869.
20. Liang K-Y, Self SG. On the asymptotic behavior of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society, Series B* 1996; **58**: 785–796.
21. Troxel AB, Lipsitz SR, Harrington DP. Marginal models for the analysis of longitudinal measurements subject to nonignorable non-monotone missing data. *Biometrika* 1998; **85**: 661-672.
22. Fitzmaurice G, Lipsitz SR, Molenberghs G, Ibrahim JG. A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *Journal of the Royal Statistical Society, Series A* 2005; **168**: 723-735.
23. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**: 1-26.
24. Brown CH. Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 1990; **46**: 143-155.
25. Wei LJ, Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985; **72**: 359-364.
26. Bloch DA, Moses LE. Nonoptimally weighted least squares. *The American Statistician* 1988; **42**: 50-53.
27. Schabenberger O. Introducing the GLIMMIX procedure for generalized linear mixed models. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. SAS Institute Inc: Cary, NC, 2005; Paper 196-30.
28. Plackett RM. A class of bivariate distributions. *Journal of the American Statistical Association* 1965; **60**: 526-22.
29. Bahadur RR. A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, Solomon H (ed). Stanford University Press, 1961; 158-68.
30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the *EM* algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1–38.
31. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* 1983; **45**: 212–18.
32. Nordheim EV. Inference from nonrandomly missing categorical data: an example from a genetic study in Turner's syndrome. *Journal of the American Statistical Association* 1984; **79**: 772–80.
33. Scharfstein D, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-

response models (with discussion). *Journal of the American Statistical Association* 1999; **94**: 1096-1146.

Table I. Number of subjects seen at each occasion in AIDS data

Weeks from baseline	Number of Subjects	Percent (out of $n = 431$)
0	431	100.00
1	383	88.86
2	345	80.05
3	324	75.17
4	306	71.00
5	285	66.13

Table II. Parameter Estimates for β for the AIDS Data

Effect	Approach	$\hat{\beta}$	SE	Z-statistic	P-value
INTERCEPT	Protect	-2.410	0.307	-7.86	0.000
	Pseudo	-2.771	0.242	-11.43	0.000
	WLS	-2.643	0.234	-11.28	0.000
	GEE	-2.579	0.232	-11.12	0.000
AGE	Protect	0.338	0.408	0.83	0.407
	Pseudo	1.112	0.289	3.85	0.000
	WLS	1.011	0.270	3.74	0.000
	GEE	0.902	0.280	3.22	0.001
TIME	Protect	-0.203	0.049	-4.18	0.000
	Pseudo	-0.116	0.041	-2.85	0.004
	WLS	-0.145	0.034	-4.24	0.000
	GEE	-0.101	0.042	-2.38	0.016
AZT*TIME	Protect	0.282	0.111	2.55	0.011
	Pseudo	0.272	0.192	1.42	0.157
	WLS	0.196	0.075	2.62	0.009
	GEE	0.070	0.073	0.95	0.338

Table III. Simulation Results. The marginal logistic model has parameters $(\beta_\tau, \beta_x) = (0.2, 0.5)$, and $\rho_{ist} = \rho$ (exchangeable)

n		APPROACH	$\rho = 0.10$ $\beta_\tau = 0.2$	$\rho = 0.10$ $\beta_x = 0.5$	$\rho = 0.25$ $\beta_\tau = 0.2$	$\rho = 0.25$ $\beta_x = 0.5$	$\rho = 0.40$ $\beta_\tau = 0.2$	$\rho = 0.40$ $\beta_x = 0.5$
150	Simulation	Protect	0.220	0.511	0.224	0.507	0.209	0.515
	Average	Pseudo-Ind	0.206	0.499	0.206	0.497	0.204	0.507
	Estimate	WLS	0.204	0.490	0.191	0.488	0.191	0.496
		ML	0.202	0.496	0.203	0.499	0.204	0.511
		GEE	0.364	0.389	0.353	0.408	0.336	0.424
	Simulation	Protect	73.915	0.131	6.136	0.145	0.164	0.154
	Average	Pseudo-Ind	0.046	0.067	0.044	0.075	0.039	0.084
	Variance	WLS	0.043	0.096	0.026	0.069	0.016	0.072
		ML	0.039	0.064	0.024	0.068	0.015	0.069
		GEE	0.017	0.055	0.014	0.065	0.012	0.075
	Empirical	Protect	1.872	0.136	0.357	0.153	0.034	0.158
	Simulation	Pseudo-Ind	0.042	0.066	0.037	0.072	0.033	0.082
	Variance	WLS	0.038	0.063	0.026	0.068	0.019	0.076
		ML	0.039	0.066	0.024	0.065	0.017	0.070
		GEE	0.016	0.055	0.014	0.066	0.011	0.077
	Coverage	Protect	99.6	94.6	98.9	95.5	96.9	94.6
	Probability	Pseudo-Ind	90.4	94.6	90.8	95.1	91.6	94.6
		WLS	90.3	94.7	91.6	94.4	93.1	93.2
		ML	90.6	93.6	93.9	95.1	93.4	94.7
		GEE	75.7	91.5	75.8	93.7	76.8	93.6
300	Simulation	Protect	0.253	0.501	0.216	0.494	0.203	0.501
	Average	Pseudo-Ind	0.203	0.495	0.206	0.496	0.203	0.498
	Estimate	WLS	0.202	0.488	0.198	0.491	0.195	0.490
		ML	0.202	0.499	0.200	0.501	0.204	0.502
		GEE	0.362	0.397	0.353	0.403	0.334	0.421
	Simulation	Protect	76.292	0.065	0.159	0.072	0.014	0.077
	Average	Pseudo-Ind	0.023	0.033	0.022	0.038	0.018	0.042
	Variance	WLS	0.023	0.053	0.013	0.034	0.008	0.037
		ML	0.020	0.032	0.012	0.034	0.007	0.034
		GEE	0.008	0.028	0.007	0.033	0.006	0.037
	Empirical	Protect	1.404	0.064	0.086	0.073	0.014	0.074
	Simulation	Pseudo-Ind	0.022	0.032	0.020	0.038	0.018	0.041
	Variance	WLS	0.020	0.031	0.013	0.035	0.009	0.038
		ML	0.019	0.031	0.012	0.033	0.008	0.036
		GEE	0.008	0.027	0.007	0.032	0.006	0.038
	Coverage	Protect	99.6	94.5	96.2	94.4	97.4	97.7
	Probability	Pseudo-Ind	92.9	94.6	93.4	94.0	96.5	97.3
		WLS	92.9	94.2	94.2	94.1	96.8	97.7
		ML	93.0	94.6	94.0	95.4	95.2	94.8
		GEE	56.8	90.8	55.0	91.2	56.7	92.4
450	Simulation	Protect	0.250	0.504	0.204	0.504	0.203	0.504
	Average	Pseudo-Ind	0.199	0.502	0.204	0.500	0.206	0.499
	Estimate	WLS	0.198	0.497	0.198	0.495	0.198	0.494
		ML	0.200	0.502	0.201	0.502	0.202	0.502
		GEE	0.361	0.398	0.350	0.404	0.332	0.424
	Simulation	Protect	12.435	0.043	0.043	0.048	0.009	0.051
	Average	Pseudo-Ind	0.016	0.022	0.014	0.025	0.012	0.028
	Variance	WLS	0.016	0.032	0.009	0.023	0.006	0.025
		ML	0.013	0.021	0.008	0.022	0.005	0.023
		GEE	0.005	0.018	0.005	0.022	0.004	0.025
	Empirical	Protect	1.008	0.041	0.027	0.048	0.009	0.049
	Simulation	Pseudo-Ind	0.015	0.021	0.014	0.024	0.012	0.026
	Variance	WLS	0.014	0.020	0.009	0.022	0.006	0.024
		ML	0.013	0.022	0.008	0.023	0.005	0.023
		GEE	0.006	0.019	0.005	0.022	0.004	0.025
	Coverage	Protect	99.4	95.6	97.2	95.6	95.9	95.1
	Probability	Pseudo-Ind	93.2	95.7	93.4	95.3	93.1	94.2
		WLS	93.4	95.0	94.2	95.0	94.0	94.1
		ML	94.2	94.0	94.8	94.2	94.0	94.4
		GEE	40.7	88.0	38.9	89.6	42.1	92.6