Published in final edited form as:

Stat Med. 2009 August 30; 28(19): 2451–2472. doi:10.1002/sim.3617.

Estimation of average treatment effect with incompletely observed longitudinal data: Application to a smoking cessation study

Hua Yun Chen¹ and Shasha Gao²

¹Division of Epidemiology and Biostatistics School of Public Health, University of Illinois at Chicago
1603 West Taylor Street Chicago, IL 60612

²Center for Health Equity Research and Promotion VA Pittsburgh Healthcare System 7180 Highland
Drive Pittsburgh, PA 15206

Abstract

We study the problem of estimation and inference on the average treatment effect in a smoking cessation trial where an outcome and some auxiliary information were measured longitudinally, and both were subject to missing values. Dynamic generalized linear mixed effects models linking the outcome, the auxiliary information, and the covariates are proposed. The maximum likelihood approach is applied to the estimation and inference on the model parameters. The average treatment effect is estimated by the G-computation approach, and the sensitivity of the treatment effect estimate to the nonignorable missing data mechanisms is investigated through the local sensitivity analysis approach. The proposed approach can handle missing data that form arbitrary missing patterns over time. We applied the proposed method to the analysis of the smoking cessation trial.

Keywords

Causal effect; Potential outcomes; Robust estimator; Surrogate outcome

1 Introduction

The smoking cessation trial to be analyzed is a part of “IT’S TIME” study, a 2-year longitudinal study consisting of clinic-based interventions delivered to woman smokers of child-bearing age to help them quit smoking. In this study, 1706 eligible smokers were accrued in 12 Chicago area clinics. Subjects were randomized either to the treated arm receiving an educational program helping them to quit smoking or to the controlled arm where no such a program was offered. Follow-up telephone interviews were conducted at 2, 6, 12, and 18 months into the study. The primary outcome of interest in the study is QUIT, a binary variable denoting a woman’s report on whether she was abstinent (had not smoked for at least 7 days) at a specific wave. The primary objective of the statistical analysis of the smoking cessation data is to estimate the average treatment effects of the intervention relative to that of the control for a given subpopulation. One challenging problem in estimating the treatment effect is the incompletely observed data. Only about 20% of the subjects completed all follow-up measures on the primary outcome. See Table 1 for more details.

When no missing data are involved, the average treatment effect can be easily obtained. With missing data, estimating the average treatment effect is no longer straightforward. When data are missing at random (MAR) [1], two different approaches may be followed. The first estimates the missing data probability from the data and then uses the weighted estimating

equation approach based on the marginal model to obtain the estimate. The other approach uses the maximum likelihood to estimate a joint model on the outcome and potential confounders and then marginalize to obtain the estimate. The former approach is simple but can be very inefficient when the proportion of complete cases is low. Methods for gaining efficiency can be very difficult to implement with non-monotonic missing data patterns [2], [3]. The second approach is more efficient when the joint model is correctly specified. When data are not missing at random (NMAR), the second approach can be taken by modeling the missing data mechanism and the outcome and confounders jointly and then marginalizing the joint model for the full data to obtain the average treatment effect. In this article, we follow the idea of the second approach to estimating the average treatment effect.

Modeling longitudinal data has been extensively discussed in the literature, perhaps, starting from [4] for the continuous outcomes. The discrete outcomes are more challenging to model. Generalized linear mixed models [5],[6], generalized estimating equations [7],[8], and the Markov transitional models [9],[10] are usually considered the three main models of choice. Further development combines features of the three types of models [11],[12],[13]. We propose a mixture of the random effects model and the Markov transitional model, which we call the dynamic generalized linear mixed model for categorical outcomes. For continuous outcomes, the dynamic generalized linear mixed model is known to be equivalent to the linear mixed effects model with autocorrelated errors [14]. For discrete outcomes, no such correspondence exists. In comparison to the dynamic latent trait models [13], such dynamic models are more transparent and stable, and can be used to directly model the relationship among multiple outcomes measured longitudinally.

In the smoking cessation study, in addition to the primary outcome: QUIT, a self-reported auxiliary variable was also measured when the subject had not quit at a given wave. This auxiliary variable measures a subject's readiness to quit at the given time. This variable was called STAGE. A subject classified herself into one of the 5 ordinal stages with respect to her readiness to quit, with 1 corresponding to the least likely to quit and 5 to the most likely to quit. It is speculated that the current stage of readiness to quit smoking of a subject may be informative in predicting the quitting status of that subject in the future. From this point of view, STAGE may be loosely regarded as a surrogate for the future outcome QUIT [15],[16], [17]. Since the auxiliary variable may provide useful information on the mechanism leading to the event [18]: quit, it is of practical importance to include the auxiliary information in the statistical model. Furthermore, including the auxiliary covariate in the model may help to correct bias and gain efficiency in the parameter estimation and make a MAR assumption more plausible. It is therefore of substantial interest to jointly model both the primary outcome and the auxiliary information when data are subject to missing values. In the smoking cessation study, both QUIT and STAGE were subject to missing values and the missingness occurred synchronized. We model them jointly using the dynamic generalized linear mixed models and evaluate the usefulness of the auxiliary information in helping estimate the average treatment effects.

The incompletely observed data are a challenging issue to handle in analyzing the smoking cessation data because the missing data formed complex non-monotonic missing patterns. Handling missing outcomes in longitudinal studies has been studied by many authors. Among them are [19], [20], [21], [22] for general models; [2], [3] for the generalized estimating equations, [23] for linear mixed models. More recently, [24],[25], [26],[27] discussed handling missing outcomes in various longitudinal data models. The missing data mechanism models proposed in the literature usually cannot accommodate complex missing data patterns that are nonmonotone and nonsaturated (not all potential missing patterns occur). When data are missing at random, the modeling approach we take is robust against misspecification of the MAR missing data mechanism. When data are not missing at random, rather than carrying out

a global sensitivity analysis of the missing data mechanism models on the treatment effect estimate, we take the simpler approach by performing a local sensitivity analysis [28],[29], [30],[31],[32]. The general form of the logistic regression models we proposed for modeling the missing data mechanism accommodates the nonmonotone nonsaturated missing data patterns. The proposal reduces to a series of logistic regression models usually used for modeling the missing data mechanism when the missing data form monotonic patterns.

The remainder of this article is organized in the following way. In section 2, we discuss a general framework of the dynamic mixed-effects generalized linear models. The framework is then specified for the smoking cessation data. Arbitrary missing data patterns are then modeled by the general logistic models at individual time points. The maximum likelihood approach is proposed for estimating and making inference on the model parameters under the MAR assumption. The approach for estimating the average treatment effect for a subpopulation is discussed in section 3. Sensitivity of the average treatment effect estimate to the possible nonignorable missing data mechanism models is studied through the local sensitivity analysis procedure in section 4. Section 5 presents a detailed analysis of the smoking cessation data using the framework discussed in sections 2, 3, 4. Section 6 concludes the article with a discussion on the merits and limitations of the proposed approach along with possible extensions of the proposed approach in future works.

2 Statistical method for the data analysis

2.1 The modeling framework for statistical analysis

Note that the main objective of the statistical analysis of the smoking cessation trial is to determine if the intervention group has a higher rate of quitting compared with that of the control group. More specifically, compared with smokers in the control group, it is of interest to know whether those in the intervention group would be more likely to have actually quit smoking over time. The statistical problem in analyzing the data can be formulated in the following way. Let Y_{it} , ($t = 1, 2, \dots, T$) denote the outcome of subject i measured at time t . In the smoking cessation study, Y_{it} is the quitting status measured at time t for subject i . Let $\mathbf{X}_i$ denote the collection of the treatment assignment, the baseline covariates, and possibly time-varying covariates that are not affected by the treatment. We use time-dependent notation to denote all of them until section 3 where we need to distinguish the time-varying and the time-independent covariates when we discuss the average causal effect estimator. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{iT})$ denote all the covariates. In addition to the outcome and covariates, a time-dependent auxiliary variable, denoted by $\mathbf{V}_i = (V_{i1}, \dots, V_{iT})$, was measured in the study. In contrast to the time-varying covariates, the auxiliary variables are those that may be affected by the treatments under study. In the smoking cessation study, $\mathbf{V}_i$ recorded the variable STAGE over time. When data are completely observed, the auxiliary variable appears not very useful in the estimation of the treatment effect. When the primary outcome is subject to missing values, modeling the auxiliary variable in the analysis may help correct bias and gain efficiency in the estimation of the treatment effect.

There are many ways to decompose the joint distribution of $(\mathbf{Y}, \mathbf{V})$ given $\mathbf{X}$. One natural and useful way to decompose the joint distribution is by the following sequence of conditional distributions in the order of time.

$$P(\mathbf{Y}_i, \mathbf{V}_i | \mathbf{X}) = \prod_{t=1}^T P(V_{it} | Y_{it}, V_{i(t-1)}, Y_{i(t-1)}, \dots, V_{i1}, Y_{i1}, \mathbf{X}_i) P(Y_{it} | V_{i(t-1)}, Y_{i(t-1)}, \dots, V_{i1}, Y_{i1}, \mathbf{X}_i),$$

where V_0 is assumed to be a constant if no such measure is recorded. The long-dependence on the recorded history in the conditional distributions can be a problem both in interpreting and in fitting the model. On the other hand, a short-dependence is easier to interpret in practice. This leads us to consider models with a short chain of dependence plus random effects. The idea behind the modeling strategy is that it is anticipated that the random effects capture the average long-term dependence while the short chain dependence captures the short-term dependence in addition to the average long-term dependence. Such a model appears as,

$$P(\mathbf{Y}_i, \mathbf{V}_i, a_i, b_i | \mathbf{X}_i) = \left[\prod_{t=1}^T P(V_{it} | Y_{it}, Y_{i(t-1)}, X_{it}, a_i) P(Y_{it} | Y_{i(t-1)}, Y_{i(t-1)}, X_{it}, b_i) \right] P(a_i, b_i)$$

We assume further that $P(V_{it} | Y_{it}, Y_{i(t-1)}, X_{it}, a_i)$ is known up to an unknown parameter ψ , $P(Y_{it} | V_{i(t-1)}, Y_{i(t-1)}, X_{it}, b_i)$ is known up to an unknown parameter β , and $P(a_i, b_i)$ is known up to an unknown Σ , which is a function of parameter θ .

For simplicity of presentation, we assume that covariates are always observed. Further, either both (Y, V) are observed or both are missing. Let $\mathbf{R}_i = (R_{i1}, \dots, R_{iT})'$ be the missing data indicator, whose t th component, R_{it} , equals 1 if (Y_{it}, V_{it}) is observed for individual i and is 0 if (Y_{it}, V_{it}) is missing. When data are missing at random, it is well known that the maximum likelihood method can yield valid inferences concerning the parameters β , ψ , and θ . In general, the missingness may depend on the missing values resulting in nonignorable missing data. In this case, a parametric model on the missing data probabilities may allow the identification of the model parameter. Under this situation, the maximum likelihood approach can still be used to obtain the parameter estimates and inference. However, if the model for the missing data probabilities is sufficiently large, the model may become unidentifiable. Therefore, nonignorable models for the missing data probabilities tend to serve as a framework for sensitivity analysis rather than as a way to identify the model parameter. We have the same purpose in mind here in specifying such a model. We specify the model for the missing data probabilities sequentially and assume parametric models as

$$\begin{aligned} P(\mathbf{R}_i | \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i) &= \prod_{t=1}^T P(R_{it} | R_{i(t-1)}, \dots, R_{i1}, Y_i, V_i, \mathbf{X}_i) \\ &= \prod_{t=1}^T P\{R_{it} | R_{i(t-1)}, \dots, R_{i1}, (Y_{it}, V_{it}, X_{it}), \dots, (Y_{i1}, V_{i1}, X_{i1}), \alpha\}, \end{aligned}$$

where α is the unknown parameter.

With the foregoing model specification, the marginal likelihood for the observed data $\{\mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i\}$, $i = 1, \dots, n$, can be written as

$$\prod_{i=1}^n \int P(R_i | \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i, \alpha) \cdot \int \int P(\mathbf{Y}_i, \mathbf{V}_i, a_i, b_i | \mathbf{X}_i, \beta, \psi, \theta) da_i db_i d\bar{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \quad (1)$$

where $R_i(\mathbf{Y}_i, \mathbf{V}_i)$ and $\bar{R}_i(\mathbf{Y}_i, \mathbf{V}_i)$ denote respectively the observed and missing parts of $(\mathbf{Y}_i, \mathbf{V}_i)$ determined by $\mathbf{R}_i$. For any categorical component of $(\mathbf{Y}_i, \mathbf{V}_i)$, the integration with respect to that component should also be understood as summation over the range. Assume that the parameters are identifiable either from the model restriction or from the fixation of some parameter values in the model for the missing data probabilities. Estimation and inference on the model parameters can be carried out by the maximum likelihood approach. The maximum

likelihood estimator can be obtained by the EM algorithm. More details are given in the next subsection when we deal with the specific application.

2.2 Model specification for the application

For the smoking cessation data, Y_{it} is a binary variable taking value 1 if the subject is considered to have quit smoking at time t , 0 otherwise. V_{it} , the readiness to quit, takes value 6 if $Y_{it} = 1$ and takes values 1 to 5 indicating the level of readiness. It can be seen that V_{it} predicts Y_{it} perfectly for $t = 1, \dots, T$ and we can express V_{it} as

$$V_{it} = 6Y_{it} + (1 - Y_{it})V_{it}^0,$$

where V_{it}^0 takes categories 1 to 5 when $Y_{it} = 0$. We have that $Y_{it} = 1_{\{V_{it}=6\}}$. Although modeling the distribution of $\{V_{it}, t = 1, \dots, T\}$ alone is equivalent to modeling $\{Y_{it}, V_{it}^0\}$ jointly for the estimation of the treatment effect, the fact that the event $V_{it} = 6$ carries so much weight in this clinical trial makes it very much worthwhile to treat it separately. As a result of this consideration, we model $\{V_{it}, t = 1, \dots, T\}$, the mathematical equivalent of $\{Y_{it}, V_{it}^0, t = 1, \dots, T\}$ jointly rather than directly. The models for $\{Y_{it}, V_{it}^0, t = 1, \dots, T\}$ are

$$P(V_{it}|Y_{it}, V_{i(t-1)}, Y_{i(t-1)}, \dots, V_{i1}, Y_{i1}, \mathbf{X}_i, a_i, b_i) = \{P(V_{it}^0|V_{i(t-1)}, \dots, V_{i1}, \mathbf{X}_i, a_i, b_i)\}^{1-Y_{it}}.$$

$$P(Y_{it}|V_{i(t-1)}, Y_{i(t-1)}, \dots, V_{i1}, Y_{i1}, \mathbf{X}_i, a_i, b_i) = P(Y_{it}|V_{i(t-1)}, \dots, V_{i1}, \mathbf{X}_i, a_i, b_i).$$

The short-chain dependence mixed effects model that was proposed in the last subsection becomes

$$P(V_{it}^0|V_{i(t-1)}, \dots, V_{i1}, \mathbf{X}_i, a_i, b_i) = P(V_{it}^0|V_{i(t-1)}, X_{it}, a_i, \psi),$$

and

$$P(Y_{it}|V_{i(t-1)}, \dots, V_{i1}, \mathbf{X}_i, a_i, b_i) = P(Y_{it}|V_{i(t-1)}, X_{it}, b_i, \beta).$$

Note that V_{it}^0 takes one of the $K(= 5)$ discrete values. We consider in this paper the baseline logit model for $P(V_{it}^0|V_{i(t-1)}, X_{it}, a_i)$ which specifies

$$\log \frac{P(V_{it}^0 = k|V_{i(t-1)}, X_{it}, a_i)}{P(V_{it}^0 = K|V_{i(t-1)}, X_{it}, a_i)} = \psi_{0k} + a_i + \psi_{1k} V_{i(t-1)} + \psi_{2k} X_{it},$$

for $k = 1, \dots, K - 1$. Note, however, that other models such as the proportional odds ratio model specified as

$$\log \frac{P(V_{it}^0 > k | V_{i(t-1)}, X_{it}, a_i)}{P(V_{it}^0 \leq k | V_{i(t-1)}, X_{it}, a_i)} = \psi_{0k} + a_i + \psi_1 V_{i(t-1)} + \psi_2 X_{it},$$

for $k = 1, \dots, K-1$, may also be used. The model for the distribution of Y_{it} given X_{it} and b_i is specified by the logistic regression as

$$\log \frac{P(Y_{it}=1 | V_{i(t-1)}, X_{it}, b_i)}{P(Y_{it}=0 | V_{i(t-1)}, X_{it}, b_i)} = \beta_0 + b_i + \beta_1 V_{i(t-1)} + \beta_2 X_{it}.$$

The random effects distribution for (a_i, b_i) is specified as $N(0, \Sigma(\theta))$.

Specification of models for the missing data mechanism can be accomplished by using the general logistic regression models as follow. Let $\pi_{itr} = P\{R_{it} = 1 | R_{i(t-1)}, \dots, R_{i1}, (Y_{it}, V_{it}, X_{it}), \dots, (Y_{i1}, V_{i1}, X_{i1})\}$, where r denote the missing data pattern formed by $(R_{i1}, \dots, R_{i(t-1)})$. Logistic model

$$\log \frac{\pi_{itr}}{1 - \pi_{itr}} = \alpha_{t0r} + \alpha_{t1r}(Y_{i1}, V_{i1}, X_{i1})' + \dots + \alpha_{itr}(Y_{it}, V_{it}, X_{it})'$$

may be fit when patterns $(R_{i1}, \dots, R_{i(t-1)}, R_{it} = 1)$ and $(R_{i1}, \dots, R_{i(t-1)}, R_{it} = 0)$ are both observed, where $\alpha_{tjr} = (\alpha_{tjry}, \alpha_{tjrv}, \alpha_{tjrx})$, $j = 1, \dots, t$, are respectively coefficients for each component of (Y_{ij}, V_{ij}, X_{ij}) . When either $(R_{i1}, \dots, R_{i(t-1)}, R_{it} = 1)$ or $(R_{i1}, \dots, R_{i(t-1)}, R_{it} = 0)$ is unobserved, parameter $\alpha_{tr} = (\alpha_{ttr}, \dots, \alpha_{t1r}, \alpha_{t0r})$ may become inestimable when fitting the logistic regression model even under missing completely at random assumption. Those probabilities are estimated by the degenerated probabilities 1 or 0 in this paper. This can be seen as a direct extension of the modeling strategy routinely used for monotonic missing data to nonmonotonic missing data. The probability for the observable pattern $(R_{i1}, \dots, R_{it})$ for subject i is

$$\prod_{t=1}^T \pi_{itr}^{R_{it}} (1 - \pi_{itr})^{1-R_{it}},$$

where r is conformable with the observed pattern $(R_{i1}, \dots, R_{iT})$.

With the foregoing model specification, we can now assemble the likelihood in a more concrete way as

$$L(\alpha, \beta, \psi, \theta) = \prod_{i=1}^n \left[\int \prod_{t=1}^T \pi_{itr}^{R_{it}}(\alpha_{tr}) \{1 - \pi_{itr}(\alpha_{tr})\}^{1-R_{it}} \int \int P(Y_{it} | V_{i(t-1)}, X_{it}, a_i, \beta) \right. \\ \left. \left\{ P(V_{it}^0 | V_{i(t-1)}, X_{it}, b_i, \psi) \right\}^{1-Y_{it}} P(a_i, b_i | \theta) da_i db_i dR_i(Y_i, V_i) \right].$$

where α is the collection of α_{tr} for $t = 1, \dots, T$ and for all the missing patterns where π_{tr} is not degenerated. Note that, for any of the models specified and used so far, by virtually the same approach, more complex models can be specified and used in place of the relatively simple ones whenever the simple models do not fit the data well. For example, more than one-step dependence on history may be added to the model. We suppress these developments for simplicity of presentation.

2.3 Parameter estimation and inference

We use the maximum likelihood approach to estimate and make inferences on the parameters. The EM algorithm is adopted for maximizing the likelihood function by augmenting the observed data $\{\mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i\}$ naturally to $(\mathbf{R}_i, \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i, a_i, b_i)$. The likelihood for the augmented data is

$$L^F(\alpha, \beta, \psi, \theta) = \prod_{i=1}^n \left[\prod_{t=1}^T \pi_{itr}^{R_{it}}(\alpha_{tr}) \{1 - \pi_{itr}(\alpha_{tr})\}^{1-R_{it}} P(Y_{it} | V_{it-1}, X_{it}, b_i, \beta) \right. \\ \left. \{P(V_{it}^0 | V_{it-1}, X_{it}, a_i, \psi)\}^{1-Y_{it}} P\{a_i, b_i | \Sigma(\theta)\} \right].$$

Let $l_i^F = \log L_i^F$ and $l_i = \log L_i$ which are respectively the contribution of the i th subject to the augmented data likelihood and to the observed data likelihood. Let $l^F = \sum_{i=1}^n l_i^F$ and $l = \sum_{i=1}^n l_i$. The objective function to be maximized in the EM algorithm is

$$Q(\alpha, \beta, \psi, \theta | \alpha^*, \beta^*, \psi^*, \theta^*) = \sum_{i=1}^n E \left[l_i^F(\alpha, \beta, \psi, \theta) | \{\mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i\}, \alpha^*, \beta^*, \psi^*, \theta^* \right].$$

The conditional distribution used for the calculation of the contribution to the objective function from the i th subject is

$$P\{\mathbf{R}_i, \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i, a_i, b_i | \mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \beta^*, \psi^*, \theta^*\} = \frac{L_i^F(\alpha^*, \beta^*, \psi^*, \theta^*)}{L_i(\alpha^*, \beta^*, \psi^*, \theta^*)}.$$

To compute the expectations under this distribution, we need to resolve the issue of computing integrations. We applied Gauss quadrature approximation to the integrations. Other numeric approaches or the Monte Carlo simulation approach may be applied. Such methods usually require intensive computation.

The EM algorithm applied to the example does not yield a closed-form solution to the maximization. Iterative approaches, such as the Newton-Raphson method, need to be applied. Such algorithms require computing both the first and the second derivatives of the augmented data log-likelihood. Score functions from the augmented data log-likelihood for the parameter α_{tr} , β , ψ , and θ are given in the Appendix.

Before applying the computational method to the data, we need to sort out the missing data patterns observed with degenerated probabilities. For longitudinal data with T time points, the total number of potential missing patterns is 2^T . However, we seldom observed all the missing data patterns in a single data set. This was the case for the smoking cessation data. For the smoking cessation data, there are potentially $2^4 = 16$ missing data patterns, of which only 14 were observed in this data set, as is seen from Table 1. To determine the missing patterns with degenerated probabilities, we first sort the missing data indicators in a dictionary order as is done in Table 2. Determination of the degenerated probabilities that are not required to model can then be done forwardly in time. For any given missing data pattern up to time $t-1$, $r = (r_1, \dots, r_{t-1})$, if both $R_t = 0$ and 1 occur in the data at time t , π_{tr} is not degenerated. Otherwise, it is degenerated. Whether $\pi_{tr} = 0$ or 1 for the degenerated case can be easily determined by whether $R_t = 0$ or 1. Finally, patterns associated with 0 probability of a degenerated pattern is also degenerated. When we apply this approach of determination to the observed missing data patterns in Table 2, it can be seen that the degenerated probabilities are $\pi_{3(00)} = 0$, $\pi_{4(100)} = 0$, and all the patterns associated with the zero probabilities. Such patterns are $(0, 0, *, *)$, and

(1,0,0,*), where * can be filled by either 0 or 1. Altogether, there are $2^1 + 2^0 = 3$ patterns whose probabilities are either 0 or 1 (degenerated). In other words, we need $2^4 - 3 = 13$ logistic models for the missing data probabilities. Table 3 showed the regression coefficient estimates from the fit of 13 logistic regression models with different intercepts and common coefficients. The missingness appears to be significantly affected by the education level and the number of years of smoking. The result suggests that those who did not complete high school education and/or have long smoking history are more likely to stay in the trial.

3 Estimating the average treatment effect

We use G-computation approach [33], [34], [35], to estimate the average treatment effect for a subpopulation defined by the baseline covariates and possibly by the time-dependent covariate processes that are not affected by the treatment. In the smoking cessation study, the covariates include the treatment assignment and the baseline covariates, such as race and educational levels. They are all time-independent covariates. To simplify the notation, from now on, we use Z to denote the treatment assignment and X to denote the baseline covariates. There are two treatment conditions: one is treated ($Z = 1$) and the other is control ($Z = 0$). Let $Y_t^{(0)}$ and $Y_t^{(1)}$ respectively denote the potential outcomes [36], [37], [38], if control or treatment is applied to a unit. Similarly, $V_t^{(0)}$ and $V_t^{(1)}$ denote respectively the potential auxiliary variable values if control or treatment is applied to the unit. We are interested in estimating the average treatment effect in the form of

$$ACE(X) = E\{Y_T^{(1)} - Y_T^{(0)}|X\}$$

for a subpopulation defined by the covariates that are not affected by treatment. Note that the average causal effect is defined at time T primarily for notational simplicity. It is easy to see that the definition can be extended to any time point. If there were no missing data, the data observed from the trial would be $(Y_{it}^{(Z_i)}, V_{it}^{(Z_i)}, t=1, \dots, T; X_i, Z_i)$ for $i = 1, \dots, n$. Suppose that treatment assignment is independent of the potential outcomes. More precisely,

$$P(Z|Y_t^{(0)}, Y_t^{(1)}, V_t^{(0)}, V_t^{(1)}, t=1, \dots, T; a^{(0)}, a^{(1)}, b^{(0)}, b^{(1)}, X) = P(Z|X). \quad (2)$$

It follows from the no-unmeasured-confounder assumption (2) that

$$\begin{aligned} P(Y_t^{(u)}|V_{t-1}^{(u)}, Y_{t-1}^{(u)}, \dots, V_1^{(u)}, Y_1^{(u)}, a^{(u)}, X) &= P(Y_t^{(Z)}|V_{t-1}^{(Z)}, Y_{t-1}^{(Z)}, \dots, V_1^{(Z)}, Y_1^{(Z)}, a^{(Z)}, X, Z=u), \\ P(V_{t-1}^{(u)}|Y_{t-1}^{(u)}, V_{t-2}^{(u)}, \dots, V_1^{(u)}, Y_1^{(u)}, b^{(u)}, X) &= P(V_{t-1}^{(Z)}|Y_{t-1}^{(Z)}, V_{t-2}^{(Z)}, \dots, V_1^{(Z)}, Y_1^{(Z)}, b^{(Z)}, X, Z=u), \end{aligned}$$

for $t = 1, \dots, T$, and

$$P(a^{(u)}, b^{(u)}|X) = P(a^{(Z)}, b^{(Z)}|Z=u),$$

because the random effects are independent of the covariates. This leads to

$$\begin{aligned}
E \left\{ h(Y_T^{(u)}) | X \right\} &= \int \cdots \int h(Y_T^{(u)}) \prod_{t=1}^T \left\{ P(Y_t^{(u)} | V_{t-1}^{(u)}, Y_{t-1}^{(u)}, \dots, V_1^{(u)}, Y_1^{(u)}, a^{(u)}, X) \right. \\
&\quad \left. P(V_{t-1}^{(u)} | Y_{t-1}^{(u)}, V_{t-2}^{(u)}, \dots, V_1^{(u)}, Y_1^{(u)}, b^{(u)}, X) \right\} dY_t^{(u)} dV_{t-1}^{(u)} \cdots dV_1^{(u)} dY_1^{(u)} P(a^{(u)}, b^{(u)} | X) da^{(u)} db^{(u)}, \\
&= \int \cdots \int h(Y_T^{(Z)}) \prod_{t=1}^T \left\{ P(Y_t^{(Z)} | V_{t-1}^{(Z)}, Y_{t-1}^{(Z)}, \dots, V_1^{(Z)}, Y_1^{(Z)}, a^{(Z)}, X, Z=u) \right. \\
&\quad \left. P(V_{t-1}^{(Z)} | Y_{t-1}^{(Z)}, V_{t-2}^{(Z)}, \dots, V_1^{(Z)}, Y_1^{(Z)}, b^{(Z)}, X, Z=u) \right\} P(a^{(Z)}, b^{(Z)} | Z=u) \\
&\quad dY_t^{(Z)} dV_{t-1}^{(Z)} \cdots dV_1^{(Z)} dY_1^{(Z)} da^{(Z)} db^{(Z)},
\end{aligned} \tag{3}$$

where $u = 0, 1$ and h is an integrable function. For the smoking cessation data, suppose that all the models specified in the previous section are true for the observed data when Y_{it} and V_{it} are replaced by $Y_{it}^{(Z_i)}$ and $V_{it}^{(Z_i)}$.

$$\begin{aligned}
E \left\{ h(Y_T^{(u)}) | X \right\} &= \int \cdots \int h(Y_T^{(Z)}) \prod_{t=1}^T \left[P(Y_t^{(Z)} | V_{t-1}^{(Z)}, a^{(Z)}, X, Z=u, \beta) \right. \\
&\quad \left. \left\{ P(V_{t-1}^{(Z)} | V_{t-1}^{(Z)}, b^{(Z)}, X, Z=u, \psi) \right\}^{1-Y_{t-1}^{(Z)}} \right] P(a^{(Z)}, b^{(Z)} | Z=u, \theta) \\
&\quad dY_t^{(Z)} dV_{t-1}^{(Z)} \cdots dV_1^{(Z)} dY_1^{(Z)} da^{(Z)} db^{(Z)},
\end{aligned} \tag{4}$$

Now suppose that the assumptions on the missing data mechanism are correct and parameter (β, ψ, θ) can be estimated from the likelihood consistently. Denote the estimate by $(\hat{\beta}, \hat{\psi}, \hat{\theta})$. If we use the distribution estimated from the maximum likelihood in computing the expectation in (4) with $h(y) = y$, denoted by $\widehat{E} \{ Y_T^{(u)} | X \}$ for $u = 0, 1$, the maximum likelihood estimator of the average causal effect can be obtained as

$$\widehat{ACE}(X) = \widehat{E}(Y_T^{(1)} | X) - \widehat{E}(Y_T^{(0)} | X).$$

The variance of the estimator can be obtained by the asymptotic approximation using the δ -method. See Appendix for more details. Finally, the average causal effect, $E(Y_T^{(1)} - Y_T^{(0)})$, for the population where the sample was drawn can be estimated by

$$\widehat{ACE} = \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{E}(Y_T^{(1)} | X_i) - \widehat{E}(Y_T^{(0)} | X_i) \right\}.$$

The variance of the estimator can be estimated by

$$\frac{1}{n} \sum_{i=1}^n V_{Ai} + \frac{1}{n-1} \sum_{i=1}^n \left\{ \widehat{ACE}(X_i) - \widehat{ACE} \right\}^2,$$

where $V_{Ai} = V_A(X_i)$ and $V_A(x)$ is the estimated variance of the limit of $\sqrt{n} \{ \widehat{ACE}(x) - ACE(x) \}$ in distribution.

4 Sensitivity analysis with nonignorable missing data

In the sensitivity analysis of the estimates to nonignorable missing data mechanism models, we concentrate on the analysis of the influence of nonignorable missing data mechanisms on the estimate of the average causal effect of the treatment versus the control. Although the framework proposed in the previous section can be used to obtain a global sensitivity analysis

of the average treatment effect to the nonignorable missing data mechanisms, the computation involved in such an analysis can be very intensive. We choose to perform a local sensitivity analysis which approximates the global sensitivity analysis. Note that the average causal effect can be expressed as

$$m(\beta, \psi, \theta) = E(Y_r^{(1)} | X, \beta, \psi, \theta) - E(Y_r^{(0)} | X, \beta, \psi, \theta).$$

To simplify notation, we use $\eta = (\beta, \psi, \theta)$ in this section. In the local sensitivity analysis, we first estimate η using the likelihood approach by ignoring the nonignorable missing data mechanism model. The estimated η is thus biased and is an implicit function of the parameter α . As a result, $m(\hat{\eta})$ is an implicit function of α through η , and using $m(\hat{\eta})$ as the estimate of $m(\eta)$ is subject to bias. The idea of the local sensitivity analysis is to try to correct the bias. We consider the set of missing data mechanisms that are modeled in the previous section. From now on, we use (γ, α) to denote all the parameters in the missing data mechanism model with α denoting the part of the parameters that define a NMAR mechanism if $\alpha \neq \alpha^*$, where α^* is often 0. γ denotes the rest of the parameters in the missing data mechanism model.

For a fixed α , let the estimate of (γ, η) be denoted by $\{\hat{\gamma}(\alpha), \hat{\eta}(\alpha)\}$. Suppose that the functions of m on η and η on α are continuously differentiable. To compute the local sensitivity approximation to $m(\eta)$, we need first to obtain the local sensitivity approximation to η . Note that the local sensitivity approximation to η is

$$\widehat{\eta}(\alpha) \approx \widehat{\eta}(\alpha^*) + (\alpha - \alpha^*)' \left\{ \frac{\partial \widehat{\eta}(\alpha^*)}{\partial \alpha} \right\}.$$

It then follows that

$$m\{\widehat{\eta}(\alpha)\} \approx m\{\widehat{\eta}(\alpha^*)\} + (\alpha - \alpha^*)' \left\{ \frac{\partial \widehat{\eta}(\alpha^*)}{\partial \alpha} \right\} \frac{\partial m}{\partial \eta} \{\widehat{\eta}(\alpha^*)\}.$$

Note that when $\alpha = \alpha^*$, the missing data mechanism is ignorable for any γ . From Appendix B, we see that

$$\frac{\partial \widehat{\eta}(\alpha^*)}{\partial \alpha} = \frac{\sum_{i=1}^n \frac{\partial^2 l}{\partial \alpha \partial \eta} \{\mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*)\}}{\left[-\sum_{i=1}^n \frac{\partial^2 l}{\partial \eta^2} \{\mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*)\} \right]^{-1}}.$$

The second term on the right-hand side is the inverse of the observed information matrix for the incomplete data, which can be obtained by fitting an ignorable model to the observed data. Since

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha \partial \eta} \{\mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*)\} &= E \left\{ \frac{\partial^2 l^F}{\partial \alpha \partial \eta} \{\mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*)\} \right. \\ &\quad \left. + Cov \left\{ \frac{\partial l^F}{\partial \alpha}, \frac{\partial l^F}{\partial \eta} \mid \mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i), \mathbf{X}_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \right\}, \end{aligned}$$

and $\frac{\partial^2 l^F}{\partial \alpha \partial \eta} = 0$ when α and η are variation independent parameters, the first term on the right-hand side becomes the conditional covariance between the score for the nonignorable part of

the missing data mechanism model and the score from the full data model given the observed data.

From Appendix A, we see that

$$\begin{aligned}\frac{\partial m}{\partial \eta} \{\widehat{\eta}(\alpha)\} &= \text{Cov} \left[\frac{\partial l^F}{\partial \eta} \{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\}, Y_T^{(1)} - Y_T^{(0)} | X, \widehat{\eta}(\alpha) \right] \\ &= \text{Cov} \left[\frac{\partial l^F}{\partial \eta} \{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\}, Y_T^{(Z)} | X, Z=1, \widehat{\eta}(\alpha) \right] \\ &\quad - \text{Cov} \left[\frac{\partial l^F}{\partial \eta} \{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\}, Y_T^{(Z)} | X, Z=0, \widehat{\eta}(\alpha) \right].\end{aligned}$$

Note that (α, γ) and η are in general variation-independent. As a result, $\frac{\partial l^F}{\partial \eta}$ does not involve the missing data mechanism model. The sensitivity of $m(\eta)$ with respect to α can then be computed by the linear extrapolation.

When the local sensitivity of the average causal effect, $E(Y_T^{(1)} - Y_T^{(0)})$, to the nonignorable missing data models is of interest, it follows that

$$\frac{1}{n} \sum_{i=1}^n m_i \{\widehat{\eta}(\alpha)\} \approx \frac{1}{n} \sum_{i=1}^n m_i \{\widehat{\eta}(\alpha^*)\} + (\alpha - \alpha^*)' \left\{ \frac{\partial \widehat{\eta}(\alpha^*)}{\partial \alpha} \right\} \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i}{\partial \eta} \{\widehat{\eta}(\alpha^*)\}.$$

where $m_i(\beta, \psi, \theta) = E(Y_T^{(1)} | X_i, \beta, \psi, \theta) - E(Y_T^{(0)} | X_i, \beta, \psi, \theta)$. The computation of the local sensitivity direction can be carried out by the foregoing approach.

5 Analysis of the smoking cessation data

Table 1 lists the rates of abstinence at each wave respectively for the control group and the intervention group. It shows that the abstinence rates have improved for both groups over the 18-month period. The intervention group appears to have a higher rate of abstinence than that of the control group. Note, however, that the rates were computed based on the actual number of subjects under observation. In view of the large number of subjects with missing outcomes, this improvement may be due to the selection bias. It is possible that the rate of abstinence at a wave could be increased over the previous wave even when no new subject became abstinent if subjects who were not going to quit smoking were more likely to drop out of the study. Table 2 displays the missing data patterns for the outcome: QUIT, and the auxiliary variable: STAGE. Note that QUIT and STAGE are either both missing or both observed. As a result, only one indicator of the missing values at a given time point was used. It can be seen from the table that the missing values form nonmonotone patterns and the total number of missing patterns (14) is less than the number of all potential patterns (16). In the following analysis, we aim at addressing the issue of missing outcomes in estimating the average treatment effect.

Before focusing on the issue of missing outcomes, we note that incomplete covariate information also occurred in some subjects. Among those 1706 women eligible at baseline, 14 did not report their race, 6 did not provide information on their educational level. Variables on years since started smoking and the number of daily cigarettes smoked had 81 and 22 missing values, respectively. The fraction of missing covariates is relatively small in comparison to that of missing outcomes. Rather than modeling the covariate distribution and maximizing the joint likelihood, which can be computationally much more challenging to be carried out, we took two simplified approaches: One is to exclude subjects with incomplete covariate information from the analysis; the other is to impute the incomplete covariate information by regression models fit to the available cases. Since the analytical results for the deleted data are

close to those for the imputed data, we only present the analysis of the imputed data set in this article.

Table 1 shows that the rates of abstinence change over time for both the treatment group and the control group. Based on this observation, we model the logit of the abstinent rate (QUIT) as a function of time and treatment condition after accounting for the effect of pre-treatment conditions. Specifically, two types of models were fit. The first one is a simple model that does not include the auxiliary variable into the model. This type of models can be schematically expressed as

$$\log \frac{P(\text{QUIT}=1 \text{ at time } t)}{P(\text{QUIT}=0 \text{ at time } t)} = \text{intercept}(t) + \text{treatment}(t) + \text{pre-treatment covariates} + \text{lag} - 1 \text{ QUIT} + \text{random effect.} \quad (5)$$

The other type includes the auxiliary variable in the models. This type of models can be schematically expressed as two joint models with the logistic model,

$$\log \frac{P(\text{QUIT}=1 \text{ at time } t)}{P(\text{QUIT}=0 \text{ at time } t)} = \text{intercept}(t) + \text{treatment}(t) + \text{pre-treatment covariates} + \text{lag} - 1 \text{ QUIT} + \text{lag} - 1 \text{ STAGE} + \text{random effect,} \quad (6)$$

for QUIT, and the baseline logit model,

$$\log \frac{P(\text{STAGE}=k \text{ at time } t)}{P(\text{STAGE}=5 \text{ at time } t)} = \text{intercept}(k) + \text{treatment}(k) + \text{pre-treatment covariates}(k) + \text{lag} - 1 \text{ QUIT} \& \text{STAGE}(k) + \text{random effect,} \quad (7)$$

for the STAGE, where $k = 1, \dots, 5$.

We fitted one model of the first type and two models of the second type to the observed data under MAR assumption. Model I had the form (5) and did not include V . Models II and III included V and had the joint model form of (6) and (7). All three models included the same baseline covariates. The variables lag-1 QUIT and lag-1 STAGE were combined as one variable and were treated as an ordinal variable in Model II. That is, $\text{STAGE}=6$ when $\text{QUIT}=1$. Model III treated $1_{\{\text{STAGE}=6\}}$ differently from those of stages 1 to 5, which were again treated as an ordinal variable.

Results on parameter estimation for the outcome from Models I, II, and III are listed in Table 4. Results on parameter estimates for the auxiliary variable from Models III are listed in Table 5. Note that we did not list parameter estimates for the auxiliary variables from Model II because they are very close to those of Model III. Results in Table 4 suggest that the treatment has significant effects on the QUIT status at months 2 and 6, but not significant effects on the QUIT status at months 12 and 18 when both pre-treatment covariates and the auxiliary variable STAGE at the previous time are adjusted. It can also be seen that the current QUIT status strongly depends on the STAGE at the previous time point even after adjusted for the pretreatment covariates. Subjects with more advanced stage at the previous time point tends to have a higher probability of quitting smoke in the current time point, which is consistent with intuition. In addition, the strong positive association between the current quitting status and the quitting status at the previous time suggests that a quitter at the previous time point is much more likely to remain quitting than to relapse to smoking in the current time point. Results in Table 5 also show that the current STAGE strongly depends on STAGE at the previous time point. Although the coefficient estimate for the lag-1 STAGE is negative, nevertheless, the result should be interpreted as such that subjects were more likely to advance their stage of

readiness to quit over time because the base category used in the baseline logit model was STAGE=5 rather than STAGE=1.

Since STAGE is a post-treatment variable, it is important to remember that the treatment effect may be diluted when STAGE is included in the adjustment. This issue was addressed by computing the average treatment effect using the G-computation approach. The results are given on the left panel of Table 6 along with the naive estimate of treatment effect, which is the difference of point-prevalence rates for treated and control groups. Those results indicate that, on average, treatment has significant effects on QUIT at months 2 and 6, but has insignificant effects on QUIT at months 12 and 18. The conclusion drawn from the average treatment effect estimates is consistent with those from the conditional model estimates and the naive rate estimates.

The treatment effect estimates are consistent based on different methods under MAR. It is of substantial interest to know how the conclusion changes if the MAR assumption is violated. Assume that missing values on Y and V were generated from the logistic models

$$\log \frac{\pi_{itr}}{1 - \pi_{itr}} = \text{intercept}(r) + \text{Baseline covariates} + \alpha_{1Y} Y_{it} + \alpha_{2Y} Y_{i(t-1)} + \alpha_{1V} V_{it} + \alpha_{2V} V_{i(t-1)},$$

where t denotes the time and r denotes the missing pattern. Under this model, the local sensitivity analysis was performed. The results are displayed on the right panel of Table 6. The numbers under each variable indicate the amount of change in the average treatment effect under the corresponding model at the given time point when the coefficient of that variable in the missing data mechanism model is changed by one unit. The results allow us to assess the impact of the nonignorable missing data on the average treatment effect estimation for any given combination of the coefficients in the missing data mechanism model. For example, if the actual missing data mechanism has $(\alpha_{1Y}, \alpha_{2Y}, \alpha_{1V}, \alpha_{2V}) = (0, 0.5, 0.1, 0)$, approximately, the average treatment effects at 2, 6, 12, and 18 months based on Model II are 5.28%, 4.10%, 0.68%, and -1.03% respectively. The results in Table 6 suggest that the average treatment effect at month 6 is much more sensitivity to the dependence of the missing data mechanism on the STAGE variable than at month 18 under both models II and III. The results also suggest that, when missingness depends on the current or previous QUIT or STAGE, the treatment observed under MAR can be substantially altered if the dependence is relatively strong. Overall, if Quitter or subjects in the advance stage of readiness to quit are more likely to dropout, the average treatment effect estimate will be much smaller than estimated under the MAR assumption.

The Fortran code for performing the analysis can be obtained from the authors by email to hychen@uic.edu.

6 Discussion

We presented a framework for the analysis of the longitudinal data with missing outcome and auxiliary variables. The proposed framework was used in the analysis of the smoking cessation data. Although the example we treated in this paper is special in that both the outcome and the auxiliary variables are observed or missing simultaneously, nevertheless, the framework is easily extended to handle the case when they are not observed or missing simultaneously. The auxiliary variable can also be viewed as a time-dependent covariate. In this way, the proposed method can be regarded as a joint modeling approach to the missing data problem with both the longitudinal outcomes and longitudinal covariates subject to missing values. Since the missing data form non-monotonic patterns, modeling the time-dependent covariate is necessary in the application of the likelihood approach. This framework can also handle nonmonotone

and unsaturated missing data. One possible argument against our treatment of the unsaturated missing data patterns is that we implicitly assume that the missing data patterns not observed cannot happen in the future. But we view a pattern that is not observed is so unlikely to happen that we estimate the probability of its happening zero rather than an estimated nonzero probability (not necessarily small as one may think it is) based on a potentially saturated model. In addition, a saturated logistic model can behave erratically when fitting to the data with observed unsaturated missing patterns. Our modeling approach naturally reduces to the consecutive logistic models usually used for modeling missing data probability when the missing data form monotonic patterns.

We applied the G-computation approach to obtain the average treatment effect estimate, which is of primary interest in trials of the similar type. We proposed a sensitivity analysis of the average treatment effect to the missing data mechanism models through the local sensitivity analysis approach. Since one may not know whether the observed data deviated from the MAR assumption, and if it does, how strong the deviation is, the results only provided us a way to quantify the potential impact on the estimated average causal effect of the deviation from the MAR assumption rather than a decisive conclusion, which may not be achieved without additional information from subject-matter experts. Note that the asymptotic unbiasedness of the proposed estimators in this article relies on the correct model assumptions, which need to be carefully evaluated. Alternative approaches that provide some protection against the bias of the parameter estimator, such as the doubly robust approach, may be applied under MAR assumption. However, when the MAR assumption on the missing data mechanism does not hold, the local sensitivity analysis can provide useful information while keeping the computation simple.

When the average causal effect of a subpopulation defined by the baseline covariates is of interest, the proposed approach can be modified to give an estimator of the effect. However, if the subpopulation is defined by only a subset of the covariates in the model, a model for the covariates not used in defining the subpopulation given the covariates used in defining the subpopulation is required to be specified. We leave those issues for further studies in the future.

Acknowledgments

We thank reviewers for helpful comments on the earlier versions of the paper, which led to substantial improvements on the presentation. The research is partially supported by NIH/NCI grant R01 CA106355.

Appendix A: Derivatives for the log-likelihood

The first derivatives are for the augmented data are

$$\begin{aligned}\frac{\partial l^F}{\partial \alpha_{ir}} &= \sum_{i=1}^n \frac{R_{it} - \pi_{it}}{(1 - \pi_{it})} \frac{\partial \log \pi_{it}}{\partial \alpha_{ir}}, \\ \frac{\partial l^F}{\partial \beta} &= \sum_{i=1}^n \sum_{t=1}^T \frac{\partial}{\partial \beta} \log P(Y_{it} | V_{i(t-1)}, X_{it}, b_i, \beta), \\ \frac{\partial l^F}{\partial \psi} &= \sum_{i=1}^n \sum_{t=1}^T (1 - Y_{it}) \frac{\partial}{\partial \psi} \log P(V_{it}^0 | V_{i(t-1)}, X_{it}, a_i, \psi), \\ \frac{\partial l^F}{\partial \theta} &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \log P(a_i, b_i | \Sigma(\theta)).\end{aligned}$$

The second derivatives can be obtained similarly. The first derivatives for the observed data are the conditional expectations of the foregoing derivatives given the observed data. The variance estimates for the maximum likelihood estimator can be obtained by the inverse of the observed data information matrix, which can be computed by Louis' (1984) formula as

$$\frac{\partial^2 l(\alpha, \beta, \psi, \theta)}{\partial^2(\alpha, \beta, \psi, \theta)} = \sum_{i=1}^n E \left\{ \frac{\partial^2 l_i^F(\alpha, \beta, \psi, \theta)}{\partial^2(\alpha, \beta, \psi, \theta)} \mid \{\mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i\}, \alpha^*, \beta^*, \psi^*, \theta^* \right\} \\ + \sum_{i=1}^n Var \left\{ \frac{\partial l_i^F(\alpha, \beta, \psi, \theta)}{\partial(\alpha, \beta, \psi, \theta)} \mid \{\mathbf{R}_i, \mathbf{R}_i(\mathbf{Y}_i, \mathbf{V}_i), \mathbf{X}_i\}, \alpha^*, \beta^*, \psi^*, \theta^* \right\}.$$

Appendix B: Variances of the causal effect estimators

We first derive a variance formula for $\widehat{ACE}(X) = \widehat{E}(Y_t^{(1)}|X) - \widehat{E}(Y_t^{(0)}|X)$. Use δ -method,

$$\widehat{ACE}(X) = ACE(X) + (\widehat{\eta} - \eta_0)^T \frac{\partial}{\partial \eta} \{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)\} + o_p(\|\widehat{\eta} - \eta_0\|),$$

where $\widehat{\eta} - \eta = (\widehat{\beta} - \beta, \widehat{\psi} - \psi, \widehat{\theta} - \theta)$. This implies that $\sqrt{n}\{\widehat{ACE}(X) - ACE(X)\}$ is asymptotically normal with variance

$$V_A = \frac{\partial}{\partial \eta} \{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)\} \Omega \left[\frac{\partial}{\partial \eta} \{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)\} \right]^T,$$

where Ω is the asymptotic variance for $\{\sqrt{n}(\widehat{\beta} - \beta_0), \sqrt{n}(\widehat{\psi} - \psi_0), \sqrt{n}(\widehat{\theta} - \theta_0)\}$. Note that,

$$\frac{\partial}{\partial \eta} E(Y_t^{(u)}|X) = Cov \left\{ Y_t^{(u)}, \frac{\partial l^F}{\partial \eta} \mid X \right\},$$

for $u=0,1$, which can be obtained by G-computation. It follows that

$$\frac{\partial}{\partial \eta} \{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)\} = Cov \left\{ Y_t^{(1)} - Y_t^{(0)}, \frac{\partial l^F}{\partial \eta} \mid X \right\}.$$

The variance of the estimator of $ACE(X)$ can be obtained based on the following expansion.

$$\widehat{ACE}(X) = ACE(X) + (\widehat{\eta} - \eta_0)^T \frac{\partial}{\partial \eta} \left\{ \frac{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)}{E(Y_t^{(0)}|X)} \right\} + o_p(\|\widehat{\eta} - \eta_0\|).$$

Therefore, the variance has a similar sandwich form as

$$V_R = \frac{\partial}{\partial \eta} \left\{ \frac{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)}{E(Y_t^{(0)}|X)} \right\} \Omega \left[\frac{\partial}{\partial \eta} \left\{ \frac{E(Y_t^{(1)}|X) - E(Y_t^{(0)}|X)}{E(Y_t^{(0)}|X)} \right\} \right]^T,$$

where

$$\frac{\partial}{\partial \eta} \left\{ \frac{E(Y_T^{(1)} - Y_T^{(0)} | X)}{E(Y_T^{(0)} | X)} \right\} = \frac{\text{Cov}(Y_T^{(1)} - Y_T^{(0)}, \frac{\partial l^F}{\partial \eta} | X)}{E(Y_T^{(0)} | X)} - \frac{E(Y_T^{(1)} - Y_T^{(0)} | X)}{E(Y_T^{(0)} | X)} \frac{\text{Cov}(Y_T^{(0)}, \frac{\partial l^F}{\partial \eta} | X)}{E(Y_T^{(0)} | X)}.$$

The estimated variances can be obtained by replacing Ω by its estimate obtained from the inverse of the observed information matrix and computing the covariance using the estimated models for the full data.

We can obtain the variance of $\widehat{ACE}$ similarly in the following way. Note that

$$\begin{aligned} \sqrt{n} \{ \widehat{ACE} - ACE \} &= \sqrt{n} (\widehat{\eta} - \eta_0)^T \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \eta} \{ E(Y_T^{(1)} - Y_T^{(0)} | X_i) \} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \{ ACE(X_i) - ACE \} + o_p(\sqrt{n} \|\widehat{\eta} - \eta_0\|). \end{aligned}$$

Note that

$$\sqrt{n} (\widehat{\eta} - \eta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n s\{\mathbf{R}_i, \mathbf{R}_i(Y_i, \mathbf{V}_i, \mathbf{X}_i)\} + o_p(1),$$

where $E\{s(\mathbf{Y}, \mathbf{V}, \mathbf{X}) | \mathbf{X}\} = 0$. It is now easy to see that the variance estimate is a consistent estimate of the asymptotic variance of $\sqrt{n}(\widehat{ACE} - ACE)$.

Appendix C: Derivatives in local sensitivity analysis

Note that the estimate of (γ, η) satisfies

$$\frac{\partial}{\partial (\gamma, \eta)} l\{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\} = 0,$$

for any α . By differentiating with respect to α on both sides of the equation, it follows that

$$\frac{\partial \{ \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \}}{\partial \alpha} = \frac{\partial^2 l\{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\}}{\partial \alpha \partial (\gamma, \eta)} \left\{ - \frac{\partial^2 l\{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\}}{\partial^2 (\gamma, \eta)} \right\}^{-1}.$$

Note that the observed data likelihood score can be written as conditional expectations of the full data likelihood scores, which separate the score for γ from the score for η as

$$\begin{aligned} \frac{\partial l}{\partial (\gamma, \eta)} \{\alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha)\} &= \left(E \left[\frac{\partial l^F}{\partial \gamma} \{\alpha, \widehat{\gamma}(\alpha)\} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right], \right. \\ &\quad \left. E \left[\frac{\partial l^F}{\partial \eta} \{\widehat{\eta}(\alpha)\} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right] \right), \end{aligned}$$

where l_1^F and l_2^F are respectively the likelihood for missing data mechanism model and the likelihood for the full data when (R, V, Y, X) is observed. It can be seen that

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha \partial \gamma} \{ \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \} = & E \left\{ \frac{\partial^2 l_1^F}{\partial \alpha \partial \gamma} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + \frac{\partial \widehat{\gamma}(\alpha)}{\partial \alpha} E \left\{ \frac{\partial^2 l_1^F}{\partial \gamma^2} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + Cov \left\{ \frac{\partial l_1^F}{\partial \gamma}, \frac{\partial l_1^F}{\partial \alpha} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + \frac{\partial \widehat{\gamma}(\alpha)}{\partial \alpha} Cov \left\{ \frac{\partial l_1^F}{\partial \gamma}, \frac{\partial l_1^F}{\partial \gamma} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + \frac{\partial \widehat{\eta}(\alpha)}{\partial \alpha} Cov \left\{ \frac{\partial l_1^F}{\partial \gamma}, \frac{\partial l_2^F}{\partial \eta} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\}. \end{aligned}$$

Similarly, it follows that

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha \partial \eta} \{ \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \} = & \frac{\partial \widehat{\eta}(\alpha)}{\partial \alpha} E \left\{ \frac{\partial^2 l_2^F}{\partial \gamma^2} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + Cov \left\{ \frac{\partial l_2^F}{\partial \eta}, \frac{\partial l_1^F}{\partial \alpha} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + \frac{\partial \widehat{\gamma}(\alpha)}{\partial \alpha} Cov \left\{ \frac{\partial l_2^F}{\partial \eta}, \frac{\partial l_1^F}{\partial \gamma} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} \\ & + \frac{\partial \widehat{\eta}(\alpha)}{\partial \alpha} Var \left\{ \frac{\partial l_2^F}{\partial \eta} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\}. \end{aligned}$$

When α^* is chosen such that the missing data mechanism is ignorable, it can be seen that

$$\frac{\partial l_1^F}{\partial \gamma} - E \left\{ \frac{\partial l_1^F}{\partial \gamma} | R, R(Y, V), X, \alpha, \widehat{\gamma}(\alpha), \widehat{\eta}(\alpha) \right\} = 0.$$

This implies that

$$\begin{aligned} \frac{\partial \widehat{\gamma}(\alpha^*)}{\partial \alpha} &= \sum_{i=1}^n E \left\{ -\frac{\partial^2 l_1^F}{\partial \alpha \partial \gamma} | R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \\ &\quad \left[\sum_{i=1}^n E \left\{ \frac{\partial^2 l_1^F}{\partial \gamma^2} | R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \right]^{-1} \\ &= - \sum_{i=1}^n \frac{\partial^2 l_{1i}^F}{\partial \alpha \partial \gamma} \{ \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \} \left[\sum_{i=1}^n \frac{\partial^2 l_{1i}^F}{\partial \gamma^2} \{ \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \} \right]^{-1} \end{aligned}$$

and that

$$\begin{aligned} \frac{\partial \widehat{\eta}(\alpha^*)}{\partial \alpha} &= \sum_{i=1}^n Cov \left\{ \frac{\partial l_2^F}{\partial \eta}, \frac{\partial l_1^F}{\partial \alpha} | R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \\ &\quad \left[\sum_{i=1}^n E \left\{ -\frac{\partial^2 l_2^F}{\partial \eta^2} | R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \right. \\ &\quad \left. - \sum_{i=1}^n Var \left\{ \frac{\partial l_2^F}{\partial \eta} | R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \right\} \right]^{-1} \\ &= \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \eta \partial \alpha} \{ R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \} \\ &\quad \left[- \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \eta^2} \{ R_i, R_i(Y_i, V_i), X_i, \alpha^*, \widehat{\gamma}(\alpha^*), \widehat{\eta}(\alpha^*) \} \right]^{-1}. \end{aligned}$$

References

1. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Values. 2nd edition. Wiley; New York: 2002.

2. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* 1995;90:106–121.
3. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of American Statistical Association* 1998;93:1321–1339.
4. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982;38:963–974. [PubMed: 7168798]
5. Stiratelli R, Laird N, Ware JH. Random-Effects Models for Serial Observations with Binary Response. *Biometrics* 1984;40:961–971. [PubMed: 6534418]
6. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *The Journal of American Statistical Association* 1993;88:9–25.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:11–22.
8. Liang KY, Zeger SL, Qaqish B. Multivariate regression analysis for categorical data (with discussion). *Journal of the Royal Statistical Society, Ser. B* 1992;54:3–40.
9. Diggle, PJ.; Heagerty, PJ.; Liang, KY.; Zeger, SL. *Analysis of Longitudinal Data*. 2nd edition. Clarendon Press; Oxford: 2002.
10. Albert PS. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* 2000;56:602–608. [PubMed: 10877323]
11. Daniels MJ, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 2002;89:553–566.
12. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* 2002;58:342–351. [PubMed: 12071407]
13. Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* 2003;98:555–563.
14. Pourahmadi M, Daniels M. Dynamic conditionally linear mixed models. *Biometrics* 2002;58:225–231. [PubMed: 11890319]
15. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* 1989;8:431–440. [PubMed: 2727467]
16. Buyse M, Molenburghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998;54:1014–1029. [PubMed: 9840970]
17. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002;58:21–29. [PubMed: 11890317]
18. Aalen OO, Frigessi A. What can statistics contribute to a causal understanding? *Scandinavian Journal of Statistics* 2007;34:155–168.
19. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* 1994;43:49–93.
20. Little RJA. Modeling the drop-Out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995;90:1112–1121.
21. Robins JM. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine* 1997;16:21–37.
22. Robins JM. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* 1997;16:39–56.
23. Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. 2001;88:551–564.
24. Cole BF, Bonetti M, Zaslavsky AM, Gelber RD. A multistate Markov chain model for longitudinal, categorical quality-of-life data subject to nonignorable missingness. *Statistics in Medicine* 2006;24:2317–2334. [PubMed: 15977292]
25. Parzen M, Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Troxel A. Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in Medicine* 2006;25:2784–2796. [PubMed: 16345018]
26. Kurland BF, Heagerty PJ. Marginalized transition models for longitudinal binary data with ignorable and non-ignorable drop-out. *Statistics in Medicine* 2004;23:2673–2695. [PubMed: 15316952]

27. Stubbendick AL, Ibrahim JG. Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica* 2006;16:1143–1167.
28. Copas JB, Eguchi S. Local sensitivity approximations for selection bias. *Journal of the Royal Statistical Society* 2001;67:459–513.
29. Copas JB, Eguchi S. Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society* 2005;59:55–95.
30. Troxel AB, Ma G, Heitjan DF. An index of sensitivity to nonignorability. *Statistica Sinica* 2004;14:1221–1237.
31. Ma GG, Troxel AB, Heitjan DF. An index of local sensitivity to nonignorable drop-out in longitudinal modelling. *Statistics in Medicine* 2005;24:2129–2150. [PubMed: 15909292]
32. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics* 2001;57:7–14. [PubMed: 11252620]
33. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods —application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986;7:1393–1512.
34. Robins JM. Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods — application to control of the healthy worker survivor effect”. *Computers and Mathematics with Applications* 1987;14:923–945.
35. Robins JM. Causal inference from complex longitudinal data. In: Berkane, M., editor. *Lecture notes in Statistics: Latent Variable Modeling and Applications to Causality*. Vol. 120. 1997. p. 67-117.
36. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974;66:688–701.
37. Rubin DB. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 1990;5:1151–1172.
38. Rubin DB. Causal inference using potential outcomes: Design, modeling and decisions. *The Journal of American Statistical Association* 2005;100:322–331.
39. Holland PW. Statistics and causal inference (with discussion). *The Journal of American Statistical Association* 1986;81:945–970.

Table 1

Unadjusted point-prevalence of QUIT at each wave by treatment condition

Wave	Rate of Abstinence	
	Control	Intervention
Baseline	0.00% (n=891)	0.00% (n=815)
2 month	8.23% (n=510)	15.50% (n=484)
6 month	11.92% (n=453)	20.90% (n=378)
12 month	18.71% (n=342)	22.02% (n=286)
18 month	24.73% (n=279)	26.94% (n=219)

Table 2

Patterns of Missing Outcome

Index	2 month	6 month	12 month	18 month	Frequency
1	O	O	O	O	336
2	O	O	O	M	135
3	O	O	M	O	65
4	O	O	M	M	156
5	O	M	O	O	35
6	O	M	O	M	46
7	O	M	M	M	221
8	M	O	O	O	38
9	M	O	O	M	21
10	M	O	M	O	18
11	M	O	M	M	62
12	M	M	O	O	6
13	M	M	O	M	11
14	M	M	M	M	556
Total					1706

Table 3

Estimation of MAR missing data mechanism model

Parameter	Estimate	St.err
Intercepts*	-	-
Observed lag-1 QUIT	-0.050	0.06
Observed lag-1 STAGE	-0.035	0.04
Treatment	-0.067	0.06
Race Black	-0.048	0.07
High School Education	-0.25	0.06
Daily Cigarettes Smoked	-0.003	0.004
Year of Smoking	0.022	0.004
Baseline Stage	-0.021	0.03

* Estimates of the 13 intercepts are omitted.

Table 4

Parameter estimates for outcome models under MAR assumption

Parameter	Model I		Model II		Model III	
	Estimate	St.err	Estimate	St.err	Estimate	St.err
Intercept						
2 month	-2.15	0.42	-4.11	0.47	-3.94	0.53
6 month	-1.76	0.42	-3.68	0.50	-3.57	0.51
12 month	-0.91	0.43	-2.88	0.51	-2.80	0.51
18 month	-0.30	0.43	-2.28	0.52	-2.23	0.52
Treatment						
2 month	0.99	0.30	0.89	0.30	0.87	0.29
6 month	0.91	0.37	0.74	0.33	0.71	0.31
12 month	0.06	0.38	-0.004	0.35	-0.03	0.34
18 month	-0.08	0.37	-0.18	0.37	-0.18	0.35
Race Black	-0.24	0.25	-0.38	0.25	-0.37	0.24
High School Education	-0.049	0.23	-0.021	0.22	-0.016	0.21
Daily Cigarettes Smoked	-0.068	0.017	-0.069	0.016	-0.067	0.016
Year of Smoking	-0.079	0.017	-0.077	0.016	-0.073	0.017
Lag-1 Quit	1.21	0.33	-	-	3.60	0.50
Lag-1 Stage	-	-	0.59*	0.08	0.55	0.09
Random Effects Variance	2.23	0.23	2.08	0.18	1.97	0.26

* In this model, STAGE=6 when QUIT=1.

Table 5

Parameter estimates for auxiliary covariate in model III under MAR assumption

categories Parameter	1 vs. 5		2 vs. 5		3 vs. 5		4 vs. 5	
	Est.	Std.err	Est.	Std.err	Est.	Std.err	Est.	Std.err
Intercept	6.77	0.57	3.47	0.68	4.68	0.53	3.48	0.51
Treatment	-0.68	0.23	-0.78	0.29	-0.39	0.19	-0.20	0.18
Race Black	-0.56	0.28	-0.46	0.34	-0.31	0.24	-0.20	0.23
High School Education	0.006	0.23	0.52	0.29	0.17	0.19	0.14	0.18
Daily Cigarettes Smoked	0.028	0.018	0.026	0.021	0.046	0.015	0.048	0.014
Year of Smoking	0.012	0.017	0.056	0.021	0.016	0.014	0.025	0.013
Lag-1 Quit&Stage*	-1.87	0.12	-1.33	0.13	-1.00	0.10	-0.57	0.10
Variance								
Std.err.								
Random Effects	1.21		1.21		1.21		0.19	

* Quit&Stage means STAGE=6 when QUIT=1.

Table 6
Estimated average treatment effects and their sensitivity to nonignorable missing data models

Time	Model	Treatment Effect			Local Sensitivity		
		Estimate	Std.err	Quit	Lag-1 Quit	Stage	Lag-1 Stage
2 month	Naive	7.26%	2.0%	-	-	-	-
	I	6.59%	2.9%	-2.37%	-2.03%	-	-
	II	6.41%	3.0%	-1.85%	-1.77%	-2.45%	-2.97%
	III	6.45%	3.1%	-1.90%	-1.84%	-2.49%	-3.08%
6 month	Naive	8.98%	2.6%	-	-	-	-
	I	7.77%	3.1%	-4.98%	-4.87%	-	-
	II	7.37%	3.3%	-4.73%	-4.87%	-8.39%	-8.57%
	III	7.42%	3.4%	-4.84%	-4.92%	-8.39%	-8.60%
12 month	Naive	3.31%	3.2%	-	-	-	-
	I	2.17%	3.0%	-4.39%	-2.22%	-	-
	II	2.31%	3.2%	-5.54%	-2.16%	-5.49%	-2.34%
	III	2.33%	3.2%	-5.39%	-2.20%	-4.93%	-2.33%
18 month	Naive	2.21%	4.0%	-	-	-	-
	I	0.21%	3.5%	-2.74%	-1.47%	-	-
	II	-0.10%	3.7%	-3.20%	-1.57%	-1.48%	-0.98%
	III	-0.12%	3.7%	-3.30%	-1.66%	-1.41%	-1.08%

Naive: The crude estimates of the rate difference.

Model I: The model without V and with all covariates.

Model II: The model with V and all covariates.

Model III: Model II with the effects of $V=6$ and $V<6$ different for the outcome.