Title
Biased and unbiased estimation in longitudinal studies with informative visit processes

Journal
Biometrics, 72(4)

ISSN
0006-341X

Authors
McCulloch, Charles E
Neuhaus, John M
Olin, Rebecca L

Publication Date
2016-12-01

DOI
10.1111/biom.12501

Peer reviewed

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 December 27.

Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1315–1324. doi:10.1111/biom.12501.

Biased and unbiased estimation in longitudinal studies with informative visit processes

Charles E. McCulloch[1,*], **John M. Neuhaus**[1,**], and **Rebecca L. Olin**[2,***]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco San Francisco, California, U.S.A.

[2]Division of Hematology/Oncology, University of California, San Francisco San Francisco, California, U.S.A.

Summary

The availability of data in longitudinal studies is often driven by features of the characteristics being studied. For example, clinical databases are increasingly being used for research to address longitudinal questions. Because visit times in such data are often driven by patient characteristics that may be related to the outcome being studied, the danger is that this will result in biased estimation compared to designed, prospective studies. We study longitudinal data that follow a generalized linear mixed model and use a log link to relate an informative visit process to random effects in the mixed model. This device allows us to elucidate which parameters are biased under the informative visit process and to what degree. We show that the informative visit process can badly bias estimators of parameters of covariates associated with the random effects, while allowing consistent estimation of other parameters.

Keywords

missing not at random; mixed effects model; bias; multiplicative link

1. Introduction

Calls have been made recently to utilize the wealth of information in large clinical databases to drive biomedical research. For example, the Patient-Centered Outcomes Research Institute (PCORI) recently funded 11 clinical data research networks with the goal (PCORI, 2014) to "...integrate data from ... networks that originate in healthcare systems such as hospitals, health plans, or practise-based networks and securely collect 'real-time,' 'real-world' health information during the routine course of patient care" in order to have, in part, the "capacity to support large-scale comparative effectiveness trials, as well as observational studies of multiple research questions, including prevention and treatment." In contrast to

* chuck@biostat.ucsf.edu. ** john@biostat.ucsf.edu. *** rebecca.olin@ucsf.edu.

Supplemental material

Web Appendices, Tables, and Figures referenced in Sections 2.2, 4.2, and 5 are available with this paper at the Biometrics website on Wiley Online Library. In addition, the code used to conduct the simulation studies, as well as the data and code for the analysis of the hemoglobin data are available.

well-designed longitudinal studies, the availability of data in clinical databases is often driven by patient characteristics. For example, in a study of decline in kidney function in the elderly, a patient may visit their doctor because they are feeling ill and that visit might generate a measurement of kidney function that would be included in the analysis. If the likelihood of a visit is higher among those with lower kidney function – an informative visit process – then it is clear that standard statistical analyses will yield a biased estimate of the average kidney function in the population served by the clinic. But what about trends over time in kidney function? Or the differences between levels of a covariate in trends over time in kidney function? Can those features reliably be estimated from the clinical database? In general the question we set out to answer is the following: using standard statistical analyses that ignore the informative visit process, which parameters are estimated with bias and which are not? Though our motivation has been clinical databases and biomedical research, the results apply more broadly to data collected longitudinally subject to informative visit processes, e.g., data collected through internet sampling.

Previous work has tended to follow two avenues. First, specification of joint models for the visit process and the outcome process (e.g., Lin et al., 2004; Sun et al., 2007; Wulfsohn and Tsiatis, 1997). These require specification of a model for the visit process (which is not of scientific interest) as well as a model for the association between the outcome and visit processes. The joint models specified in the literature have been quite simple. More realistic models would be problematic because they would require accurate specification of the model for reasons why visits occur or do not occur. The variables that govern the presence of visits are often not measured and attempts to fit models to correct for missing data have typically been extremely sensitive to model specification errors (Kenward, 1998).

The second avenue of argument has appealed to the missing data literature, typically requiring the assumption that the data are “missing at random,” or MAR (e.g., Lipsitz et al., 2002; Rathouz, 2004; Fitzmaurice et al., 2006). In the situations we consider, only a very small percentage of possible visits are present and hence a large percentage of data are “missing.” In such a situation the assumption that the data are MAR becomes less and less tenable with the importance of the assumptions becoming more and more critical, and hence the characterization becomes less and less useful. Intuitively, if we have less than 5% missing data in a well-conducted longitudinal study with yearly visits, then it might be reasonable and innocuous to assume that the data are MAR. Contrast this with a study in which data are also scheduled to be collected approximately yearly, but that patients come in and get measured more frequently when they are feeling ill. Because of the haphazard timing of visits we might consider time on a scale of a week, resulting in 95% of the data being missing.

Figure 1 shows the actual visit process from 20 patients being cared for by the University of California, San Francisco bone-marrow transplant clinic. The patients’ hemoglobin levels are being monitored following transplant because of concerns of anemia. The goal is to achieve normal hemoglobin levels (defined as 12 to 15.5 in women and 13.6 to 17.5 in men, as marked on the figure in light gray) and planned follow-up would be scheduled within reasonable windows of 30, 90 and 180 days.

We fit a linear mixed model to the data in Figure 1 with an outcome of hemoglobin, predictors of sex, days post-transplant, whether a visit fell within the scheduled window or not, interactions of days post-transplant with sex and whether the visit was scheduled, and we allowed for correlated random intercepts and slopes (with days post-transplant). The interaction between days post-transplant and whether the visit fell within the scheduled window was statistically significant (p=0.019) with a statistically-significant decline in hemoglobin of 0.025 per day (95% CI −0.041 to −0.010) for the unscheduled visits and a not-statistically-significant decline of 0.011 per day (95% CI −0.023 to 0.0002) for the scheduled visits. This is suggestive that the unscheduled visits may be concentrated in patients with declining hemoglobin levels.

Several features of the process are noteworthy. First, many of the visits are unplanned and are driven by feeling ill or physician concern. Second, there are missed visits. And third, the inter-visit timing is highly irregular and would be quite difficult to model. In many realistic situations such as this one, there will be little to no information available from which to model the informative data process and hypothesizing simplistic modeling approaches may be more harmful than helpful. Instead, our goal is to study the influence of an informative visit process on longitudinal data analysis and the consequences of naive estimation from such data assuming the data were collected in a non-informative manner.

We begin by assuming the outcome process follows a generalized linear mixed model with random effects, a very flexible class of models for a variety of outcome types. Our approach to modeling the informative visit process starts in a fashion similar to other work (e.g., Sun et al., 2007; Liu et al., 2008). Namely, we build in an association between the outcome process and visit process through shared random effects. Because we wish to avoid assuming a specific model for how the random effects are shared, we model this quite flexibly using a novel, log link relationship and derive general results under a wide class of models. In contrast with previous approaches, we do not use information on the visit time process and instead focus on the impact on the distribution of the outcome process, conditioning on the data being observed. Our log link approximations and simulations evaluate the case where many visit times are possible but only a small fraction of the possible visits are observed and, essentially, all the visits are unplanned as in electronic health record data. Our models and simulations hence generate irregular visit patterns.

After defining our models and notation we derive the distribution of the outcome under informative visit processes and a log link. The next section elaborates the effect for a special case of a random slopes and intercepts model and then we evaluate the effects under a more realistic logistic link visit model via simulations.

2. Models and Notation

2.1 Longitudinal outcome process

We begin by defining the model for the outcome, which is assumed to be a generalized linear mixed model, a commonly used model for longitudinal data analysis. We describe our models in terms of “subjects” for the correlated clusters of data and “times” for the observations within a cluster though, of course, the results apply more broadly.

Let Y_{it} represent the measurement at time t on subject i. Our model assumes that observations are conditionally independent given the random effects and that our outcome process follows a generalized linear mixed model with normally distributed random effects, b_i:

$$\begin{aligned} Y_{it}|b_i &\sim \text{independent} \quad f_Y \quad i=1,\ldots,m;t=1,\ldots,n_i \\ g\left(\mathrm{E}\left[Y_{it}|b_i\right]\right) &= x_{it}^T\beta + z_{it}^T b_i \end{aligned} \tag{1}$$

$$b_i \sim \text{i.i.d.} \quad \mathcal{N}\left(0,\Sigma_b\right), \tag{2}$$

In this model, x_{it} represents the covariates associated with subject i at time t, β is the vector of covariate effects, z_{it} is the model matrix for the random effects, and $g(\cdot)$ is the link function.

2.2 Informative visit process

Let R_{it} be a binary indicator with $R_{it} = 1$ indicating that Y_{it} is observed and is 0 otherwise. We assume that, conditional on the random effects, Y_{it} and R_{it} are independent (and independent from one another) and that the probability that $R_{it} = 1$ is dependent on the random effects via a log link:

$$\mathrm{P}\left(Y_{it} \quad \text{is} \quad \text{observed}|b_i\right) \equiv \mathrm{P}\left(R_{it}=1|b_i\right) = exp\left\{\mu_{it}+\gamma_{it}^T b_i\right\}. \tag{3}$$

In this model, γ_{it} governs the strength and directionality of the association between the random effects and whether or not data are observed. The model in (3) is a flexible specification; for example we can allow dependence on where subjects start (i.e., through dependence on their random intercept), their trend over time (i.e., through dependence on a random slope) or on a subject's true mean value (i.e., the mean conditional on the random effects) at the current or previous time points. Importantly, and to allow more realistic models, both μ_{it} and γ_{it} can depend arbitrarily on either fixed or time-varying covariates.

Because the visit process depends on the unobserved random effects, it is a “missing not at random” (Fitzmaurice et al., 2001) process and even methods such as maximum likelihood based fits, which are consistent under “missing at random” assumptions, may be biased. In Section 3.6 we extend our results to the case where we allow dependence of the visit process on *unobserved* previous *outcomes*. This is often a realistic scenario under which virtually no approaches that attempt to model the joint process can succeed.

There is a technical issue with using a log link in (3) for a probability, along with the specification of a normal distribution for the random effects in (2) since the value of the probability is not constrained to be less than 1. To be technically correct the distribution needs to be truncated so that the probability in (3) is less than or equal to 1, That is, the distribution of the random effects needs to be a multivariate normal, truncated so that

$\gamma_{it}^T b_i < -\mu_{it}$. For now we ignore this truncation to exploit the simplification that results in the theoretical calculations. In Section 5 we give simulation results under a logit link informative visit process (which naturally constrains probabilities to be less than one) and in the online supplemental material we give exact calculations under truncation as well as more detailed simulation results under both log and logit link informative visit processes.

3. Observed data model

3.1 Conditional distributions of random effects for the observed data

Our goal is to elucidate the consequences of analyzing the available data, so we are led to consideration of the conditional distribution of the outcome process, given $R_{it} = 1$. This, in turn, leads to consideration of the distribution of the random effects conditional on $R_{it} = 1$.

From (3) and (2) we can obtain the joint probability of b_i being less than s and $R_{it} = 1$:

$$\begin{aligned} \mathrm{P}\left(R_{it}=1, b_i \leq s\right) &= \mathrm{E}\left[\mathrm{P}\left(R_{it}=1, b_i \leq s | b_i=u\right)\right] \\ &= \mathrm{E}\left[\mathrm{P}\left(R_{it}=1 | b_i=u\right) I_{\{u<s\}}\right] \\ &= \int_{-\infty}^{s} exp\left\{\mu_{it}+\gamma_{it}^T u\right\} \frac{exp\left\{-u^T \Sigma_b^{-1} u/2\right\}}{(2\pi)^{-q/2}|\Sigma_b|^{-1/2}} \mathrm{d}u, \end{aligned} \quad (4)$$

where q is the dimension of b_i.

The conditional *density* of b_i given $R_{it} = 1$ can be derived by differentiating (4) with respect to s and dividing by the marginal probability that $R_{it} = 1$:

$$f_{b_i|R_{it}=1}(b) = exp\left\{\gamma_{it}^T b\right\}\ exp\left\{-b^T \Sigma_b^{-1} b/2\right\} / \int_{-\infty}^{\infty} exp\left\{\gamma_{it}^T u\right\}\ exp\left\{-u^T \Sigma_n^{-1} u/2\right\} du \quad (5)$$

The denominator in (5) is constant in b so completing the square in the numerator leads to

$$f_{b_i|R_{it}=1}(b) \propto exp\left\{-\left(b-\Sigma_b\gamma_{it}\right)^T \Sigma_n^{-1}\left(b-\Sigma_b\gamma_{it}\right)/2\right\}. \quad (6)$$

We thus have that the conditional distribution of b_i given $R_{it} = 1$ is given by

$$b_i | R_{it}=1 \sim \mathcal{N}\left(\Sigma_b\gamma_{it}, \Sigma_b\right). \quad (7)$$

That is, the distribution is multivariate normal with variance Σ_b (both the same as the unconditional distribution), but with a mean given by $\Sigma_b\gamma_{it}$ instead of 0.

3.2 Conditional distributions for the observed data

We are now in a position to derive the conditional distribution of Y_{it}, conditional on being observed. Our calculations are for the "marginal" distribution of each individual Y_{it} (as opposed to the joint distribution of Y). The result concerning the random effects above

generalizes to Y_{it}, namely that the distribution of Y_{it} conditional on $R_{it} = 1$ is the same as the unconditional distribution of Y_{it} except that the mean of Y_{it} is affected by the mean of b_i given in (7). To see this, first recall that the distributions of Y_{it} and R_{it} are defined independently of one another, conditional on b_i. Therefore, using bracket notation for distributions, we have $[Y_{it} | R_{it} = 1, b_i] = [Y_{it} | b_i]$, and the conditional distribution of Y_{it} given $R_{it} = 1$ is

$$\begin{aligned}[Y_{it}|R_{it}=1] = & \int [Y_{it}|R_{it}=1, b]\,[b|R_{it}=1]\,\mathrm{d}b \\ = & \int [Y_{it}|b_i]\,[b_i|R_{it}=1]\,\mathrm{b}b_i. \end{aligned} \quad (8)$$

That is, the distribution of Y_{it} given $R_{it} = 1$ is a convolution of the same, non-informative, conditional distribution of Y_{it} given b from (1) with the conditional distribution of b given $R_{it} = 1$. Since the conditional distribution of b given $R_{it} = 1$ is the same as its unconditional distribution (except for its mean), the only influence conditioning on $R_{it} = 1$ has is to modify the mean of the random effects distribution.

Furthermore, because both the random and fixed effects enter the linear predictor in (1), we can move the mean of the conditional distribution of b_i into the fixed effects portion of the model and re-center the distribution of b_i given $R_{it} = 1$ to have mean 0. The importance of this result is that the distribution of the observed data is exactly the same as that of a non-informative visit process with fixed effects given by $x_{it}^T\beta + z_{it}^T\Sigma_b\gamma_{it}$. This allows calculation of the exact distribution of the outcome or aspects of that distribution in a number of special cases that we consider in the following subsections. One important result is immediate from the form of differences induced in the marginal distribution. Namely that μ_{it} in (3) does not impact the marginal distribution under the log link informative visit process. Since μ_{it} can depend arbitrarily on the covariates, dependence of the informative visit process on the covariates alone does *not* influence the marginal distribution.

3.3 Linear mixed model

Using the result above, and for a linear mixed model with Y_{it} having a distribution (conditional on b_i) that is normal with variance σ_ϵ^2, we immediately have the distribution of Y_{it} conditional on being observed:

$$Y_{it}|R_{it}=1 \sim \mathcal{N}\left(x_{it}^T\beta + z_{it}^T\Sigma_b\gamma_{it}, \sigma_e^2 + z_{it}^T\Sigma_b z_{it}\right).$$

This is practical because we can determine the asymptotic limit of many estimation methods (for example, ordinary least squares or generalized estimating equations with an independence working correlation structure) by simply examining the mean. To determine consistency for estimating β, we can often simply compare $x_{it}^T\beta + z_{it}^T\Sigma_b\gamma_{it}$ with $x_{it}^T\beta$. Coefficients of covariates that do not enter in the second term, $z_{it}^T\Sigma_b\gamma_{it}$, may be consistently estimated.

3.4 Mean under log link models

For models with a log link for the outcome and arbitrary distributions (conditional on b_i) we can easily calculate the mean when the random effects are normally distributed. Without selection the mean is given by $\mathrm{E}[Y_{it}] = exp\left\{x_{it}^T\beta + z_{it}^T\Sigma_b z_{it}/2\right\}$. On the other hand, with selection and using the results above, we have $\mathrm{E}[Y_{it}|R_{it}=1] = exp\left\{x_{it}^T\beta + z_{it}^T\Sigma_b z_{it}/2 + z_{it}^T\Sigma_b\gamma_{it}\right\}$. Therefore the difference in the log of the mean with and without selection is given by $log\,\mathrm{E}[Y_{it}|R_{it}=1] - log\,\mathrm{E}[Y_{it}] = z_{it}^T\Sigma_b\gamma_{it}$.

3.5 Probit models

For probit models, the mean of Y_{it} without selection is (McCulloch et al., 2008) $\mathrm{E}[Y_{it}] = \Phi\left(\lambda x_{it}^T\beta\right)$, where $\lambda = 1/\sqrt{1+z_{it}^T\Sigma_b z_{it}}$. With selection we have $\mathrm{E}[Y_{it}|R_{it}=1] = \Phi\left(\lambda x_{it}^T\beta + \lambda z_{it}^T\Sigma_b\gamma_{it}\right)$. Therefore the difference, on the probit scale, of the mean with and without selection is given by $\Phi(\mathrm{E}[Y_{it}|R_{it}=1]) - \Phi(\mathrm{E}[Y_{it}]) = \lambda z_{it}^T\Sigma_b\gamma_{it}$. Since the outcome is binary, the outcome model under selection is still a probit model but with a shift of the mean (on the probit scale) of $\lambda z_{it}^T\Sigma_b\gamma_{it}$. Given that probit models closely approximate logistic models, this suggests that logit models will exhibit similar patterns of bias. We report simulation results for a logistic outcome model in Section 5.

3.6 Linear mixed model with dependence on previous outcomes

The results of Section 3.2 can be extended to dependence on previous outcomes by essentially the same arguments in the case of a linear mixed model. Suppose the probability of observing an outcome is dependent on the value of the outcome lagged by τ time units:

$$\begin{aligned}\mathrm{P}(Y_{it} \text{ is observed}|Y_{i,t-\tau}) &= exp\{\alpha+\delta Y_{i,t-\tau}\} \\ &= exp\left\{\alpha+\delta\left(x_{i,t-\tau}^T\beta + z_{i,t-\tau}^T b_i + \epsilon_{i,t-\tau}\right)\right\}, \end{aligned} \quad (9)$$

with ε_{it} being the error term in the linear mixed model for subject i at time t. Then the joint probability of b_i being less than s, $\varepsilon_{i,t-\tau}$ being less than s_ε, and $R_{it} = 1$ is given by:

$$\begin{aligned}\mathrm{P}(R_{it}=1, b_i \le s, \epsilon_{i,t-\tau} \le s_\epsilon) &= \mathrm{E}[\mathrm{P}(R_{it}=1, b_i \le s, \epsilon_{i,t-\tau} \le s_\epsilon|b_i=u, \epsilon_{i,t-\tau}=\epsilon)] \\ &= \mathrm{E}[\mathrm{P}(R_{it}=1|b_i=u, \epsilon_{i,t-\tau}=\epsilon)\, I\{u \le s, \epsilon \le s_\epsilon\},] \\ &= \int_{-\infty}^{s_\epsilon}\int_{-\infty}^{s} exp\left\{\alpha+\delta\left(x_{i,t-\tau}^T\beta + z_{i,t-\tau}^T u + \epsilon\right)\right\} \times \frac{exp\left\{-u^T\Sigma_b^{-1}u/2\right\}}{(2\pi)^{-q/2}|\Sigma_b|^{-1/2}} \frac{exp\left\{-\epsilon^2/2\sigma_\epsilon^2\right\}}{\sqrt{2\pi\sigma_\epsilon^2}} \mathrm{d}u\mathrm{d}\epsilon. \end{aligned}$$

(10)

Integrating out ε from (10) and using the same argument as before shows that the conditional density of b_i given $R_{it} = 1$ is $b_i|R_{it}=1 \sim \mathcal{N}(\delta\Sigma_b z_{i,t-\tau}, \Sigma_b)$.

Also, using assumed independence of errors, ε_{it} is independent of $\varepsilon_{i,t-\tau}$ and b_i, so that the conditional distribution of ε_{it} is unchanged by conditioning on $\{R_{it}=1\}$. We thus have the following result for a linear mixed model:

$$Y_{it}|R_{it}=1 \sim \mathcal{N}\left(x_{it}^T\beta+\delta z_{it}^T\Sigma_b z_{i,t-\tau}, \sigma_e^2+z_{it}^T\Sigma_b z_{it}\right).$$

The results derived previously thus carry over to this situation. That is, the marginal distribution of Y_{it} in the observed data is unchanged except for a modification of the mean which depends on the model matrix of the random effects. So for covariates that are unrelated to the random effects we can expect little or no bias.

4. Consequences of the observed data process for a random intercepts and slopes model

4.1 A random intercepts and slopes linear mixed model

We next consider the impact of selection for a common mixed model used in the longitudinal context: a random intercepts and slopes linear mixed model. In this model Z consists of subject-specific intercepts and slopes for one of the variables in X. To better understand the consequences of the calculations above we work out the details for the random intercept and slope model under three informative visit process models: dependence on random intercept only, random slope only, and conditional mean. In both this section and the next we use a common longitudinal data model, incorporating a subject-specific "time" variable, x_1, a treatment variable, x_2, which is 1 for a treatment group and zero otherwise, and a time by treatment interaction variable, $x_3 = x_1 \times x_2$. We incorporate the random slopes as slopes over time (and so associated with x_1):

$$Y_{it}|b_i \sim \text{independent} \quad \mathcal{N}\left(\mathrm{E}\left[Y_{it}|b_i\right], \sigma_\epsilon^2\right) \quad i=1,\ldots,m; t=1,\ldots,n_i$$

$$\begin{aligned} \mathrm{E}\left[Y_{it}|b_i\right] = &\quad (\beta_0+b_{0i})+(\beta_1+b_{1i})\,x_{1it}+\beta_2 x_{2i}+\beta_3 x_{3it} \\ b_i \sim &\quad \text{i.i.d} \quad \mathcal{N}\left(0,\Sigma_b\right), \end{aligned} \tag{11}$$

with $var\left(b_{0i}\right)=\sigma_0^2$, $var\left(b_{1i}\right)=\sigma_1^2$, and cov($b_{0i}$, b_{1i}) = σ_{01}. For this model, the t^{th} row of of Z_i is given by $z_{it}^T=(1 \quad x_{1it})$. In most of this section we consider only the linear mixed model given in (11) but indicate generalizations in Section 4.4.

4.2 Dependence on the conditional mean

In this section we consider a model where the probability of observing an observation is dependent on the true state of the subject at time t. In the Supplementary Material we also give results for an informative visit process that depends directly on the random intercepts and slopes. Dependence on the true mean is incorporated through the conditional value of the linear predictor and a log link:

$$\begin{aligned} \mathrm{P}\left(R_{it}=1|b_i\right) = & \; exp\left\{\mu+\delta \mathrm{E}\left[Y_{it}|b_i\right]\right\} \\ = & \; exp\left\{\mu+\delta\left(\beta_0+\beta_1 x_{1it}+\beta_2 x_{2i}+\beta_3 x_{3it}\right)+\delta b_{0i}+\delta x_{1it} b_{1i}\right\} \\ \equiv & \; exp\left\{\mu_{it}+\delta b_{0i}+\delta x_{1it} b_{1i}\right\}, \end{aligned} \quad (12)$$

so this fits into the general formulation, (3), with $\mu_{it} = \mu+\delta(\beta_0+\beta_1 x_{1it}+\beta_2 x_{2i}+\beta_3 x_{3it})$, $\gamma_{1it} = \delta$, and $\gamma_{2it} = \delta x_{1it}$. For this model, the result of Section 3.3 gives the expected value of the linear predictor conditional on being observed as

$$\begin{aligned} x_{it}^T\beta+z_{it}^T\Sigma_b\gamma_{it} = & \; \beta_0+\beta_1 x_{1it}+\beta_2 x_{2i}+\beta_3 x_{3it}+\delta\sigma_0^2+\delta\sigma_{01}x_{1it}+\delta\sigma_{01}x_{1it}+\delta\sigma_1^2 x_{1it}^2 \\ = & \; \left(\beta_0+\delta\sigma_0^2\right)+\left(\beta_1+2\delta\sigma_{01}\right)x_{1it} \end{aligned} \quad (13)$$

$$+\beta_2 x_{2i}+\beta_3 x_{3it} \quad (14)$$

$$+\delta\sigma_1^2 x_{1it}^2. \quad (15)$$

Under this informative visit model, the results in (13)-(15) indicate that:

- The estimators of β_0 and β_1 will be biased due to the extra terms in (13);
- The functional dependence of the mean on x_2 and x_3 will be unaffected by the informative visit process (from (14));
- An additional functional relationship (quadratic in x_1) is introduced, which could further bias estimation of the βs if it is not accommodated.

Because the functional dependence on x_2 and x_3 is unaffected by the selection process, these results also indicate that the estimation of β_2 and β_3 may be unbiased if the additional spurious relationship with x_1 is accommodated in the model.

4.3 Multiplicative link functions

In this section we show that some of the results hold for more general link functions than that specified by (3). We hypothesize a model where the conditional probability of a visit is the product of two terms with the first term depending on both the random effects and x_1 and any dependence on other covariates is in the second term:

$$\mathrm{P}\left(R_{it}=1|b_i\right) \equiv p_1\left(x_{1it}, b_i\right) p_2\left(x_{2i}, x_{3it}\right). \quad (16)$$

This class encompasses models that may be more realistic than the log link model, for example both $p_1(\cdot)$ and $p_2(\cdot)$ could be logistic in form, constraining the probabilities to be between 0 and 1. As before, we can obtain the joint probability of b_i being less than s and $R_{it} = 1$:

$$\begin{aligned} \mathrm{P}\left(R_{it}{=}1, b_i \leq s\right) &= \mathrm{E}\left[\mathrm{P}\left(R_{it}{=}1, b_i \leq s | b_i{=}u\right)\right] \\ &= \mathrm{E}\left[\mathrm{P}\left(R_{it}{=}1 | b_i{=}u\right) I_{\{u<s\}}\right] \end{aligned} \quad (17)$$

$$= \int_{-\infty}^{s} p_1\left(x_{1it}, u\right) p_2\left(x_{2i}, x_{3it}\right) f_b\left(u\right) \mathrm{d}u. \quad (18)$$

The joint density of b_i and $R_{it} = 1$ is therefore:

$$f_{b_i, R_{it}=1}\left(b\right) = p_1\left(x_{1it}, b\right) p_2\left(x_{2i}, x_{3it}\right) f_b\left(b\right), \quad (19)$$

and the marginal probability of $R_{it} = 1$ is the integral of (19) with respect to b. Therefore the conditional distribution of b_i given $R_{it} = 1$ is given by

$$\begin{aligned} f_{b_i | R_{it}=1} &= p_1\left(x_{1it}, b\right) p_2\left(x_{2i}, x_{3it}\right) f_b\left(b\right) / \int_{-\infty}^{\infty} p_1\left(x_{1it}, u\right) p_2\left(x_{2i}, x_{3it}\right) f_b\left(u\right) \mathrm{d}u \\ &= p_1\left(x_{1it}, b\right) f_b\left(b\right) / \int_{-\infty}^{\infty} p_1\left(x_{1it}, u\right) f_b\left(u\right) \mathrm{d}u \end{aligned} \quad (20)$$

That is, there is no functional dependence of the conditional distribution of b on either x_2 or x_3. Therefore the mean of Y in a linear mixed model, conditional on $R_{it} = 1$ has the same functional dependence on x_2 and x_3 as the unconditional mean. We therefore expect that fitting a model while ignoring the selection process will yield consistent estimation of β_2 and β_3, perhaps after accommodating spurious relationships in x_1.

4.4 Probit and log link outcome models

The results derived in the above subsections for the linear mixed model generalize in a straightforward way to the log and probit link models in Sections 3.4 and 3.5, albeit on the log or probit scales. For example, under the conditional mean dependence of Section 4.2, the probit model will have (on the probit scale) additional terms associated with the intercept and x_1 as well as a spurious quadratic relationship in x_1.

How these results will affect maximum likelihood fitting, which is based on the entire *joint* distribution rather than the marginal distribution under selection, is not immediate. However, the performance of fitting methods such as generalized estimating equations with a working independence structure *will* be governed by the marginal mean structure and the bias results from Section 4.2 will apply. This then suggests that maximum likelihood fits will similarly be affected, which we check using simulations, described in the next section.

5. Simulations

Since the log link theoretical results are only an approximation and apply to the marginal distribution, we conducted a simulation study to verify that they held under visit processes

with reasonable degrees of informativeness and to compare the results to a more natural logit link instead of the log link in (3), namely

$$\mathrm{P}\left(R_{it}=1|b_i\right)=1/\left[1+exp\left\{-\left(\mu_{it}+\gamma_{it}^{T}b_i\right)\right\}\right]. \quad (21)$$

To do so, we simulated data with two different outcome distributions and used the linear predictor in Section 4.2, namely a model with an intercept (β_0), time effect (β_1), group effect (β_2), and a group by time interaction (β_3) and random intercepts, b_{0i}, and slopes, b_{1i}, with time. The first model was a linear mixed model, (11), with covariances for b_{0i}, b_{1i}, and ε_{it} of $\sigma_0^2=\sigma_1^2=\sigma_\epsilon^2=1$, σ_{01} = 0 or 0.5, and fixed effect parameters $\beta_k = k$. Using common random numbers, we simulated informativeness using both (3) and (21) and using the informative visit models in Sections 4.2 and 4 of the Supplemental Material. We simulated 3000 subjects with up to 25 visits per subject, though the number of subjects in any simulation replication was much lower because many subjects have no visits.

We also simulated data from a logistic model, that is a logit link and Bernoulli distribution in (1), again under both a logit and log link informative visit process and using parameters β_0 = –1, β_1 = 0.5, β_2 = 1, β_3 = 0.5, $\sigma_0^2=\sigma_1^2=1$, σ_{01} = 0 or 0.5, and using 3000 subjects and up to 10 visits.

We simulated data ranging from no “informativeness” to a high degree of informativeness. To determine the upper range of informativeness we aimed to have about a five-fold ratio of $P(R_{it} = 1 \mid b_i)$ as the random effect distribution in (3) ranged from its 25^{th} to 75^{th} percentiles. This would lead to more than a 10-fold ratio going from an observation that is one standard deviation below normal compared to an observation that is one standard deviation above normal and more than a 100-fold ratio comparing observations two standard deviations below to two standard deviations above normal.

Our first informative visit model has dependence on the conditional mean of Y:

$$log\ P\left(R_{it}=1|b_i\right)=\alpha+\delta\mathrm{E}\left[Y_{it}|b_i\right]. \quad (22)$$

Using the outcome model described above, the standard deviation of $\mathrm{E}[Y_{it}|b_i]$ is a little less than 2.5. To achieve the five-fold ratio would require δ of about 0.6. Accordingly we simulated values of δ of 0, 0.25, 0.5, and 0.75 and used α = –5 for the linear mixed model and α = –1 for the Bernoulli outcome model.

Our second informative visit model, which we used only for the linear mixed model, has dependence directly on the random effects:

$$log\ P\left(R_{it}=1|b_i\right)=\mu+\gamma_0 b_{0i}+\gamma_1 b_{1i}. \quad (23)$$

In this model, if $\gamma_0 = \gamma_1 = \gamma$ and if the random effects were uncorrelated then the value of γ giving a five-fold difference would be 0.84. Accordingly, for this model we simulated values of $\gamma_1 = 0$, 0.5 or 1, $\sigma_{01} = 0$ or 0.5 and we used $\mu = -4$.

Under each of these scenarios we fit a random intercepts and slopes model (allowing separate variances and a covariance) using maximum likelihood and also fit an independence generalized estimating equations approach (i.e., ordinary least squares fit for the linear mixed model or logistic regression fit for the Bernoulli model). For the simulations under the conditional mean informative model we also included a quadratic term in x_1 to accommodate the functional dependence noted in Section 4.2. All simulations were conducted in Stata 13.1 (StataCorp, College Station, TX) and used 500 replications. We report illustrative results in figures below, but the full set of simulation results (including standard errors) is available in tabular form in the online supplemental material.

Figure 2 shows the results for the linear mixed model under the conditional mean informative visit model, (22), as well what is predicted by theory based on the log link model (ignoring truncation), and are as expected. The estimators of β_0 and β_1 are badly biased as the degree of informativeness increases. Further, within the range of low to moderate informativeness, the estimators of β_2 and β_3, the parameters unconnected to the random effects, exhibit little bias. There is slight bias at the upper ranges of informativeness for both the mixed model and GEE independence fits. Under strong degrees of informativeness, the approximations needed for the log link theory to apply (namely that the probabilities are less than 1) are violated. For the log link (shown in the supplemental material) there is still slight bias for the mixed model fit under $\delta = 0.75$, but the GEE independence fit does not exhibit bias, as predicted by the theory. The difference between ML and GEE may be because our results apply to the marginal distribution (on which GEE depends) but the ML fits utilize the entire joint distribution.

Figure 3 shows the results for the linear mixed model under the random effects informative visit model (23). The results again match well with the theory (see Section 4 of the Supplemental Material), especially qualitatively: large degrees of bias in the estimators of β_0 and β_1 and little or no bias for β_2 and β_3. There are minor discrepancies between the quantitative results predicted by the theory in the most extreme degrees of informativeness (again due to the simulation being conducted under the logit link).

Figure 4 shows the results for the logistic outcome model under the conditional mean informative visit model, (22). The results are similar to that of the linear mixed model. The estimators of β_0 and β_1 exhibit a large degree of bias as the degree of informativeness increases. Further, within the range of low to moderate informativeness, the estimators of β_2 and β_3, the parameters unconnected to the random effects, exhibit little bias. There is slight bias at the upper ranges of informativeness for both the mixed model and GEE independence fits. Note that on Figure 4 the “true” value represents the subject-specific parameter from the generalized linear mixed model and the GEE independence estimator is instead estimating the population-averaged parameter.

As noted at the end of Section 3.2 and under the log link visit model, the marginal distribution of an individual Y_{it} is unchanged with arbitrary dependence of the visit process on covariates, which can be absorbed as part of μ_{it} in (3). To check to see if this also held under the logistic link, (21), we redid the simulation of Figure 1 but allowing additional dependence of the visit process on the group variable, x_2. The results were virtually unchanged even under strong dependence on x_2, as predicted by the log link theory. Details are given in the Supplementary Material.

6. Discussion

In this paper we developed theory for the marginal distribution of data generated under a generalized linear mixed model but subject to a novel log link informative visit process. That visit process allowed dependence of the probability of a visit on the random effects in the mixed model. We used the log link because it allowed approximate, theoretical quantification of the bias under the informative visit process; this was supplemented by simulation results using a logit link that gave very similar results.

Broadly speaking, the theory and simulation studies indicate that estimators of parameters associated with the random effects will be badly biased but that those not associated with the random effects will be estimated with little or no bias. The lack of bias with estimated covariate effects unconnected to the random effects is similar to results we and others have demonstrated in the informative cluster size literature (e.g., Williamson et al., 2003; Neuhaus and McCulloch, 2011). However, we did not see the severe degree of bias in that context that we see here.

Our work is similar to the investigation of bias in selection models in the econometric literature. For example, Heckman (1979) studied the effect of selection on the mean in linear models using the equivalent of a probit link instead of our log link. This gives bias for the mean in terms of ratios of normal p.d.f.s and c.d.f.s (inverse Mill's ratios) which are, in turn, complicated functions of the variance components and covariates. Our approach has two main advantages: 1) the log link gives easy to understand bias terms as contrasted with the effects imbedded in inverse Mill's ratios and 2) in dealing with generalized linear mixed models in which the variance-covariance structure influences the marginal mean, we must derive the impact of selection on the entire marginal distribution, not just the mean.

Because the log link visit process does not constrain the visit probabilities to be less than 1 and because the theoretical results apply only to the marginal distribution, we also conducted simulations of the performance of standard analyses (mixed model fits and GEE independence fits) to longitudinal data under a more realistic logit link informative visit process. To a large degree the results mirrored the theory. First, estimators of parameters associated with the random effects were badly biased. Second, under low to moderate degrees of informativeness, the estimators of parameters unassociated with the random effects exhibited little or no bias. However, for large degrees of informativeness, both methods exhibited slight bias.

The results herein indicate that analysis of data that may be subject to informative visit processes should be undertaken with care. Investigation of the random effects structure can provide guidance as to which parameter estimates may be biased (those for covariates associated with the random effects) and therefore interpreted with caution. On the positive side, parameter estimates not associated with the random effects are not likely to be biased due to the informative visit process.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Support was provided by NIH grant R01 CA82370 and PCORI contract ME-1306-01466.

References

Fitzmaurice GM, Laird NM, Shneyer L. An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. Stat Med. 2001; 20:1009–1021. [PubMed: 11276032]

Fitzmaurice GM, Lipsitz SR, Ibrahim JG, Gelber R, Lipshultz S. Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. Biostatistics. 2006; 7:469–485. [PubMed: 16428260]

Heckman JJ. Sample selection bias as a specification error. Econometrica. 1979; 47:153–161.

Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. Statistics in Medicine. 1998; 23:2723–32. [PubMed: 9881418]

Lin H, McCulloch CE, Rosenheck RA. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics. 2004; 60:295–305. [PubMed: 15180654]

Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. Bio-metrics. 2002; 58:621–630.

Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observation times and a dependent terminal event, with application to medical cost data. Biometrics. 2008; 64:321–327.

McCulloch, C.; Searle, S.; Neuhaus, J. Generalized, Linear and Mixed Models. 2nd Ed.. Wiley; New York: 2008.

Neuhaus JM, McCulloch CE. Estimation of covariate effects in generalised linear mixed models with informative cluster sizes. Biometrika. 2011; 98:147–162. [PubMed: 23049125]

PCORI. Patient-centered outcomes research institute cooperative agreement funding announcement: Improving infrastructure for conducting patient-centered outcomes research. 2014. (http://www.pcori.org/assets/pcori-cdrn-funding-announcement-042313.pdf)

Rathouz PJ. Fixed effects models for longitudinal binary data with drop-outs missing at random. Statistica Sinica. 2004; 14:969–988.

Sun J, Sun L, Liu D. Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times. Journal of the American Statistical Association. 2007; 102:1397–1406.

Sun J, Tong X, He X. Regression Analysis of Panel Count Data with Dependent Observation Times. Biometrics. 2007; 63:1053–1059. [PubMed: 18078478]

Williamson J, Datta S, Satten G. Marginal analyses of clustered data when cluster size is informative. Biometrics. 2003; 59:36–42. [PubMed: 12762439]

Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997; 53:330–339. [PubMed: 9147598]

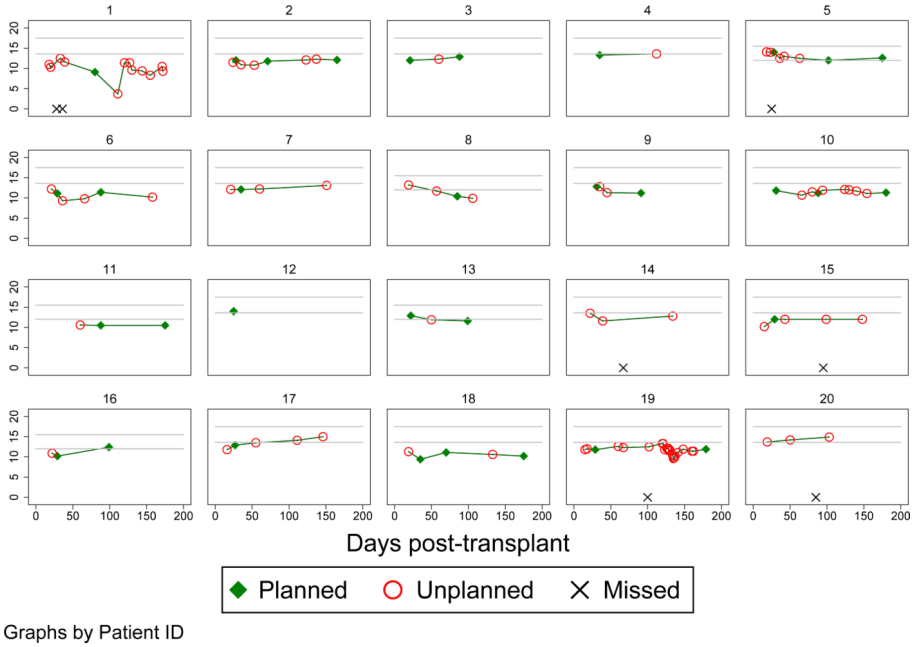

Figure 1.
Hemoglobin levels versus days post bone-marrow transplant with different types of visits.

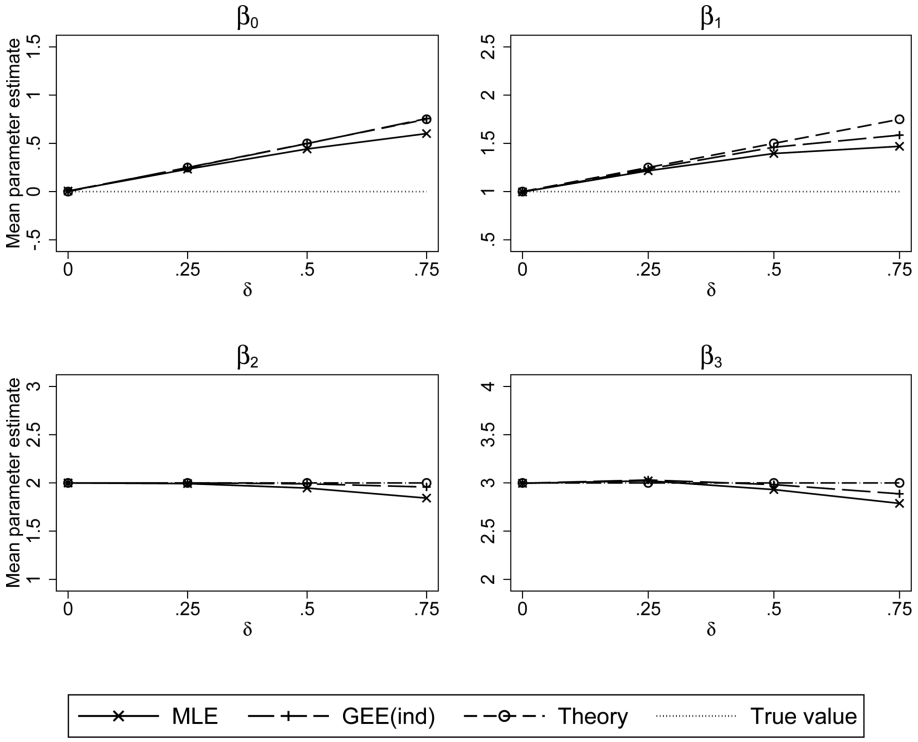

Figure 2.

Simulated mean values of the maximum likelihood (MLE) and GEE-independence regression coefficient estimators. Simulated under a conditional mean informative visit process with a logit link, i.e., logit($P(R_{it} = 1)$) = –5 + δE[Y | b], and linear mixed outcome model with random intercepts and slopes.

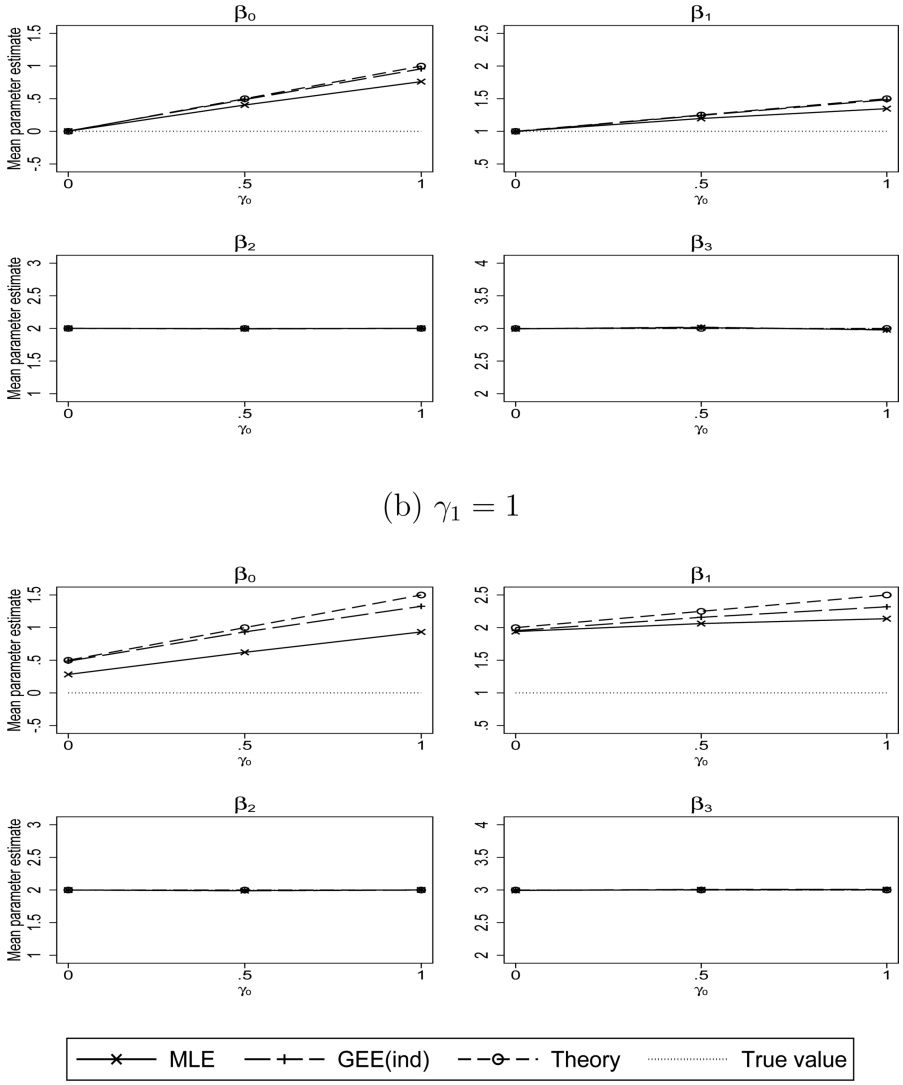

Figure 3.

Simulated mean values of the maximum likelihood (MLE) and GEE-independence regression coefficient estimators. Simulated under a random effects informative visit process with a logit link, i.e., logit($P(R_{it} = 1)$) = –5 + $\gamma_0 b_0$ + $\gamma_1 b_1$, and linear mixed outcome model with random intercepts, b_0, and random slopes, b_1.

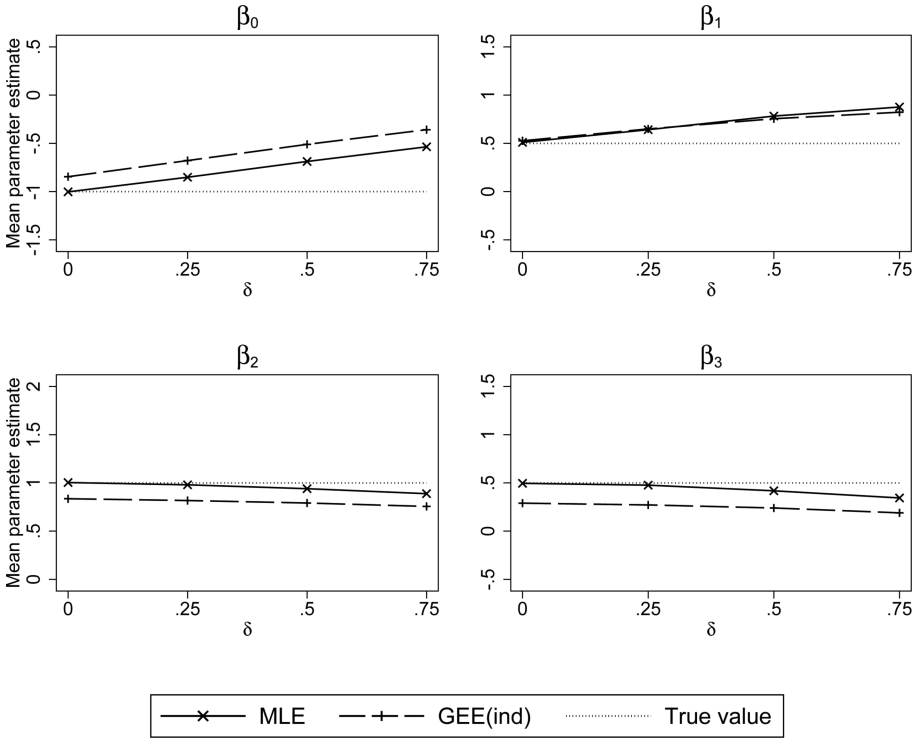

Figure 4.

Simulated mean values of the maximum likelihood (MLE) and GEE-independence regression coefficient estimators. Simulated under a conditional mean informative visit process with a logit link, i.e., logit($P(R_{it} = 1)$) = –1 + δE[Y | b], and logistic mixed outcome model with random intercepts and slopes.