STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:4359–4380
Published online 9 July 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3352

Marginalized models for longitudinal ordinal data with application to quality of life studies

Keunbaik Lee[1] and Michael J. Daniels[2, 3, *, †]

[1]*Biostatistics Program, School of Public Health, Louisiana State University Health Science Center, New Orleans, LA 70112, U.S.A.*
[2]*Department of Epidemiology and Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*
[3]*Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.*

SUMMARY

Random effects are often used in generalized linear models to explain the serial dependence for longitudinal categorical data. Marginalized random effects models (MREMs) for the analysis of longitudinal binary data have been proposed to permit likelihood-based estimation of marginal regression parameters. In this paper, we introduce an extension of the MREM to accommodate longitudinal ordinal data. Maximum marginal likelihood estimation is implemented utilizing quasi-Newton algorithms with Monte Carlo integration of the random effects. Our approach is applied to analyze the quality of life data from a recent colorectal cancer clinical trial. Dropout occurs at a high rate and is often due to tumor progression or death. To deal with progression/death, we use a mixture model for the joint distribution of longitudinal measures and progression/death times and principal stratification to draw causal inferences about survivors. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: marginalized likelihood-based models; ordinal data models; dropout

1. INTRODUCTION

Longitudinal data are repeated measurements from the same subject observed over time. The within-subject measurements (over time) are typically not independent. Although serial correlation may not be of primary interest but it must be taken into account to make proper inferences. In marginal models, the population-averaged effect of covariates on the longitudinal response is directly specified [1, 2] and the regression coefficients have interpretation for the population

*Correspondence to: Michael J. Daniels, Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.
†E-mail: mdaniels@stat.ufl.edu

Contract/grant sponsor: Publishing Arts Research Council; contract/grant number: 98-1846389
Contract/grant sponsor: NIH; contract/grant numbers: CA85295, HL079457

Received 25 September 2007
Accepted 6 May 2008

rather than for any individual [3]. In conditional models, the effect of covariates on responses is specified conditional on random effects or previous history of responses. Hence, the population-averaged effect of covariates is indirectly specified [4, 5]. In this paper, we consider marginal model approaches.

Properly specified probability models lead to efficient estimation even under missing at random (MAR) [6] and nested models can be compared using likelihood ratio tests and non-nested models by penalized criteria such as Akaike information criterion (AIC) [7] or Bayesian information criterion [8]. Recently, marginalized likelihood-based models have been developed for the analysis of longitudinal categorical data [9–13]. Heagerty [9, 10] proposed marginally specified logistic-normal models and marginalized transition models (MTM) for longitudinal binary data. In both models, a marginal logistic-regression model was used for explaining the average response. The models were specified by introducing random effects in the logistic-normal models and Markov dependence for MTM to explain the within-subject dependence. Miglioretti and Heagerty [12] developed marginalized multilevel models for longitudinal binary data in the presence of time-varying covariates. Lee and Daniels [13] extended Heagerty's work to accommodate longitudinal ordinal data using Markov dependence. Marginalized models have advantages over conditional models. First, the interpretation of regression coefficients does not depend on specification of the dependence in the model unlike in conditional models. In addition, estimation of covariate effects has been shown to be more robust to mis-specification of dependence [10, 11, 13].

Models for correlated ordinal data typically fall into two classes based on how the dependence is modeled: via the global odds ratio [14–17] or via random effects [18–20]. A general overview of models for ordinal categorical data can be found in Liu and Agresti [21]. The main contribution of this paper will be to introduce a new marginalized model for longitudinal ordinal data.

A common issue in inference from longitudinal studies is potential biases introduced by missing data. Classes of models to accommodate longitudinal data with dropout are summarized in Hogan *et al.* [22]. Standard approaches to handle missing data implicitly 'impute' values of response after dropout. For quality of life (QOL) data, if a subject drops out due to death, the QOL will not be defined after the dropout time. One way to address the type of dropout is to model the joint distribution of the longitudinal responses and progression/death times [23–25]. Hogan and Laird [23] used a mixture model for the joint distribution of longitudinal measures and progression/death times. Pauler *et al.* [24] proposed a pattern mixture model (PMM) for longitudinal QOL data with non-ignorable missingness due to dropout and censorship by death. Recently, Kurland and Heagerty [25] explored regression models conditioning on being alive as a valid target of inference. They used regression models that condition on survival status rather than a specific survival time. We will use the ideas in Hogan and Laird [23] similar to the previous work by Pauler *et al.* [24].

We implement a principal stratification approach [26, 27] here to make inference on the causal effect of the treatment on QOL among (potential) survivors on both treatment arms. Frangakis and Rubin [26] discussed causal effects in studies where the outcome was recorded and unobserved due to death. Rubin [28] and Hayden *et al.* [29] referred to the estimand in Frangakis and Rubin [26] as 'the survivors average causal effect (SACE)'. Egleston *et al.* [27] proposed assumptions to identify the SACE and implemented a sensitivity analysis for some of those assumptions. Rubin [30] introduced the causal effect of a treatment on a outcome that is censored by death in QOL studies. In this paper, we describe an approach that can be used to obtain the causal effect of treatment in the presence of death based on principal stratification for longitudinal ordinal outcomes.

DOI: 10.1002/sim

This paper is arranged as follows. In Section 2, we describe the motivating example. We briefly review marginalized random effects models (MREMs) for longitudinal binary data [9] in Section 3. In Section 4, we propose an ordinal MREM (OMREM) for the longitudinal data with ignorable dropout. In Section 5, we conduct a simulation study to examine bias and efficiency in estimation of marginal mean parameters. In the context of QOL data collected in a recent colorectal cancer clinical trial [31], we propose models for the OMREM under dropout due to progression/death and illustrate them on this data in Section 6.

2. MOTIVATING EXAMPLE

We analyzed QOL data from a recent colorectal cancer clinical trial [31]. A total of 795 patients with colorectal cancer were randomly assigned to one of the three treatments (FOLFOX, IFL (control) and IROX) between May 1999 and April 2001. The main objective of this trial was to find a better treatment for colorectal cancer. The median survival for patients receiving IFL was 15.0 months compared with 19.5 months for those receiving FOLFOX and 17.4 months for those receiving IROX. Survival for patients receiving FOLFOX did not differ from those receiving IROX (see [31]).

However, given that the toxicity profiles were quite different on the three treatment arms, it was of interest to see if there was a negative impact of 'better' treatments on patients QOL. We focus on one QOL measure, fatigue. Fatigue is measured on a 5-point ordinal scale. Additional complications for analyzing QOL are posed by patients dying during the trial.

The models and analysis to follow focus on two treatments, FOLFOX and IFL (control) and address appropriate analysis of QOL data in the presence of death.

3. REVIEW OF MARGINALLY SPECIFIED LOGISTIC-NORMAL MODELS FOR LONGITUDINAL BINARY DATA

Now we review MREMs for longitudinal binary data [9]. Define $\mu_{it}^M = P(Y_{it}=1|x_{it})$. The marginalized logistic-normal model (MLNM) is specified using the following regressions:

mean model:

$$\operatorname{logit}\mu_{it}^M = x_{it}^{\mathrm{T}}\beta \tag{1}$$

dependence model:

$$\operatorname{logit}\mu_{it}^c(b_{it}) = \Delta_{it} + b_{it} \tag{2}$$

where β is the $p\times 1$ vector of regression coefficients and $\mu_{it}^c(b_{it}) = E(Y_{it}|b_{it}, x_{it})$. We assume that the response vector Y_i is conditionally independent given $b_i = (b_{i1},\ldots,b_{in_i})^{\mathrm{T}}$ and that

$$b_i \sim \mathrm{N}(0, A) \tag{3}$$

The covariance matrix A is assumed to have a simple structure. Conditional on x_i and b_i, the responses Y_{it} are assumed to be conditionally independent. Parameters in A provide measures of random variation both across individuals and over time.

DOI: 10.1002/sim

The parameters Δ_{it} in (2) are functions of both the marginal mean parameters and the random effects variance and can be obtained using the following identity:

$$P(Y_{it}=1|x_{it})=\int P(Y_{it}=1|b_{it},x_{it})f(b_{it})\,\mathrm{d}b_{it} \tag{4}$$

where f is the univariate normal density function. This model has several desirable features. First, the mean model is specified separately from the dependence model. As a result, the interpretation of the regression parameter β does not change as we modify assumptions regarding the dependence in equation (2). This is not true for classical generalized random effects models, which parameterize μ_{it}^c directly as a function of covariates. In addition, parameters in $\mathrm{cov}(b_i)$ provides measures of random variation both across individuals and over time. Second, the MLNM can be used with data where subjects have variable lengths of follow-up, permitting the likelihood analysis in settings where data may be MAR. Further details on MLNM are given in [9].

4. MREMS FOR LONGITUDINAL ORDINAL DATA

In this section, we extend Heagerty's MLNM to accommodate longitudinal ordinal data.

4.1. Proposed models

Let $Y_i=(Y_{i1},\ldots,Y_{in_i})$ be a vector of longitudinal K-category ordinal responses on subject $i=1,\ldots,N$ at times $t=1,\ldots,n_i$, $n_i\leqslant T$. We assume that associated exogenous but possibly time-varying covariates, $x_{it}=(x_{it1},\ldots,x_{itp})$, are recorded for each subject at each time, and that the regression model properly specifies the full covariate conditional probability such that $P(Y_{it}=y_{it}|X_{it})=P(Y_{it}=y_{it}|X_{i1},\ldots,X_{in_i})$. The MREM for longitudinal ordinal data, also called OMREM, is specified using the following two regressions:

mean model:

$$\log\frac{P(Y_{it}\leqslant k|x_{it})}{1-P(Y_{it}\leqslant k|x_{it})}=\beta_{0k}+x_{it}^{\mathrm{T}}\beta \tag{5}$$

dependence model:

$$\log\frac{P(Y_{it}\leqslant k|b_i,x_{it})}{1-P(Y_{it}\leqslant k|b_i,x_{it})}=\Delta_{itk}+b_{it} \tag{6}$$

where $b_i^{\mathrm{T}}=(b_{i1},\ldots,b_{in_i})\sim\mathrm{N}(0,\Sigma_i)$ for $i=1,\ldots,N$ and $k=1,\ldots,K-1$. We assume that the variance–covariance matrix Σ_i of b_i has an autoregressive covariance structure

$$\Sigma_i=\sigma_i^2\begin{pmatrix}1 & \mathrm{e}^{-\alpha} & \mathrm{e}^{-2\alpha} & \cdots & \mathrm{e}^{-(n_i-1)\alpha}\\ \mathrm{e}^{-\alpha} & 1 & \mathrm{e}^{-\alpha} & \cdots & \mathrm{e}^{-(n_i-2)\alpha}\\ \mathrm{e}^{-2\alpha} & \mathrm{e}^{-\alpha} & 1 & \cdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathrm{e}^{-(n_i-1)\alpha} & \mathrm{e}^{-(n_i-2)\alpha} & \mathrm{e}^{-(n_i-3)\alpha} & \cdots & 1\end{pmatrix}\overset{\text{let}}{=}\sigma_i^2\Sigma^* \tag{7}$$

DOI: 10.1002/sim

where $\log \sigma_i = z_i^{\mathrm{T}} \lambda$, z_i is a $c \times 1$ vector, and λ is a $c \times 1$ coefficient vector of z_i. The regression model for σ_i allows heterogeneity to depend on subject-level covariates such as treatment or gender.

In the dependence model, (6), we need to ensure the monotonicity of Δ in k to guarantee the validity of the mean model in (5).

Theorem 1
For the OMREM given by (5) and (6), if $\beta_{01} < \cdots < \beta_{0K-1}$, then $\Delta_{it1} < \cdots < \Delta_{itK-1}$.

Proof 1
See Appendix. □

The marginal mean model is a proportional odds model [32]. Serial dependence is captured by the random effects. The regression parameters in (5) have a marginal interpretations unlike generalized linear mixed models [5]. An advantage of this model is the ability to use conditional models for association (6) while still structuring the marginal mean as a function of covariates directly (5). As a result, the interpretation of the regression coefficients, (β_0, β) does not depend on the specification of the dependence model.

For longitudinal data with random effects b_i, the marginal probability captures the systematic variation in the marginal probability that is due to x_{it}, whereas parameters in $\mathrm{cov}(b_i)$ provide a measure of random variation both across individuals and over time. Heagerty and Kurland [33] investigated the impact on the estimates of regression coefficients of incorrect assumptions regarding the random effects in generalized linear mixed models and marginalized models and found that marginalized regression models are much less susceptible to bias resulting from random effects model misspecification.

The marginal and conditional probabilities in (5) and (6) are related as follows:

$$P_{itk}^{M} = \int P_{itk}^{c}(b_{it}) f(b_{it}) \, \mathrm{d}b_{it} \tag{8}$$

where $P_{itk}^{M} = P(Y_{it} \leqslant k | x_{it})$, $P_{itk}^{c}(b_{it}) = P(Y_{it} \leqslant k | b_{it}, x_{it})$ and $f(\cdot)$ is a univariate normal distribution with mean 0 and variance $\mathrm{var}(b_{it})$. We use (8) to solve for Δ_{itk} given β_{0k}, β and λ.

Reparametrization of the random effects and their covariance matrix: From a computational perspective, it is convenient to orthogonalize the random effects by setting $b_i = \sigma_i \Sigma^{*1/2} a_i$ where $\Sigma^{*1/2}$, a lower triangular matrix with positive diagonal elements, is the Cholesky factor of the $n_i \times n_i$ matrix Σ^{*} [34] and a_i is a $n_i \times 1$ vector of independent standard normals. The reparameterized conditional model is then given by

$$\log \frac{P(Y_{it} \leqslant k | a_i, x_{it})}{1 - P(Y_{it} \leqslant k | a_i, x_{it})} = \Delta_{itk} + \sigma_i s^{(t)} a_i$$

$$a_i \sim \mathrm{N}(0, I)$$

where $s^{(t)}$ is the tth row of $\Sigma^{*1/2}$ and I is the identity matrix of order n_i. This transformation allows us to estimate the Cholesky factor $\Sigma^{*1/2}$ instead of the covariance matrix Σ^{*}. Since the Cholesky factor is the square root of the covariance matrix, it allows more stable estimation of near-zero variance terms [35].

DOI: 10.1002/sim

4.2. Maximum likelihood estimation

The likelihood function is the integral over random effects of a product of multinominals,

$$L(\theta; y) = \prod_{i=1}^{N} \int \prod_{t=1}^{n_i} \prod_{k=1}^{K} (P_{itk}^c(a_i) - P_{itk-1}^c(a_i))^{y_{itk}} \phi(a_i)\, \mathrm{d}a_i \tag{9}$$

where $P_{itk}^c(a_i) = P(Y_{it} \leqslant k | a_i, x_{it})$, $P_{it0}^c(a_i) = 0$, y_{itk} is the set of K indicators with $y_{itk} = 1$ if $y_{it} = k$; $y_{itk} = 0$ otherwise, for $k = 1, \ldots, K$, $\phi(\cdot)$ is a multivariate normal density with mean vector 0 and variance–covariance matrix I, and $\theta = (\beta_0, \beta, \lambda, \alpha)$. The marginalized likelihood in (9) is not available in the closed form. There are several approaches to (numerically) integrate out the random effects. Gauss–Hermite quadrature is popular for low-dimensional random effects models such as random intercept models [35]; adaptive versions [36, 37] can increase the efficiency. Monte Carlo methods are often used for models with higher-dimensional integrals and use randomly sampled points to approximate the integrals. In our model, we use Monte Carlo methods to evaluate the integral in (9) as the dimension of a_i is high.

Maximizing the log likelihood with respect to θ yields the likelihood equation

$$\sum_{i=1}^{N} \frac{\partial \log L(\theta; y_i)}{\partial \theta} = \sum_{i=1}^{N} L^{-1}(\theta; y_i) \int \frac{\partial L(\theta, a_i; y_i)}{\partial \theta} \phi(a_i)\, \mathrm{d}a_i = 0$$

where

$$\begin{aligned} L(\theta; y_i) &= \int \prod_{t=1}^{n_i} \prod_{k=1}^{K} (P_{itk}^c(a_i) - P_{itk-1}^c(a_i))^{y_{itk}} \phi(a_i)\, \mathrm{d}a_i \\ L(\theta, a_i; y_i) &= \prod_{t=1}^{n_i} \prod_{k=1}^{K} \{P_{itk}^c(a_i) - P_{itk-1}^c(a_i)\}^{y_{itk}} \end{aligned} \tag{10}$$

The $(K-1+p+c+1)$-dimensional likelihood equations are given in the Appendix.

The matrix of second derivatives of the observed data log likelihood has a very complex form. Fortunately, the sample empirical covariance matrix of the individual scores in any correctly specified model is a consistent estimator of the information and involves only the first derivatives. Therefore, the quasi-Newton method can be used to solve the likelihood equations, using

$$\theta^{(m+1)} = \theta^{(m)} + [I_e(\theta^{(m)}; y)]^{-1} \frac{\partial \log L}{\partial \theta^{(m)}}$$

where $I_e(\theta)$, an empirical and consistent estimator of the information matrix at step m, is given by

$$I_e(\theta; y) = \sum_{i=1}^{N} \frac{\partial L(\theta; y_i)}{\partial \theta} \frac{\partial L(\theta; y_i)}{\partial \theta^{\mathrm{T}}}$$

At convergence, the large-sample variance–covariance matrix of the parameter estimates is then obtained as the inverse of $I_e(\hat{\theta}; y)$.

For the explicit forms of the terms in the quasi-Newton algorithm including the derivatives $\partial \Delta_{itk} / \partial \beta_0$, $\partial \Delta_{itk} / \partial \beta$ and $\partial \Delta_{itg} / \partial \lambda$ calculated using (8), see the Appendix.

DOI: 10.1002/sim

The intercepts Δ_{itk} are a function of β_{0k}, β, α and λ and must be obtained within the quasi-Newton algorithm. Let $f(\Delta_{itk}) = \int P^c_{itk}(b_{it})\phi(b_{it})\,\mathrm{d}b_{it} - P^M_{itk}$. Estimates of Δ_{itk} can be obtained using the Newton–Raphson algorithm as follows:

$$\Delta^{(l+1)}_{itk} = \Delta^{(l)}_{itk} - \left(\frac{\partial f(\Delta_{itk})}{\partial \Delta_{itk}}\right)^{-1} f(\Delta_{itk})$$

where

$$\frac{\partial f(\Delta_{itk})}{\partial \Delta_{itk}} = \int P^c_{itk}(b_{it})(1 - P^c_{itk}(b_{it}))h(b_{it})\,\mathrm{d}b_{it} \tag{11}$$

Note that the integral in (11) is one dimensional and we use Gauss–Hermite quadrature to evaluate this integral.

5. SIMULATION STUDY

We conducted a simulation to examine the bias and efficiency for estimation of the marginal mean parameters in the setting of misspecification of the dependence model under no missing data and under MAR missingness (common in longitudinal data). We also compare the efficiency of the OMREM to the independence proportional odds model (IPOM), given in (5).

We simulated the longitudinal ordinal data under an OMREM. Covariates were time and group (two levels). The marginal probability for the OMREM was specified as

$$\log\left(\frac{P(Y_{it} \leqslant k|x_{it})}{1 - P(Y_{it} \leqslant k|x_{it})}\right) = \beta_{0k} + \beta_1 \cdot \mathrm{time}_{it} + \beta_2 \cdot \mathrm{group}_i$$

$$\beta_0 = (\beta_{01}, \beta_{02}, \beta_{03}) = (-1.0, 0.5, 1.0); \quad \beta = (\beta_1, \beta_2) = (-0.5, 0.5)$$

where $t = 1, \ldots, 6$, $\mathrm{time}_{it} = (t-1)/6$, and group_i equals 0 or 1 with an approximately equal sample size per group. The conditional probabilities were specified from (6) and (7) with $(\sigma_i, \alpha) = (1.1, 0.2)$ if $\mathrm{group}_i = 0$; $(\sigma_i, \alpha) = (1.5, 0.2)$ if $\mathrm{group}_i = 1$. Note that $\alpha = 0.2$ corresponds to a lag one correlation of $\exp(-0.2) = 0.819$. We simulated 500 data sets each with a sample size of 300. We then fit the IPOM and the OMREM.

For the MAR missingness, we specified the following MAR dropout model:

$$\mathrm{logit}\, P(\mathrm{dropout} = t|\mathrm{dropout} \geqslant t) = \begin{cases} -1.5 + 0.3Y_{it-1} & \mathrm{grp}_i = 0 \\ -1.5 + 0.1Y_{it-1} & \mathrm{grp}_i = 1 \end{cases} \tag{12}$$

where $Y_{it-1} \in \{0, 1, 2, 3\}$.

Table I presents the point estimates, root mean-squared error ($\sqrt{\mathrm{MSE}}$), and 95 per cent Monte Carlo error intervals of the marginal mean parameters. When there were no missing data, the estimates were essentially unbiased for both the OMREM and the IPOM. The root MSEs in the IPOM were also similar to those in the OMREM but sometimes slightly larger (e.g. for β_1 and β_2). In the presence of MAR missingness, the estimates in the OMREM were still essentially unbiased. However, for the IPOM, we saw considerable biases; for example, the relative bias for the coefficient of time, β_1 was 34 per cent $((-0.33 + 0.5)/(-0.5))$. In addition, the root MSEs were larger in

DOI: 10.1002/sim

Table I. Bias of maximum likelihood estimators.

		Complete data				MAR dropout			
		OMREM		IPOM		OMREM		IPOM	
Parameter	Truth	Mean	$\sqrt{\text{MSE}}$	Mean	$\sqrt{\text{MSE}}$	Mean	$\sqrt{\text{MSE}}$	Mean	$\sqrt{\text{MSE}}$
β_{01}	-1.00	-0.99	0.01	-0.99	0.01	-1.01	0.01	-0.98	0.02
		$(-1.00, -0.98)$		$(-1.00, -0.98)$		$(-1.02, -0.99)$		$(-0.99, -0.97)$	
β_{02}	0.50	0.51	0.01	0.51	0.01	0.49	0.01	0.51	0.01
		$(0.50, 0.51)$		$(0.50, 0.52)$		$(0.48, 0.50)$		$(0.50, 0.52)$	
β_{03}	1.00	1.00	0.00	1.00	0.00	0.99	0.01	1.01	0.01
		$(0.99, 1.01)$		$(1.00, 1.01)$		$(0.98, 1.00)$		$(1.00, 1.02)$	
β_1 (Time)	-0.50	-0.51	0.01	-0.52	0.02	-0.48	0.02	-0.33	0.17
		$(-0.53, -0.49)$		$(-0.54, -0.50)$		$(-0.52, -0.45)$		$(-0.36, -0.30)$	
β_2	0.50	0.50	0.00	0.51	0.01	0.51	0.01	0.47	0.03
		$(0.49, 0.51)$		$(0.49, 0.52)$		$(0.49, 0.52)$		$(0.46, 0.48)$	

Displayed are the average regression coefficient estimates, the $\sqrt{\text{MSE}}$ and 95 per cent Monte Carlo error interval $(\bar{\beta} \pm 1.96\sqrt{\text{var}(\bar{\beta})/500})$. The true conditional probabilities were specified with $\sigma_i = 1.1$ and $\alpha = 0.2$ if $\text{group}_i = 0$; $\sigma_i = 1.5$ and $\alpha = 0.2$ if $\text{group}_i = 1$. The fitted conditional probabilities were specified with random effects, $b_i \sim \text{N}(0, \sigma_i^2 \Sigma^*)$ with $\log \sigma_i = \lambda_1 + \text{group}_i \lambda_2$ for the OMREM.

the IPOM; for β_1 and β_2, the root MSEs were eight and three times as large as the OMREM, respectively.

Overall, the simulation shows the increased efficiency of the OMREM over the independence model (IPOM) in complete data and the large biases that can occur in the marginal mean parameters when the dependence is mis-modeled in the presence of MAR missingness.

6. QOL EXAMPLE WITH DROPOUT DUE TO PROGRESSION/DEATH

We focus on one QOL measure, fatigue, measured on a 5-point ordinal scale (1, I am usually not tired at all; 2, I am occasionally rather tired; 3, there are frequently periods when I am quite tired; 4, I am usually very tired; 5, I feel exhausted most of the time). Because very few patients reported category 5, we combined categories 4 and 5 into one category. We used 707 subjects without missing data at baseline and focused on two treatments, FOLFOX and IFL (control).

To examine treatment differences in fatigue levels, we included the type of treatment, Tx,

$$\text{Tx}_i = \begin{cases} 0 & \text{if subject } i \text{ is assigned to IFL (control)} \\ 1 & \text{if subject } i \text{ is assigned to FOLFOX (active)} \end{cases}$$

and visit number $(\text{TIME} = 0.0, 0.1, \ldots, 0.5)$ re-scaled. The patient's visit corresponds to the time period during which the survey was filled out (0, baseline; 1, 1–84 days after going on study; 2, 85–168 days after going on study;...; 5, 337–420 days after going on study). We will analyze the data from the first five windows (up to about 1 year). We assumed that the missing responses (mostly due to dropout) were MAR in our initial analysis. In Section 6.2, we more carefully handled the missingness related to the reason for dropping out (including death).

DOI: 10.1002/sim

The quasi-Newton algorithm is not trivial computationally due to the need to obtain estimates and derivatives of Δ_{it} using the Gauss–Hermite quadrature for all subjects and at all times, within each quasi-Newton step. Each quasi-Newton step (in which all the Δ_{it} need to be computed) on a Pentium with a 1.6 GHz processor took about 3 min for the OMREM with MC sample size of 10 000 and 40 point Gauss–Hermite quadrature. Using good initial values based on fitting an IPOM in standard software results in a minimal number of iterations until convergence. For example, in our analysis below, we obtained convergence in 20 iterations using a fairly strict convergence criterion, $|\hat{\theta}^{\text{old}} - \hat{\theta}^{\text{new}}| \leqslant 10^{-4}$ where $\hat{\theta}^{\text{new}}$ and $\hat{\theta}^{\text{old}}$ are current and previous estimates of the parameters, respectively.

6.1. Model fit

We first fitted three OMREMs and one IPOM under an assumption of ignorable dropout. OMREM-1 allowed the random effects variance to depend on treatment, $\log \sigma_i = \lambda_0 + \lambda_1 \times \text{Tx}_i$. OMREM-2 was a simpler model, with a constant variance, $\log \sigma_i = \lambda_0$. Both had autoregressive variance–covariance structures. OMREM-3 was the OMREM-2 with $b_{it} = b_{i0} \sim \text{N}(0, \sigma^2)$. Table II gives the maximum likelihood estimates (MLEs) for all four models.

The inferences for some of the coefficients under the IPOM were very different from those for the dependence models. For example, the interaction of treatment and visit was highly significant under the IPOM, indicating an *increase* in fatigue over time for FOLFOX relative to IFL, whereas the dependence models indicated a non-significant *decrease* in fatigue over time.

Comparison of AIC for IPOM and OMREM-3 indicated that the OMREM-3 fit much better than the IPOM (2561.154 for OMREM-3, 2745.116 for IPOM). Point estimates and standard

Table II. Maximum likelihood estimates for marginalized random effects models under ignorable missingness.

	OMREM-1	OMREM-2	OMREM-3	IPOM
Mean parameters				
Int1	−1.099* (0.126)	−1.093* (0.126)	−1.119* (0.124)	−1.163 (0.636)
Int2	0.839* (0.126)	0.839* (0.127)	0.816* (0.123)	0.796* (0.149)
Int3	2.280* (0.160)	2.277* (0.160)	2.261* (0.156)	2.277* (0.982)
Visit	−0.334 (0.620)	−0.367 (0.609)	−0.166 (0.472)	−0.424* (0.113)
Tx	−0.073 (0.168)	−0.072 (0.167)	−0.069 (0.161)	−0.003 (0.039)
Visit × Tx	−0.607 (0.922)	−0.469 (0.917)	−0.541 (0.725)	1.059* (0.063)
Dependence parameters				
Int	1.086* (0.634)	1.251 (0.816)	0.664* (0.144)	
Tx	0.280 (0.897)			
α	0.243* (0.076)	0.264* (0.081)		
Max. $\log L$	−1266.112	−1267.623	−1273.577	−1366.558
AIC	2550.224	2551.246	2561.154	2745.116

OMREM-1 and OMREM-2 are OMREMs with random effects variance $\log \sigma_i = \lambda_0 + \lambda_1 \times \text{Tx}_i$ and $\log \sigma_i = \lambda_0$, respectively. Both had autoregressive covariance structures. OMREM-3 is the OMREM-2 with $b_{it} = b_{i0} \sim \text{N}(0, \sigma^2)$. IPOM is an independent proportional odds model. Tx is an indicator for FOLFOX. Visit is the patient's visit corresponding to the time period.
*Significance under 95 per cent confidence level.

DOI: 10.1002/sim

errors for marginal mean parameters for the OMREM-1 and OMREM-2 were similar. To compare the fit of the two models under ignorability, we computed the likelihood ratio test. Comparison of deviances for OMREM-2 and OMREM-1, which were nested yielded $\Delta D_{12}=2\times(1267.623-1266.112)=3.022$, p-value$=0.082$ on 1 d.f. This comparison indicated that the OMREM-1 did not provide a significantly better fit than the OMREM-2. We also computed the likelihood ratio test to compare OMREM-2 and OMREM-3. The deviance difference was $\Delta D_{23}=2\times(1273.577-1267.623)=11.908$ p-value<0.01 on 1 d.f. This indicated that the OMREM-2 fit better than the OMREM-3.

The estimate of the correlation parameter ρ in OMREM-2 was significant and corresponded to an estimated correlation of $\hat{\rho}=\exp(-\hat{\alpha})=\exp(-0.264)=0.768$. The MLE for σ ($\exp(1.251)=3.49$) indicated large subject-to-subject variation in the odds of the cumulative probability of fatigue. The coefficients of treatment and the interaction of treatment and visit in the marginal mean model were not significant indicating that fatigue (and its trajectory over time) did not differ between the two treatments.

6.2. *Dropout due to progression/death*

QOL responses not measured due to participant death do not exist whereas scheduled measurements due to dropout for other reasons can be viewed as existing but unobserved. If we fit models using methods in Section 4 (as we did in Section 6.1), the missing data due to death are implicitly imputed (under MAR). We outline a PMM approach to address this next.

6.2.1. PMM approach. Define S_i to be death time for subject i. To address dropout due to death, we can specify the OMREMs conditional on death time, S. The models are given by

$$\begin{aligned}
\log\frac{P(Y_{it}\leqslant k|x_{it},S_i=j)}{1-P(Y_{it}\leqslant k|x_{it},S_i=j)}&=\beta_{0k}(j)+x_{it}^{\mathrm{T}}\beta(j)\\
\log\frac{P(Y_{it}\leqslant k|b_i,x_{it},S_i=j)}{1-P(Y_{it}\leqslant k|b_i,x_{it},S_i=j)}&=\Delta_{itk}(j)+b_{it}(j)\\
b_i(j)^{\mathrm{T}}=(b_{i1}(j),\ldots,b_{iT}(j))&\sim\mathrm{N}(0,\Sigma_i(j))
\end{aligned}\tag{13}$$

where x_{it} is a vector of covariates including treatments, $j=1,\ldots,J$, and J is the number of patterns ($J\leqslant T$). This approach implicitly assumes that for a given death time (pattern) that missing data before the death time is MAR (conditional on pattern).

Now, suppress i without loss of clarity. One target of inference is $P(Y_t>k|S>j,x)$ [25], which is recovered from the PMM by summing over the survival distribution [38],

$$P(Y_t>k|S>j,x)=\frac{\sum_{g>j}P(Y_t>k|S=g,x)P(S=g|x)}{\sum_{l>j}P(S=l|x)}\tag{14}$$

where $P(Y_t>k|S=g,x)$ is estimated from (13) and is only defined for $t<g$. Unfortunately, this approach only uses the 'survivors' under each treatment and the corresponding inference is not the causal effect of the treatment.

6.2.2. PMM approach applied. A large number of subjects dropped out of this study due to tumor progression or death (41 per cent). Based on consultation with Mayo investigators for

DOI: 10.1002/sim

Table III. Break down of completers and dropouts by treatment groups for QOL data.

		Dropouts by reason		
Treatment	Completer	Progression/death	Other	Total
IFL	10 (0.04)	117 (0.50)	108 (0.46)	235
FOLFOX	4 (0.02)	68 (0.29)	161 (0.69)	233
Total	14 (0.03)	185 (0.41)	269 (0.57)	468

Proportions are given in parentheses.

Table IV. Break down of progression/death windows by treatment groups for QOL data.

	Progression/death window						
Treatment	1	2	3	4	5	6	Total
IFL	35 (0.15)	32 (0.14)	38 (0.16)	26 (0.11)	18 (0.08)	86 (0.37)	235
FOLFOX	16 (0.07)	28 (0.12)	24 (0.10)	17 (0.07)	19 (0.08)	129 (0.55)	233
Total	51 (0.11)	60 (0.13)	62 (0.13)	43 (0.09)	37 (0.08)	215 (0.46)	468

Window 6 means 'after study.' Proportions are given in parentheses.

assessing QOL, we grouped tumor progression and death together. When the dropout rates due to progression/death are broken down by treatment arms, the rates were marginally higher in the IFL arm. For subjects randomized to the IFL arm, 50 per cent dropped out due to progression/death, whereas 29 per cent dropped out due to progression/death for subjects with the FOLFOX arm (see Table III).

The number of subjects by progression/death windows by treatment groups is given in Table IV. For a subject who dropped out for reason unrelated to death/progression, we still know when the subject died or progressed if it happened before the study ended. For example, if a subject dropped out for reason unrelated to death/progression after the second visit and was alive until the study was finished, the subject belonged to progression/Death window 6.

We specify the OMREMs conditional on progression/death time, S as outlined in Section 6.2.1. Due to the small sample sizes when conditioning on individual times, we assumed that the parameters were the same for $S=1,\ldots,5$ (those who progressed/died before the end of the study) but different for those who did not, $S=6$.

In our analysis here, we compared three models, an OMREM (OMREM-2), an independent proportional odds model (IPOM) under ignorable missingness and a mixture model (PMM) as outlined in Section 6.2.1. In the PMM, the target probabilities given in (14) evaluated at $j=5$ correspond to those who did not progress/die before the end of the study. Table V presents the estimated target probabilities on the two treatment arms. Figure 1 indicates MLEs of $P(Y_t>k|S>5,\mathrm{Tx})$ for the PMM and those of $P(Y_t>k|\mathrm{Tx})$ for the MREM (OMREM) and the independent proportional odds model (IPOM) under ignorable missingness, respectively. In the PMM and OMREM, $P(Y_t>k|S>5,\mathrm{Tx})$ and $P(Y_t>k|\mathrm{Tx})$, which, respectively, evaluate the probability of fatigue for those who did not progress/die before the end of study and that for all patients were calculated. In the PMM, the target probabilities for the FOLFOX arm increased over time, whereas that for the IFL arm did not

DOI: 10.1002/sim

Table V. Maximum likelihood estimates of $P(Y_t>k|S>5,\text{Tx})$ (standard errors) where S is progression/death time.

		Visit(t)					
Trt	k	1	2	3	4	5	6
IFL	1	0.753	0.754	0.754	0.755	0.755	0.756
		(0.039)	(0.034)	(0.037)	(0.047)	(0.061)	(0.077)
	2	0.287	0.288	0.288	0.288	0.289	0.290
		(0.044)	(0.037)	(0.040)	(0.051)	(0.066)	(0.084)
	3	0.086	0.086	0.087	0.087	0.087	0.087
		(0.019)	(0.017)	(0.019)	(0.023)	(0.028)	(0.035)
FOLFOX	1	0.731	0.757	0.781	0.803	0.823	0.842
		(0.033)	(0.028)	(0.032)	(0.039)	(0.047)	(0.054)
	2	0.264	0.291	0.320	0.350	0.381	0.413
		(0.034)	(0.031)	(0.038)	(0.054)	(0.073)	(0.095)
	3	0.078	0.088	0.099	0.112	0.126	0.141
		(0.016)	(0.016)	(0.019)	(0.026)	(0.036)	(0.049)

$P(Y_t>k|S>5,\text{Tx})$ evaluates probability of fatigue stress for those who did not progress/die before the end of the study.

change. However, there were no significant differences over time between the treatment arms. In the OMREM, we have a pattern similar to that in the PMM. In the IPOM, the target probabilities for the FOLFOX were higher than those for the IFL unlike the other models as time increases because the estimate of coefficient of interaction between visit and arm was positive and large compared with those in the other models (1.059 for the IPOM and 0.469 for the OMREM). Because the PMM handled the missingness due to progression/death properly, we focus on the PMM and conclude that patients' fatigue was not affected by treatment like in the inappropriate ignorable models. However, because we restrict the analyses for the QOL data to subgroup of patients who survived, the resulting treatment comparisons will no longer have a causal interpretation. In the following section, we present a principal stratification approach to address this.

6.2.3. Principal stratification. Before defining the causal effect of treatment, we introduce potential outcomes. Potential outcomes are all the outcomes that would be observed if both treatments had been applied to each of the subjects [39–41]. Frangakis and Rubin [26] used the concept of potential outcomes in an approach called 'principal stratification'. Principal stratification partitions subjects into sets with respect to post-treatment variables. The principal strata are not affected by treatment assignment and therefore can be used as any pre-treatment covariate. Causal effects are defined within these principal strata. In the following, the post-treatment variable of interest is progression/death.

Let Tx be the treatment indicator as defined earlier. Let $Y_t(\text{Tx})$, the potential outcome at time t, be an ordinal QOL response. We only observe either $Y_t(1)$ or $Y_t(0)$ for each subject. Now, let $D_t(\text{Tx})$ be tumor progression/death indicator at visit t on treatment Tx,

$$D_t(\text{Tx})=\begin{cases}0 & \text{alive}\\ 1 & \text{tumor progressed or dead}\end{cases}$$

DOI: 10.1002/sim

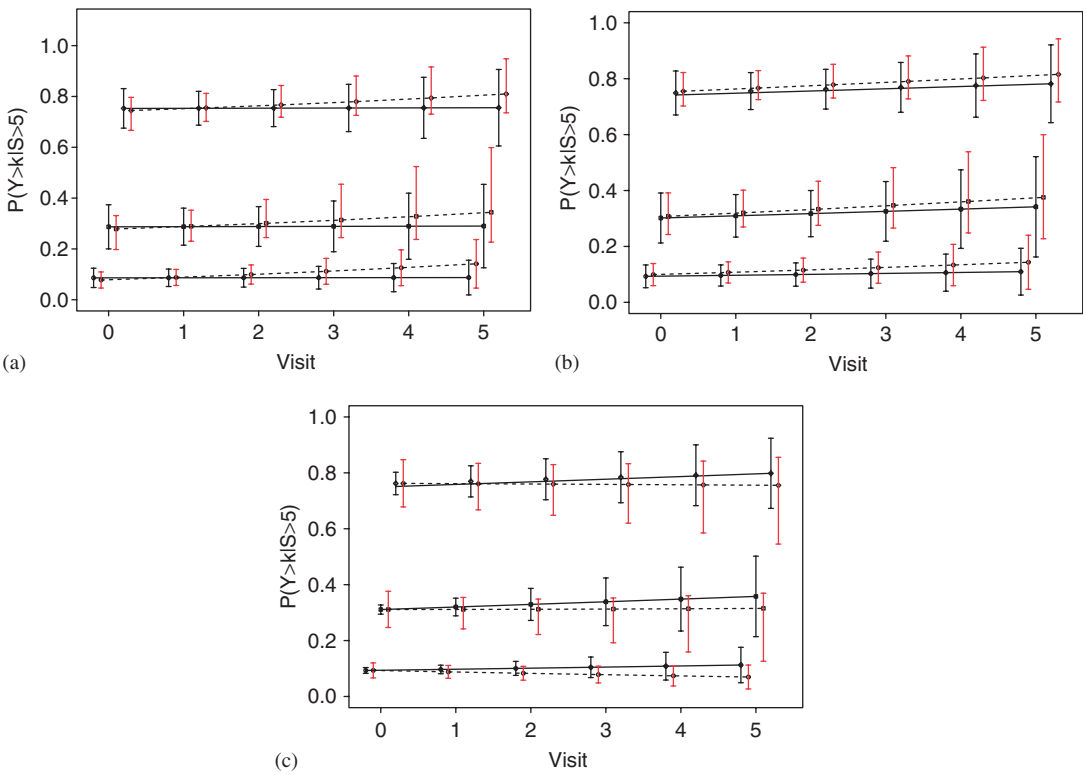

Figure 1. Maximum likelihood estimates of $P(Y_t>k|S>5, \mathrm{Tx})$ for the pattern mixture model (PMM) and those of $P(Y_t>k|\mathrm{Tx})$ for the marginalized random effects model (OMREM) and the independent proportional odds model (IPOM) under ignorable missingness, respectively, and 95 per cent confidence intervals, where S is progression/death time. $P(Y_t>k|S>5, \mathrm{Tx})$ evaluates the probability of fatigue for those who did not progress/die before the end of study. Solid line is for IFL; dashed line is for FOLFOX. ∘, □, and ◇ are for $k=3, 2, 1$, respectively. (a) PMM for $k=1, 2, 3$; (b) OMREM for $k=1, 2, 3$; and (c) IPOM for $k=1, 2, 3$.

Obviously, if $D_t(\mathrm{Tx})=1$, then $D_s(\mathrm{Tx})=1$ for $s>t$. In this case, there are four principal strata defined by the pairs of potential values of $D_t(\mathrm{Tx})$:

1. $A_t(0)=\{i|(D_t(0), D_t(1))=(0, 0)\}$: the subjects who would be alive under both arms at visit t;
2. $A_t(1)=\{i|(D_t(0), D_t(1))=(1, 0)\}$: the subjects who would be alive under the treatment arm but not alive under the control arm at visit t;
3. $A_t(2)=\{i|(D_t(0), D_t(1))=(0, 1)\}$: the subjects who would be alive under control arm but not alive under the treatment arm at visit t;
4. $A_t(3)=\{i|(D_t(0), D_t(1))=(1, 1)\}$: the subjects who would not be alive under either treatment at visit t.

As we are interested in subjects who were alive until the end of the study, $A_t(3)$ is not of direct interest. The full set of potential outcomes at visit t is

$$\mathscr{P}_t=\{D_t(0), (Y_t(0); D_t(0)=0), D_t(1), (Y_t(1); D_t(1)=0)\}$$

DOI: 10.1002/sim

where for treatment tx, $Y_t(\text{tx})$ is the potential response if a subject is alive at visit t ($D_t(\text{tx})=0$). If the patient is alive ($D_{it}(\text{Tx}_i)=0$), define R_{it} to be the indicator that $Y_{it}=Y_{it}(\text{Tx}_i)$ is observed. In the following, we assume monotone dropout ($R_{it}=1 \Rightarrow R_{it-1}=1$). Hence, the observed data for individual i is

$$\mathcal{O}_{it}=\{\text{Tx}_i, D_{it}, (R_{it}; D_{it}=0), (Y_{it}; D_{it}=0, R_{it}=1)\}$$

We formally define the causal effects of interest and then state some additional assumptions that are necessary to estimate them.

Suppress i here for clarity and let T be the last visit in the study. We define the causal effect of interest [27] as

$$\text{SACE}_k(1,0)=\frac{\text{odds}\{Y_T(1)>k|(D_T(0),D_T(1))=(0,0)\}}{\text{odds}\{Y_T(0)>k|(D_T(0),D_T(1))=(0,0)\}}$$

$$=\frac{\text{odds}\{Y_T(1)>k|A_T(0)\}}{\text{odds}\{Y_T(0)>k|A_T(0)\}}$$

$\text{SACE}_k(1,0)$ is the odds ratios based on the probability that response at time T is larger than k for subjects who would be alive under both treatments between the treatment arm (FOLFOX) and the control arm (IFL).

Now, define $\overline{\mathscr{P}}_t=(\mathscr{P}_1,\ldots,\mathscr{P}_t)$ to be the history of potential outcomes up to and including the outcomes at visit t. We first make the Stable Unit Treatment Value Assumption (SUTVA, [40]) which states that a patient's potential outcomes are unrelated to the treatment status of other patients. Next, we list the additional assumptions necessary to identify the causal effect of interest:

Assumption 1 (*Monotonicity*)
If $D_t(1)=1$, then $D_t(0)=1$; if $D_t(0)=0$, then $D_t(1)=0$.

Assumption 1 assumes that the active treatment is effective. As such, if a patient is alive under the control arm, then the patient will also be alive under the active arm. Note that $A_t(2)$ is empty from Assumption 1.

Assumption 2 (*Ignorability*)
$\text{Tx} \perp \overline{\mathscr{P}}_T$.

Note that $\overline{\mathscr{P}}_t$ is all the potential outcome up to and including time t. Assumption 2 states that the treatment arm is unrelated to the set of potential outcomes.

Assumption 3
$R_t \perp Y_t | \overline{\mathscr{P}}_{t-1}$.

This assumption states that missingness of the outcome is independent of the value of the outcome given all the potential outcome up to and including time $t-1$. It is similar to an assumption of sequential MAR [42].

Assumption 4 (*Proportional odds*)

$$\text{logit}P(Y_T(1)>k|a_1)=\alpha_{0k}^{(1)}+\alpha_1^{(1)}a_1$$

DOI: 10.1002/sim

where $\alpha_{01}^{(1)} \geqslant \alpha_{02}^{(1)} \geqslant \cdots \geqslant \alpha_{0K-1}^{(1)}$ and

$$a_1 = \begin{cases} 1 & \text{if a subject is in } A_T(1) \\ 0 & \text{otherwise} \end{cases}$$

This assumption can be re-written as

$$\frac{\text{odds}(Y_T(1)>k|A_T(1))}{\text{odds}(Y_T(1)>k|A_T(0))} = e^{\alpha_1^{(1)}} \stackrel{\text{let}}{=} \tau \tag{15}$$

We use this to reduce the number of sensitivity parameters. This will become apparent in what follows.

In general, the goal here is to use the observed data, $\mathcal{O}_t$, along with these assumptions to draw inference about SACE. To identify SACE, we need to identify $P(Y_T(0)>k|A_T(0))$ and $P(Y_T(1)>k|A_T(0))$. We provide the details of this identification using Assumptions 1–4 in the Appendix.

6.2.4. Principal stratification applied. In our study, even though there is a known survival difference between the groups, differential toxicity between the treatment arms led us to examine QOL in a manner not influenced by the survival difference.

The causal estimand of interest, SACE is a function of $g_T(\text{Tx})$ and $h_{T,\text{Tx}}(k)$ (see below and Appendix for more details). For $\text{Tx}=0, 1$,

$$\sum_{i=1}^{N(\text{tx})} 1 - D_{iT}|\text{Tx}=\text{tx} \sim \text{Bin}(N(\text{tx}), g_T(\text{tx}))$$

where $N(\text{tx})$ is the number of subjects who had the treatment $\text{Tx}=\text{tx}$. The estimate of $g_T(\text{tx})$ is given by

$$\hat{g}_T(\text{tx}) = N_T(\text{tx})/N(\text{tx}) \tag{16}$$

where $N_T(\text{tx})$ is the number of subjects who were alive after the study was complete under $\text{Tx}=\text{tx}$.

To define $h_{T,\text{Tx}}(k)$, we define S to be the progression/death time and specify the OMREM conditional on progression/death time as in (13). Within this framework, $h_{T,\text{Tx}}(k)$ is given by

$$\begin{aligned} h_{T,\text{Tx}}(k) &= P(Y_T>k|S>T-1,\text{Tx}) \\ &= \frac{\sum_{j>T-1} P(Y_T>k|S=j,\text{Tx})P(S=j|\text{Tx})}{\sum_{j>T-1} P(S=j|\text{Tx})} \end{aligned} \tag{17}$$

where $P(S=j|\text{Tx})$ is a multinomial mass function for S, and $P(Y_T>k\,|\,S=j,\text{Tx})$ is calculated from (14) given Tx and is only defined for $t<j$. To calculate standard errors for SACE, we use the delta method.

The identification of the SACE relies on untestable assumptions. We implemented a sensitivity analysis procedure to draw inference about SACE by varying the sensitivity parameter τ in (15) (see Assumption 4). This is similar to the approach in [27]. The sensitivity parameter τ is the odds ratio of the probability of fatigue under the FOLFOX arm among the group that is alive under

DOI: 10.1002/sim

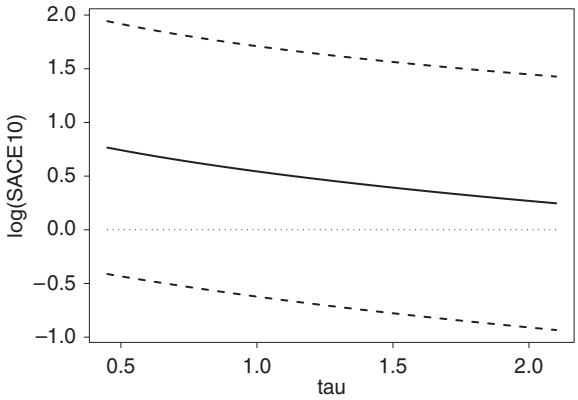

Figure 2. $\mathrm{SACE}_{K-1}(1,0)$ as a function of τ (Solid line). Dashed lines are 95 per cent confidence intervals.

the FOLFOX arm but not alive under the control treatment at visit T when compared with the group that is alive under both arms. A large value of τ means that the probability of fatigue under the FOLFOX arm among patients in stratum $A_T(1)$ is higher than that among patients in stratum $A_T(0)$. The Mayo investigator told us it was very unlikely that τ was outside the range $(0.5, 2.0)$, corresponding to a two-fold or less change in either direction. Thus, we used this as the range for our sensitivity analysis.

The causal effect of treatment may be estimated for principal stratum $A_T(0)$, patients who would survive to the end of the study period regardless of the treatment. Table IV shows the proportions of patients who were alive after the study was finished (progression/Death window $=6$). Using the data in Table IV and (16), the MLEs of $g_T(\cdot)$ were $\hat{g}_T(0)=0.366$ and $\hat{g}_T(1)=0.554$.

To estimate $\mathrm{SACE}_{K-1}(1,0)$ we also need to estimate $P(D_T(0)=0|D_T(1)=0)$, $P(Y_T(0)>K-1|A_T(0))$ and $P(Y_T(1)>K-1|A_T(0))$. All these probabilities are estimable given Assumptions 1–4. The estimate of $P(D_T(0)=0|D_T(1)=0)$, the probability of surviving under the control arm given survival under the FOLFOX arm, was 0.661. The estimate of $P(Y_T(0)>K-1|A_T(0))$ was $\hat{h}_{T,0}(K-1)=0.087$ and $P(Y_T(1)>K-1|A_T(0))$ was $\hat{h}_{T,1}(K-1)=0.141$.

The estimated values of $\mathrm{SACE}_{K-1}(1,0)$ and associated confidence intervals over τ ranging from 0.5 to 2.0 are given in Figure 2. As τ increased, the estimated values of $\mathrm{SACE}_{K-1}(1,0)$ decreased. The estimated values of $\mathrm{SACE}_{K-1}(1,0)$ were always larger than 1. This indicated that the probability of a patient's fatigue on the FOLFOX treatment was larger than the probability of a patient's fatigue on the IFL treatment. However, there was no significant difference between the two arms (at 95 per cent level). Thus, we conclude that patients' fatigue was not affected by the treatment.

7. CONCLUSION

We have proposed MREMs for longitudinal ordinal data that directly model marginal probabilities as a function of covariates while accounting for the longitudinal correlation via random effects. To evaluate the marginalized likelihood, we used Monte Carlo integration. Parameter estimation was based on ML using a quasi-Newton algorithm.

DOI: 10.1002/sim

As discussed in Section 6, about 40 per cent of the patients dropped out due to tumor progression or death. We adjusted for this using a pattern mixture approach with patterns defined based on the observed progression/death times and then using a principal stratification approach to estimate treatment effects. We plan on extending this to a fully Bayesian approach and as such, using informative priors for the sensitivity parameters to obtain a single inference that characterizes our uncertainty about the sensitivity parameters. We also will consider an extension to three treatments (two active and one placebo) and a weakening of the monotonicity assumptions.

We can extend marginalized ordinal models to allow both serial dependence via a Markov structure and random effects [43]. These models are particularly useful in longitudinal analyses with a moderate to large number of repeated measurements per subject. We are also working on extensions to multivariate longitudinal ordinal responses which would be applicable to QOL data [44].

APPENDIX A

A.1. Proof of Theorem 1

We know that $\beta_{01}<\cdots<\beta_{0K-1}$ implies that

$$P^M_{itk-1}<P^M_{itk}$$

where

$$P^M_{itk}=\int P(Y_{it}\leqslant k|b_{it},x_i)\,f(b_{it})\,\mathrm{d}b_{it}$$

Hence,

$$\int P(Y_{it}\leqslant k|b_{it},x_i)\,f(b_{it})\,\mathrm{d}b_{it}>\int P(Y_{it}\leqslant k-1|b_{it},x_i)\,f(b_{it})\,\mathrm{d}b_{it}$$

for all k. We can rewrite this as

$$\int\left(\frac{\mathrm{e}^{\Delta_{itk}+b_{it}}}{1+\mathrm{e}^{\Delta_{itk}+b_{it}}}-\frac{\mathrm{e}^{\Delta_{itk-1}+b_{it}}}{1+\mathrm{e}^{\Delta_{itk-1}+b_{it}}}\right)f(b_{it})\,\mathrm{d}b_{it}>0 \tag{A1}$$

We now claim that $\Delta_{itk}>\Delta_{itk-1}$ for all k. We will show that if it is not true (i.e, $\Delta_{itk}\leqslant\Delta_{itk-1}$ for some k), then condition (A1) cannot be satisfied.

(i) If $\Delta_{itk}=\Delta_{itk-1}$ for some k, then the left-hand side term of (A1) is zero. This is a contradiction to (A1).

(ii) If $\Delta_{itk}<\Delta_{itk-1}$ for some k, then let $\Delta_{itk-1}=\Delta_{itk}+\varepsilon$ for some $\varepsilon>0$. Therefore, (A1) is given by

$$0<\int\left(\frac{\mathrm{e}^{\Delta_{itk}+b_{it}}}{1+\mathrm{e}^{\Delta_{itk}+b_{it}}}-\frac{\mathrm{e}^{\Delta_{itk}+\varepsilon+b_{it}}}{1+\mathrm{e}^{\Delta_{itk}+\varepsilon+b_{it}}}\right)f(b_{it})\,\mathrm{d}b_{it}$$

$$=\int\frac{\mathrm{e}^{\Delta_{itk}}(\mathrm{e}^{b_{it}}-\mathrm{e}^{\varepsilon}\mathrm{e}^{b_{it}})}{(1+\mathrm{e}^{\Delta_{itk}+b_{it}})(1+\mathrm{e}^{\varepsilon}\mathrm{e}^{\Delta_{itk}+b_{it}})}f(b_{it})\,\mathrm{d}b_{it}$$

DOI: 10.1002/sim

$$\leqslant \int e^{\Delta_{itk}}(e^{b_{it}} - e^{\varepsilon} e^{b_{it}}) f(b_{it})\, db_{it}$$

$$= e^{\Delta_{itk}} \int e^{b_{it}}(1 - e^{\varepsilon}) f(b_{it})\, db_{it}$$

As $E_b\{e^{b_{it}}(1-e^{\varepsilon})\}<0$, this is a contradiction to (A1).

A.2. Detailed calculations of quasi-Newton under ignorability

The contribution of subject i to the log likelihood is given by

$$\log L(\theta; y_i) = \log \int \prod_{t=1}^{n_i} \prod_{k=1}^{K} (P^c_{itk}(a_i) - P^c_{itk-1}(a_i))^{y_{itk}} \phi(a_i)\, da_i$$

$$= \log \int \exp\left[\left\{\sum_{t=1}^{T} \sum_{k=1}^{K-1} (R_{itk}\phi_{itk} - R_{itk+1} g(\phi_{itk}))\right\}\right] \phi(a_i)\, da_i$$

where $R_{itk} = Y_{it1} + \cdots + Y_{itk}$, $\phi_{itk} = \log(P^c_{itk}/P^c_{itk+1} - P^c_{itk})$, and $g(a) = \log\{1+\exp(a)\}$.

The forms of the derivatives for quasi-Newton algorithm are

$$\frac{\partial \log L}{\partial \beta_{0j}} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta, a_i | y_i) \left\{ \sum_{t=1}^{n_i} \sum_{k=1}^{K-1} \left(R_{itk} - R_{itk+1} \frac{e^{\phi_{itk}}}{1+e^{\phi_{itk}}} \right) \right.$$

$$\left. \times \left(\frac{\partial \phi_{itk}}{\partial P^c_{itk}(a_i)} \frac{\partial P^c_{itk}(a_i)}{\partial \Delta_{itk}} \frac{\partial \Delta_{itk}}{\partial \beta_{0j}} + \frac{\partial \phi_{itk}}{\partial P^c_{itk+1}(a_i)} \frac{\partial P^c_{itk+1}(a_i)}{\partial \Delta_{itk+1}} \frac{\partial \Delta_{itk+1}}{\partial \beta_{0j}} \right) \right\} \phi(a_i)\, da_i$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta, a_i | y_i) \left\{ \sum_{t=1}^{n_i} \sum_{k=1}^{K-1} \left(R_{itk} - R_{itk+1} \frac{e^{\phi_{itk}}}{1+e^{\phi_{itk}}} \right) \right.$$

$$\left. \times \left(\frac{\partial \phi_{itk}}{\partial P^c_{itk}(a_i)} \frac{\partial P^c_{itk}(a_i)}{\partial \Delta_{itk}} \frac{\partial \Delta_{itk}}{\partial \beta} + \frac{\partial \phi_{itk}}{\partial P^c_{itk+1}(a_i)} \frac{\partial P^c_{itk+1}(a_i)}{\partial \Delta_{itk+1}} \frac{\partial \Delta_{itk+1}}{\partial \beta} \right) \right\} \phi(a_i)\, da_i$$

$$\frac{\partial \log L}{\partial \lambda} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta, a_i | y_i) \left[\sum_{t=1}^{n_i} \sum_{k=1}^{K-1} \left(R_{itk} - R_{itk+1} \frac{e^{\phi_{itk}}}{1+e^{\phi_{itk}}} \right) \right.$$

$$\times \left\{ \frac{\partial \phi_{itk}}{\partial P^c_{itk}(a_i)} \left(\frac{\partial P^c_{itk}(a_i)}{\partial \Delta_{itk}} \frac{\partial \Delta_{itk}}{\partial \lambda} + P^c_{itk}(a_i)(1 - P^c_{itk}(a_i)) s^{(t)} a_i z_i \right) \right.$$

$$\left. \left. + \frac{\partial \phi_{itk}}{\partial P^c_{itk+1}(a_i)} \left(\frac{\partial P^c_{itk+1}(a_i)}{\partial \Delta_{itk+1}} \frac{\partial \Delta_{itk+1}}{\partial \lambda} + P^c_{itk}(a_i)(1 - P^c_{itk}(a_i)) s^{(t)} a_i z_i \right) \right\} \right] \phi(a_i)\, da_i$$

DOI: 10.1002/sim

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta, a_i | y_i) \left[\sum_{t=1}^{n_i} \sum_{k=1}^{K-1} \left(R_{itk} - R_{itk+1} \frac{e^{\phi_{itk}}}{1+e^{\phi_{itk}}} \right) \right.$$
$$\times \left\{ \frac{\partial \phi_{itk}}{\partial P^c_{itk}(a_i)} P^c_{itk}(a_i)(1 - P^c_{itk}(a_i))\sigma_i \frac{\partial s^{(t)}}{\partial \alpha} a_i \right.$$
$$\left. \left. + \frac{\partial \phi_{itk}}{\partial P^c_{itk+1}(a_i)} P^c_{itk}(a_i)(1 - P^c_{itk}(a_i))\sigma_i \frac{\partial s^{(t)}}{\partial \alpha} a_i \right\} \right] \phi(a_i)\, \mathrm{d}a_i$$

where $L(\theta, a_i | y_i)$ is given by (10) and reexpressed as

$$L(\theta, a_i | y_i) = \exp\left[\left\{ \sum_{t=1}^{\mathrm{T}} \sum_{k=1}^{K-1} (R_{itk}\phi_{itk} - R_{itk+1} g(\phi_{itk})) \right\} \right]$$

for $j = 1, \ldots, K-1$. The integrals are estimated using Monte Carlo integration. We simply generate 10 000 random vectors from multivariate standard normal distributions to compute the integrals.

To make the derivatives simpler, (8) can be reexpressed as

$$P^M_{itk} = \int P^c_{itk}(\sigma_i a)\phi(a)\, \mathrm{d}a \tag{A2}$$

where $\phi(\cdot)$ is the standard normal density function. Note that the integral in (A2) is one dimensional. To compute the score vector and information matrix, we also need derivatives of Δ_{it} with respect to β_0, β and λ. They can be obtained from relationship (A2),

$$\frac{\partial P^M_{itk}}{\partial \beta} = \int \frac{\partial P^c_{itk}(\sigma_i a)}{\partial \Delta_{itk}} \frac{\partial \Delta_{itk}}{\partial \beta} \phi(a)\, \mathrm{d}a$$
$$\Rightarrow \frac{\partial \Delta_{itk}}{\partial \beta} = \frac{\partial P^M_{itk}/\partial \beta}{\int \frac{\partial P^c_{itk}(\sigma_i a)}{\partial \Delta_{itk}} \phi(a)\, \mathrm{d}a}$$

Similarly, we have

$$\frac{\partial \Delta_{itk}}{\partial \beta_{0j}} = \frac{\partial P^M_{itk}/\partial \beta_{0j}}{\int \frac{\partial P^c_{itk}(\sigma_i a)}{\partial \Delta_{itk}} \phi(a)\, \mathrm{d}a}$$
$$\frac{\partial \Delta_{itk}}{\partial \lambda} = -\frac{\int P^c_{itk}(\sigma_i a)(1 - P^c_{itk}(\sigma_i a)) a \sigma_i z_i \phi(a)\, \mathrm{d}a}{\int \frac{\partial P^c_{itk}(\sigma_i a)}{\partial \Delta_{itk}} \phi(a)\, \mathrm{d}a}$$

DOI: 10.1002/sim

A.3. Identification of the causal effects

Based on observed data and Assumptions 1–4 in Section 6.2.3, we show how to estimate the causal effects SACE. Assumption 1 implies that

$$P(Y_t(\text{tx})>k|A_t(0))=P(Y_t(\text{tx})>k|D_t(0)=0)$$

for $\text{tx}=0,1$. Assumption 2 implies that

$$P(D_t(\text{tx})=0)=P(D_t=0|\text{Tx}=\text{tx})\stackrel{\text{let}}{=}g_t(\text{tx}) \tag{A3}$$

and

$$P(Y_t(\text{tx})>k|D_t(\text{tx})=0,\bar{D}_{t-1}(\text{tx}))=P(Y_t>k|D_t=0,\bar{D}_{t-1},\text{Tx}=\text{tx})$$

for $\text{tx}=0,1$. From Assumptions 2 and 3, we have that

$$\begin{aligned}P(Y_t(\text{tx})>k|D_t(\text{tx})=0,\bar{D}_{t-1}(\text{tx}))&=P(Y_t>k|D_t=0,\bar{D}_{t-1},\text{Tx}=\text{tx},R_t=1)\\&=P(Y_t>k|D_t=0,D_{t-1},\text{Tx}=\text{tx},R_t=1)\end{aligned}$$

and

$$P(Y_t(\text{tx})>k|D_t(\text{tx})=0)=P(Y_t>k|D_t=0,\text{Tx}=\text{tx},R_t=1)\stackrel{\text{let}}{=}h_{t,\text{tx}}(k)$$

for $\text{tx}=0,1$.

Recall we need to estimate $P(Y_T(1)>k|A_T(0))$ to estimate $\text{SACE}_k(1,0)$. To do this, we show that $P(Y_T(1)>k|D_T(1)=0)$ can be expressed as

$$\begin{aligned}&P(Y_T(1)>k|D_T(1)=0)\\&=P(Y_T(1)>k|A_T(0))P(A_T(0)|D_T(1)=0)+P(Y_T(1)>k|A_T(1))P(A_T(1)|D_T(1)=0)\\&=P(Y_T(1)>k|A_T(0))P(D_T(0)=0,D_T(1)=0|D_T(1)=0)\\&\quad+P(Y_T(1)>k|A_T(1))P(D_T(0)=1,D_T(1)=0|D_T(1)=0)\\&=P(Y_T(1)>k|A_T(0))P(D_T(0)=0|D_T(1)=0)\\&\quad+P(Y_T(1)>k|A_T(1))P(D_T(0)=1|D_T(1)=0)\end{aligned} \tag{A4}$$

The original expressions $P(Y_T(\text{tx})>k|D_T(\text{tx})=0)$ for $\text{tx}=0,1$ are identifiable. However, only some of the factors in (A4) are identified. The mixing probabilities, $P(D_T(0)=0|D_T(1)=0)$ and $P(D_T(0)=1|D_T(1)=0)$, are identifiable by (A3) as

$$\begin{aligned}P(D_T(0)=0|D_T(1)=0)&=\frac{g_T(0)}{g_T(1)}\\P(D_T(0)=1|D_T(1)=0)&=1-P(D_T(0)=0|D_T(1)=0)\end{aligned}$$

However, $P(Y_T(1)>k|A_T(0))$ and $P(Y_T(1)>k|A_T(1))$ in (A4) are not identified. Given the identified components, to identify these unidentified quantities, we only need to know their ratios. All three ratios are identified via Assumption 4,

$$\frac{\text{odds}(Y_T(1)>k|A_T(1))}{\text{odds}(Y_T(1)>k|A_T(0))}=e^{\alpha_1^{(1)}}\stackrel{\text{let}}{=}\tau \tag{A5}$$

DOI: 10.1002/sim

From (A4) and (15), we can identify $P(Y_T(1)>k|A_T(0))$ by solving the following quadratic equation:

$$g_T(0)(1-\tau)x^2-\{(1-\tau)(g_T(0)+h_{T,1}(k)g_T(1))+\tau g_T(1)\}x+h_{T,1}(k)g_T(1)=0$$

where $x=P(Y_T(1)>k|A_T(0))$. When $\tau=1$,

$$P(Y_T(1)>k|A_T(0))=h_{T,1}(k)$$

When $\tau\neq 1$, we have

$$P(Y_T(1)>k|A_T(0))=-\frac{b(\tau)+\sqrt{b^2(\tau)-4a(\tau)c(1)}}{2a(\tau)}$$

where $a(\tau)=(1-\tau)g_T(0)$, $b(\tau)=(\tau-1)(h_{T,1}(k)g_T(1)+g_T(0))-\tau g_T(1)$, and $c(1)=h_{T,1}(k)g_T(1)$. Note that the solution is a decreasing function of τ.

ACKNOWLEDGEMENTS

We would like to thank Dr Daniel Sargent and Ms Erin Green of the Mayo clinic for providing the data, for their help in data collection, and for clarifying some issues with the data. The authors gratefully acknowledge two referees for their helpful comments on this paper. This project was supported by NIH grants CA85295 and HL079457.

REFERENCES

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
2. Fitzmaurice GM, Laird NM. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 1993; **80**:141–151.
3. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
4. Zeger SL, Karin MR. Generalized linear model with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**:79–86.
5. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:125–134.
6. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988; **7**:305–315.
7. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.
8. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**:461–464.
9. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 1999; **55**: 688–698.
10. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* 2002; **58**:342–351.
11. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference (with Discussion). *Statistical Science* 2000; **15**:1–26.
12. Miglioretti D, Heagerty P. Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* 2004; **5**:381–398.
13. Lee K, Daniels M. A class of Markov models for longitudinal ordinal data. *Biometrics* 2007; **63**:1060–1067.
14. Dale JR. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* 1986; **42**:909–917.
15. Molenberghs G, Lesaffre E. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* 1994; **89**:633–644.
16. Williamson JM, Kim K, Lipsitz SR. Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* 1995; **90**:1432–1437.

DOI: 10.1002/sim

17. Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* 1996; **91**:1024–1036.
18. Gibbons R, Hedeker D. Random effects probit and logistic regression models for three-level data. *Biometrics* 1997; **53**:1527–1537.
19. Todem D, Kim K, Lesaffre E. Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine* 2006; **26**:1034–1054.
20. Liu LC, Hedeker D. A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics* 2006; **62**:261–268.
21. Liu I, Agresti A. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test* 2005; **14**:1–73.
22. Hogan JW, Roy J, Korkontzelou C. Tutorial in biostatistics handling drop-out in longitudinal studies. *Statistics in Medicine* 2004; **23**:1455–1497.
23. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **16**:239–257.
24. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine* 2003; **22**:795–809.
25. Kurland BF, Heagerty PJ. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* 2005; **6**:241–258.
26. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**:21–29.
27. Egleston B, Scharfstein DO, Freeman ER, West SK. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics* 2007; **8**:526–545.
28. Rubin DB. Comment on causal inference without counterfactuals by A. P. Dawid. *Journal of the American Statistical Association* 2000; **6**:34–58.
29. Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* 2005; **61**:305–310.
30. Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with 'censoring' due to death. *Statistical Science* 2006; **21**:299–309.
31. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK, Findlay BP, Pitot HC, Albets SR. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology* 2004; **22**:23–30.
32. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society*, *Series B* 1980; **42**:109–142.
33. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001; **88**:973–985.
34. Gibbons R, Bock R. Trend in correlated proportions. *Psychometrika* 1987; **52**:113–124.
35. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
36. Liu Q, Pierce DA. A note on Gauss–Hermite quadrature. *Biometrika* 1994; **81**:624–629.
37. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
38. Fitzmaurice GM, Laird NM. Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* 2000; **1**:141–156.
39. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
40. Rubin DB. Bayesian inference for causal effects. *Annals of Statistics* 1978; **6**:34–58.
41. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–961.
42. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
43. Schildcrout J, Heagerty PJ. Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics* 2007; **63**:322–331.
44. Ilk O, Daniels M. Marginalized transition random effects models for multivariate longitudinal binary data. *Canadian Journal of Statistics* 2007; **35**:105–123.

DOI: 10.1002/sim