Published in final edited form as:

Stat Methods Med Res. 2019 March ; 28(3): 858–870. doi:10.1177/0962280217738141.

A joint model for multivariate hierarchical semicontinuous data with replications

Wondwosen Kassahun-Yimer¹, Paul S Albert², Leah M Lipsky³, Tonja R Nansel³, and Aiyi Liu¹

¹Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, USA

²Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

³Health Behavior Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, USA

Abstract

Longitudinal data are often collected in biomedical applications in such a way that measurements on more than one response are taken from a given subject repeatedly overtime. For some problems, these multiple profiles need to be modeled jointly to get insight on the joint evolution and/or association of these responses over time. In practice, such longitudinal outcomes may have many zeros that need to be accounted for in the analysis. For example, in dietary intake studies, as we focus on in this paper, some food components are eaten daily by almost all subjects, while others are consumed episodically, where individuals have time periods where they do not eat these components followed by periods where they do. These episodically consumed foods need to be adequately modeled to account for the many zeros that are encountered. In this paper, we propose a joint model to analyze multivariate hierarchical semicontinuous data characterized by many zeros and more than one replicate observations at each measurement occasion. This approach allows for different probability mechanisms for describing the zero behavior as compared with the mean intake given that the individual consumes the food. To deal with the potentially large number of multivariate profiles, we use a pairwise model fitting approach that was developed in the context of multivariate Gaussian random effects models with large number of multivariate components. The novelty of the proposed approach is that it incorporates: (1) multivariate, possibly correlated, response variables; (2) within subject correlation resulting from repeated measurements taken from each subject; (3) many zero observations; (4) over dispersion; and (5) replicate measurements at each visit time.

Corresponding author: Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6710B Rockledge Dr, Room 3230, Bethesda, MD 20817, USA., liua@mail.nih.gov.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Keywords

Beta-binomial; joint model; many zeros; multivariate; semicontinuous

1 Introduction

Longitudinal data are common in biomedical applications whereby measurements on more than one response are taken from a given subject repeatedly over time. Some applications may demand analysis of these outcomes jointly to get insight on the joint evolution and/or association of several responses over time. In practice, all or part of these multiple longitudinal responses can be characterized by many zeros. For example, in dietary intake study, as we focus in this paper, some food components are eaten daily by almost every study participant, while others are consumed episodically, so that food record data are characterized by many zeros for the latter.

There is an extensive body of literature on analysis of each single outcome separately. For a Gaussian longitudinal outcome, for example, the linear mixed model is very popular.¹ On the other hand, non-Gaussian repeated data (e.g. binary and count) are often modeled using generalized linear mixed model (GLMM). Details can be found, for example, in Engel and Keen² and Molenberghs and Verbeke.³ These modeling frameworks make use of subject-specific random effects, often assumed to be Gaussian, to account for the correlation between observations taken from the same subject repeatedly over time. Several extensions to the basic GLMM have been proposed to account for excess zeros in hierarchical count data.^{4–6} Tooze et al.⁷ studied a two-part model in the context of longitudinal semicontinuous data. Additional information on the two-part model in longitudinal semicontinuous data can be found in Olsen and Schafer.⁸ Albert and Shen⁹ proposed a two-part latent process model to simultaneously deal with repeated semicontinuous data with many zeros, serial correlation as well as missing observations. Kipnis et al.¹⁰ formulated a model to analyze data with many zeros and measurement error and applied it to evaluate relationships between episodically consumed foods and health outcomes. A survey of methods to analyze non-negative continuous and discrete univariate data having a substantial proportion of zeros in cross-sectional and repeated measures extensions have been reviewed by Min and Agresti.¹¹

It is not uncommon to find situations where several outcomes are collected simultaneously over time. When separate analysis of each outcome does not adequately address the underlying research question, joint analysis of the multiple responses is often needed. However, modeling of several longitudinal profiles jointly is complex and poses computational challenges. This led Fieuws and Verbeke¹² to propose a pairwise fitting approach to reducing the high dimensionality in estimating multivariate random effects models. Some or all of the multiple longitudinal responses can be characterized by many zeros, and this is an additional challenge for the statistical modeler. The zero and positive amounts are distinct and can each be influenced by covariates. It has been documented, based on simulation studies and real data analysis, that failure to appropriately account for many zeros leads to biased results, and hence erroneous inferences and conclusions.⁷ Zhang et al.¹³ proposed a measurement error model for multivariate cross-sectional semicontinuous

data characterized by many zeros and applied to a dietary intake study. The authors employed survey-weighted Monte Carlo computations to fit the joint distribution.

In this paper, we propose a joint model to analyze multivariate hierarchical semicontinuous data characterized by many zeros and multiple replicate observations at each measurement occasion. The novelty of the proposed approach is that it incorporates: (1) multivariate, possibly correlated, response variables; (2) within subject correlation resulting from repeated measurements taken from each subject; (3) many zero observations; (4) over dispersion; and (5) replicate measurements at each visit time.

The rest of the paper is organized as follows. In Section 2, a motivating case study is described. Section 3 focuses on a review of mixed models and their extension for many zeros to analyze a single longitudinal outcome. Section 4 proposes a zero-inflated joint model for multivariate semicontinuous longitudinal data. A likelihood-based estimation approach along the lines of Fieuws and Verbeke¹² is proposed in Section 5. Section 6 presents an analysis of longitudinal semicontinuous data from a dietary behavioral intervention study. Section 7 presents results of a simulation study. Some concluding remarks are given in Section 8.

2 Motivating case study

The Cultivating Healthful Environments in Families of Youth with Type 1 Diabetes (CHEF) study is an 18-month randomized trial to evaluate the efficacy of a family-based behavioral intervention that integrated motivational interviewing, active learning, and applied problem-solving to increase intake of whole plant foods (fruits, vegetables, whole grains, legumes, nuts, and seeds) among youth with type 1 diabetes. The study was conducted at an outpatient, free standing, multidisciplinary diabetes center in Boston, MA. Children aged 8 to 16 years were included in the study. Families in the intervention condition received sessions on healthy eating, with a focus on increasing intake of whole plant foods. Sessions subsequently applied intervention content to each meal time and other eating contexts. Participants in the control condition received no additional dietary advice beyond that provided as part of the standard type 1 diabetes care. Details of the study design, randomization procedures and treatment conditions can be found in Nansel et al.¹⁴ There were 66 and 70 children randomized to the intervention and control group, respectively. The child's usual dietary intake was estimated using three-day food records; these daily records are treated as replicate measurements at each visit. Data on intake of various food groups including total fruit, whole fruit, whole grains, among others were collected from each child at six time points, with a maximum of four replicate measurements per subject per visit, and a median number of replicates of four combining visits and subjects. Some of these components were consumed by almost every child daily, while others are episodically consumed, and hence characterized by many zeros. The latter include, for example, total fruit, whole fruit, whole grain, dark green, orange vegetables, and legumes. The longitudinal profile of each of these food components can be studied separately with little efficiency loss relative to the joint model. If interest is, however, on the association among the food components and/or global tests of covariate effects, joint modeling of the several food intake components is necessary. Table 1 shows summary of percent of zeros, as well as food group

mean with and without zeros at each measurement occasion. Evidently, there is strong evidence of many zeros having a noticeable impact on the average values across all of the four responses. The mean number of nonzero replicates is almost the same across the visits. This paper focuses on the joint modeling of total fruit, whole fruit, whole grain, dark green, orange vegetables, and legumes intake profiles characterized by many zeros and well known to be clinically important food components from nutritional viewpoint.

3 Model formulation

3.1 Mixed model for semicontinuous data with replicates

Let Y_{ijk} be the semicontinuous outcome of the k th replicate for subject i at time j . Also, let O_{ijk} be the occurrence indicator variable for the i th subject at the j th time point for the k th replicate such that

$$O_{ijk} = \begin{cases} 0 & \text{if } Y_{ijk} = 0, \\ 1 & \text{if } Y_{ijk} > 0 \end{cases}$$

The binary responses O_{ijk} 's are clustered due to the repeated measurements per subject and the replicates at each measurement occasion. One can consider a random effects approach to account for the data hierarchy stemming from replicate-level and the subject-level correlation in a unified way.

Since replicates are all exchangeable, we model the sum of replicates at each measurement time $O_{ij} = \sum_{k=1}^{n_{ij}} O_{ijk}$. The outcome O_{ij} can be modeled with a overdispersed (relative to a binomial distribution) count distribution to allow for extra-variation in the replicate sum across measurements.^{15,16} We can incorporate both replication and subject-specific variation by using a beta-binomial model for the replicate and a Gaussian random effects model for the subject-specific variation. This is in contrast to considering the normal random effects alone, which may be too restrictive.^{17,18} Alternatively, we could incorporate two levels of nested random effects (one for subject-level and one for replicate-level); however, this is more computationally expensive to fit. Our modeling strategy considered

$$O_{ij} \sim \text{Binomial}(n_{ij}, \kappa_{ij})$$

where

$$\begin{aligned} \kappa_{ij} &= \theta_{ij} p_{ij}, \quad \theta_{ij} \sim \text{Beta}(\alpha, \beta), \\ p_{ij} &= \frac{\exp(x'_{1ij} \gamma + z'_{1ij} \mathbf{b}_{1i})}{1 + \exp(x'_{1ij} \gamma + z'_{1ij} \mathbf{b}_{1i})} \end{aligned}$$

x_{1ij} are the regressors corresponding to γ , $\mathbf{b}_{1i}$ are random effects with design matrix z_{1ij} , p_{ij} are the binomial probability expressed in terms of regressors, and θ_{ij} are overdispersion random effects assumed to follow beta distribution with α and β as shape and scale parameters, respectively. It follows that

$$E(O_{ij}) = n_{ij}E(\theta_{ij})E(p_{ij}),$$

$$\text{Var}(O_{ij}) = n_{ij}E(\theta_{ij})E(p_{ij})[1 - n_{ij}E(\theta_{ij})E(p_{ij})] + n_{ij}(n_{ij} - 1)E(\theta_{ij})^2E(p_{ij})^2$$

The resulting partially marginalized density, over the beta random effects, still conditioning on the subject-level random effects, is given by

$$f_{1ij}(o_{ij} | n_{ij}, \mathbf{b}_{1i}) = \sum_{t=0}^{n_{ij}} (-1)^t \kappa_{ij}^{o_{ij}+t} \frac{n_{ij}!}{o_{ij}!(n_{ij}-o_{ij}-t)!} \frac{B(o_{ij}+t+\alpha, \beta)}{B(\alpha, \beta)} \quad (1)$$

where B is the beta function. Details on partial marginalization can be found in Molenberghs et al.¹⁷

Turning to the continuous part, suppose that S_{ijk} denotes the nonzero values of the response defined as $S_{ijk} \equiv [Y_{ijk}|O_{ijk} = 1]$. As with the O_{ijk} 's, since replicates are exchangeable, it is natural to sum them

$$S_{ij} = \sum S_{ijk}$$

We assume a Gaussian model for S_{ij} such that $(S_{ij} | \mathbf{b}_{2i}, \boldsymbol{\beta}) \sim N(n_{ij}^* \mu_{ij}, n_{ij}^* \sigma^2)$, where

$\mu_{ij} = \mathbf{x}'_{2ij} \boldsymbol{\beta} + \mathbf{z}'_{2ij} \mathbf{b}_{2i}$, n_{ij}^* is the number of nonzero replicates for subject i at visit j , $\mathbf{x}_{2ij}$ are the regressors corresponding to $\boldsymbol{\beta}$, and $\mathbf{b}_{2i}$ are random effects with design matrix $\mathbf{z}_{2ij}$. The random effects $\mathbf{b}_i = (\mathbf{b}_{1i}, \mathbf{b}_{2i})'$ are assumed jointly normal and possibly correlated. The regressors $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ij}$ can be overlapping, a subset of the regressors in the probability part, or entirely different sets of regressors. The likelihood is given by

$$L(\boldsymbol{\beta}, \gamma, D, \sigma^2) = \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_{1ij}(o_{ij} | \mathbf{b}_i) f_{2ij}(S_{ij} | \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i$$

where f_{1ij} is the beta binomial density function for O_{ij} as defined in equation (1), and f_{2ij} is the Gaussian density function for the continuous response S_{ij} .

4 Joint model for multivariate hierarchical semicontinuous data

Suppose there are m longitudinal semicontinuous responses characterized by many zeros and each can be modeled as described in Section 3.1. A joint model formulation follows from combining each single model of Y_{ij} in Section 3.1 through random effects that are assumed to follow multivariate normal distribution.

The resulting joint model takes the following general expression

$$f(y_1, \dots, y_m) = \int \dots \int \prod_{\ell=1}^m f(y_\ell | \mathbf{b}_\ell) f^*(\mathbf{b}_1, \dots, \mathbf{b}_m) d\mathbf{b}_1 \dots d\mathbf{b}_m \quad (2)$$

where $f(y_\ell | \mathbf{b}_\ell)$ is the density of the ℓ^{th} response, expressed in terms of f_{1ij} and f_{2ij} of Section 3 and f^* is a $2 \times m$ -dimensional Gaussian density for the random effects $(\mathbf{b}_1, \dots, \mathbf{b}_m)$, such that, for subject i , each $\mathbf{b}_\ell$ consists of two random effects $(\mathbf{b}_{\ell i}, \mathbf{b}_{2\ell i})$, one for the binomial part and one for the continuous part, as described in Section 3.1. The likelihood contribution of subject i is $L_i(\mathbf{Y}_{1i}, \dots, \mathbf{Y}_{mi} | \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the vector of fixed effect and covariance parameters in the joint model and constitutes components of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, D , α , and β . Estimation of the full joint model and hence inference for $\boldsymbol{\Omega}$ is not feasible in the standard statistical software due to the computational difficulty resulting from the high dimensionality. To overcome this computational challenge and reduce the dimensionality in the m responses, we adopted pairwise fitting approach proposed by Fieuws and Verbeke.¹²

5 Estimation

As mentioned in Section 4, one possible strategy to reduce the dimensionality in equation (2) and estimate the model parameters is pairwise bivariate fitting. For simplicity, in this section, we consider outcome-specific random-intercept models that are shared among the binomial part and the continuous part for a given response. More complicated models can also be done. For subject i , consider a vector of the r th and s th bivariate responses $\mathbf{Y}_i = (\mathbf{Y}'_{ri}, \mathbf{Y}'_{si})$, where $\mathbf{Y}_{ri} = (y_{ri1}, \dots, y_{rin_i})$ and $\mathbf{Y}_{si} = (y_{si1}, \dots, y_{sin_i})$, $r = 1, \dots, m-1$, $s = r+1, \dots, m$. We assume that $\mathbf{Y}_{ri}$ and $\mathbf{Y}_{si}$ are conditionally independent given $\mathbf{b}_i = (\mathbf{b}_{ri}, \mathbf{b}_{si})'$, where $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, D)$, and D is given by

$$D = \begin{pmatrix} d_r^2 & \rho d_r d_s \\ \rho d_r d_s & d_s^2 \end{pmatrix} \quad (3)$$

The likelihood to be maximized corresponding to the r th and s th bivariate two-part mixed models takes the form

$$\prod_{i=1}^N \int \int \left\{ \prod_{j=1}^{n_{ri}} f(y_{rij} | \mathbf{b}_{ri}) \right\} \left\{ \prod_{j=1}^{n_{si}} f(y_{sij} | \mathbf{b}_{si}) \right\} f^*(\mathbf{b}_{ri}, \mathbf{b}_{si}) d\mathbf{b}_{ri} d\mathbf{b}_{si} \quad (4)$$

where N is the total number of subjects, f and f^* are as defined in equation (2). The total number of possible pairs from m responses is $P = \frac{m(m-1)}{2}$, and an equivalent way of writing the loglikelihood to be maximized is $\sum_{p=1}^P l(\mathbf{Y}_p | \boldsymbol{\Omega}'_p)$, where $p = 1, \dots, P$, and $\boldsymbol{\Omega}'_p$ is the vector of parameters corresponding to the p^{th} bivariate model. Let $\boldsymbol{\Omega}'$ be the vector of parameters obtained by combining all pair-specific parameter vectors $\boldsymbol{\Omega}'_p$. Assuming that $\hat{\boldsymbol{\Omega}}'_p$

maximizes $l(Y_p | \Omega'_p)$, $\widehat{\Omega}'$ would maximize $l(\Omega')$. Then, a single estimate for each parameter in Ω is obtained by averaging all pair-specific estimates in $\widehat{\Omega}'$. The resulting averages are asymptotically normal. However, the standard errors cannot be directly computed by simple averaging of their corresponding pair-specific estimates. The pairwise fitting approach in equation (4) is equivalent to maximizing a pseudo-likelihood function of pairs of responses at different time points.^{12,19,20} As a result, the asymptotic normal distribution for $\widehat{\Omega}'$ can be written as

$$\sqrt{N}(\widehat{\Omega}' - \Omega') \sim MVN(\mathbf{0}, \mathbf{J}^{-1} \mathbf{K} \mathbf{J}^{-1}) \quad (5)$$

where $\mathbf{J}$ is a block-diagonal matrix with diagonal blocks $\mathbf{J}_{pp}$, and $\mathbf{K}$ is a symmetric matrix containing blocks $\mathbf{K}_{pq}$, i.e.

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_{11} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{J}_{pp} \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} \mathbf{K}_{11} & \cdots & \mathbf{K}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{p1} & \cdots & \mathbf{K}_{pp} \end{pmatrix}$$

where

$$\mathbf{J}_{pp} = -\frac{1}{N} \sum_{i=1}^N E \left(\frac{\partial^2 l_{pi}}{\partial \Omega'_p \partial \Omega'^T_p} \right),$$

$$\mathbf{K}_{pq} = \frac{1}{N} \sum_{i=1}^N E \left(\frac{\partial l_{pi}}{\partial \Omega'_p} \frac{\partial l_{qi}}{\partial \Omega'^T_q} \right), \quad p, q = 1, \dots, P$$

For a matrix $\mathbf{A}$, such that $\Omega = \mathbf{A}\Omega'$, then $\sqrt{N}(\widehat{\Omega} - \Omega) \sim MVN(\mathbf{0}, \mathbf{A}\Sigma\mathbf{A}^T)$, where $\Sigma = \mathbf{J}^{-1} \mathbf{K} \mathbf{J}^{-1}$, and $\widehat{\Omega} = \mathbf{A}\widehat{\Omega}'$. Hence, as mentioned in the earlier sections, estimates for Ω are obtained by averaging the corresponding elements in $\widehat{\Omega}'$, whereby $\mathbf{A}$ is used as an appropriate matrix of coefficients, and the standard errors follow from $\mathbf{A}\Sigma\mathbf{A}^T$. The components θ_{ijk} are assumed independent, and for identifiability reasons, $\frac{\beta}{\alpha}$ is constrained to be constant, as α and β are not simultaneously identifiable, in the presence of a linear predictor, because there is aliasing effect with the intercept.^{18,21} There is no closed-form expression for the pairwise components of the likelihood in equation (4). It is advisable to provide user-defined starting values for stability of the iterative process. A SAS macro has been used to fit the P bivariate two-part mixed models and combine the results whereby each pair as shown in equation (4) is fitted with the SAS procedure NLMIXED using adaptive Gaussian quadrature based on quasi-Newton optimization technique.

6 Analysis of the CHEF study

We analyze the CHEF study data described in Section 2. The four energy-adjusted food intake components, total fruit (TF), whole fruit (WF), whole grain (WG), dark green, orange

vegetables and legumes (DGOVL), will be studied ($m = 4$). Let $O_{\ell j}$, $\ell = 1, \dots, 4$ be the number of times a specific outcome takes nonzero values out of $n_{\ell j}$ replicates. Hence, for the logit

$$O_{\ell ij} | (b_i) \sim \text{Binomial}(n_{\ell ij}, \theta_{\ell ij} p_{\ell ij}), \quad \theta_{\ell ij} \sim \text{Beta}(\alpha, \beta),$$

$$\text{logit}(p_{\ell ij}) = \gamma_{\ell 0} + \gamma_{\ell 1} T_{ij} + \gamma_{\ell 2} G_i + \gamma_{\ell 3} IG_i + \gamma_{\ell 4} IG_i \times T_{ij} + b_{\ell i}$$

and given that a particular food component is consumed, the mean energy-adjusted amount is modeled as

$$\mu_{\ell ij} = \beta_{\ell 0} + \beta_{\ell 1} T_{ij} + \beta_{\ell 2} G_i + \beta_{\ell 3} IG_i + \beta_{\ell 4} IG_i \times T_{ij} + b_{\ell i}$$

where T_{ij} is the j th measurement time for subject i and defined as the difference of age at the j th measurement and age at baseline, G_i is a gender indicator coded as (1: girl; 0: boy), IG_i is an intervention group indicator for subject i coded as (1: intervention; 0: control), b_i are shared random effects among the two parts. All fixed effects parameters in the two parts are allowed to be outcome specific. The ratio $\frac{\beta}{\alpha}$ is constrained to be constant, as mentioned in Section 5. The four random intercepts are assumed correlated and jointly normally distributed. Correlation between any two pair of responses is captured through the corresponding random intercept correlations whereby random effects are shared among the probability and amount parts within each outcome.

The pairwise model fitting as outlined in Section 5 is employed. A total of six pairs of bivariate mixed models are fitted for the four responses. Results are summarized in Table 2. We observe considerable correlations among the corresponding random effects, except for WF and WG. Standard deviations of random intercepts and overdispersion are statistically significant across all responses, implying strong evidence of these phenomena. Among fixed-effect covariates, gender is found to be significant both in the amount and probability parts, except for WF and in the occurrence probability (γ_2) part of WG. However, energy-adjusted intake of TF, WF, WG, and DGOVL are not different among intervention and control groups.

One interest in the CHEF data analysis would be to test whether or not there is a joint treatment by time interaction in the four energy-adjusted food intake components. To achieve this, we employed a multivariate Wald test for the hypothesis $H_0 = \mathbf{L}(\hat{\beta}_4)$, where $\mathbf{L}$ is a 4×4 identity matrix corresponding to $\hat{\beta}_{\ell 4}$, $\ell = 1, \dots, 4$. The elements of $\mathbf{L}$ are extracted from the 62×62 matrix $\text{cov}(\hat{\beta})$. Thus, $W^2 = (\mathbf{L}\hat{\beta})'(\mathbf{L}\widehat{\text{COV}}(\hat{\beta})\mathbf{L}')^{-1}(\mathbf{L}\hat{\beta}) \sim \chi^2_{(4)}$. This test suggested a nonsignificant group by time interaction ($p = 0.46$).

Similar data could arise from household medical expenditure, in which the zeros represent patients who do not use health services, while the level of expenditures among the users can be described by the continuous distribution. Categories of expenditure include payments for medical provider visits, non-physician services, hospital inpatient stays, emergency room

services, dental visits, home health care, and prescription medications. In this example, the statistical modeler may be also interested in the joint analysis of two or more of these expenditure categories.

7 Simulation

A simulation study was conducted to investigate the performance of the proposed method. Section 7.1 focuses on bias in parameter estimates and coverage probabilities in the more general model versus a model that ignores the many zeros and replicates. Section 7.2 explores the method with missing data.

7.1 Simulation I

The aim of this section is to examine the effect of ignoring excess zeros and replication effects while both features are actually present. For this purpose 1000 data points were generated from the most general model for 200 subjects at seven time points with four replicates on the basis of the CHEF study. Three correlated longitudinal outcomes were considered. The occurrence probability for each response in the i th subject for the j th visit of the k th replicate is generated as

$$O_{ijk} \sim \text{Bernoulli}(\kappa_{ijk} = \theta_{ijk} p_{ijk}), \quad \theta_{ijk} \sim \text{Beta}(\alpha, \beta),$$

$$p_{ijk} = \frac{\exp(\gamma_0 + \gamma_1 \text{time}_{ij} + b_i)}{1 + \exp(\gamma_0 + \gamma_1 \text{time}_{ij} + b_i)}$$

Given that $O_{ijk} = 1$, the continuous part is generated as $S_{ijk} = \beta_0 + \beta_1 \text{time}_{ij} + b_i + \varepsilon_{ijk}$, and when generating data, the true parameter values for the three responses were $\gamma_1 = (-1.5, -0.1)'$, $\gamma_2 = (-1.2, -0.04)'$, $\gamma_3 = (-1.6, 0.1)'$, $\beta_1 = (-1.2, 0.05)'$, $\beta_2 = (-1.5, -0.08)'$, $\beta_3 = (-1.7, 0.01)'$, the residual errors ε_{ijk} are assumed independent, each generated from normal distribution with mean 0 and standard deviations 0.8, 1, and 0.5. We also included beta distributed random components θ_{ijk} in the linear predictor to induce additional heterogeneity in order to investigate whether they could be separable. We consider the ratio of the two positive parameters of beta distribution $\frac{\beta}{\alpha}$ to be a constant of 0.5. The random intercepts (b_{1i}, b_{2i}, b_{3i}) among the three responses are assumed correlated and generated from multivariate normal distribution with mean $\mathbf{0}$, standard deviations 1.8, 2, and 1.5, and correlation matrix

$$\rho = \begin{pmatrix} 1 & 0.8 & 0.3 \\ 0.8 & 1 & 0.6 \\ 0.3 & 0.6 & 1 \end{pmatrix}$$

Table 3 shows results of the simulation under the two-part multivariate model where all features are simultaneously dealt with. Clearly, all of the mean estimates are very close to the true values suggesting that point estimates are unbiased. Furthermore, model fitting converged for all simulations, implying that the model can be implemented and fitted in standard software regardless of its complexity. In addition, as shown in Table 4, Monte Carlo standard errors and average standard errors are quite similar across all model parameters suggesting that the standard errors are estimated well.

We also fitted the ordinary mixed model that treats the many zeros as continuous observations and replicates as independent observations. As presented in Table 5, omitting the many zeros as well as replications lead to severe impact in model fit whereby a non-negligible bias is observed on all parameters. In addition, the two models were compared in terms of 95% coverage probabilities. While the coverage probabilities under the two-part model are very close to the expected, the ordinary mixed model performed very poorly where the coverage rates for most of the parameters estimated to be 0%. The 95% coverage probabilities for parameter estimates in the two-part model and the ordinary mixed model are summarized in Table 6. These results underscore the necessity of appropriately accounting for excess zeros and replications in modeling multivariate longitudinal semicontinuous data.

Simulations were conducted to investigate the effect of omitting the replicates. This was made by taking the first observations of the data generated from the two-part model. The beta-binomial model in the O_{jj} component now reduces to the Bernoulli model, as a special case. The results suggest that parameter estimates are still unbiased but less efficient based on mean square error comparison (data not shown).

Furthermore, slightly modified versions of our earlier simulation were adopted to investigate the power gained in joint testing of the covariate effects. To achieve this, the global test procedure employed in Section 6 for the CHEF data analysis is applied to test the null hypothesis $H_0 = \beta_{11} = \beta_{12} = \beta_{13} = 0$ for the covariate time in the amount part of the model. The so obtained results are compared with Bonferroni corrected individual tests $H_{0,i} = \beta_{1i} = 0$. Three settings were considered by varying values of β_{1i} 's. Joint and individual models were fitted for the three responses. The resulting *p-values* were compared with $\alpha = 0.05$ and $\frac{\alpha}{3} = 0.0167$ for joint and individual tests, respectively. For the latter, at least one significant test is considered favoring the alternative hypothesis. The percentages of significant tests for joint (Bonferroni correction) for everyone of the three settings are 100(86.05), 95.16(64), 69.11(26). The proportions based on joint testing are consistently higher implying better statistical power relative to Bonferroni-corrected univariate tests.

7.2 Simulation II

The proposed method is studied under missing completely at random (MCAR) as well as missing at random (MAR) setting. In MCAR, the missingness mechanism is independent of both observed and unobserved data, while MAR assumes that missingness is independent of unobserved given the observed data.²² An average of 2%, 5%, 10%, 15%, 25%, and 30% observations were randomly selected and set missing at visits 2, 3, 4, 5, 6, and 7, respectively, from data simulated as in Section 7.1, with no observation missing at baseline. This defines the MCAR. For the MAR situation, a similar proportion of missingness is induced at each time point in such a way that observed data below certain baseline value (in this case 0.2) were randomly deleted from each of the first two responses (Y_1 and Y_2), while the third variable (Y_3) was set not to have any missing value. For example, at $t = 2$, to get 2% of the data missing, we draw a number from a uniform distribution for every observation in the dataset, and if it is < 0.02 and meets the cut-off point requirement, we label the value to be missing. Data were deleted from the complete case of Section 7.1 in such a way that

some pattern of missingness was induced through the cut-off point for the MAR mechanism, while the MCAR was purely random at each time point. Furthermore, simulations were conducted to study the behavior of the method in the context of a dropout model that often occurs in longitudinal studies. The simulation setting is similar to Section 7.1, but with the following dropout mechanism. For multivariate longitudinal data, say, $Y_{\ell j}$, where ℓ is response ($\ell = 1, 2, 3$), i is subject ($i = 1, \dots, 200$), j is time ($j = 1, \dots, 7$), an individual i drops out before the first follow-up ($j = 1$) with binomial probability p_{i1} , then Y_{1i1} , Y_{2i1} , Y_{3i1} and all future measurements for subject i are missing. Assuming individual i does not dropout at time 1, its drop out probability at time 2 is $p_{i2} = \text{logit}^{-1}(\xi_0 + \xi_1 Y_{1i1} + \xi_2 Y_{2i1} + \xi_3 Y_{3i1})$, then Y_{1i2} , Y_{2i2} , and Y_{3i2} and all future measurements for subject i are missing. In general, for $j = 3, \dots, 7$, assuming individual i does not dropout at time $(j-1)$, its dropout probability at time j is $p_{ij} = \text{logit}^{-1}(\xi_0 + \xi_1 Y_{1i(j-1)} + \xi_2 Y_{2i(j-1)} + \xi_3 Y_{3i(j-1)})$, then Y_{1ij} , Y_{2ij} and Y_{3ij} and all future measurements for subject i are missing.

The results for MCAR are summarized in Table 7. As expected, the parameter estimates do not show bias. Also, the simulation results under MAR in Table 8 suggest good performance of the method except for a small bias in time effect estimates corresponding to the amount (β_1) and occurrence probability (γ_1) in Y_1 and Y_2 . We observe that estimates corresponding to Y_3 are very similar in both MAR and MCAR as well as the results in Table 3.

The results for the dropout model are shown in Table 9. For $p_{i1} = 5\%$ and $1\xi = (-3, -0.2, -0.2, -0.2)'$, on average, nearly 70% of the subjects remain until end of the study. These results suggest that parameter estimates are approximately unbiased.

In addition, a simulation study to examine the efficiency loss of our approach relative to a full likelihood approach was conducted assuming MAR. This was done exactly in the same spirit as the simulation in Table 8, except the outcome specific random intercepts in Section 7.1 are now replaced by a shared random intercept common across the outcomes. This choice was made mainly because the full likelihood model with separate random effects for each outcome is computationally cumbersome. Mean square error (MSE) was computed for the full likelihood model and pairwise fitting, as shown in Table 10. These results suggest that pairwise approach does not automatically lead to efficiency loss. This is in line with Fieuws and Verbeke¹² who also suggested no evidence of efficiency loss when outcome-specific parameters are considered. Furthermore, similar simulations in MCAR and without missing data suggest slight gain in efficiency of random effect variances of the two-part full likelihood model as compared to the corresponding estimates from the pairwise or separate two-part analysis of each component, while this is not the case for the fixed effect parameter estimates with or without missing data. The random effect variance estimation in the two-part model slightly benefits from the joint modeling as this allows both the intensity and probability parts to correlate within and between the components.

In longitudinal studies, especially with many repeated measurements per subject, random slope could be important as well. Given the complexity of the proposed multivariate model and the need for random effects to be included in both model parts, adding random slope in the simulation leads to a more complex correlation structure, and hence, model fitting

becomes increasingly difficult. Even in the context of univariate semicontinuous longitudinal data, fitting two-part models in the presence of both random intercept and random slope, poses several computational challenges. An account of this can also be found, for example, in Tooze et al.⁷ We considered separate random intercept models in our simulations for simplicity and computational reasons.

8 Concluding remarks

In this paper, we proposed a joint modeling framework for multivariate longitudinal semicontinuous data characterized by many zeros and replications. A random effect approach is considered whereby correlation among multiple response profiles is captured by allowing the random effects to be correlated with each other. To account for zeros, a two-part model is employed in such a way that the first part models whether or not a response takes the zero value and the second part models the positive measurement given a nonzero.

In terms of estimation, we have focused on pairwise fitting of bivariate two-part mixed models. An advantage of this approach is that it reduces dimensionality of the random-effects structure and extremely simplifies implementation in software, as compared to the full multivariate problem. Apart from computational flexibility, the proposed approach helps to estimate correlation among individual outcomes. Furthermore, joint inference of covariate effects can be made, as shown in the CHEF data analysis.

Our simulation study demonstrated that the proposed approach leads to unbiased parameter estimates and coverage probabilities in very close agreement with targeted interval. However, failure to properly account for the zeros may severely impact on parameter estimates as well as associated inferences. In addition, joint modeling of multiple correlated responses increases power for global tests of covariate effects. The pseudo-likelihood approach has the benefit of computational simplicity, but theoretically may lose the advantage of a full likelihood in that estimation may be robust to MAR outcome data. Based on limited simulations, we observe that the pseudo-likelihood approach is approximately unbiased under MAR in practical situations. The results can not necessarily be generalized to all MAR configurations.

Extensions can be made to the proposed method. When interest is on the marginal or population-averaged inference of covariate effects in the probability part, one can make use of the so-called connector functions to link conditional and marginal models, as proposed by Heagerty and Zeger.²³ For longitudinal data with more repeated measurements per subject, random slopes can be included. In addition, separate but correlated random effects can be allowed for probability of a zero and amount parts, as shown in Tooze et al.⁷ in the context of univariate semicontinuous longitudinal responses. However, such extensions may result in computational difficulties in our setting as the covariance matrix of the random effects becomes increasingly complex.

Acknowledgements

The authors thank the Editor and the referees for their helpful and constructive comments. They also thank the Center for Information Technology, National Institutes of Health, for providing access to the high performance computational capabilities of the Biowulf cluster computer system.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development and National Cancer Institute of the National Institutes of Health.

References

1. Laird N and Ware J. Random-effects models for longitudinal data. *Biometrics* 1982; 38: 963–974. [PubMed: 7168798]
2. Engel B and Keen A. A simple approach for the analysis of generalized linear mixed models. *Statist Neer-landica* 1994; 48: 1–22.
3. Molenberghs G and Verbeke G. *Models for discrete longitudinal data*. New York: Springer, 2005.
4. Min Y and Agresti A. Random effect models for repeated measures of zero-inflated count data. *Statist Model* 2005; 5: 1–19.
5. Lee AH, Wang K, Scott J, et al. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statist Meth Med Res* 2006; 15: 47–61.
6. Kassahun W, Neyens T, Molenberghs G, et al. Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Stat Med* 2014; DOI: 10.1002/sim.6237.
7. Tooze JA, Grunwald GK and Jones RH. Analysis of repeated measures data with clumping at zero. *Statist Meth Med Res* 2002; 11: 341–355.
8. Olsen MK and Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc* 2001; 96: 730–745.
9. Albert PS and Shen J. Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial. *Appl Stat* 2005; 54: 707–720.
10. Kipnis V, Midthune D, Buckman DW, et al. Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* 2009; 65: 1003–1010. [PubMed: 19302405]
11. Min Y and Agresti A. Modeling nonnegative data with clumping at zero: A survey. *J Iran Stat Soc* 2002; 1: 7–33.
12. Fieuw S and Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* 2006; 62: 424–431. [PubMed: 16918906]
13. Zhang S, Midthune D, Guenther PM, et al. A new multivariate measurement error model with zero-inflated dietary data, and its application in dietary assessment. *Ann Appl Stat* 2011; 5: 1456–1487. [PubMed: 21804910]
14. Nansel TR, Laffel LMB, Haynie DL, et al. Improving dietary quality in youth with type 1 diabetes: Randomized clinical trial of a family-based behavioural intervention. *Int J Behav Nutr Phys Act* 2015; DOI: 10.1186/s12966-015-0214-4.
15. Hinde J and Demétrio CGB. Over dispersion: Models and estimation. *Comput Stat Data Anal* 1998; 27: 151–170.
16. Hinde J and Demétrio CGB. Over dispersion: Models and estimation. XIII Sinape: São Paulo, 1998.
17. Molenberghs G, Verbeke G, Iddi S, et al. A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *J Multivar Anal* 2012; 111: 94–109.
18. Kassahun W, Neyens T, Molenberghs G, et al. Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Arch Public Health* 2012 DOI: 10.1186/0778-7367-70-7.

19. Arnold B and Strauss D. Pseudolikelihood estimation: Some examples. *Sankhyā* 1991; 53: 233–243.
20. Geys H, Molenberghs G and Ryan L. Pseudolikelihood inference for clustered binary data. *Commun Stat: Theory Meth* 1997; 26: 2743–2767.
21. Molenberghs G, Verbeke G, Demétrio C, et al. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Stat Sci* 2010; 25: 325–347.
22. Little RJA and Rubin DB. *Statistical analysis of missing data*. New York: Wiley, 2002.
23. Heagerty PJ and Zeger SL. Marginalized multilevel models and likelihood inference (with discussion). *Statist Sci* 2000; 15: 1–26.

Table 1.

Percent of zero observations, mean with (without) zeros of energy-adjusted total fruit (TF), whole fruit (WF), whole grain (WG), and dark green and orange vegetables and legumes (DGOVL).

Outcome	Statistics	Visit ₁	Visit ₂	Visit ₃	Visit ₄	Visit ₅	Visit ₆
TF	% zeros	42.3	45.2	40.8	45.2	41.5	45.2
	Mean with zeros	0.43	0.34	0.37	0.37	0.39	0.36
	Mean without zeros	0.75	0.62	0.63	0.67	0.67	0.65
WF	% zeros	51.8	56.4	50.6	56.1	55.1	56.4
	Mean with zeros	0.32	0.21	0.25	0.24	0.24	0.23
	Mean without zeros	0.66	0.48	0.51	0.54	0.54	0.53
WG	% zeros	39.7	33.1	37.1	36.7	37.9	37.3
	Mean with zeros	0.65	0.80	0.73	0.64	0.77	0.69
	Mean without zeros	1.08	1.20	1.17	1.01	1.23	1.09
DGOVL	% zeros	69.9	68.3	68.8	67.7	64.9	62.9
	Mean with zeros	0.12	0.11	0.11	0.11	0.12	0.13
	Mean without zeros	0.39	0.35	0.34	0.33	0.35	0.36
Mean no. of nonzero replicates		2.24	2.17	2.32	2.13	2.29	2.07

Table 2.

Parameter estimates and standard errors.

Effect	Parameter	TF	WF	WG	DGOVL
		Estimate(SE)	Estimate(SE)	Estimate(SE)	Estimate(SE)
Intercept	β_0	0.162(0.329)	1.411(0.232)	0.641(0.645)	-0.155(0.186)
Time	β_1	-0.027(0.121)	-0.183(0.087)	0.250(0.382)	0.076(0.082)
Gender	β_2	0.809(0.185)	-0.091(0.131)	1.193(0.382)	0.408(0.095)
Group	β_3	0.317(0.226)	0.397(0.221)	0.818(0.425)	0.018(0.129)
Group× Time	β_4	0.022(0.195)	-0.356(0.201)	-0.262(0.482)	-0.029(0.147)
Inflation intercept	γ_0	-0.691(0.395)	0.988(0.396)	0.304(0.303)	-1.477(0.325)
Inflation time	γ_1	-0.066(0.121)	-0.198(0.169)	-0.108(0.099)	0.202(0.135)
Inflation gender	γ_2	0.643(0.211)	0.023(0.175)	0.197(0.167)	0.625(0.175)
Inflation group	γ_3	0.385(0.238)	0.383(0.284)	0.141(0.193)	-0.130(0.191)
Inflation Group× Time	γ_4	-0.039(0.197)	-0.238(0.272)	0.168(0.142)	-0.049(0.205)
Std dev. RE	d	0.983(0.080)	0.345(0.070)	0.841(0.078)	0.460(0.075)
Std dev. error	σ	1.396(0.213)	1.417(0.246)	3.906(0.552)	0.746(0.103)
Ratio	$\frac{\beta}{\alpha}$	0.059(0.027)	0.616(0.086)	0.042(0.019)	0.243(0.097)
Correlation					
TF		1	0.44(0.15)	0.55(0.09)	0.47(0.13)
WF			1	0.08(0.14)	0.40(0.17)
WG				1	0.51(0.10)
DGOVL					1

TF: total fruit; WF: whole fruit; WG: whole grain; DGOVL: dark green, orange vegetables and legumes.

Table 3.

Summary of simulation using the two-part model with pairwise fitting.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		True	Mean(RB)	True	Mean(RB)	True	Mean(RB)
Intercept	β_0	-1.2	-1.196(0.003)	-1.5	-1.492(0.005)	-1.7	-1.700(0.000)
Time	β_1	0.05	0.050(0.000)	-0.08	-0.080(0.000)	0.01	0.010(0.000)
Std dev. error	σ	0.8	0.796(-0.005)	1	0.998(-0.002)	0.5	0.499(-0.002)
Std dev. random effect	d	1.8	1.780(-0.011)	2	1.976(-0.012)	1.5	1.485(-0.010)
Occurrence intercept	γ_0	-1.5	-1.538(0.025)	-1.2	-1.229(0.024)	-1.6	-1.654(0.033)
Occurrence time	γ_1	-0.1	-0.098(-0.02)	-0.04	-0.039(-0.025)	0.1	0.100(0.000)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.479(-0.042)	0.5	0.488(-0.024)	0.5	0.465(-0.070)
Correlation (γ_1, γ_2)	ρ_{12}	0.8	0.804(0.005)				
Correlation (γ_1, γ_3)	ρ_{13}	0.3	0.307(0.023)				
Correlation (γ_2, γ_3)	ρ_{23}	0.6	0.605(0.008)				

Table 4.

Monte Carlo standard error (MC.SE) and average standard error (AVG.SE) using the two-part model with pairwise fitting.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		MC.SE	AVG.SE	MC.SE	AVG.SE	MC.SE	AVG.SE
Intercept	β_0	0.144	0.153	0.155	0.158	0.110	0.116
Time	β_1	0.014	0.014	0.015	0.015	0.008	0.008
Std dev. error	σ	0.028	0.029	0.031	0.030	0.016	0.015
Std dev. random effect	d	0.115	0.116	0.114	0.111	0.092	0.082
Occurrence intercept	γ_0	0.185	0.191	0.187	0.189	0.163	0.164
Occurrence time	γ_1	0.027	0.026	0.024	0.025	0.024	0.023
Ratio	$\frac{\beta}{\alpha}$	0.091	0.090	0.067	0.067	0.090	0.089
Correlation (γ_1, γ_2)	ρ_{12}	0.030	0.031				
Correlation (γ_1, γ_3)	ρ_{13}	0.071	0.073				
Correlation (γ_2, γ_3)	ρ_{23}	0.051	0.047				

Table 5.

Summary of simulation using the ordinary mixed model with pairwise fitting.

Effect	Parameter	Response 1 (γ_1)		Response 2 (β_2)		Response 3 (γ_3)	
		True	Mean(RB)	True	Mean(RB)	True	Mean(RB)
Intercept	β_0	-1.2	0.071(-1.059)	-1.5	0.038(-1.025)	-1.7	-0.069(-0.959)
Time	β_1	0.05	0.009(-0.982)	-0.08	-0.013(-0.838)	0.01	-0.009(-1.900)
Std dev. error	σ	0.8	0.558(-0.303)	1	0.721 (-0.279)	0.5	0.565(0.130)
Std dev. random effect	d	1.8	0.389(-0.784)	2	0.407(-0.797)	1.5	0.249(-0.834)
Correlation (γ_1, γ_2)	ρ_{12}	0.8	0.667(-0.166)				
Correlation (γ_1, γ_3)	ρ_{13}	0.3	0.148(-0.507)				
Correlation (γ_2, γ_3)	ρ_{23}	0.6	0.399(-0.335)				

Table 6.

95% coverage probabilities for the two-part model and ordinary mixed model with pairwise fitting.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		FM	RM	FM	RM	FM	RM
Intercept	β_0	0.962	0.000	0.964	0.000	0.964	0.000
Time	β_1	0.940	0.000	0.968	0.000	0.946	0.000
Std dev. error	σ	0.956	0.000	0.948	0.000	0.936	0.050
Std dev. random effect	d	0.944	0.000	0.944	0.000	0.908	0.000
Occurrence intercept	γ_0	0.958	–	0.956	–	0.944	–
Occurrence time	γ_1	0.944	–	0.952	–	0.938	–
Ratio	$\frac{\beta}{\alpha}$	0.940	–	0.952	–	0.932	–
Correlation (γ_1, γ_2)	ρ_{12}	0.934	0.704				
Correlation (γ_1, γ_3)	ρ_{13}	0.938	0.682				
Correlation (γ_2, γ_3)	ρ_{23}	0.929	0.508				

Table 7.

Summary of simulation using the two-part model with pairwise fitting for MCAR.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		True	Mean(RB)	True	Mean(RB)	True	Mean(RB)
Intercept	β_0	-1.2	-1.190(-0.008)	-1.5	-1.492(-0.005)	-1.7	-1.691(-0.005)
Time	β_1	0.05	0.050(0.000)	-0.08	-0.079(-0.013)	0.01	0.010(0.000)
Std dev. error	σ	0.8	0.796(-0.005)	1	0.998(-0.002)	0.5	0.499(-0.002)
Std dev. random effect	d	1.8	1.779(-0.012)	2	1.981(-0.010)	1.5	1.481(-0.013)
Occurrence intercept	γ_0	-1.5	-1.537(0.024)	-1.2	-1.223(0.019)	-1.6	-1.646(0.028)
Occurrence time	γ_1	-0.1	-0.098(-0.02)	-0.04	-0.039(-0.025)	0.1	0.099(-0.010)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.477(-0.048)	0.5	0.488(-0.024)	0.5	0.464(-0.078)
Correlation (γ_1, γ_2)	ρ_{12}	0.8	0.804(0.005)				
Correlation (γ_1, γ_3)	ρ_{13}	0.3	0.307(0.023)				
Correlation (γ_2, γ_3)	ρ_{23}	0.6	0.605(0.008)				

Table 8.

Summary of simulation using the two-part model with pairwise fitting for MAR.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		True	Mean(RB)	True	Mean(RB)	True	Mean(RB)
Intercept	β_0	-1.2	-1.194(-0.005)	-1.5	-1.497(-0.002)	-1.7	-1.690(-0.006)
Time	β_1	0.05	0.058(0.148)	-0.08	-0.071(-0.126)	0.01	0.010(0.000)
Std dev. error	σ	0.8	0.796(-0.005)	1	0.999(-0.001)	0.5	0.499(-0.002)
Std dev. random effect	d	1.8	1.778(-0.012)	2	1.983(-0.008)	1.5	1.481(-0.013)
Occurrence intercept	γ_0	-1.5	-1.579(0.050)	-1.2	-1.236(0.029)	-1.6	-1.640(0.024)
Occurrence time	γ_1	-0.1	-0.084(-0.190)	-0.04	-0.037(-0.081)	0.1	0.099(-0.010)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.455(-0.100)	0.5	0.481(-0.040)	0.5	0.466(-0.073)
Correlation (γ_1, γ_2)	ρ_{12}	0.8	0.805(0.006)				
Correlation (γ_1, γ_3)	ρ_{13}	0.3	0.306(0.020)				
Correlation (γ_2, γ_3)	ρ_{23}	0.6	0.604(0.007)				

Table 9.

Summary of simulation using the two-part model with pairwise fitting for dropout missing data.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		True	Mean(RB)	True	Mean(RB)	True	Mean(RB)
Intercept	β_0	-1.2	-1.186(-0.012)	-1.5	-1.486(-0.008)	-1.7	-1.689(-0.007)
Time	β_1	0.05	0.051(0.020)	-0.08	-0.078(0.026)	0.01	0.010(0.000)
Std dev. error	σ	0.8	0.797(-0.004)	1	0.997(-0.003)	0.5	0.499(-0.002)
Std dev. random effect	d	1.8	1.775(-0.014)	2	1.978(-0.011)	1.5	1.482(-0.012)
Occurrence intercept	γ_0	-1.5	-1.525(0.016)	-1.2	-1.217(0.014)	-1.6	-1.639(0.024)
Occurrence time	γ_1	-0.1	-0.099(-0.010)	-0.04	-0.040(0.000)	0.1	0.100(0.000)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.484(-0.033)	0.5	0.489(-0.022)	0.5	0.470(-0.064)
Correlation (γ_1, γ_2)	ρ_{12}	0.8	0.804(0.005)				
Correlation (γ_1, γ_3)	ρ_{13}	0.3	0.308(0.026)				
Correlation (γ_2, γ_3)	ρ_{23}	0.6	0.602(0.003)				

Table 10.

Mean and mean square error (MSE) of the two-part model with full likelihood and pairwise fitting under MAR simulation for shared random effects.

Effect	Parameter	Response 1 (γ_1)		Response 2 (γ_2)		Response 3 (γ_3)	
		True	Mean(MSE)	True	Mean(MSE)	True	Mean(MSE)
	Full likelihood						
Intercept	β_0	-1.2	-1.203(0.0140)	-1.5	-1.502(0.0143)	-1.7	-1.699(0.0132)
Time	β_1	0.05	0.057(0.0003)	-0.08	-0.072(0.0003)	0.01	0.010(0.0001)
Std dev. error	σ	0.8	0.798(0.0007)	1	1.002(0.0009)	0.5	0.499(0.0002)
Occurrence intercept	γ_0	-1.5	-1.603(0.0462)	-1.2	-1.267(0.0336)	-1.6	-1.650(0.0295)
Occurrence time	γ_1	-0.1	-0.085(0.0008)	-0.04	-0.038(0.0006)	0.1	0.098(0.0005)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.429(0.0183)	0.5	0.450(0.0110)	0.5	0.462(0.0086)
Std dev. random effect	d	1.5	1.493(0.0066)				
	Pairwise fitting						
Intercept	β_0	-1.2	-1.199(0.0149)	-1.5	-1.501(0.0149)	-1.7	-1.696(0.0131)
Time	β_1	0.05	0.057(0.0003)	-0.08	-0.072(0.0003)	0.01	0.010(0.0001)
Std dev. error	σ	0.8	0.799(0.0007)	1	1.003(0.0010)	0.5	0.499(0.0002)
Occurrence intercept	γ_0	-1.5	-1.599(0.0475)	-1.2	-1.274(0.0341)	-1.6	-1.647(0.0284)
Occurrence time	γ_1	-0.1	-0.086(0.0009)	-0.04	-0.037(0.0006)	0.1	0.099(0.0006)
Ratio	$\frac{\beta}{\alpha}$	0.5	0.428(0.0186)	0.5	0.446(0.0110)	0.5	0.463(0.0086)
Std dev. random effect	d	1.5	1.489(0.0074)				