Published in final edited form as:
Stat Methods Med Res. 2015 April ; 24(2): 194–205. doi:10.1177/0962280211414620.

A likelihood-based two-part marginal model for longitudinal semi-continuous data

Li Su[*], **Brian D. M. Tom**, and **Vernon T. Farewell**
Medical Research Council Biostatistics Unit, Robinson Way, Cambridge CB2 0SR, UK

Abstract

Two-part models are an attractive approach to analyzing longitudinal semicontinuous data consisting of a mixture of true zeros and continuously distributed positive values. When interest lies in the population-averaged (marginal) covariate effects, two-part models that provide straightforward interpretation of the marginal effects are desirable. Presently, the only available approaches for fitting two-part marginal models to longitudinal semicontinuous data are computationally difficult to implement. Therefore there exists a need to develop two-part marginal models that can be easily implemented in practice. We propose a fully likelihood-based two-part marginal model that satisfies this need by using the bridge distribution for the random effect in the binary part of an underlying two-part mixed model; and its maximum likelihood estimation can be routinely implemented via standard statistical software such as the SAS NLMIXED procedure. We illustrate the usage of this new model by investigating the marginal effects of pre-specified genetic markers on physical functioning, as measured by the Health Assessment Questionnaire (HAQ), in a cohort of psoriatic arthritis (PsA) patients from the University of Toronto Psoriatic Arthritis Clinic. An added benefit of our proposed marginal model when compared to a two-part mixed model is the robustness in regression parameter estimation when departure from the true random effects structure occurs. This is demonstrated through simulation.

Keywords

Bridge distribution; Logit link; Repeated measures; Random effects

1 Introduction

Over the last decade or so, the two-part modelling framework has become increasingly popular when analysing ‘semicontinuous’ response data measured either cross-sectionally or repeatedly over time1–10. By semicontinuous data we refer to data generated from a response which is a mixture of true zeros and continuously distributed positive values4. For this type of data, it is natural to view the response observed as the result of two processes, one determining whether the response is zero and the other determining the actual value if it is non-zero; and for convenience, we refer to the data arising from these two processes as the ‘binary part’ and the ‘continuous part’ of the original data, respectively.

[*]Address for correspondence: Li Su, MRC Biostatistics Unit, Robinson Way, Cambridge CB2 0SR, UK. Email: li.su@mrc-bsu.cam.ac.uk; phone: 44-1223-760722; fax: 44-1223-760729.

Europe PMC Funders Author Manuscripts

In the case of longitudinal semicontinuous data, two approaches have been proposed within this two-part modelling framework. The first is based on two-part mixed models with correlated random effects in both parts of the model4–6,11. The other is based on two-part marginal models7. The approach adopted will depend on the aims of the study and the intended purposes for the results obtained. If the objective is to investigate the effects of covariates at the subject-specific level (conditional effects) then the two-part mixed modelling approach is appropriate. For example, in the fitted two-part mixed model from the analysis reported in Su *et al.*11, the regression coefficients in the binary part for explanatory variables, such as disease activity and disease damage, can be interpreted as the log odds ratios representing the change in the probability of being functionally disabled for any specific patient who had one unit increase in disease activity or disease damage over time. The corresponding regression coefficients in the continuous part represented the expected change in any observed (non-zero) disability level for a patient with one unit increase in disease activity or disease damage over time. On the other hand, if, as is the case in this article, straightforwardly interpreted covariate effects at the population-averaged level (marginal effects) are required, then two-part marginal models are needed. For example, it would be interesting to investigate whether on average the patients with certain genetic markers had different odds of being functionally disabled or different mean disability level than those patients without those genetic markers. A subject-specific interpretation for the genetic marker effects would be less attractive here as genetic markers are time-invariant within the same patients. It is worth noting that in generalized linear models for longitudinal data, marginal and conditional effects will differ in magnitude unless linear models with an identity link are used (see detailed discussion in Chapter 7 of Diggle *et al.*12).

Currently, when interpretation of population-averaged covariate effects is of interest, there are only moment-based two-part modelling approaches available for fitting longitudinal semicontinuous data. In particular, Hall and Zhang7 have described both a direct estimation method based on generalized estimating equations (GEE) for the observed semicontinuous responses alone and an Expectation-Solution (ES) algorithm with GEE in the S-step for estimating the marginal covariate effects. However, because of the complexity of the estimating equations and algorithms, both methods require specialized programs that are not readily available to analysts and would require considerable statistical programming skills for implementation. Therefore it would seem advantageous to develop a two-part marginal model which can be conveniently and routinely implemented in practice.

In this article, we propose a likelihood-based approach to the two-part marginal modelling of longitudinal semicontinuous data. Specifically, our two-part marginal model is derived from an underlying two-part mixed model where the random intercept in the conditional logistic model for the binary part and the random intercept in the linear mixed model for the continuous part are assumed to be correlated and follow a bridge distribution (instead of a Normal distribution as per usual) and a Normal distribution, respectively13,14. The marginal covariate effects are directly specified in both parts of the model because the marginal expectations in both parts preserve the logit and identity links after integration over the random effects. The integration can be achieved using adaptive Gaussian quadrature techniques and the likelihood is then maximized by performing quasi-Newton optimization6.

In Section 2 we describe the work on the association between genetic markers and physical functioning in psoriatic arthritis which partly motivated this research. Section 3 describes formally our two-part marginal model for longitudinal semicontinuous data. We conduct a simulation study to evaluate how the two-part marginal model performs under a plausible departure from the true underlying random effect structure in Section 4 and the psoriatic arthritis data are then analyzed in Section 5 to illustrate the methods. We conclude the article in Section 6.

2 Motivating example

This research on developing an appropriate two-part marginal model was partly motivated by work on a dataset from The University of Toronto Psoriatic Arthritis (PsA) Clinic15. The Health Assessment Questionnaire (HAQ) is a self-report functional status (disability) measure that has become the dominant instrument in many disease areas, including arthritis16. It produces a measure that has a point mass at zero, whilst non-zero values vary "continuously" in the range zero (no disability) to three (completely disabled). Since June 1993, the HAQ has been administered annually to patients in the PsA Clinic and, as of March 2005, 382 patients had completed at least two HAQs with 2107 observations in total for analyses17.

In the earlier work on HAQ11,17, our objective was to examine whether the effects of disease activity and disease damage on physical functioning (as measured by the HAQ) were changing over the PsA disease duration. On examining these data (see Figure 1), a notable feature was the relatively high preponderance of zeros (i.e. observation cluster at zero of 645⁄2107 = 30.6%), which presented a challenge in characterizing the relationship between the HAQ scores and covariates. Our use of two-part mixed models allowed us to overcome this challenge and investigate the changing relationship of disease activity and damage with physical functioning; both in terms of distinguishing a PsA patient when no functional disability (HAQ score = 0) occurs to when at least mild difficulty (HAQ >0) occurs, and in determining the impact on the actual level of difficulty (represented by positive HAQ scores), given that the patient had at least mild difficulty. These effects of interest were at the subject-specific level and therefore mixed models were deemed appropriate. Moreover, it was found to be important to allow the random effects in both parts of the two-part mixed model to be correlated (rather than wrongly assumed independent) otherwise bias would ensue in the parameter estimators obtained for the continuous part11.

In a study characterizing the relationship between genetic markers and disease progression in psoriatic arthritis18, a number of alleles that code for HLA antigens were found to be associated with progression of clinical damage. HLA-B27 in the presence of HLA-DR7, HLA-B39, and HLA-DQw3 in the absence of HLA-DR7 were predictive of progression of clinical damage, whereas HLA-B22 was protective. Here the marginal effects of the various genetic markers on disease progression were of interest, that is, we aimed to investigate whether on average the patients with genetic markers present had more clinical damage than those without genetic markers.

In more recent follow-up work, we are interested in investigating the relationship of the aforementioned HLA alleles with physical functioning, as measured by the HAQ. The question to be answered is whether patients with those specific HLA alleles had on average different levels of physical functioning over time than others. The marginal effects of these genetic markers were again of interest, but the HAQ data to be used were repeatedly measured over time and had, as described earlier, a large number of observations clustered at zero. To analyze these data, two-part marginal models which would provide straightforward interpretation were found to be needed. However no such easily implementable method to achieve this was available in practice.

In the next section, we propose a two-part marginal model that is easily implementable and interpretable and will allow us to analyze the above HAQ data.

3 Model

We build our two-part marginal model based on the original two-part mixed models introduced in Olsen and Schafer4 and Tooze *et al.*6 and the random effects specifications in Lin *et al.*14. Let Y_{ij} be a semicontinuous variable for the ith ($i = 1, \ldots, N$) subject at time t_{ij} ($j = 1, \ldots, n_i$). This response variable can be represented by two variables, the occurrence variable

$$Z_{ij} = \begin{cases} 0 & \text{if } Y_{ij} = 0 \\ 1 & \text{if } Y_{ij} > 0 \end{cases}$$

and the intensity variable $g(Y_{ij})$ given that $Y_{ij} > 0$, where $g(\cdot)$ is a transformation that makes $Y_{ij} \mid Y_{ij} > 0$ approximately Normally distributed with a subject-time-specific mean.

Instead of focusing on the marginal distribution of Y_{ij}, in a two-part model we are interested in both the distribution for the occurrence variable Z_{ij} and the conditional distribution of the intensity variable $g(Y_{ij})$ given that $Y_{ij} > 0$. Specifically, it is assumed that Z_{ij} follows a random effects logistic regression model with

$$\text{logit}\left\{\Pr\left(Z_{ij}=1 \middle| B_i\right)\right\} = \mathbf{X}_{ij}\tilde{\boldsymbol{\theta}} + B_i, \quad (3.1)$$

where $\mathbf{X}_{ij}$ is a $1 \times q$ covariate vector, $\tilde{\boldsymbol{\theta}}$ is a $q \times 1$ regression coefficient vector and B_i is the subject-level random intercept. The intensity variable $g(Y_{ij})$ given $Y_{ij} > 0$ follows a linear mixed model

$$g(Y_{ij}) | V_i, Y_{ij} > 0 = \mathbf{X}^*_{ij}\boldsymbol{\beta} + V_i + \epsilon_{ij}, \quad (3.2)$$

where $\mathbf{X}^*_{ij}$ is a $1 \times p$ covariate vector, $\boldsymbol{\beta}$ is a $p \times 1$ regression coefficient vector and V_i is again a subject-level random intercept. The error term ϵ_{ij} is assumed to be distributed as $N(0, \sigma^2_e)$. Note that the covariate vectors $\mathbf{X}_{ij}$, $\mathbf{X}^*_{ij}$ may coincide, but this is not required.

Further, we assume that B_i, the random intercept in the binary part, follows the bridge density of Wang and Louis13

$$f_B(b_i|\phi)=\frac{1}{2\pi}\frac{sin(\phi\pi)}{cosh(\phi b_i)+cos(\phi\pi)}(-\infty<b_i<\infty)$$

with unknown parameter φ ($0 < \varphi < 1$). This bridge distribution is symmetric with mean zero and variance $\sigma_b^2=\pi^2(\phi^{-2}-1)/3$. It is slightly heavy tailed and more concentrated than the Normal distribution with the same variance. The key characteristic of this bridge density is that after integration over the random intercepts, (B_i, V_i), the marginal probability $\Pr(Z_{ij} = 1)$ relates to the linear predictors through the same logit link function as for the corresponding conditional probability. In addition, if we specify the marginal regression structure of the binary part as

$$\text{logit}\left\{\Pr\left(Z_{ij}=1\right)\right\}=\mathbf{X}_{ij}\boldsymbol{\theta},$$

then the marginal covariate effects $\boldsymbol{\theta}$ are proportional to the subject-specific conditional covariate effects $\tilde{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}=\phi\tilde{\boldsymbol{\theta}}$. Therefore, we could rewrite (3.1) as

$$\text{logit}\left\{\Pr\left(Z_{ij}=1|B_i\right)\right\}=\mathbf{X}_{ij}\boldsymbol{\theta}/\phi+B_i. \quad (3.3)$$

Based on marginalization of random effects models, Heagerty19 and Heagerty and Zeger20 proposed full likelihood-based methods of estimating marginal regression parameters for longitudinal binary data. In their models, random effects are assumed to be Normally distributed and the marginal probability and the conditional probability given the random effects are matched by an intercept term Δ_{ij}. Similarly, in our model we have

$$\Pr(Z_{ij}=1)=\int\Pr(Z_{ij}=1|b_i)f_B(b_i)db_i=\int\text{logit}^{-1}(\Delta_{ij}+b_i)f_B(b_i)db_i,$$

and the intercept term is actually $\Delta_{ij}=\mathbf{X}_{\text{ij}}\tilde{\boldsymbol{\theta}}$.

For the continuous part of the model, we let V_i be Normally distributed with mean zero and variance σ_v^2. Therefore, $g(Y_{ij}) \mid Y_{ij} > 0$ given the random intercepts (B_i, V_i) follows a Normal linear mixed model with mean $\mathbf{X}_{ij}^*\boldsymbol{\beta}+V_i$ and variance σ_e^2. It follows that the marginal mean of $g(Y_{ij}) \mid Y_{ij} > 0$ integrated over (B_i, V_i) is $\mathbf{X}_{ij}^*\boldsymbol{\beta}$, the fixed effects part of the linear mixed model.

It is natural to conjecture that the two processes that generate semicontinuous data may be related, especially if the response is observed at multiple time points. Therefore, we construct a bivariate joint distribution for the random intercepts (B_i, V_i) from a pair of Normal random variables

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma_\upsilon \\ \rho\sigma_\upsilon & \sigma_\upsilon^2 \end{bmatrix} \right), \quad (3.4)$$

and use the probability integral transformation

$$B_i = F_B^{-1}\{\Phi(U_i)\}$$

to obtain B_i13,14. Here $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal, and $F_B^{-1}(\cdot)$ is the inverse cumulative distribution function,

$$F_B^{-1}(x) = \frac{1}{\phi} log\left[\frac{sin(\phi\pi x)}{sin\{\phi\pi(1-x)\}}\right]$$

of the bridge density for $0 < x < 1$. Lin *et al.*14 found that the correlation for (B_i, V_i) is approximately the same as the correlation ρ for (U_i, V_i).

In this two-part marginal model, we consider the primary targets of inference to be the marginal covariate effects $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, while variance components σ_b^2 (or equivalently φ), σ_υ^2, σ_e^2 and the correlation parameter ρ are treated as nuisance parameters. The estimation of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, σ_b^2, σ_υ^2, ρ and σ_e^2 is based on maximization of the likelihood

$$\begin{aligned} L &= \prod_{i=1}^{N} \int_{b_i} \int_{\upsilon_i} \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}, \boldsymbol{\beta}, b_i, \upsilon_i, \sigma_e^2) f(b_i, \upsilon_i|\sigma_b^2, \sigma_\upsilon^2, \rho) d\upsilon_i db_i \\ &= \prod_{i=1}^{N} \int_{b_i} \int_{\upsilon_i} \prod_{j=1}^{n_i} \{1 - Pr(Z_{ij}=1|\boldsymbol{\theta}, b_i)\}^{(1-z_{ij})} \{Pr(Z_{ij}=1|\boldsymbol{\theta}, b_i)\}^{z_{ij}} \\ &\quad \times \{f(g(y_{ij})|\boldsymbol{\beta}, \upsilon_i, \sigma_e^2)\}^{z_{ij}} f(b_i, \upsilon_i|\sigma_b^2, \sigma_\upsilon^2, \rho) \, d\upsilon_i db_i, \end{aligned} \quad (3.5)$$

which can be implemented in the SAS NLMIXED procedure by quasi-Newton optimization with adaptive Gaussian quadrature techniques6.

There are three advantages of this marginally specified two-part model. First, compared with alternative two-part marginal modelling specifications, it can be conveniently implemented using standard software procedures such as SAS NLMIXED. Second, compared with the moment-based approaches in Hall and Zhang7, it can deal with unbalanced longitudinal data either by design or due to ignorable missingness (such as ‘Missing at Random’ (MAR)) because it is fully likelihood-based12,21. Third, compared with the two-part mixed model, it can offer some degree of robustness in regression parameter estimation when departure from the true underlying random effect structure occurs. For generalized linear mixed models (GLMM), it has been shown that even point estimates, under certain conditions, can be sensitive to assumptions made regarding the random effect structure19,20,22–28. In particular, Heagerty and Kurland25 showed that substantial bias can arise for the subject-

Europe PMC Funders Author Manuscripts

specific conditional covariate effects in a GLMM when the true random effect structure includes both a random intercept and a random slope but the specified model includes only the random intercept, whereas the marginally specified regression structure can be more robust to this violation of the random effect structure assumption. The situation for longitudinal semicontinuous data is analogous: because of the computational burden, a random intercept is often assumed in practice for the conditionally specified regression structure in the binary part of a two-part mixed model and this could give rise to biased point estimates of the conditional covariate effects when an additional true random slope is ignored. In this scenario, a marginally specified two-part model, with marginal interpretation of covariate effects, might be preferable although this would be dependent on the purpose of the study. We will conduct a simulation study to further investigate this issue in Section 4.

4 Simulation Study

Here we describe and report the findings from our simulation study to investigate the performance of our proposed two-part marginal model and the original two-part mixed model with bivariate Normal random intercepts, when the underlying random effects assumption is violated. We shall explicitly focus on the scenario in which the true random effect structure in the binary part include both random intercept and random slope but the models to be fitted incorporate a random intercept only in this part. The true random effect structure in the continuous part includes only the random intercept and the models to be fitted will include the random intercept alone in the continuous part. The true random effects are generated from the trivariate Normal distribution in (4.1). Our objective is to investigate the relative biases in the marginal covariate effects and conditional covariate effects under this misspecification of the random effect structure in the binary part. The setups for investigating these biases for marginal and conditional effects are described next.

4.1 Setup for marginal covariate effects

Let the marginal covariate vector $\mathbf{X}_{ij} = (1, G_i, t_{ij}, G_i t_{ij})$ follow a group by time design, where $G_i \in (0, 1)$ is a group membership indicator, $t_{ij} = (j-1)/(n_i - 1)$, $j = 1, \ldots, n_i$ and $n_i = 5 \; \forall \; i$. Further, for illustration, we assume that subjects have equal probability of being in the two groups, in other words, $\Pr(G_i = g) = 1/2$ $(g = 0, 1)$. The response variables Y_{ij}, Z_{ij} are defined in the same way as in Section 3, and data are simulated from a logistic-lognormal mixture distribution with

$$\begin{aligned} logit\left\{\Pr\left(Z_{ij}=1\right)\right\} &= \theta_0+\theta_1 t_{ij}+\theta_2 G_i+\theta_3 G_i t_{ij}, \\ logit\left\{\Pr\left(Z_{ij}=1 \middle| U_{0i},\ U_{1i}\right\}\right. &= \Delta_{ij}+U_{0i}+U_{1i}t_{ij}, \\ \left[log(Y_{ij}) \middle| V_i, Y_{ij}>0\right] &\sim N(\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 t_{ij}+\boldsymbol{\beta}_2 G_i+\boldsymbol{\beta}_3 G_i t_{ij}+V_i, \sigma_\varepsilon^2), \end{aligned}$$

and with correlated random effects

$$\begin{bmatrix} U_{0i} \\ U_{1i} \\ V_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & 0 & \rho_0\sigma_{u_0}\sigma_v \\ 0 & \sigma_{u_1}^2 & \rho_1\sigma_{u_1}\sigma_v \\ \rho_0\sigma_{u_0}\sigma_v & \rho_1\sigma_{u_1}\sigma_v & \sigma_v^2 \end{bmatrix}\right). \quad (4.1)$$

Note that Δ_{ij} satisfies

$$\text{logit}^{-1}(\theta_0+\theta_1 t_{ij}+\theta_2 G_i+\theta_3 G_i t_{ij})=\int\int \text{logit}^{-1}(\Delta_{ij}+u_{0i}+u_{1i}t_{ij}) f(u_{0i},u_{1i}) du_{0i} du_{1i},$$

and we use Newton-Raphson algorithm with two-dimensional Gaussian quadrature to compute Δ_{ij}19,29. We then generate 500 datasets with N = 500 subjects using the set of parameter values given in Section 4.3. The two-part marginal model described in Section 3 is then fitted with the marginal mean regression structures correctly specified, but assuming that the random effect structure in the binary part only includes a random intercept from the bridge distribution. We also fit a two-part mixed model with correlated Normal random intercepts and with conditional mean structures for the fixed effects following the group by time design. To obtain the approximate marginal covariate effects in the binary part, we use the methods in Zeger *et al.*30, and multiply the conditional covariate effects by an attenuation factor $[1+\{16\sqrt{3}/(15\pi)\}^2\sigma^2_{u_0}]^{-1/2}$.

4.2 Setup for conditional covariate effects

Similarly, for conditional covariate effects, we simulate data from a logistic-lognormal mixture distribution with

$$\begin{aligned}\text{logit}\{\Pr(Z_{ij}=1|U_{0i},U_{1i})\} &= \tilde{\theta}_0+\tilde{\theta}_1 t_{ij}+\tilde{\theta}_2 G_i+\tilde{\theta}_3 G_i t_{ij}+U_{0i}+U_{1i}t_{ij},\\ [\log(Y_{ij})|V_i, Y_{ij}>0] &\sim N(\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 t_{ij}+\boldsymbol{\beta}_2 G_i+\boldsymbol{\beta}_3 G_i t_{ij}+V_i,\sigma^2_e),\end{aligned}$$

and with the random effects structure in (4.1).

Five hundred datasets with N = 500 subjects are generated for each set of parameter values given in Section 4.3. Again, the two-part marginal model described in Section 3 and a two-part mixed model with correlated Normal random intercepts are fitted to the simulated data. The conditional mean structures for the fixed effects are both correctly specified and we focus on their estimated conditional covariate effects.

4.3 Simulation Results

Table 1 displays the Monte Carlo relative bias ($100 \times (\theta_* - \theta_0)/\theta_0$, θ_* is the estimate and θ_0 is the true value) for marginal and conditional covariate effects in both the binary and continuous parts of the two-part models as functions of the random intercept variance, $\sigma^2_{u_0}$, and random slope variance, $\sigma^2_{u_1}$ of the true random effect structure in the binary part. The true values of the parameters are set as follows: the true marginal covariate effects in the binary part are $\boldsymbol{\theta}$ = (0.5, log 2, −1, 0.5)$^{\mathrm{T}}$; the true conditional covariate effects in the binary part are $\tilde{\boldsymbol{\theta}}=(0.5, log2-1, 0.5)^{\mathrm{T}}$; the true marginal/conditional covariate effects in the continuous part are $\boldsymbol{\beta}$ = (1, 0.5, −1, 0.5)$^{\mathrm{T}}$; the random intercept variance in the continuous part is σ^2_v=0.2; the error variance in the continuous part is σ^2_e=0.08; the correlation between random intercepts in the two parts is ρ_0 = 0.5; the correlation between random slopes in the binary part and random intercepts in the continuous part is ρ_1 = 0.5.

The top part of Table 1 shows the relative bias of marginal and conditional covariate effects in the binary part. Similar to Heagerty and Kurland25, for both models, the (relative) bias in the estimated conditional interaction term between group and time, $\hat{\theta}_3^C$, in the binary part was found to be as large as 23 – 26% when the random intercept variance component, $\sigma_{u_0}^2$, was small relative to the random slope variance component, $\sigma_{u_1}^2$. Conversely, when $\sigma_{u_0}^2$ was large relative to $\sigma_{u_1}^2$, the bias of $\hat{\theta}_3^C$ reduced to 15 – 20%. The biases for other conditional covariate effects (i.e. the intercept, $\hat{\theta}_0^C$, and the main effects of time, $\hat{\theta}_1^C$, and group, $\hat{\theta}_2^C$) were similar across a range of values for the random intercept and random slope variance components, and were found to be relatively small (less than 6%). The biases for all marginal covariate effects from our two-part marginal model were less than 5% for the range of values chosen for the variance components in the binary part. However, the biases for the corresponding approximate marginal parameter estimates associated with the original two-part mixed model tended to be larger and were observed to be as large as 13% for the marginal group by time interaction effect.

As expected, the (relative) biases of marginal and conditional covariate effects from the continuous part for both our two-part marginal model and the original two-part mixed model (the bottom part of Table 1) were small (less than 3%) and similar because the true random effect structure of the continuous part included a random intercept only and this was correctly specified in the models.

Overall, our simple simulation study shows that incorrectly assuming only a random intercept in a random coefficient model may lead to moderate bias in the estimated conditional covariate effects in the binary part, while under the same situation it has much less impact on marginal covariate effect estimation using the two-part marginal model.

5 Investigation of the association between HLA alleles and HAQ

In this section we use the proposed model to investigate the relationship between the alleles that code for HLA antigens (identified in earlier work as associated with clinical damage) and physical functioning as measured by the HAQ. Recall that our objective is to examine the marginal effects of these alleles on physical functioning in a cohort of psoriatic arthritis patients from the Toronto Psoriatic Arthritis Clinic.

To both parts of our two-part marginal model for HAQ we initially included the main effects of HLA-B27, HLA-DR7, HLA-B39, HLA-DQw3 and HLA-B22, and the interaction of HLA-B27 with HLA-DR7, and the interaction of HLA-DQw3 with HLA-DR7. Additionally, we controlled for age at onset of PsA (standardized), sex and PsA disease duration in years (standardized). After model selection, we arrived at a final two-part marginal model which included in both parts the genetic markers that in either of the two parts had statistically significant main effects or interactions. In this final model, age at onset of PsA, sex and PsA disease duration were also controlled for in both parts. Thus $\mathbf{X}_{ij}$ in (3.1) and $\mathbf{X}_{ij}^*$ in (3.2) coincided. Because residual plots suggested a symmetric error distribution for the continuous part, no transformation was applied to the non-zero HAQ scores11. For

estimation, the SAS NLMIXED procedure was used with the maximum number of points in the adaptive Gaussian quadrature procedure for the quasi-Newton algorithm held at thirty-one (default option in the SAS NLMIXED procedure). A sample SAS program for the final HAQ analysis is provided in the Supplementary Material.

The results for marginal effects of genetic markers are given in Table 2. Note that the conditional estimates associated with the binary part of the underlying two-part mixed model, from which our two-part marginal model is derived, are also shown in this table. These conditional effect estimates are obtained by inflating the corresponding marginal covariate effects in the binary part by the reciprocal of $\varphi = 0.4861$ (95% CI: 0.4256–0.5465). The corresponding standard errors are calculated using the delta method.

From Table 2 we observe that the presence of HLA-B27 significantly increases both the odds of the presence of functional disability ($p = 0.0324$) and the actual level of physical functioning given that one has functional disability ($p = 0.0294$). The (marginal) odds ratio associated with HLA-B27 is 1.605 (95% CI: 1.041–2.476) and the population-averaged difference in the mean (non-zero) HAQ scores between PsA patients with HLA-B27 present compared to PsA patients with HLA-B27 absent, but all else the same, is 0.1652 (95% CI: 0.0166–0.3138). Furthermore, there is statistically significant evidence ($p = 0.0358$) for an interaction effect between HLA-DQw3 and HLA-DR7 on the probability of having functional disability, with an apparent detrimental effect of having HLA-DQw3 present (compared to absent) whilst in the presence of HLA-DR7. There are no statistically significant effects of HLA-DQw3, HLA-DR7 or their interaction on the level of physical functioning once functional disability occurs.

The estimate of ρ is 0.9801 and in the context of this HAQ analysis, ρ can be interpreted as the presence of disability at one occasion being strongly positively related to the level of disability at that and other occasions. Note also that since the estimated correlation between the random intercepts in the underlying two-part mixed model is close to one, this suggests that there might be a single unmeasured latent process which influences the two processes of the HAQ data, corresponding to perfectly correlated random intercepts11. In various analyses of the PsA HAQ data we found that the estimates of the correlation parameter ρ were usually positive and close to one. Since the two-part model described is essentially for a single response process, it is not surprising to observe high correlation between the random effects for the two parts of the longitudinal semicontinuous data. In practice, the estimates of the correlation parameter can be anywhere in the range $(-1, 1)$ as evidenced in other contexts5,6.

6 Conclusion

In this article we have proposed a likelihood-based two-part marginal model for longitudinal semi-continuous data. Building upon the original two-part mixed models of Olsen and Schafer4, we specified the bridge distribution in Wang and Louis13 for the random intercept in the binary part and a Normal distribution for the random intercept in the continuous part, where the two random intercepts were allowed to be correlated. Under this specification, the marginal and conditional expectations in both the binary and continuous parts had the

logistic and linear forms, respectively. Thus this allowed us to obtain the marginal covariates effects directly through the model, with the benefit of preserving the straightforward interpretations of covariate effects in terms of odds ratios and mean differences. Our work here is in a similar spirit to that of Lin *et al.*14 on clustered mixed-type bivariate responses.

Some of the benefits of our two-part marginal model over those presented by Hall and Zhang7 are its easy implementation in standard statistical software packages such as SAS and it being readily extendable to more complicated data structures such as semicontinuous data with additional artificial zeros due to left-censoring5. Moreover, as our two-part marginal model is fully likelihood-based, all the advantages that this brings are present. For example, the ability to construct likelihood ratio tests and deal with unbalanced longitudinal data that result either by design or due to MAR. These advantages are not all available for other two-part marginal models based on GEE methodology.

For the HAQ data used in Section 5, we also fit the original two-part mixed model (with Normal random intercepts in both parts)4–6,11 and the conditional estimates and standard errors obtained are similar to those obtained in Table 2 (results not shown). The estimate of the variance component corresponding to the random intercept in this model is found to be smaller than the estimate obtained for σ_b^2 in the bridge distribution. This is because the bridge distribution is more peaked than the Normal distribution when they have equal variances13. Despite this difference in the variance component estimates between the two models, if the scientific questions of interest were targeted at the subject-specific level then the conclusions arrived at from both models would be the same as long as the random effects and mean structures are correctly specified. However, if the random effect structures are misspecified, for example, if we assume a random intercept only in the binary part of the model when both a random intercept and random slope should be included, then this may lead to bias in the estimated conditional covariate effects in the binary part, while having a lesser impact on the corresponding estimated marginal effects in the binary part. These findings have been verified through the simulation study in Section 4 and are supported by the work of Heagerty and Kurland25 on generalized linear mixed models. Thus in practice when there is some evidence to suggest that a simple random intercept structure for the binary part of the underlying two-part mixed model may be incorrect, if interest is focused on the marginal effects of the covariates in this model, rather than the conditional effects, then there will be minimal impact of this misspecification on estimation and interpretation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by grant MC_US_A030_0022 from Medical Research Council (UK).

References

1. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. Journal of Business and Economic Statistics. 1983; 1:115–126.

2. Zhou XH, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. Biometrics. 1999; 55:645–651. [PubMed: 11318228]
3. Tu W, Zhou XH. A Wald test comparing medical costs based on log-normal distributions with zero valued costs. Statistics in Medicine. 1999; 18:2749–2761. [PubMed: 10521864]
4. Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. Journal of the American Statistical Association. 2001; 96:730–745.
5. Berk KN, Lachenbruch PA. Repeated measures with zeros. Statistical Methods in Medical Research. 2002; 11(4):303–316. [PubMed: 12197298]
6. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. Statistical Methods in Medical Research. 2002; 11(4):341–355. [PubMed: 12197301]
7. Hall DB, Zhang Z. Marginal models for zero inflated clustered data. Statistical Modelling. 2004; 4:161–180.
8. Li N, Elashoff D, Robbins W, Xun L. A hierarchical zero-inflated log-normal model for skewed responses. Statistical Methods in Medical Research. 2008; doi: 10.1177/0962280208097372
9. Liu L, Ma JZ, Johnson BA. A multi-level two-part random effects model, with application to an alcohol-dependence study. Statistics in Medicine. 2008; 27:3528–3539. [PubMed: 18219701]
10. Neelon B, O'Malley AJ, Sharon-Lise TN. A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. Biometrics. 2011; 67:280–289. DOI: 10.1111/j.1541-0420.2010.01439.x [PubMed: 20528856]
11. Su L, Tom BD, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics. 2009; 10:374–389. [PubMed: 19136448]
12. Diggle, P.; Heagerty, P.; Liang, KY.; Zeger, S. Analysis of Longitudinal Data. Oxford University Press; New York: 2002.
13. Wang Z, Louis T. Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution function. Biometrika. 2003; 90:765–775.
14. Lin L, Bandyopadhyay D, Lipsitz SR, Sinha D. Association models for clustered data with binary and continuous responses. Biometrics. 2010; 66:287–293. [PubMed: 19432772]
15. Gladman DD, Shuckett R, Russell ML, Thorne J, Schachter RK. Psoriatic arthritis (PsA) - An analysis of 220 patients. The Quarterly Journal of Medicine. 1987; 62:127–141. [PubMed: 3659255]
16. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: Dimensions and practical applications. Health and Quality of Life Outcomes. 2003; 1:1–20. [PubMed: 12605709]
17. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? Arthritis & Rheumatism. 2007; 56(3):840–849. [PubMed: 17328058]
18. Gladman DD, Farewell VT, Kopciuk M, Cook R. HLA markers and progression in psoriatic arthritis. Journal of Rheumatology. 1998; 25:730–733. [PubMed: 9558177]
19. Heagerty PJ. Marginally specifed logistic-normal models for longitudinal binary data. Biometrics. 1999; 55:688–698. [PubMed: 11314994]
20. Heagerty PJ, Zeger SL. Marginalized multilevel model and likelihood inference (with Discussion). Statistical Science. 2000; 15:1–26.
21. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics. 2002; 58:342–351. [PubMed: 12071407]
22. Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika. 1992; 79:755–762.
23. Molenberghs G, Declerck L, Aerts M. Misspecifing the likelihood for clustered binary data. Computational Statistics and Data Analysis. 1998; 26:327–349.
24. Ten Have TR, Kunselman RA, Tran L. A comparison of mixed effects logistic regression models of binary response data with two nested levels of clustering. Statistics in Medicine. 1999; 18:947–960. [PubMed: 10363333]
25. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika. 2001; 88:973–985.

26. Litiere S, Abad AA, Molenberghs G. Type I and type II error under random-effects misspecification in generalized linear mixed models. Biometrics. 2007; 63:1038–1044. [PubMed: 17425642]
27. Litiere S, Abad AA, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Statistics in Medicine. 2008; 27:3125–3144. [PubMed: 18069726]
28. Abad AA, Litiere S, Molenberghs G. Testing for misspecification in generalized linear mixed models. Biostatistics. 2010; 11(4):771–786. [PubMed: 20407039]
29. Stroud, AH.; Secrest, D. Gaussian Quadrature Formulas. Prentice-Hall; Englewood Cliffs, NJ: 1966.
30. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: A generalized estimating equaton approach. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

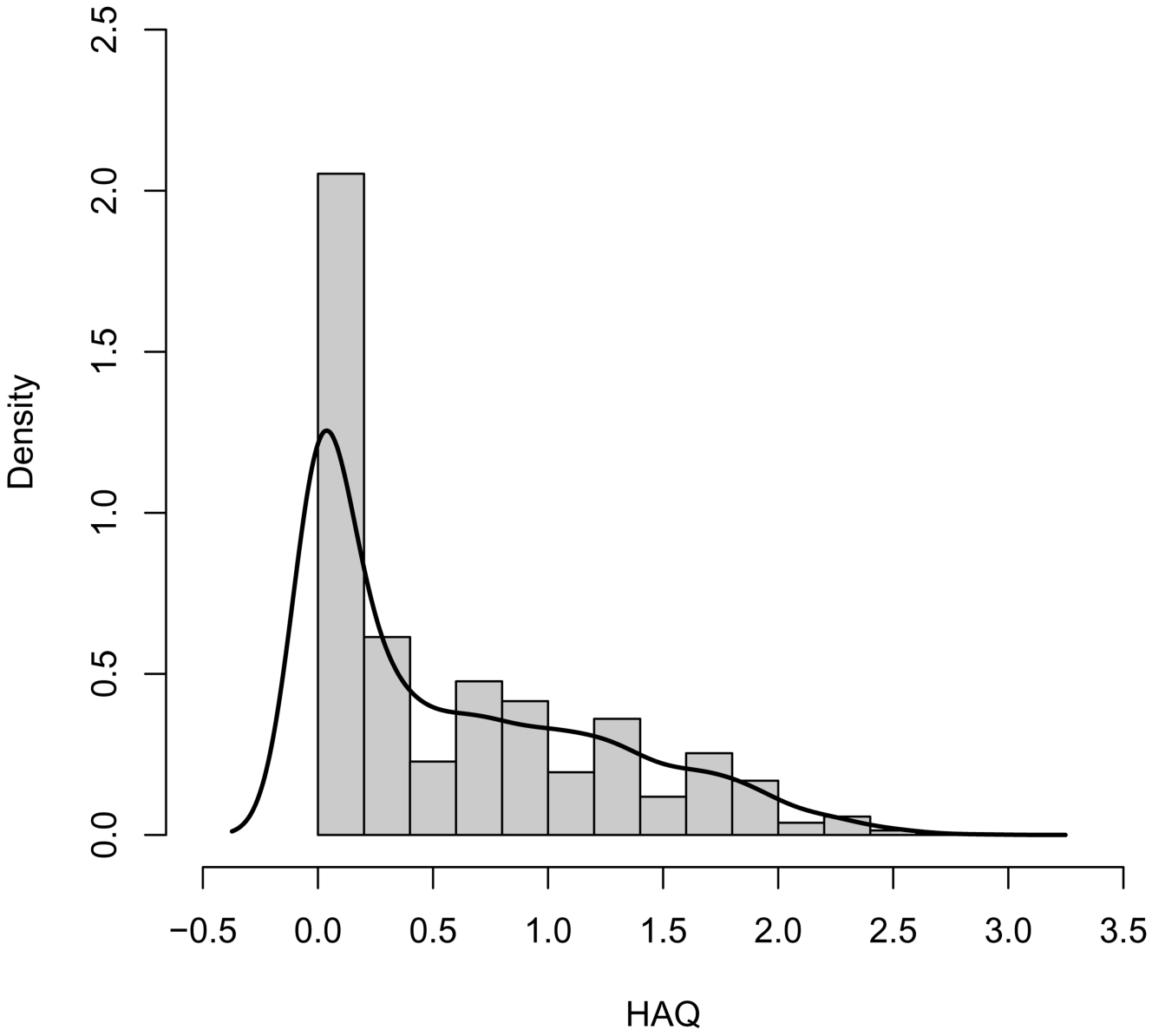

Figure 1: Histogram and kernel density estimates (dark line) for the HAQ data in Section 1

Table 1:

Monte Carlo relative bias, $100 \times (\theta_* - \theta_0)/\theta_0$ (θ_* is the estimate and θ_0 is the true value), for the marginal and conditional covariate effects in the simulation study.

		two-part marginal model analysis				**two-part mixed model analysis**			
		binary part				binary part			
	simulated marginal effect	$\hat{\theta}_0^M$	$\hat{\theta}_1^M$	$\hat{\theta}_2^M$	$\hat{\theta}_3^M$	$\hat{\theta}_0^M$	$\hat{\theta}_1^M$	$\hat{\theta}_2^M$	$\hat{\theta}_3^M$
$\sigma^2_{u_1} = 1$ $\sigma^2_{u_0} = 1$		1.5	−2.8	1.3	1.8	6.4	1.6	5.1	6.6
$\sigma^2_{u_0} = 3$		−0.2	−2.1	1.0	4.8	7.0	3.4	7.4	13.2
	simulated conditional effect	$\hat{\theta}_0^C$	$\hat{\theta}_1^C$	$\hat{\theta}_2^C$	$\hat{\theta}_3^C$	$\hat{\theta}_0^C$	$\hat{\theta}_1^C$	$\hat{\theta}_2^C$	$\hat{\theta}_3^C$
$\sigma^2_{u_1} = 1$ $\sigma^2_{u_0} = 1$		−1.0	0.0	−4.6	−25.8	0.5	0.7	−3.7	−22.7
$\sigma^2_{u_0} = 3$		−3.3	1.6	−5.8	−19.8	−1.0	−0.1	−4.4	−15.4
		continuous part				continuous part			
	simulated marginal effect	$\hat{\boldsymbol{\beta}}_0^M$	$\hat{\boldsymbol{\beta}}_1^M$	$\hat{\boldsymbol{\beta}}_2^M$	$\hat{\boldsymbol{\beta}}_3^M$	$\hat{\boldsymbol{\beta}}_0^M$	$\hat{\boldsymbol{\beta}}_1^M$	$\hat{\boldsymbol{\beta}}_2^M$	$\hat{\boldsymbol{\beta}}_3^M$
$\sigma^2_{u_1} = 1$ $\sigma^2_{u_0} = 1$		−1.5	1.6	−1.8	−2.4	−1.0	1.4	−1.7	−1.9
$\sigma^2_{u_0} = 3$		−1.3	1.5	−1.3	−1.4	−1.0	1.6	−1.3	−1.4
	simulated conditional effect	$\hat{\boldsymbol{\beta}}_0^C$	$\hat{\boldsymbol{\beta}}_1^C$	$\hat{\boldsymbol{\beta}}_2^C$	$\hat{\boldsymbol{\beta}}_3^C$	$\hat{\boldsymbol{\beta}}_0^C$	$\hat{\boldsymbol{\beta}}_1^C$	$\hat{\boldsymbol{\beta}}_2^C$	$\hat{\boldsymbol{\beta}}_3^C$
$\sigma^2_{u_1} = 1$ $\sigma^2_{u_0} = 1$		−1.6	0.3	−1.6	−1.6	−0.9	0.6	−1.4	−1.7
$\sigma^2_{u_0} = 3$		−1.1	0.9	−1.0	−0.6	−0.9	1.0	−1.0	−0.6

Europe PMC Funders Author Manuscripts

Table 2:

Parameter estimates in the binary and continuous parts from the two-part marginal model for the HAQ data: marginal/conditional estimates in the binary part and the continuous part are in the form of log odds ratio and difference in means, respectively.

	Binary part				Continuous part	
	marginal estimate (SE)	***p***	**conditional estimate(SE*)**	***p***	**marginal/conditional estimate(SE)**	***p***
Intercept	0.6245(0.1788)	0.0005	1.2848(0.3665)	0.0005	0.4563(0.0630)	< .0001
HLA-B27	0.4732(0.2203)	0.0324	0.9736(0.4535)	0.0325	0.1652(0.0756)	0.0294
HLA-DQw3	−0.2246(0.2182)	0.3040	−0.4620(0.4465)	0.3015	0.1075(0.0762)	0.1589
HLA-DR7	−0.4757(0.2860)	0.0972	−0.9786(0.5869)	0.0964	−0.0158(0.1023)	0.8775
HLA-DQw3:HLA-DR7	0.8089(0.3839)	0.0358	1.6642(0.7860)	0.0350	0.0256(0.1344)	0.8489
Age at onset of PsA	0.3988(0.0881)	< .0001	0.8206(0.1838)	< .0001	0.1071(0.0289)	0.0002
PsA disease duration	0.1878(0.0694)	0.0072	0.3863(0.1415)	0.0067	0.0488(0.0205)	0.0182
Sex (Female)	1.2184(0.1929)	< .0001	2.5067(0.4093)	< .0001	0.3388(0.0630)	< .0001
σ_b^2	10.6352(1.7621)	< .0001				
φ	0.4861(0.0308)	< .0001				
σ_v^2	0.2851(0.0261)	< .0001				
σ_e^2	0.0907(0.0040)	< .0001				
ρ	0.9801(0.0151)	< .0001				

*Obtained using the delta method