Mixed models for correlated compositional data: applied to microbiome studies in Indonesia

Martin, I.

5

The joint mixture model for the effect of multivariate count on the continuous outcome subject to measurement error

Abstract

In modelling the association between exposure and multiple outcomes from a hierarchical setting, one needs to take into account the correlation structure between these observations. When outcomes are a mixture of continuous and discrete types, modelling becomes complex because joint multivariate distribution cannot be formulated. Specifically, here the outcomes are of a continuous type and multivariate counts with a fixed total. In addition, the multivariate data are overdispersed and measured with errors. For this purpose, we developed a joint regression model in which the multivariate count data are assumed to be multinomially distributed given the random effects. A set of random effects are

This chapter is prepared for a submission as: Ivonne Martin, Renaud Tissier, Jeanine J. Houwing-Duistermaat. The joint mixture model for the effect of multivariate count on the continuous outcome subject to measurement error.

incorporated to account for the measurement errors in the multivariate counts as well as for the correlation between two different types of outcomes and are assumed to follow multivariate normal distribution. The model was also extended to account for a repeated - measurement setting, where additional latent variables are needed. Different covariance structures were explored. The performance of the proposed method was assessed via simulation studies which show that the joint model outperformed the model that ignores the measurement errors (the so-called naive model) in estimating the effect size of the covariate of interest. Data from a repeated measurement study of gut microbiome and cytokine responses carried out in helminth-endemic areas were analyzed.

5.1 Introduction

Biomedical studies often collect multiple outcomes from the same subject to reveal complex underlying biological mechanisms. One of the interests might be to model the association between a specific outcome with regard to the presence or absence of a disease or treatment. A straightforward method is to analyze the association for each outcome separately. However, such an approach might reduce the statistical power since observations from the same subject are potentially highly-correlated. In addition, one might be interested in the association between predictor and both outcomes. Here the randomness of both outcome variables needs to be modelled since ignoring these randomness yields biased estimates. A joint regression model is the approach for this purpose and also increases the statistical power to estimate effects of covariates on outcomes by incorporating the correlation between observations from the same subject via random effects. This approach is however challenging when the observations are from different types, for instance a mixture between continuous and discrete outcomes. The reason is that a multivariate distribution of these outcomes cannot be formulated [McCulloch (2008); Geys et al. (2008)]. In addition, biomedical studies have often a cluster or a longitudinal design which induces a correlation between observations from the same unit.

Our study is motivated by the repeated measurements of gut microbial community and whole blood cytokine responses on subjects in helminth-endemic area in Indonesia. The gut microbiome compositions are obtained from sequencing of 16S rRNA gene. The processsed data consists of counts of taxonomical data with a unit constraint for all taxonomical abundance with additional heterogeneity in the data due to measurement error or variability in sampling or individual. The observation on whole blood cytokine response are continuous data representing the response of this cytokine to certain antigen. Separate studies have shown that the interaction between treatment and helminth infection alter the microbiome composition (**Chapter 4**) as well as the whole blood cytokine re-

sponses [Wammes et al. (2016)]. A straighforward method was used to model the cytokine responses as an outcome with infection, treatment and microbiome composition expressed as a relative abundance for each bacteria taxa as covariate. It was shown that the proportion of *Bacteroidetes* has a significant association with the interleukin-10 (IL-10) response to lipopolysaccharide (LPS) in an uninfected subjects and when the subjects were helminth infected, the association between *Bacteroidetes* and IL-10 response to LPS are significantly different. This result suggests a role of helminth in changing the association between microbiome composition and cytokine responses, however, the model assumes that the microbiome composition are fixed and hence does not account for the randomness due to measurement error. Microbiome data obtained through sequencing of 16S rRNA gene is observed with errors [Schloss et al. (2011)], adding an extra variation in the resulting data [Rosenthal et al. (2014)]. Furthermore, the joint effect of infection status on both outcomes cannot be assessed in this simple model. Thus, our objectives in this paper are to characterize the association between covariates of interest and two outcomes and to quantify the correlation between these two outcomes.

Several works on development of statistical model in the joint model between continuous and discrete type outcomes in the biomedical research have been published, namely between continous and count data [Kassahun et al. (2013); Yang and Kang (2010)], between continuous and time to event (reviewed in Neuhaus et al. (2009)), and continuous type with binary data [Iddi and Molenberghs (2012); Catalano and Ryan (1992); Catalano et al. (1993)] but less on multinomial type data. Here, we are dealing with the mixture of continuous and multivariate discrete outcome with a constraint that the total count is fixed. Review on formulating the joint model is discussed in Verbeke et al. (2014). Typically, when the objective is on modelling the association between covariates and multiple predictors and quantification of the correlation between outcomes, shared random effect is used to account for the correlation between multiple outcomes from the same subject [Geys et al. (2008)]. When dataset has a complex correlation structure as in our study, the model needs to be extended. In our motivating data, three types of correlation structures need to be accounted for, namely the correlation between multiple categories at the same time, the correlation between multiple observation at each type of outcome over time and the correlation between two types of outcome. First of all, we consider the mixed model for each outcomes separately and for each type of outcome, a random effect for outcome-specific is introduced. Several distributions for a random effect to model the overdispersion in the multinomial data has been discussed in literature [Li (2015)]. Here, we proposed to use a normally distributed random effect to allow for a more flexible covariance structure. Secondly, as two outcomes were observed from the same subjects, we incorporated a random shared effect to account for the correlation between two types of outcomes. Estimation and inference were done using the

maximum likelihood approach [Gueorguieva (2016)]. The marginal model was obtained by integrating over the random effect distribution using Gauss-Hermite quadrature.

The rest of the manuscript is organized as follows. In Section 5.2, we described the proposed joint method in modelling the association of binary covariate with mixture types of outcomes. We carried out the investigation of the performance of the proposed method in comparison with the naive method in Section 5.3. The proposed method is then applied to the motivating dataset in Section 5.4 and we conclude and discuss the proposed method in Section 5.5.

5.2 Statistical methods

Suppose for subject $i, i = 1,\ldots,N$, two types of outcomes were collected at time points $t, t = 1,\ldots,N$, namely a continuous type of outcome $Y_i^{(t)}$, and a J dimensional vector of multivariate counts $\boldsymbol{C}_i^{(t)} = \left\{C_{i1}^{(t)},\ldots,C_{iJ}^{(t)}\right\}$, with a fixed total count $C_{i+}^{(t)}$. In addition, let $\boldsymbol{X}_i^{(t)}$ be the covariate values for subject i at time point t. Our aim is to model the relationship between these two outcomes while taking into account the effects of covariates on the outcomes and the presence of measurement error in the multivariate counts. We start with the cross-sectional setting and then extend the model to the longitudinal setting. Note that the superscript t in the cross-sectional setting will be eliminated.

In the cross-sectional setting, a simple linear regression model can be used to assess the relationship between the continuous outcome Y_i and the variable $\frac{C_{ij}}{C_{i+}} = \pi_{ij}$, i.e. the proportion of counts in category j while adjusting for the covariate $\boldsymbol{X}$. Specifically,

$$Y_i = \boldsymbol{X}_i\boldsymbol{\xi} + \gamma_j\pi_{ij} + \varepsilon_i. \tag{5.1}$$

Note that interaction terms between the covariates $\boldsymbol{X}_i$ and the proportion π_{ij} can also be included. This model however ignores that the multivariate count data are subject to measurement error. Further, it is also often of interest to estimate the effect of the covariate $\boldsymbol{X}_i$ on both outcomes. A joint model for the continuous outcome and for the multivariate count outcome addresses these two issues while potentially increasing the power to detect association between $\boldsymbol{X}$ and the two outcomes. The correlation between these two outcomes can be modelled by random shared effects. We first describe the regression model for the multivariate count data and then describe the joint model.

5.2.1 The multinomial logistics mixed model

Let the random effect $\boldsymbol{u}_i^C$ represents the measurement error which is present in the count data. Following the generalized linear framework, the multivariate count outcome conditioned on $\boldsymbol{u}_i^C$ is assumed to follow a multinomial distribution with parameter $\boldsymbol{\pi}_i = \{\pi_{i1}, \ldots, \pi_{iJ}\}$ [Hartzel et al. (2016); Hedeker (2003)]. One could specify the random effect $\boldsymbol{u}_i^C$ to follow the conjugate distribution as introduced by Chen and Li (2013). Although this approach yields a closed form formula for the marginal distribution, the correlation structure between the categories is modelled by only one parameter. In order to make the model more flexible, we assumed that the vector $\boldsymbol{u}_i$ follows a multivariate normal distribution. Note that the measurement error for counts in different categories observed for the same person might be correlated. Let ρ be the correlation between u_{ij}^C and u_{ik}^C. The corresponding regression model is defined as follows.

$$\text{logit}\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \boldsymbol{X}_i \boldsymbol{\xi}^C + u_{ij}^C, \quad j = 2, \ldots, J. \tag{5.2}$$

with the first category as a reference. Here, $\boldsymbol{u}_i^C = \left\{u_{i2}^C, \ldots, u_{iJ}^C\right\}$ are the random effects for each logit, which follow a multivariate normal distribution with zero mean and a symmetric covariance matrix Σ^C which is defined as follows.

$$\Sigma = \begin{pmatrix} \sigma_{u_{C_2}}^2 & \cdot & \cdot & \cdot \\ \rho\,\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma_{u_{C_3}}^2 & \cdot & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ \rho\,\sigma_{u_{C_2}}\sigma_{u_{C_J}} & \rho\,\sigma_{u_{C_3}}\sigma_{u_{C_J}} & \cdots & \sigma_{u_{C_J}}^2 \end{pmatrix}.$$

The marginal distribution for $\boldsymbol{C}_i$ is

$$\begin{aligned} \Pr\left(\boldsymbol{C}_i = \{C_{i1}, \ldots, C_{iJ}\}\right) &= \int \Pr\left(C_{i1}, \ldots, C_{iJ} | \boldsymbol{U}_i^C\right) \Pr\left(\boldsymbol{U}_i^C\right) d\boldsymbol{U}_i^C \\ &= \int C_{i+}! \prod_{j=1}^{J} \left(\frac{1}{C_{ij}!}\right) (\pi_{ij})^{C_{ij}} \Pr\left(\boldsymbol{U}_i^C\right) d\boldsymbol{U}_i^C \end{aligned} \tag{5.3}$$

In our data example, since we assume only three bacterial categories, we have the following formulation:

$$\begin{aligned} \log\left(\frac{\pi_{i2}}{\pi_{i1}}\right) &= \boldsymbol{X}_i \boldsymbol{\xi}_2^C + u_{i2}^C \\ \log\left(\frac{\pi_{i3}}{\pi_{i1}}\right) &= \boldsymbol{X}_i \boldsymbol{\xi}_3^C + u_{i3}^C \end{aligned}$$

and the random effect $\boldsymbol{u}_i^C = \{u_{i2}^C, u_{i3}^C\} \sim \text{MVN}(\Sigma)$ where

$$\begin{pmatrix} \sigma^2_{u_{C_2}} & \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} \\ \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma^2_{u_{C_3}} \end{pmatrix}$$

5.2.2 The joint model in the cross-sectional setting

To model the association between the two types of outcomes in the cross-sectional setting, we introduce a vector of normally distributed shared random effects $\boldsymbol{u}_S$. These random effects represent all unobserved factors having an effect on both outcomes. Note that for the count data, the overdispersion feature may include a measurement error which is modelled by the random effects $\boldsymbol{u}_i^C = \{u_{i2}^C, \ldots, u_{iJ}^C\}$. Now the joint model for both outcomes in the cross-sectional setting is as follows.

$$\begin{aligned} \log\left(\frac{\pi_2}{\pi_1}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_2^C + u_{i2}^C + u_{i2}^S \\ \log\left(\frac{\pi_3}{\pi_1}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_3^C + u_{i3}^C + u_{i3}^S \\ Y_i &= \boldsymbol{X}_i\boldsymbol{\xi}^{(Y)} + u_{i2}^S + u_{i3}^S + \varepsilon_i. \end{aligned} \tag{5.4}$$

We define $\boldsymbol{u}_i^* = \boldsymbol{u}_i^C + \boldsymbol{u}_i^S$. Therefore, $\boldsymbol{u}_i^*$ follows the multivariate normal distribution

$$\begin{aligned} \boldsymbol{u}_i^* &= \begin{pmatrix} u_{i2}^C + u_{i2}^S \\ u_{i3}^C + u_{i3}^S \\ u_{i2}^S + u_{i3}^S + \varepsilon_1 \end{pmatrix} \sim \text{MVN}(\boldsymbol{0}_3, \Sigma), \\ \Sigma &= \begin{pmatrix} \sigma^2_{u_{C_2}} + \sigma^2_{u_{S_2}} & \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma^2_{u_{S_2}} \\ \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma^2_{u_{C_3}} + \sigma^2_{u_{S_3}} & \sigma^2_{u_{s_3}} \\ \sigma^2_{u_{S_2}} & \sigma^2_{u_{S_3}} & \sigma^2_{u_{s_2}} + \sigma^2_{u_{s_3}} + \sigma^2_{\varepsilon_1} \end{pmatrix}. \end{aligned} \tag{5.5}$$

As there might be not sufficient information to estimate all parameters, we could assume that the variance for both shared effects are the same, i.e. $\sigma^2_{u_{S_2}} = \sigma^2_{u_{S_3}}$ or that also the shared random effect themselves are equal, i.e. for both logits we have u_i^S. This latter model can be formulated as follows

$$\begin{aligned} \log\left(\frac{\pi_2}{\pi_1}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_2^C + u_{i2}^C + u_i^S \\ \log\left(\frac{\pi_3}{\pi_1}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_3^C + u_{i3}^C + u_i^S \\ Y_i &= \boldsymbol{X}_i\boldsymbol{\xi}^Y + u_i^S + \varepsilon_i \end{aligned} \tag{5.6}$$

and the covariance structure for the random effect Σ_2:

$$\Sigma = \begin{pmatrix} \sigma^2_{u_{C_2}} + \sigma^2_{u_S} & \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} + \sigma^2_{u_S} & \sigma^2_{u_S} \\ \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} + \sigma^2_{u_S} & \sigma^2_{u_{C_3}} + \sigma^2_{u_S} & \sigma^2_{u_S} \\ \sigma^2_{u_S} & \sigma^2_{u_S} & \sigma^2_{u_S} + \sigma^2_{\varepsilon_1} \end{pmatrix} \tag{5.7}$$

More information about the variances of the random effects is available in a longitudinal study design.

5.2.3 The joint model for mixture of outcomes in a longitudinal setting

In modelling the association between covariates and both outcomes simultaneously in a repeated measurements setting, we need to account for the additional correlation structure in the data. For each type of outcomes, observations from the same subject at different time points will be correlated. A linear mixed effect model with one subject-specific random effect u_Y is used for continuous outcome [Laird and Ware (1982)]. The correlation between two different type of outcomes will be incorporated using the random shared effect $\boldsymbol{U}_i^{(S)}$. Thus, for each subject i we may formulate the following model.

$$\begin{aligned}
\log\left(\frac{\pi_{21}}{\pi_{11}}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_2^C + u_{i2}^C + u_{i2}^S \\
\log\left(\frac{\pi_{31}}{\pi_{11}}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_3^C + u_{i3}^C + u_{i3}^S \\
Y_{i1} &= \boldsymbol{X}_i\boldsymbol{\xi}^Y + u_{i2}^S + u_{i3}^S + u_y + \varepsilon_1 \\
\log\left(\frac{\pi_{22}}{\pi_{12}}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_2^C + u_{i2}^C + u_{i2}^S \\
\log\left(\frac{\pi_{32}}{\pi_{12}}\right) &= \boldsymbol{X}_i\boldsymbol{\xi}_3^C + u_{i3}^C + u_{i3}^S \\
Y_{i2} &= \boldsymbol{X}_i\boldsymbol{\xi}^Y + u_{i2}^S + u_{i3}^S + u_y + \varepsilon_2
\end{aligned} \tag{5.8}$$

Thus, the vector of random effect $\boldsymbol{u}_i^*$ can be defined as follows.

$$\boldsymbol{u}_i^* = \begin{pmatrix} u_{i2}^C + u_{i2}^S \\ u_{i3}^C + u_{i3}^S \\ u_{i2}^S + u_{i3}^S + u_y + e_1 \\ u_{i2}^S + u_{i3}^S + u_y + e_2 \end{pmatrix} \sim \text{MVN}\left(\boldsymbol{0}_4, \Sigma\right),$$

$$\Sigma = \begin{pmatrix} \sigma^2_{u_{C_2}} + \sigma^2_{u_{s_2}} & \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma^2_{u_{s_2}} & \sigma^2_{u_{s_2}} \\ \rho\sigma_{u_{C_2}}\sigma_{u_{C_3}} & \sigma^2_{u_{C_3}} + \sigma^2_{u_{s_3}} & \sigma^2_{u_{s_3}} & \sigma^2_{u_{s_3}} \\ \sigma^2_{u_{s_2}} & \sigma^2_{u_{s_3}} & \sigma^2_{u_{s_2}} + \sigma^2_{u_{s_3}} + u^2_y + \sigma^2_{\varepsilon_1} & \sigma^2_{u_{s_2}} + \sigma^2_{u_{s_3}} + u^2_y \\ \sigma^2_{u_{s_2}} & \sigma^2_{u_{s_3}} & \sigma^2_{u_{s_2}} + \sigma^2_{u_{s_3}} + u^2_y & \sigma^2_{u_{s_2}} + \sigma^2_{u_{s_3}} + u^2_y + \sigma^2_{\varepsilon_2} \end{pmatrix} \tag{5.9}$$

Note that just as in the cross sectional setting we can assume that we have just one shared effect per subject, i.e. $u^S_{i2} = u^S_{i3} = u^S_i$.

The marginal distribution for multiple longitudinal outcomes is now the joint distribution of these outcomes. We assume that conditionally on $\boldsymbol{U}^S_i$, the outcomes Y_i and $\boldsymbol{C}_i$ are independent.

$$\begin{aligned} \Pr(\boldsymbol{C}_i, \boldsymbol{Y}_i) &= \int \Pr\left(\boldsymbol{C}_i, \boldsymbol{Y}_i | \boldsymbol{U}^S_i\right) \Pr\left(\boldsymbol{U}^S_i\right) d\boldsymbol{U}^S_i \\ &= \int \Pr\left(\boldsymbol{C}_i | \boldsymbol{U}^S_i\right) \Pr\left(\boldsymbol{Y}_i | \boldsymbol{U}^S_i\right) \Pr\left(\boldsymbol{U}^S_i\right) d\boldsymbol{U}^S_i \\ &= \int \left[\int \Pr\left(\boldsymbol{C}_i, \boldsymbol{U}^C_i, \boldsymbol{U}^S_i\right) d\boldsymbol{U}^C_i\right] \left[\int \Pr\left(\boldsymbol{Y}_i, \boldsymbol{U}^Y_i, \boldsymbol{U}^S_i\right) d\boldsymbol{U}^Y_i\right] \Pr\left(\boldsymbol{U}^S_i\right) d\boldsymbol{U}^S_i \end{aligned} \tag{5.10}$$

Estimates of all parameters are obtained by maximizing the likelihood of the joint distribution (5.10). Since this likelihood does not have a closed form formula, numerical approximations, such as Gauss-Hermite quadrature need to be utilized.

The variance of the shared effect u_S represents the correlation between two types of outcome. This value is hard to interpret and the marginal correlation between two different types of outcomes might be more interesting. This correlation is given by

$$\text{Corr}(C_{ij}, Y_i) = \frac{\sigma_{C_{ij}, Y_i}}{\sqrt{\sigma^2_{C_{ij}} \sigma^2_{Y_i}}}.$$

The marginal correlation can be computed from Monte-Carlo estimates of the first and second moments.

5.3 Simulation studies

A simulation study was conducted to investigate the performance of the proposed methods. We considered both the cross-sectional and the longitudinal study design. With regard to the random effects structure, we considered models with one univariate shared random effect (equation (5.6)) and models with multivariate random effects in equations (5.4) and (5.8). We considered various

values for the standard deviations of these random effects. Our aims were firstly to investigate the performance of the proposed method in estimating the fixed effects parameters and the variances of the random effects. We also studied the robustness when using the simpler univariate shared effects structure while the multivariate random effect structure is the correct one. Performance was depicted by box plots of the distribution of the parameter estimates across the replicates. Secondly, we compared the performance of our advanced method to the naive method in equation (5.1) in estimating the effects of covariates on the continuous outcomes. Finally, we assessed the efficiency of testing for the presence of a relationship between the multivariate count outcome and the continuous one. This was done by assessing the significance of the shared effect in the joint model by using a likelihood ratio test and of the proportion of bacteria in the naive method by using a t-test.

The integral over the normally distributed random effects was numerically approximated using the Gauss-Hermite quadrature. The simulation study was performed in R statistical software. The SAS software with proc NLMIXED was used for the data application.

5.3.1 Simulation setting

We first generated datasets following the joint model with fixed effect parameters as follows: $\boldsymbol{\xi} = \{\xi_0^Y, \xi_1^Y, \xi_{02}^C, \xi_{12}^C, \xi_{03}^C, \xi_{13}^C\} = \{-2.3, 0.1, -3.5, 0.8, -1.3, -0.15\}$. These parameters represent the intercepts and covariate effects for continuous outcome (ξ_0^Y, ξ_1^Y) and for each category logits $(\xi_{02}^C, \xi_{12}^C, \xi_{03}^C, \xi_{13}^C)$. We also fixed the following random effect standard deviations: $\left\{\sigma_{u_{C_2}}, \sigma_{u_{C_3}}, \sigma_\varepsilon\right\} = \{1, 0.7, 0.1\}$ and the correlation between the random effects for measurement errors $\rho = -0.2$. The values of these parameters are chosen to represent the estimated parameters from the dataset. We considered two sets of standard deviations for the shared random effect, namely $\left\{\sigma_{u_{S_2}}, \sigma_{u_{S_3}}\right\} = \{(0.5, 0.6), (1, 0.8)\}$. For the model with a univariate random effect the standard deviation of the shared effect u_S could take the value 0.5 or 1. Finally we considered $N = 100$ subjects and a total count for the multivariate outcome are the same $C_{i+} = 2000$.

Datasets were generated using the following procedure.

1. Based on the fixed effects parameters and the standard deviations of the random effects, we generated a multivariate normal random effect $\boldsymbol{u}_i^*$ with covariance matrix Σ which is defined in equation (5.5).

2. Using the parameterization of conditional mean given in (5.4), we generated the normally distributed and the multinomial count outcomes for a subject.

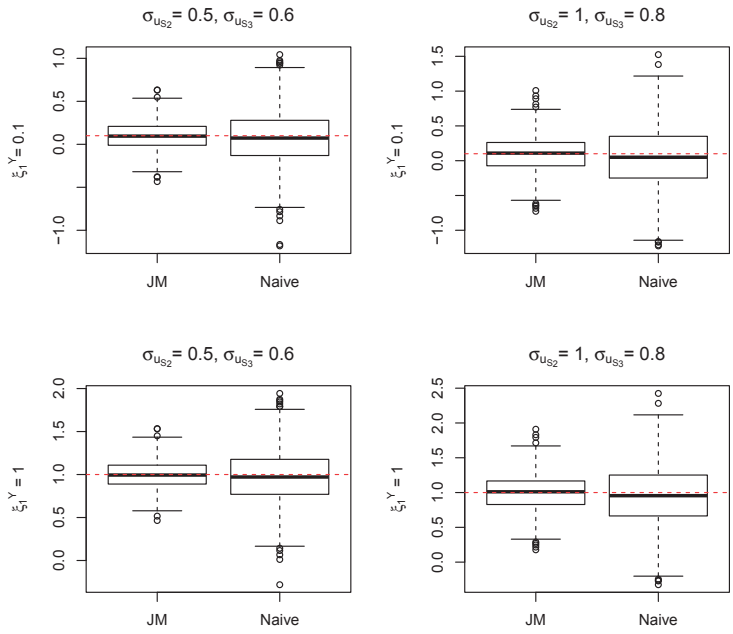

Figure 5.1: Simulation results: the point estimate of covariate of interest from joint model and naive model at the cross-sectional setting. The datasets were generated using a joint model in a cross-sectional setting with logit-dependent random shared effects. The horizontal lines represent the true value.

A similar procedure was used to generate a dataset following a joint model in the longitudinal setting. The used fixed effects parameters are the same as in the case of cross-sectional setting. The standard deviations of the random effects were fixed as follows $\left\{\sigma_{u_{C_2}}, \sigma_{u_{C_3}}, \sigma_{u_Y}, \sigma_\varepsilon\right\} = \{1, 0.8, 0.9, 0.7\}$ and a correlation coefficient between the measurement errors of $\rho = 0.1$ was used. The parameters for the distribution of the shared random effects were the same as in the cross-sectional setting. For each scenarios mentioned above, 1000 replicates were used.

5.3.2 Simulation results

For the cross-sectional model and logit-dependent shared random effects, the results are given in Figure 5.1 and Figure 5.2A. The estimators of all parameters are unbiased. However there are quite some outliers for the estimates of the standard deviations of the random shared effects (Figure 5.2B) especially for small values of standard deviations of the random effects. The same conclusions hold for the

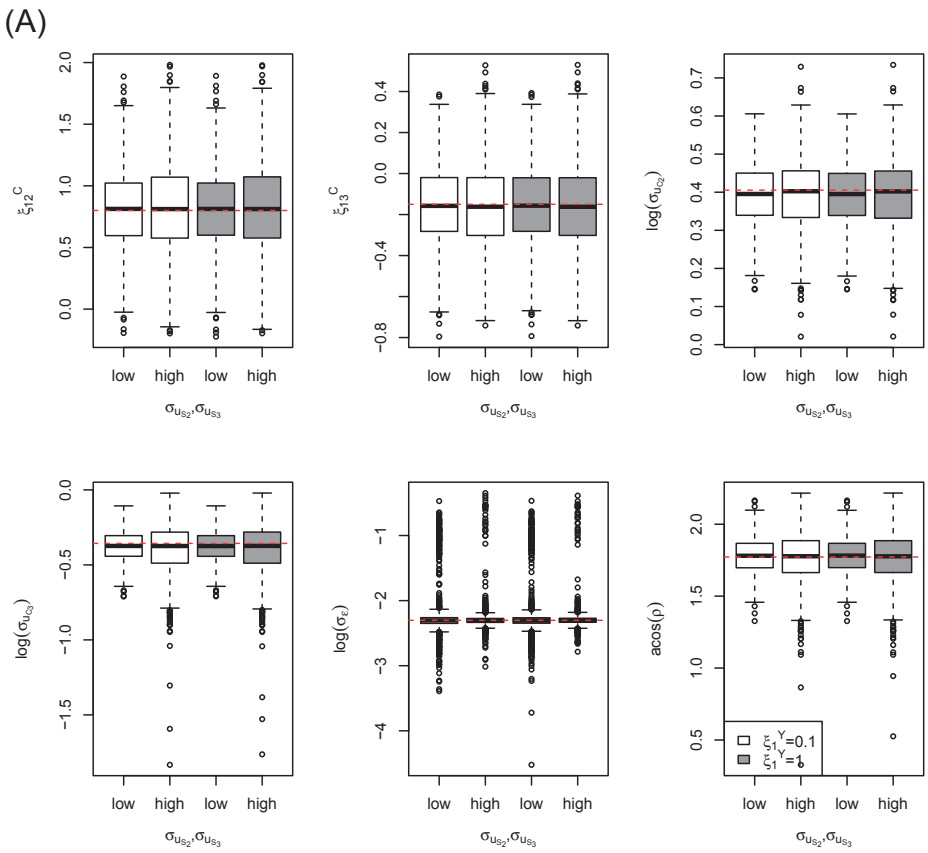

Figure 5.2: Simulation results from a joint model at the cross-sectional setting with logit-dependent random shared effect (A) the point estimates of covariate of interest as well as random effect at different values of shared effect standard deviations, and (B) standard deviations of shared effects. The boxplots in grey represents the distribution when the effect size of covariate of interest is higher ($\xi_1^Y = 1$). The horizontal lines represent the true value.
low represents the combination of $\boldsymbol{\sigma}_S = \{0.5; 0.6\}$.
high represents the combination of $\boldsymbol{\sigma}_S = \{1; 0.8\}$ (first part; continued on next page)

longitudinal design (Figures 5.3 and 5.4). With regards to the joint models with a univariate shared effect (Figure S5.6.1), we noticed that although the obtained distributions for the standard deviations of the random effects do not show outliers, the estimates are biased. The estimators for the fixed effect parameters were unbiased (Figure S5.6.2).

We analyzed the robustness of the parameter estimators for the situation where datasets are generated from the joint model with two-dimensional shared ran-

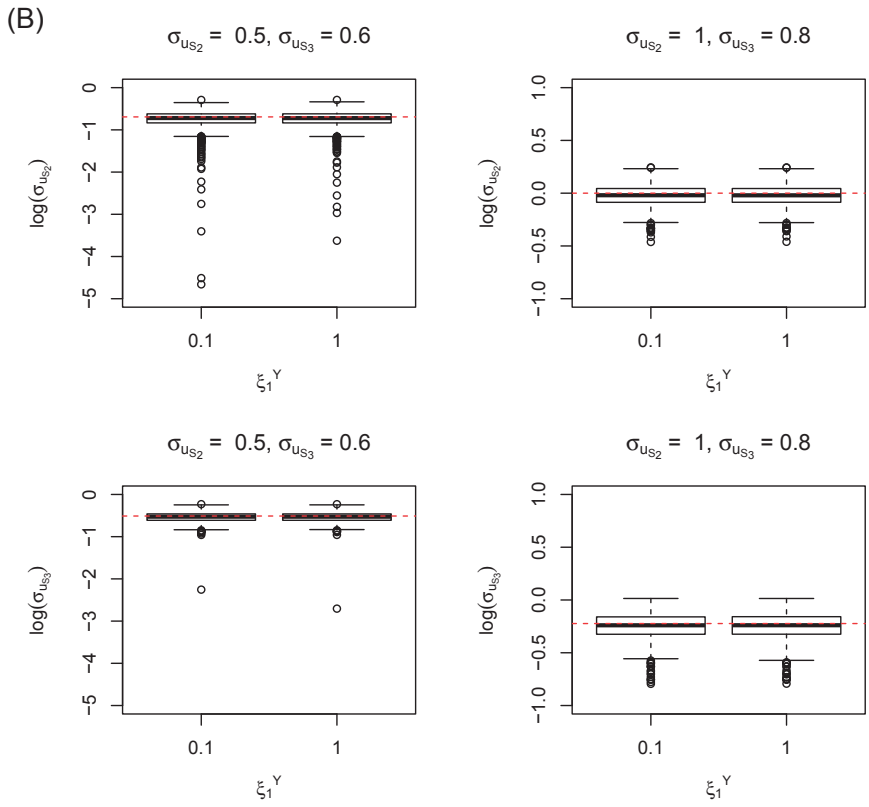

Figure 5.2: (cont.) Simulation results from a joint model at the cross-sectional setting with logit-dependent random shared effect (A) the point estimates of covariate of interest as well as random effect at different values of shared effect standard deviations, and (B) standard deviations of shared effects. The boxplots in grey represents the distribution when the effect size of covariate of interest is higher ($\xi_1^Y = 1$). The horizontal lines represent the true value.
low represents the combination of $\boldsymbol{\sigma}_S = \{0.5; 0.6\}$.
high represents the combination of $\boldsymbol{\sigma}_S = \{1; 0.8\}$

dom effects while a simpler joint model with a univariate random shared effect was used for analysis. While the estimated fixed effects parameters were not affected, the estimated covariance was (Figure 5.5A). In addition, Figure 5.5B illustrates the distribution of the estimated variability of a random shared effect for the situation where the dataset was generated following the joint model with two dimensional random shared effect while a joint model with a univariate random effect was fitted. This showed the effect of uncorrectly reducing the number of parameters in modelling the variability of multiple categories. When the shared effects for both categories had about the same variability ($\sigma_{u_{S_2}} = 0.5, \sigma_{u_{S_3}} = 0.6$),

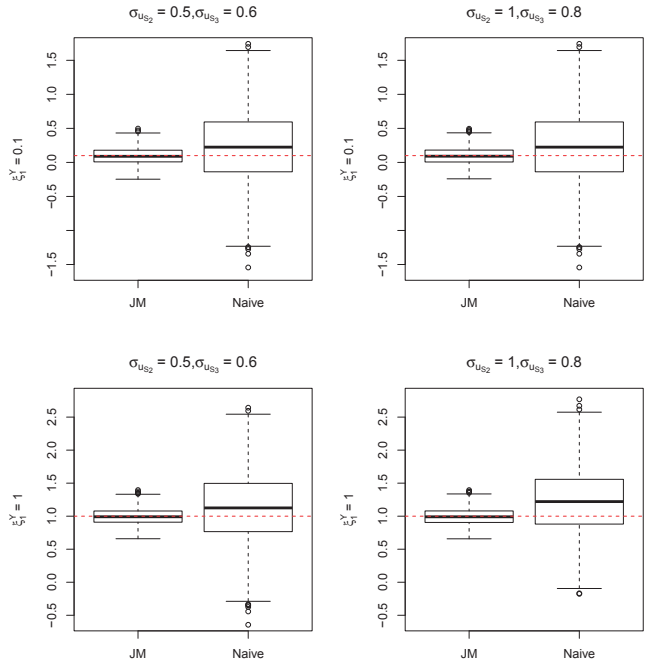

Figure 5.3: Simulation results: The point estimates of the covariate of interest from joint model and naive approach in longitudinal setting.

the estimated standard deviation for the shared effect using a univariate random effect was closer to the true value.

Finally we compared the true marginal correlation of the multivariate outcomes data with the covariance structure corresponding to the joint model with various covariance structure and of the simpler model with univariate shared effects. The covariances corresponding to the models were estimated using the Monte-Carlo method. Table 5.1 gives the estimates of the marginal correlation for the two models. It appears that the absolute correlations between the multivariate outcomes and the continuous were overestimated when using the simpler model, namely for the first category -0.503 instead of -0.475, for the second category 0.161 instead of 0.10 and for the third category 0.426 instead of 0.417 (Table 5.1A). Similar case was also observed in the case of higher standard deviations of random shared effect (Table 5.1B).

For the longitudinal setting and logit-dependent shared random effects, the results are depicted in Figure 5.3. When using the joint model with logit-dependent random shared effect to generate the data, the naive method showed a bias.

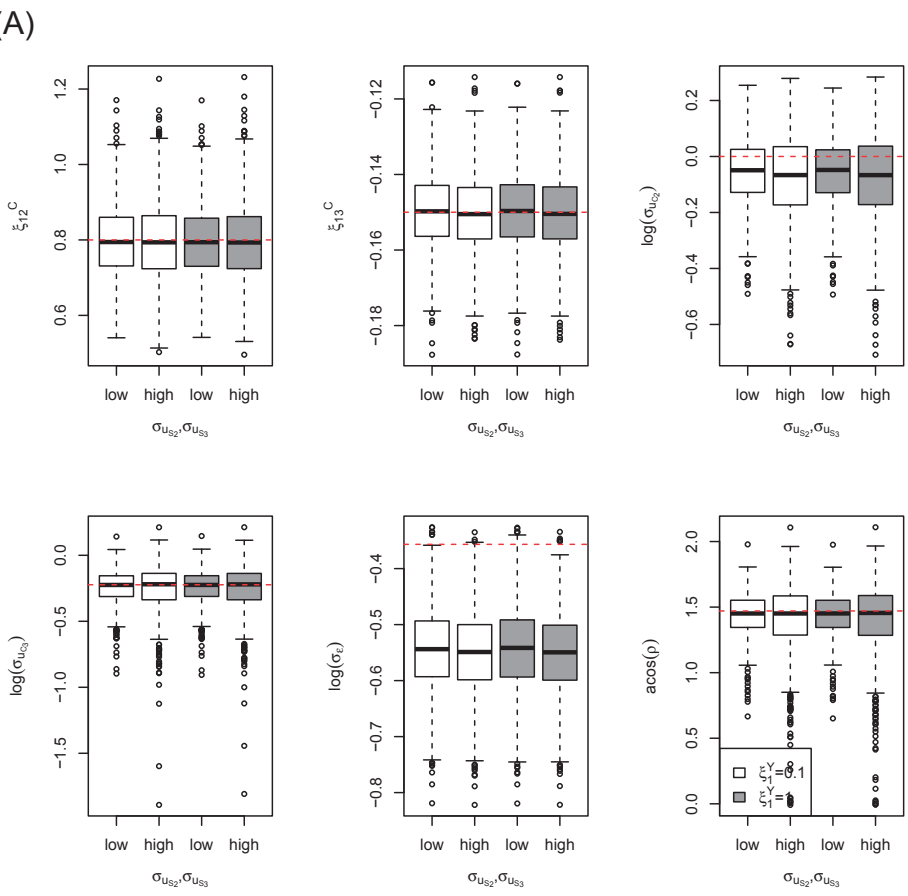

Figure 5.4: Simulation results: the point estimates of (A) categorical covariate effects as well as random effects (excluding the shared effects) for different standard deviations of shared effects, and (B) standard deviations of the shared effect at different effect size from joint model in longitudinal setting with logit-dependent random effect. Details of low and high are similar as Figure 5.2. (first part; continued on next page)

Furthermore, the naive method gave a larger standard deviation of the estimates compared to the true joint model as in the cross-sectional setting. In Figure 5.6 the distributions of the estimated σ_{uY} is given for the joint model and the naive model. It appeared that the estimator based on the naive method was biased.

Finally, we evaluated the power to detect a relationship between the two outcomes by comparing the rejection rate of the null hypothesis of a zero standard deviation of the shared random effect in the joint model with the rejection rate of the null hypothesis of a zero effect of the proportion of categorical outcomes

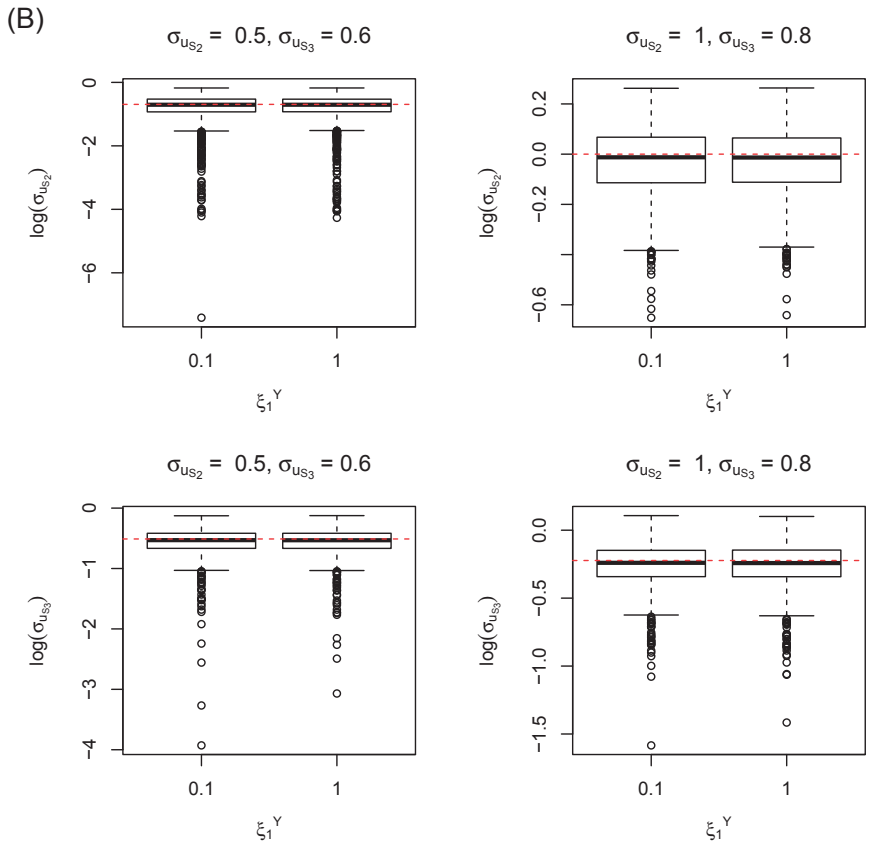

Figure 5.4: (cont.) Simulation results: the point estimates of (A) categorical covariate effects as well as random effects (excluding the shared effects) for different standard deviations of shared effects, and (B) standard deviations of the shared effect at different effect size from joint model in longitudinal setting with logit-dependent random effect. Details of low and high are similar as Figure 5.2.

on the continuous outcome in the naive approach. The results are given in (Table 5.2). It appears that for the cross-sectional setting the joint model only had power when the standard deviation was large and for the univariate shared effects (85%), while the naive methods showed sufficient power for all models. For the longitudinal setting the joint model outperformed the naive method with a power of 86% for small standard deviations of the shared effects compared to 77% and of 100% for large standard deviations of the shared effects compared to 97%.

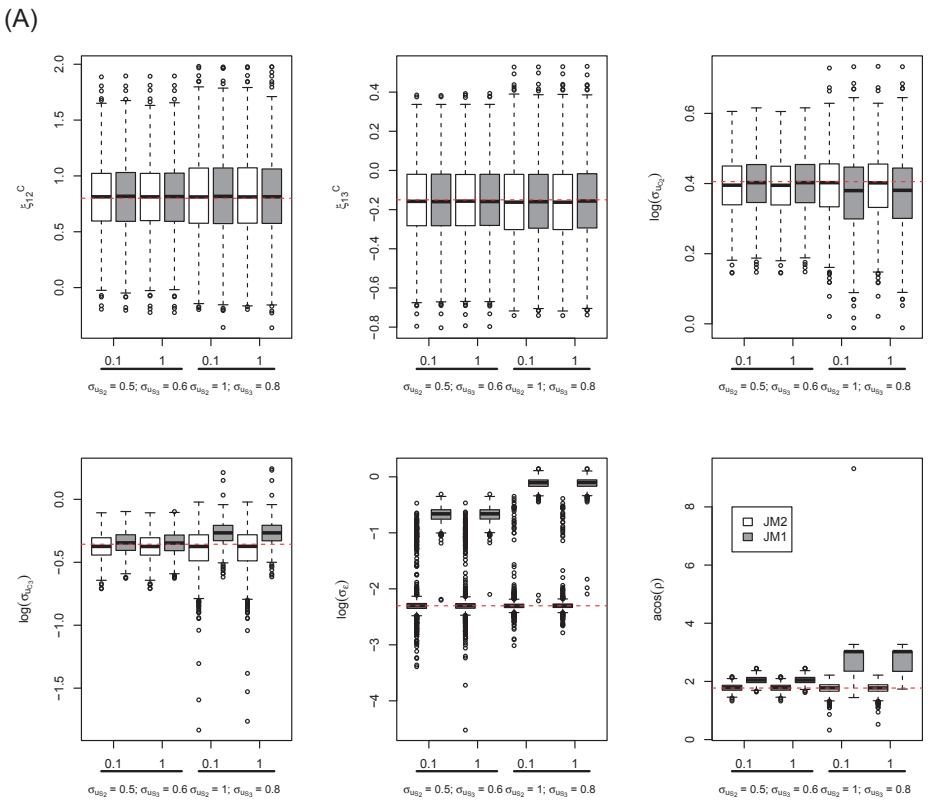

Figure 5.5: The robustness of (A) fixed effect and standard deviation of the random effect parameters and (B) standard deviations of shared effects in joint model in cross-sectional setting. Datasets were generated using the joint model with logit-dependent shared effect. The estimates were obtained from fitting these datasets with joint model logit-dependent shared effect (JM2) and univariate random effect (JM1). The horizontal lines represent the true value. (first part; continue on next page)

5.4 Data analysis

The dataset considered here was measured in a subset of randomized controlled trial in a helminth-endemic area in Indonesia to assess the influence of helminth infection on inflammatory diseases Wiria et al. (2010). Households were randomized for a 400 mg albendazole or placebo for a period of one and half year. Yearly stool samples were collected on a voluntary basis, to detect the presence of helminth infections as well as obtaining genomic material of gut microbial community. Blood samples were drawn for immunological examinations.

Trichuris trichiura infection was detected only by microscopy, while the DNA of hookworms (*Ancylostoma duodenale* and *Necator americanus*) and *Ascaris lum-*

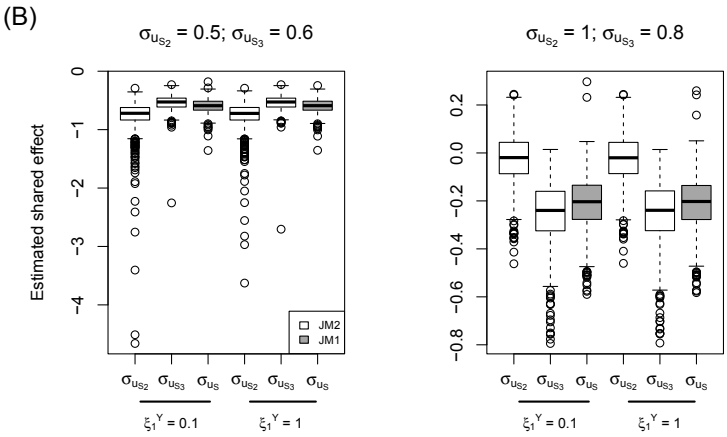

Figure 5.5: (cont.) The robustness of (A) fixed effect and standard deviation of the random effect parameters and (B) standard deviations of shared effects in joint model in cross-sectional setting. Datasets were generated using the joint model with logit-dependent shared effects. The estimates were obtained from fitting these datasets with joint model logit-dependent shared effect (JM2) and univariate random effect (JM1). The horizontal lines represent the true value.

	$\sigma_{u_{S_2}} = 0.5, \sigma_{u_{S_3}} = 0.6$				$\sigma_{u_S} = 0.5$			
	C_1	C_2	C_3	Y	C_1	C_2	C_3	Y
C_1	1	-0.425	-0.712	-0.475	1	-0.498	-0.702	-0.503
C_2	·	1	-0.334	0.1	·	1	-0.267	0.161
C_3	·	·	1	0.417	·	·	1	0.426
Y	·	·	·	1	·	·	·	1
	$\sigma_{u_{S_2}} = 1, \sigma_{u_{S_3}} = 0.8$				$\sigma_{u_S} = 1$			
	C_1	C_2	C_3	Y	C_1	C_2	C_3	Y
C_1	1	-0.463	-0.690	-0.563	1	-0.497	-0.812	-0.776
C_2	·	1	-0.321	0.267	·	1	-0.103	0.298
C_3	·	·	1	0.383	·	·	1	0.688
Y	·	·	·	1	·	·	·	1

Table 5.1: The estimated marginal correlations from the joint model in a cross-sectional setting from different covariance structures.

bricoides were observed via multiplex real-time PCR. A subject was regarded as helminth-infected if it was infected with at least one helminth species. The pyrosequencing process of 16S rRNA gene to obtain the bacterial data has been described in Martin et al. (2018). Here, we focus on two specific phyla, namely *Bacteroidetes* and *Firmicutes* and pooled the remaining phyla into pooled category.

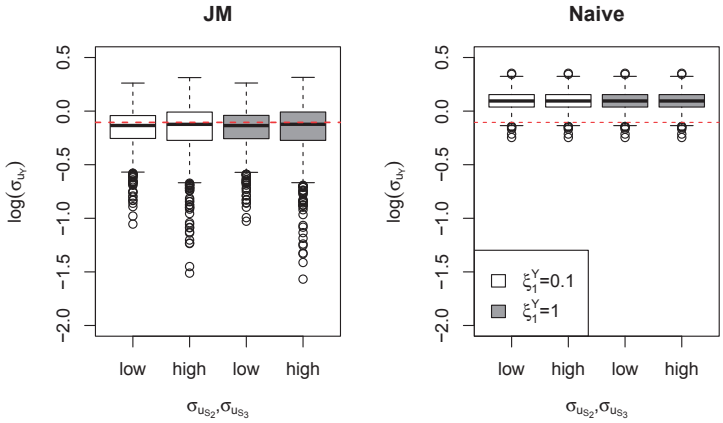

Figure 5.6: The estimates for random effect's variability of continuous outcome from joint model and naive approach in longitudinal setting. The horizontal lines represents the true value. Details about low and high are the same as in Figure 5.2

Shared effect	Cross-sectional			Longitudinal		
	JM	Naïve		JM	Naïve	
		π_2	π_3		π_2	π_3
low	0.2	75.5	99.2	86.3	19	76.6
high	84.7	97.7	100	100	83	96.6

Table 5.2: **Statistical Power.** The rejection rate of shared effect in joint model and proportion of bacteria in naive approach. The computation was done for the fixed effect $\xi_1^Y = 0.1$. The joint model in cross-sectional setting uses the univariate random shared effect and the logit-dependent shared effect for longitudinal case. Low represents the shared effect of $\sigma_S = 0.5$ or $\boldsymbol{\sigma}_S = \{0.5, 0.6\}$ and high represents $\sigma_S = 1$ or $\boldsymbol{\sigma}_S = \{1, 0.8\}$

The blood cultures were stimulated to assess the innate and adaptive immune responses, characterized by cytokine responses. In **Chapter 4**, among all analyzed cytokine responses, only the innate interleukin(IL)-10 response to lipopolysaccharide (LPS) that was significantly associated with *Bacteroidetes* proportion. In this analysis we aim to reanalyze these outcomes simultaneously in relation with helminth-infections. Thus, we focus on the continuous type observation IL-10 response to LPS. Our data consists of 62 subjects who have complete measurements on microbiome composition and cytokine responses at before and 21 months after the first treatment (Table 5.3).

To assess the relationship between the IL-10 response and the microbiome compositions, we first applied the naive approach in a cross-sectional setting by analyzing only the observations at the first time point. Specifically, a linear

Characteristics		albendazole (N = 23)	placebo (N = 39)
Gender, female (n (%))		12 (52.17)	22 (56.41)
Age (mean(SD))		27.03 (15.80)	26.53 (15.86)
Helminth infections (N(%))			
A. lumbricoides		9 (39.13)	8 (20.51)
Hookworm		10 (43.48)	10 (25.64)
	N. americanus	9 (39.13)	10 (25.64)
	A. duodenale	2 (8.69)	2 (5.13)
T. trichiura		5 (21.74)	10 (25.64)
Any helminths			16 (69.57)
23 (58.97)			
Abundance of bacterial phyla, mean % (SD)			
Firmicutes		73.21 (10.76)	71.54 (12.94)
Actinobacteria		9.73 (5.84)	9.40 (7.75)
Bacteroidetes		6.70 (9.97)	7.27 (12.19)
pooled		10.35 (7.29)	11.79 (8.10)
Cytokine responses (median, IQR)			
LPS	IL-10	250 (137.5, 400.5)	221 (137, 381.5)

Table 5.3: The characteristics of participants at pre-treatment.

model with the IL-10 response to LPS as a continuous outcome and bacterial proportion, infection status and their interaction as covariates. The results are given in Table 5.4A. It appears that infection has no significant effect on IL-10 to LPS (estimated effect of 0.202 (s.e. of 1.121), p-value of 0.858). The *Bacteroidetes* proportion showed a trend of association with the IL-10 response to LPS antigen (estimated effect of -1.812 (s.e of 1.024), p-value of 0.082). For subjects who are helminth-infected, this association seems to disappear while for subjects who are helminth-uninfected, the relationship is stronger (**Chapter 4**). When using all data in the longitudinal setting, the estimated parameters are given in Table 5.4B. The association between helminth infections and IL-10 response remains not significant, but the association between *Bacteroidetes* proportion and IL-10 to LPS are significantly different depending on infection status. When subjects were helminth-uninfected, the cytokine responses and *Bacteroidetes* proportion are negatively associated while this association disappears when subjects where helminth-infected. This suggests that microbiome composition is likely to correlate with cytokine response.

Next, we fitted the joint models to these data. These models take into account the measurement error of the microbiome proportions and analyze the joint ef-

	(A) Cross-sectional		(B) Longitudinal			
	Estimate (s.e)	p-values	Estimate (s.e)	p-values	Group name	Variance
(Intercept)	2.588 (0.757)	0.001	2.337 (0.450)	< 0.001	individual	0.022
inf	0.202 (1.121)	0.858	-0.796 (0.626)	0.206	Residuals	0.109
p.Actinobacteria	-0.301 (1.224)	0.806	-0.442 (0.749)	0.556		
p.Bacteroidetes	-1.812 (1.024)	0.082	-2.139 (0.733)	0.004		
p.Firmicutes	-0.284 (0.874)	0.746	0.022 (0.514)	0.967		
inf:p.Actinobacteria	-1.399 (1.928)	0.471	1.306 (1.093)	0.235		
inf:p.Bacteroidetes	1.392 (1.377)	0.316	2.831 (0.902)	0.002		
inf:p.Firmicutes	-0.001 (1.265)	1.000	0.849 (0.713)	0.237		

Table 5.4: Data analysis: The estimates of the fixed effect and random effect parameters from the naive approach for the cross-sectional and the longitudinal setting.

fect of infection on microbiome composition and cytokine response simultaneously. We used model (5.4) with as covariate $\boldsymbol{X}_i$ the infection status and as random effect $\boldsymbol{u}_i^* = \{u_{C_2} + u_{S_2}, u_{C_3} + u_{S_3}, u_{S_2} + u_{S_3} + u_Y\}$ following a multivariate normal distribution with mean of zero and covariance matrix Σ, where Σ is defined in equation (5.5). The estimated parameters of the fixed effects and standard deviations of the random effects (and their corresponding standard error and significance) are given in Table 5.5A. Infection has no significant association with neither microbiome composition nor the cytokine responses. In contrast to the naive approach, we observed that the two outcomes are not correlated, i.e. the estimates of the variances of the random shared effects $\sigma_{u_{S_2}}$ and $\sigma_{u_{S_3}}$ are almost zero ($\sigma_{u_{S_2}}{}^2 = 0.002$,(s.e of 0.010), p-value of 0.796; $\sigma_{u_{S_3}}{}^2 = 0.006$, (s.e. of 0.015), p-value of 0.628).

We further analyzed the dataset with the simplified joint model where the shared random effects in the logits are the same (u_S) as in equation (5.6) and (5.7). Table 5.5B lists the estimated parameters for the fixed effects and variances of the random effects. The estimated parameters for the fixed effect were similar to the joint model with two shared random effects. Again the estimated standard deviation of the univariate random shared effect appears to be small, namely $\sigma_{u_S}{}^2 = 0.002$ (s.e of 0.007). When assessing the marginal correlation between multivariate counts and continuous outcome, we observed that the marginal correlation based on the fitted joint models do not fit the data properly (Table 5.6). The second bacteria category is negatively correlated with the continuous outcome ($\text{cor}(C_1, Y_1)$ = -0.089, Table 5.6A), while the estimated correlation using the joint model with univariate and logit dependent random effect is positive (Table 5.6B and C).

Next, we investigated the correlation between two outcomes when subjects

(A) The Joint Model with logit dependent shared effects

Fixed Effects	**Estimate** (95% CI)	*p*-value	**Random Effects**	Estimate (s.e)	*p*-value
Intercepts					
ξ_0^Y	2.25 (2.11, 2.39)	<.0001	$\sigma_{u_{C_2}}^2$	2.372 (0.427)	<.0001
ξ_{02}^C	-3.71 (-4.33, -3.09)	<.0001	$\sigma_{u_{C_3}}^2$	0.463 (0.084)	<.0001
ξ_{03}^C	-1.32 (-1.59, -1.04)	<.0001	$\sigma_{u_{S_2}}^2$	0.002 (0.010)	0.796
Infection			$\sigma_{u_{S_3}}^2$	0.006 (0.015)	0.628
ξ_1^Y	0.15 (-0.03, 0.32)	0.103	σ_ε^2	0.110 (0.026)	0.000
ξ_{12}^C	0.61 (-0.18, 1.41)	0.129	ρ	-0.271 (0.118)	0.027
ξ_{13}^C	-0.11 (-0.47, 0.24)	0.513			

(B) The Joint model with univariate shared effect.

Fixed Effects	Estimate (95% CI)	*p*-value	**Random Effects**	Estimate (s.e)	*p*-value
Intercepts					
ξ_0^Y	2.25 (2.11, 2.39)	<.0001	$\sigma_{u_{C_2}}^2$	2.371 (0.427)	<.0001
ξ_{02}^C	-3.70 (-4.32, -3.08)	<.0001	$\sigma_{u_{C_3}}^2$	0.466 (0.083)	<.0001
ξ_{03}^C	-1.32 (-1.59, -1.04)	<.0001	$\sigma_{u_S}^2$	0.002 (0.007)	0.796
Infection					
ξ_1^Y	0.15 (-0.03, 0.32)	0.103	σ_ε^2	0.117 (0.022)	0.000
ξ_{12}^C	0.61 (-0.18, 1.41)	0.129	ρ	-0.276 (0.117)	0.027
ξ_{13}^C	-0.11 (-0.47, 0.24)	0.513			

Table 5.5: Data analysis: The parameter estimates from the joint model with two-dimensional random shared effects (A) and a univariate random effect (B) in the cross-sectional setting. Fitted with SAS PROC NLMIXED with 10 quadratures.

were helminth-uninfected. For this purpose, we selected helminth-uninfected subjects at pre-treatment (N = 23) and fitted the joint model with only intercepts and random shared effects. We used the model with two shared random effects with the assumption that both shared effects have the same variance ($\sigma_{u_{S_2}}^2 = \sigma_{u_{S_3}}^2 = \sigma_{u_S}^2$). The results are given in Table 5.7. The variance of this shared effect is again very small ($\sigma_{u_S}^2 = 0.003$, (s.e. of 0.012)) suggesting that there was not enough evidence to conclude that both outcomes were correlated even in subjects who were helminth-uninfected.

The observed marginal correlation of the 23 helminth-uninfected subjects are given in Table S5.6.1 as well as the estimated marginal correlation obtained from the joint model. It appears that the correlation between the second category and the continuous outcome of the model disagrees with the observed marginal correlations, i.e. for the first category -0.092 for the model and 0.056 observed and for the second category 0.018 for the model and -0.355 observed.

(A) The observed marginal correlation				
	C_1	C_2	C_3	Y_1
C_1	1.000	-0.545	-0.530	0.075
C_2	-0.545	1.000	-0.422	-0.089
C_3	-0.530	-0.422	1.000	0.009
Y_1	0.075	-0.089	0.009	1.000
(B) The Joint Model with logit dependent shared effect				
C_1	1.000	-0.515	-0.589	-0.032
C_2	-0.515	1.000	-0.389	0.028
C_3	-0.589	-0.389	1.000	0.008
Y_1	-0.032	0.028	0.008	1.000
(C) The Joint Model with univariate shared effect				
C_1	1.000	-0.518	-0.586	-0.018
C_2	-0.518	1.000	-0.389	0.031
C_3	-0.586	-0.389	1.000	-0.010
Y_1	-0.018	0.031	-0.010	1.000

Table 5.6: The marginal correlation between multivariate count and continuous outcome. Observed and based on the joint models in the cross-sectional setting.

Parameters	Estimate (95% CI)	*p*-value
$\xi_0{}^Y$	2.26 (2.08, 2.44)	<.0001
$\xi_{02}{}^C$	-3.77(-4.42, -3.12)	<.0001
$\xi_{03}{}^C$	-1.30 (-1.65,-0.95)	<.0001
Random Effect	Estimate (s.e)	*p*-value
$\sigma_{u_{C_2}}{}^2$	2.173 (0.666)	0.004
$\sigma_{u_{C_3}}{}^2$	0.638 (0.191)	0.003
$\sigma_{u_S}{}^2$	0.003 (0.012)	0.819
$\sigma_\varepsilon{}^2$	0.166(0.055)	0.006
ρ	-0.392 (0.180)	0.042

Table 5.7: The estimated parameters using joint model in selected helminth-uninfected subjects at pre-treatment (N = 23)

When analyzing the joint model in the longitudinal setting with logit-dependent random effect, we noticed that infection status was significantly associated with the increasing odds of *Bacteroidetes* to *Firmicutes* ($\xi_{12}^{C} = 0.79$, (s.e. of 0.03) ,Table 5.9) although the estimated variances of shared effect between discrete and continuous outcomes remained small. We notice however, that the magnitude of the correlation is slightly increased in the longitudinal setting. To investigate the estimated variance of shared effect in subjects who remained uninfected, we selected 16 subjects who were helminth-uninfected at pre-treatment and remained uninfected at 21 months after the first treatment. The estimated parameters are listed in Table S5.6.2. We observed that the estimated variance of the shared effect was getting larger in the subjects who were uninfected and measured longitudinally.

(A) The observed marginal correlation				
	C_1	C_2	C_3	Y_1
C_1	1.000	-0.264	-0.741	0.056
C_2	-0.264	1.000	-0.451	-0.355
C_3	-0.741	-0.451	1.000	0.195
Y_1	0.056	-0.355	0.195	1.000
(B) The Joint model with logit dependent shared effect.				
C_1	1.000	-0.175	-0.849	-0.092
C_2	-0.175	1.000	-0.371	0.018
C_3	-0.849	-0.371	1.000	0.077
Y_1	-0.092	0.018	0.077	1.000

Table 5.8: Data analysis: The observed and the estimated marginal correlation from joint model in the cross-sectional setting. The joint model was fitted on datasets consists of only helminth-uninfected subjects at pre-treatment (N =23).

Finally, we fitted a joint model for the cytokines and only two bacterial categories, namely the Bacteroidetes and pooled category consisted of the remaining taxa. The estimated covariate effects as well as the standard deviation of the random effect were given in Table S5.6.3. Again, we observed that there is no correlation between the two outcomes.

The estimates of the parameters of interest from the joint model in the longitudinal setting are listed in Table 5.9. It is shown that helminth infection is only associated with the microbiome composition and not the cytokine response.

5.5 Discussion

We proposed a joint model to analyze simultaneously the effect of a specific covariate on multiple outcomes collected from the same subject and to model the

Fixed effects	Estimate (95% CI)	*p*-values	**Random effects**	Estimate (s.e)	*p*-values
Intercepts					
ξ_0^Y	2.19 (2.08,2.30)	<.0001	${\sigma_{u_{C_2}}}^2$	1.877 (0.348)	<.0001
ξ_{02}^C	-3.46 (-3.81, -3.11)	<.0001	${\sigma_{u_{C_3}}}^2$	0.308 (0.059)	<.0001
ξ_{03}^C	-0.96 (-1.10, -0.82)	<.0001	${\sigma_{u_{S_2}}}^2$	-0.016 (0.050)	0.754
Infection			${\sigma_{u_{S_3}}}^2$	-0.0002 (0.021)	0.99
ξ_1^Y	0.09 (-0.05, 0.23)	0.209	${\sigma_{u_Y}}^2$	0.035 (0.059)	0.559
ξ_{12}^C	0.79 (0.73, 0.86)	<.0001	${\sigma_\varepsilon}^2$	0.128 (0.023)	<.0001
ξ_{13}^C	-0.33 (-0.37, -0.30)	<.0001	ρ	0.074 (0.127)	0.562

Table 5.9: Data analysis: the estimated parameters of joint model in the longitudinal setting. Fitted with SAS with 10 quadrature points.

relationship between the outcomes. Specifically our work was motivated by data on the association between helminth infection status as covariate and microbiome composition and cytokine responses as outcomes while taking into account the correlation structure in the data as well as the presence of measurement errors in the microbiome data. We used a linear mixed effect model for the continuous outcome and a multinomial logistics mixed model approach introduced by Hartzel et al. (2016) for the microbiome data. To model extra variation due to measurement error or unobserved heterogeneity in the multinomial type data, a conjugate or normally distributed random effect can be used. However, there has been a discussion with regard to the choice of the random effect distribution in multinomial type data. While the conjugate random effect has an advantage of having a closed form formula for the marginal distribution, the correlation between categories is described with a single parameter representing overdispersion Li (2015). On the other hand, the multinomial logistics mixed model with normally distributed logit-dependent random effect provides more flexibilities in modelling measurement error present in microbiome data. To model the correlation between multiple outcomes from the same subject, different covariance structure for the random shared effect were considered, namely random shared effect for each categorical logit and the continuous outcome and a single random shared effect for each categorical logit and the continuous outcome.

We compared our model with a naive approach which includes bacterial proportions as a covariate in a linear model ignoring the measurement error in the microbiome data. Our simulation study in the cross-sectional setting showed that the joint model with either with logit-dependent or univariate random shared effect gives the unbiased estimate of the parameter modelling the effect of covariates on the continuous outcomes as well as smaller standard deviation compared to the estimate obtained using the naive model. Overall, the fixed effect parameters and the variability of the random effect were better estimated in the model

with logit-dependent random shared effect.

In the longitudinal setting, we noticed that the estimator of the parameter modelling the effect of the covariate on the continuous outcome in the naive approach was biased in all cases of simulation setting. This was probably caused by the additional correlation structure in the repeated measurement of the cytokine responses. Finally when testing for the presence of a relation between the outcomes, the joint model had more power than the naive approach in the longitudinal setting. However this was not the case for the cross-sectional setting, probably due to lack of information to estimate all the variance components in this design. Overall the joint model is preferred over the naive method in the longitudinal setting.

In our data application of the proposed joint model in the cross-sectional setting, helminth infection was not significantly associated with both cytokine response and microbiome composition. In the absence of helminth infection, the estimated average value of cytokine response was positive, while there was a decreasing ratio of *Bacteroidetes* to *Firmicutes* and pooled category to *Firmicutes*, indicating there was an inverse relationship between cytokine response and gut microbiome composition when subjects were helminth-uninfected. In the proposed joint model in the longitudinal setting, we observed a significant association between helminth infection on microbiome composition but not in cytokine response. With regard to the estimated fixed effect, our proposed method is in line with the inference in the naive approach where in helminth-infected subjects, the *Bacteroidetes* proportion was negatively associated with cytokine response. With regard to the estimated correlation between discrete and continuous outcomes, we also observed small correlation (estimated variance of shared effect was ${\sigma^2}_{u_{S_2}}$ 0.002 (s.e of 010) and ${\sigma_{u_{S_3}}}^2$ of 0.006 (s.e. of 0.015)) while the measurement errors were relatively large and significant for both the bacterial count outcomes. With regard to the marginal correlation within the multivariate count our model gives similar correlations as observed. However the correlation between the two outcomes was not well represented by our model.

Our results of the data analysis indicated the importance of our proposed method over the naive method. First of all, the naive method considered the effect of single bacterial phyla, which ignores the correlation structure between multiple phyla imposed by the compositional structure of microbiome data. Secondly, the measurement errors in the microbiome data were ignored in the naive method. In our dataset, the variances of the measurement error for the ratio of *Bacteroidetes* and *Firmicutes* were relatively high. When this is not modelled properly, it may result in biased estimates as shown by our simulations. On the other hand the observed marginal correlation appeared not to be well modelled by our joint approach. It might be that the proposed normal distribution used for the random structure did not fit the data well. The advantage of the normal distribution is that complex structures can be easily modelled. Future work will be to

develop goodness of fit measures for our models.

We proposed here the joint model between multivariate count of three categories and continuous outcome. In general, the model could be extended to higher dimensional of multivariate outcome although the computational burden increases. Future research will be needed to develop statistical method which reduce the computational burden.

To conclude, although the joint model are challenging to fit when the outcomes are from different types, they might give more insight on three way relationships between a covariate and two outcomes. The joint model proposed here is an alternative for model with conjugate distribution which gives more flexibility in modelling the covariance structure, especially in the presence of measurement errors. However the marginal correlation between the two different outcomes is not well represented by this model.

5.6 Supplementary Materials

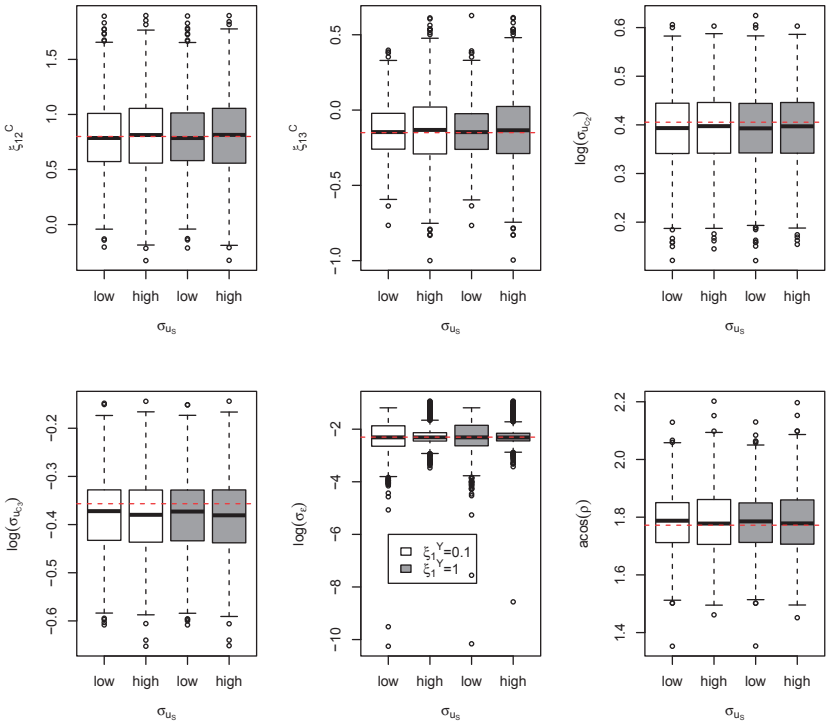

Figure S5.6.1: Simulation study at the cross-sectional setting: the distribution for all parameters under joint model with univariate random shared effect. Details of low and and high are the same as in Figure 5.2.

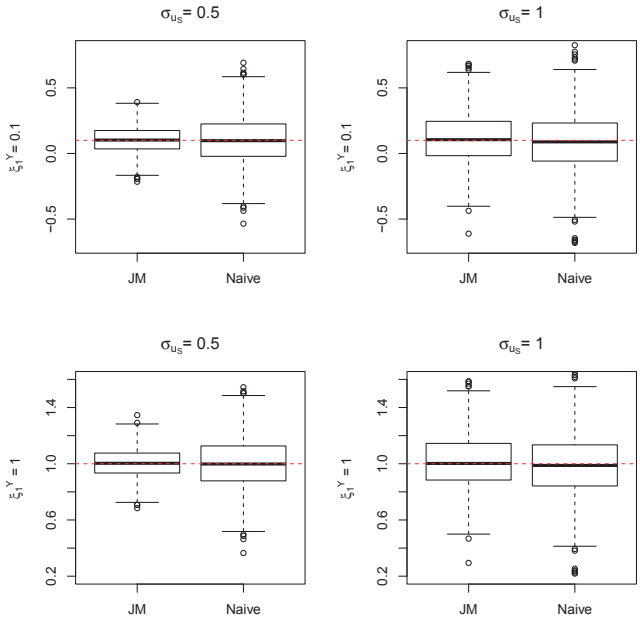

Figure S5.6.2: Simulation study at the cross-sectional setting: the point estimate for the effect of covariate of interest when dataset were generated using the joint model with a univariate random shared effects.

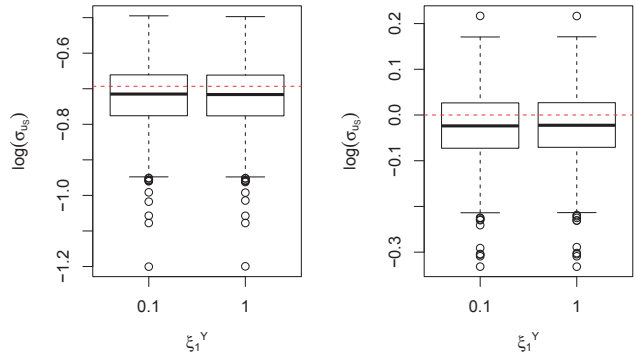

Figure S5.6.3: Simulation study at the cross-sectional setting: the distribution of the variability of random shared effect under the joint model with a univariate random shared effect

(A) The observed marginal correlation				
	C_1	C_2	C_3	Y_1
C_1	1.000	-0.264	-0.741	0.056
C_2	-0.264	1.000	-0.451	-0.355
C_3	-0.741	-0.451	1.000	0.195
Y_1	0.056	-0.355	0.195	1.000
(B) The Joint model with logit dependent shared effect.				
C_1	1.000	-0.175	-0.849	-0.092
C_2	-0.175	1.000	-0.371	0.018
C_3	-0.849	-0.371	1.000	0.077
Y_1	-0.092	0.018	0.077	1.000

Table S5.6.1: The observed and the estimated marginal correlation from joint model in the cross-sectional setting. The joint model was fitted on datasets consists of only helminth-uninfected subjects at pre-treatment (N =23).

Fixed Effects	Estimate (95%CI)	*p*-value
Intercepts		
ξ_1^Y	2.12 (1.93, 2.31)	<.0001
ξ_{02}^C	-3.02 (-3.65, -2.38)	<.0001
ξ_{03}^C	-1.01 (-1.26, -0.77)	<.0001
Random effects	Estimate (s.e)	*p*-value
${\sigma_{u_{C_2}}}^2$	1.499 (0.544)	0.016
${\sigma_{u_{C_3}}}^2$	0.204 (0.082)	0.028
${\sigma_{u_{S_2}}}^2$	-0.140 (0.107)	0.216
${\sigma_{u_{S_3}}}^2$	-0.0004 (0.039)	0.992
${\sigma_{u_Y}}^2$	0.156 (0.143)	0.294
${\sigma_\varepsilon}^2$	0.208 (0.052)	0.002
ρ	0.314 (0.207)	0.207

Table S5.6.2: Data analysis: the joint model in the longitudinal setting in subjects who were helminth-uninfected at pre-treatment and remained uninfected at 21 months after the first treatment (N=16). The model fitting used SAS with 10 quadrature points.

Parameter	Estimate (95%CI)	p-value
Infection		
Bacteroidetes	-1.04 (-1.11,-0.97)	<.0001
IL10-LPS	0.06 (-0.08, 0.20)	0.402
Time		
Bacteroidetes	-0.21 (-0.25, -0.18)	<.0001
IL10-LPS	-0.21 (-0.32, -0.09)	0.001
$\log(\sigma_\varepsilon)$	-1.12 (-1.30, -1.12)	<.0001
Random Effects		
$\sigma^2_{u_C}$	1.882 (1.183, 2.582)	<.0001
$\sigma^2_{u_S}$	0.016 (-0.085, 0.118)	0.754
$\sigma^2_{u_Y}$	0.040 (-0.044, 0.124)	0.343

Table S5.6.3: Data analysis: the joint model with two bacterial categories