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

Martin, I.

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

Ivonne Martin

Cover design: Ivonne Martin
Printing: Off Page, the Netherlands (www.offpage.nl)

ISBN: 978-94-6182-959-7

Research leading to this thesis was financially funded by the Royal Netherlands Academy of Arts and Sciences and the Joint Scholarship of Directorate General of Resources for Science Technology and Higher Education (DGRSTHE) of Indonesia and Leiden University.

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

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 8 oktober 2019
klokke 13.45 uur

door

Ivonne Martin

geboren te Jakarta, Indonesië
in 1982

Promotores:	Prof. Dr. J. J. Houwing-Duistermaat
	Prof. Dr. M. Yazdanbakhsh
Leden promotiecommissie:	Prof. Dr. E. J. Kuijper
	Prof. Dr. M. Luz Calle
	· *University of Vic, Spain*
	Prof. Dr. M. J. C. Eijkemans
	· *University Medical Centre Utrecht, Utrecht*

We may encounter many defeats but we must not be defeated.

—Maya Angelou, *The Art of Fiction no 119, The Paris Review*, 1990

Table of Contents

1 Introduction

1.1 The human gut microbiome

For more than a decade, studying human microbes and their collective genome (termed as microbiome) has became an additional method to study human function beyond protein coding genes [Grice and Segre (2012); Yadav et al. (2018)]. The whole human body is inhabited by microbes with the gastro-intestinal tract harboring the most abundant and diverse species [Ley et al. (2006); Gupta et al. (2017)]. The bacterial community interacts with the host and contributes to processes varying from metabolic homeostasis [Yadav et al. (2018)] to the development of the immune system [Gensollen et al. (2016); Thursby and Juge (2017)]. Hence disruption of the microbial community is linked to development of a variety of diseases, for instance obesity and several metabolic disorders [Sonnenburg and Bäckhed (2016)]. With the advent of high-throughput sequencing technologies, the goal of microbiome studies has shifted from mapping and cataloguing genes related to bacteria to characterizing the microbial community in relation to health and diseases [Rodrigues Hoffmann et al. (2016)]. However, these technologies have not been accompanied by the development of statistical tools necessary to analyze the data generated. This thesis presents epidemiological work related to human gut microbiome, accompanied by the development of statistical methods to analyze them. The remaining of this introduction section provides information on available features of microbiome data, microbiome related epidemiological studies and statistical methodologies.

The microbial data analyzed in this thesis were obtained using high-throughput sequencing technology. This technology worked by targeting the specific region of 16S rRNA gene that is unique to bacteria. Robinson et al. (2016) have reviewed the whole process of sequencing the 16S rRNA gene to extract the microbial data. The process begins by Polymerase Chain Reaction (PCR) amplification of these rRNA genes, followed by sequencing of the PCR products and alignment to a reference database, for instance the Ribosomal Database Project (RDP) [Cole et al. (2014)]. The sequence reads are then clustered into operational taxonomical units (OTU) which is usually based on 97% similarity (a proxy of species) [Mysara et al. (2017)]. The procedure then continues with the taxonomy annotation in which the dataset can be viewed in the format shown in Table 1.1 where rows represent samples and columns represent taxa. The total reads for each sample are usually different due to technical difficulties in loading the same molar amount to an instrument [Gloor et al. (2017)]. Usually a normalization or rarefaction is done to obtain the same number of total reads [Weiss et al. (2017)].

Taxon / Samples	**1**	**2**	...	***J***	**Total reads**
1	c_{11}	c_{12}	...	c_{1J}	c_{1+}
2	c_{21}	c_{22}	...	c_{2J}	c_{2+}
⋮	⋮	⋮	⋱	⋮	⋮
N	c_{N1}	c_{N2}	...	c_{NJ}	c_{N+}

Table 1.1: The format of the taxonomical count of microbiome data.

As a result of high-throughput sequencing, several important features of the microbiome data can be highlighted. First of all, it is a compositional data [Weiss et al. (2017)]; it consists of multiple proportions of various organisms that sums up to a constant [Gloor et al. (2017)]. The total reads per sample are up to the number of molar concentration loaded into the instrument. Thus, it is not possible to assume independence between different taxa as decreasing number of one bacterial taxa increases the other, or vice versa. Secondly, microbiome data vary highly due to unknown reasons [Turnbaugh et al. (2007)]. This could be due to sampling or individual heterogeneity, which needs to be taken into account in the statistical model. Finally, a variation in microbial taxa could occur as a result of measurement error [Li (2015)].

1.2 Soil-transmitted helminthiasis and immunity

According to the World Health Organization, approximately 1.5 billion of people are infected with soil-transmitted helminths worldwide [Collender et al. (2015)],

majority species of *Ascaris lumbricoides*, *Necator americanus*, *Ancylostoma duodenale* and *Trichuris trichiura*. Infected individuals live mostly in low and middle income countries. Due to poor level of hygiene and sanitation, these parasites enter through the skin or through orofecal route and reach the gastro-intestinal tract, which then cause malnutrition, growth stunting and physical impairment [Crompton and Nesheim (2002); Hall et al. (2008); Albonico et al. (2008)]. Aside from this negative impact, there seems also a positive impact; helminths are associated with lower incidence of metabolic disorders [Wiria et al. (2012)]. This may be related to the ability of helminths to modulate the host's immune system [Wiria et al. (2012); McSorley and Maizels (2012); Gazzinelli-Guimaraes and Nutman (2018)] or due to their parasitism, consuming energy from their host [Wiria et al. (2012)].

Given the fact that helminths reside in the same niche, it has been hypothesized that gut microbiota and the helminth parasites may interact and thus modulate the immune system (reviewed in Leung et al. (2018)). This three-way relationship has been largely analyzed in laboratory animal models, and less in humans [Wegener Parfrey et al. (2017); Reynolds et al. (2015)].

1.3 The randomized controlled trial in a repeated measurement setting

The randomized controlled trial design is ideal to analyze the causal effect of treatment on outcomes of interest [Hernán and Robins (2006a); Hernan and Robins (2018)]. The dataset used in this thesis was obtained from a household-based cluster-randomized, double-blind, placebo-controlled trial conducted in an area endemic for helminth infections in Indonesia with the main purpose of analyzing the association between immunological responses and helminth infections [Wiria et al. (2010)]. Briefly, irrespective of their infection status, subjects were randomized into treatment or placebo arm. An anthelminthic treatment was administered every three months for a total period of 21 months. The stool samples were collected at two different time-points, namely before and 21 months after the first treatment to identify helminth infection status and collect microbial data. The outcomes modelled in this thesis are the microbiome composition and immune responses. Using this randomized treatment design, the causal effect of anthelminthic treatment on helminth infections, microbiome composition and immune responses is assessed.

In addition, the longitudinal design enables one to study the dynamic association between helminth infections and outcomes of interests over time. At 21 months after the first treatment, irrespective of the treatment allocation, subjects may either get a new infection, remain infected, or may be helminth-uninfected. Specifically, this design enables one to take into account gut microbiome changes

over time due to changing lifestyles. The challenge in the statistical analyses of data from a repeated measurements design is the fact that observations from the same individuals are correlated. For valid inference these correlations have to be addressed. The following sections introduce statistical methodologies for correlated data.

1.4 Methodology of compositional data

The sequencing process of 16S rRNA to obtain the microbiome data has been described earlier and the format of the dataset used in the analysis has been tabulated in Table 1.1. Each sample has counts of sequence reads that are clustered in multiple categories with arbitrary total reads imposed by the instrument. As it has been described above, the important feature of this data is the compositional structure, in which counts for each category cannot be considered as an independent realization. Thus, it is important to analyze this data using a multivariate approach.

Typically, analyses in microbiome studies aim to characterize the relationship between the microbiome composition and biological, clinical or environmental features [Xia and Sun (2017)]. For this purpose, either a nonparametric or parametric approach can be used. For instance, nonparametric ecological distances or dissimilarity measures, such as alpha and beta diversity measures are commonly used for comparisons between groups. By doing so, no parameters are estimated. Note that while no specific distribution needs to be assumed for the outcome variable, the hypothesis testing may still have assumptions which might be violated. For instance, the assumption that observations are obtained from the distribution that has the same shape and are independent [Lumley et al. (2002)]. In addition, for small sample sizes, the statistical power might be limited to detect differences. The parametric approach may provide a solution for these issues. In this thesis, the parametric approach is deployed because: the interest is in modelling complex relationships in a relatively small study.

When considering the microbiome data in Table 1.1, each cell's entry c_{ij} represents the count for subject i belonging to taxa j. Count data are typically modelled using the Poisson distribution. When considering multivariate count data with a fixed total count, the multinomial distribution is used. These parametric distributions have a restricted assumption, namely the variances of the responses are specified by the means. This appears not to hold for microbiome data. As it has been pointed out in Section 1.1, microbiome data vary highly due to presence of many zeros, outliers or heterogeneity caused by sampling mechanisms and differences between individuals. An extra variation is thus observed, which is known as overdispersion.

To account for overdispersion in modelling multivariate count data, a compound distribution is used, i.e., when observations are drawn from multinomial

or Poisson distribution, its mean parameters are assumed to be random variable following a specific distribution. Two often used distributions are the conjugate distribution (Gamma for Poisson and Dirichlet for multinomial) and the normal distribution. The advantage of using the conjugate distribution is that the marginal distribution has a closed form formula which consequently reduces computational time and directly take the measurement error into account [Li (2015)]. However, the drawback of this model is the restriction of the number of parameters to model the covariance structure [Li (2015)]. As an alternative, multivariate normally distributed random effects can be used [Hartzel et al. (2016); Hedeker (2003)]. The trade off between these options is in the computational burden.

Finally, as studies in this thesis were done in a repeated measurement setting, the compositional data are observed at two time-points. Hence, a correlation structure is imposed on the data by multiple sources, namely the correlation between different bacterial categories at the same time-point and the correlation within the same categories at different time-points.

1.5 General mixed models

In a repeated measurement setting, one is interested in analyzing the progression of outcomes over time in relation to predictors. In the simplest case, this outcome-predictor relationship is assumed to be linear, and linear regression can be used. However, it should be noted that under a repeated measurement setting, observations are not independent since observations from the same subjects are likely to be more similar than observations from others. Thus the variability in the observations could be due to variability of the observations within subjects and between subjects. The formulation of this linear regression needs to be extended to account for this extra variation.

Starting from linear model context, it is assumed that for subject i, a n_i numbers of observations are collected $\boldsymbol{Y}_i = \{Y_{i1}, \ldots, Y_{in_i}\}$. The observations from subject i, $\boldsymbol{Y}_i$ are modelled as individual response trajectory $\boldsymbol{\mu}_i$ and an independent residual term $\boldsymbol{e}_i$.

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{e}_i \tag{1.1}$$

with $\boldsymbol{e}_i$ is a random variable assumed to follow the multivariate normal distribution with zero mean and variance of $\Sigma_e = \sigma_e^2 \mathbf{I}_{n_i}$ (denoted as $\boldsymbol{e}_i \sim \text{MVN}(\mathbf{0}, \Sigma_e)$). The profile response $\boldsymbol{\mu}_i$ is then linked to the design matrix $\boldsymbol{X}_i$ via the following equation

$$\boldsymbol{\mu}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{A}_i \boldsymbol{b}_i, \tag{1.2}$$

in which $\boldsymbol{X}_i$ is the design matrix for the fixed effect part (population) and $\boldsymbol{A}_i$ is the design matrix for the random effect part (subject-specific trajectory). The parameter $\boldsymbol{\beta}$ represents the population part and $b_i \sim \text{N}\{0, \sigma_b^2\}$ are the subject-specific

effect. This is known as a linear mixed model which consists of a fixed population parameter β and a subject-specific effect b_i [Laird and Ware (1982)]. The estimate of the regression parameters $\boldsymbol{\beta}$ and subject-specific variability σ_b^2 are obtained by maximizing the full likelihood which has a closed form formula.

The above formulation of the model is for the continuous outcome, where the response conditional on the covariates follows the normal distribution. For other outcomes, the generalized linear model needs to be extended to account for subject-specific effects [Molenberghs and Verbeke (2005)]. The expected value of the responses $\boldsymbol{\mu}_i$ in equation (1.1) is linked with the predictor via the so-called link function g. For instance, in the case of a count response, the function of its expected value μ_i becomes

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}_i$$

in which $\boldsymbol{\beta}_i$ is defined as in equation (1.1) and the logarithm function is used as a link function. For the multinomial distribution often used for microbiome data, the logit link function is used.

1.6 Outline of the thesis

This thesis is a collection of five articles which are published or submitted for publication. It is organized in two major parts. Each of this part consists of an epidemiological and a methodological paper to aid in understanding the underlying biological mechanisms which motivate development of the statistical method in the topic related to microbiome and immunity in relation to helminth infections. The last chapter discusses all results of these two parts.

The focus in **Part I** is modelling the association between helminth infections and the gut microbiome composition in a randomized-controlled trial setting. In **Chapter 2**, an advanced statistical method was utilized to analyze this association while addressing the multivariate structure of the compositional data. The correlation between bacterial taxa is accounted for by introducing a conjugate distributed random effect. This method does not account for the correlation between observations. As the data were repeatedly measured, the correlation of the observation between different time points was left unaccounted for in the statistical model.

In **Chapter 3**, a statistical framework to analyze microbiome data was developed to account for the additional correlation due to repeated measurements. We extended the Dirichlet - multinomial regression model used in **Chapter 2** to include an extra normally distributed random effect. Several covariance structures for the normally distributed random effects were considered. The conjugate distributed random effect serves as an estimate for the correlation between categories. In addition, a loglinear model approach is introduced to aid interpretation of the regression estimates.

It has been described earlier in this chapter that there is a hypothesized three-way relationship between helminth infection, gut microbiome and human immune response. Thus, the aim of **Part II** of this thesis is to analyze these complex relationships. In **Chapter 4**, the linear mixed model with subject-specific random effect is utilized to study the role of gut microbiome on human immune response in the presence or absence of helminth infections. Stimulated cytokine responses were considered as a marker of immune response while bacterial proportion and helminth infection were included as covariates. Since helminth infection might be a confounder for both microbiome and cytokine responses, it was included as covariate in the model, however the relationship between helminth infection and microbiome composition was not modelled. In addition, the presence of measurement error in the microbiome data was not taken into account in the model.

In **Chapter 5** of this thesis, the challenges in modelling the relationships between helminth infection, gut microbiome and immune response were addressed by using a joint model. The correlation between observation of the same subject as well as the measurement error present in the microbiome sequencing data were taken into account. A flexible covariance structure was used for this purpose, namely by including multivariate normally distributed random effects in the model. The proposed mixed regression model is able to model the association between helminth infection on both outcomes and the association between the different outcomes, while accounting for the correlation over time-points and between bacterial categories. Finally, **Chapter 6** discusses the relationship between helminth infections, gut microbiome and cytokine responses by using the directed acyclic graph. Using results from our and other published studies, the potential biases of estimated associations are identified and discussed. Possible solutions are presented.

Part I

Gut microbiome composition and helminth infections

2

Dynamic changes in human-gut microbiome in relation to a placebo-controlled anthelminthic trial in Indonesia

Abstract

Background. Microbiome studies suggest the presence of an interaction between the human gut microbiome and soil-transmitted helminth. Upon deworming, a complex interaction between the anthelminthic drug, helminths and microbiome composition might occur. To dissect this, we analyse the changes that take place in the gut bacteria profiles in samples from a double blind placebo controlled trial conducted in an area endemic for soil transmitted helminths in Indonesia.

This chapter has been published as: Ivonne Martin, Yenny Djuardi, Erliyani Sartono, Bruce A. Rosa, Taniawati Supali, Makedonka Mitreva, Jeanine J. Houwing-Duistermaat, Maria Yazdanbakhsh (2018). Dynamic changes in human-gut microbiome in relation to a placebo-controlled anthelminthic trial in Indonesia. *PLoS Neglected Tropical Diseases* 12 (8):e0006620.

Methods Either placebo or albendazole were given every three months for a period of one and a half years. Helminth infection was assessed before and at 3 months after the last treatment round. In 150 subjects, the bacteria were profiled using the 454 pyrosequencing. Statistical analysis was performed cross-sectionally at pre-treatment to assess the effect of infection, and at post-treatment to determine the effect of infection and treatment on microbiome composition using the Dirichlet-multinomial regression model.

Results At a phylum level, at pre-treatment, no difference was seen in microbiome composition in terms of relative abundance between helminth-infected and uninfected subjects and at post-treatment, no differences were found in microbiome composition between albendazole and placebo group. However, in subjects who remained infected, there was a significant difference in the microbiome composition of those who had received albendazole and placebo.

This difference was largely attributed to alteration of Bacteroidetes. Albendazole was more effective against *Ascaris lumbricoides* and hookworms but not against *Trichuris trichiura*, thus in those who remained infected after receiving albendazole, the helminth composition was dominated by *T. trichiura*.

Discussion We found that overall, albendazole does not affect the microbiome composition. However, there is an interaction between treatment and helminths as in subjects who received albendazole and remained infected there was a significant alteration in Bacteroidetes. This helminth-albendazole interaction needs to be studied further to fully grasp the complexity of the effect of deworming on the microbiome.

Trial registration ISRCTN Registy, ISRCTN83830814.

2.1 Introduction

Shortly after birth, the human body is colonized by a community of bacteria [Zaiss and Harris (2016),Macpherson and Harris (2004)] with relatively simple composition which increase in number and complexity with age [Ursell et al. (2012)]. The densest colonization with commensal microbes of the human body is found in the intestine [Savage (1977)] which has a beneficial impact on gastrointestinal function and host health by providing support for host metabolism, protection against pathogenic microbes, integrity of intestinal mucosa, and modulation of the immune system [Macpherson and Harris (2004), Ursell et al. (2012), Eckburg et al. (2005)]. Furthermore, it has been shown that intestinal microbiota is associated with dietary habits [Turnbaugh et al. (2009), Conlon and Bird (2014)], physiological factors such as age, gender and BMI [Haro et al. (2016), Yatsunenko

et al. (2012)] as well as diseases, such as inflammatory bowel disease and obesity [Zaiss and Harris (2016), Eckburg et al. (2005), Turnbaugh et al. (2006)].

Apart from intestinal microbiota, certain pathogens such as soil-transmitted helminths (STH) may coexist in the human intestine. It is estimated that STH, largely represented by *Ascaris lumbricoides*, hookworm such as *Necator americanus* and *Ancylostoma duodenale*, and whipworm *Trichuris trichiura*, infect 2 billion people in the majority of developing countries and mostly children [Zaiss and Harris (2016), Hotez et al. (2008)]. These infections have been reported to cause impairments in physical, intellectual, and cognitive development [Bethony et al. (2006)]. At the same time, these parasitic worms have a long co-evolutionary interaction with their host. The result of this co-evolutionary trajectory, seems to be that helminths lead to immune regulatory responses that allow their long term survival within their host [McSorley and Maizels (2012), Wammes et al. (2016)]. Since intestinal microbiota and helminths share the same niche in their host, it is hypothesized that the presence or absence of intestinal helminths may affect their interaction with each other within the host. In an interesting study, evidence was provided for the beneficial effects of the microbiome on successful completion of whipworm life cycle [White et al. (2018)]. Currently, there is also much interest to determine whether helminth infections affect the gut microbiome and whether the effects of worms on human health is mediated via alteration in the microbiome composition. It is becoming increasingly clear that the gut microbiota has important link to the immune system and several disease outcomes. With the mass drug administration programs underway to eliminate intestinal helminths in many endemic regions, it is essential to fully understand the consequences of deworming on community health by characterizing the effect on the gut microbial composition.

Recently, several studies investigated the relationship between the intestinal microbiome and intestinal helminth infections. In swines, a statistically significant association between *Trichuris* infection and the gut microbiome composition was shown [Li et al. (2012), Holm et al. (2015)], evident from the altered abundance of the genus Paraprevotella and phylum Deferribacteres in the infected pigs. The chronic infection of *Trichuris muis* in C57BL/6 wild-type mice increased the relative abundance of Lactobacilli [Holm et al. (2015)], while giving *T. trichiura* ova to macaques with chronic diarrhea increased the phylum Tenericutes and resulted in clinical improvement [Broadhurst et al. (2012). Therefore, in animal models, *Trichuris* infection seems to be associated with alternation in the gut microbiome. However, in humans, findings are not consistent. In an observational study in Ecuador, comparing the gut microbiome of infected and uninfected school children, no significant differences at various taxonomical levels were found [Cooper et al. (2013)]. On the contrary, two other observational studies in rural villages of Malaysia [Lee et al. (2014)] and Zimbabwe [Kay et al. (2015)] found a significant increase in diversity and abundance of certain bacteria

taxa in infected compared to uninfected subjects. An increase in *Paraprevotellaceae* was seen in the Malaysian study, which seemed to be associated with Trichuris infection while an increase in *Prevotella* was reported in the study in Zimbabwe that was attributed to *S. haematobium* infection. Furthermore, in an interventional study carried out in another rural village in Malaysia [Ramanan et al. (2016)], a significant change in order *Bacteroidales* and *Clostridiales* was observed after deworming while deworming of *S. haematobium* in an interventional study in Zimbabwe [Kay et al. (2015)] did not seem to alter the microbiome.

The study designs which were used to investigate the human-gut microbiome in relation to helminth infections were either observational [Cooper et al. (2013), Lee et al. (2014), and Kay et al. (2015)] or interventional without a control group [Cooper et al. (2013), Lee et al. (2014), Kay et al. (2015), and Ramanan et al. (2016)] hampering the estimation of the true relationship between helminth infection and the microbiome composition. Motivated by the findings from previous studies of helminths on microbiome, we used samples from a larger randomized placebo-controlled trial of albendazole treatment in a population living in an area endemic for soil-transmitted helminth infections [Wiria et al. (2010)] to further characterize the effect of helminth infection and treatment at before and 21 months after treatment. The study design allowed the investigation of the effect of helminths on the fecal microbial community through comparing helminth infected and uninfected at baseline and subsequently assessing the effect of treatment with albendazole. We also explored the effect of the interaction between treatment and infection status on the faecal microbiome. In addition, we used the opportunity to assess whether albendazole has a direct effect on the microbiome by analyzing those who received albendazole and were uninfected throughout the study. The placebo group enables the estimation of the effect of deworming on the microbiome composition in the absence of anthelminthic treatment which itself could affect the microbiome.

The analyses carried out in this study aim to characterize the joint effects of several predictors, such as helminth infection and treatment on each bacterial category. For comparing the gut microbiome of premature infants with different severities of necrotizing enterocolitis, a Dirichlet – multinomial model was used [Barron et al. (2017)]. Here, we consider the same approach for modelling and hypothesis testing for the association between treatment and helminth infection on microbial composition at the phylum level. Our approach addresses the possible correlation between bacteria categories, the compositional feature of the microbiome data [Chen and Li (2013)], and the multiple testing issue.

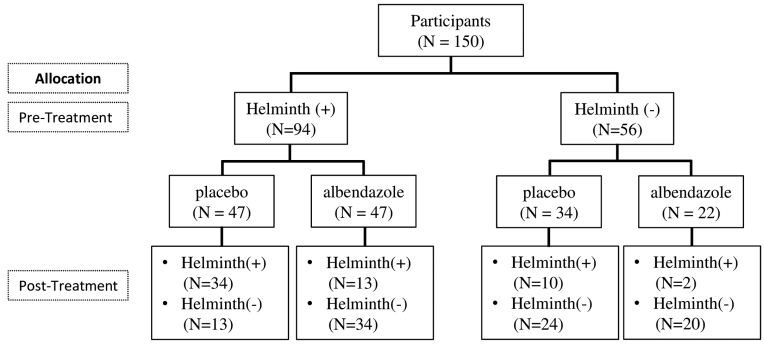


Figure 2.1: **The profile of the microbiome study**. The chart shows the number of subjecs infected with at least one of the prevalent soil transmitted helminths (Helminth (+)) or free of helminth infection (Helminth (-)) that belonged to either the placebo or albendazole treatment group, at pre-treatment and 21 months after treatment.

2.2 Methods

2.2.1 Ethics statement

This study was nested within the ImmunoSPIN study, a double blind placebo-controlled trial conducted in Flores Island, Indonesia [Wiria et al. (2010)]. The ImmunoSPIN study has been approved by the Ethical Committee of Faculty of Medicine, Universitas Indonesia, ref:194/PT02.FK/Etik/2006 and has been filed by ethics committee of the Leiden University Medical Center. The clinical trial was registered with number: ISRCTN83830814 in which the protocol for the trial and supporting CONSORT checklist are available elsewhere [Wiria et al. (2013)]. The subjects gave their informed consent either by written signature or thumb print. Parental consent was obtained for children below 15 years old.

2.2.2 Sample populations and detection of soil-transmitted helminth (STH) infection.

Households were randomized to receive either a single dose of 400 mg albendazole or placebo once every 3 months for 2 years. To assess the effect of treatment on the prevalence of soil transmitted helminth infection, yearly stool samples were collected on a voluntary basis. *T. trichiura* infection was detected by microscopy and a multiplex real time PCR was used for detection of hookworm (*A. duodenale*, *N. americanus*), *A. lumbricoides* and *Strongyloides stercoralis* DNA. For the current study, paired DNA samples before and at 21 months after treatment from 150 inhabitants in Nangapanda were selected based on the treatment allocation and infection status as well as the availability of complete stool data at pre and

post-treatment (Figure 2.1). The procedure for sample collection and processing is already described elsewhere [Wiria et al. (2010)].

Briefly, prior to DNA isolations, approximately 100 mg unpreserved faeces (kept at -20°C) were suspended in 200µl PBS containing 2% polyvinylpolypyrolidone (PVPP;Sigma, Steinheim, Germany). Suspensions were heated at 100oC for 10 min and were treated subsequently with sodium dodecylsulphate-proteinase K at 55°C for 2 h. DNA was isolated using QIAamp DNeasy Tissue Kit spin columns (QIAgen, Venlo, The Netherlands). The whole procedure of DNA isolations and setup of PCR plates were performed using a custom-made automatic liquid handling station (Hamilton, Bonaduz, Switzerland). As published already, sequences of the *A. lumbricoides* and *N. americanus*-specific primers and probes as well as the *A. duodenale* specific XS-probes were used to accommodate the specific fluorophor combinations of the CFX real-time PCR system (Table S1[Wiria et al. (2010), Verweij et al. (2007)] . The real-time PCRs were optimized first as monoplex assays with 10-fold dilution series of *A. duodenale*, *N. americanus* and *A. lumbricoides* DNA, respectively. The monoplex realtime PCRs were thereafter compared with the multiplex PCR with the PhHV internal control. The cycle threshold (Ct) values obtained from testing the dilution series of each pathogen in both the individual assay and the multiplex assay were similar, and the same analytical sensitivity was achieved. Amplification reactions were performed in white PCR plates in a volume of 25µl with PCR buffer. Amplification consisted of 15 min at 95°C followed by 50 cycles of 15 s at 95°C, 30 s at 60°C, and 30 s at 72°C. Amplification, detection, and analysis were performed with the CFX real-time detection system (Bio-Rad laboratories). The PCR output from this system consists of a cycle threshold (Ct) value, representing the amplification cycle in which the level of fluorescent signal exceeds the background fluorescence and reflecting the parasite-specific DNA load in the sample tested. In this manuscript, we set the ct value 30 as a threshold for the infection status i.e. subjects with PCR lower than 30 was identified as clearly infected and PCR above 30 as uninfected or very low infection. The analyses were carried with regard to the infection status and we do not consider the analysis in the level of infection.

2.2.3 Sequencing of 16S rRNA gene

Genomic DNA samples were isolated from 100 mg of fresh stool, which were also used for detection of helminth infection by real time PCR. The DNA amplification and pyrosequencing followed the protocols developed by the Human Microbiome Project (HMP) [Group HDGW (Group HDGW)] at the McDonnell Genome Institute, Washington University School of Medicine in St. Louis. Briefly, The V1-V3 hypervariable region of the 16S rRNA gene was amplified by PCR and the PCR products were purified and sequenced on the Genome Sequencer Titanium FLX (Roche Diagnostics, Indianapolis, Indiana), generating on average

6,000 reads per sample. The filtering and analytical processing of 16S rRNA data for this cohort has been previously described in details [Rosa et al. (2018)]. The assembled contigs count data as a result of RDP classification was organized in matrix format with taxa in columns and subjects in row. The entries in the table represent the number of reads for each phyla for each subject. Rarefaction to 2000 reads was performed using an R package (vegan) [Oksanen et al. (2017)]. We obtained the count data of 609 bacterial genera and 18 bacteria phyla. In the analysis at phylum level, we retained the 5 most prevalent phyla (Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, and Unclassified Bacteria) and pooled the remaining phyla into a pooled category such that there are only 6 phyla categories. The Unclassified bacteria represents the category where all the sequences cannot be assigned into a phylum. We conducted further analyses by decomposing the statistically significant phylum (Bacteroidetes) into the two most prevalent genera (*Bacteroides* and *Prevotella*) and the remaining genera into a pooled Bacteroidetes category and combining the Proteobacteria, Unclassified Bacteria and Pooled in a Pooled Phyla category. In total we have six categories since we also selected Actinobacteria and Firmicutes at phylum level.

2.2.4 Statistical methods

The within sample diversity (Shannon and richness diversity) indices as well as the between sample diversity (Bray-Curtis distance) were computed at baseline and follow-up using the dataset at genera level. Clustering of samples and bacteria was studied by plotting a heat map of bacteria genera which were present in at least one sample and which had an average relative abundance of more than 1%. This cutoff was chosen to exclude rare genera. Unless stated otherwise, the rest of the analyses were done at the phylum level. A Pearson's chi-squared test statistic was used to test for differences of infection prevalence between the two treatment groups at pre and at post-treatment. Although the study design allows for the pairwise analysis, unfortunately no method is available for multivariate categorical count data. For this reason, we used the Dirichlet-multinomial regression where the characterization of infection and treatment are similar to the interpretation in loglinear model. Each count outcome within a category was assumed to follow the negative binomial distribution. This distribution is the result of a Poisson distribution for counts with the additional assumption that the underlying parameter is a random variable which follows the conjugate distribution (Gamma). By assuming that the underlying parameter was random, the presence of overdispersion due to multiple counts observed within a sample was modelled. To incorporate the fact that the total count is fixed per sample, we conditioned the probability of the multivariate count outcome on the total count per sample. This model is equivalent to the approach of Guimarães and Lindrooth (2007), i.e. the Dirichlet-multinomial regression model. The model parameters

are log of odds ratios which compare the prevalence rate of each bacteria phyla associated with the covariates with the reference category. In all analyses, Firmicutes was used as reference since it has the highest abundance among the phyla. The covariates were infection status and treatment allocation which are both binary variables.

The likelihood ratio statistic was used to test the null hypothesis of no effect of the covariate on the microbiome composition. The test statistic follows asymptotically a χ^2 distribution with J degrees of freedom, representing the $J-1$ bacterial comparison with the reference and one overdispersion parameter. As the Dirichlet – multinomial regression is available for cross-sectional setting, we modelled the association between microbiome composition and covariates including treatment at 21 months after treatment. First, we modelled the association between treatment and microbiome composition by including all study participants. Next, we selected subjects who were infected with at least one single helminth at baseline and included a categorical variable representing the four combinations of treatment allocation and infection status at post-treatment in the model. The R package MGLM [Zhang and Zhou (2017)] was used for analyses. The results were reported in terms of odds ratios, 95% confidence intervals and p-values.

To confirm our finding with this method, we used the univariate pairwise analysis for single bacterial categories of interest in albendazole arm. For this purpose, the inverted beta binomial test was applied to test the null hypothesis that the relative abundance of certain bacteria category at pre-treatment is similar to the relative abundance at post-treatment. Note that the inverted beta-binomial regression model is only defined for two categories and is equivalent to the Dirichlet multinomial. The R package ibb [Pham and Jimenez (2012), Pham (2013)] was used for this test. All computations were conducted in R version 3.1.0 [R Core Team (R Core Team)].

2.3 Results

2.3.1 Characteristics of the study subjects

At baseline, 94 out of 150 (62.7%) individuals were infected with one or more helminth species, and hookworm was the most dominant species (52.1%) followed by *T. trichiura* (44.7%) and *A. lumbricoides* (37.2%). The baseline characteristics such as age, gender, and helminth prevalence were similar between the two treatment arms although the prevalence of *N. americanus* was slightly higher in albendazole group, but not statistically significant (Table 2.1). The additional relevant characteristics of the participants are listed in Table S2. With regard to the microbiome composition, the proportions of each bacterial phyla were also similar between two treatment arms with the highest abundance at the phylum level being Firmicutes followed by Actinobacteria, Proteobacteria and Bacteroidetes.

Characteristics	pre-treatment		post-treatment	
	albendazole arm	placebo arm	albendazole arm	placebo arm
	(N = 69)	(N = 81)	(N = 69)	(N = 81)
Age,mean(SD) (in years)	27.38 (16.5)	27.85 (16.9)		
Sex, female, n(%)	39 (56.5)	45 (55.6)	39 (56.5)	45 (55.6)
Helminth Infections, n(%)				
Single infection				
A. lumbricoides	17 (24.6)	18 (22.2)	1 (1.4)	7 (8.6)
Hookworm	26 (37.7)	23 (28.4)	3 (4.3)	11 (13.6)
N. americanus	25 (36.2)	23 (28.4)	3 (4.3)	10 (12.3)
A. duodenale	2 (2.9)	2 (2.5)	0 (0)	1 (1.2)
T. trichiura	20 (28.9)	22 (27.2)	9 (13.0)	9 (11.1)
Multiple infection•)				
A. lumbricoides	17 (24.7)	18 (22.2)	3 (4.3)	23 (28.4)
Hookworm	26 (37.7)	23 (28.4)	3 (4.3)	20 (24.7)
T. trichiura	20 (28.9)	22 (27.1)	11 (15.9)	23 (28.4)
Any helminth	47 (68.12)	47 (58.0)	15 (21.7)	44 (54.3)
Proportion (in %) of the 6 most abundant bacteria phyla, mean(SD)				
Actinobacteria	12.5 (8.9)	11.0 (7.9)	13,2 (8.4)	11.8 (8.5)
Bacteroidetes	7.4 (11.3)	6.4 (11.0)	5.7 (9.5)	6.2 (12.5)
Firmicutes	66.8 (13.5)	70.0 (13.7)	66.0 (13.8)	68.1 (14.2)
Proteobacteria	9.8 (7.9)	9.2 (8.4)	11.7 (11.0)	10.1 (8.6)
Unclassified*)	2 (2.22)	2.7 (3.2)	2.1 (1.6)	2.6 (2.7)
Pooled#)	1.5 (3. 7)	0.7 (1.2)	1.3 (2.2)	1.2 (2.4)

Table 2.1: **Characteristics of the study subjects at baseline and at 21 months after the initial treatment**.
•) Species is indicated that is in combination with one or more of the other helminth species.
*)Unclassified represents sequences that cannot be assigned to a phyla.
#)Pooled category consists of the remaining 13 phyla having average relative abundance among samples less than 1%.

At 21 months after treatment, the prevalence of STH infection was 21.7% in the albendazole arm and 54.3% in placebo arm (p-value < 0.001). Albendazole had the greatest effect on hookworm (24.7% (placebo) vs 4.3% (albendazole)) followed by *A. lumbricoides* (28.4% (placebo) vs 4.3% (albendazole)) and lastly *T. trichiura* (28.4% (placebo) vs 15.9% (albendazole)). These percentages are similar to what was seen in the whole ImmunoSPIN trial [Wiria et al. (2010)]. These data show

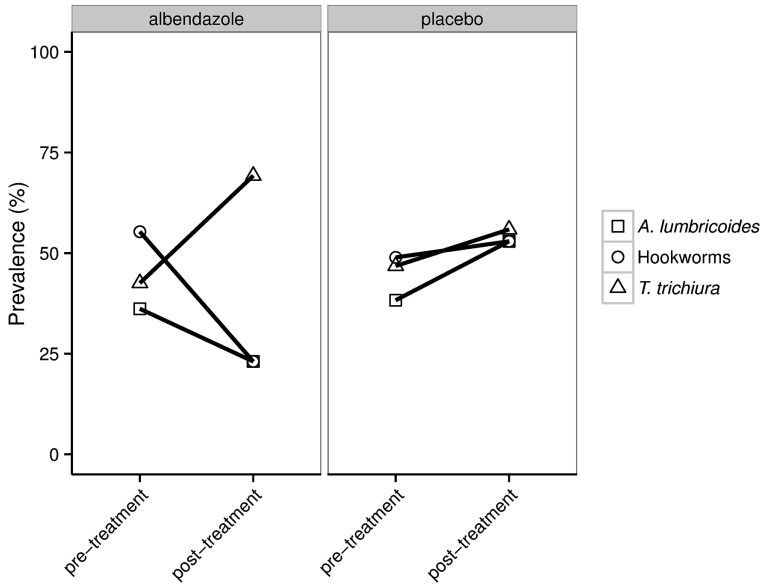


Figure 2.2: **The prevalence of helminth coinfections in two randomization arms for subjects who were infected at pre-treatment and remained infected at post-treatment**. For each helminth species depicted in the plot, square represents the percentage of subjects infected with *A. lumbricoides* (with or without other helminth species), circle represents hookworm (with or without other helminths) and triangle represents *T. trichiura* (with or without other helminths).

that while infections with *A. lumbricoides* and with hookworms decrease at post-treatment, the infections with *T. trichiura* was not affected much by albendazole and therefore the proportion of individuals infected with *T. trichiura* increased when considering those that remained infected at post-treatment (Figure 2.2). In the placebo group, there was no such difference in the composition of helminth species at post-treatment. It was noted that 12 (2 from albendazole and 10 from placebo) out of 56 uninfected subjects at baseline (21.4%) gained helminth infection over the study time period.

2.3.2 Effects of helminths and treatment on microbiome diversity

Using bacterial data at the genus level (a total of 609 genera), we calculated the within sample diversity (richness and Shannon index) and between sample diversity (Bray-Curtis dissimilarity). We observed a similar within-sample diversity at pre and post-treatment as evident from the Shannon diversity index (2.99 vs 2.96) and the richness index (66.17 vs 62.16). The Bray-Curtis dissimilarity measures the percentage of similarities between two samples in a community and the val-

ues range from 0 (completely similar) to 1 (completely dissimilar). As reported earlier [Rosa et al. (2018)], the Bray-Curtis dissimilarities calculated from 150 subjects at pre-treatment was 0.61 and the same average was obtained when calculating the Bray-Curtis dissimilarities at post-treatment, indicating that in average there was 61% dissimilatory percentages between each pairs of samples. When stratifying all samples based on infection status at pre-treatment and on randomization arm at post-treatment, again we observed similar beta-diversities, indicating that neither infection nor treatment induced a shift in diversity. When analyzing the genera in relation to infection status rather than treatment, the average Shannon diversity index as well as the average richness was similar between the infected and the uninfected group at pre-treatment and post-treatment (Figure S1).

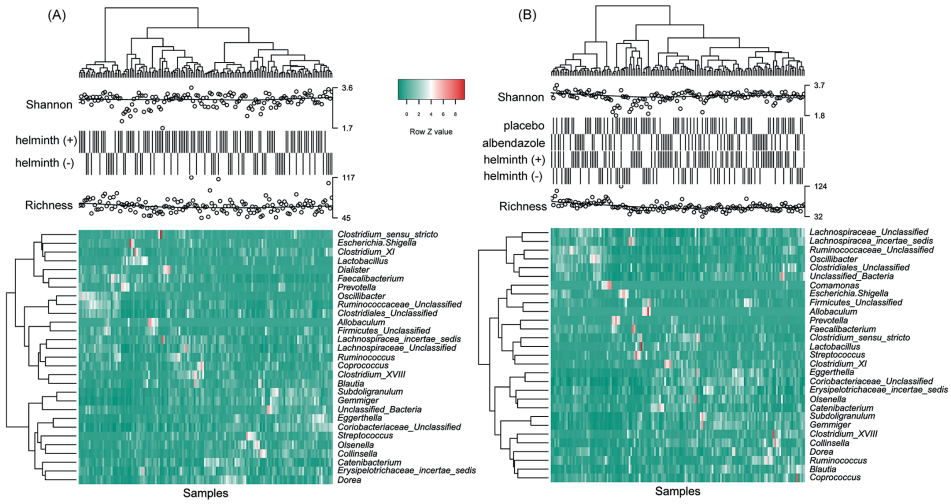


Figure 2.3: **Heatmaps showing the relative abundance of the 29 most abundant genera of each sample at pre-treatment (A) and post-treatment (B)**. Each column in the heatmap represents a specific sample and each row represents a genera. Colors represent the scaled relative abundance of genera with green and red representing low and high abundance, respectively. Samples and genera were clustered hierarchically (using the Ward method [Murtagh and Legendre (2014)]) based on Euclidean distance of the relative abundance profiles and were depicted on the top and left dendrogram, respectively. The infection status, treatment allocation, Shannon and richness indices for all samples were annotated above the heatmap. Circles in Shannon and richness represents the diversity indices for each sample. There is no clustering of samples or genera based on infection or treatment status.

The average relative abundances of all bacterial genera at both time-points were below 10%, with the highest being in the phylum Firmicutes, specifically the genus *Catenibacterium* (6.7% at pre-treatment) and the unclassified genus belonging to the family *Ruminococcaceae* (5.6% at post-treatment). The relative abundance at the genus level as well as the dominant genera vary between populations

as observed in studies where samples in rural Ecuador [Cooper et al. (2013)] or Malaysia were compared with the US [Lee et al. (2014)] or in studies where samples of healthy European and American adults were analysed [Arumugam et al. (2011)]. To illustrate the bacterial genera profile in relation to infection and treatment status, we selected the 29 genera (at pre and post-treatment) with an average of relative abundance across all samples larger than 1%. Genera from phylum Firmicutes are the most dominant (21 of 29 genera belongs to Firmicutes). As shown in heatmaps based on composition of the most prevalent genera, no significant clustering could be seen, neither at the level of bacteria nor at the level of individuals (Figure 2.3A and B) in relation to helminth infection or treatment, which indicates that neither helminths nor treatment affected the predominant genera in the gut.

2.3.3 The association of infection and treatment with the microbiome composition at the phylum level

Using the Dirichlet-multinomial regression model, we observed that there was no difference on the microbiome composition at the phylum level when subjects with any helminth infection were compared with uninfected ones either at pre (Figure 2.4A) or at post treatment (Figure 2.4B) time points. The same was the case when infection with a specific helminth species was considered (Figure2.4A and B).

The Dirichlet-multiomial regression model was also used to discern the effect of helminths and treatment on the microbiome data at post treatment. Six bacterial categories were considered in the analyses with Firmicutes used as a reference. The effect of treatment on microbiome composition in all individuals irrespective of whether they were infected or not at post-treatment was not significant. No differences were observed between placebo and albendazole at post-treatment (p-value = 0.305, Table 2.2A, likelihood ratio test). We further selected subjects who were infected at baseline (N=94) and characterized their microbiome composition at post-treatment with regard to their infection status and treatment arm, namely: subjects who lost their infection either in the albendazole (group 1, N= 34) or placebo arm (group 2, N = 13), and subjects who remained infected in either the albendazole (group 3,N = 13) or placebo arm (group 4, N= 34). We compared the microbiome composition of the first three groups to the group of remained infected in the placebo arm (group 4) as the latter group were neither influenced by treatment nor the changing of infection status. When subjects who were infected at pre-treatment and lost their infection in the albendazole arm were compared to subjects who remained-infected in placebo group, no differences were observed (p-value of 0.371, Table 2.2B), indicating that removing helminths with albendazole did not change the microbiome profile at a phylum level. Furthermore, in subjects who lost their infection in the placebo arm, there was a trend for de-

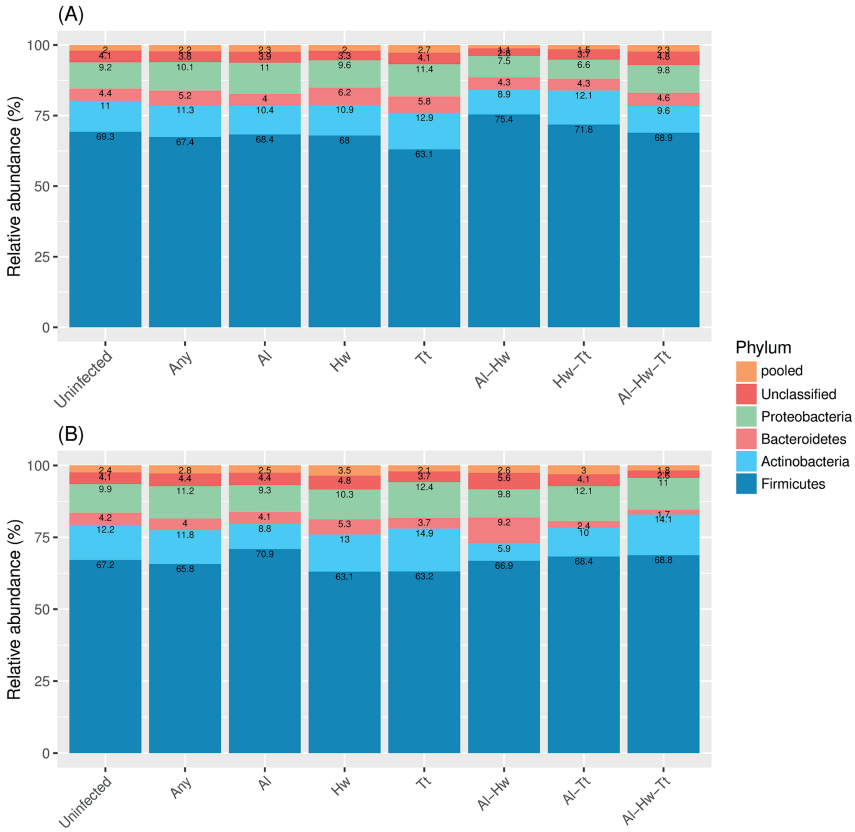


Figure 2.4: **The microbiome composition at pre-treatment (A) and post-treatment (B) stratified by helminth infection**. The stacked bar plots represent the relative abundance for each of the most abundant phyla where the Unclassified represents the category of sequences that could not be assigned to a phyla, and the pooled category consists of the remaining 13 phyla with average relative abundance less than 1%. The numbers inside the stacked bar plots show the relative abundance of the specific taxa. The microbiome compositions were depicted for group of helminth-uninfected (Uninfected), any helminth infected (Any), single helminth infection (*A. lumbricoides* (Al), hookworm (Hw) or T. trichiura (Tt)), double infection (Al – Hw, Al - Tt and Hw - Tt) or triple infections (Al – Hw - Tt).

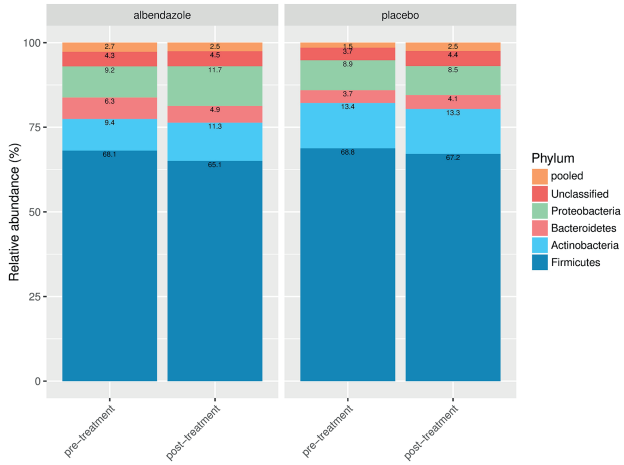


Figure 2.5: **Direct treatment effect on microbiome composition**. Details for the stacked barplots were as given in Figure 2.4. The microbiome composition is shown at pre and post-treatment for subjects who were uninfected at pre-treatment and remained uninfected at post-treatment in albendazole and placebo arm.

crease in Bacteroidetes and pooled category (OR 0.49, 95% CI:(0.27,0.91) and OR 0.47, 95% CI:(0.23,0.96), respectively, Table 2.2B), moreover, the whole composition in this group did not differ significantly from that in the group of remained infected in the placebo arm (p-value of 0.069). These two comparisons suggest that removing helminths regardless of treatment did not alter the microbiome composition when analysed at a phylum level. Interestingly, the comparison of microbiome composition between subjects who remained infected in the albendazole group was significantly different from the microbial composition in subjects who remained infected in the placebo group (p-value of 0.004,Table 2.2B). This difference was driven by the increasing odds of having Actinobacteria (OR 1.57, 95% CI of (1.05, 2.35)) and the decreasing odds of having Bacteroidetes (OR 0.35, 95% CI: (0.18,0.70)). To further analyse the direct treatment effect without the influence of helminth infection, we selected subjects who were uninfected at baseline and remained-uninfected at post-treatment (N = 44). For these subjects, we compared the microbial composition at post-treatment of subjects who received albendazole versus those who received placebo. No difference was observed (the estimate odds ratios range from 0.88, 95% CI: (0.56, 1.39) to 1.42, 95% CI: (0.88, 2.29), p-value = 0.666, illustrated in Figure 2.5), indicating that albendazole alone does not seem to affect the microbiome composition in uninfected subjects when compared at a phylum level.

As neither treatment alone nor the infection affected the microbial composition, we further hypothesized that the significant difference in microbiome com-

position in subjects who remained infected and received albendazole compared to the group that remained infected in the placebo arm was caused by the alteration of the abundance of Actinobacteria and Bacteroidetes during the treatment period. To test this hypothesis, we used the inverted beta-binomial test to compare the relative abundance of Actinobacteria and Bacteroidetes in subjects who remained infected in albendazole group at pre-treatment to the relative abundances of these bacterial phyla at post-treatment. While the relative abundance of Actinobacteria did not change significantly between pre and post-treatment (p-value of 0.155, inverted beta binomial test), the relative abundance of Bacteroidetes was estimated to be 1.88 fold higher at pre-treatment compared to post-treatment (p-value of 0.012, inverted beta binomial test). This result indicates that there is a complex interaction between helminths and treatment, which induces a change in bacterial composition during the treatment period. Using the same analysis, the direct effect of albendazole was assessed by comparing subjects who were uninfected but received albendazole at pre treatment and remained uninfected at post treatment. Although some differences were seen in the microbiome composition between pre and post-treatment, specifically in the phyla Actinobacteria, Bacteroidetes and Proteobacteria, these differences were not statistically significant (p-values of 0.149, 0.267 and 0.064, respectively). This is in line with the finding when we used the Dirichlet-multinomial regression model where no direct effect of albendazole on the microbiome composition was found. In addition, similar microbiome composition was seen in subjects free of helminth infection at baseline who received placebo and remained uninfected at post-treatment, which suggests that the microbiome was stable over time.

2.3.4 The association of Bacteroidetes genera with infection and treatment

In the Dirichlet – multinomial regression analysis carried out at the phylum level, Bacteroidetes was the phyla that showed significant differences in subjects who remained infected in the albendazole arm compared to those who remained infected in the placebo arm. We dissected this further to assess which Bacteroidetes genera accounted for this difference using the Dirichlet-multinomial regression model on 6 bacterial categories which were obtained as follows. The phylum Bacteroidetes was divided into three categories, namely the *Bacteroides*, *Prevotella* and pooled Bacteroidetes. The first two genera were chosen as they were the two most abundant in the phylum Bacteroidetes. In the analyses, as 6 categories are needed, we included another three phyla, i.e., Actinobacteria, Firmicutes and pooled remaining phyla (pooled Phyla). As for the modelling at the phylum level, Firmicutes was used as a reference. Similar to the analyses at the phylum level, we characterized the association of infection and treatment on these 6 bacterial categories that comprised the genera belonging to Bacteroidetes.

Predictor			N	OR (95% CI)					*p*-values
(A)									
				Actinobacteria	Bacteroidetes	Proteobacteria	Unclassified	Pooled	
placebo			81	reference					
albendazole			69	1.15 (0.91,1.45)	0.97 (0.70,1.35)	1.18 (0.92,1.51)	0.94 (0.68,1.29)	1.11(0.78,1.58)	0.305
(B)									
placebo	Helminth(+)	group 4	34	reference					
	Helminth(-)	group 2	13	0.95 (0.62,1.46)	**0.49 (0.27,0.91)**	0.71 (0.45,1.10)	0.80 (0.45,1.43)	**0.47 (0.23,0.96)**	0.069
albendazole	Helminth(+)	group 3	13	**1.57 (1.05,2.35)**	**0.35 (0.18,0.70)**	1.01 (0.66,1.55)	0.76 (0.41,1.38)	0.75 (0.39,1.47)	**0.004**
	Helminth(-)	group 1	34	1.18 (0.54,2.57)	0.79 (0.24,2.54)	0.89 (0.40,1.10)	0.79 (0.27,2.36)	0.83 (0.25,2.74)	0.371

Table 2.2: **The association between each bacteria phylum with treatment and infection at post-treatment**. The estimated OR (odds ratio) and 95% CI (confidence interval) were obtained from the Dirichlet – multinomial regression and the *p*-values were obtained from the likelihood ratio test. Firmicutes is used as a reference for bacterial category. (A) The regression was fitted on all subjects irrespective of their infection status to assess the significant effect of treatment on microbiome composition. (B) The regression was fitted on subjects who were infected at baseline (N=94) to assess the significance of microbiome composition of each group compared to placebo infected (group 4). Bold represents the significant association between specific bacteria phylum with predictors. Bold represents the significant association. OR: odds ratio, CI: confidence interval.

Predictor			N	OR (95% CI)					*p*-values
(A)									
				Actinobacteria	Bacteroidetes			Pooled phyla	
					Bacteroides	*Prevotella*	Pooled Bacteroidetes		
placebo			81	reference					
albendazole			69	1.16 (0.90,1.49)	1.00 (0.61,1.62)	1.02 (0.69,1.50)	0.97 (0.65,1.45)	1.08 (0.85,1.37)	0.7
(B)									
placebo	helminth(+)	group 4	34	reference					
	helminth(-)	group 2	13	0.92 (0.59,1.43)	0.49 (0.18,1.34)	**0.44 (0.21,0.90)**	0.49 (0.24,1.03)	0.73 (0.48,1.10)	0.086
albendazole	helminth(+)	group 3	13	**1.54 (1.00,2.35)**	0.79 (0.32,1.97)	**0.44 (0.21,0.94)**	0.40 (0.18,0.89)	0.92 (0.61,1.40)	**0.016**
	helminth(-)	group 1	34	1.17 (0.51,2.67)	0.78 (0.15,4.13)	0.78 (0.21,2.93)	0.74 (0.18,3.00)	0.83 (0.38,1.80)	0.478

Table 2.3: **The association between combination of bacteria taxa with treatment and infection at post-treatment**. The Dirichlet – multinomial regression was fitted on 6 categories consists of three genera under Bacteroidetes phyla and three phyla (Actinobacteria, Firmicutes and Pooled phyla). Here, the Pooled phyla consists of Proteobacteria, Unclassified and Pooled category as in Table 2.2. Firmicutes is used as a reference. Similar as in Table 2.2, (A) the regression was fitted on all subjects irrespective of their infection status to assess the significance effect of treatment on the bacterial composition. (B) The regression was fitted on infected subjects at baseline (N=94) to assess the significance of each group compared to placebo infected (group 4). Bold represents the significant association. OR: odds ratio, CI: confidence interval.

When considering the whole study subjects irrespective of infection status, there was no difference between albendazole and placebo (Table 2.3A). When 94 infected subjects at pre-treatment were selected and 6 bacterial categories as above were analysed with regard to infection and treatment, we observed a decrease in odds of having *Prevotella* in subjects who lost their helminth infection in placebo group (OR 0.44, 95% CI: (0.21,0.90)) compared to subjects who remained infected in placebo group although this fell short of statistically significant (p-value of 0.086, Table 2.3B). Furthermore, in line with the finding at the phylum level, we also observed a significant difference in microbial composition of subjects who remained infected with albendazole compared to the microbial composition of subjects who remained infected in the placebo group (p-value of 0.016). This alteration was mainly due to the increase in odds of having Actinobacteria (OR 1.54, 95% CI: (1.00, 2.35)) and a decrease in odds of having *Prevotella* (OR 0.44, 95% CI: (0.21, 0.94)), suggesting that the decrease in Bacteroidetes at the phylum level observed in Table 2.2B was driven by *Prevotella*.

2.4 Discussion

There are two unique aspects to the current study on the effect of helminths on the gut microbiome in subjects living in rural areas of Indonesia, namely the combination of the study design and the statistical approach. The statistical parametric or nonparametric approaches are typically used to test the hypothesis whether the microbiome compositions are significantly different between groups [Cooper et al. (2013), Lee et al. (2014), Kay et al. (2015), and Ramanan et al. (2016)]. While the nonparametric approach suffers from lack of statistical power when the sample size is small [Whitley and Ball (2002)], available parametric approaches consider the abundance of each bacterial categories separately, hence requiring multiple testing corrections. The previous studies in Zimbabwe, Malaysia and Ecuador relating microbiome and helminths compared the difference of abundance of certain bacteria category between groups by using the standard or paired t-test and addressed multiple testing by Bonferonni corrections or False Discovery Rate [Cooper et al. (2013), Lee et al. (2014), and Kay et al. (2015)]. The clustering of bacteria has been investigated before using descriptive nonparametric approaches such as PCA or NMDS. When we applied these method to our genera data, no clustering was observed; neither by infection status nor by randomization arm. This might be an indication that PCA or NMDS were unable to capture the correlation between genera. We further analysed the multivariate data composed of 6 phyla (Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria, Unclassified bacteria and pooled category) simultaneously in relation to helminth infection status and treatment using a parametric approach. This multivariate approach takes into account the nature of metagenomics data, such as the abundance of all phyla forming the compositional structure and that these

abundances are known to vary highly between subjects [Li (2015)]. Our methods is able to quantify the relationship between the whole bacteria community with regard to the presence/absence of helminths or antihelminthic treatment while taking into account the correlational structure between bacterial categories imposed by the compositional nature. As bacterial categories are correlated, the decrease of one category should cause the increase of other categories and vice versa [Conlon and Bird (2014), Wu et al. (2011)]. Several microbiome studies have reported the change of the ratio Firmicutes to Bacteroidetes [de Filippo et al. (2010), Koliada et al. (2017)]. Thus, inference with regard to the decrease or increase of certain bacteria only makes sense when all bacterial categories are considered.

The reparameterization of Dirichlet – multinomial in the data analyses provided an interpretation in terms of odds ratios on how bacterial categories were affected by the helminth infection or treatment allocation. To obtain odds ratios, a reference category needs to be selected. In this study, we used Firmicutes as a reference due to its high abundance among bacterial categories as well as its presence in all samples. The high abundance of Firmicutes remained relatively stable, which had the advantage of allowing us to reveal subtle differences in other bacterial categories.

One potential limitation of our multivariate method is that the number of bacterial categories to be modelled was limited. As a consequence, taxa had to be pooled. Such a procedure assumes that the effect of the underlying taxa are captured in one single parameter. On the other hand, pooling can be viewed as a practical way to deal with sequencing error by providing a more robust model [Chen and Li (2013)]. Instead of pooling, one might use a shrinkage method as proposed by Chen and Li (2013) to deal with multiple rare taxa. As an alternative to biostatistical regression methods, machine learning methods are typically used for analysis of microbiome data. However, such methods require larger samples to allow the split into a training and a validation set. Our dataset is too small for such a method. Moreover, this method ignores the correlation structure, such as overdispersion.

It should be noted that the coverage depth in our study is relatively low (in average of 6000 reads per sample) as a result of using pyrosequencing platform (454) compared to more recent deep sequencing technologies (Illumina). We noted that two microbiome studies have reported similar average reads per samples as in our study [Cooper et al. (2013), Lin et al. (2013)]. As a consequence, rare taxa or taxa with low abundance might not be detected [Torbati et al. (2016)], and it is also possible that the similar diversity that we observed could be caused by the use of this platform. However, a direct comparison between Illumina MiSeq and the 454 platform has revealed that the limitation of the 454 is at the genus and family level, while at the higher taxonomic level (such as order, class and phylum level), the 454 platform is able to detect the same number of bacterial categories as the Illumina platform [Rosa et al. (2018)]. This could be considered as an advantage

of this approach allowing the analysis at the phylum level. Another unique aspect regarding our study was that a placebo-controlled anthelminthic trial design was used, while other studies were either observational or used an intervention without a placebo group. A control group that did not get the anthelminthic treatment (received placebo) has the advantage of controlling for confounders and estimating a direct treatment effect [Fisher (1925), Hall (2007)]. There were no significant differences in the microbiome composition, analyzed at the phylum level, of subjects with and without helminth infection at baseline, nor at the 21 months time point. One possibility is the low resolution of the bacterial data at phylum level. It is also possible that the similarity in microbiome composition between infected and uninfected subjects is due to infection history [Lee et al. (2014)]. Surprisingly, we observed a significant difference in the microbiome composition between placebo and albendazole-treated subjects at post-treatment in those who remained infected (Table 2.2B). This difference seemed to be represented by an increase in relative abundance of Actinobacteria and a decrease in relative abundance of Bacteroidetes. This difference in the relative abundance of Bacteroidetes was confirmed by comparing paired samples at pre and post-treatment in the albendazole group who were infected at baseline and remained infected at post-treatment. No significant difference in microbiome composition was found when comparing the albendazole and placebo arms in subjects who remained uninfected, or when comparing pre and post-treatment in those who received albendazole but remained uninfected. These data indicate that first of all, microbiome composition is stable over time and second, albendazole has no direct effect on microbiome composition. Together, our results suggest that the interplay between anthelminthic treatment and helminths in the gut has a complex effect on the microbiome composition. We observed that deworming is more effective against certain helminth species but not others. Indeed, *T. trichiura* infection was dominant after treatment in our study. This means that infected subjects who had received placebo harboured different helminth species than those who had received albendazole. However, at pre-treatment, there was no difference between the microbiome associated specifically with *T. trichiura*, *A. lumbricoides* or hookworm and therefore the effect of albendazole on the microbiome at post-treatment in infected subjects can not only be due to the dominance of *T. trichiura* but possibly the result of a combination of *Trichuris* and albendazole on the microbiome composition. It should be noted that in a recent study taking a different approach from us by using machine learning techniques, considering all taxonomic levels, and large sample size from not only Indonesia but also Liberia, differences in certain taxa were found to be worm-specific [Rosa et al. (2018)]. Therefore, to confirm whether *T. trichiura* has a different effect on microbiome composition after albendazole treatment compared to other helminth species, further and larger studies are needed. With regard to the treatment effect, a study in Malaysia reported the increasing abundance of *Bacteroidales* (an order of Bacteroidetes) and

the decreasing abundance of *Clostridiales* (an order of Firmicutes) after treatment [Ramanan et al. (2016)]. This result might be confounded as there was no control group to assess the treatment effect. Another interventional study was carried out in Zimbabwe, but it did not provide information on the effect of treatment in those who remained infected since the microbiome composition was only measured in subjects who completely cleared their helminths.

A longitudinal setting in microbiome studies has the advantage of analysing the microbiome composition at different time points in the same population. However, the studies using longitudinal approach differed in the length of follow-up time. The studies in Malaysia [Ramanan et al. (2016)] had a follow-up time of 21 days, the study in Zimbabwe [Kay et al. (2015)] examined the microbiome composition at 12 weeks after treatment while our study had the longest follow-up time of 21 months (with treatment given every three months). Thus, so far the previous studies have examined the effect of short term removal of helminths on microbiota [Ramanan et al. (2016)], while in our study, we used a longer follow-up time to ensure succesful and long lasting deworming of the subjects. Differences in study design and techniques used for collection and analysis of samples hamper comparison across studies.

The regression model used in this study is only applicable in a cross-sectional manner and assumes a simple correlation structure between bacterial categories. Such a method could be extended to more complex correlation structures. One is the correlation between bacterial categories or between the microbiome composition of the same subject measured at different time points. A statistical test for paired two categorical counts is available, however to model the change in microbiome composition over time we would need to extend our model. To conclude, the microbiome composition is likely to change due to interactions between helminth and anthelminthic treatment, but a direct impact of treatment on microbiome composition has not been observed. Larger studies are needed to dissect these effects of treatment and also to take into account the history of helminth infection. Furthermore, new statistical methods that allow longitudinal analysis of changes in the microbiome composition need to be developed.

2.5 Supplementary Materials

Supplementary materials are available online at the *PLoS Neglected Tropical Diseases* website.

Figure S1 Bacterial diversity in relation with helminth at pre and post-treatment.

Table S1 List of primers and probe sequences used in detecting the helminth species.

Table S2 The characteristics of participants of current study and total population in Nangapanda. The number of subject (n) of the total participants who were surveyed (N).

3

The mixed model for the analysis of a repeated-measurement multivariate count data

Abstract

Clustered overdispersed multivariate count data are challenging to model due to the presence of correlation within and between samples. Typically, the first source of correlation needs to be addressed but its quantification is of less interest. Here we focus on the correlation between time-points. In addition, the effects of covariates on the multivariate counts distribution need to be assessed. To fulfill these requirements, a regression model based on the Dirichlet-multinomial distribution for association between covariates and the categorical counts is extended by using random effects to deal with the additional clustering. This model is the Dirichlet - multinomial mixed regression model. Alternatively, a negative binomial regression mixed model can be deployed where the corresponding likelihood is conditioned on the total count. It appears that these two approaches

This chapter has been published as: Ivonne Martin, Hae-Won Uh, Taniawati Supali, Makedonka Mitreva, Jeanine J. Houwing-Duistermaat (2019). The mixed model for the analysis of a repeated measurement multivariate count data. *Statistics in Medicine*, 38(12): 2248 - 2268.

are equivalent when the total count is fixed and independent of the random effects. We consider both subject-specific and categorical-specific random effects. However, the latter has a larger computational burden when the number of categories increases. Our work is motivated by microbiome datasets obtained by sequencing of the amplicon of the bacterial 16S rRNA gene. These data have a compositional structure and are typically overdispersed. The microbiome dataset is from an epidemiological study carried out in a helminth-endemic area in Indonesia. The conclusions are: time has no statistically significant effect on microbiome composition, the correlation between subjects is statistically significant, and treatment has a significant effect on the microbiome composition only in infected subjects who remained infected.

3.1 Introduction

Microbiome data are overdispersed multivariate counts; for each sample, counts across multiple taxa are observed. If one is interested in the change of the microbiome composition over time, subjects are measured longitudinally [Ramanan et al. (2016)]. Such data are subject to two sources of correlation, namely the correlation between the counts of a sample and between multiple samples across time of a subject. For this type of data, the available statistical models are still limited.

The microbiome dataset considered in this paper is obtained by sequencing the amplicon of the bacterial 16S rRNA gene, where the sequencing procedure follows the HMP standardized protocol [HMP (2012)]. Chimeric sequences were filtered out and the resulting sequences are either categorized based on similarity into Operational Taxonomical Units (OTUs) followed by annotation, or directly annotated using relevant databases (e.g. Ribosomal Database Project, Greengenes or Silva). The counts for a specific category represent the abundances of the bacteria at a biological taxonomy level. Datasets generated through this sequencing process comprise features that have not been adequately accounted for by currently available statistical methods [Li (2015)]. Firstly, the dataset might be represented by a matrix of taxonomical counts with a compositional structure, which imposes a correlation between taxa [Gloor et al. (2017)]. Secondly, overdispersion might exist due to unobserved heterogeneity in the sampling procedure, the presence of taxa with rare abundance (zero-inflation), and pooling of categories. Another source might be differences in total sequence reads per sample, which might be caused by technical difficulties or by sampling or individual variability. This is commonly addressed by dividing the bacteria for each categories with the total count of the smallest reads (normalization), which results in a constant total bacterial count for all samples. Alternatively, an offset can be used in the model.

Our work is motivated by the microbiome measurements from an epidemio-

logical study carried out in a helminth endemic rural area in Indonesia [Martin et al. (2018)]. The primary research question of this study is to analyze the joint effect of helminth infections and albendazole treatment on the microbial composition comprising multiple bacterial taxa. It has been hypothesized that the presence of helminths is linked with the microbial dysbiosis. However, recent findings report inconsistencies, probably due to limitation in the study design [Ramanan et al. (2016); Cooper et al. (2013); Lee et al. (2014)]. For our study, the stool samples were collected and measured on a subset of subjects participating in a randomized placebo-controlled trial. Thus, we included the microbiome data from infected subjects who received placebo, which makes our study unique. The bacterial count and the helminth infection status were assessed in samples before and 21 months after the first treatment. Details of the study can be found elsewhere [Wiria et al. (2010)]. In a previous paper [Martin et al. (2018)], we identified an effect of treatment on the microbiome composition in subjects who were infected at baseline and at follow up. This relationship was studied in the post treatment samples, whereas the microbiome composition at baseline was not used. Here, we model all the available data simultaneously and hence need to address the correlation structure.

The objective of this paper is to develop a parametric model for the analysis of the overdispersed multivariate count data in the repeated measurement setting. To date, several statistical parametric methods for analysis of microbiome data are available, which take into account the features of the data such as overdispersion and the presence of rare taxa. One approach is to consider a univariate taxa of interest and model the association of this taxa with biological covariates. Several regression models for this simplified problem exist. Zero-inflated models or hurdle models have been proposed to deal with rare taxa [Xu et al. (2015)]. These models are also available for longitudinal studies. This approach however ignores the multivariate structure of the data. A second approach which considers the compositional feature of the microbiome data, models the multivariate count outcome across taxa by a multinomial distribution. To deal with overdispersion, the underlying parameters are assumed to follow the conjugate distribution [Chen and Li (2013)]. This formulation has an advantage that the marginal distribution has a closed form formula.

The correlation due to repeated measurements within the same person is often modelled by including a normally distributed random effect in the linear predictor, i.e., generalized linear mixed model. The overdispersion is typically accounted for by the conjugate distribution Chen and Li (2013); Zhang and Zhou (2017); Guimarães and Lindrooth (2007). Molenberghs et al. (2007, 2010) and Booth et al. (2003) introduced a combined model, where the conjugate distribution for the overdispersion is used and the correlation over time is modelled by normally distributed random effects, i.e., generalized linear mixed model. The authors only consider single categorical count data; hence these models cannot

be directly applied to our data, where we have to acknowledge the compositional feature. Therefore, in spirit of the combined model, we propose an extension of the Dirichlet - multiomial regression model with random effects to incorporate the correlation due to repeated measurements. We will use the reparameterization of the Guimarães and Lindrooth (2007), in which the overdispersion is a function of the covariates and the random effects.

This manuscript is organized as follows. In Section 3.2, we briefly describe the formulation of the loglinear model in the setting of multivariate count data and derive the likelihood of the multinomial distribution obtained by conditioning on the total count. We show the derivation of this method in the case where the count is overdispersed. The model is then extended to include the correlation due to repeated measurements over time. In Section 3.3, simulation studies are described to investigate the performance of the proposed methods and the results of the analyses of the motivating dataset are presented in Section 3.4. In Section 3.5, we conclude and discuss the proposed method.

3.2 Methods

A novel mixed model is considered for the relationship between counts of six phyla categories and the binary variables of infection status and treatment allocation before and after the first treatment round. Due to the normalization, the total count per sample is fixed at 2000 at each time point. Before introducing our new model, we will review various models for categorical count data in the cross-sectional setting: namely for independent count data (the loglinear and the multinomial logistic regression model), and for count data subject to overdispersion (the negative binomial and the Dirichlet-multinomial model) [Agresti (2013); Tutz (2012)].

We first introduce the following notations. Let $C_i^{(t)} = \left\{C_{i1}^{(t)}, \ldots, C_{iJ}^{(t)}\right\}$ be the J dimensional vector of the multivariate microbial count with $C_{ij}^{(t)}$ the abundance of bacteria taxa $j\,(j = 1, \ldots, J)$ for subject $i\,(i = 1, \ldots, N)$ at time point t. The total count for each subject i at time-point t is fixed and denoted as $C_{i+}^{(t)} = \sum_{j=1}^{J} C_{ij}^{(t)}$. Let P be the number of categorical covariates and $X_i^{(t)}$ be the P dimensional vector of covariate values for subject i at time point t. When modelling microbiome data as described above, either the sequence count itself can be considered, or the normalized count related to the total sequence read, i.e. compositional data. Multiple counts distributed over categories are usually represented by a contingency table. We briefly review models for the cross-sectional setting and therefore suppress the superscript t in the model formulation in Subsection 3.2.1.

3.2.1 Cross-sectional setting

The loglinear model for two categorical variables

The loglinear model is commonly used to model the association between multivariate categorical count data and predictors of categorical or continuous value. In the case where all variables are categorical, the data can be represented by a contingency table. Consider two categorical variables E and F, with J and K levels, respectively. The count outcome c_{jk} is associated with the jth level of predictor E and kth level of predictor F, which could be described in a $J \times K$ contingency table (Table 3.1) as follows.

		F		
	j \ k	**1**	…	***K***
E	**1**	c_{11}	…	c_{1K}
	2	c_{21}	…	c_{2K}
	⋮	⋮	⋮	⋮
	J	c_{J1}	…	c_{JK}
Marginal		c_{+1}	…	c_{+K}

Table 3.1: The $J \times K$ Contingency Table

Each cell's count outcome c_{jk} is assumed to follow a Poisson distribution with a mean μ_{jk}. Here, the saturated loglinear model for such contingency table is given by

$$\log\left(\mu_{jk}\right) = \lambda_0 + \lambda_j^E + \lambda_k^F + \lambda_{jk}^{EF}, \tag{3.1}$$

where λ_0, $\lambda_0 + \lambda_j^E$, $\lambda_0 + \lambda_k^F$, and $\lambda_0 + \lambda_k^F + \lambda_j^E + \lambda_{jk}^{EF}$ represent the overall mean, the marginal mean of categorical variable E at the jth level, the marginal mean of variable F at the kth level, and the mean when variables E and F taking the value j and k, respectively. Because there are $J \times K$ cells, the $J + K + JK + 1$ parameters of the saturated loglinear model (3.1) are not uniquely identifiable and thus constraints are needed to ensure the model identifiability. Two sets of constraints are commonly used, namely the baseline and the symmetrical constraint given by

$$\lambda_1^E = \lambda_1^F = \lambda_{j1}^{EF} = \lambda_{1k}^{EF} = 0$$

and

$$\sum_{j=1}^{J} \lambda_j^E = \sum_{k=1}^{K} \lambda_k^F = \sum_{j=1}^{J} \lambda_{j1}^{EF} = \sum_{k=1}^{K} \lambda_{1k}^{EF} = 0, \qquad \text{for all } j,k,$$

respectively. In this manuscript, we use the baseline constraint.

Note that in model (3.1) the response (bacterial categories) E, and predictor F, are exchangeable. The loglinear model (3.1) could be written in the regression format for the bacterial outcome as follows.

$$\log\left(\mu_{jk}\right) = \xi_{0j} + \xi_{1jk}[F = k], \quad j = 1,\ldots,J, \quad k = 1,\ldots,K,$$

with [.] the indicator function. To show the equivalence between two models, note the following j runs over the category and k runs over the predictor levels. For a subject with their predictor in category $k = 1$, the regression model $\{\xi_{01}, \xi_{02}, \ldots \xi_{0J}\}$ with ξ_{0j} for $j = 2,\ldots,J$ corresponds to $\lambda_0 + \lambda_j^E$. For subjects with their predictor in other categories k, the regression model $\{\xi_{01} + \xi_{11k}, \xi_{02} + \xi_{12k}, \ldots, \xi_{0J} + \xi_{1Jk}\}$ where ξ_{1jk} for $j = 2,\ldots J$ corresponds to $\lambda_k^F + \lambda_{jk}^{EF}$. Thus, in the context of regression, the λ_{jk}^{EF} represents the effect of the categorical variable F on outcome category j relative to the reference category.

To estimate the parameters, we assume that each cell's entry represents a realization from the Poisson distribution. The maximum likelihood estimate of $\boldsymbol{\lambda}$ or of $\boldsymbol{\xi}$ can be obtained by maximizing the following likelihood function. Specifically, for subject i, it is given by

$$\begin{aligned} L_i(\boldsymbol{\lambda}) &= \prod_j f_{\text{Pois}}\left(\boldsymbol{\lambda}; c_{ijk}\right) \\ &= \prod_j \frac{\exp\left(-\mu_{jk}\right)\mu_{jk}^{c_{ijk}}}{c_{ijk}!}, \end{aligned} \tag{3.2}$$

where person i belongs to category k and has counts in each bacteria category j. The model could be straightforwardly generalized to incorporate more categorical covariates which results into more than two-way contingency table. For instance, when incorporating the infection and treatment status we will have a three way contingency table. As before, the categorical variable E corresponds to the bacteria category, variable F to the treatment randomization arm and G to the infection status. The corresponding loglinear model can be written as follows

$$\begin{aligned} \log\left(\mu_{jkl}\right) &= \lambda_0 + \lambda_j^E + \lambda_k^F + \lambda_{jk}^{EF} + \lambda_l^G + \lambda_{jl}^{EG} + \lambda_{kl}^{FG} + \lambda_{jkl}^{EFG}, \quad j = 2,\ldots,J; k = l = 2 \\ &\quad \text{or in the regression format as} \\ &= \xi_{0j} + \xi_{1j}\text{Treatment} + \xi_{2j}\text{Infection} + \xi_{3j}\text{Treatment} \times \text{Infection}, \quad j = 1,\ldots,J. \end{aligned} \tag{3.3}$$

Here the baseline constraint is applied on the first equation, while for the second equation this is not needed since there are only $J \times P$ parameters. This last equation represents the loglinear model written in terms of regression coefficients $\boldsymbol{\xi}$ and covariate values, where Treatment and Infection are binary variables. To

assess the statistical significance of the pth covariate ($p = 1, \ldots, P$) on the multivariate count distribution, the null hypothesis $\boldsymbol{\xi}_p = \mathbf{0}$ should be tested. We will use the standard Likelihood Ratio Test which follows a χ^2 distribution with J degrees of freedom.

Multinomial logistic regression

In our data example, the total bacterial count is fixed to a constant for all samples. Under this constraint of a fixed total count, it is sufficient to model the counts for $J-1$ categories and $(J-1) \times P$ parameters are uniquely identified. Guimarães and Lindrooth (2007) showed that the distribution of the multivariate counts under the constraint that the total is a constant could be derived from the distribution of the unconstrained multivariate counts above by using the conditional log likelihood given the total count. When the counts in each category are independently Poisson distributed with mean μ_{jkl}, the total count c_{+kl} follows a Poisson distribution with mean $\sum_{j=1}^{J} \mu_{jkl} = \mu_{+kl}$. The distribution of the multivariate counts conditional on the total for each subject i is therefore given by

$$\begin{aligned}
\Pr\left(\boldsymbol{c}_i = \{c_{1kl}, \ldots, c_{Jkl}\} \,|c_{+kl}\right) &= \frac{\Pr\left(c_{1kl}, \ldots, c_{Jkl}, c_{+kl}\right)}{\Pr\left(c_{+kl}\right)} \\
&= \frac{\prod_{j=1}^{J} f_{\text{Pois}}\left(c_{jkl}; \mu_{jkl}\right)}{f_{\text{Pois}}\left(c_{+kl}; \mu_{+kl}\right)} = c_{+kl}! \prod_{j=1}^{J} \left(\frac{1}{c_{jkl}!}\right) \left(\frac{\mu_{jkl}}{\mu_{+kl}}\right)^{c_{jkl}} \\
&\sim \text{Multinomial}\left(c_{+kl}; \pi_{1kl}, \ldots, \pi_{Jkl}\right), \\
&\quad \text{where } \pi_{jkl} = \frac{\mu_{jkl}}{\mu_{+kl}}. \qquad (3.4)
\end{aligned}$$

Thus, under the baseline constraint and the constraint that the total count is fixed, the distribution of the multivariate count is equivalent to the multinomial distribution with parameter $\pi_j = \dfrac{\mu_j}{\mu_+}$. This model is the multinomial logistic regression model. Note that the parameters $\boldsymbol{\lambda}$ of the loglinear model (3.1) cancel out. In the multinomial logistic regression model, the parameters of the reference category are typically assumed to be equal to zero, although other constraints can be used as well.

Overdispersed count data

When the count data are overdispersed, the variance of the cell count is no longer equal to its expected value and the Poisson distribution cannot be used. A common approach to deal with overdispersion is to assume that the conditional mean

of the count outcome is a random variable following the conjugate distribution. Consider a count at category j and let $\exp(\eta_{ij})$ be the random effect for overdispersion following the Gamma distribution (conjugate for Poisson) with parameter θ. Guimarães and Lindrooth (2007) formulated the model for an overdispersed count outcome as follows:

$$\begin{aligned}
&C_{ij}|\exp(\eta_{ij}) \sim \text{Pois}(\widetilde{\mu}_{ij}), \quad j = 1,\ldots,J \\
&\widetilde{\mu}_{ij} = \exp(\eta_{ij})\,\mu_{ij}, \quad \text{where } \exp(\eta_{ij}) \sim \Gamma\left(\text{shape} = \theta^{-1}\mu_{ij}, \text{rate} = \theta^{-1}\mu_{ij}\right) \\
&\widetilde{\mu}_{ij} = \exp(\eta_{ij})\,\mu_{ij} \sim \Gamma\left(\text{shape} = \theta^{-1}\mu_{ij}, \text{rate} = \theta^{-1}\right).
\end{aligned}$$

Here, μ_{ij} corresponds to the mean of the count in the non-overdispersed model. Now the marginal distribution for the count at category j in person i, C_{ij} can be obtained by integrating out the random effect $\exp(\eta_{ij})$ as

$$\begin{aligned}
\Pr(C_{ij}) &= \int_0^\infty \Pr(C_{ij}|\exp(\eta_{ij}))\, g(\exp(\eta_{ij}))\, d\exp(\eta_{ij}) \\
&= \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{C_{ij}!\Gamma(\theta^{-1}\mu_{ij})} \left(\frac{1}{\theta^{-1}+1}\right)^{C_{ij}} \left(\frac{\theta^{-1}}{\theta^{-1}+1}\right)^{\theta^{-1}\mu_{ij}}.
\end{aligned}$$

This corresponds to a negative binomial distribution with parameters $\left(\theta^{-1}\mu_{ij}, \frac{\theta^{-1}}{1+\theta^{-1}}\right)$. By the properties of the negative binomial random variable, the total count for subject i also follows the negative binomial distribution

$$C_{i+} \sim \text{NB}\left(\theta^{-1}\mu_{i+}, \frac{\theta^{-1}}{1+\theta^{-1}}\right).$$

The likelihood for subject i in this setting is given by

$$\begin{aligned}
L_i(\theta, \boldsymbol{\lambda}) &= \prod_j f_{\text{NB}}(\boldsymbol{\lambda}, \theta; C_{ij}) \\
&= \prod_j \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{C_{ij}!\Gamma(\theta^{-1}\mu_{ij})} \left(\frac{1}{\theta^{-1}+1}\right)^{C_{ij}} \left(\frac{\theta^{-1}}{\theta^{-1}+1}\right)^{\theta^{-1}\mu_{ij}}. \qquad (3.5)
\end{aligned}$$

Note that in this setting, the parameter θ which models the overdispersion and the intercept λ_0 are both not identifiable. An often used solution is to absorb the overdispersion parameter into the grand mean λ_0, i.e. $\theta^{-1}\exp(\lambda_0) = \delta_0^{-1}$ Guimarães and Lindrooth (2007).

Overdispersed multinomial

We briefly review the overdispersed count data introduced by Guimarães and Lindrooth (2007) as follows. To guarantee that the parameters of the count for

each category follows a Gamma distribution with the same rate parameter, the overdispersion parameter $\exp(\eta_{ij})$ needs to be a function of the linear predictor μ_{ij}. For such a distribution, Theorem 1 of Mosimann (1962) can be applied. This theorem states that if $\boldsymbol{C}_i = \{C_{i1}, C_{i2}, \ldots, C_{iJ}\}$ are independently Gamma distributed random variables with parameters $(\widetilde{\mu}_{i1}, \widetilde{\mu}_{i2}, \ldots, \widetilde{\mu}_{iJ})$ with the same scale parameter θ^{-1}, then the random variables $\boldsymbol{\Pi}_i = \{\Pi_{i1}, \Pi_{i2}, \ldots, \Pi_{iJ}\}$ with $\Pi_{ij} = \dfrac{C_{ij}}{\sum_{j=1}^{J} C_{ij}}$ have a multivariate beta distribution (Dirichlet distribution) with parameters $\{\widetilde{\mu}_{i1}, \widetilde{\mu}_{i2}, \ldots, \widetilde{\mu}_{iJ}\}$. Note that the Dirichlet distribution is the conjugate for the multinomial distribution. Hence, the marginal distribution for the random variable $\boldsymbol{\Pi}_i$ is obtained by integrating out the Dirichlet random effects. Now, the corresponding Dirichlet - multinomial distribution is given by

$$\Pr(\boldsymbol{\Pi}_i) = \frac{\Gamma(\widetilde{\mu}_{i+})\, C_{i+}!}{\Gamma(\widetilde{\mu}_{i+} + C_{i+})} \prod_{j=1}^{J} \frac{\Gamma(\widetilde{\mu}_{ij} + C_{ij})}{\Gamma(\widetilde{\mu}_{ij})\, C_{ij}!}. \tag{3.6}$$

Alternatively, we consider the conditional likelihood of the multivariate negative binomial given the total count. The contribution for the ith subject is given by

$$\begin{aligned} L_i(\boldsymbol{\lambda}, \theta) = \Pr(\boldsymbol{C}_i | C_{i+}) &= \frac{\prod_{j=1}^{J} f_{\mathrm{NB}}(C_{ij}; \widetilde{\mu}_{ij})}{f_{\mathrm{NB}}(C_{i+}; \widetilde{\mu}_{i+})} \\ &= \frac{\Gamma\left(\theta^{-1}\mu_{i+}\right) C_{i+}!}{\Gamma(\theta^{-1}\mu_{i+} + C_{i+})} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{\Gamma(\theta^{-1}\mu_{ij})\, C_{ij}!}. \end{aligned} \tag{3.7}$$

By $\widetilde{\mu}_{ij} = \theta^{-1}\mu_{ij}$, it follows that the likelihood (3.7) is equivalent to the the Dirichlet-multinomial distribution (3.6). Here, the parameter θ is unidentifiable. Similar to (3.5), we apply the parameterization in Guimarães and Lindrooth (2007) where the overdispersion is absorbed in the grand mean λ_0 such that $\theta^{-1}\exp(\lambda_0) = \delta_0^{-1}$ in the reference category. In contrast to the non-overdispersed multinomial model, the intercepts of the overdispersed multinomial model do not cancel out.

3.2.2 Repeated measurement of overdispersed count

In addition to the overdispersion due to the presence of multiple bacteria within one sample, there is also correlation between measurements of the same person at the two time-points, i.e. at the pre- and post-treatment. To deal with this correlation, we propose to include a random effect u_i in the linear predictor of the model and assume that conditional on this random effect the observations of the two time points are independent. We further assume that the random effect u_i follows a normal distribution with zero mean and variance σ_u^2. The idea of using different distributions for the random effects representing overdispersion and

correlation was introduced by Molenberghs et al. (2007, 2010) and Booth et al. (2003). Molenberghs and Booth modelled the mean of an outcome as a multiplication of overdispersion and the linear predictor. However, to guarantee that the Theorem 1 of Mossimann holds, i.e. that the proportion of each bacterial category has Dirichlet-multinomial distribution, we need to model the overdispersion as a function of the linear predictor.

In the rest of this section, we describe three different mixed models for multivariate count data with overdispersion in the repeated measurement setting using random effects: conditional on the random effect u_i, the counts follow the multivariate negative binomial distribution; the counts follow the conditional multivariate negative binomial distribution given the total count; the proportions (cell's count divided by total count) follow the Dirichlet-multinomial distribution. In all models, we will add the random effect u_i to the linear predictor. These models are therefore extensions of the models for overdispersed multivariate count given in Subsection 3.2.1. Specifically for the first model, we assume that conditional on the random effects $\exp\left(\eta_{ij}^{(t)}\right)$ and u_i, the count $C_{ij}^{(t)}$ follows a Poisson distribution with mean equal to

$$\begin{aligned}
\mathrm{E}\left[C_{ij}^{(t)} \,|\, \exp\left(\eta_{ij}^{(t)}\right), u_i\right] &= \exp\left(\eta_{ij}^{(t)}\right) \exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \\
\text{where } \widetilde{\mu}_{ij}^{(t)} &= \boldsymbol{X}_i \boldsymbol{\xi}_j + u_i, \quad j = 1, \ldots, J. \\
\exp\left(\eta_{ij}^{(t)}\right) &\sim \Gamma\left(\text{shape} = \theta^{-1} \exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \text{rate} = \theta^{-1} \exp\left(\boldsymbol{X}_i \boldsymbol{\xi}_j + u_i\right)\right) \qquad (3.8)
\end{aligned}$$

Thus, given the random effect u_i, the two vectors of counts $C_i^{(t)}$ for $t = 1$ and $t = 2$ are independently distributed and follow the negative binomial distribution. The corresponding likelihood can be written as follows

$$\begin{aligned}
L_{\text{UNBM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) = \prod_i \Pr\left(\boldsymbol{C}_i^{(t)}\right) &= \prod_i \int_{u_i} \Pr\left(C_{i1}^{(t)}, \ldots, C_{iJ}^{(t)}, u_i\right) d\, u_i \\
&= \prod_i \int_{u_i} \prod_{t=1}^{2} \prod_{j=1}^{J} \Pr\left(C_{ij}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i \qquad (3.9)
\end{aligned}$$

and we denote the regression model under this likelihood to be the unconstrained negative-binomial mixed model (UNBM).

For the second approach, we consider the counts follow the conditional multivariate distribution given the total count. When each categorical count conditional on the total count follows the negative binomial with the same rate parameter, the total count $C_{i+}^{(t)} | u_i$ follows the negative binomial distribution with

parameters $\left(\sum_{j=1}^{J}\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \frac{\theta^{-1}}{1+\theta^{-1}}\right)$. Thus, the corresponding conditional likelihood is given by

$$\begin{aligned} L_{\text{CNBM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) &= \prod_i \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) \\ &= \prod_i \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)} | u_i\right) \Pr\left(u_i\right) d\, u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\, u_i} \\ &= \prod_i \frac{\int_{u_i} \prod_{t=1}^{2} \prod_{j=1}^{2} \Pr\left(C_{ij}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i}{\int_{u_i} \prod_{t=1}^{2} \Pr\left(C_{i+}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i}. \end{aligned} \tag{3.10}$$

The model corresponding to this likelihood is denoted as the conditional negative-binomial mixed model (CNBM). However, when the total counts depends on u_i the total count should be a random variable. This is not the case in our dataset. Therefore, we propose the third method with the assumption that the total count is independent of u_i.

In the third approach, we model the multivariate counts in terms of the relative abundance. We assume that the vector of proportions $\Pi_i^{(t)}$ conditional on the random effect u_i follows the Dirichlet multinomial distribution, i.e.

$$\begin{aligned} \left\{\frac{C_{i1}^{(t)}}{C_{i+}^{(t)}}, \ldots, \frac{C_{i1}^{(t)}}{C_{i+}^{(t)}}\right\} | \left\{\alpha_{i1}^{(t)}, \ldots, \alpha_{ij}^{(t)}\right\}, u_i &\sim \text{Mult}\left(\tilde{\pi}_{i1}^{(t)}, \ldots, \tilde{\pi}_{iJ}^{(t)}\right) \\ \left\{\tilde{\pi}_{i1}^{(t)}, \ldots, \tilde{\pi}_{iJ}^{(t)}\right\} &\sim \text{Dir}\left(\alpha_{i1}^{(t)}, \ldots, \alpha_{iJ}^{(t)}\right) \\ \alpha_{ij}^{(t)} &= \theta^{-1}\mu_{ij}^{(t)} \end{aligned} \tag{3.11}$$

where the $\mu_{ij}^{(t)}$ is the linear predictor as in the loglinear model for the Poisson count. With this parameterization, the expected multinomial parameter becomes

$$\widetilde{\pi}_{ij}^{(t)} = \frac{\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right)}{\sum_{j=1}^{J}\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right)}.$$

The likelihood for each subject i is then formulated as follows

$$
\begin{aligned}
L_{\mathrm{DMM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) &= \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}\right) \\
&= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}} | u_i\right) \Pr\left(u_i\right) d\, u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \frac{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)}\right) C_{i+}^{(t)}!}{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)} + C_{i+}^{(t)}\right)} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)} C_{ij}^{(t)}\right)}{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)}\right) C_{ij}^{(t)}!} \Pr\left(u_i\right) d\, u_i. \quad (3.12)
\end{aligned}
$$

The corresponding regression model under this likelihood is denoted as the Dirichlet - multinomial mixed model (DMM). It is shown in the Appendix A, that in the case where the total count does not depend on the random effect u_i, the likelihoods (3.10) and (3.12) are equivalent.

The variance of the random effect u (σ_u^2) represents the correlation between the samples of the same subject across time. However, this value is hard to interpret and the marginal correlation between categorical count outcomes might be more interesting. This correlation is given by

$$
\mathrm{Corr}\left(C_{ij}{}^{(t)}, C_{ij*}^{(t^*)}\right) = \frac{\sigma_{C_{ij}{}^{(t)}, C_{ij*}^{(t^*)}}}{\sqrt{\sigma^2_{C_{ij}^{(t)}} \cdot \sigma^2_{C_{ij*}^{(t(*))}}}}.
$$

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

The program language R is used for all the computations except for data application with categorical-specific random effects. When maximizing the likelihoods the integrals are approximated by the adaptive Gauss-Hermite quadrature method [Liu and Pierce (1994)], and we used the functions available in the ecoreg package [Jackson et al. (2008)] to compute the integral. R implementations are available in github (https://github.com/IvonneMartin/CombinedMultinomial)

3.2.3 The categorical-specific random effect

In the above parameterization, we assume that the subject-specific effect u_i is univariate and is the same for all bacteria categories and time-points. Alternatively, a J dimensional vector of random effects can be used. Equation (3.8) now becomes

$$
\begin{aligned}
\widetilde{\mu}_{ij}^{(t)} &= \boldsymbol{X}_i \boldsymbol{\xi}_j + u_{ij}, \quad j = 1, \ldots, J. \\
u_{ij} &\sim \mathrm{MVN}\left(\mathbf{0}, \Delta_{J\times J}\right) \quad (3.13)
\end{aligned}
$$

Here, each bacterial category has its own realization of the random effect and the random effects solely model correlation between the categories over time. The vector $\boldsymbol{u}_i$ of length J follows a multivariate normal distribution with a J by J diagonal variance matrix Δ with σ_j^2 as diagonal elements. In addition to the general model (3.13), we consider a model with common variance $\sigma_j^2 = \sigma_u^2$, for all j to reduce the parameter space. Since the overdispersion already takes care of the correlation among the categories, this model might be better interpretable. However, a drawback of this model is that computation of the likelihood function involves an intractable J dimensional integral.

3.3 Simulation study

3.3.1 Simulation setting

Three sets of simulation studies were conducted to evaluate the performance of the proposed methods. With regard to estimation of the fixed effect parameters and variance components, we first investigated the performance of the DMM models for a subject- and categorical-specific random effects. We reported the bias and MSE as well as the sensitivity and specificity for these parameters. The sensitivity and the specificity of the likelihood ratio test statistics were computed for the following pairs of hypotheses (for fixed and variance of random effect, respectively).

$$\begin{aligned} &\mathrm{H}_0 : \boldsymbol{\xi}_p = \mathbf{0} \quad \text{vs} \quad \mathrm{H}_1 : \text{at least one of } \boldsymbol{\xi}_p \neq 0, \\ &\mathrm{H}_0 : \sigma_u^2 = 0 \quad \text{vs} \quad \mathrm{H}_1 : \sigma_u^2 > 0. \end{aligned}$$

In the second set, we want to estimate the marginal correlation given the distribution of the random effect. The purpose of this study is to verify whether the marginal intraclass correlation observed in our motivating dataset can be represented by our models (UNBM and DMM). For this purpose, we vary the standard deviation of the random effect and we used 10,000 Monte-Carlo simulation for estimating the marginal intraclass correlation.

In the third set, we aimed to study the robustness of the parameter estimates by fitting the DMM models when the true model is UNBM. For this purpose, we generated datasets with three categories from the UNBM model.

Dataset generation

To reduce the computational burden, datasets with only three categories at two different time-points t were considered. The total count per sample S was 25, 50 or 2000, and the number of samples N was 150 or 500. Two sets of parameters were

used, namely $\boldsymbol{\lambda}$ was fixed at $\{\lambda_2^F, \lambda_2^E, \lambda_3^E, \lambda_{22}^{EF}, \lambda_{32}^{EF}\} = \{0.5, -1, 0.1, 0.8, -2\}$ as well as the parameters from the dataset (results are given in Supporting Information Table S1). To increase the power, the parameter values of the first set are relatively larger. Note that the parameter λ_0 is fixed at zero to guarantee identifiability of the overdispersion parameter. The overdispersion parameter was fixed at $\theta = 0.1$. For the standard deviation of the random effects, we considered values σ_u of 0.5, 0.8 and 1.

Specifically, for the Dirichlet-multinomial mixed (DMM) model with a univariate random effect, multivariate counts were generated as follows.

1. For each subject $i, i = 1, \ldots, N$, we randomly generate binary covariates X_i^t for each time point t and a random effect $u_i \sim N\left(0, \sigma_u^2\right)$.

2. The mean for each category j is computed as $\tilde{\mu}_{ij}^{(t)} = \theta^{-1} \exp\left(\boldsymbol{\lambda} + u_i\right)$ where the $\boldsymbol{\lambda}$ correspond to ξ.

3. A multivariate count with mean $\tilde{\mu}_{ij}^{(t)}$ is generated.

For the DMM model with multivariate random effect, a similar procedure was used except that the random effects in step (1) are now generated from the multivariate normal distribution with a diagonal covariance matrix Σ. We considered three sets of values for the standard deviations of random effects, namely $\boldsymbol{\sigma}_u = \left(\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3}\right)$ is (0.5,0.6,0.5), (0.8,0.9,0.8) or (1,0.9,1).

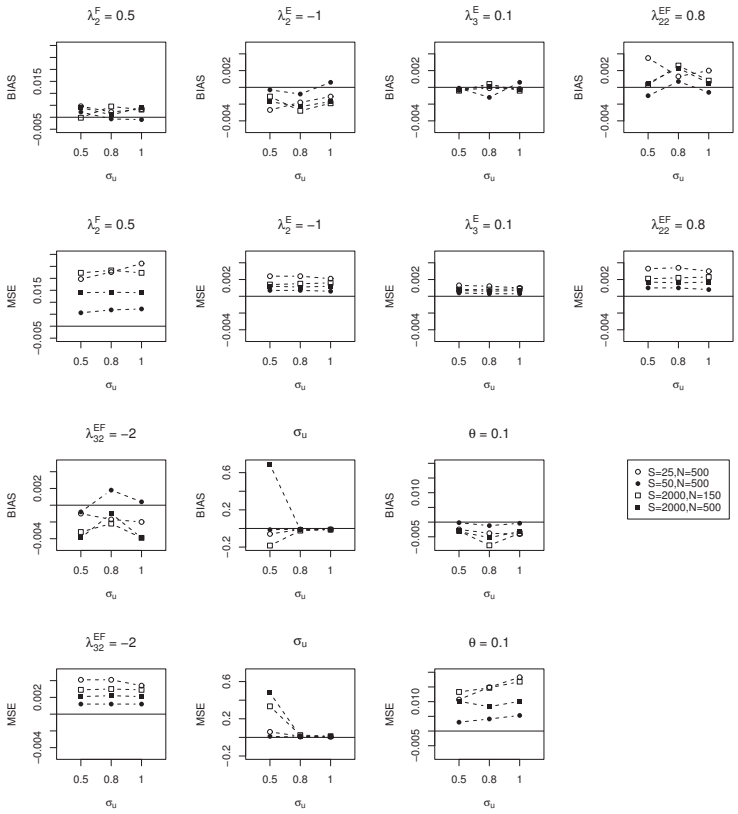
For the second set of simulation, 6 bacterial categories are used and parameters for the simulation are obtained from the dataset. Finally for the unconstrained negative binomial mixed (UNBM) model, the second step was replaced by computation of the expected count outcome for each category j of $\tilde{\mu}_{ij} = \theta^{-1} \exp\left(\log(S) + \boldsymbol{\lambda} + u_i\right)$. Here the offset $\log(S)$ is incorporated to take into account the total bacteria count S. For each scenario mentioned above, 1000 replicates were generated. The models were fitted to each of the replicates.

3.3.2 Simulation results

Evaluation of DMM model

The performance of the method in estimating the parameters is described in Figure 3.1. Overall, the bias and MSE appears to be improved when either the total bacterial count (from $S = 25$ to $S = 50$ and the sample size was $N = 500$), or the sample size was increased (from $N = 150$ to $N = 500$ and the total count was $S = 2000$). For small value of σ_u, both the bias and the MSE of this estimate are relatively large. Similar results are obtained for the model with categorical-specific random effects (Figure S1). The sensitivity of the likelihood ratio test for the fixed effects parameters that are obtained from the dataset are very low for all scenarios except when the total sample size is large (Table S2A). For testing the zero

variance component, the likelihood ratio test has a high sensitivity and specifity when the sample size and variance component are large (Table S2B).



$\boldsymbol{\lambda}$: a vector of parameters in loglinear model.

σ_u: the standard deviation of the between individual variation.

θ : the overdispersion.

Figure 3.1: Bias and MSE of datasets generated from the DMM model with subject-specific random effect.

	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-1.309	$\log(\theta)$ = -2.302	Loglik
Subj-sp	0.418(0.130)	-0.959(0.059)	0.096(0.050)	0.765(0.071)	-1.900(0.075)	-1.680(0.754)	-1.886(0.094)	-3918.581(18.333)
Cat-sp	0.438(0.172)	-1.007(0.116)	0.108(0.109)	0.809(0.084)	-2.017(0.086)	-0.704(0.005)	-2.366(0.125)	-3961.971(14.865)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-0.693	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.359(0.130)	-0.882(0.080)	0.096(0.077)	0.695(0.088)	-1.739(0.095)	-0.973(0.334)	-1.320(0.099)	-4064.461(18.503)
Cat-sp	0.462(0.170)	-1.000(0.122)	0.110(0.117)	0.794(0.085)	-1.997(0.083)	-0.607(0.028)	-2.253(0.129)	-3988.959(16.392)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-0.223	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.300(0.132)	-0.766(0.099)	0.092(0.105)	0.602(0.11)	-1.509(0.121)	-0.766(0.196)	-0.698(0.112)	-4171.966(17.482)
Cat-sp	0.455(0.194)	-1.004(0.173)	0.099(0.17)	0.795(0.089)	-1.985(0.096)	-0.237(0.049)	-2.235(0.165)	-4011.262(19.136)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u) = 0$	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.270(0.129)	-0.691(0.112)	0.092(0.117)	0.541(0.122)	-1.376(0.133)	-0.699(0.187)	-0.367(0.117)	-4193.591(19.342)
Cat-sp	0.449(0.200)	-1.003(0.213)	0.100(0.197)	0.795(0.088)	-1.985(0.101)	-0.020(0.046)	-2.225(0.177)	-3993.517(23.146)

Each rows started with Sub-sp represents the estimates (standard deviation) when datasets were fitted with the DMM model with subject-specific random effect and rows started with Cat-sp represents the estimates (standard deviation) when datasets were fitted with DMM model with categorical-specific random effect having common variance.
$\boldsymbol{\lambda}, \sigma_u, \theta$ are as explained in Figure 3.1.
Loglik represents the loglikelihood value obtained using the corresponding model.
Rows in gray represent the estimation when the standard deviation of the normally distributed random effect is small.

Table 3.2: The mean estimates (standard deviation) over 1000 replicates when datasets were generated from the DMM model with categorical-specific random effect with common variance.

Since the model with the categorical-specific random effect is time consuming to fit, we also investigate the robustness of assuming a subject-specific random effect while the datasets were generated by using a vector of random effects following the multivariate normal distribution. The results are given in Table 3.2. It appears that for a random effect with smaller standard deviation ($\log(\sigma_u)$ of -1.309), the biases of the estimates of fixed effect parameters and of $\log(\sigma_u)$ are relatively small, while for a random effect with larger standard deviation $\log(\sigma_u) = 0$ (σ_u of 1) the biases are relatively large.

In Table S3, the marginal correlations are given for the subject-specific random effects. It appears that the correlation between categories are all negative and the correlation between samples across time are very small. These results are not affected by the standard deviation of the random effect for our considered values. Table S4 lists the marginal correlations using categorical-specific random effects where each category-specific random effect has the same standard deviations σ_u. We notice that a part of the correlations between categories is now positive and the correlation between the same categories across time are larger. Moreover, these correlations tend to increase with a larger variance of the random effects.

Simulations under the UNBM model

The marginal correlations for the UNBM with a subject-specific random effect are listed in Table S5. It appears that the correlations between categories are positive as well as negative. The correlations of the same category between time points are all positive and increase with σ_u. A similar result is observed for the UNBM model with categorical-specific random effects (Table S6) although here the correlation varies more across categories.

Next, we investigated the robustness of the models. Datasets were generated using the multivariate negative binomial mixed model without conditioning on the total count (UNBM model). The results of fitting the unconditional multivariate negative binomial mixed model (UNBM), the multivariate negative binomial mixed model conditional on the total (CNBM) and the Dirichlet-multinomial mixed model (DMM) are given in Figure 3.2 for the fixed effect parameters and Figure 3.3 for the variance component.

In general, the fixed effect parameters obtained from these three different models are unbiased except the estimates of the intercepts (λ_2^F) for the CNBM model and the DMM model. Since the model used for analysis and generating the data are the same, the estimates of the fixed effect parameters in Figure 3.2 are unbiased and the variance of the estimator decreases when the total count was increased (from $S = 25$ to $S = 50$) or the sample size is increased (from $N = 150$ to $N = 500$). When using the conditional distribution given the total, the estimates of the fixed effect parameters in Figure 3.2 are biased when the total bacterial count is small ($S = 25$ and $S = 50$). When the total count is relatively large ($S = 2000$), the estimates of the fixed effects (including the intercept λ_2^F) are less biased. When

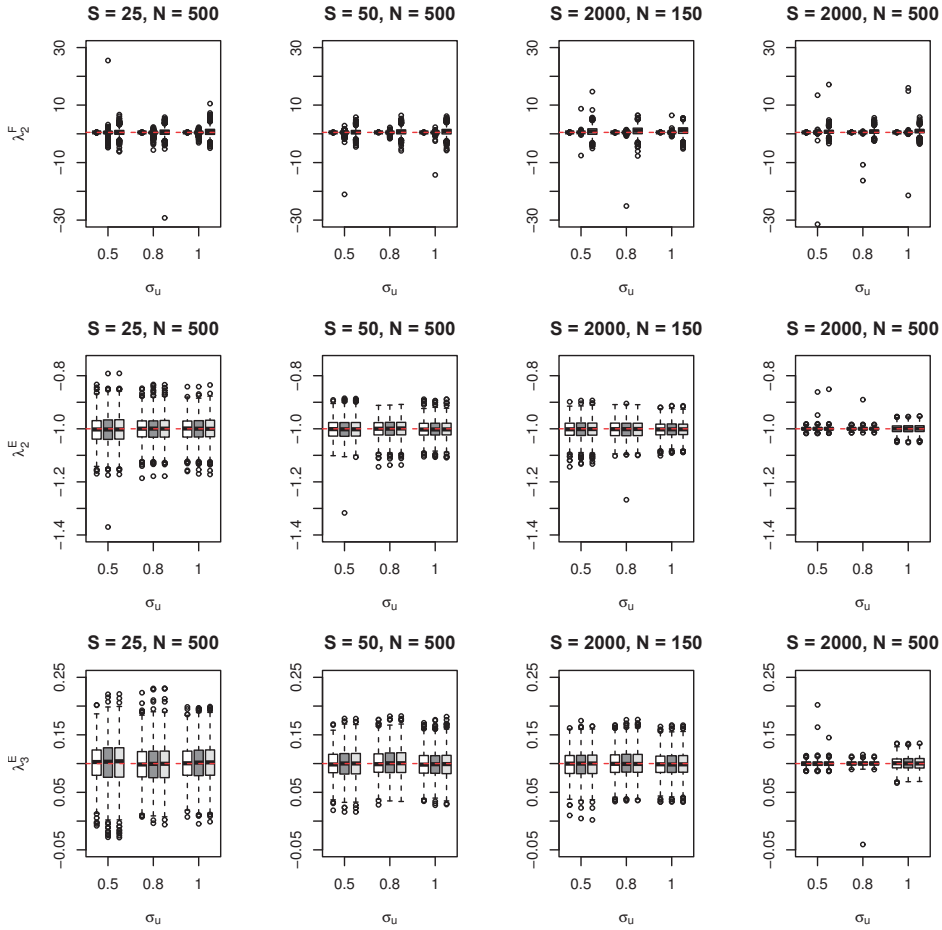


Figure 3.2: Estimates of the fixed effect and overdispersion parameters obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model. (first part)

estimating the fixed effect parameters using the DMM model, the estimate of the fixed effects are unbiased except for the intercept term λ_2^F and increasing the sample size does not improve the estimation.

The estimates of the random effect parameters in the UNBM model are unbiased and by increasing the total bacterial count or the sample size improves the precision. In the CNBM model, when the total bacterial count is small ($S = 25$ and $S = 50$), we observe that the standard deviation of u_i is overestimated and that the bias in the estimate of the overdispersion parameter is small. When the total count is large $S = 2000$, the estimate of the standard deviation of u_i appears

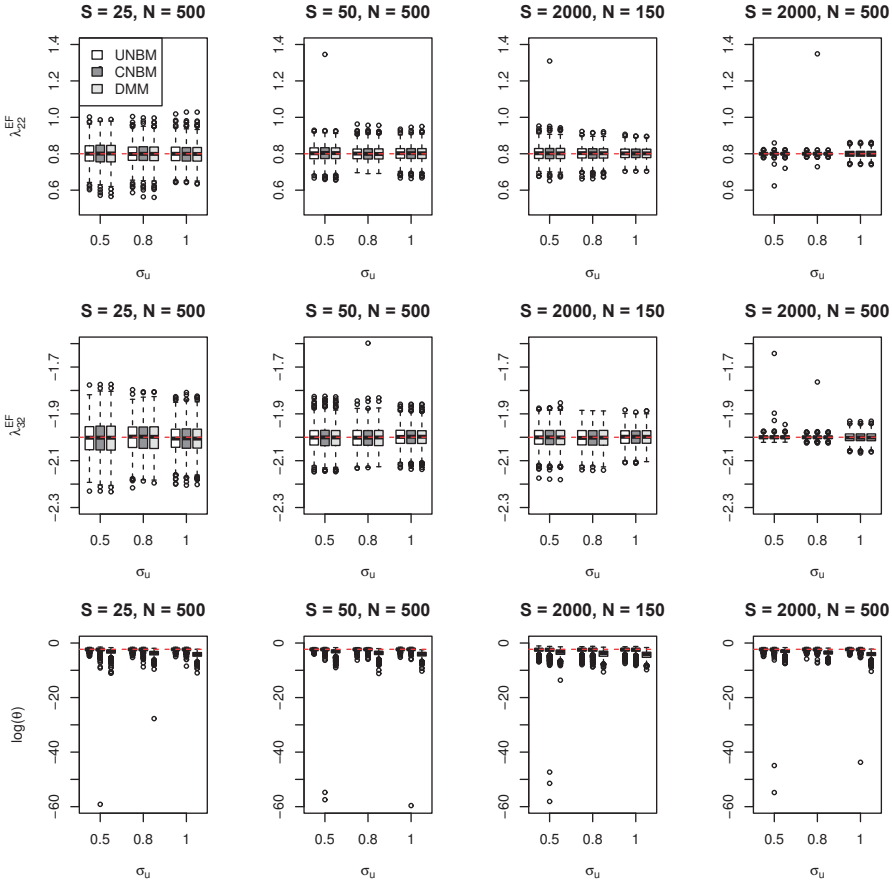


Figure 3.2: Estimates of the fixed effect and overdispersion parameters obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model. (cont.)

to be less biased while the overdispersion parameters is underestimated. When fitting the DMM model to the data, the estimates of the random effect parameters are biased in all scenarios.

3.4 Data Application

We used the DMM models to analyze the effect of helminth infections and treatment on microbiome composition. For this purpose, we first consider the fixed effect structure and fitted several DMM models to our dataset assuming (com-

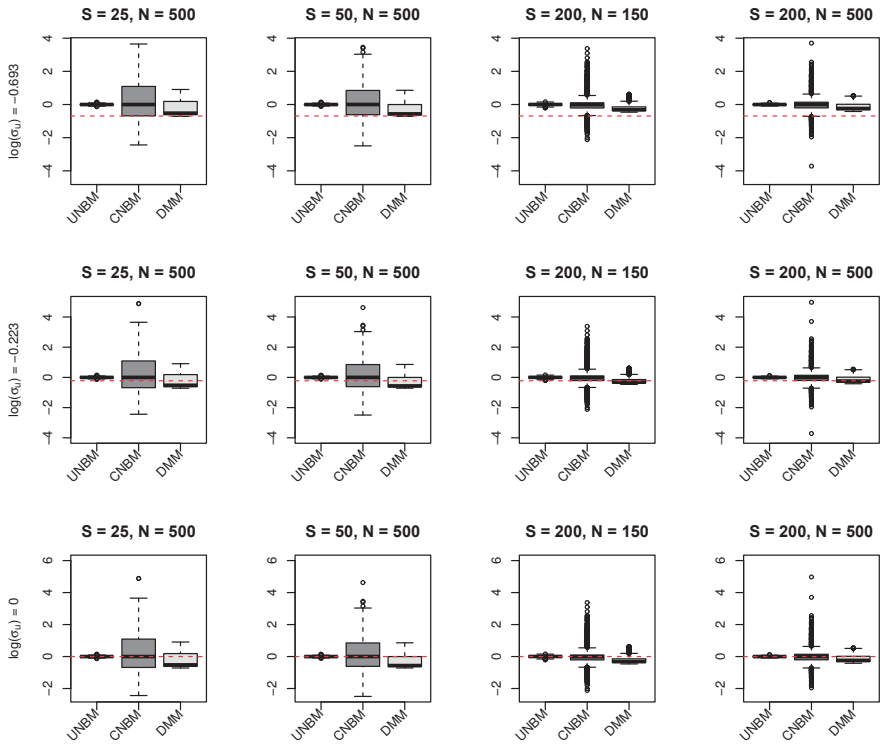


Figure 3.3: Estimates for the variance components obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model.

mon) random effect for each category. Next, we will investigate the best random effect structure and we will verify whether the parameter estimates of the fixed effects are affected by the random effect structure.

The microbiome dataset considered here was measured in a subset of a randomized clinical trial performed in a helminth-endemic area in Nangapanda subdistrict, Indonesia, described elsewhere [Wiria et al. (2010)] and is publicly available at Nematode.net (http://nematode.net/Data/Indonesia_16S/S1_Table.xlsx). In brief, households were randomized to receive either a single dose of 400 mg albendazole or placebo, once every three months for a period of one and a half years. To assess the effect of treatment on the prevalence of soil transmitted helminth infections, yearly stool samples were collected on a voluntary basis. *T. trichiura* infection was detected by microscopy and a multiplex real time PCR was used to detect the DNA of hookworm (*Ancylostoma duodenale* or *Necator americanus*) and *Ascaris lumbricoides*. A subject was regarded as infected if it was in-

fected with at least one helminth species.

For the current study, paired DNA samples before and at 21 months after the first treatment round from 150 inhabitants in Nangapanda were selected based on the treatment allocation and infection status, as well as the availability of complete stool data at pre- and post-treatment. The procedure for sample collection and processing was already described in Wiria et al. (2010). The 16s rRNA gene from the stool samples were processed through the 454 pyrosequencing technique, and the classification of the sequence resulted in counts of 18 bacterial phyla. For the current analyses, we retained the 5 most prevalent phyla and pooled the remaining into one category, resulting in six phyla categories: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, unclassified, and pooled category.

The description of relative abundance of each bacterial phyla at each time points are given in Table S7. *Firmicutes* has the highest relative abundance at each time points (around 68%), followed by *Actinobacteria* (around 12%), *Proteobacteria* (around 10%), *Bacteroidetes* (around 6 %) and Unclassified and pooled category (each around 1%). The dispersions are estimated by the ratio between the variance and mean. All bacteria counts show dispersion larger than 1 indicating the presence of overdispersion. Since zero-inflation might lead to overdispersion, we investigated the number of the samples with zero counts for the six categories at the two time points. Only for the following three categories, a small number of samples with zero counts was observed: *Bacteroidetes* (5 samples at post-treatment), Unclassified bacteria (1 at pre-treatment and post-treatment), and the pooled category (15 at pre-treatment and 6 at post-treatment). The corresponding histograms can be found in Figure S2. From this, we conclude that zero-inflation is not present, hence the overdispersion is probably caused by other sources. We will therefore account for overdispersion by additional random effects.

Table 3.3 gives the observed correlations between categories and of categories between time points. The order j for $C_j^{(t)}$ are *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, Unclassified and pooled category. The observed correlations between *Firmicutes* and the three most abundant bacteria (*Actinobacteria*, *Proteobacteria* and *Bacteroidetes*) are relatively high and negative (around -0.50), indicating an increase of *Firmicutes* corresponds to the decrease of these bacterial categories. These correlations are relatively similar for both time points, except for the correlation between *Firmicutes* and *Actinobacteria* which becomes smaller at the second time point (-0.27). The correlations between *Firmicutes* and Unclassified, and the pooled category, are relatively small. The intraclass correlations of bacterial categories between the two time points are always positive. *Firmicutes* and *Actinobacteria* show the highest correlation between two time points (0.14 and 0.17).

The baseline characteristics of the study participants were given in Table 3.4. In each of the randomization arms, there are four possible combinations of infection status at pre- and post-treatment. Namely, uninfected subjects who either

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	1	-0.46	-0.43	-0.48	-0.12	-0.23						
$C_2^{(1)}$	·	1	-0.29	0.13	0.02	0						
$C_3^{(1)}$	·	·	1	-0.27	-0.19	0						
$C_4^{(1)}$	·	·	·	1	0.1	0.06						
$C_5^{(1)}$	·	·	·	·	1	0.01						
$C_6^{(1)}$	·	·	·	·	·	1						
$C_1^{(2)}$	**0.14**	-0.11	-0.05	-0.01	0	-0.13	1	-0.27	-0.53	-0.57	0.04	-0.14
$C_2^{(2)}$	-0.14	**0.17**	0.04	0.03	-0.01	-0.05	·	1	-0.27	-0.15	-0.05	0.01
$C_3^{(2)}$	0.04	0.05	**0.01**	-0.07	-0.08	-0.1	·	·	1	-0.07	-0.22	-0.11
$C_4^{(2)}$	-0.11	-0.02	0.01	**0.07**	0.05	0.3	·	·	·	1	0.02	0.09
$C_5^{(2)}$	0.06	-0.25	0.09	0.01	**0.05**	0.01	·	·	·	·	1	-0.05
$C_6^{(2)}$	-0.07	0.08	-0.06	-0.01	0.23	**0.17**	·	·	·	·	·	1

$C_j^{(t)}$ represents the bacterial phyla $j, j = 1, \ldots, 6$ at time point t. The order of j are *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, Unclassified and pooled category.

Table 3.3: The observed marginal correlation of the motivating dataset.

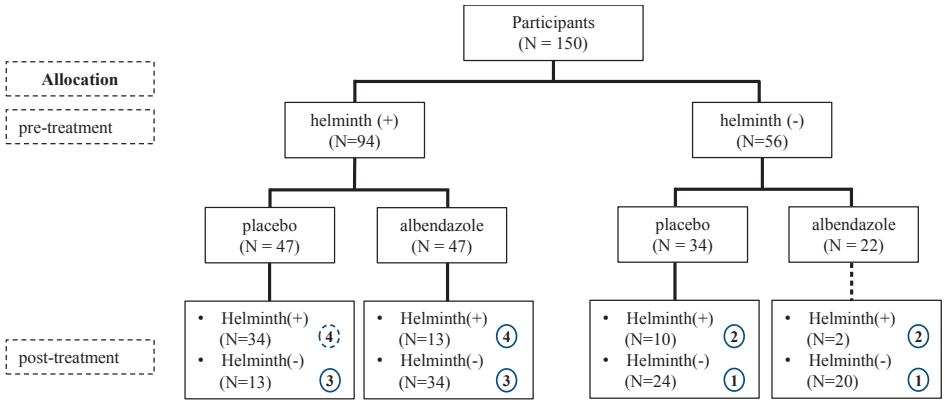


Figure 3.4: The profile of the microbiome study. The chart shows the number of subjecs infected with at least one of the prevalent soil transmitted helminths (Helminth (+)) or free of helminth infections (Helminth (-)) that belonged to either the placebo or albendazole treatment group, at pre-treatment and 21 months after the first treatment round. The circled number represents the condition explained in Section 3.4.

remained uninfected (condition 1) at post-treatment or became infected at post-treatment (condition 2) and infected subjects who either became uninfected at post-treatment (condition 3) or remained infected at post-treatment (condition 4). The number of samples in each conditions at pre- and post-treatment are given in Figure 3.4. It has been shown previously that treatment had an effect on the

Characteristics	albendazole arm (N = 69)	placebo arm (N = 81)
Age (in years) ,mean(SD)	27.38 (16.5)	27.85 (16.91)
Sex, female, n(%)	39 (56.5)	45 (55.6)
Helminth Infections, n(%)		
A. lumbricoides	17 (24.6)	18 (22.2)
Hookworm	26 (37.7)	23 (28.4)
N. americanus	25 (36.2)	23 (28.4)
A. duodenale	2 (2.9)	2 (2.5)
T. trichiura	20 (28.9)	22 (27.2)
Any helminth	47 (68.12)	47 (58.0)
Proportion (in %) of the 6 most abundant bacteria phyla, mean (SD)		
Actinobacteria	12.5 (8.9)	11.0 (7.9)
Bacteroidetes	7.4 (11.3)	6.4 (11.0)
Firmicutes	66.8 (13.5)	70.0 (13.7)
Proteobacteria	9.8 (7.9)	9.2 (8.4)
Unclassified*)	2 (2.22)	2.7 (3.2)
Pooled#)	1.5 (3. 7)	0.7 (1.2)

Table 3.4: **Characteristics at baseline for study participants**.
*)Unclassified represents sequences that cannot be assigned to a phyla.
#)Pooled category consists of the remaining 13 phyla having average relative abundance among samples less than 1%.

composition at post-treatment in infected subjects who remained infected (condition 4)[Martin et al. (2018)]. Here, we want to reanalyze this dataset by using a joint model for the microbiome data at pre- and post-treatment to assess the treatment effect in the infected subjects who remained infected. Additionally, we want to estimate the time effect, while adjusting for other variables such as infection status and treatment allocation. The following loglinear model is considered. Let D, E, F, G, H represent the categorical variables: bacterial taxa, infection (INF), treatment (TRT), baseline infection status (BHelm), and time (t) with J, K, L, M, N levels for each variable. For bacterial phyla, the *Firmicutes* was considered as a reference category. Now the following model was fitted to the data

$$\begin{aligned}\log\left(\mu_{ijklm}^{DEFGH}\right) = &\left(\log\left(\delta_0^{-1}\right) + \lambda_j^D\right) + \left(\lambda_k^E + \lambda_{jk}^{DE}\right) + \left(\lambda_n^H + \lambda_{jn}^{DH}\right) + \\ &\left(\lambda_{ln}^{FH} + \lambda_{jln}^{DFH}\right) + \left(\lambda_{mn}^{GH} + \lambda_{jmn}^{DGH}\right) + \left(\lambda_{lmn}^{FGH} + \lambda_{jlmn}^{DFGH}\right) + \\ &\left(\lambda_{klmn}^{EFGH} + \lambda_{jklmn}^{DEFGH}\right) + u_i, \\ &\text{with the baseline constraint at } J = K = L = M = N = 1, \\ &u_i \sim N\left(0, \sigma_u^2\right) \end{aligned} \tag{3.14}$$

Alternatively the model could be written in terms of regression coefficients as follows.

$$\begin{aligned}\log\left(\mu_{ij}^{(t)}\right) = &\xi_{0j} + \xi_{1j}\text{INF} + \xi_{2j}t + \xi_{3j}\text{TRT} \times t + \xi_{4j}\text{BHelm} \times t + \\ &\xi_{5j}\text{BHelm} \times \text{TRT} \times t + \xi_{6j}\text{INF} \times \text{BHelm} \times \text{TRT} \times t + u_i\end{aligned}$$

where $\xi_{0j} = \log\left(\delta_0^{-1}\right) + \lambda_j^D$, $\xi_{1j} = \lambda_l^F + \lambda_{jl}^{DF}$, and so forth. In this model, there are 6×7 estimable covariate effects on each bacterial phyla. In condition 4, the difference in the microbiome composition between the albendazole and placebo arm is represented by $\xi_{3j} + \xi_{5j} + \xi_{6j}$, while in condition 3, the difference in the microbiome composition between two arms by $\xi_{3j} + \xi_{5j}$. In the subjects who are uninfected at baseline the treatment effect is represented by ξ_{3j}, irrespective of their infection status at post-treatment. The change of microbiome composition, when subjects were uninfected at baseline, remained uninfected at post-treatment, and received placebo, is modelled by ξ_{2j}. Two interaction terms with BHelm were included in this model (3.14) (i.e. the coefficient ξ_{4j} and ξ_{5j}) to model the effect of having infection at pre-treatment and still being infected at follow up, irrespective of treatment by albendazole. The coefficient ξ_{4j} represents the effect of having infection at pre-treatment in the placebo group. We first included a subject-specific random effect u_i in the model. Statistical significance for each covariate was assessed by the likelihood ratio test with 6 degrees of freedom and the significance of the random effect was assessed using the likelihood ratio test with mixture of $\chi^2_{[0,1]}$ distribution.

The parameter estimates from the loglinear model with subject-specific random effects (3.14) are given in Table S8. The between subject variation over time is estimated by the standard deviation σ_u of 0.269 (s.e. of 0.053). The variance of this random effect is significantly different from zero (p-value < 0.001, LRT with mixture of $\chi^2_{[0,1]}$ distribution), indicating that the microbiome counts of a person over time are correlated. The regression coefficients for the covariates BHelm$\times t$ (ξ_{4j}) and BHelm$\times$TRT$\times t$ (ξ_{5j}) appear not to be significantly associated

with the microbiome ($p-$values > 0.05), indicating that having infection at pre-treatment does not influence the microbiome composition. These two covariates were present at the second time point for subjects in condition 3 and 4. Being the terms $\xi_{4j} + \xi_{5j}$ almost zero for all categories, the change of microbiome in these conditions appears to be not affected by these two covariates.

To obtain a model with less parameters, we first eliminated the covariate BHelm$\times$TRT$\times t$. The covariate BHelm$\times t$ was also not significant in this reduced model ($p-$value of 0.795). Hence, we reduced the model (3.14) further by eliminating this covariate. In this updated model, BHelm$\times$TRT$\times t$ was still not significant ($p-$value of 0.843). Finally, we fitted the following model

$$\log\left(\mu_{ij}^{(t)}\right) = \xi_{0j} + \xi_{1j}\text{INF} + \xi_{2j}t + \xi_{3j}\text{TRT} \times t + \xi_{4j}\text{INF} \times \text{BHelm} \times \text{TRT} \times t + u_i. \tag{3.15}$$

In this final model for fixed effects assuming a subject-specific random effect (3.15), 6×4 parameters represent the covariate effects on the microbiome composition. The treatment effect is modelled by ξ_{3j} for all conditions except for condition 4. The difference in the microbiome composition in condition 4 between the albendazole and placebo arm is represented by $\xi_{3j} + \xi_{4j}$. The estimated log odds ratio for each bacterial category compared to *Firmicutes* is given in Table S9. Also for this model the standard deviation of random subject-specific effect u_i is significantly greater than zero ($p-$value < 0.001). Albendazole has no direct effect in subjects who remained uninfected as the odds ratios for each bacterial category are approximately 1. On the other hand, when subjects remained infected, the odds of *Actinobacteria* to *Firmicutes* at the second time point compared to the first time point increases about 55% while the odds ratio for *Bacteroidetes* to *Firmicutes* decreases about 62%.

Next we considered a 6 dimensional random effects structure for this data. We fitted DMM model (3.15). The results are listed in Tables 3.5 and S10. Overall, the estimates of the fixed effects and overdispersion are very similar for these random effect structures. This is in line with the result of the simulation study. However, when we fitted the DMM model with categorical-specific random effects, we observed the following; while the estimated variance component over time for the first three categories are relatively large ($\sigma^2_{u_1} = 0.369$ to $\sigma^2_{u_3} = 0.536$), for the last three categories (*Proteobacteria*, Unclassified and Pooled) are small and hence the random effects for these categories can be omitted.

Finally, we investigated whether the correlations induced by the model correspond to the observed correlations; the marginal correlation induced by the DMM model with a subject-specific random effect (Table S11A),a categorical specific random effect with common variance (Table 3.6) and with categorical-dependent variance for the random effects (Table S11B). For all DMM models, the pairwise correlations at each time points between *Firmicutes* and the other three preva-

Categories	INF	t	TRT$\times t$	Bhelm$\times$INF$\times$TRT$\times t$
Actinobacteria	-0.006 (-0.218, 0.207)	0.050 (-0.155, 0.256)	0.046 (-0.235, 0.326)	0.326 (-0.042, 0.694)
Bacteroidetes	0.220 (-0.056, 0.496)	-0.119 (-0.395, 0.157)	-0.012 (-0.381, 0.356)	-0.916 (-1.573, -0.259)
Protobacteria	0.171 (-0.054, 0.396)	0.056 (-0.161, 0.273)	0.035 (-0.256, 0.326)	0.026 (-0.376, 0.427)
Unclassified	-0.024 (-0.304, 0.257)	0.129 (-0.149, 0.407)	-0.099 (-0.476, 0.277)	-0.159 (-0.727, 0.410)
Pooled	0.166 (-0.158, 0.490)	0.195 (-0.124, 0.515)	-0.030 (-0.449, 0.388)	-0.180 (-0814, 0.454)
Loglik	-8285.5	$\widehat{\theta}$ (s.e)	0.08 (0.01)	
$\hat{\sigma}_u$ (s.e)	0.22 (0.03)			

*Fitted with SAS procedure NLMIXED with 3 quadrature points of Adaptive Gauss-Hermite approximation.

Table 3.5: The log odds ratio (95% CI) when dataset were fitted with DMM with categorical-specific random effect having common variance.*

lent bacterial phyla are relatively high and similar to the observed marginal correlations (Table 3.6, Table S11A-B). With regard to the correlation of categories between the two time points, the DMM model with categorical-specific random effects with common variance showed a similar correlation structure to the observed one (Table 3.6). For the DMM model with categorical-specific random effect, the correlation between the same category over time seems to be too high compared to the dataset (Table S11B). Therefore, we concluded that the DMM model with a categorical specific random effect having common variance across categories is the model which describes our data best.

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	1	-0.55	-0.35	-0.51	-0.3	-0.22						
$C_2^{(1)}$	·	1	-0.06	-0.09	-0.05	-0.03						
$C_3^{(1)}$	·	·	1	-0.04	-0.03	-0.02						
$C_4^{(1)}$	·	·	·	1	-0.05	-0.03						
$C_5^{(1)}$	·	·	·	·	1	-0.02						
$C_6^{(1)}$	·	·	·	·	·	1						
$C_1^{(2)}$	**0.19**	-0.12	-0.06	-0.1	-0.05	-0.04	1	-0.57	-0.3	-0.51	-0.3	-0.23
$C_2^{(2)}$	-0.12	**0.13**	0.03	0.03	0.02	0.02	·	1	-0.07	-0.08	-0.06	-0.04
$C_3^{(2)}$	-0.05	0.02	**0.05**	0.02	0.01	0.01	·	·	1	-0.05	-0.02	-0.01
$C_4^{(2)}$	-0.1	0.02	0.02	**0.12**	0.02	0.01	·	·	·	1	-0.05	-0.03
$C_5^{(2)}$	-0.05	0.02	0.01	0.02	**0.05**	0.01	·	·	·	·	1	-0.02
$C_6^{(2)}$	-0.04	0.02	0.01	0.01	0.01	**0.03**	·	·	·	·	·	1

Table 3.6: The estimated marginal correlation of the dataset obtained by DMM model with categorical-specific random effect having common variance across categories.

3.5 Discussion

We proposed a novel parametric multivariate method to model microbiome data from an epidemiological study using a repeated measurements design. Current parametric models that account simultaneously for overdispersion and repeated measurements use a combination of a conjugate and a normal distribution. This method was introduced by Booth et al. (2003) for count data. Molenberghs et al. (2010) reviewed the combined model for the binary [Molenberghs et al. (2012)] and time-to-event data [Efendi et al. (2014)]. The multinomially-distributed data were however not considered in these papers. The rationale of this combined model is the simplification to the parent distribution when overdispersion is absent and furthermore, the conditional distribution given the normally distributed random effect has a closed-form formula which reduces computational time. Thus, this model has an advantage over the generalized linear mixed models where multivariate normal distributions were used to model correlation due to overdispersion and repeated measurements. Our proposed model is also an extension of the econometrics model for the analysis of choice probabilities in the cross-sectional setting [Guimarães and Lindrooth (2007)]. We considered three models for the analysis of repeatedly measured microbiome data, namely models corresponding to the unconditional distribution and to conditional distribution given the total count of a sample. For the latter distribution, we considered the situations where the total counts either vary or are fixed. We showed that for the last situation, i.e. total count is fixed, the likelihood is equivalent to the likelihood of the multinomial logistic model. Since in our dataset the total number of counts per sample is constant we prefer to use the DMM model.

In a simulation study, we showed that the DMM model provides unbiased estimates for the fixed and random effects independent of the used random effect structure to model the correlation between subjects across time. The sensitivity of the likelihood ratio test for the fixed and random effect components are relatively high when the sample size is large as in the case of our data application. We also showed that the models provided similar estimates for the fixed and random effects when datasets were generated from DMM model with different random effect structure. Two structures of the random effects were considered in the DMM model; one is the simplest subject-specific random effect where the variation of each categorical count outcomes is the same, and the second is to assume a diagonal covariance structure with the same variance for each category. With regard to the marginal correlation for each category between time points, we observed that different correlations can be obtained by changing the random effect structure. The simple random effect structure provides small correlations while for the model with categorical-specific random effect, the correlations are larger and increase with the size of variance component. Hence, if the interest is solely on the fixed effects and random effect estimates, the simple model with subject-specific

random effect can be used. On the other hand, when the correlation structure between the same category across time is of interest, a more complex DMM model with categorical-specific random effects should be used.

For our data application, we were interested in the parameters modelling the variability between subjects and the effect of covariates on microbiome composition therefore we used a subject - specific random effect. Following the generalized linear mixed model framework, the random effect $\boldsymbol{u}_i$ is linked to the expected outcome and measures the variation of the count outcome for certain category between subjects. The variability of the categorical count between subjects is then captured by a single estimate of the standard deviation of the random effect and its significance reveals that the variability between subjects should be taken into account in the model. The estimate of the standard deviation in our data analysis is 0.269 (s.e. of 0.053, p-value < 0.001) which is relatively small hence our assumption of a subject-specific random effect is justified. The standard deviation although small is significant hence our extensive model is necessary for this data. With regard to the fixed effects, their estimates describe the contribution of the covariate to the odds ratio of two bacterial categories. One advantage from our model is to model the change of microbiome in different strata over time. For instance, we showed in the motivating dataset that the change of microbiome over time in subject who remained uninfected in the placebo arm could be inferred from the estimate of the time coefficients. Using the same model, we could also infer the change of microbiome when subjects remained unifected in the albendazole arm as well as the change of microbiome when subjects remained infected. In the previous analyses, we selected subjects who were infected at pre-treatment and fit the Dirichlet-multinomial regression at post-treatment to observe the effect of having long term infection and treatment on microbiome composition. The statistical test using that model showed that subjects who remained infected and received albendazole harbored significantly different microbiome composition compared to subjects who remained infected and received placebo. This result is confirmed by the analysis in this manuscript.

On the other hand, for the data application, when the interest is on the marginal correlation, the random effect structure has to be correctly modelled. For our dataset, we considered three structures, namely subject-specific random effects, categorical-specific random effects with common variance and with categorical-dependent variances. The second correlation structure represents our data best, suggesting that the first structure is too restricted and in the third structure, there were too many parameters for which there is not sufficient information in our data to estimate all of them.

Several challenges in modelling the microbiome data using this method exist. Firstly, in our data application, we were able to fit a categorical-specific random effects structure however the computational burden was large. More research is needed to obtain computational efficient methods. The second challenge is

related to the number of categorical count outcome involved in the study. Typically, categories with rare count (bacteria only presence in the small number of samples) are pooled. One might argue that this rare count might be due to systematic error rather than sequencing error and thus pooling could be viewed as losing the information. Future research should address the issue of the number of categories included in the analyses and consequently a development of computationally efficient method is needed to take into account the category-dependent random effect.

Several alternatives for our approach can be considered. Although modelling overdispersion with the conjugate distribution has computational advantages, it might be too simple since all correlation is modeled by one additional parameter. Extensions to more complex correlation structures would be of interest. Secondly, the choice of six categories is arbitrary. More categories can be analyzed if the dimension of the parameter space is reduced, for example using penalization [Chen and Li (2013); Xia et al. (2013)]. Thirdly, the interpretation of the fixed effect parameters are all conditional on the random effects. In practice, one might be interested in marginal parameters [Heagerty (1999); Tsonaka et al. (2015)]. To this end, marginalized models for multivariate counts need to be developed. Finally, it is of interest to analyze the microbiome data jointly with other outcomes such as diseases or immunological markers. For example, we would like to model the effect of helminths and treatment on microbiome composition and cytokines. This is a topic of an ongoing research.

3.6 Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 Bias and MSE of datasets generated from DMM model with subject-specific random effect when parameters are obtained from the true dataset.

Figure S1 Bias and MSE of datasets generated from the DMM model with categorical-specific random effect having categorydependent variances.

Table S2 The sensitivity and specificity of the hypothesis testing for (A) covariate effect and (B) variability of random effect.

Table S3 The estimated marginal correlations based on the DMM model with subject-specific random effect across different standard deviation of random effects using Monte-Carlo.

Table S4 The estimated marginal correlations based on the DMM model with categorical-specific random effects with common variance across categories using Monte-Carlo.

Table S5 The estimated marginal correlations based on the UNBM model with subject-specific random effect across different standard deviation of random effects using Monte-Carlo.

Table S6 The estimated marginal correlation based on the UNBM model with categorical-specific random effect having common variance across categories using Monte-Carlo.

Table S7 The description of bacterial count data at each time-points.

Figure S2 The distribution of bacterial phyla when zero count presents.

Table S8 **The starting model**. The estimate (95% CI) of the log odds ratio for each covariates in the microbiome dataset.

Table S9 **Final Model**. The estimate of the log odds ratio (95% CI) for each covariates in the microbiome dataset.

Table S10 The log odds ratio (95% CI) when dataset were fitted with DMM with categorical-specific random effect having category-dependent variance across categories.

Table S11 The estimated marginal correlation of the dataset obtained by DMM models.

A Derivation of the joint multivariate distribution

A.1 Joint multivariate distribution for proportions

Conditioned on the random effect $\boldsymbol{u}_i$, the relative abundances are independent. Thus, the joint distribution for the multivariate relative abundance for subject i could be formulated as follows.

$$\begin{aligned}
\Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}\right) &= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}, u_i\right) d\,u_i \\
&= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}} | u_i\right) \Pr\left(\frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}} | u_i\right) \Pr(u_i)\, d\,u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \Pr\left(\frac{\boldsymbol{C}_i^{(t)}}{C_{i+}^{(t)}}\right) \Pr(u_i)\, d\,u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \frac{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)}\right) C_{i+}^{(t)}!}{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)} + C_{i+}^{(t)}\right)} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)} + C_{ij}^{(t)}\right)}{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)}\right) C_{ij}^{(t)}!} \Pr(u_i)\, d\,u_i
\end{aligned} \tag{3.16}$$

with $\mu_{i\cdot}^{(t)}$ is the loglinear mean.

A.2 Joint multivariate distribution under condition on total count.

We will show that the distribution given in equation (3.9) and (3.10) are in general not equivalent. The distribution in the equation (3.10) and (3.12) are not equivalent except for the situation where the total count is fixed.

We denote the $\boldsymbol{C}_i^{(t)}$ as the multivariate count outcome at time t for subject i and the total count to be $C_{i+}^{(t)}$. Thus the multivariate count outcome for subject i conditional on their total is as follows.

$$\begin{aligned}
\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) &= \frac{\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}\right)}{\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}\right)} \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i} \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}
\end{aligned} \tag{3.17}$$

The probability $\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right)$ could be rewritten as follows.

$$\begin{aligned}
\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) &= \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)} | u_i\right) \Pr(u_i) \\
\text{(by conditional independence)} &= \Pr\left(C_{i+}^{(1)} | u_i\right) \Pr\left(C_{i+}^{(2)} | u_i\right) \Pr(u_i) \\
\text{(conditional distribution)} &= \frac{\Pr\left(C_{i+}^{(1)}, u_i\right)}{\Pr(u_i)} \frac{\Pr\left(C_{i+}^{(2)}, u_i\right)}{\Pr(u_i)} \Pr(u_i) \\
&= \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right)}{\Pr(u_i)} \frac{\Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)} \Pr(u_i) \\
&= \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)}
\end{aligned}$$

Thus, the joint probability of multivariate count outcome given in equation (3.17) can be written as follows.

$$\begin{aligned}
&\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)} d u_i}{\int_{u_i} \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)}} \\
&= \int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \left[\frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(u_i | C_{i+}^{(2)}\right)}{\Pr(u_i) \int_{u_i} \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(u_i | C_{i+}^{(2)}\right)}{\Pr(u_i)} d u_i} \right] d u_i
\end{aligned} \tag{3.18}$$

Since $\Pr(u_i)$ is not equal to the term in bracket in equation (3.18) then equation (3.9) and (3.10) are not equivalent. However, when the total count is fixed, the following equation holds: $\Pr\left(u_i | C_{i+}^{(t)}\right) = \Pr(u_i)$. Now, using the last equation in (3.18), the joint distribution becomes

$$\begin{aligned}
\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) &= \int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)} | C_{i+}^{(1)}, u_i\right) \Pr\left(\boldsymbol{C}_i^{(2)} | C_{i+}^{(2)}, u_i\right) \Pr(u_i) \, d u_i \\
&= \int_{u_i} \frac{\Pr\left(\boldsymbol{C}_i^{(1)}, C_{i+}^{(1)} | u_i\right) \Pr\left(\boldsymbol{C}_i^{(2)}, C_{i+}^{(2)} | u_i\right)}{\Pr\left(C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(2)}\right)} \Pr(u_i) \, d u_i.
\end{aligned} \tag{3.19}$$

Since the count at each category $C_{ij}^{(t)}|u_i \sim \mathrm{NB}\left(\theta^{-1}\mu_{ij}^{(t)}, \frac{\theta^{-1}}{1+\theta^{-1}}\right)$ where $\log\left(\mu_{ij}^{(t)}\right)$ $= \boldsymbol{X}_i\boldsymbol{\xi}_j + u_i$, we obtain similar formulation as the conditional likelihood at the cross-sectional setting. Thus, in the case where the total count is fixed, the formulation is equivalent to the distribution of the multivariate relative abundance (3.16).

Part II

Helminth infections on gut microbiome and immune responses

4

The effect of gut microbiome composition on human immune responses - interference of helminth infections

Abstract

Background. Soil transmitted helminths have been shown to have immune regulatory capacity, which they use to enhance their long term survival within their host. As these parasites reside in the gastro-intestinal tract, they might modulate the immune system through altering the gut bacterial composition. Although the relationships between helminth infections or the microbiome with the immune system have been studied separately, their combined interactions are largely unknown. In this study we aim to analyze the relationship between bacterial com-

This chapter is submitted for publication as: Ivonne Martin, Maria MM Kaisar, Aprilianto E Wiria, Firdaus Hamid, Yenny Djuardi Erliyani Sartono, Bruce A Rosa, Makedonka Mitreva, Taniawati Supali, Jeanine J Houwing-Duistermaat, Maria Yazdanbakhsh, Linda J Wammes. The effect of gut microbiome composition on human immune response-interference of helminth infections.

munities with cytokine response in the presence or absence of helminth infections.

Results. For 66 subjects from a randomized placebo-controlled trial, stool and blood samples were available at both baseline and 21 months after starting three-monthly albendazole treatment. The stool samples were used to identify the helminth infection status and fecal microbiota composition, while whole blood samples were cultured to obtain cytokine responses to innate and adaptive stimuli. When subjects were free of helminth infection (helminth-negative), increasing proportions of *Bacteroidetes* was associated with lower levels of IL-10 response to LPS (estimate (95% confidence interval (CI)) -1.96 (-3.05, -0.87)). This association was significantly diminished when subjects were helminth-infected (helminth positive) (p-value for the difference between helminth-negative versus helminth-positive was 0.002). Higher diversity was associated with greater IFN-γ responses to PHA in helminth-negative (0.95 (0.15, 1.75); versus helminth-positive -0.07 (-0.88, 0.73), p-value = 0.056) subjects. Albendazole treatment showed no direct effect in the association between bacterial proportion and cytokine responses, although the *Bacteroidetes'* effect on IL-10 responses to LPS was lower in the albendazole-treated group (-1.74 (-4.08, 0.59) versus placebo (-0.11 (-0.84, 0.62); p-value = 0.193).

Conclusion. The differences that we observed in groups of helminth-positive versus helminth-negative supports the hypothesis that helminth are able to modulate the immune system and specifically may alter the relationship between bacterial communities and cytokine response.

Trial registration: ISRCTN, ISRCTN83830814. Registered 27 February 2008 - Retrospectively registered, http://www.isrctn.com/ISRCTN83830814

4.1 Introduction

Diseases of modernity, such as allergy, autoinflammatory and metabolic diseases are increasingly observed in industrialized countries. It has been speculated that this growing rate was caused by changes in lifestyle, diet and environmental factors, such as pollutant exposure or hygiene. Hygiene improvement has dramatically decreased the prevalence of certain infectious agents such as parasitic helminths while these may have protective effects against autoinflammatory diseases [Wammes et al. (2016)]. Studies analysing the capacity of helminths to modulate the immune system have been carried out in recent decades. However, it has become clear that this is an interplay with several other factors, such as diet, environment and also other gut inhabitants, such as the microbiota.

Early studies showed that gut microbiota are involved in developmental aspects of the immune system and that disturbance can lead to autoinflammatory disorders [Round and Mazmanian (2009)]. Already in 1963 it was reported that the immune system of germ-free mice failed to respond to molecular patterns of pathogenic and beneficial microorganisms, causing morphological tissue defects in the intestinal wall [Abrams et al. (1963)]. In healthy humans, the role of gut microbiota and immune response was studied more recently. It was found that certain bacteria are beneficial for development and function of the immune system and simultaneously the immune system can influence the composition or function of gut microbiota, all relating to inflammatory disorders (reviewed in Belkaid and Hand (2014)).

The presence of parasitic helminths in the gastro-intestinal tract may exert a direct influence on host's gut microbiome as they share the same niche. Although in animal models helminths were shown to increase microbial abundance and diversity [Reynolds et al. (2015)], the findings in human studies are not consistent. Several studies analysing the effect of helminth on gut microbiota have indicated higher diversity of gut microbiota in helminth-positive subjects compared to helminth-negative subjects [Lee et al. (2014), Ramanan et al. (2016)]. A study in Ecuador showed that this difference in diversity might be related to specific helminth species, since they did not find any alterations in *Trichuris trichiura*-infected children [Cooper et al. (2013)]. This might be influenced by different factors among which are different bacterial profiling techniques or confounders such as ethnicity, anthelminthic treatment and environmental differences.

As it has been shown that changes in both gut microbiota and helminth infection status might affect the host's immune response, it is suspected that the presence of helminth might directly or indirectly affect the immune system by altering the gut microbial community [Zaiss et al. (2015)]. For instance, transfer of the microbiota of *Heligmosomoides polygyrus bakeri*-infected mice to uninfected mice induced similar protection against allergic airway inflammation as observed with helminth infection [Wilson et al. (2005)]. In humans, studies on the triangular relation between helminth with microbiome and immune system are still in infancy. To our knowledge, the number of longitudinal studies analysing the association between gut microbiota and immune responses in helminth-endemic areas is still limited. To understand the interaction of the gut microbial community and helminths and their common effect on immune responses, we used data from a household cluster-randomized, double blind, placebo-controlled trial of albendazole treatment in a helminth-endemic area. In this study, it has been shown that deworming reduced helminth prevalence and consequently increased several whole blood cytokine responses [Wammes et al. (2016)]. Helminth infection and anthelminthic treatment separately did not change the gut microbiota [Martin et al. (2018)]. However, when subjects remained infected while treated with albendazole, a decrease of *Bacteroidetes* : *Firmicutes* ratio and an increase of *Acti-*

nobacteria : *Firmicutes* ratio were observed, leading to the hypothesis that there is a cross-talk between microbiome composition and immune response which is disrupted by the presence of helminths and that removing helminth by anthelminthic might affect this communication. Our aim was to characterize the association between bacterial relative abundance with the whole blood cytokine responses and the effect of helminth infections and deworming on this interaction.

4.2 Methods

4.2.1 Participants

Stool samples from 150 subjects from immunoSPIN study [Wiria et al. (2010)] were analyzed for the fecal microbiome. From these, 66 subjects were included in this study based on the complete stool data and available cytokine measurements before and 21 months after the first treatment. Four different helminth species were found namely *Ascaris lumbricoides*, hookworms (*Necator americanus* and *Ancylostoma duodenale*) and *Trichuris trichiura*. Details on sample collection and measuring the infection status using PCR are described elsewhere [Wiria et al. (2010)]. *Trichuris trichiura* infection was assessed only by microscopy, since at that time there was no real-time PCR data available for this species. For this manuscript, we defined a helminth-infected subject as participant with a positive real-time PCR (cycle of threshold (Ct) value $\leqslant 30$) and/or positive microscopy for one or more species of helminths, as described previously [Martin et al. (2018)]. Subjects with a positive real-time PCR with a Ct above 30 were regarded as uninfected.

In addition, from the 66 subjects, 20 subjects of 18 years old or older at pretreatment who were helminth-negative were selected from Nangapanda (Ende) area, as well as 16 subjects who had migrated to Jakarta more than 10 years before and 14 people from the USA (healthy adults from the Human Microbiome Project (HMP) to illustrate the microbiome profile from subjects residing in different geographical environments.

4.2.2 Microbiome composition

The amplification and pyrosequencing of the 16S rRNA gene followed the protocols developed by the Human Microbiome Project (HMP) [HMP (2012)] at the McDonnell Genome Institute, Washington University School of Medicine in St. Louis and have been described previously elsewhere [Martin et al. (2018), Rosa et al. (2018)]. Briefly, the V1 – V3 hypervariable region was PCR – amplified and the PCR products were sequenced on the Genome Sequencer Titanium FLX (Roche Diagnostics, Indianapolis, Indiana), generating on average 6,000 reads per

sample. Details of the filtering and analytical processing of 16S rRNA data for this cohort has been previously described in Rosa et al. (2018). The assembled contigs count data as a result of RDP classification was organized in matrix format with taxa in columns and subjects in row. The entries in the table represent the number of reads for each taxa for each subject. Our work is focused at a phylum level of gut bacterial. Five bacterial phyla have average relative abundances larger than 1%, namely *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria* and an unclassified category, which consists of sequences which could not be categorized into a phylum. The remaining bacterial phyla which had lower relative abundance were pooled together into a pooled category. In the analysis, we retained the count for the three most abundant bacterial phylum proportions, namely *Actinobacteria*, *Bacteroidetes* and *Firmicutes*. The proportion for each phylum was obtained by dividing each sequence count by the total sequence per person at each time point. Along with bacterial proportions, we computed at a phylum level the bacterial diversity within samples (Shannon index) and between samples (Bray-Curtis dissimilarity) using R package vegan [Oksanen et al. (2017)]. The Shannon index represents not only the presence of taxa but also the abundance of corresponding taxa. The higher diversity index means that there was not a single taxa dominating the community and the total bacterial abundance is spread out over all taxa. The Bray-Curtis dissimilarity measures the percentage of similarity between one sample from the other with values range from 0 (completely similar) to 1 (completely dissimilar).

4.2.3 Whole blood cytokine responses

The method to obtain and assess the cytokines responses were described elsewhere [Wiria et al. (2010)]. In brief, heparinized blood was diluted 1:4 and cultured in 96-well plates. Plates were incubated for 24 (innate responses) or 72 (adaptive responses) hours, after which supernatants were harvested and stored in freezers. Cytokine levels were measured by Luminex bead technology in samples obtained at before and 21 months after start of treatment. The analyses carried in this manuscript are limited to innate responses (interleukin (IL)-10 and tumor necrosis factor-alpha (TNF-α)) to lipopolysaccharide (LPS) from *E. coli* and adaptive responses (interferon-gamma (IFN-γ) and IL-5) to Ascaris antigen (AscAg) and general T cell stimulator phytohemagglutinin (PHA). The AscAg was a homogenate of adult worm *A. lumbricoides* obtained from infected human [Wammes et al. (2014)].

4.2.4 Statistical Methods

The microbiome composition for each group of the different demographical areas was assumed to follow a Dirichlet – multinomial distribution with 6 cate-

gories which represents the 6 most abundant phyla (*Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, unclassified bacteria and pooled). The difference in the microbiome composition between groups was tested using the likelihood ratio test statistic with 6 degrees of freedom.

All parameters of interest were described as means or frequency ($\pm$ standard deviation). Prevalence rates were calculated and compared using the Pearson chi-square test, while the Student *t*-test was used to compare continuous variables.

To study the relationship between cytokines and microbiome over the two time points, a linear mixed effect regression model was fitted with helminth status and treatment as covariates. All models have been adjusted with age and sex, however, since both covariates were not significantly associated with the cytokine responses in any model, they are not included in the final analysis. The correlation between observations from the same subjects was modelled by including a subject-specific random effect. The microbiome was included in the model either as a bacterial proportion or by the Shannon diversity index. The cytokine responses were $\log_{10}$-transformed ($\log_{10}$(concentration + 1)) to obtain normally distributed variables. First, we analyzed the main effect of bacterial proportion and diversity on cytokine responses. Second, to allow for different effect sizes of bacterial proportion or diversity on cytokine responses in helminth-positive versus -negative subjects, an interaction term between bacterial categories and infection was included in the model. The *p*-value for this interaction term indicated the statistical evidence for different effect sizes in helminth-positive or –negative groups.

To allow the estimation of the treatment effect on the relationship between bacterial proportion and cytokine responses, the randomized controlled trial design was used. Since the sample size is too small, we only stratified based on randomization arm. Hence, the effect of treatment cannot be distinguished from the effect of helminth infection. Therefore, we explored the relationship between cytokines and microbiome after anthelminthic treatment. A linear mixed effect model was fitted with bacterial proportion or diversity, and treatment as covariates. This model was able to characterize three different associations, namely the association between bacterial proportion or diversity on cytokines at pre-treatment, the difference of the association at pre-treatment and at post-treatment in the placebo group (time effect), as well as the difference of the association at post-treatment between albendazole and placebo group.

For each outcome separately, these models were fitted on subjects who at least had an observation at pre-treatment. The lme4 package in statistical software R was used for model fitting. The significance of the covariate effect was obtained from the likelihood ratio test. Bonferroni correction was used to adjust for multiple testing. The statistical analyses were performed in R [R Core Team (R Core Team)] with mainly lme4 and lmerTest packages [Bates et al. (2015), Kuznetsova et al. (2017)]. The full record was created using the knitr package in R [Xie (2018)]

and is available online at https://github.com/Helminths_GutMicrobes_Cytokine/Ch4_PhDThesis_StatisticalAnalysisinR.pdf.

4.3 Results

4.3.1 Geographical differences in microbiome composition in a rural to urban gradient

From a subpopulation participating in the ImmunoSPIN study in Flores island, Indonesia, 66 individuals were selected for analysis. To illustrate the difference in gut-bacterial community between different geographical areas, we first compared the microbial composition from a sub-selection of helminth-negative subjects who were 18 years or older from Ende (n=20) with subjects from the same area who had moved to Jakarta (n=16) and healthy adults from the USA (n=14) which were considered as residents of rural, urban area of Indonesia, and Western urban area, respectively. The age ranges were from 18 to 62 years old (Jakarta samples) and 18 to 40 years old according to HMP protocol (USA samples) [HMP (2012)]. The proportions of the five main bacterial phyla and a pooled category of remaining bacteria are depicted in Figure 4.1. Bacteroidetes was dominating in the more urban areas (mean ± SD 53.23 $\pm$ 2.38% in US to 4% $\pm$ 0.22% in Ende) while Firmicutes were the most prominent in the rural area (72.45% $\pm$ 1.16% in Ende to 32.11 $\pm$ 2.08% in US) (Figure 4.1A). The microbiome compositions among these three geographical areas were significantly different (p-value < 0.001). Furthermore, the distribution of alpha and beta diversity (Shannon index and Bray Curtis dissimilarity) in samples from three different geographical areas were relatively similar (Shannon diversity index: mean $\pm$ SD 0.85 $\pm$ 0.21 in Ende, 0.91 $\pm$ 0.14 in Jakarta and 0.78 $\pm$ 0.16 in US (Figure 4.1B); Bray-Curtis dissimilarity index mean $\pm$ SD 0.21 $\pm$ 0.09 in Ende, 0.15 $\pm$ 0.08 in Jakarta and 0.31 $\pm$ 0.20 in US; Figure 4.1C).

4.3.2 The effect of bacterial proportions and diversity on *in vitro* cytokine responses

We observed a difference in microbial profiles in rural compared to urban areas. Since it is hypothesized that gut bacteria are associated with certain cytokine responses and thereby possibly immune disorders, we went on to explore this relationship by using data from the ImmunoSPIN trial. For 66 subjects, cytokine responses were measured at pre-treatment and 21 months after the start of anthelmintic treatment. At baseline, 40 out of 66 (60.6%) individuals in Ende were infected with one or more helminth species, and hookworm was the most dominant species (31.8%) followed by *A. lumbricoides* (25.7%) and *T. trichiura* (22.7%). The baseline characteristics such as age, gender, BMI and helminth prevalence

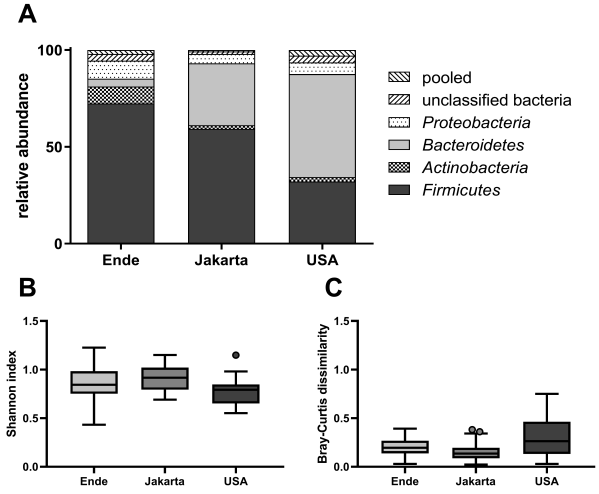


Figure 4.1: **The microbiome composition and diversity for subjects in three different geographical areas.** Composition and diversity of the fecal microbiota was assessed for subjects from different geographical areas: Ende (n=20), Jakarta (n=16) and USA (n=14). In panel (A) the microbiome composition is depicted in percentages of the six categories, where unclassified bacteria represents the category of sequences that could not be assigned to a phyla, and the pooled category consists of the remaining 13 phyla with average relative abundance less than 1%. Panel B and C show the average ± SD of the Shannon diversity index and the Bray-Curtis dissimilarity index, respectively, in the different areas.

were similar between the two treatment arms (Table 4.1). Three-monthly albendazole treatment for 21 months reduced the infection prevalence from 65.4% to 19.2% versus a slight increased of helminth infections from 57.5% to 65% in placebo group (Table S4.5.1).

We analyzed proportions of three bacterial phyla (*Actinobacteria*, *Bacteroidetes* and *Firmicutes*) as these were most abundant in our study population. As we analyzed two cytokines for each antigens, we applied conventional Bonferroni correction and used a cut-off level for significance (α) of 0.025. When fitting the linear mixed model, no direct effect was observed of bacterial proportions or Shannon diversity on whole blood cytokine responses (Table 4.2).

4.3.3 Interference by helminth infection in the effect of bacterial proportions and diversity on in vitro cytokine responses

To elucidate the possible role of helminth infections in the interplay of bacteria and immune responses, we conducted analyses in helminth-positive and -negative groups. For this purpose, we used observations at both pre-treatment

Characteristics		albendazole		placebo	
		N	Result	N	Result
Gender, female (N (%)		26	12 (46.1)	40	22 (55.0)
Age (mean (SD))		26	27.3 (16.1)	40	26.7 (15.7)
Children (<= 18 years old; N (%))			10 (38.4)		17 (42.5.0)
Adults (>18 years old; N (%))			16 (61.5)		23 (57.5)
zBMI (mean (SD))		10	-0.52 (0.98)	17	-0.83 (0.64)
BMI (mean (SD))		16	23.39 (3.44)	23	23.49 (4.89)
Parasite infection (N (%))					
A. lumbricoides[a]		26	9 (34.6)	40	8 (20.0)
Hookworm		26	11 (42.3)	40	10 (25.0)
N. americanus[a]		26	10 (38.5)	40	10 (25.0)
A. duodenale[a]		26	2 (7.7)	40	2 (5.0)
T. trichiura[b]		26	5 (19.2)	40	10 (25.0)
Any helminth		26	17 (65.4)	40	23 (57.5)
Abundance of bacterial phyla (mean % (SD))					
Actinobacteria		26	8.2 (5.3)	40	8.6 (6.9)
Bacteroidetes			7.6 (10.1)		6.7 (11.5)
Firmicutes			66.4 (11.8)		65.0 (13.5)
Proteobacteria			7.3 (5.6)		7.4 (4.6)
unclassified bacteria#)			1.3 (0.7)		2.1 (2.4)
pooled*)			9.2 (6.0)		10.1 (5.8)
Diversity Index, median (IQR)					
Shannon index		26	0.85 (0.71, 0.99)	40	0.84 (0.73, 1.00)
Bray-Curtis			0.19 (0.12, 0.26)		0.19 (0.13, 0.28)
Cytokine responses (pg/mL, median, IQR)					
LPS	IL-10	25	242.00 (132.00, 400.00)	40	213.50 (142.00, 380.20)
	TNF-α	25	664.00 (294.00,1029.00)	40	550.50 (343.00, 840.00)
AscAg	IL-5	22	32.55 (9.55, 58.42)	37	18.90 (12.00, 62.00)
	IFN-γ	23	28.50 (12.10, 111.80)	37	17.40 (7.74, 60.90)
PHA	IL-5	23	490.00 (276.00, 747.50)	37	515.00 (333.00, 870.00)
	IFN-γ	23	2449.00 (354.00, 5424.00)	37	2299.00 (997.00, 3829.00)

Table 4.1: **Characteristics of the participants at baseline**. [a] diagnosed by real-time PCR; [b]diagnosed by microscopy; unclassified bacteria represents the category of sequences that could not be assigned to a phyla, and the; *pooled category consists of the remaining 13 phyla with average relative abundance less than 1%.

		Estimated effect (95 % CI)			
		Actinobacteria	*Bacteroidetes*	*Firmicutes*	Shannon
LPS	IL-10	0.20 (-0.58, 0.98)	-0.39 (-0.90, 0.12)	0.24 (-0.23, 0.71)	-0.22 (-0.51, 0.07)
	TNF-α	0.55 (-0.35, 1.44)	-0.06 (-0.66, 0.54)	-0.14 (-0.70, 0.41)	0.03 (-0.31, 0.37)
AscAg	IL-5	-1.02 (-2.78, 0.74)	0.09 (-1.10, 1.28)	0.39 (-0.74, 1.52)	-0.48 (-1.16, 0.20)
	IFN-γ	-1.03 (-2.45, 0.39)	0.15 (-0.80, 1.10)	-0.20 (-1.13, 0.74)	0.14 (-0.44, 0.71)
PHA	IL-5	-0.04 (-1.55, 1.46)	0.32 (-0.67, 1.32)	-0.85 (-1.82, 0.11)	0.61 (0.02, 1.20)
	IFN-γ	-0.57 (-2.12, 0.98)	-0.26 (-1.28, 0.75)	-0.03 (-1.05, 0.99)	0.45 (-0.18, 1.08)

Table 4.2: The association between bacterial proportion and diversity on cytokine responses.

and post-treatment. Regardless of randomization arm, we fitted the linear mixed model on each cytokine responses as outcomes. The predictors were bacterial proportions and its interaction with helminth infection. A similar analysis was performed to estimate the association between bacterial diversity and cytokine responses. Table 4.3 illustrates the associations between bacterial proportions or diversity and cytokine responses when subjects were helminth-positive or -negative.

In the innate immune response to LPS, the *Bacteroidetes* proportion showed a significant negative association with IL-10 levels in helminth-negative subjects (estimated effect (95% confidence interval (CI)): -1.96 (-3.05, -0.87), *p*-value = 0.001; Table 4.3). This association was significantly different from that of helminth-negative subjects (*p*-value for the difference = 0.002, Figure 2A) in which the association was absent (-0.03 (-0.59, 0.53), Table 4.3). The bacterial diversity had no significant association with IL-10 response to LPS (Table 3, Figure 2B). With regard to the helminth-specific cytokine responses, none of IFN-γ and IL-5 responses to AscAg were significantly associated with bacterial proportions or diversity (Table 4.3). In the adaptive responses (PHA), none of the cytokine responses were significantly associated with the bacterial proportion in uninfected subjects (Table 3). Although not significant, we noticed lower levels of IFN-γ to PHA with higher *Firmicutes* proportions (-1.57 (-3.08, -0.05), *p*-value = 0.045; Table 4.3). This association between *Firmicutes* proportion with IFN-γ to PHA in uninfected subjects was however significantly different from that in subjects who were infected (*p*-value for the difference = 0.009, Figure 2C). At the same time, there was a significantly increasing concentration of IFN-γ to PHA among those who were uninfected when bacterial diversity was higher (0.95 (0.15, 1.75), *p*-value 0.022; Table 3), although this association was not significantly different from the helminth-positive group (-0.07 (-0.88, 0.73), *p*-value for the difference = 0.056; Table 3, Figure 2D). A similar negative association of *Firmicutes* was observed in IL-5 responses to PHA in uninfected subjects (-1.52 (-3.02, -0.02), *p*-value = 0.05; Table 4.3). Conversely increasing bacterial diversity led to slightly higher levels

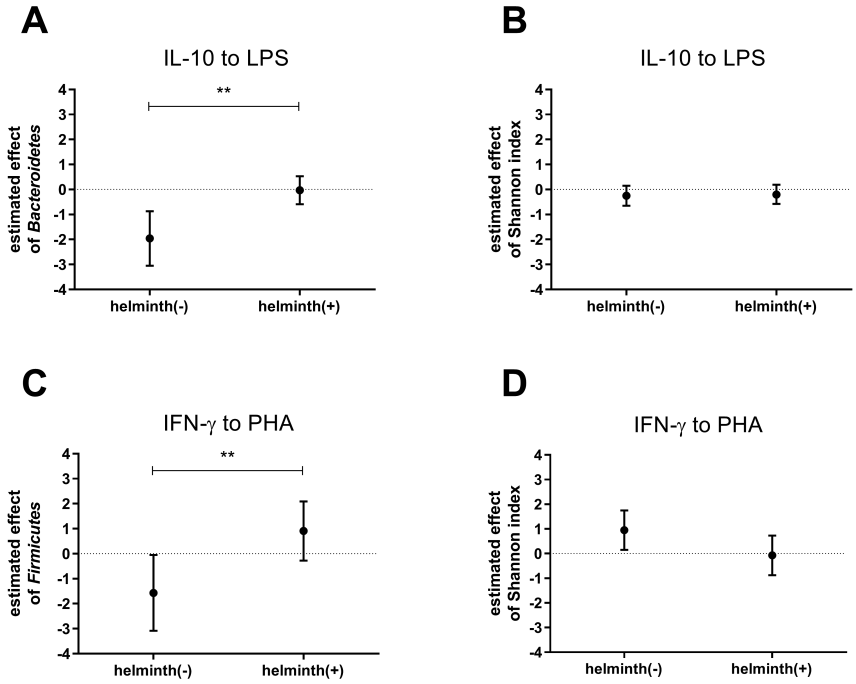


Figure 4.2: **The association between bacterial proportion and diversity on certain cytokines in helminth-negative and -positive subjects.** The effect of bacterial proportions on cytokine responses was analyzed for helminth-negative (helminth(-)) and helminth-positive (helminth(+)) groups by a linear mixed model. Estimated effects $\pm$ 95% CI are shown for the effect of *Bacteroidetes* proportion on IL-10 responses to LPS (A), diversity on IL-10 to LPS (B) and for the effect of *Firmicutes* (C) and diversity (D) on IFN-γ responses to PHA. For assessing statistical significance conventional Bonferroni correction was applied; *p-value $\leqslant$ 0.025, **p-value $\leqslant$ 0.01.

of IL-5 to PHA in the uninfected subjects (0.85 (0.07, 1.63), p-value = 0.034; Table 3). Both observations were not significantly different from the effects in those who were helminth-positive.

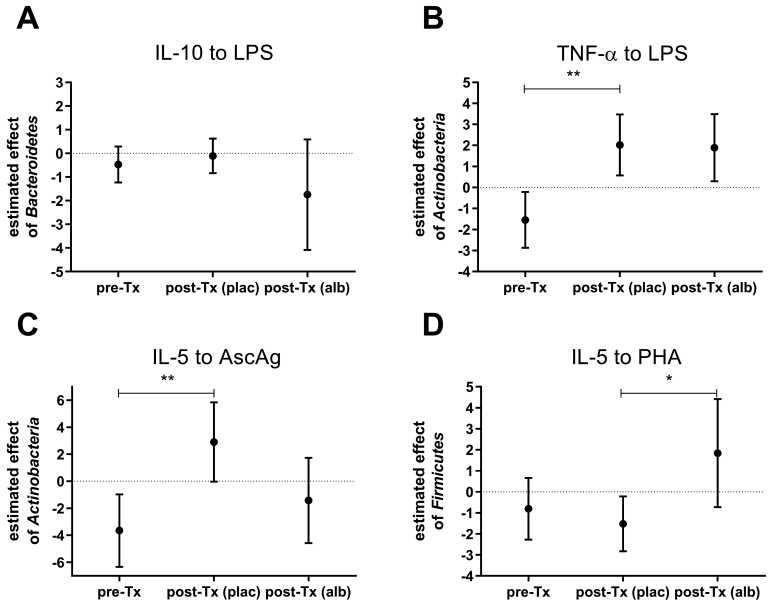


Figure 4.3: **The association between bacterial proportion and diversity on certain cytokines at pre-treatment and post-treatment in two randomization arms.** After deworming, comparisons were made for all subjects pre-treatment versus post-treatment (placebo group) and post-treatment placebo versus albendazole groups. Estimated effects of a linear mixed model $\pm$ 95% CI are depicted. In panel A, the effect of *Bacteroidetes* proportion on IL-10 responses to LPS is shown for the different groups. Panel B and C depict the effect of *Actinobacteria* on TNF-α levels to LPS and on IL-5 responses to AscAg, respectively. The effect of *Firmicutes* on PHA-induced IL-5 is shown in panel D. *p-value $\leqslant$ 0.025, **p-value $\leqslant$ 0.01.

		Estimated effect (95 % CI)				
		Actinobacteria	*Bacteroidetes*	*Firmicutes*	Shannon	Infection status
LPS	IL-10	0.25 (-0.90, 1.41)	-0.03 (-0.59, 0.53)*	0.14 (-0.45, 0.73)	-0.20 (-0.58, 0.19)	helminth(+)
		0.28 (-0.77, 1.34)	**-1.96 (-3.05, -0.87)**	0.42 (-0.31, 1.14)	-0.25 (-0.65, 0.15)	helminth(-)
	TNF-α	0.68 (-0.64, 2.00)	-0.18 (-0.86, 0.50)	-0.04 (-0.73, 0.65)	-0.09 (-0.54, 0.36)	helminth(+)
		0.50 (-0.72, 1.71)	0.35 (-0.97, 1.67)	-0.29 (-1.14, 0.55)	0.16 (-0.31, 0.63)	helminth(-)
AscAg	IL-5	-1.69 (-4.34, 0.97)	-0.04 (-1.41, 1.32)	0.60 (-0.81, 2.00)	-0.76 (-1.70, 0.17)	helminth(+)
		-0.46 (-2.78, 1.86)	0.38 (-2.08, 2.84)	0.10 (-1.69, 1.89)	-0.21 (-1.12, 0.71)	helminth(-)
	IFN-γ	-1.48 (-3.58, 0.63)	-0.03 (-1.12, 1.05)	0.35 (-0.78, 1.48)	-0.43 (-1.18, 0.33)	helminth(+)
		-0.71 (-2.58, 1.16)	0.78 (-1.15, 2.70)	-1.11 (-2.56, 0.33)	0.67 (-0.07, 1.41)	helminth(-)
PHA	IL-5	-1.68 (-3.87, 0.52)	0.33 (-0.81, 1.46)	-0.45 (-1.62, 0.73)	0.34 (-0.45, 1.13)	helminth(+)
		1.36 (-0.58, 3.31)	0.18 (-1.83, 2.20)	-1.52 (-3.02, -0.02)	0.85 (0.07, 1.63)	helminth(-)
	IFN-γ	-1.92 (-4.14, 0.31)	-0.71 (-1.85, 0.43)	0.91 (-0.28, 2.09)*	-0.07 (-0.88, 0.73)	helminth(+)
		0.54 (-1.48, 2.55)	1.20 (-0.85, 3.25)	-1.57 (-3.08, -0.05)	**0.95 (0.15, 1.75)**	helminth(-)

Table 4.3: **The association between bacterial proportion and diversity on cytokine of helminth-infected and -uninfected subjects regardless of treatment allocation**. Bold printed numbers represent significant association between corresponding bacterial proportion or diversity and cytokine (p-value $\leqslant 0.025$). *Significant difference in Helminth(-) versus Helminth(+) (p-value $\leqslant 0.025$).

		Estimated effect (95 % CI)				
		Actinobacteria	*Bacteroidetes*	*Firmicutes*	Shannon	group
LPS	IL-10	-0.36 (-1.60, 0.87)	-0.47 (-1.23, 0.29)	0.44 (-0.28, 1.16)	-0.32 (-0.72, 0.08)	pre
		0.61 (-0.74, 1.96)	-0.11 (-0.84, 0.62)	0.06 (-0.63, 0.75)	-0.14 (-0.57, 0.30)	post-pla
		0.09 (-1.40, 1.58)	-1.74 (-4.08, 0.59)	0.27 (-0.82, 1.35)	-0.20 (-0.91, 0.50)	post-alb
	TNF-α	-1.55 (-2.87, -0.22)*	0.10 (-0.80, 1.00)	0.44 (-0.39, 1.27)	-0.41 (-0.87, 0.05)	pre
		2.02 (0.57, 3.47)**	-0.24 (-1.11, 0.62)	-0.35 (-1.14, 0.45)	0.27 (-0.22, 0.76)	post-pla
		1.89 (0.29, 3.49)	1.15 (-1.61, 3.91)	-0.98 (-2.22, 0.27)	0.75 (-0.06, 1.55)	post-alb
AscAg	IL-5	-3.65 (-6.34, -0.97)*	-0.23 (-2.10, 1.64)	1.52 (-0.19, 3.24)	-0.89 (-1.83, 0.05)	pre
		2.90 (-0.03, 5.84)**	-0.06 (-1.69, 1.58)	-0.76 (-2.29, 0.78)	0.17 (-0.81, 1.16)	post-pla
		-1.42 (-4.58, 1.73)	2.26 (-2.89, 7.41)	1.84 (-1.18, 4.86)	-1.48 (-3.24, 0.28)	post-alb
	IFN-γ	0.76 (-1.42, 2.95)	0.13 (-1.37, 1.63)	-0.84 (-2.24, 0.56)	0.49 (-0.29, 1.27)	pre
		-1.61 (-4.00, 0.77)	0.11 (-1.22, 1.43)	-0.21 (-1.46, 1.04)	0.06 (-0.75, 0.87)	post-pla
		-2.95 (-5.52, -0.38)	0.73 (-3.46, 4.91)	1.92 (-0.54, 4.38)	-0.94 (-2.40, 0.51)	post-alb
PHA	IL-5	0.67 (-1.66, 3.00)	-0.20 (-1.74, 1.35)	-0.80 (-2.27, 0.66)	0.59 (-0.22, 1.40)	pre
		1.66 (-0.90, 4.21)	0.32 (-1.04, 1.69)	-1.52 (-2.83, -0.22)	1.03 (0.18, 1.87)	post-pla
		-2.54 (-5.29, 0.20)	3.16 (-1.14, 7.47)	1.84 (-0.73, 4.41)***	-0.79 (-2.30, 0.72)	post-alb
	IFN-γ	0.32 (-2.09, 2.72)	0.41 (-1.18, 1.99)	-0.72 (-2.25, 0.82)	0.63 (-0.23, 1.48)	pre
		0.90 (-1.72, 3.53)	-1.03 (-2.42, 0.36)	-0.06 (-1.42, 1.31)	0.64 (-0.23, 1.52)	post-pla
		-3.54 (-6.36, -0.71)	2.60 (-1.80, 7.00)	2.19 (-0.50, 4.88)	-0.95 (-2.53, 0.64)	post-alb

Table 4.4: **The association between bacterial proportion and diversity on cytokine responses irrespective of infection status at pre- and post-treatment in two randomization arms**. pre = pre-treatment (regardless of helminth infection status); post-pla = post-treatment placebo arm; post-alb = post-treatment albendazole arm. *p-value $\leqslant$ 0.025, **p-value $\leqslant$ 0.025 in pre versus post-pla, ***p-value $\leqslant$ 0.025 in post-alb versus post-pla.

4.3.4 The effect of albendazole on the relationship between *in vitro* cytokine responses and bacterial proportion and diversity

We further investigated whether deworming affects the relationship between bacterial proportions or diversity and cytokine responses. For this purpose, we fitted the linear mixed model on all subjects (n = 66) to characterize the association between bacterial proportions and cytokine responses at two time points and in the two randomization arms. These analyses were irrespective of the infection status. A similar model was applied for the diversity index.

Table 4.4 lists the associations between the proportions of three major bacterial phyla and diversity with cytokine responses, before and after anthelminthic treatment. With regard to the relationship between *Bacteroidetes* and IL-10 response to LPS, no significant differences were observed between pre- versus post-treatment or between treatment groups (Table 4.4). While the estimated association between *Bacteroidetes* proportion and IL-10 to LPS at pre-treatment (estimate (95% CI): -0.47 (-1.23, 0.29)) and post-treatment in placebo group (-0.11 (-0.84, 0.62)) were close to zero, the association at post-treatment in albendazole group was clearly lower (-1.74 (-4.08, 0.59); *p*-value for the difference between placebo and albendazole was 0.193, Table 4.4 Figure 4.3A). The association between IFN-γ in response to PHA and bacterial diversity was also not significant at post-treatment either in placebo or in albendazole group (Table 4.4).

The association between higher *Actinobacteria* proportion with decreasing response of TNF-α to LPS was borderline significant at pre-treatment (estimate (95% CI): -1.55 (-2.87, -0.22), *p*-value = 0.024; Table 4.4). This association was significantly different to the effect of *Actinobacteria* at post-treatment when subject received placebo (2.02 (0.57, 3.47); *p*-value < 0.001; Figure 4.3B), however no difference was observed when comparing placebo and albendazole group (1.89 (0.29, 3.49), *p*-value for the difference = 0.907; Figure 4.3B). A similar result was obtained from the association between *Actinobacteria* with IL-5 responses to AscAg. At pre-treatment, the increasing *Actinobacteria* proportions were significantly associated with less IL-5 production in response to AscAg (-3.65: (-6.34, -0.97), *p*-value = 0.009; Table 4.4). This association was significantly different to the effect of *Actinobacteria* at post-treatment in placebo group (2.90 (-0.03, 5.84), *p*-value = 0.002; Figure 4.3C). Although the estimated association in albendazole group was lower (-1.42 (-4.58, 1.73), this was not significantly different between the treatment groups (*p*-value = 0.052; Figure 4.3C).

On the other hand, while the association between *Firmicutes* and IL-5 response to PHA at pre-treatment was not significantly different compared to the association at post-treatment in placebo group, there was a significant difference of this association between albendazole and placebo group at post-treatment (estimate (95% CI) for placebo -1.52 (-2.83, -0.22) versus albendazole 1.84 (-0.73, 4.41), *p*-

value = 0.024; Table 4.4, Figure 4.3D).

4.4 Discussion

This study aimed to analyze the effect of helminth infections on the relationship between gut microbiota and the immune system. Examination of the microbiome composition in rural and urban area of Indonesia as well as USA showed that there were clear gradients in *Bacteroidetes* to *Firmicutes* proportion. This was one of the reason we focused on the three bacterial phyla in this study besides the result of the previous study on these subjects which reveals the associations between helminth infection and the odds of *Bacteroidetes* to *Firmicutes* as well as *Actinobacteria* to *Firmicutes*. When focusing on samples from Ende, we found a negative association between proportions of *Bacteroidetes* and IL-10 response to LPS in helminth-negative subjects and the presence of helminths was shown to dampen this effect. Anthelminthic treatment partly recovered this effect, although not statistically significant. To our knowledge, this is the first time that the association between gut microbiome, presence of parasitic helminths and whole blood cytokine responses was analyzed in a longitudinal study using a randomized placebo-controlled anthelminthic trial.

IL-10 was already marked as a key anti-inflammatory cytokine involved in induction of immune suppression by helminths [Yazdanbakhsh et al. (2002)]. Our observation that helminths counteract the suppressed IL-10 response to LPS in subjects with higher *Bacteroidetes* proportions supports the so called "old friends hypothesis" [Rook (2009)], stating that certain infectious agents such as helminths may have protective effects against immune dysfunction and inflammatory diseases, possibly through IL-10. This is strengthened by our observed gradient of the relative abundance of *Bacteroidetes* from rural to urban areas, where immune-related diseases are more prevalent [Bach (2002)]. In contrast, a recent meta-analysis indicated that inflammatory bowel disease (IBD) patients displayed lower proportions of *Bacteroidetes* [Zhou and Zhi (2016)], however this was only found when measuring by real-time quantitative PCR (not by conventional culture) and mainly in Asian studies. Furthermore, a member of the *Bacteroidetes* family, gut inhabitant *Bacteroides fragilis*, was shown to protect mice from experimental colitis, mediated by polysaccharide A (PSA) possibly through IL-10 induction [Mazmanian et al. (2008)]. However, although *B. fragilis* is the most well-known pathogen of the *Bacteroidetes*, it is the least common species in the *Bacteroidetes* phylum in the human gut [Wexler (2007)]. It could therefore be that other factors or species play a dominant role in the general effect of *Bacteroidetes* on IL-10 responses. Further studies are therefore needed to assess the translation of our findings to a clinical setting, for example prevalence or activity of IBD or other auto-immune diseases. Moreover, since we have measured systemic whole blood cytokine responses, we are not sure whether this is representative for the gut re-

sponses.

A trend of negative association between *Firmicutes* and concentration of IFN-γ to PHA was seen in helminth-negative subjects only. In subjects with helminth, this association was positive, although this difference fell short of statistical significance. Parallel to this trend, the bacterial diversity was positively associated with IFN-γ responses to PHA in subjects who did not carry helminths, and in helminth-positive subjects this association was dampened. Since a similar opposite trend was observed in the relationship between *Firmicutes* compared to bacterial diversity on IL-5 responses to PHA, we may conclude that not the proportion of *Firmicutes*, but the total bacterial diversity drove this association. *Firmicutes* was the most abundant phyla in this population and the increasing proportion of *Firmicutes* will obviously reduce diversity. This indicates that analyzing single bacterial phyla without considering the remaining phyla may lead to biased results as microbiome data is compositional and thus correlated between phyla.

Although deworming removed most helminths, treatment did not significantly alter the effects of bacterial proportions on cytokine responses. Regarding the *Bacteroidetes* effect on LPS to IL-10, we did observe a lower effect in the albendazole group compared to placebo. Although not significant, this might point towards the idea that anthelminthic treatment could restore the -possibly detrimental- interaction of bacteria with immune responses. Surprisingly, we found differences in immune modulation by *Actinobacteria* in the pre- versus post-treatment group. Although there was a significant association of time (in subjects receiving placebo), these associations were not significantly different in the albendazole group. The effect of time could be explained by other factors such as diet and possibly improved hygiene, resulting from increased awareness during the presence of our medical team in the study area. In the analysis of treatment's effect on the association between bacterial proportion and diversity, there was a significant difference between the association of *Firmicutes* on the IL-5 response to PHA in albendazole group compared to placebo group. In subjects receiving albendazole, *Firmicutes* proportions were positively associated with IL-5 levels, while we observed a negative (non-significant) effect in helminth-negative individuals over time. This result seems contradictory, but might be related to the fact that small numbers were analyzed and not everyone in the albendazole group lost their helminth infection. The analysis on subjects who were infected at baseline and cleared their infection would possibly reveal more clearly how the relationship between bacterial communities and immunity are affected by treatment. This analysis lacks statistical power in our study as the sample size was small (n = 12 out of 17 subjects who were successfully dewormed). Future research which involves larger sample sizes needs to be conducted. Another relevant thought in this and similar research settings is that although albendazole removes helminths effectively, the immunomodulatory effects of helminths on cytokine responses are long-lasting and cannot be easily corrected by short-term treatment. It was

previously reported by Endara et al. (2010) that the length of periodic treatment is important for altering immune responses, i.e. that studies with a longer period of treatment (up to 30 months) are more likely to show effects of deworming.

As significant associations between bacterial communities and cytokine responses were only observed when subjects were helminth-negative, clearly other factors than helminth and treatment are also involved in the alteration of the microbiome community and their interaction with the immune system. For example, our study data lack information on diet. Dietary intake was clearly shown to affect bacterial communities in the gastro-intestinal tract [Wu et al. (2011)]. This might also be related to changes in social economic status leading towards a more high-fat diet when moving from rural to urbanized areas. Recent articles reported inconsistencies with regard to the direction of *Bacteroidetes* to *Firmicutes* ratio in rural to urban comparisons of microbiome profiles from different geographical areas. Studies comparing children from Bangladesh to USA children showed direction of increasing *Bacteroidetes* : *Firmicutes* in USA, as observed in our data [Lin et al. (2013)], while studies in elderly Korean and children in Burkina Faso showed opposing results, i.e. decreasing *Bacteroidetes* : *Firmicutes* ratios from rural to urban [Park et al. (2015), de Filippo et al. (2017)]. This could be caused by different genera under *Bacteroidetes* or *Firmicutes* phyla which might be affected by certain type of diet. Therefore, it will be beneficial for the future studies to also include dietary factors from the study participants.

A further limitation is related to the statistical tools available in analyzing this relationship. Here, we characterized the association of three single bacterial proportions on cytokine response in the helminth-positive and -negative group. Using this approach, we first ignore the effect of compositional structure in the microbiome data, namely when computing the p-value we assumed that these bacterial categories are independent while they are correlated. Secondly, the current statistical model ignores the fact that microbiome is a variable measured with errors at a different scale than the cytokine responses [Teixeira-Pinto et al. (2009)]. In addition, we might as well ignore the possible unobserved confounders. It is therefore important for the future studies in this field to develop a statistical method to characterise the effects of helminth infection on both outcomes simultaneously by accounting these unobserved errors with a joint model.

To conclude, our findings supports the hypothesis for a role of helminths in modulating the immune response, which might be related to bacterial proportion and diversity. Deworming did not show a particular effect on the observed associations. It is therefore important to repeat such studies with a larger sample size as well as using more advanced statistical models to further analyze this relationship by considering the complex structure of microbiome data and other possible confounders.

4.5 Supplementary Materials

Characteristics	albendazole		placebo	
	N	Result	N	Result
Parasite infection (%)				
A. lumbricoides	26	3 (11.5)	40	17 (42.5)
Hookworm	26	0 (0)	40	11 (27.5)
N. americanus	26	0 (0)	40	11 (27.5)
A. duodenale	26	0 (0)	40	2 (5.0)
T. trichiura	26	4 (15.4)	40	13 (32.5)
Any helminths	26	5 (19.2)	40	26 (65.0)
Proportion (in %) of the 6 most abundant bacteria phyla, mean(SD)				
Actinobacteria	26	14.1 (8.9)	40	9.2 (7.4)
Bacteroidetes		3.6 (5.8)		7.7 (14.1)
Firmicutes		60.1 (13.7)		59.2 (16.7)
Proteobacteria		9.0 (6.5)		8.7 (6.8)
Unclassified		1.8 (1.1)		2.6 (2.8)
Pooled		11.5 (6.7)		12.5 (7.4)
Diversity Index, median(IQR)				
Shannon index	26	0.90 (0.85, 1.08)	40	0.97 (0.79, 1.05)
Bray-Curtis		0.19 (0.14, 0.26)		0.24 (0.17, 0.36)

Table S4.5.1: The characteristics of the participants at 21 months after treatment

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.

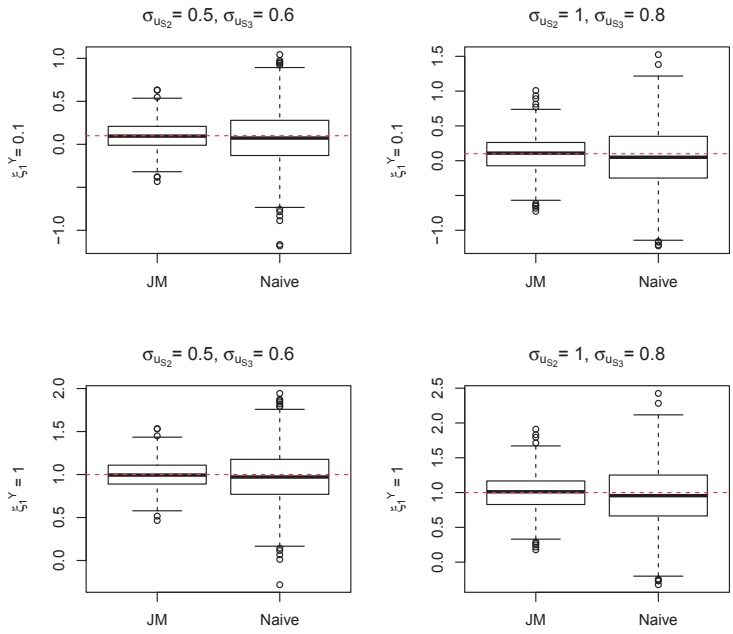


Figure 5.1: Simulation results: the point estimate of covariate of interest from joint model and naive model at the cross-sectional setting. The datasets were generated using a joint model in a cross-sectional setting with logit-dependent random shared effects. The horizontal lines represent the true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

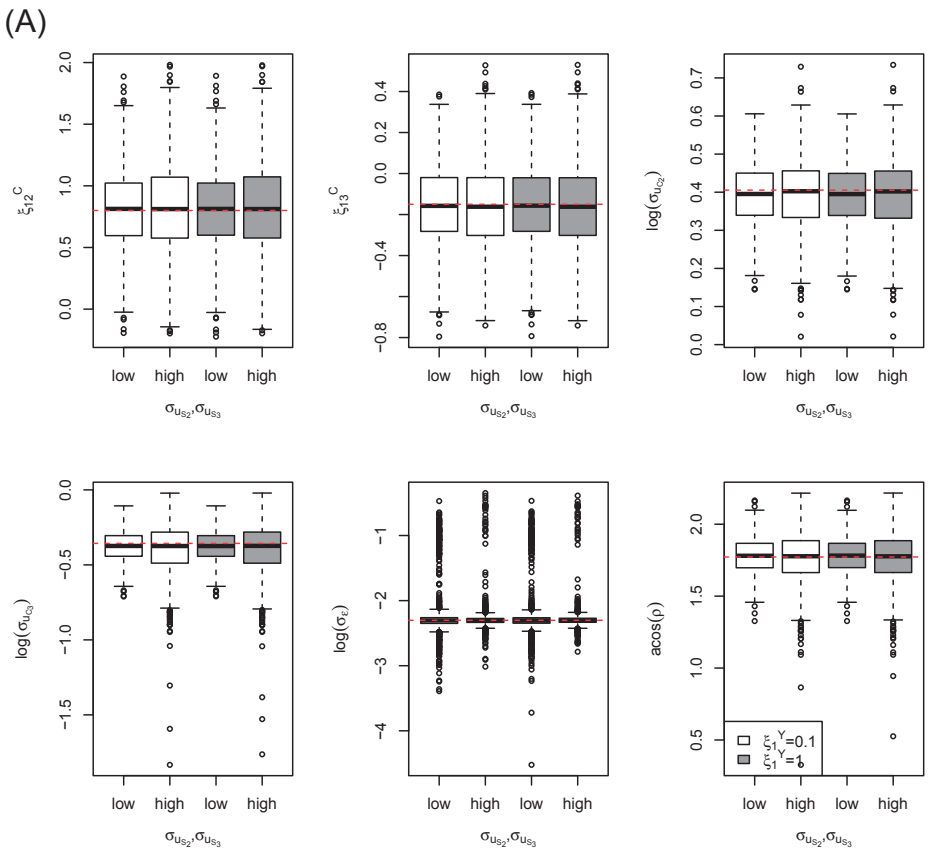


Figure 5.2: Simulation results from a joint model at the cross-sectional setting with logit-dependent random shared effect (A) the point estimates of covariate of interest as well as random effect at different values of shared effect standard deviations, and (B) standard deviations of shared effects. 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-

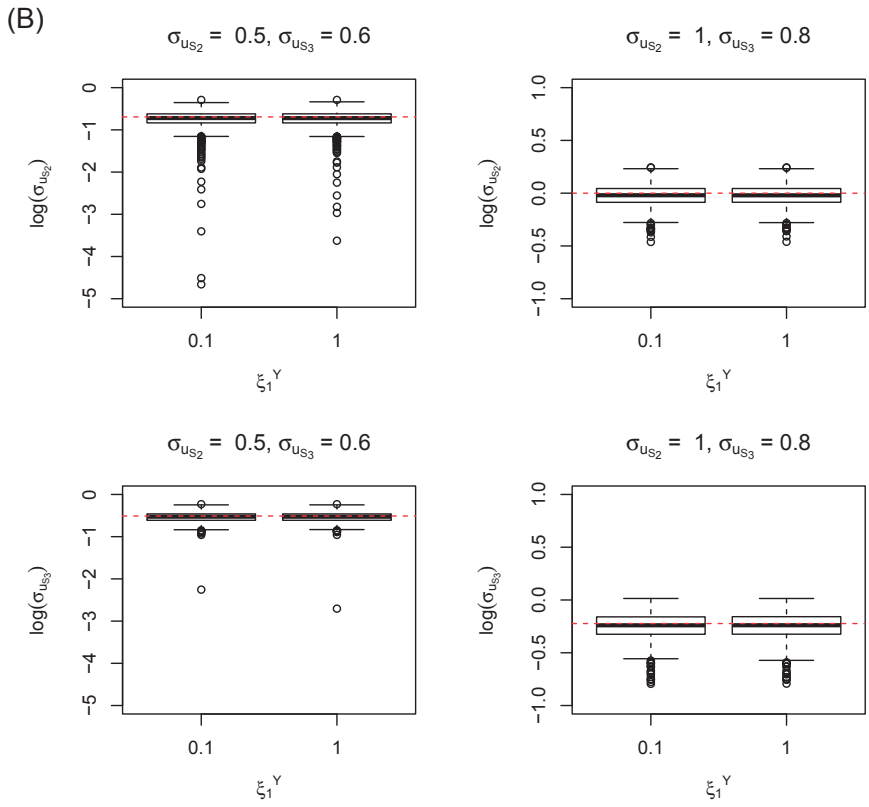


Figure 5.2: (cont.) Simulation results from a joint model at the cross-sectional setting with logit-dependent random shared effect (A) the point estimates of covariate of interest as well as random effect at different values of shared effect standard deviations, and (B) standard deviations of shared 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 ($\sigma_{u_{S_2}} = 0.5, \sigma_{u_{S_3}} = 0.6$),

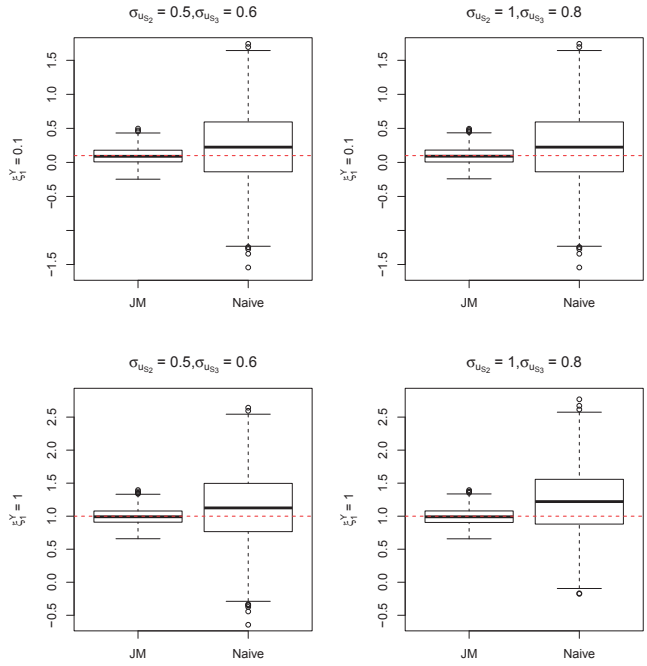


Figure 5.3: Simulation results: The point estimates of the covariate of interest from joint model and naive approach in longitudinal setting.

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.

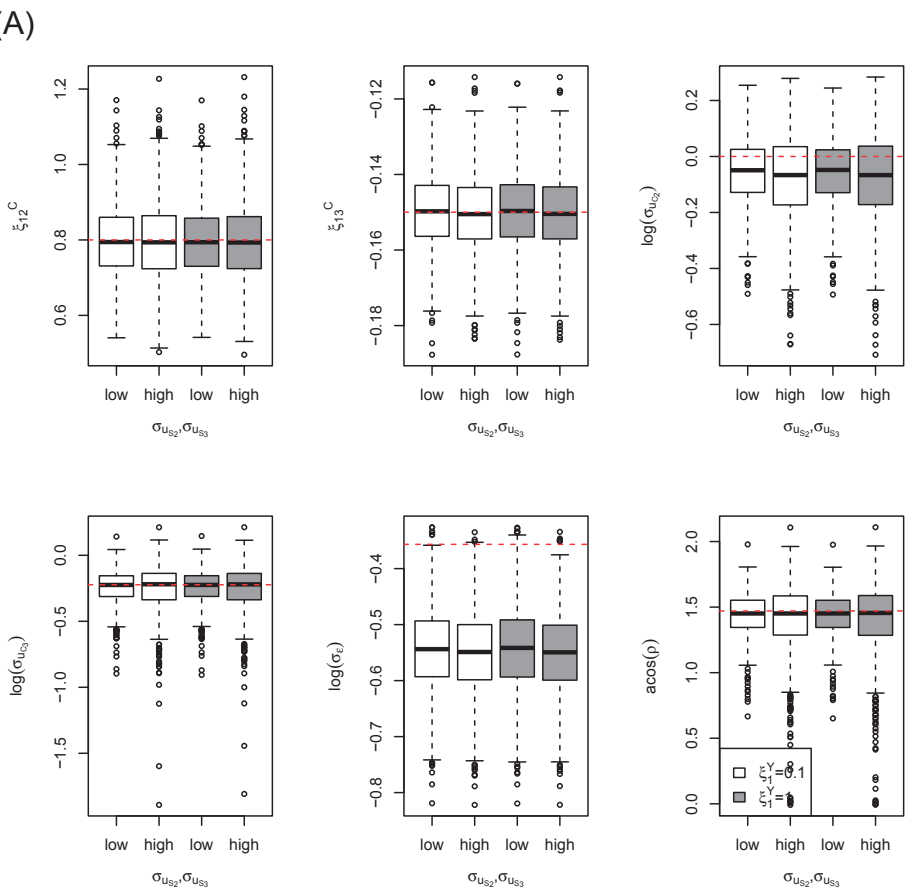


Figure 5.4: Simulation results: the point estimates of (A) categorical covariate effects as well as random effects (excluding the shared effects) for different standard deviations of shared effects, and (B) standard deviations of the shared effect at different effect size from joint model in longitudinal 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

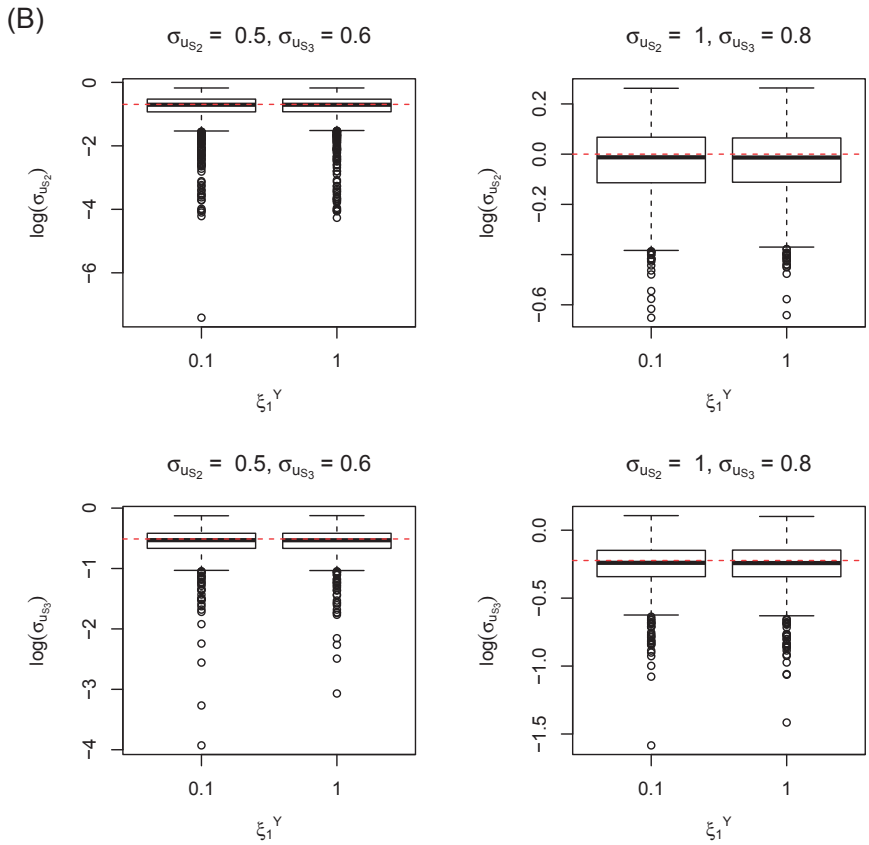


Figure 5.4: (cont.) Simulation results: the point estimates of (A) categorical covariate effects as well as random effects (excluding the shared effects) for different standard deviations of shared effects, and (B) standard deviations of the shared effect at different effect size from joint model in 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%.

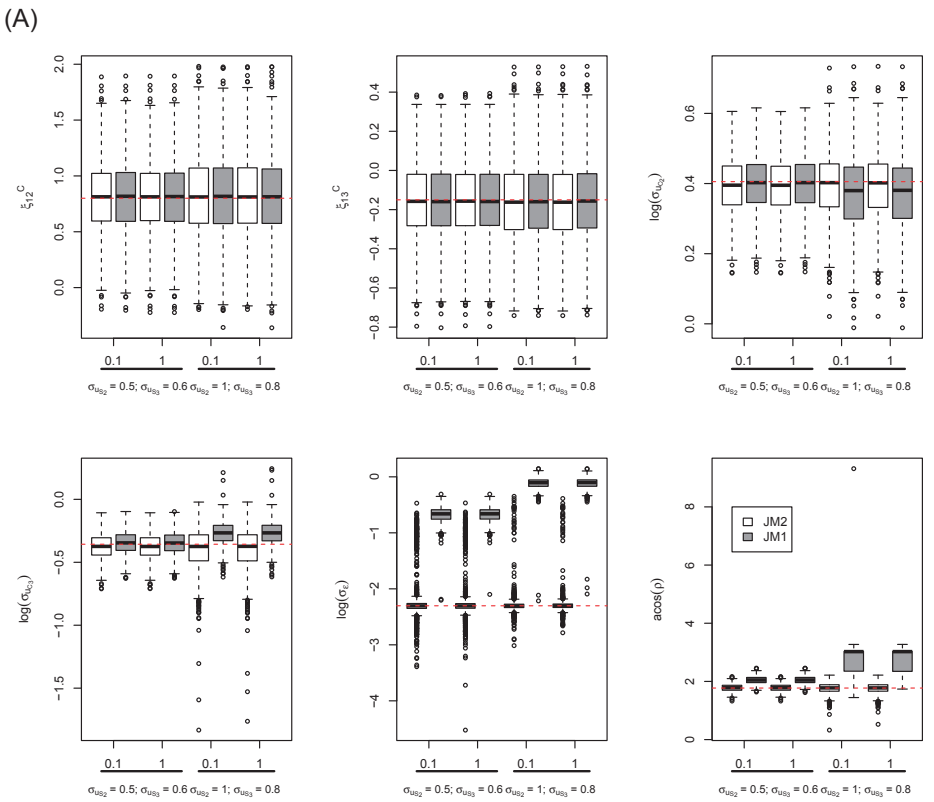


Figure 5.5: The robustness of (A) fixed effect and standard deviation of the random effect parameters and (B) standard deviations of shared effects in joint model in cross-sectional setting. Datasets were generated using the joint model with logit-dependent shared effect. The estimates were obtained 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-*

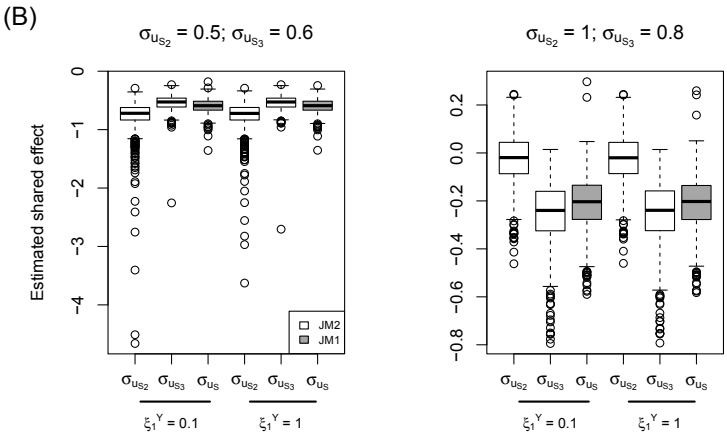


Figure 5.5: (cont.) The robustness of (A) fixed effect and standard deviation of the random effect parameters and (B) standard deviations of shared effects in joint model in cross-sectional setting. Datasets were generated using the joint model with logit-dependent shared effects. The estimates were 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.

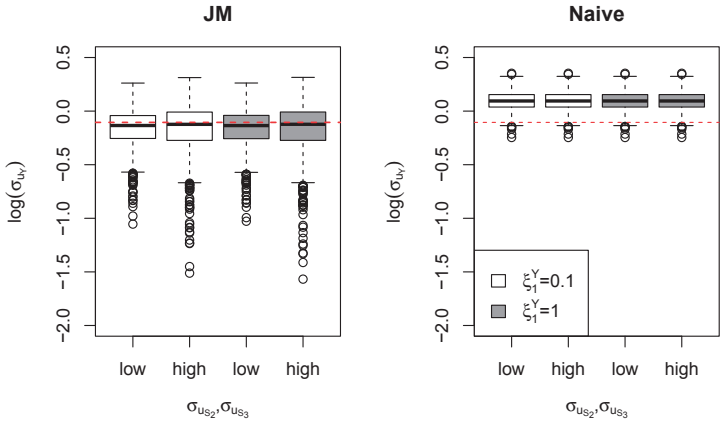


Figure 5.6: The estimates for random effect's variability of continuous outcome from joint model and naive approach in longitudinal setting. The horizontal lines represents the true value. Details about low and high are the same as in Figure 5.2

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

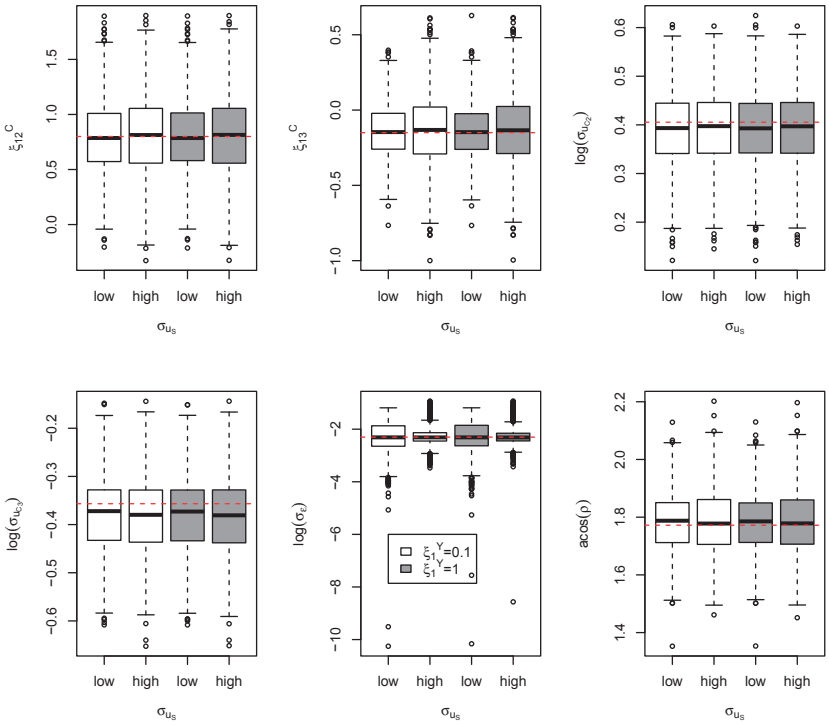


Figure S5.6.1: Simulation study at the cross-sectional setting: the distribution for all parameters under joint model with univariate random shared effect. Details of low and and high are the same as in Figure 5.2.

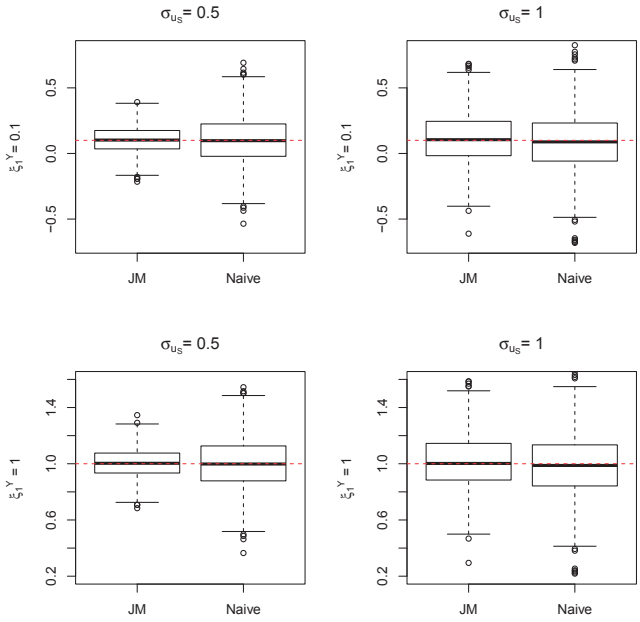


Figure S5.6.2: Simulation study at the cross-sectional setting: the point estimate for the effect of covariate of interest when dataset were generated using the joint model with a univariate random shared effects.

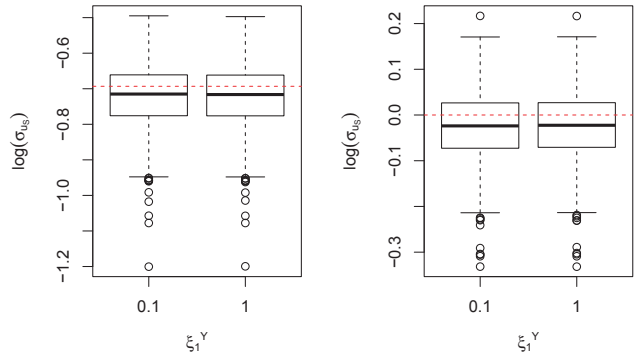


Figure S5.6.3: Simulation study at the cross-sectional setting: the distribution of the variability of random shared effect under the joint model with a univariate random shared effect

(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

6
General Discussion

In this thesis, several analyses of the gut microbiome composition in relation to health outcomes have been carried out. Randomized studies presented in this thesis utilized the observations of gut microbiome composition, cytokine responses and helminth infections at two different time-points, namely before and 21 months after the first treatment. The first part of the thesis deals with the analysis of gut microbiome and helminthiasis, while the second part deals with the three-way relationship between helminth infection, gut microbiome, and immune responses. The main purpose of this chapter is to assess how much evidence there is for the associations that are observed in this thesis to be causal. In line with this purpose, it is observed that many microbiome studies have been directed towards causality such as in the work of microbiota and metabolic diseases [Zhao (2013); Zhang and Zhao (2016)]. In analyzing the causal effect of certain exposure, it is important to minimize all possible biases, and to account for potential unobserved confounders or measurement errors. This chapter serves as a key to understanding whether the identified effect may be causal. The remaining of this chapter is organized as follows; the findings in epidemiological works as well as in the development of statistical methods are summarized, several basic terminologies of causal effect are briefly described, followed by a discussion of the findings. Finally, the conclusion is derived and directions for future research are listed.

6.1 Summary of the findings

In **Chapter 2**, treatment was significantly associated with microbiome composition only in subjects who had helminth infections and remained infected at 21 months after the first treatment. This significant association is also confirmed using a newly developed statistical method outlined in **Chapter 3**. In addition, the stability of gut microbiome composition over time is also confirmed by analyzing the microbiome composition of subjects who remained uninfected and did not receive albendazole at two time-points. When analysing the relationship between gut microbiome composition and immune responses, the microbiome composition is significantly associated with an immune response when subjects were helminth-uninfected but this association was not observed when subjects were helminth-infected (**Chapter 4**). When analyzing the association between helminth infection and both microbiome composition and immune responses jointly (**Chapter 5**), only gut microbiome composition is significantly associated with helminth infections.

In relation to statistical methodologies, this thesis contributes to the development of appropriate statistical models which address the features of compositional data and the collection design. The features of microbiome data are addressed, namely the compositional artifact, the presence of extra variation (overdispersion) due to unobserved causes and measurement errors. The compositional feature is addressed by multivariate approach, i.e. jointly modelling all bacterial taxa. This is done to avoid multiple testing correction when analyzing each bacteria taxa separately. The overdispersion is taken into account by introducing random effect in the model. When considering a distribution for the random effect of overdispersion, one could opt for a conjugate [Chen and Li (2013); Guimarães and Lindrooth (2007)] as it is done in **Chapter 3** or normal distribution [Hartzel et al. (2016); Hedeker (2003)] as it is done in **Chapter 5**. The measurement error is accounted for in the model by introducing additional normally distributed random effect. Finally, it has been shown in **Chapter 5** that modelling the association between helminth infection and different type of outcomes jointly in a hierarchical setting provides unbiased estimates. Another advantage from this joint modelling is enhancing the statistical power as multiple correction is not needed.

6.2 Basic terminologies of causal inference

Before conferring causal relationship in this thesis, basic terminologies of causal inference [Hernan and Robins (2018)] are briefly reviewed. In principle, a predictor has a causal effect on an outcome if the presence or absence of this predictor yields different responses [Rubin (1974)]. In a randomized controlled trial setting, as is the case in the study described in this thesis, the significant association be-

tween treatment and outcome is indeed causal since the counterfactual response can be quantified through a control group. When the randomized study is not possible, researchers rely on observational studies. The causal effect in observational design still can be estimated by utilizing an instrumental variable, i.e. a variable that has an effect on an outcome only via a predictor [Burgess and Small (2016)]. In fact, the method of instrumental variable is also useful for inferring total effect of predictor on outcome even in the presence of confounder [Hernán and Robins (2006b)]. To understand these terminologies as well as to identify the causal effect of variables involved in these analyses, directed acyclic graphs (DAGs) are used to visualize the relationship between variables of interests in this thesis. In these DAGs, vertices represent variables and arrows represent the direction from a cause to an effect.

In making inferences about causation from association study, one needs to be aware of the presence of confounders, colliders and measurement errors as these will strengthen or weaken the observed associations [Pourhoseingholi et al. (2012)]. A confounding bias is caused by the presence of a confounder, i.e. a variable that affects both predictor and outcome simultaneously. In the presence of a confounder, the association between predictor and outcome is no longer caused only by the predictor. This bias can be eliminated by conditioning (stratification or regression adjustment) on the confounder. Conversely, the presence of collider, i.e. variable that is affected by both predictor and outcome, will block an association between them. One needs to cautiously assess this relationship as conditioning on the collider will introduce bias [Hernan and Robins (2018)], i.e. observing a significant association while it actually does not exist. Finally, errors in measuring the variables need to be taken into account in the model.

6.3 Synthesis of findings

Suppose the associations observed in this thesis are indeed causal, then the relationship between anthelminthic treatment, helminth infections, gut microbiome and immune responses characterized by stimulated cytokine responses is illustrated in Figure 6.1. Note that it is assumed that treatment affect gut microbiome composition and cytokine are completely mediated via infection.

Here, it is considered that treatment as a covariate and the other variables (helminth infection, gut-microbiome and cytokine response) as outcomes. Since anthelminthic treatment was randomized, the causal effect of treatment on these three variables separately can be assessed, since association in randomized design is indeed causal. Let us focus on the relationship between infection and gut microbiome. As infection is not randomized, the causal effect of infection on gut microbiome cannot be assessed. However, treatment can be used as a proxy for this causal relationship under certain assumptions. Suppose that treatment has no effect on gut microbiome and treatment is only associated with gut micro-

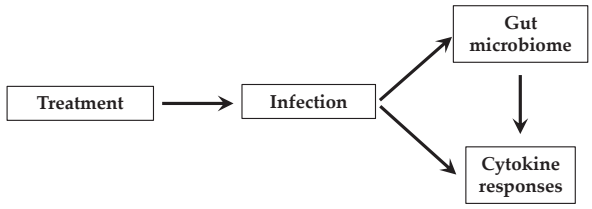


Figure 6.1: The hypothesized relationship based on the findings of our analyses.

biome via helminth infection, then treatment is an instrumental variable for the relationship between infection and gut microbiome. Thus, the causal effect of infection can be assessed via this instrumental variable [Burgess and Small (2016)]. In a similar way, it can be hypothesized that treatment is an instrumental variable in assessing the effect of helminth infection on cytokine response. However this is not true since a previous study by Wammes et al. (2016) showed that treatment was significantly associated with cytokine responses.

The assumption of treatment as an instrumental variable in the relationship between helminth infections and gut microbiome is hard to infer. The mechanism of albendazole on gut microbiome directly has not been fully analyzed [Leung et al. (2018)]. In our study, the relatively small sample size results in a lack of statistical power to identify a direct effect of albendazole on gut microbiome. Thus, at this moment treatment is not considered as an instrumental variable for this relationship.

Since we do not have an instrumental variable, we need to consider possible confounders for the relationship. In animal studies where mostly experimental in which helminth-free animals were introduced to the helminth parasite and other factors that could affect their gut microbiome were controlled (reviewed in Reynolds et al. (2015)). Animal models ensure that any changes in gut microbiome due to helminth exposure can be clearly quantified (reviewed in Zaiss and Harris (2016)). These studies conclude that helminth infections has a causal effect on gut microbiome. However for human studies, the sample size is either to0 small (this thesis) or the design is interventional or observational. Any alterations that were observed in gut microbiome composition might be confounded by other factors.

When considering the confounders that affect the gut microbiome composition in humans, dietary consumption and hygiene are major candidates [Gilbert et al. (2018)]. Dietary intake may also affect weight gain, and thus in Figure 6.2, the relationship with these additional variables (weight gain and hygiene) are added. As illustrated in Figure 6.2, hygiene affects both helminth infection and gut microbiome, thus it is a confounder for both helminth infection and gut microbiome. It is necessary to adjust for hygiene when quantifying the effect of

helminth infection on gut microbiome. However, in general, confounders may be difficult to measure or it may be unobserved. This will add an extra randomness in the exposure for each subjects. For this purpose, the inclusion of random effect subject-specific in the statistical model in a longitudinal setting takes care of this extra variation due to unobserved confounder.

In addition to confounders, there are several factors that could affect both helminth infections and microbiome composition. As can be seen in Figure 6.2, helminth infection is known to cause reduction of food intake and thus affect the body mass index (BMI) [Crompton and Nesheim (2002)]. Here, BMI plays a role as a mediator for the relationship between helminth infection and gut microbiome. Assessing both direct and indirect effects of helminth infection on microbiome composition is needed to identify the role of mediator and understand the underlying biology. Usually this indirect effect through a mediator is analyzed within the framework of linear structural equation models (LSEMs) [MacKinnon et al. (2007)].

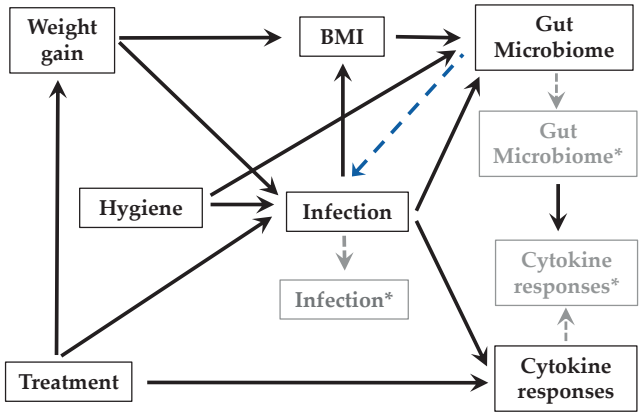


Figure 6.2: The DAG representing the relationship of all variables when measurement errors were included. The grey variables represent the observed variables with errors and blue line represents the possible causal direction.

In **Chapter 2** and **3**, treatment appeared to be significantly associated with gut microbiome only in subjects who had helminth infections. A current review describes the potential influence of gut microbiome on the presence of helminths in human intestinal tract by altering the immune system although the exact mechanism is still unknown [Rapin and Harris (2018)]. If this is indeed the case, as both gut microbiome and treatment influence helminth infections, thus infection becomes a collider. The association path between treatment and gut microbiome is blocked. This association is not causal as treatment is associated with gut mi-

crobiome given the subjects is helminth-infected. This could be the reason the effect of treatment is not observed in subjects who were helminth-uninfected.

Another concern in this randomized study is a possibility that the longer the time frame of the study, the more individual and contextual changes could occur [Wunsch et al. (2010)]. It has been reported that administration of albendazole in schoolchildren in Kenya [Stephenson et al. (1993)], Indonesia [Hadju et al. (1998)], and Uganda [Alderman et al. (2006)] for a period of more than 4 months increases the appetite and eventually weight gain. These may lead to lack of compliance. More importantly study in Ghana [Humphries et al. (2017)] reported the efficacy of albendazole treatment on removing helminth was strongly improved by nutrition factor. This shows that the effect of treatment in removing helminth may be mediated via the weight gain. As a consequence, in the long run, the assumption of randomized treatment is no longer held.

6.4 Measurement errors

Biomedical data are measured with errors. Firstly, helminth infection status was measured by PCR or microscopy. Microscopic examination as a conventional method to identify helminth infections potentially gives unreliable results especially in the case of light infection [Llewellyn et al. (2016); Khurana and Sethi (2017)]. On the other hand, researchers often classifying infection status based on PCR which is a reliable measurement, have to use a threshold as is the case in this thesis which can bring about error. Secondly, microbiome data was obtained through sequencing process which is not free of noise [Goodrich et al. (2014)]. The procedure undergoes the clustering process until the taxonomical count data is obtained [Robinson et al. (2016)]. Thirdly, the data generated from assays that measure cytokine levels may be censored by detection limit and as a result data might be skewed. To deal with this caveat, transformation of the data using logarithm transformation was done so that the transformed data conform with normal distribution. However, such a transformation might not reduce the variability in the data.

In practice, researchers only observe variables which are measured with errors, as depicted by the relationship in grey in Figure 6.2. These measurement errors could occur in any study design [Hernan and Robins (2018)] and when it is left unaccounted for in the analyses, it weakens or strengthens the association between outcome and predictor. In **Chapter 4** of this thesis, the relationship between helminth infection, gut microbiome and cytokine responses were analyzed by ignoring the measurement error. It is shown in the simulation study in **Chapter 5** that ignoring the measurement error might give biased regression estimates.

Considering the above discussions with regard to the observed significant associations and their possible confounders. Firstly we believe that the effect of treatment on helminth infections is causal as treatment is randomized and the

effect of the long time frame via gain in weight is likely to be small. Secondly, we believe that the effect of helminth infection on microbiome composition and on cytokine responses are causal, because we assume that the random effects used in modelling the repeated measurements takes care of most of the confounders (Figure 6.3). The relationship between gut microbiome and cytokine responses is not discussed here since it is shown in Chapter 5 that these outcomes are not correlated.

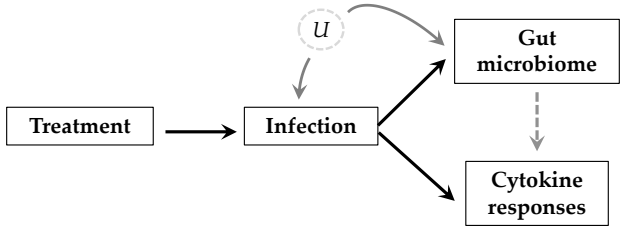


Figure 6.3: The concluded causal effect. The variable U represents latent variable to account for unobserved confounders.

6.5 Future directions

To conclude, this general discussion highlights the critical considerations when moving from association to causation in microbiome studies. Researchers should specify the relationship of the studied variables, identify potential biases and use proper statistical methods that account for these challenges. The study design used in this thesis is key for causal inferences and the statistical methods developed in this thesis illustrates a solution to obtain unbiased estimates of the relationship between variables.

The findings that gut microbiome is related to obesity and several metabolic diseases have shown that the relationship might be causal. With regard to this direction, it is important to understand the biological mechanism that underlying the relationship between infection, gut microbiome, and cytokine response. It has been shown in the above DAGs that gut microbiome could be a potential mediator for the relationship between infection and cytokine responses. To this end, work on mediation analysis is limited on single variable and not in the compositional variable and the statistical analysis framework for this purpose is still limited. This could be another direction for future research.

The framework developed in **Chapter 5** can be extended to include multiple omics type data to unravel the complex mechanism of gut microbiota. Recent findings show that gut microbiota produces metabolites that regulate the

immune-homeostasis [Thorburn et al. (2014)]. Thus, to understand the relationship between gut microbiome and immune system, more research with regard to this metabolite is needed.

In relation to the development of appropriate statistical model which account for the unobserved confounders, two distributional assumptions were made in this thesis, namely the conjugate and normal distribution. However, there is still lack of method to assess models' goodness of fit. A statistical method needs to be developed for that purpose. Further research is needed in this direction.

In the joint model in **Chapter 5**, the random effect describing the measurement error is assumed to be the same for two time-points due to computational burden. This assumption may not be true. More research is needed to analyzed different random effect structure to model the measurement error.

Bibliography

(2012). A framework for human microbiome research. *Nature 486*(7402), 215–221.

Abrams, G. D., H. BAUER, and H. SPRINZ (1963). Influence of the normal flora on mucosal morphology and cellular renewal in the ileum. a comparison of germ-free and conventional mice. *Laboratory investigation; a journal of technical methods and pathology 12*, 355–364.

Agresti, A. (2013). *Categorical data analysis* (3rd ed. ed.), Volume 792 of *Wiley series in probability and statistics*. Hoboken, NJ: Wiley.

Albonico, M., H. Allen, L. Chitsulo, D. Engels, A.-F. Gabrielli, and L. Savioli (2008). Controlling soil-transmitted helminthiasis in pre-school-age children through preventive chemotherapy. *PLoS neglected tropical diseases 2*(3), e126.

Alderman, H., J. Konde-Lule, I. Sebuliba, D. Bundy, and A. Hall (2006). Effect on weight gain of routinely giving albendazole to preschool children during child health days in uganda: cluster randomised controlled trial. *BMJ (Clinical research ed.) 333*(7559), 122.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, M. Antolín, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, and P. Bork (2011). Enterotypes of the human gut microbiome. *Nature 473*(7346), 174–180.

Bach, J.-F. (2002). The effect of infections on susceptibility to autoimmune and allergic diseases. *The New England journal of medicine 347*(12), 911–920.

Barron, L. K., B. B. Warner, P. I. Tarr, W. D. Shannon, E. Deych, and B. W. Warner (2017). Independence of gut bacterial content and neonatal necrotizing enterocolitis severity. *Journal of pediatric surgery 52*(6), 993–998.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1).

Belkaid, Y. and T. W. Hand (2014). Role of the microbiota in immunity and inflammation. *Cell 157*(1), 121–141.

Bethony, J., S. Brooker, M. Albonico, S. M. Geiger, A. Loukas, D. Diemert, and P. J. Hotez (2006). Soil-transmitted helminth infections: Ascariasis, trichuriasis, and hookworm. *The Lancet 367*(9521), 1521–1532.

Booth, J. G., G. Casella, H. Friedl, and J. P. Hobert (2003). Negative binomial loglinear mixed models. *Statistical Modelling 3*(3), 179–191.

Broadhurst, M. J., A. Ardeshir, B. Kanwar, J. Mirpuri, U. M. Gundra, J. M. Leung, K. E. Wiens, I. Vujkovic-Cvijin, C. C. Kim, F. Yarovinsky, N. W. Lerche, J. M. McCune, and P. Loke (2012). Therapeutic helminth infection of macaques with idiopathic chronic diarrhea alters the inflammatory signature and mucosal microbiota of the colon. *PLoS pathogens 8*(11), e1003000.

Burgess, S. and D. S. Small (2016). Predicting the direction of causal effect based on an instrumental variable analysis: A cautionary tale. *Journal of Causal Inference 0*(0), 1525.

Catalano, P. J. and L. M. Ryan (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association 87*(419), 651.

Catalano, P. J., D. O. Scharfstein, L. M. Ryan, C. A. Kimmel, and G. L. Kimmel (1993). Statistical model for fetal death, fetal weight, and malformation in developmental toxicity studies. *Teratology 47*(4), 281–290.

Chen, J. and H. Li (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics 7*(1).

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje (2014). Ribosomal database project: data and tools for high throughput rrna analysis. *Nucleic acids research 42*(Database issue), D633–42.

Collender, P. A., A. E. Kirby, D. G. Addiss, M. C. Freeman, and J. V. Remais (2015). Methods for quantification of soil-transmitted helminths in environmental media: Current techniques and recent advances. *Trends in parasitology 31*(12), 625–639.

Conlon, M. A. and A. R. Bird (2014). The impact of diet and lifestyle on gut microbiota and human health. *Nutrients 7*(1), 17–44.

Cooper, P., A. W. Walker, J. Reyes, M. Chico, S. J. Salter, M. Vaca, and J. Parkhill (2013). Patent human infections with the whipworm, trichuris trichiura, are not associated with alterations in the faecal microbiota. *PloS one 8*(10), e76573.

Crompton, D. W. T. and M. C. Nesheim (2002). Nutritional impact of intestinal helminthiasis during the human life cycle. *Annual review of nutrition 22*, 35–59.

de Filippo, C., D. Cavalieri, M. Di Paola, M. Ramazzotti, J. B. Poullet, S. Massart, S. Collini, G. Pieraccini, and P. Lionetti (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proceedings of the National Academy of Sciences of the United States of America 107*(33), 14691–14696.

de Filippo, C., M. Di Paola, M. Ramazzotti, D. Albanese, G. Pieraccini, E. Banci, F. Miglietta, D. Cavalieri, and P. Lionetti (2017). Diet, environments, and gut microbiota. a preliminary investigation in children living in rural and urban burkina faso and italy. *Frontiers in microbiology 8*, 1979.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman (2005). Diversity of the human intestinal microbial flora. *Science (New York, N.Y.) 308*(5728), 1635–1638.

Efendi, A., G. Molenberghs, and S. Iddi (2014). A marginalized combined gamma frailty and normal random-effects model for repeated, overdispersed, time-to-event outcomes. *Communications in Statistics - Theory and Methods 43*(22), 4806–4828.

Endara, P., M. Vaca, M. E. Chico, S. Erazo, G. Oviedo, I. Quinzo, A. Rodriguez, R. Lovato, A.-L. Moncayo, M. L. Barreto, L. C. Rodrigues, and P. J. Cooper (2010). Long-term periodic anthelmintic treatments are associated with increased allergen skin reactivity. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology 40*(11), 1669–1677.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Gazzinelli-Guimaraes, P. H. and T. B. Nutman (2018). Helminth parasites and immune regulation. *F1000Research 7*.

Gensollen, T., S. S. Iyer, D. L. Kasper, and R. S. Blumberg (2016). How colonization by microbiota in early life shapes the immune system. *Science (New York, N.Y.) 352*(6285), 539–544.

Geys, H., P. Catalano, and C. Faes (2008). Joint models for continuous and discrete longitudinal data. In G. Verbeke, M. Davidian, G. Fitzmaurice, and G. Molenberghs (Eds.), *Longitudinal Data Analysis*, Volume 20085746 of *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pp. 327–348. Chapman and Hall/CRC.

Gilbert, J. A., M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight (2018). Current understanding of the human microbiome. *Nature medicine 24*(4), 392–400.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in microbiology 8*, 2224.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight, and R. E. Ley (2014). Conducting a microbiome study. *Cell 158*(2), 250–262.

Grice, E. A. and J. A. Segre (2012). The human microbiome: our second genome. *Annual review of genomics and human genetics 13*, 151–170.

Group HDGW. Generalized draft form of hmp data generation working group 16s 454 default protocol version 4.2- pilot study p.1 (15).

Gueorguieva, R. (2016). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling: An International Journal 1*(3), 177–193.

Guimarães, P. and R. C. Lindrooth (2007). Controlling for overdispersion in grouped conditional logit models: A computationally simple application of dirichlet-multinomial regression. *The Econometrics Journal 10*(2), 439–452.

Gupta, V. K., S. Paul, and C. Dutta (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology 8*, 1162.

Hadju, V., L. S. Stephenson, H. O. Mohammed, D. D. Bowman, and R. S. Parker (1998). Improvements of growth, appetite, and physical activity in helminth-infected schoolboys 6 months after single dose of albendazole. *Asia Pacific journal of clinical nutrition 7*(2), 170–176.

Hall, A., G. Hewitt, V. Tuffrey, and N. de Silva (2008). A review and meta-analysis of the impact of intestinal worms on child growth and nutrition. *Maternal & child nutrition 4 Suppl 1*, 118–236.

Hall, N. S. (2007). R. a. fisher and his advocacy of randomization. *Journal of the History of Biology 40*(2), 295–325.

Haro, C., O. A. Rangel-Zúñiga, J. F. Alcalá-Díaz, F. Gómez-Delgado, P. Pérez-Martínez, J. Delgado-Lista, G. M. Quintana-Navarro, B. B. Landa, J. A. Navas-Cortés, M. Tena-Sempere, J. C. Clemente, J. López-Miranda, F. Pérez-Jiménez, and A. Camargo (2016). Intestinal microbiota is influenced by gender and body mass index. *PloS one 11*(5), e0154090.

Hartzel, J., A. Agresti, and B. Caffo (2016). Multinomial logit random effects models. *Statistical Modelling: An International Journal 1*(2), 81–102.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics 55*(3), 688–698.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in medicine 22*(9), 1433–1446.

Hernán, M. A. and J. M. Robins (2006a). Estimating causal effects from epidemiological data. *Journal of epidemiology and community health 60*(7), 578–586.

Hernán, M. A. and J. M. Robins (2006b). Instruments for causal inference: an epidemiologist's dream? *Epidemiology (Cambridge, Mass.) 17*(4), 360–372.

Hernan, M. A. and J. M. Robins (2018). *Causal inference*. Chapman & Hall/CRC monographs on statistics & applied probability. Boca Raton, Fla. and London: CRC and Taylor & Francis [distributor].

Holm, J. B., D. Sorobetea, P. Kiilerich, Y. Ramayo-Caldas, J. Estellé, T. Ma, L. Madsen, K. Kristiansen, and M. Svensson-Frej (2015). Chronic trichuris muris infection decreases diversity of the intestinal microbiota and concomitantly increases the abundance of lactobacilli. *PloS one 10*(5), e0125495.

Hotez, P. J., P. J. Brindley, J. M. Bethony, C. H. King, E. J. Pearce, and J. Jacobson (2008). Helminth infections: the great neglected tropical diseases. *The Journal of clinical investigation 118*(4), 1311–1321.

Humphries, D., S. Nguyen, S. Kumar, J. E. Quagraine, J. Otchere, L. M. Harrison, M. Wilson, and M. Cappello (2017). Effectiveness of albendazole for hookworm varies widely by community and correlates with nutritional factors: A cross-sectional study of school-age children in ghana. *The American journal of tropical medicine and hygiene 96*(2), 347–354.

Iddi, S. and G. Molenberghs (2012). A joint marginalized multilevel model for longitudinal outcomes. *Journal of Applied Statistics 39*(11), 2413–2430.

Jackson, C. H., N. G. Best, and S. Richardson (2008). Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society, Series A: Statistics In Society 171*(1), 159–178.

Kassahun, W., T. Neyens, G. Molenberghs, C. Faes, and G. Verbeke (2013). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation 85*(3), 552–571.

Kay, G. L., A. Millard, M. J. Sergeant, N. Midzi, R. Gwisai, T. Mduluza, A. Ivens, N. Nausch, F. Mutapi, and M. Pallen (2015). Differences in the faecal microbiome in schistosoma haematobium infected children vs. uninfected children. *PLoS neglected tropical diseases 9*(6), e0003861.

Khurana, S. and S. Sethi (2017). Laboratory diagnosis of soil transmitted helminthiasis. *Tropical parasitology 7*(2), 86–91.

Koliada, A., G. Syzenko, V. Moseiko, L. Budovska, K. Puchkov, V. Perederiy, Y. Gavalko, A. Dorofeyev, M. Romanenko, S. Tkach, L. Sineok, O. Lushchak, and A. Vaiserman (2017). Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. *BMC microbiology 17*(1), 120.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software 82*(13).

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963.

Lee, S. C., M. S. Tang, Y. A. L. Lim, S. H. Choy, Z. D. Kurtz, L. M. Cox, U. M. Gundra, I. Cho, R. Bonneau, M. J. Blaser, K. H. Chua, and P. Loke (2014). Helminth colonization is associated with increased diversity of the gut microbiota. *PLoS neglected tropical diseases 8*(5), e2880.

Leung, J. M., A. L. Graham, and S. C. L. Knowles (2018). Parasite-microbiota interactions with the vertebrate gut: Synthesis through an ecological lens. *Frontiers in microbiology 9*, 843.

Ley, R. E., D. A. Peterson, and J. I. Gordon (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell 124*(4), 837–848.

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application 2*(1), 73–94.

Li, R. W., S. Wu, W. Li, K. Navarro, R. D. Couch, D. Hill, and J. F. Urban (2012). Alterations in the porcine colon microbiota induced by the gastrointestinal nematode trichuris suis. *Infection and immunity 80*(6), 2150–2157.

Lin, A., E. M. Bik, E. K. Costello, L. Dethlefsen, R. Haque, D. A. Relman, and U. Singh (2013). Distinct distal gut microbiome diversity and composition in healthy children from bangladesh and the united states. *PloS one 8*(1), e53838.

Liu, Q. and D. A. Pierce (1994). A note on gauss-hermite quadrature. *Biometrika 81*(3), 624.

Llewellyn, S., T. Inpankaew, S. V. Nery, D. J. Gray, J. J. Verweij, A. C. A. Clements, S. J. Gomes, R. Traub, and J. S. McCarthy (2016). Application of a multiplex quantitative pcr to assess prevalence and intensity of intestinal parasite infections in a controlled clinical trial. *PLoS neglected tropical diseases 10*(1), e0004380.

Lumley, T., P. K. Diehr, S. S. Emerson, and L. Chen (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health 23*, 151–69.

MacKinnon, D. P., A. J. Fairchild, and M. S. Fritz (2007). Mediation analysis. *Annual review of psychology 58*, 593–614.

Macpherson, A. J. and N. L. Harris (2004). Interactions between commensal intestinal bacteria and the immune system. *Nature reviews. Immunology 4*(6), 478–485.

Martin, I., Y. Djuardi, E. Sartono, B. A. Rosa, T. Supali, M. Mitreva, J. J. Houwing-Duistermaat, and M. Yazdanbakhsh (2018). Dynamic changes in human-gut microbiome in relation to a placebo-controlled anthelminthic trial in indonesia. *PLoS neglected tropical diseases 12*(8), e0006620.

Mazmanian, S. K., J. L. Round, and D. L. Kasper (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature 453*(7195), 620–625.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical methods in medical research 17*(1), 53–73.

McSorley, H. J. and R. M. Maizels (2012). Helminth infections and host immune regulation. *Clinical microbiology reviews 25*(4), 585–608.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis 13*(4), 513–531.

Molenberghs, G., G. Verbeke, C. G. B. Demétrio, and A. M. C. Vieira (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science 25*(3), 325–347.

Molenberghs, G., G. Verbeke, S. Iddi, and C. G. Demétrio (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis 111*, 94–109.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β- distribution, and correlations among proportions. *Biometrika 49*(1/2), 65.

Murtagh, F. and P. Legendre (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification 31*(3), 274–295.

Mysara, M., P. Vandamme, R. Props, F.-M. Kerckhof, N. Leys, N. Boon, J. Raes, and P. Monsieurs (2017). Reconciliation between operational taxonomic units and species boundaries. *FEMS microbiology ecology 93*(4).

Neuhaus, A., T. Augustin, C. Heumann, and D. Daumer (2009). A review on joint models in biometrical research. *Journal of Statistical Theory and Practice 3*(4), 855–868.

Oksanen, J., F. G. Blanchet, M. Friendly, and R. Kindt (2017). R-package: Community ecology package.

Park, S.-H., K.-A. Kim, Y.-T. Ahn, J.-J. Jeong, C.-S. Huh, and D.-H. Kim (2015). Comparative analysis of gut microbiota in elderly people of urbanized towns and longevity villages. *BMC microbiology 15*, 49.

Pham, T. V. (2013). ibb: The (inverted) beta-binomial test for count data. r package.

Pham, T. V. and C. R. Jimenez (2012). An accurate paired sample test for count data. *Bioinformatics (Oxford, England) 28*(18), i596–i602.

Pourhoseingholi, M. A., A. R. Baghestani, and M. Vahedi (2012). How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench 5*(2), 79–83.

R Core Team. R: A language and environment for statistical computing.

Ramanan, D., R. Bowcutt, S. C. Lee, M. S. Tang, Z. D. Kurtz, Y. Ding, K. Honda, W. C. Gause, M. J. Blaser, R. A. Bonneau, Y. A. L. Lim, P. Loke, and K. Cadwell (2016). Helminth infection promotes colonization resistance via type 2 immunity. *Science (New York, N.Y.) 352*(6285), 608–612.

Rapin, A. and N. L. Harris (2018). Helminth-bacterial interactions: Cause and consequence. *Trends in immunology 39*(9), 724–733.

Reynolds, L. A., B. B. Finlay, and R. M. Maizels (2015). Cohabitation in the intestine: Interactions among helminth parasites, bacterial microbiota, and host immunity. *Journal of immunology (Baltimore, Md. : 1950) 195*(9), 4059–4066.

Robinson, C. K., R. M. Brotman, and J. Ravel (2016). Intricacies of assessing the human microbiome in epidemiologic studies. *Annals of epidemiology 26*(5), 311–321.

Rodrigues Hoffmann, A., L. M. Proctor, M. G. Surette, and J. S. Suchodolski (2016). The microbiome: The trillions of microorganisms that maintain health and cause disease in humans and companion animals. *Veterinary pathology 53*(1), 10–21.

Rook, G. A. W. (2009). Review series on helminths, immune modulation and the hygiene hypothesis: the broader implications of the hygiene hypothesis. *Immunology 126*(1), 3–11.

Rosa, B. A., T. Supali, L. Gankpala, Y. Djuardi, E. Sartono, Y. Zhou, K. Fischer, J. Martin, R. Tyagi, F. K. Bolay, P. U. Fischer, M. Yazdanbakhsh, and M. Mitreva (2018). Differential human gut microbiome assemblages during soil-transmitted helminth infections in indonesia and liberia. *Microbiome 6*(1), 33.

Rosenthal, M., A. E. Aiello, C. Chenoweth, D. Goldberg, E. Larson, G. Gloor, and B. Foxman (2014). Impact of technical sources of variation on the hand microbiome dynamics of healthcare workers. *PloS one 9*(2), e88999.

Round, J. L. and S. K. Mazmanian (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature reviews. Immunology 9*(5), 313–323.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual review of microbiology 31*, 107–133.

Schloss, P. D., D. Gevers, and S. L. Westcott (2011). Reducing the effects of pcr amplification and sequencing artifacts on 16s rrna-based studies. *PloS one 6*(12), e27310.

Sonnenburg, J. L. and F. Bäckhed (2016). Diet-microbiota interactions as moderators of human metabolism. *Nature 535*(7610), 56–64.

Stephenson, L. S., M. C. Latham, E. J. Adams, S. N. Kinoti, and A. Pertet (1993). Physical fitness, growth and appetite of kenyan school boys with hookworm, trichuris trichiura and ascaris lumbricoides infections are improved four months after a single dose of albendazole. *The Journal of nutrition 123*(6), 1036–1046.

Teixeira-Pinto, A., J. Siddique, R. Gibbons, and S.-L. Normand (2009). Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric Annals 39*(7), 729–735.

Thorburn, A. N., L. Macia, and C. R. Mackay (2014). Diet, metabolites, and western-lifestyle inflammatory diseases. *Immunity 40*(6), 833–842.

Thursby, E. and N. Juge (2017). Introduction to the human gut microbiota. *The Biochemical journal 474*(11), 1823–1836.

Torbati, M. E., M. Mitreva, and V. Gopalakrishnan (2016). Application of taxonomic modeling to microbiota data mining for detection of helminth infection in global populations. *Data 1*(3).

Tsonaka, R., D. van der Woude, and J. Houwing-Duistermaat (2015). Marginal genetic effects estimation in family and twin studies using random-effects models. *Biometrics 71*(4), 1130–1138.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon (2007). The human microbiome project. *Nature 449*(7164), 804–810.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature 444*(7122), 1027–1031.

Turnbaugh, P. J., V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight, and J. I. Gordon (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine 1*(6), 6ra14.

Tutz, G. (2012). *Regression for categorical data*. Cambridge series in statistical and probabilistic mathematics. Cambridge and New York: Cambridge University Press.

Ursell, L. K., J. L. Metcalf, L. W. Parfrey, and R. Knight (2012). Defining the human microbiome. *Nutrition reviews 70 Suppl 1*, S38–44.

Verbeke, G., S. Fieuws, G. Molenberghs, and M. Davidian (2014). The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research 23*(1), 42–59.

Verweij, J. J., E. A. T. Brienen, J. Ziem, L. Yelifari, A. M. Polderman, and L. van Lieshout (2007). Simultaneous detection and quantification of ancylostoma duodenale, necator americanus, and oesophagostomum bifurcum in fecal samples using multiplex real-time pcr. *The American journal of tropical medicine and hygiene 77*(4), 685–690.

Wammes, L. J., F. Hamid, A. E. Wiria, L. May, M. M. M. Kaisar, M. A. Prasetyani-Gieseler, Y. Djuardi, H. Wibowo, Y. C. M. Kruize, J. J. Verweij, S. E. de Jong, R. Tsonaka, J. J. Houwing-Duistermaat, E. Sartono, A. J. F. Luty, T. Supali, and M. Yazdanbakhsh (2016). Community deworming alleviates geohelminth-induced immune hyporesponsiveness. *Proceedings of the National Academy of Sciences of the United States of America 113*(44), 12526–12531.

Wammes, L. J., H. Mpairwe, A. M. Elliott, and M. Yazdanbakhsh (2014). Helminth therapy or elimination: Epidemiological, immunological, and clinical considerations. *The Lancet Infectious Diseases 14*(11), 1150–1162.

Wegener Parfrey, L., M. Jirků, R. Šíma, M. Jalovecká, B. Sak, K. Grigore, and K. Jirků Pomajbíková (2017). A benign helminth alters the host immune system and the gut microbiota in a rat model system. *PloS one 12*(8), e0182205.

Weiss, S., Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome 5*(1), 27.

Wexler, H. M. (2007). Bacteroides: the good, the bad, and the nitty-gritty. *Clinical microbiology reviews 20*(4), 593–621.

White, E. C., A. Houlden, A. J. Bancroft, K. S. Hayes, M. Goldrick, R. K. Grencis, and I. S. Roberts (2018). Manipulation of host and parasite microbiotas: Survival strategies during chronic nematode infection. *Science advances 4*(3), eaap7399.

Whitley, E. and J. Ball (2002). *Critical Care 6*(6), 509.

Wilson, M. S., M. D. Taylor, A. Balic, C. A. M. Finney, J. R. Lamb, and R. M. Maizels (2005). Suppression of allergic airway inflammation by helminth-induced regulatory t cells. *The Journal of experimental medicine 202*(9), 1199–1212.

Wiria, A. E., Y. Djuardi, T. Supali, E. Sartono, and M. Yazdanbakhsh (2012). Helminth infection in populations undergoing epidemiological transition: a friend or foe? *Seminars in immunopathology 34*(6), 889–901.

Wiria, A. E., F. Hamid, L. J. Wammes, M. M. M. Kaisar, L. May, M. A. Prasetyani, S. Wahyuni, Y. Djuardi, I. Ariawan, H. Wibowo, B. Lell, R. Sauerwein, G. T.

Brice, I. Sutanto, L. van Lieshout, A. J. M. de Craen, R. van Ree, J. J. Verweij, R. Tsonaka, J. J. Houwing-Duistermaat, A. J. F. Luty, E. Sartono, T. Supali, and M. Yazdanbakhsh (2013). The effect of three-monthly albendazole treatment on malarial parasitemia and allergy: a household-based cluster-randomized, double-blind, placebo-controlled trial. *PloS one 8*(3), e57899.

Wiria, A. E., M. A. Prasetyani, F. Hamid, L. J. Wammes, B. Lell, I. Ariawan, H. W. Uh, H. Wibowo, Y. Djuardi, S. Wahyuni, I. Sutanto, L. May, A. J. F. Luty, J. J. Verweij, E. Sartono, M. Yazdanbakhsh, and T. Supali (2010). Does treatment of intestinal helminth infections influence malaria? background and methodology of a longitudinal study of clinical, parasitological and immunological parameters in nangapanda, flores, indonesia (immunospin study). *BMC infectious diseases 10*, 77.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.) 334*(6052), 105–108.

Wunsch, G., F. Russo, and M. Mouchart (2010). Do we necessarily need longitudinal data to infer causal relations? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 106*(1), 5–18.

Xia, F., J. Chen, W. K. Fung, and H. Li (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics 69*(4), 1053–1063.

Xia, Y. and J. Sun (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & diseases 4*(3), 138–148.

Xie, Y. (2018). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.21.

Xu, L., A. D. Paterson, W. Turpin, and W. Xu (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one 10*(7), e0129606.

Yadav, M., M. K. Verma, and N. S. Chauhan (2018). A review of metabolic potential of human gut microbiome in human nutrition. *Archives of microbiology 200*(2), 203–217.

Yang, Y. and J. Kang (2010). Joint analysis of mixed poisson and continuous longitudinal data with nonignorable missing values. *Computational Statistics & Data Analysis 54*(1), 193–207.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon (2012). Human gut microbiome viewed across age and geography. *Nature 486*(7402), 222–227.

Yazdanbakhsh, M., P. G. Kremsner, and R. van Ree (2002). Allergy, parasites, and the hygiene hypothesis. *Science (New York, N.Y.) 296*(5567), 490–494.

Zaiss, M. M. and N. L. Harris (2016). Interactions between the intestinal microbiome and helminth parasites. *Parasite immunology 38*(1), 5–11.

Zaiss, M. M., A. Rapin, L. Lebon, L. K. Dubey, I. Mosconi, K. Sarter, A. Piersigilli, L. Menin, A. W. Walker, J. Rougemont, O. Paerewijck, P. Geldhof, K. D. McCoy, A. J. Macpherson, J. Croese, P. R. Giacomin, A. Loukas, T. Junt, B. J. Marsland, and N. L. Harris (2015). The intestinal microbiota contributes to the ability of helminths to modulate allergic inflammation. *Immunity 43*(5), 998–1010.

Zhang, C. and L. Zhao (2016). Strain-level dissection of the contribution of the gut microbiome to human metabolic disease. *Genome medicine 8*(1), 41.

Zhang, Y. and H. Zhou (2017). R-package: Multivariate response generalized linear models.

Zhao, L. (2013). The gut microbiota and obesity: from correlation to causality. *Nature reviews. Microbiology 11*(9), 639–647.

Zhou, Y. and F. Zhi (2016). Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: A meta-analysis. *BioMed research international 2016*, 5828959.

Summary

Rapid urbanization is almost always accompanied by a transition from infectious diseases to noncommunicable inflammatory disorders as a dominant cause of morbidity. This is likely due to changing lifestyle and environmental factors. To understand a precise mechanism of this transition, a population study was done in Nangapanda, Ende district in Indonesia. This area was chosen as chronic parasitic worm infections were endemic and lifestyle changes occurred at a rapid pace. Despite its detrimental effect on human health, parasitic helminth infections are associated with a strong modulation of immune responses which explains a low prevalence of inflammatory disorders in areas endemic for parasitic worms. To investigate the complex biological mechanism underlying this association, clinical and biomedical data were collected from subjects using a household-based cluster-randomized, double-blind, placebo-controlled trial. Specifically, studies in this thesis focus on analyzing the relationships between helminth infections, gut microbiome composition, and immune responses. Considering the complexities of the data gathered in this study, the available methods appeared to be limited, hence the development of statistical methodology is another focus of this thesis.

Chapter 1 provides a general introduction to the thesis with regard to the collected data, the research questions and the available statistical methods for the analysis purpose. The pyrosequencing procedure to obtain microbiome profiles for each sample is briefly described. Such a process imposes a compositional structure on the microbiome data which needs to be accounted for in the modeling. In addition, multiple observations from the same subject were collected, which yields a correlation structure between measurements. Statistical tools used to analyze microbiome data are reviewed and challenges for proper modeling of this type of data in a repeated measurement design are discussed.

Chapter 2 describes the application of a recently developed statistical method to model the microbiome data collected for this study. This model assumes that microbiome data are realizations of a multinomial distribution. In order to account for the presence of extra variation, the parameters of this multinomial distribution are assumed to be random effects following a conjugate distribution. However, the method is only valid for independent multinomial observations

and thus the correlated structure in our data due to repeated measurements is left unaccounted for in the modeling. Therefore, the analyses were carried out at each time point separately. The effect of helminth infection on the gut microbiome composition is analyzed using the data at pre-treatment while the effect of anthelminthic treatment on microbiome composition is assessed at post-treatment. To investigate whether the treatment has a different effect on subjects who were helminth-infected compared to helminth-uninfected, an interaction term between infection status and treatment is included in the model. It appears that only in subjects who received anthelminthic treatment and remained infected at both pre- and post-treatment, the ratio of Bacteroidetes to Firmicutes and the ratio of Actinobacteria to Firmicutes significantly differed compared to other groups. The method here is limited to the analysis of data from one time point, hence the alteration of microbiome composition over time cannot be analyzed. Chapter 3 attempts to develop a model to address this.

In Chapter 3, a statistical method for modeling repeatedly measured microbiome data is developed which addresses the correlation structure between a subject's multiple observations. This is done by introducing a normally distributed random effect. Three different covariance structures for the normally distributed random effects are considered. Firstly, we assume a univariate subject-specific random effect where the random effect for each bacterial category at different time points is the same. Secondly, each category has a different random effect with category-specific variance. Finally, it is assumed that the multivariate random effects have a common variance for all categories. A simulation study was conducted to investigate the performance of the proposed method in estimating the fixed effects and standard deviations of the random effects. It appeared that the estimates of the fixed effects are not affected by the choice of the covariance structure of the normally distributed random effect. For our application, the conclusion based on the analysis in Chapter 2 with regard to the fixed effect is confirmed, i.e. subjects who were infected at baseline and remained infected at post-treatment showed also an alteration in their Bacteroidetes to Firmicutes ratio in our extended model. To assess model fit, we computed the marginal correlations between and within categories over time. It appears that the marginal correlation in the data is well captured using the model with a multivariate random effect having a common variance for all categories.

In the next two chapters, the interplay between helminth infection, the gut microbiome composition, and immune responses which were characterized by the whole blood cytokine responses to antigens is studied. It is known that the removal of helminth infection by anthelminthic treatment restores immune responsiveness. It is also hypothesized that certain gut bacteria influences immune responses. Our aim in Chapter 4 is to gain insights into the mechanism underlying this interplay using the observations at both time points. A linear mixed

model is applied to the data with a cytokine response to specific antigen as an outcome variable. For this model, the predictors are bacterial proportion or diversity and their interaction term with helminth infection status. We restricted our analysis to three bacterial categories, namely Actinobacteria, Bacteroidetes and Firmicutes as these bacterial phyla were associated with helminth infection in the analysis of Chapter 2 and 3.

In this study, we observed that a gain in the proportion of Bacteroidetes is significantly associated with a decrease in concentration of IL-10 to LPS in helminth-uninfected subjects. This association is dampened in helminth-infected subjects. This finding confirms the hypothesized relationship that the removal of helminth infections restores immune responsiveness and that gut bacteria influences immune responses. Several limitations of this analysis can be noted: each bacterial proportion is assumed to be independent and its association with cytokine responses is analyzed separately. Thus, the compositional feature of microbiome data is ignored. In addition, the measurement error of the microbiome data is left unaccounted for in this model which potentially leads to biased estimators of the regression parameters. Furthermore, the association between helminth infection and bacterial proportion is not quantified in this model. Therefore, a statistical method developed in Chapter 5 which attempts to address these limitations.

In Chapter 5, our aim is to build a statistical model for the association between helminth infections and both microbiome and cytokine responses simultaneously by considering all sources of variability in the data. First of all, cytokine responses as continuous outcomes and gut microbiome as multivariate count outcomes observed from the same individuals are correlated. Secondly, the correlation between the same type of observations at different time points is expected. Finally, specific to microbiome data, there is an additional variability due to overdispersion and measurement error. The cytokine response and the microbiome composition are assumed to follow a normal and a multinomial distribution, respectively. A set of latent variables which is assumed to follow a multivariate normal distribution is incorporated to account for the additional variabilities in the data. The measurement error is modeled with multidimensional normally distributed random effect, i.e., each category has a different random effect which is assumed to be correlated. This is done to allow for more flexibility in modeling the extra variation in each category. The joint probability distribution is formulated and parameters are estimated by maximizing the joint likelihood with numerical quadratures. A simulation study is carried out to investigate the performance of the estimator for the fixed effects as well as random effect parameters in comparison with the method introduced in Chapter 4 (the naive method). The joint model outperforms the naive method. In the data application, it is shown that the correlation between microbiome composition and cytokine responses are small. When analyzing the marginal correlation using the proposed method, it appears that the marginal correlation does not fit to the observed one.

Chapter 6 summarizes and describes the results of the analyses performed and the statistical methods used in Chapter 2 to 5. The aim of this chapter is to evaluate the evidence for causality of the identified associations among helminth infection, the gut microbiome composition, and cytokine responses using data from the randomized controlled trial. We found that treatment has an effect on both the gut microbiome composition and cytokine responses via removing helminth infection and that the gut microbiome has a direct effect on cytokine responses. The directed acyclic graphs (DAGs) are used to visualize the direction of the causal effect and several potential sources of biases are included in this DAGs, namely unobserved confounders and measurement errors. The statistical methods developed in this thesis account for the additional variation due to unobserved confounders and measurement errors via the inclusion of the random effect. Specific to microbiome data, there are two possibilities of distributional assumption for a random effect, namely using the conjugate and normally distributed. In this thesis, we have explored these assumptions. It appears that models with random effects having a conjugate distribution fit the microbiome data well when considering how well its marginal correlation capture the observed correlation. Using the findings from the literature as well as from our analyses, we conclude that treatment has a causal effect on helminth infection and that helminth infection has direct effects on both the gut microbiome and the cytokine responses. It appears that the correlation between the gut microbiome and cytokine responses is small, hence the evaluation of their effect is not carried out.

Finally, data from randomized controlled trials, as is the case in this thesis are beneficial to examine causal relationships between variables involved. Furthermore, observations which are repeatedly measured provide information on how a specific outcome evolves over time. Unfortunately, the studies in this thesis use data from small subsamples from the larger trial which possibly decreases the statistical power to detect effects. One solution to address this limitation is by integrating different sources of observation, as we did in Chapter 5.

Samenvatting

Snelle verstedelijking gaat bijna altijd gepaard met een overgang van infectieziekten naar niet-overdraagbare ontstekingsziekten als dominante oorzaak van morbiditeit. Dit is zeer waarschijnlijk toe te wijzen aan veranderingen in leefstijl en omgevingsfactoren. Om het precieze mechanisme van deze transitie beter te begrijpen is er een populatiestudie uitgevoerd op Flores, een eiland in Indonesië. Dit gebied was geschikt voor deze studie omdat een groot aantal mensen is besmet met parasitaire wormen en er snelle veranderingen in leefstijl plaatsvinden. Hoewel worminfecties een negatieve invloed kunnen hebben op de menselijke gezondheid, wordt de aanwezigheid van wormen ook geassocieerd met een sterke regulatie van de afweerreactie en dit verklaart de lage prevalentie van ontstekingsziekten in gebieden waar veel mensen besmet zijn met wormen. In een gerandomiseerde, dubbelblind-uitgevoerde studie is er klinische en biomedische data verzameld om het complexe biologische mechanisme te onderzoeken dat aan deze associatie ten grondslag ligt. Tijdens deze studie werd de ene helft van de onderzoeksgroep behandeld tegen worminfecties, de andere helft kreeg een placebo. In dit proefschrift wordt de relatie tussen worminfecties, de samenstelling van het darmmicrobioom en de afweerreactie beschreven. Omdat de beschikbare analysemethoden niet geschikt bleken voor de complexiteit van de verzamelde data in deze populatiestudie, is een ander aandachtspunt in dit proefschrift de ontwikkeling van statistische methodologie.

Hoofdstuk 1 beschrijft een algemene introductie van het proefschrift met betrekking tot de verzamelde data, de onderzoeksvragen en de statistische methoden die beschikbaar zijn om deze data te analyseren en de vragen te beantwoorden. Er wordt ook kort beschreven hoe de microbioomprofielen zijn verkregen. De methoden onderliggend aan deze profielen leggen een compositioneel structuur op de microbioomdata op, waar rekening mee moet worden gehouden in het modelleren. Daarnaast zijn er meerdere datapunten van dezelfde persoon en dit leidt tot correlatie tussen de metingen. Tot slot wordt er een overzicht gegeven van de statistische hulpmiddelen die worden gebruikt om microbioomdata te analyseren, gevolgd door een discussie van de uitdagingen die komen kijken bij het correct modelleren van dit type data in een onderzoeksopzet met herhaalde metingen.

Hoofdstuk 2 beschrijft de toepassing van een recent ontwikkelde statistische methode om de microbioomdata, verzameld in de populatiestudie, te modelleren. Een aanname van dit model is dat de microbioomdata voortkomen uit een multinomiale verdeling. Om rekening te houden met de aanwezigheid van extra variatie, wordt de aanname gedaan dat de parameters van de multinomiale verdeling random effecten zijn die volgen uit een geconjugeerde verdeling. Echter, dit model is alleen geldig voor onafhankelijke waarnemingen van een multinomiale verdeling. Hierdoor kan er geen rekening worden gehouden met de correlatiestructuur in onze data door de herhaalde metingen. Om die reden zijn de analyses voor elk tijdspunt apart uitgevoerd. Terwijl de data verkregen voorafgaand aan de behandeling is gebruikt om het effect van worminfecties op de samenstelling van het darmmicrobioom te analyseren, is de na afloop van de behandeling verkregen data gebruikt om het effect van de anti-wormen behandeling op de samenstelling van het darmmicrobioom te bestuderen. Om te onderzoeken of de behandeling een verschillend effect heeft op personen die voor aanvang van de behandeling met wormen besmet waren, ten opzichte van personen die niet besmet waren, is er een interactieterm tussen de status van infectie en de behandeling toegevoegd aan het model. Het bleek dat alleen in de groep die behandeld was met anti-wormenmedicatie en die zowel voor aanvang van de behandeling, als na afloop van de behandeling besmet was met wormen, er een significant verschil was in de verhouding van Bacteroidetes ten opzichte van Firmicutes en de verhouding van Actinobacteria ten opzichte van Firmicutes in vergelijking met de andere groepen. De methode die hier is toegepast heeft de beperking dat er maar één tijdspunt kan worden geanalyseerd, en daarom kan de verandering van de samenstelling van het microbioom niet over het verloop van een bepaalde periode worden bestudeerd. Hoofdstuk 3 beschrijft de ontwikkeling van een model dat zich hierop richt.

In hoofdstuk 3 wordt de ontwikkeling van een statistische methode beschreven die herhaalde metingen van de samenstelling van het microbioom modelleert, waarbij rekening wordt gehouden met de correlatiestructuur tussen de herhaalde metingen van een persoon. Dit is gedaan door een normaal verdeeld random effect te introduceren. Er worden drie verschillende covariantiestructuren voor de normaal verdeelde random effecten overwogen. Ten eerste stellen we een univariate en persoons-specifieke random effect vast, waar het random effect voor elke categorie van bacteriën per verschillend tijdspunt hetzelfde is. Ten tweede, elke categorie heeft een verschillend random effect met een categoriespecifieke variantie. Als laatste wordt aangenomen dat er een multivariate random effect is met een gemeenschappelijke variantie voor alle categorieën. Aan de hand van een simulatiestudie is onderzocht in hoeverre deze voorgestelde methode in staat is om een schatting te maken van de fixed effects en de standaard deviaties van de random effects. Hieruit bleek dat de schattingen van de fixed effects niet worden beïnvloed door de keuze van de covariantiestructuur van de nor-

maal verdeelde random effecten. Met betrekking tot de dataset, de conclusie uit hoofdstuk 2 is bevestigd door het uitgebreidere model. Dat wil zeggen dat de groep die zowel voor aanvang van de behandeling, als na afloop van de behandeling besmet was met wormen, een verandering vertoonden in de verhouding Bacteroidetes ten opzichte van Firmicutes. Om de betrouwbaarheid van het model te beoordelen, hebben we de marginale correlaties uitgerekend tussen en binnen de categorieën over tijd. Hieruit blijkt dat de marginale correlaties goed worden geschat met behulp van het multivariate random effect, gebaseerd op de gemeenschappelijke variantie voor alle categorieën.

In de volgende hoofdstukken wordt de wisselwerking tussen worminfecties, de samenstelling van het darmmicrobioom en de afweerreactie, gekenmerkt door cytokinerespons in het bloed op antigenen, beschreven. Eerder onderzoek heeft aangetoond dat anti-wormenbehandeling leidt tot herstel van de verzwakte afweerreactie, geassocieerd met worminfecties. Er wordt tevens gedacht dat bepaalde darmbacteriën de afweerreactie beïnvloeden. Het doel van hoofdstuk 4 is om inzicht te verkrijgen in de mechanismen die een rol spelen bij deze wisselwerking waarbij de metingen van beide tijdspunten worden gebruikt. Een linear mixed model is toegepast op de data, met de cytokineproductie in reactie op bepaalde antigenen als uitkomstvariabele. Voor dit model zijn het bacteriële aandeel of de diversiteit en hun interactieterm met parasitaire wormen gebruikt als verklarende variabelen. We hebben onze analyse beperkt tot drie categorieën van bacteriën, namelijk Actinobacteria, Bacteriodetes en Firmicutes, omdat deze bacteriële groepen geassocieerd waren met worminfecties in de analyses in Hoofdstuk 2 en 3.

In deze studie hebben we aangetoond dat er een significante associatie is tussen de toename in het aandeel van Bacteriodetes en de afname in de concentratie van IL-10 in reactie op LPS in personen die niet besmet zijn met wormen. De associatie is minder sterk in personen die besmet zijn met wormen. Hoewel deze resultaten de hypothese bevestigen dat het verwijderen van worminfecties leidt tot een sterkere afweerreactie en dat darmbacteriën de afweerreactie beïnvloeden, zijn er ook een aantal beperkingen in de analyse te benoemen. Er wordt aangenomen dat elke bacterieel aandeel onafhankelijk is, en de associatie met de productie van cytokines wordt apart geanalyseerd. Hieruit volgt dat het kenmerk van de samenstelling van het microbioom wordt genegeerd. In aanvulling hierop wordt er ook geen rekening gehouden met de meetfout binnen de microbioomdata bij dit model. Wat weer kan leiden tot onder- of overschatten van de regressieparameters. Bovendien wordt de associatie tussen worminfecties en bacteriële aandelen niet gekwantificeerd in dit model. Om die reden is er in Hoofdstuk 5 een statistische methode ontwikkeld die probeert deze beperkingen te overwinnen.

In Hoofdstuk 5 is ons doel om een statistisch model te bouwen voor de associatie tussen helminthinfecties enerzijds en zowel microbioom als cytokinerespons simultaan anderzijds, door alle bronnen van variatie in de data mee te nemen. Allereerst zijn de cytokinereponsen als continue uitkomsten en de multivariate teluitkomst van het darmmicrobioom gecorreleerd binnen een individu. Ten tweede worden er correlaties tussen dezelfde type observaties op verschillende tijdstippen verwacht. Ten slotte is er, specifiek voor microbioomdata, extra variabiliteit door overdispersie en meetfouten. We nemen aan dat de cytokineresponsen en de microbioomcomposities respectievelijk een normale en een multinomiale verdeling volgen. Een verzameling latente variabelen, waarvan aangenomen wordt dat ze een multivariate normale verdeling volgen, is in het model opgenomenom rekening te houden met additionele variatie in de data. De meetfout is gemodelleerd met een multidimensionaal normaal verdeeld random effect, m.a.w. elke categorie heeft zijn eigen random effect waarvan aangenomen wordt dat ze zijn gecorreleerd. Dit wordt gedaan voor meer flexibiliteit in het modelleren van de extra variatie in elke categorie. De gezamenlijke kansverdeling is geformuleerd en parameters zijn geschat door het maximaliseren van de gezamenlijke waarschijnlijkheidsfunctie met numeriek bepaalde kwadraturen. Een simulatiestudie is uitgevoerd om de prestatie van de schatters te onderzoeken voor zowel de fixed als de random effecten, in vergelijking met de methode geïntroduceerd in Hoofdstuk 4 (de naïeve methode). Het gezamenlijke model presteert beter dan de naïeve model. In de datatoepassing laten we zien dat de correlatie tussen de microbioomcompositie en de cytokineresponsen klein is. Wanneer we de marginale correlaties behorende bij het voorgestelde model uitrekenen, blijken deze niet in overeenstemming te zijn met de geobserveerde correlaties.

Hoofdstuk 6 beschrijft en vat de resultaten van de analyses die zijn uitgevoerd en de methoden gebruikt in Hoofdstuk 2 tot en met 5 samen. Het doel van dit hoofdstuk is om de bewijzen voor causaliteit van de geïdentificeerde associaties te evalueren tussen helminthinfecties, darmmicrobioomcompositie en cytokineresponsen, gebruikmakende van de data uit het gerandomiseerde gecontroleerde onderzoek. We hebben gevonden dat de behandeling een effect heeft op zowel de darmmicrobioomcompositie als cytokineresponsen via het verwijderen van de helminthinfectie, en dat de darmmicrobioom een directe effect heeft op cytokineresponsen. De gerichte acyclische grafen (DAGs) zijn gebruikt om de richting van het causale effect te visualiseren, en meerdere potentiële bronnen van vertekening zijn meegenomen in deze DAGs, namelijk niet-geobserveerde confounders en meetfouten. De statistische methoden ontwikkeld in dit proefschrift nemen de extra variatie veroorzaakt door niet-geobserveerde confounders en meetfouten mee via de inclusie van random effecten. Specifiek voor microbioomdata zijn er twee voor de hand liggende mogelijkheden voor verdelingsaannames op de randomeffecten, namelijk een geconjugeerde en een normale verdeling. In dit proefschrift hebben wij deze aannames verkend. Het blijkt dat modellen met

random effecten die een geconjugeerde verdeling hebben goed bij de microbioomdata passen, als je beschouwt hoe goed de marginale correlaties van het model de geobserveerde correlaties reflecteren. Gebaseerd op resultaten uit de literatuur en onze analyses, kunnen we concluderen dat behandeling een causaal effect heeft op helminthinfectie en dat een helminthinfectie directe effecten heeft op zowel de darmmicrobioom als de cytokineresponsen. Het blijkt dat de correlatie tussen darmmicrobioom en cytokineresponsen klein is, daarom is hun effect niet verder geëvalueerd.

Ten slotte, data uit gerandomiseerde gecontroleerde onderzoeken, zoals in dit proefschrift, zijn nuttig om causale verbanden tussen betrokken variabelen te onderzoeken. Verder leveren observaties die meermaals zijn gemeten informatie over hoe een specifieke uitkomst zich ontwikkelt over tijd. Helaas gebruiken de studies in dit proefschrift data over kleine deelsteekproeven uit een groter onderzoek, wat mogelijk leidt tot een verlaagd statistisch onderscheidend vermogen om effecten te detecteren. Een oplossing om deze limitatie te beperken is om verschillende bronnen met observaties te integreren.

List of Publications

Martin I, Tissier RLM and Houwing-Duistermaat JJ. The joint mixture model for the effect of multivariate count on the continuous outcome subject to measurement error. Submitted for publication.

Martin I, Kaisar MMM, Wiria AE, Hamid F, Djuardi Y, Sartono E, Rosa BA, Mitreva M, Supali T, Houwing-Duistermaat JJ, Yazdanbakhsh M, Wammes LJ. The effect of gut microbiome composition on human immune responses - interference of helminth infections. Submitted for publication.

Martin I, Uh HW, Supali T, Mitreva M, Houwing-Duistermaat JJ (2019). The mixed model for the analysis of a repeated-measurement multivariate count data. *Statistics in Medicine*, doi: 10.1002/sim.8101

Martin I*, Djuardi Y*, Sartono E, Rosa BA, Mitreva M, Houwing-Duistermaat JJ, Yazdanbakhsh M (2018). Dynamic changes in human-gut microbiome in relation to a placebo-controlled anthelminthic trial in Indonesia. *PLoS Neglected Tropical Diseases*, 12(8):e0006620.

Tahapary DL, de Ruiter K, **Martin I**, Brienen EAT, van Lieshout L, Djuardi Y, Djimandjaja CC, Houwing-Duistermaat JJ, Soewondo P, Sartono E, Supali T, Smit JW, Yazdanbakhsh M (2017). Effect of anthelmintic treatment on leptin, adiponectin, and leptin to adiponectin ratio: a randomized controlled trial. *Nutrition & Diabetes.* 7(10):e289.

Tahapary DL*, de Ruiter K, **Martin I**, Brienen EAT, van Lieshout L, Cobbaert CM, Soewondo P, Djuardi Y, Wiria AE, Houwing-Duistermaat JJ, Sartono E, Smit JW, Yazdanbakhsh M, Supali T (2017). Effect of Anthelmintic Treatment on Insulin Resistance: A Cluster-Randomized Placebo-Controlled Trial in Indonesia. *Clinical Infectious Diseases*, 65(5):764-771.

Tahapary DL*, de Ruiter K*, **Martin I**, van Lieshout L, Guigas B, Soewondo P, Djuardi Y, Wiria AE, Mayboroda OA, Houwing-Duistermaat JJ, Tasman H, Sartono E, Yazdanbakhsh M, Smit JW, Supali T (2015). Helminth infections and type

2 diabetes: a cluster-randomized placebo controlled SUGARSPIN trial in Nangapanda, Flores, Indonesia. *BMC Infectious Diseases*, 15:133.

* These authors contributed equally to this work.

Curriculum Vitæ

Ivonne Martin was born on October, 15th 1982 in Jakarta, Indonesia. She studied Mathematics for her undergraduate in Parahyangan Catholic University where she also serves as teaching assistant during her last two years of study. After graduating in 2005, she was appointed as a lecturer in her alma mater for three years before she was awarded a scholarship from Directorate General of Higher Education (DIKTI) to pursue her Master degree in applied mathematics in Delft University of Technology. She graduated in 2010 with her thesis as a collaboration study between Academic Medical Centre and Department of Statistics in Delft University of Technology to investigate the performance of a newly developed hazard estimator for patients with myocardial infarction.

In 2013, she attended a workshop in Jakarta in which she was selected as a PhD candidate for a collaborative research project between Medical Faculty of Universitas Indonesia (FKUI) and Leiden University Medical Centre (LUMC), funded by The Royal Netherlands Academy of Arts and Sciences (KNAW), specifically the Scientific Programme Indonesia-Netherlands (SPIN). Her PhD fellowship was awarded by the Directorate of Higher Education, Republic of Indonesia (2013-2016) and Leiden University (2017). During her training, she presented her work in the International Biometric Society Channel Meeting in 2015 as well as at the International Biometric Conferences in Canada and Barcelona. In 2016, she received a travel grant from Leiden University Fund to present her work in XXVIII International Biometric Conference in Canada.

After completing her Doctoral study, she will expand her research in the area of mixed and joint modeling.

Acknowledgement

I would like to express my enormous gratitude to my promotores: Prof. Jeanine Houwing-Duistermaat and Prof. Maria Yazdanbakhsh for the opportunity to be involved in this collaboration project. Both of their immense knowledge, constructive criticism, and perpetual encouragement have driven me to excel in research. I would like to thank people in the SugarSPIN project for their guidance during this research work: Prof. Taniawati Supali, Prof. Johannes W.A. Smit, Dr. Erliyani Sartono as well as Dicky Tahapary and Karin de Ruiter who shared the PhD journey under this collaboration project.

I would like to thank all colleagues from Department of Biomedical Data Sciences with whom I learned about statistics and research. In particular, people from the group of Statistical Genetics: Dr. Hae-Won Uh, Dr. Stefan Bohringer, Dr. Mar Rodriguez-Girondo, Dr. Roula Tsonaka, Dr. Bart Mertens, Dr. Kate Xu, Brunilda Balliu, Renaud Tissier, Alexia Kakourou, Angga Fuady, Said el Bouhaddani, Giorgios Bartzis for their generous input towards my work and from whom I learn to become a team member. Also, I thank other members of the department who have shared their knowledge through formal and informal meetings: Ron Wolterbeek, Mitra Ebrahimpoor, Ningning Xu, and Mirko Signorelli. It is also through the Department of Parasitology, I have learned to improve my understanding of applying statistics into biological field. Many thanks to Yenny Djuardi, Linda Wammes, Dian Amaruddin, Mikhael Manurung, and Koen Stam.

I am grateful to several good friends and acquaintances who remembered me in their prayers for the ultimate success. In the end, life as PhD will be much more difficult without all of these people: Maria Kaisar, Abena Amoah, Chare Virakkun, Hari Nugroho and Lionov. I consider myself nothing without them. They gave me enough moral support, encouragement and motivation to accomplish the personal goals. I thank Ibu Hedi Hinzler not only for her hospitality but also her inspiring words that she has shared during my last phase of PhD.

A word of appreciation is not enough for my parents and brother for their unconditional and unwavering support throughout these years. This thesis is dedicated to them.