Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

High dimensional surrogacy in microbiome experiments: hierarchical Bayesian Approach

Edwin Kipruto

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

2017
2018

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Masterthesis

High dimensional surrogacy in microbiome experiments: hierarchical Bayesian Approach

Edwin Kipruto

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Contents

Acknowledgements

I am deeply grateful to my supervisor Prof. dr. Ziv Shkedy for his patience, dedication, encouragement and valuable support throughout my thesis period. Ziv, the meetings we conducted and the advice you provided on statistical methods were invaluable. It has been a great pleasure working with you. I would also like to thank my supervisor Owokotomo Olajumoke Evangelina for working tirelessly to make this thesis successful. Evangelina, you invested a substantial amount of your time teaching me how to use JAGS and reading my work. Your careful reading and insightful comments helped me a lot in writing this thesis.

My heartfelt gratitude also goes to all my lecturers at Hasselt University for their mentorship and VLIR for providing me with the scholarship. Also, I wish to extend my gratitude to my beloved mother, who is my best friend. She has been always there praying and supporting me. Finally, I thank all my friends who provided advice and support.

Abstract

The humoral immune system produces antibodies to protect the body against infections. Invasion by foreign substance activates the immune system, and the degree of activation can be measured using antibody levels. However, detectable levels of antibodies may take a substantial amount of time, especially during cell proliferation and maturation. Thus, a surrogate endpoint which can predict the antibody levels is preferable for the evaluation of experimental treatments.

In this study, the gastrointestinal tract bacterial family richness which can be measured more frequently and less costly due to advancement in genomic technology was evaluated as a potential surrogate endpoint for predicting immunoglobulin A. A total of 15 germ-free mice were randomized to receive either antibiotic (PAT) altered cecal content or unaltered cecal content (control) and followed up for 20 days. The measurements were recorded on day 1, 6, 12 and 20. From the exploratory data analysis using individual profile plots, it was established that the S24-7 family richness closely reflected the immunoglobulin A levels which prompted its evaluation. To answer the research questions simultaneously, the Bayesian hierarchical joint models were fitted to the data.

The results showed that the treatment effects on the family richness on day 6 and 12 were different from zero. In addition, a positive linear association between the family richness and immunoglobulin A levels were predominant towards the end of the study. Therefore, S24-7 family richness can be used as a surrogate endpoint in predicting the immunoglobulin A levels at day 12 and 20 with moderate precision. However, the sample size per treatment arm needs to be increased to confirm the stability of the results.

Keywords: *Immunoglobulin A, S24-7 family, Bayesian joint model, Gut Microbiome, longitudinal data*

1 Introduction

1.1 Human Microbiota

The human body is colonized by vast microbes that live on and inside the body, such as bacteria, fungi and viruses which are collectively called human microbiota (Hamady et al., 2009). Bacterial communities in a healthy human adult are estimated to outnumber human somatic and germ cells by a ratio of ten to one and their genomes are collectively referred to as human microbiome (Pughoeft et al., 2012; Turnbaugh et al., 2007). In the absence of intrauterine infections, human infants can acquire their initial bacteria while traveling through the maternal birth canal due to vaginal microflora and after birth through breastfeeding (Funkhouser et al., 2013; Mueller et al., 2015). However, infants delivered by cesarean section lack vaginal microbes instead their first microbes are of environmental origin and resembles microbes of the skin (Langdon et al., 2016). During child growth these microbes develop into a highly diverse ecosystem and over time, human and bacterial associations develop into beneficial relationships (Wang et al., 2017).

The commensal microbiota have co-evolved with the human for long and they have colonized different parts of the body, including the gastrointestinal tract, skin, saliva, oral mucosa, and conjunctiva with the majority found in the colon and the skin. Their persistent interaction has led to various forms of relationship including mutualistic, parasitic or commensal (Shekhar et al., 2017; Sender et al., 2016; Turnbaugh et al., 2007; Belkaid and Hand, 2014). Several studies have shown that some bacterial communities are useful in human health. For instance, the microbial communities in the skin help in protecting the body against invasion by harmful organisms and educates T cells to have immunologic memory (Grice et al., 2011). In addition, animal studies have revealed that the gastrointestinal tract microbiota plays a vital role in drug metabolism, Vitamin K production, gut development and mucosal immune system maturation (Riedel et al., 2014; Matsuki and Tanaka, 2014).

Besides the crucial role of microbiota in maintaining human health, numerous studies have demonstrated that disturbance of the gut microbiota facilitates the emergence of certain diseases. For example, mucosal biopsies of patients suffering from inflammatory bowel disease have reduced bacterial diversity with loss of commensal species such as Clostridium leptum, Eubacterium and bifidobacteria (Marchesi, 2014). Studies conducted on diabetes-prone rats and diabetes-resistant rats revealed that, type 1 diabetes progression, which is caused by insulin deficiency is associated with a higher abundance of Lactobacillus and Bifidobacterium

(Roesch et al., 2009). Further, infectious diseases such as Clostridium difficile infections occur due to an overgrowth of clostridium difficile bacteria in the gut (Wang et al., 2017).

Microbial communities vary in composition among individual. Most of these variabilities are unexplained, however, it has been linked to environmental interactions as well as variability in diet, human genotype, hygiene, delivery mode, antibiotic use, and colonization history (Huttenhower et al., 2012; Ruiz et al., 2017). For instance, rural African children who consume fiber-rich diets have higher abundance of specific Bacteroidetes, a reduced amount of Firmicutes and decreased amounts of Proteobacteria as compared to children in Europe (Nguyen et al., 2015). Such differences in microbial communities may contribute to different patterns of human diseases (Pughoeft et al., 2012).

1.1.1 Immune System and Gastrointestinal Tract (GUT)

The immune system is a network of cells, tissues, and organs that work closely together to defend the body against attacks by foreign invaders (Kelly, 2007). It is categorized into innate immunity, which is present at birth and does not distinguish between threats and adaptive immunity, which protects the body against a specific threat and mostly develops after birth. Moreover, adaptive immunity is coordinated by cellular immunity where T cells provide defenses against abnormal cells and pathogens inside the cells, while antibodies provide defenses against antigens and pathogens in body fluids with the help of B cells (Martini and Bartholomew, 2013; Marieb, 2008).

Here, our interest lies in antibodies, which are also known as immunoglobulins. Several classes of antibodies are known. This includes immunoglobulin A (IgA) which is the most predominant antibody class in the external secretion of humans with a role of protecting mucosal surfaces against infections. Mucosal surfaces are the main area of exposure to the external environment and are the point of high vulnerability to attack by pathogens (Woof, 2013). IgA protects the intestinal epithelium from the entry of pathogens by blocking their access to epithelial receptors, entrapping them in mucus and facilitating their removal (Mantis et al., 2011).

The gut is the primary site of interaction between the immune system and microorganisms, both symbiotic and pathogenic (Iebba et al., 2012). Studies have shown that the gut microbiota stimulates the production of IgA as well as maintaining the homeostasis of T-cell populations (Marchesi,2014).

1.2 The Surrogate Endpoints for Immune Response

A surrogate endpoint is a biomarker that is intended to substitute a clinically meaningful endpoint. It is also expected to predict clinical benefit or harm (Molenberghs et al., 2004). Many surrogate endpoints are usually proposed since they closely reflect the biological state of the disease (Piantadosi, 2017). Essentially, investigators often choose a surrogate endpoint when the measurements of clinical endpoints are costly, requires long follow-up time to observe the event of interest or the proposed clinical benefit requires large sample size to detect due to the low incidence of disease (Alonso and Molenberghs, 2008; Piantadosi, 2017). A surrogate endpoint may also allow early detection of safety signals in new drugs and limit potential problems with noncompliance and missing data which is associated with long studies (Buyse et al., 2016).

The humoral immune response coordinated by B cells are often faced with time delays. For instance, B cells can take a long duration of time to bind to the antigens of an invading pathogen and initiate full destruction. Similarly, a substantial delay can occur during cell proliferation of the innate immune response. Therefore, detectable levels of antibodies may take between three and four days after infection or six to seven weeks during cell proliferation and maturation (Fenton et al., 2006) which allows pathogens to cause a host cell damage and dysfunction hence the need of a surrogate endpoint. Interestingly, advances in genomic technology has rapidly increased the number of biomarkers that can potentially act as surrogate endpoints by decreasing the cost and increasing the speed of DNA sequencing thus prompting analysis of complex datasets from bacterial communities (Turnbaugh et al., 2007; Buyse et al., 2016). Here, a biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Buyse, 2007).

Contrary to the benefits of surrogate endpoints, some serious historical limitations have been reported. For example, failures of ventricular arrhythmias as a surrogate endpoint of survival in cardiovascular diseases and bone mineral density as a surrogate endpoint of bone fractures in Osteoporosis (Fleming and DeMets, 1996). The main reason behind these historical failures is the incorrect assumption that surrogacy is as a result of the association between a potential surrogate endpoint and the corresponding clinical endpoint, which is not enough for surrogacy (Buyse et al., 2016). Over the years, several criteria have been proposed to validate surrogate endpoints. The criteria postulate that the effect of the intervention on the surrogate endpoint

should predict the effect on the true endpoint (Fleming and DeMets, 1996; Molenberghs et al., 2005).

1.3 Aims and Objectives

It is well established that the humoral immune response gradually build-up over time as the number of bacteria that cause infectious diseases in humans increases. The aim of this study, therefore, is to investigate the association between bacterial family richness and the immune response in the presence of intervention. Again, our interest is to study the effects of the intervention on the bacterial family richness as well as immunoglobulin A levels. These three questions will allow us to identify a bacterial family that can be used to predict the immune response. To vividly answer the research questions simultaneously, a joint model for family richness (Poisson) and a normal endpoint was developed.

2 Data Description

The dataset analyzed in this study was obtained from mice model experiments where the effects of a single pulse of macrolide antibiotic (tylosin) administered early in life were assessed. The main interest was to determine whether the antibiotic treatment leads to modification of the intestinal microbiota, which successively can cause a permanent problem in immunological response. The experiment involved two groups of mice; the donor group and germ-free group. The donor group was randomly assigned to either one pulse of antibiotic treatment or plain drinking water from the 5^{th} to 10^{th} day after birth. On the 12^{th} day after birth, they were sacrificed, and the cecal contents transferred to 15 male and female germ-free mice. Out of these, seven received PAT altered cecal content and eight received unaltered cecal contents (control). The fecal secretory immunoglobulin A (IgA) was then measured at day 1, 6, 12 and 20 after the oral administration of donors cecal contents.

Besides the IgA levels, the family richness was also recorded. Here, the family richness was defined as the number of operational taxonomic units (OTUs) belonging to a particular bacterial family with nonzero abundance in the sample. The reason for opting for family richness is due to the high proportion of zero counts in the species level and also has the benefit of providing information about the effects of the intervention on the family level. In this analysis, the clinical endpoint of interest is Immunoglobulin A (IgA) levels while bacterial family richness is considered as the potential biomarker. Therefore, the data structure at each time point consists of 15×1 vector of IgA levels ($\boldsymbol{Y}$), 15×30 family richness matrix (X) and 15×1 vector of treatment ($\boldsymbol{Z}$) as shown below;

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{15} \end{bmatrix}, \quad \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,30} \\ X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,30} \\ X_{3,1} & X_{3,2} & X_{3,3} & \cdots & X_{3,30} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{15,1} & X_{15,2} & X_{15,3} & \cdots & X_{15,30} \end{bmatrix}, \quad \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ \vdots \\ Z_{15} \end{bmatrix}$$

The pictorial representation of the subset of the data is presented. Figure 1 shows the scatter plots of the logarithm transformed IgA levels and the two selected family richness namely S24-7 and Lachnospiraceae. The IgA levels were log transformed to achieve normality whereas the two families are selected because of their individual profile plots which resembled the longitudinal

profiles of the IgA levels (Figure 9 in appendix). From the plots, the two treatment groups are gradually disintegrating over time with discernible separation observed at time 12 and 20. Similarly, on the aforementioned time points, the IgA levels of the control group are quite higher compared to the experimental treatment. This illuminates the effects of the treatment on the family richness and IgA levels.

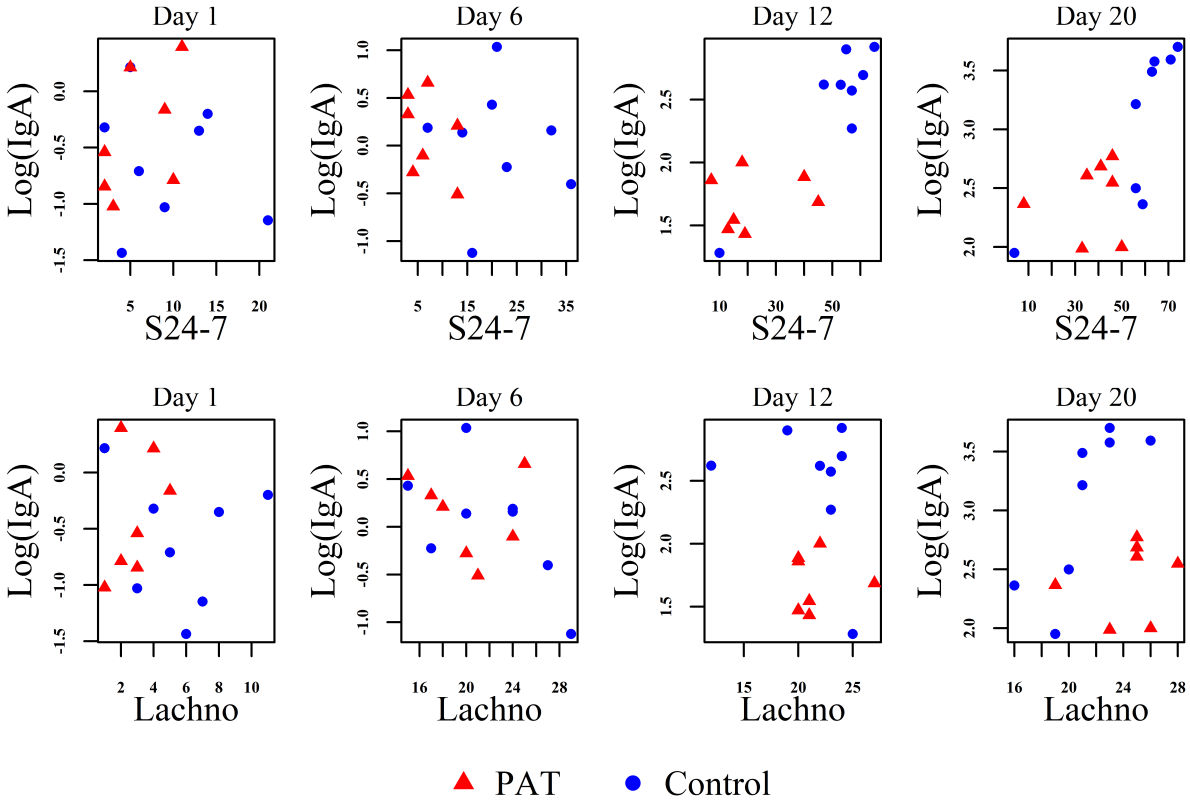

Figure 1: *Scatterplots of the observed family richness and transformed IgA levels. The upper panel shows the association between S24-7 family richness and transformed IgA levels while the lower panel shows the association between Lachnospiraceae family richness and transformed IgA levels.*

3 Methods

Over the years, statistical methods for evaluating surrogate endpoints in single and multiple trials have been proposed. In this study, we evaluate the potential surrogate marker, using the Bayesian joint modeling of Poisson and normal endpoints. The principal cause of using Bayesian method is as follows; first, the outcome variables are of different types rather than normal endpoints rendering the likelihood-based mixed models unfavorable because of computational challenge which makes it complex to provide answers to practical problems. This is not the case in the Bayesian approach since the full conditional posterior distribution is sampled using a flexible Markov Chain Monte Carlo (MCMC) algorithm (Shkedy and Barbosa, 2005; Molenberghs et al., 2010).

Secondly, unlike the frequentist paradigm where the asymptotic theory for statistical tests is more prevalent, the Bayesian approach is based on the exact inference which can be obtained by exploring the posterior distribution. Lastly, the uncertainty of all parameters in the model is taken into account by considering them as random variables rather than fixed quantities (Lesaffre and Lawson, 2012). Thus, making the Bayesian approach more appealing, especially in this study because of the small sample size. However, the weakness of Bayesian approach is that it can introduce subjectivity into the analysis through prior distribution, though, this is addressed by utilizing noninformative priors. Note that, the frequentist counterpart joint model can be fitted using SAS procedure NLMIXED (see Molenberghs and Verbeke, 2005 for details).

This chapter is organized as follows; in section 3.1 we discuss the joint modeling approach adopted, section 3.2 discusses hierarchical Bayesian joint models and finally, we close the chapter by presenting model fitting and diagnostics (section 3.3).

3.1 Joint Models for IgA and Bacterial Family Richness

Joint modeling is a statistical technique used to estimate common parameters of two or more models jointly (Lesaffre and Lawson, 2012). It offers an advantage over univariate models in several ways: the effect of the covariate(s) such as treatment on the outcomes can be evaluated simultaneously; or the association structure of the outcome variables can be assessed and it allows modeling of outcomes of different types (Molenberghs and Verbeke, 2005; Ivanova et al., 2016).

The pictorial representation of the joint model with parameters of interest is shown in Figure 2. The parameter α_{jt} measures the direct treatment effects on the j^{th} family richness at time t,

γ_{jt} characterizes the effects of the j^{th} family richness on IgA levels at time t after adjusting for the treatment effects and β_t measures the treatment effect on transformed IgA levels after adjusting for the effects of family richness at time t. For simplicity, we will use the same parameter notations across all models implemented.

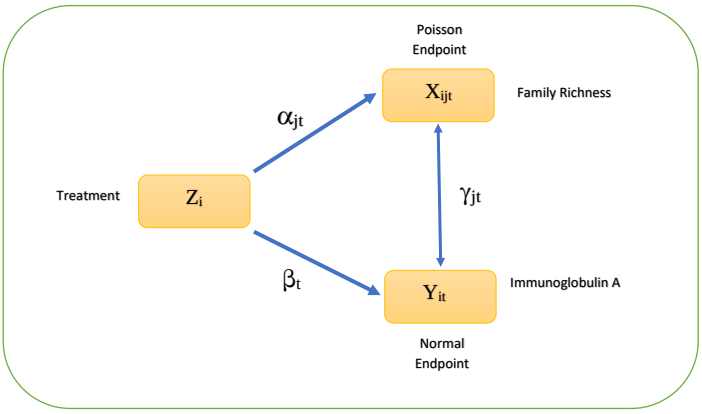

Figure 2: *Relationship between family richness (X), transformed IgA levels (Y) and treatment (Z)*

3.2 Bayesian Joint Models

This section is organized as follows. First, we start by gently introducing the Bayesian joint fixed effects models followed by random effects models for the outcome variables in conjunction with their linear predictors. Here, fixed effects models refer to models without random intercept (see section 3.2.1) and random effects models refer to models with random intercept as in section 3.2.2. We end the section by providing the motivation for choosing the priors.

3.2.1 Joint Fixed Effects Model

Let Y_{ij} and X_{ij} denote the logarithm transformed IgA levels and the family richness of the i^{th} mouse taken at time j respectively. Further, let Z_i be the treatment assigned to the i^{th} Mouse where $Z = 1$ if the Mouse received PAT altered cecal content and $Z = 0$ if it received unaltered cecal content. Since family richness is count data, the satisfactory distribution is Poisson whereas transformed IgA levels is assumed to follow a normal distribution. The two univariate distributions are joined together by taking Poisson model as a marginal model and linear model as a conditional model. In the latter, we conditioned on both the treatment and the family richness because our interest is to evaluate the relation between family richness and IgA levels while adjusting for treatment effects. The resulting Bayesian fixed effects joint model

(M1) is given by

$$\begin{cases} X_{ij}|Z_i \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = \alpha_{0j} + \alpha_{1j}Z_i \\ Y_{ij}|X_i, Z_i \sim \text{Normal}(\mu_{ij}, \sigma^2) & \text{where } \mu_{ij} = \beta_{0j} + \gamma_{1j}X_i + \beta_{1j}Z_i \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma^2 \sim \text{IG}(0.001, 0.001), & \textbf{Prior} \end{cases} \tag{1}$$

Where α_{0j} and α_{1j} are the time specific intercept and treatment effects on family richness respectively, β_{0j} and β_{1j} are the time specific intercept and treatment effects on transformed IgA levels after adjusting for the effects of X respectively, γ_{1j} is the family richness effects at time j after adjusting for the effects of Z ($j = 1, 6, 12, 20$, $i = 1, ..., 15$, $k = 0, 1$) with the assumption that the error component $\varepsilon_{ij} \sim \text{N}(0, \sigma^2)$.

The observed variances of the transformed IgA levels as shown in Table 1 suggested that the treatment groups seems to have different variances over time. In this regard, the variance of error component in model 1 is allowed to vary between treatment groups resulting to model 2 (M2) given by

$$\begin{cases} X_{ij}|Z_i \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = \alpha_{0j} + \alpha_{1j}Z_i \\ Y_{ij}|X_i, Z_i \sim \text{Normal}(\mu_{ij}, \sigma_m^2) & \text{where } \mu_{ij} = \beta_{0j} + \gamma_{1j}X_i + \beta_{1j}Z_i, \quad m = 0, 1 \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma_m^2 \sim \text{IG}(0.001, 0.001), & \textbf{Prior} \end{cases} \tag{2}$$

Where $\varepsilon_{ij} \sim N(0, \sigma_0^2)$ for control and $\varepsilon_{ij} \sim N(0, \sigma_1^2)$ for PAT group.

3.2.2 Bayesian Hierarchical Joint Models (Random effects Models)

To accommodate the longitudinal nature of the data, random effects were introduced into the models discussed in section 3.2.1. Thus, the model for the family richness is conditioned upon both the observed covariate (treatment) and unobserved random intercept where the latter accounts for the between Mice variability in the family richness. On the other hand, the model for transformed IgA level is conditioned upon the treatment, family richness and the random intercept. In this case, the random intercept captures between Mice variabilities in the IgA

levels. The Bayesian joint hierarchical model (M3) is given by

$$\begin{cases} X_{ij}|Z_i, b_{1i} \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = (\alpha_{0j} + b_{1i}) + \alpha_{1j} Z_i \\ Y_{ij}|X_i, Z_i, b_{2i} \sim \text{Normal}(\mu_{ij}, \sigma^2) & \text{where } \mu_{ij} = (\beta_{0j} + b_{2i}) + \gamma_{1j} X_i + \beta_{1j} Z_i \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma^2 \sim \text{IG}(0.001, 0.001), & \textbf{Prior} \end{cases} \tag{3}$$

Where b_{1i} and b_{2i} are the random intercepts for the i^{th} Mouse associated with the family richness and transformed IgA levels respectively. They are assumed to have a bivariate normal distribution defined as

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{D} \right] \text{ where } \text{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}, \ \hat{\rho} = \frac{d_{12}}{\sqrt{d_{11} \times d_{22}}}$$

Where d_{11} and d_{22} are the variance of random intercept associated with family richness and IgA levels respectively while d_{12} is the covariance between the two random effects. The covariance is estimated since our interest lies on inference for the correlation between the two random effects ($\hat{\rho}$). For the precision of the covariance matrix D, we assumed a non-informative (NI) Wishart hyperprior distribution given by

$$\text{D}^{-1} \sim \text{Wishart}(R_D, \nu)$$

where R_D is a 2×2 identity matrix and $\nu = 2$ is the degrees of freedom (df) which must be small for NI prior. Here, we have chosen identity matrix with two degrees of freedom which is equal to the rank of R_D as suggested by Lesaffre and Lawson (2012).

Further, we extended model 3 by assuming unequal residual variances ($\sigma_1^2 \neq \sigma_0^2$) in the treatment groups and maintaining other assumptions. The resulting model (M4) is given by

$$\begin{cases} X_{ij}|Z_i, b_{1i} \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = (\alpha_{0j} + b_{1i}) + \alpha_{1j} Z_i \\ Y_{ij}|X_i, Z_i, b_{2i} \sim \text{Normal}(\mu_{ij}, \sigma_m^2) & \text{where } \mu_{ij} = (\beta_{0j} + b_{2i}) + \gamma_{1j} X_i + \beta_{1j} Z_i, \ m = 0, 1 \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma_m^2 \sim \text{IG}(0.001, 0.001), & \textbf{Prior} \end{cases} \tag{4}$$

To assess the impact of the prior distribution on the posterior estimates, uniform priors for

the standard deviations of the random effects and correlation coefficient ($\hat{\rho}$) was chosen. In addition, equal residual variance ($\sigma_1^2 = \sigma_0^2$) is also assumed. The joint model (M5) is an extension of model 3 and is therefore given by

$$\begin{cases} X_{ij}|Z_i, b_{1i} \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = (\alpha_{0j} + b_{1i}) + \alpha_{1j}Z_i \\ Y_{ij}|X_i, Z_i, b_{2i} \sim \text{Normal}(\mu_{ij}, \sigma^2) & \text{where } \mu_{ij} = (\beta_{0j} + b_{2i}) + \gamma_{1j}X_i + \beta_{1j}Z_i \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma^2 \sim \text{IG}(0.001, 0.001), \end{cases} \tag{5}$$

With hyperpriors $\sigma_{b1i} \sim U(0, 20), \quad \sigma_{b2i} \sim U(0, 20), \quad \rho \sim U(-1, 1)$. Note that σ_{b1i} and σ_{b2i} denotes the standard deviation of b_{1i} and b_{2i} respectively, while ρ denotes the correlation coefficient between the two random effects.

Finally, model 5 is extended by assuming unequal residual variance ($\sigma_1^2 \neq \sigma_0^2$) and keeping other assumptions. The model (M6) is formulated as

$$\begin{cases} X_{ij}|Z_i, b_{1i} \sim \text{Poisson}(\lambda_{ij}) & \text{where } \log(\lambda_{ij}) = (\alpha_{0j} + b_{1i}) + \alpha_{1j}Z_i \\ Y_{ij}|X_i, Z_i, b_{2i} \sim \text{Normal}(\mu_{ij}, \sigma_m^2) & \text{where } \mu_{ij} = (\beta_{0j} + b_{2i}) + \gamma_{1j}X_i + \beta_{1j}Z_i, \quad m = 0, 1 \\ \alpha_{kj} \sim \text{N}(0, 10^5), \quad , \beta_{kj} \sim \text{N}(0, 10^5), \quad \gamma_{1j} \sim \text{N}(0, 10^5), \quad \sigma_m^2 \sim \text{IG}(0.001, 0.001), \end{cases} \tag{6}$$

$$\sigma_{b1i} \sim U(0, 20), \quad \sigma_{b2i} \sim U(0, 20), \quad \rho \sim U(-1, 1)$$

The upper bound of the uniform distribution regarding the standard deviations is arrived at after evaluating a span of a large range of values from 5 to 100. It was established that the values between 10 and 100 led to stable posterior estimates implying that the inferences are no longer sensitive to the choice of the upper bound (Gelman et al., 2014). From the joint random effects models (M3 to M6) formulated, the univariate mixed effects models are linked together in two ways. First, by conditioning on the family richness (X) as explained in section 3.2.1 and second, by allowing the random effects to be correlated which is achieved by specifying a joint bivariate normal distribution for the random effects (Fieuws and Verbeke, 2006). To select the best model, deviance information criterion (DIC) was used (see section 3.3.0.2 for details)

3.2.2.1 Basis for Selecting Prior Distributions/Specification

The Bayesian methodology utilizes both the likelihood and prior information in constructing the posterior distribution. However, controversies have surrounded the choices of prior distribution because it can introduce subjectivity into the analysis (Lesaffre and Lawson, 2012). Statisticians

have proposed the adoption of non-informative (NI) priors to express lack of knowledge (Gelman et al., 2014; Lesaffre and Lawson, 2012). For the regression coefficients, we used independent normal priors with large variance as suggested by Lesaffre and Lawson (2012) to minimize their impact on the posterior distribution since the aim is to get information from the data. The residual variance is given a non-informative prior the so-called inverse gamma prior with the shape parameter $\alpha = 0.001$ and scale parameter $\beta = 0.001$ as suggested by Shkedy et al. (2003) and Lesaffre and Lawson, (2012).

3.3 Model fitting and Diagnostics

3.3.0.1 Convergence Test

The Bayesian joint models presented in section 3.2 were fitted to the data. Three parallel chains, each of length 200,000 were initiated to allow the simulation to be representative of the target posterior distribution (Gelman et al., 2014) with a burn-in period of length 100,000 to decrease the effects of the initial values. Further, a thinning factor of 100 was applied to lower the autocorrelation which was quite high. Although thinning as the advantage of lowering autocorrelation, it increases the Monte Carlo error (Lesaffre and Lawson, 2012). However, in this analysis, the chains were run until Monte Carlo error was smaller than 5% of the posterior standard deviation to avoid loss of precision. Thus, the inference is based on three chains, each of 1,000 samples of the targeted posterior distribution.

Here, multiple chains with overdispersed starting values are preferred over a single chain since the latter might get stuck in a local mode, especially if the posterior has a multimodal distribution (Lesaffre and Lawson, 2012) hence, yielding unreliable posterior estimates. To obtain good starting values for fixed parameters to enable quick mixing rate, classical linear and Poisson regression models were fitted to the transformed IgA levels and family richness at each time point respectively. The maximum likelihood estimates with their corresponding lower and upper 95% confidence intervals were then used as starting values. Trace plots and potential scale reduction factor (PSRF) are used to assess convergence of the chains as discussed in the subsequent section.

Trace Plots

These are graphical tools that show how rapidly the chains explore the posterior distribution by inspecting the chains mixing rates. It also establishes whether the chains depend on the initial values (Lesaffre and Lawson, 2012). Trace plots are often used as the first choice for assessing

convergence before a formal assessment is conducted. If it appears as a horizontal strip and the individual moves are invisible, then this signifies stationary while upward or downward trends implies dependence on initial values (Lesaffre and Lawson, 2012). Thus, the chains were run until stationarity was attained.

Potential Scale Reduction Factor (PSRF)

PSRF is a formal convergence test proposed by Gelman and Rubin (1992). This method utilizes within and between chain variability to estimate the posterior variance ($\hat{V}$) of each parameter estimate (say β^k). For instance, let m be the number of chains, n be the length of each chain after discarding burn-in period and K be the total number of parameters in the model (both fixed and random effects). Then the potential reduction factor for the k^{th} parameter is defined as

$$\hat{R} = \frac{\frac{n-1}{n}W + \frac{1}{n}B}{W} = \frac{\hat{V}}{W} \tag{7}$$

Where B is the between chain variability and W is the within chain variability. A corrected version of $\hat{R}$ which takes sampling variability into account is proposed and is given by $\hat{R}_c = (\hat{d}+3)/(\hat{d}+1)\hat{R}$ with $\hat{d} = 2\hat{V}/\text{var}(\hat{V})$ (Gelman et al., 2014; Lesaffre and Lawson, 2012).

Gelman et al. (2014) proposed estimation of this ratio for all parameters of interest whereby $\hat{R}_c$ values near 1 or smaller than 1.1 signifies convergence to the target distribution while values greater than 1.1 implies that more iterations are necessary.

3.3.0.2 Model Selection Using DIC

Deviance information criterion (DIC) which takes into account the complexity of the hierarchical models (Spiegelhalter et al., 2002) is used to evaluate the models listed in section 3.2.1 and 3.2.2. In Bayesian, the unknown regression parameters are estimated by the posterior means. However, this posterior means are estimated using the sample means of posterior samples generated from Markov Chain Monte Carlo (MCMC) methods especially when the full conditional posterior distribution does not have a closed analytical form (Shkedy et al., 2004). From the posterior samples obtained after discarding the burn-in period, Bayesian deviance $D(\theta)$ and θ can be monitored from an MCMC run. Where θ is a random variable denoting a vector of parameters of the joint posterior distribution. Posterior mean of Bayesian deviance $\overline{D(\theta)}$ and deviance, $D(\overline{\theta})$, evaluated at the posterior expectation of θ are useful in estimation of DIC as defined in

equation 8 (Shkedy et al., 2004; Lesaffre and Lawson, 2012). DIC is calculated as follows

$$DIC = \overline{D(\theta)} + P_D, \quad P_D = \overline{D(\theta)} - D(\bar{\theta}) \tag{8}$$

Where P_D is a measure of the effective number of parameters in the model (model complexity) and a model with smallest DIC indicates a better fit to the data set (Spiegelhalter et al., 2002).

3.3.0.3 Detections of Outliers

The detection of outlying observations which have unusual response profiles is important in order to prevent their influence on the analysis. When they are found to have a great impact on statistical inference, remedial measures are required (Fitzmaurice et al., 2012). In most cases, it is not recommended to discard outlying influential cases, especially when the sample size is small (Neter et al., 1996). Posterior predictive ordinate (PPO_i) defined as the posterior predictive distribution evaluated at the outcome variable for the i^{th} observation (Lesaffre and Lawson, 2012) is used to check for outliers in both family richness and transformed IgA levels. PPO_{ij} for the i^{th} subject at time j is estimated from Monte Carlo Markov Chain (MCMC) using

$$\hat{P}(y_{ij}|y) = \frac{1}{K}\sum_{k=1}^{K} p(y_{ij}|\theta^k) \tag{9}$$

Where K is the number of iterations for the converged posterior samples and θ is the vector of posterior estimates for the regression model (Lesaffre and Lawson, 2012). Here, we adapt the formula by substituting Poisson density for the family richness and normal density for the transformed IgA levels evaluated using posterior samples $(\theta^1, ..., \theta^K)$. An observation with too low value of PPO_i is considered an outlier.

4 Results and Interpretation

4.1 Exploratory Data analysis (EDA)

EDA provides an avenue of uncovering patterns in the data which are relevant to the scientific questions of interest and serves as the foundation for data analysis (Diggle et al., 1994). Individual and mean profile plots for each bacterial family under consideration are constructed (Figure 9 in Appendix). This process allowed us to identify a potential surrogate marker which reflects the immune response in the presence of intervention. Therefore, out of 30 bacterial families, only one family is selected namely S24-7. The other remaining families are unselected because of their longitudinal profiles, which poorly reflected the immune response. Thus, special attention is given to the S24-7 family.

4.1.1 Individual Longitudinal Profile Plots

Figure 3 displays individual profile plots of the bacterial family richness and logarithm transformed IgA levels of 15 mice measured at time 1, 6, 12 and 20. An inspection of the plots reveals that in overall, all the Mice gained in both family richness and IgA levels. However, two observations in the experimental group (PAT) and one in the control group displayed outlying patterns in family richness which undoubtedly pulled the means towards their locations. Also, some observations did not maintain their family richness as well as IgA levels over time, some ended with low values while others ended with high values.

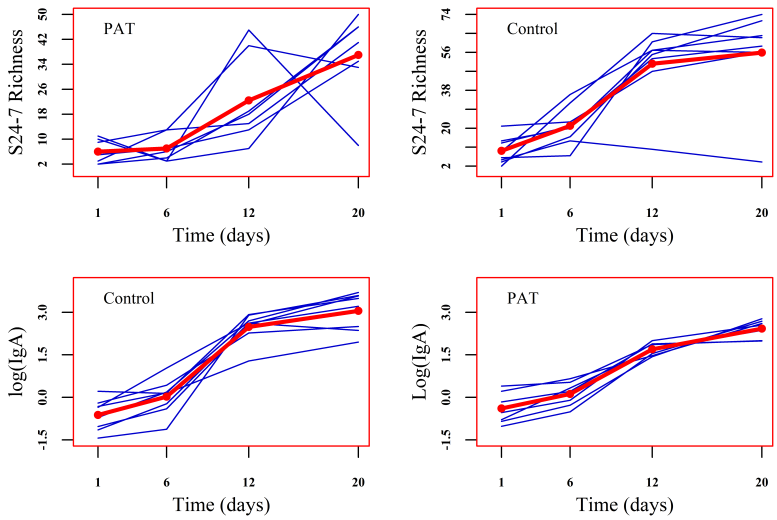

Figure 3: *Individual profile plots for S24-7 family richness and transformed IgA levels. The bold line represents the connected averages at each time point.*

Moreover, high variability is evident at the beginning and the end of the study in the family richness which can be explained by varying microbial growth rates at the individual level. The spread of IgA levels in the control group is quite high on both ends while high variability is only palpable at the beginning in the experimental group. These patterns suggest an inclusion of random intercept to capture between-subject variabilities (Verbeke and Molenberghs, 2000). In addition, the differential in variability between experimental and control group in IgA levels is accounted for by allowing distinct variances during the modeling process (see Table 1 for clarity).

4.1.2 Mean Profiles

While the individual profile plots described subject-specific patterns, it is equally necessary to explore how the mean profiles of the treatment groups evolve over time and determine whether the mean at each time point differ among the treatment groups (Fitzmaurice et al., 2004). This is feasible since all the individual measurements were taken at fixed time points and no missing data were reported.

Figure 4 displays the mean family richness and transformed IgA levels per treatment arm recorded at time 1, 6, 12 and 20. The points denote the arithmetic mean of the responses at each time point within each treatment group while the bars represent the standard errors of the means. From the plots, it is more apparent that the family richness and IgA levels are increasing over time in both treatment groups. Again, consistently high mean values of family richness and IgA levels are observed in the control group over the entire study excluding time 1 and 6 of IgA levels.

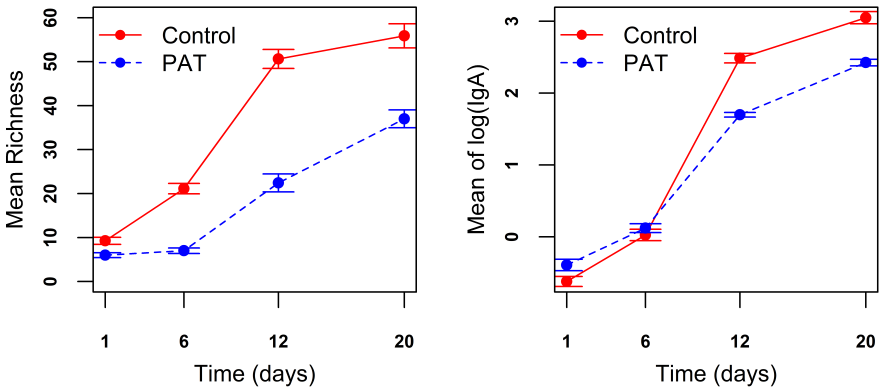

Figure 4: *The mean profile plots for family richness (left panel) and transformed IgA levels (right) split by treatment arm. The points denotes the arithmetic means and the bars represents the standard errors.*

Large standard errors are more pronounced at time 12 and 20 in the family richness, which is explained by the presence of potential outlying observations as was observed on the individual profile plots. Moreover, the plots seem to reveal that the family richness is more sensitive to the intervention than the IgA levels as shown by large mean difference at each time point.

Finally, the plots seem to suggest the adoption of unstructured mean to capture treatment differences at each time point due to the non-linear average trend. Therefore, imposing structure on the underlying mean response over time might shadow the treatment effects observed at the later stage. However, at this point, it is not yet possible to draw the conclusion about the significance of the treatment effects.

4.1.3 Covariance Structure of Transformed IgA levels

Table 1 shows the sample variance of the transformed IgA levels per treatment arm as well as the overall variance. At time one, the sample variance in both treatment groups is approximately the same (0.3). As time progresses, the variability in the two treatment groups differed with high variability in the control group. This suggests the adoption of different residual variance between the two treatment arms to investigate the effects of measurement errors. Furthermore, computing the overall variance regardless of the treatment group showed that the variances at each time point are slightly different. Thus, prompting additional assumption of equal residual variance to verify its validity in the modeling process.

Table 1: *Sample variance for PAT group, control group and overall for log-transformed IgA levels at each time point*

	Time			
	1	6	12	20
PAT	0.30	0.19	0.05	0.10
Control	0.31	0.40	0.28	0.46
Overall	0.30	0.28	0.33	0.38

4.2 Bayesian Evaluation of Surrogate Endpoints (S24-7 Family Richness)

4.2.1 Model Selection Using Deviance Information Criterion (DIC)

Table 2 provides the summary of deviance information criterion (DIC) for model selection. The models were ranked based on DIC values, from the best fit to the least fit. The Joint fixed effects models (models without random effects) have higher DIC values as compared to random effects models. This indicates that the former is not sufficient to describe the variability in the

dataset. Thus, a random effects model is required as anticipated since the individual profile plots showed high variability between subjects. Among the joint fixed effects models, the model with the lowest value of DIC assumes equal residual variance (model M1) while for joint random effects model, the model assuming equal residual variance and uniform prior distribution for standard deviations and the correlation coefficient of the random effects (model M5) has smallest DIC. In this study, we selected model M5 and therefore, was used for further inference.

Table 2: *Evaluation of joint Bayesian models using DIC*

	Model	Residual variance	Random effects	DIC	Rank
Joint fixed	M1	$\sigma_0^2 = \sigma_1^2$	-	732	5
effects	M2	$\sigma_0^2 \neq \sigma_1^2$	-	734	6
	M3	$\sigma_0^2 = \sigma_1^2$	Wishart	631	3
Joint random	M4	$\sigma_0^2 \neq \sigma_1^2$	Wishart	633	4
effects	M5	$\sigma_0^2 = \sigma_1^2$	Uniform	623	1
	M6	$\sigma_0^2 \neq \sigma_1^2$	Uniform	626	2

4.3 Results from the Joint Bayesian Hierarchical Model (M5)

Table 3 provides the summary of posterior means for the regression parameters. The parameters of interest with 95% credible intervals (CI) are visualized as shown in Figure 5 and the marginal posterior density plots in Figure 6. The treatment effects (α_1) on family richness are equal to -1.053 (95% CI: [-1.591,-0.505]) and -0.758 (95% CI: [-1.228,-0.278]) at time 6 and 12 respectively. They are both negative and different from zero as displayed in Figure 5a and 6a. This implies that the mean family richness for the experimental group (PAT) is 65% and 53% lower at time 6 and 12 respectively as compared to the control group. Indeed, this is supported by the mean profile plots (Figure 4) where high precision is observed at time 6 and large mean difference at time 12.

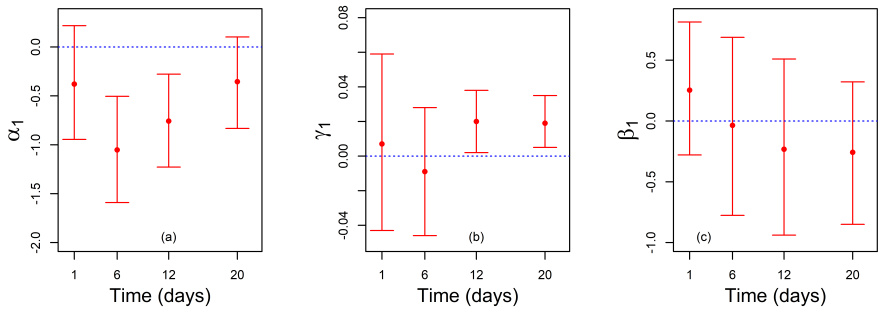

Figure 5: *95% credible interval of parameter estimates. Left: Treatment effects on family richness. middle: Family richness effects on transformed IgA levels. Right: Treatment effects on transformed IgA levels.*

On the contrary, the treatment effects on family richness at time 1 and 20 are not significant indicating that the two treatment groups are not statistically different at these time points. Furthermore, the effects of family richness (γ_1) on transformed IgA levels holding treatment constant are equal to 0.020 (95% CI: [0.002, 0.038]) and 0.019 (95% CI: [0.005, 0.035]) at time 12 and 20 respectively, which are both positive and different from zero as shown in Figure 5b and 6b. Thus, a unit increase in family richness is associated with an increase in the mean of transformed IgA levels by approximately 0.02 at the aforementioned time points. This positive linear association is also observed on the scatter plots as seen in section 2. The credible intervals are narrow implying moderate precision of the effects of family richness on transformed IgA levels. The parameter estimates at time 1 and 6 are equal to 0.007 (95% CI: [-0.043,0.058]) and -0.009 (95% CI: [-0.046,0.027]) which are well described by normal posterior density, however, their credible intervals are so wide and include zero. Thus, we can conclude that there is no linear relation between family richness and transformed IgA levels at this time points when treatment is held constant.

Moreover, the effects of treatment (β_1) on transformed IgA levels after considering the family richness is negative and not different from zero at all time points as shown by wide credible intervals in Figure 5c and 6c. This indicates that the average transformed IgA levels in the two treatment groups are not statistically different when family richness is held constant. Additionally, high between-subject variability is observed in family richness as shown by the posterior mean variance of 0.182 (95% CI: [0.046, 0.380]) while low between-subject variability is observed in transformed IgA levels as indicated by a variance of 0.016 (95% CI: [0.000, 0.062]). Further, the correlation coefficient between the two random effects is equal to -0.155 (95% CI: [-1.000,0.818]) which clearly implies no linear relation between the two random effects. The within-subject variability in the transformed IgA levels is estimated to be 0.241 (95% CI: [0.151,0.346]) which is quite high as expected since high variability is observed in the individual measurements especially the control group as shown in the individual profile plots (Figure 3) and the sample variance in Table 1.

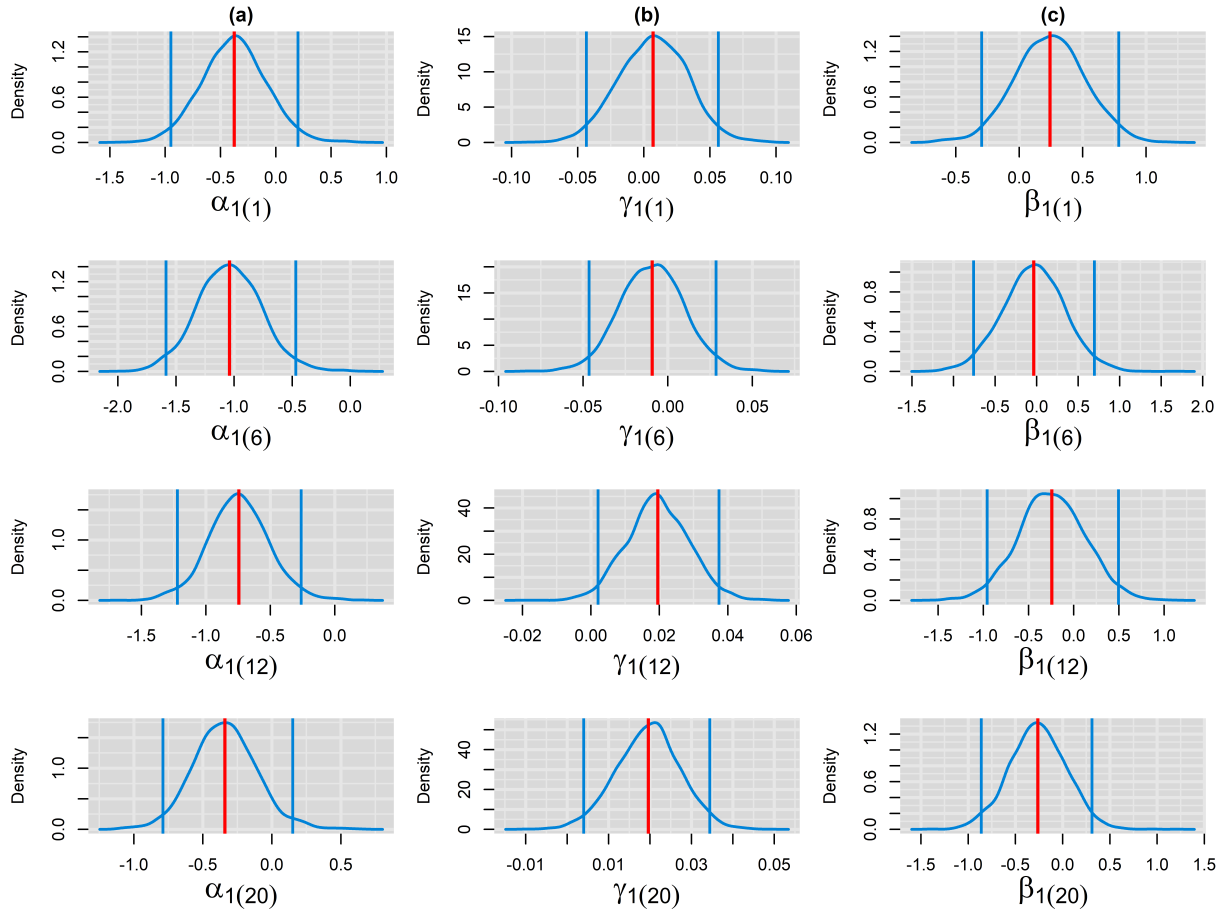

Figure 6: *Posterior density plots. The panel (a) from top to bottom represents the plot of the treatment effects on the family richness at each time point. The panel (b) shows the family richness effects on transformed IgA levels. The panel (c) displays the treatment effects on transformed IgA levels. The blue vertical lines represent the lower and upper 95% credible interval while the red vertical line denotes their corresponding posterior means.*

Table 3: *Posterior means and 95% credible interval at each time points in the joint mixed effects model.* $\alpha_{1(time)}$, $\gamma_{1(time)}$ *and* $\beta_{1(time)}$ *are the parameters of interest corresponding to the treatment effects on family richness, family richness effects on transformed IgA levels and treatment effects on transformed IgA levels respectively.*

Parameter	Mean	2.5%	97.5%
$\alpha_{1(1)}$	-0.379	-0.945	0.218
$\alpha_{1(6)}$	-1.053	-1.591	-0.505
$\alpha_{1(12)}$	-0.758	-1.228	-0.278
$\alpha_{1(20)}$	-0.355	-0.833	0.103
$\gamma_{1(1)}$	0.007	-0.043	0.059
$\gamma_{1(6)}$	-0.009	-0.046	0.028
$\gamma_{1(12)}$	0.020	0.002	0.038
$\gamma_{1(20)}$	0.019	0.005	0.035
$\beta_{1(1)}$	0.254	-0.279	0.814
$\beta_{1(6)}$	-0.036	-0.776	0.688
$\beta_{1(12)}$	-0.232	-0.939	0.510
$\beta_{1(20)}$	-0.258	-0.850	0.322
σ^2	0.241	0.151	0.346
d_{11}	0.182	0.046	0.380
d_{22}	0.016	0.000	0.062
$\hat{\rho}$	-0.155	-1.000	0.818

5 Model Fitting and Diagnostics

5.0.1 Assessing Convergence of Markov Chains

The convergence of all models was inspected using trace plots and potential scale reduction factor (PSRF) as described in section 3.3. However, after convergence, the autocorrelations of the Markov chains were pretty high which prompted the use of thinning of factor 100 in order to obtain independent samples. The chains were allowed to run again, and no further signals of strong autocorrelation were observed as shown in Figure 10 and 11 in appendix. After 100,000 iterations the trace plots exhibited a quite high mixing (Figure 12 and 13), an indication that the simulated sequences converged to a unique stationary distribution (Gelman et al., 2014). Similarly, the potential scale reduction factor (PSRF) for all parameters including random effect were close to one (Table 5). This confirms that the simulations reached convergence and therefore represents our target distribution. On that account, the posterior summary measures, including model comparisons are based on three chains, each of size 1000 posterior samples.

5.0.2 Outliers Detection in the Outcome Variables

To confirm whether the observations that displayed strange patterns in individual profile plots are indeed outliers, posterior predictive ordinate plot (Figure 7) and random effects plot (Figure 8) were constructed. The plot of random effects clearly identified one observation with a low count of family richness as an outlier (case XR01). The index plot pointed out two observations at time 12, one with extremely high counts (case XR07) and low count (case XR08) of family richness as seen in index number 6 and 7. Finally, two observations, namely case XR01 and XR07 with low PPO values are also classified as outliers at time 20. Further scrutiny of the aforementioned cases revealed that the observation number XR01 was assigned to unaltered cecal content (control treatment), observation number XR07 and XR08 were assigned to PAT altered cecal content. In general, the three outliers are clearly captured in the individual profile plots of family richness.

On the other hand, no outlying observations are observed in IgA levels since most observations appear to have a relatively moderate PPO values (Figure 8) and the random effects are concentrated around zero. This is also in agreement with the individual profile plots of transformed IgA levels. Although the three observations are considered outliers, we decided to retain them. The reason is that the sample size is so small (n=15) to warrant deletion.

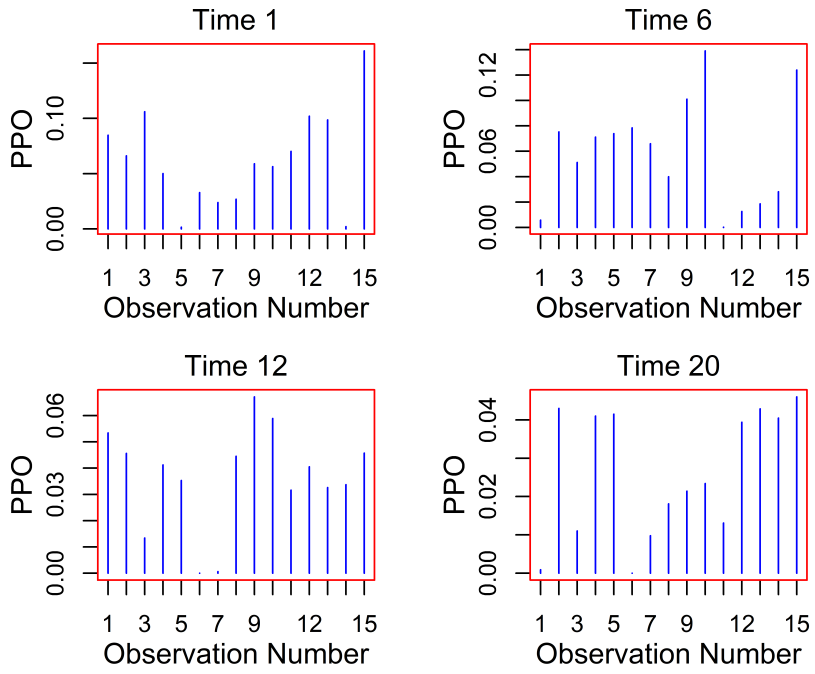

Figure 7: *Posterior predictive ordinate index plot at each time points for family richness*

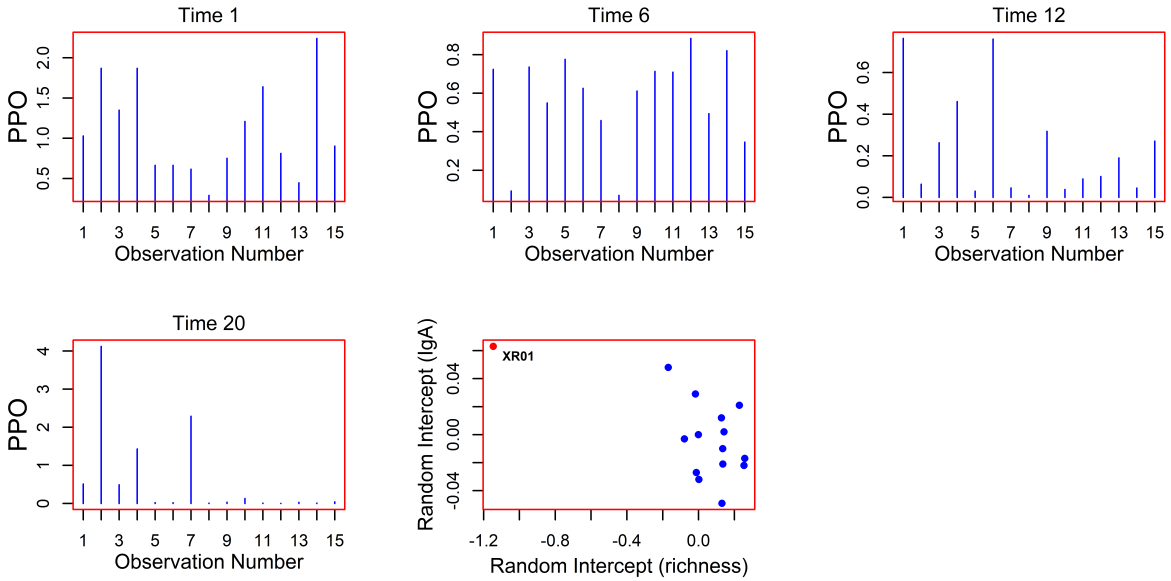

Figure 8: *Posterior predictive ordinate for transformed IgA and plot of random intercept*

6 Implementation of Bayesian Hierarchical Joint Model in R

The joint Bayesian models discussed in section 3.2 were implemented in the R runjags package which is a program developed to analyze Bayesian models using Markov Chain Monte Carlo (MCMC) methods. It offers several advantages over other BUGS and JAGS programs viz. It provides automated calculation of convergence diagnostics like potential scale reduction factor (PSRF); it allows easy access to graphical outputs and summary statistics; it provides additional distributions such as half-Cauchy which can be used as a prior for variance parameters and lastly, it allows multiple chains to be run in parallel which increases efficiency, thereby saving considerable amount of time (Denwood, 2016; Gelman et al., 2014).

The Bayesian hierarchical joint model (M5) with equal residual variance ($\sigma_0^2 = \sigma_1^2$), uniform prior distribution for the standard deviations of the random effects and the correlation coefficients ($\hat{\rho}$) is formulated as follows

```
for (i in 1:N){
for(t in 1:4){
# Poisson Likelihood
familyRichness[i,t] ~ dpois(lamda[i,t]) # Family richness follow Poisson.
log(lamda[i,t]) <- beta1[t] +b[i,1] + beta2[t]*trt[i]# Linear predictor (LP)
                                    # trt=Treatment, b=random intercept
# Normal Likelihood
log_IgA[i,t] ~ dnorm(mu[i,t],tau) # Transformed IgA follow Normal distr
mu[i,t] <-  beta3[t] +b[i,2]+beta4[t]*familyRichness[i,t]+beta5[t]*trt[i]
   }
# Prior Distribution of Random Effects (Bivariate Normal Distribution)
b[i,1:2]~dmnorm(zero,precision)
}
# Prior Distribution for Residual Variance
tau ~ dgamma(0.001,0.001) # The precision of the error term
sigma2 <- 1 / tau         # The variance of the error term
# Mean Vector of the Random Effects
zero[1]<-0 # Mean of the random intercept(family richness).
zero[2]<-0 # Mean of the random intercept (transformed IgA levels).
# Hyperprior distribution for the Elements of the D Matrix
```

```
d1  ~ dunif(0,20) # The standard deviation of random effects of richness
                  # is assumed to follow uniform distribution.
d2  ~ dunif(0,20) # The standard deviation of random effects of  IgA levels
                  # is assumed to follow uniform distribution.
corr ~ dunif(-1,1) # Correlation coefficient of random effects is assumed to
                  #follow uniform distribution.
# The D Covariance Matrix of the Random effects
d12<-d1*d2*corr    # Covariance between the two random intercepts
cov[1,1]<-pow(d1,2) # Variance of the random intercept for family richness
cov[1,2]<-d12 ;      # Covariance between the two random intercepts
cov[2,1]<-d12 ;      # Covariance between the two random intercepts
cov[2,2]<-pow(d2,2) # The Variance of the random intercept for IgA levels
precision[1:2,1:2]<-inverse(cov[,]) # The precision of the D matrix

# Prior Distribution for the Regression Coefficients
# All Time Specific Coefficients are Assumed to Follow Normal Distribution
# With Mean of Zero and  Large Variance of 100,000
for(t in 1:4){beta1[t] ~ dnorm(0,0.00001)} # intercept of richness
for(t in 1:4){beta2[t] ~ dnorm(0,0.00001)} # Treatment effects on richness
for(t in 1:4){beta3[t] ~ dnorm(0,0.00001)} # Intercept of IgA levels
for(t in 1:4){beta4[t] ~ dnorm(0,0.00001)} # Family richness effects on IgA
for(t in 1:4){beta5[t] ~ dnorm(0,0.00001)} # Treatment effects on IgA
    }"
```

It is also important to note that the model implemented here using runjags can as well be implemented in WinBUGS or any other JAGS program with little modification. For complete model formulation, the reader is directed to section 9.3. All the plots are constructed using R version 3.5.0.

7 Discussion and Conclusion

In this study, the Bayesian joint modeling approach was used to investigate whether the family S24-7 richness can be used to predict Immunoglobulin A levels. Two sets of models were fitted and evaluated namely models without random intercept and models with random intercept. Interestingly, all the models with random intercept led to the same qualitative conclusions regarding the treatment effects on the family richness with small differences observed in the parameter estimates. However, the results of joint fixed effects models were contrary in that the treatment effects at each time point were significant, whereas the former was significant at time 6 and 12. This explains the need for random effects in the model to capture the association structure for the repeated measurements to avoid underestimation of standard errors (Maas et al., 2005).

On the other hand, we observed that, assuming both equal and unequal residual variances for the treatment groups produced a different conclusion in the effects of family richness on IgA levels. In particular, the effects of family richness were not significant at time 12 in the models assuming unequal residual variance. This can be accounted for by loss of efficiency in model parameters while estimating extra residual variance (Verbeke and Molenberghs, 2000). Although it was established that all the joint random effect models assuming equal residual variance fits the data well. Additionally, negligible differences in posterior means were observed in the models assuming Uniform and Wishart distribution for the prior of the random effects, suggesting that the prior distribution selected had little impact on the posterior inference (Gelman et al., 2014).

The results of the best model for S24-7 family showed that PAT group has substantially low family richness at day 6 and 12 due to the negative treatment effects which sheds light on the negative impacts of PAT in inhibiting microbial growth and survival. Several studies have reported that different environmental conditions, as well as administration of PAT in murine animals, decreases the abundance of S24-7 family members (Ormerod et al., 2016; Ruiz et al., 2017) which concur with our current findings.

We also found that the average family richness for the two treatment groups is not different at the beginning and the end of the study. Lack of significant difference at the end of the study can be explained by declining treatment effects over time; as a result, the microbial community reverted to its original state as shown by the increasing family richness over time in the scatter plots. Similar results were also reported in a study investigating the effects of a single pulsed macrolide antibiotic treatment (PAT) in Mice; they found that, two weeks after

the administration of the treatment, microbial communities recovered to near normalcy (Ruiz et al., 2017). However, it is worth to investigate the effects of outlying observations identified at the end of the study since the mean profile plots showed a large mean difference.

Moreover, a positive linear association was observed between family richness and Immunoglobulin A levels towards the end of the study (day 12 and 20). This indicates that the impact of family richness on the IgA levels is predominant towards the end of the study. Recently, studies have explored the role of S24-7 family in immune system using animal studies and they found that some members of the family are targeted by the immune system (Ormerod et al., 2016) which supports our findings that an increase in family richness triggers production of more immunoglobulin A. Furthermore, the 95% credible intervals are slightly narrow, suggesting a moderate precision in predicting the effects of family richness on transformed IgA levels.

Alonso et al. (2016) suggested that a good surrogate endpoint should explain part of the treatment effect on the true endpoint which is clearly noticed here, because the treatment effects on the transformed IgA levels at each time point were found to be insignificant; which demonstrates that some treatment effects on IgA levels are captured by the family richness. This is further justified by the significant treatment effects on family richness observed at day 6 and 12 and supported by the mean profile plots which revealed the high sensitivity of family richness to treatment effects.

In addition, we also determined whether between-subject variation in the two endpoints are correlated by evaluating the random effects. We found high variability in family richness between subjects relative to the IgA levels, which are consistent with the studies of human and murine microbiome which reported that variations in microbial communities among individuals can be caused by several factors like hygiene, genotypes and colonization history (Huttenhower et al., 2012; Ruiz et al., 2017). However, the two random effects were found to be uncorrelated indicating that individual deviations from the population average in family richness, as well as IgA levels, are quite different which is evident in the individual profile plots. The 95% credible interval of the correlation coefficient is extremely wide, an indication of poor precision in the measure of the association which might be attributed to the small sample size.

Investigating the role of S24-7 family richness in the immune system of Mice provides insights on the human-microbe interactions because of anatomical and physiological similarities (Barré-Sinoussi et al., 2015). In this regard, the results presented have demonstrated that the IgA levels at day 12 and 20 can be predicted by the S24-7 family richness though with moderate

precision. Although these results look promising, the number of observations per treatment arm needs to be increased so as to conduct other Bayesian diagnostic test such as normality of the random effects. As such the stability of the results can be confirmed.

8 References

1. Alonso, A. and Molenberghs, G., 2008. Surrogate endpoints: hopes and perils. *Expert review of pharmacoeconomics & outcomes research*, 8(3), pp.255-259.

2. Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N.J., Shkedy, Z. and Van der Elst, W., 2016. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. CRC Press.

3. Barré-Sinoussi, F. and Montagutelli, X., 2015. Animal models are essential to biological research: issues and perspectives. *Future science OA*, 1(4).

4. Burzykowski, T., Molenberghs, G. and Buyse, M., 2005. *The evaluation of surrogate endpoints*. Springer, New York, NY.

5. Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., der Elst, W. and Burzykowski, T., 2016. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1), pp.104-132.

6. Buyse, M., 2007. Towards validation of statistically reliable biomarkers. *European Journal of Cancer Supplements*, 5(5), pp.89-95.

7. Denwood, M.J., 2016. runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9), pp.1-25.

8. Diggle, P., Liang, K.Y. and Zeger, S.L., 1994. Analysis of Longitudinal Data: Oxford Statistical Science Series.

9. Fenton, A., Lello, J. and Bonsall, M.B., 2006. Pathogen responses to host immunity: the impact of time delays and memory on the evolution of virulence. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1597), pp.2083-2090.

10. Fitzmaurice, G.M., Laird, N.M. and Ware, J.H., 2012. *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

11. Fleming, T.R. and DeMets, D.L., 1996. Surrogate end points in clinical trials: are we being misled?. *Annals of internal medicine*, 125(7), pp.605-613.

12. Funkhouser, L.J. and Bordenstein, S.R., 2013. Mom knows best: the universality of maternal microbial transmission. *PLoS biology*, 11(8), p.e1001631.

13. Fitzmaurice, G.M., Laird, N.M. and Ware, J.H., *Applied longitudinal analysis.* 2004. Hoboken Wiley-Interscience.

14. Fieuws, S. and Verbeke, G., 2006. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2), pp.424-431.

15. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B., 2014. *Bayesian data analysis.* CRC press

16. Grice, E.A. and Segre, J.A., 2011. The skin microbiome. *Nature Reviews Microbiology*, 9(4), p.244.

17. Hamady, M. and Knight, R., 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*, 19(7), pp.1141-1152.

18. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S. and Giglio, M.G., 2012. Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), p.207.

19. Ivanova, A., Molenberghs, G. and Verbeke, G., 2016. Mixed models approaches for joint modeling of different types of responses. *Journal of biopharmaceutical statistics*, 26(4), pp.601-618.

20. Iebba, V., Nicoletti, M. and Schippa, S., 2012. Gut microbiota and the immune system: an intimate partnership in health and disease.

21. Kelly, J., 2007. Understanding the immune system-how it works. *National Institute of Allergy and Infectious Diseases Science Education, United States*, p.60.

22. Langdon, A., Crook, N. and Dantas, G., 2016. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome medicine*, 8(1), p.39.

23. Lesaffre, E. and Lawson, A.B., 2012. *Bayesian biostatistics.* John Wiley & Sons.

24. Maas, C.J. and Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), p.86.

25. Mantis, N.J., Rol, N. and Corthésy, B., 2011. Secretory IgA's complex roles in immunity and mucosal homeostasis in the gut. *Mucosal immunology*, 4(6), p.603.

26. Marchesi, J.R. ed., 2014. The human microbiota and microbiome. CABI.

27. Marieb, E.N., 2008. Essentials of anatomy and physiology.

28. Martini, F.H. and Bartholomew, E.F., 2013. *Essentials of Anatomy & Physiology*: Pearson New International Edition. Pearson Higher Ed.

29. Molenberghs, G., Burzykowski, T., Alonso, A. and Buyse, M., 2004. A perspective on surrogate endpoints in controlled clinical trials. *Statistical methods in medical research*, 13(3), pp.177-206.

30. Molenberghs G., Buyse M., Burzykowski T. (2005) The History of Surrogate Endpoint Validation. In: Burzykowski T., Molenberghs G., Buyse M. (eds) The Evaluation of Surrogate Endpoints. Statistics for Biology and Health. Springer, New York, NY

31. Mueller, N.T., Bakacs, E., Combellick, J., Grigoryan, Z. and Dominguez-Bello, M.G., 2015. The infant microbiome development: mom matters. *Trends in molecular medicine*, 21(2), pp.109-117.

32. Piantadosi, S., 2017. *Clinical trials: a methodologic perspective*. John Wiley & Sons.

33. Matsuki, T. and Tanaka, R., 2014. Function of the human gut microbiota. *The Human Microbiota and Microbiome*, 90.

34. Roesch, L.F., Lorca, G.L., Casella, G., Giongo, A., Naranjo, A., Pionzio, A.M., Li, N., Mai, V., Wasserfall, C.H., Schatz, D. and Atkinson, M.A., 2009. Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. *The ISME journal*, 3(5), p.536.

35. Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A. and Buyse, M., 2010. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. *Statistical Methods in Medical Research*, 19(3), pp.205-236.

36. Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A. and Buyse, M., 2008. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *Journal of statistical planning and inference*, 138(2), pp.432-449.

37. Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W., 1996. *Applied linear statistical models* (Vol. 4, p. 318). Chicago: Irwin.

38. Nguyen, T.L.A., Vieira-Silva, S., Liston, A. and Raes, J., 2015. How informative is the mouse for human gut microbiota research?. *Disease models & mechanisms*, 8(1), pp.1-16.

39. Pflughoeft, K.J. and Versalovic, J., 2012. Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease*, 7, pp.99-122.

40. Riedel, C.U., Schwiertz, A. and Egert, M., 2014. *The stomach and small and large intestinal microbiomes* (pp. 1-19). CABI.

41. Ruiz, V.E., Battaglia, T., Kurtz, Z.D., Bijnens, L., Ou, A., Engstrand, I., Zheng, X., Iizumi, T., Mullins, B.J., Müller, C.L. and Cadwell, K., 2017. A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity. *Nature communications*, 8(1), p.518.

42. Sender, R., Fuchs, S. and Milo, R., 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8), p.e1002533.

43. Shekhar, S., Schenck, K. and Petersen, F.C., 2017. Exploring Host-Commensal Interactions in The Respiratory Tract. *Frontiers in immunology*, 8, p.1971.

44. Shkedy, Z., Torres Barbosa, F., Burzykowski, T. and Molenberghs, G., 2003. A hierarchical bayesian approach for the evaluation of surrogate endpoints in multiple randomized clinical trials.

45. Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P.,2004. Hierarchical Nonparametric Bayesian Models for the Force of Infection for Mumps and Rubella.

46. Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), pp.583-639.

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

48. Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. New York: Springer

49. Wang, B., Yao, M., Lv, L., Ling, Z. and Li, L., 2017. The human microbiota in health and disease. *Engineering*, 3(1), pp.71-82.

50. Woof, J.M., 2013. Immunoglobulin A: molecular mechanisms of function and role in immune defence. In *Molecular and Cellular Mechanisms of Antibody Activity* (pp. 31-60). Springer New York.

9 Appendix

9.1 Figures

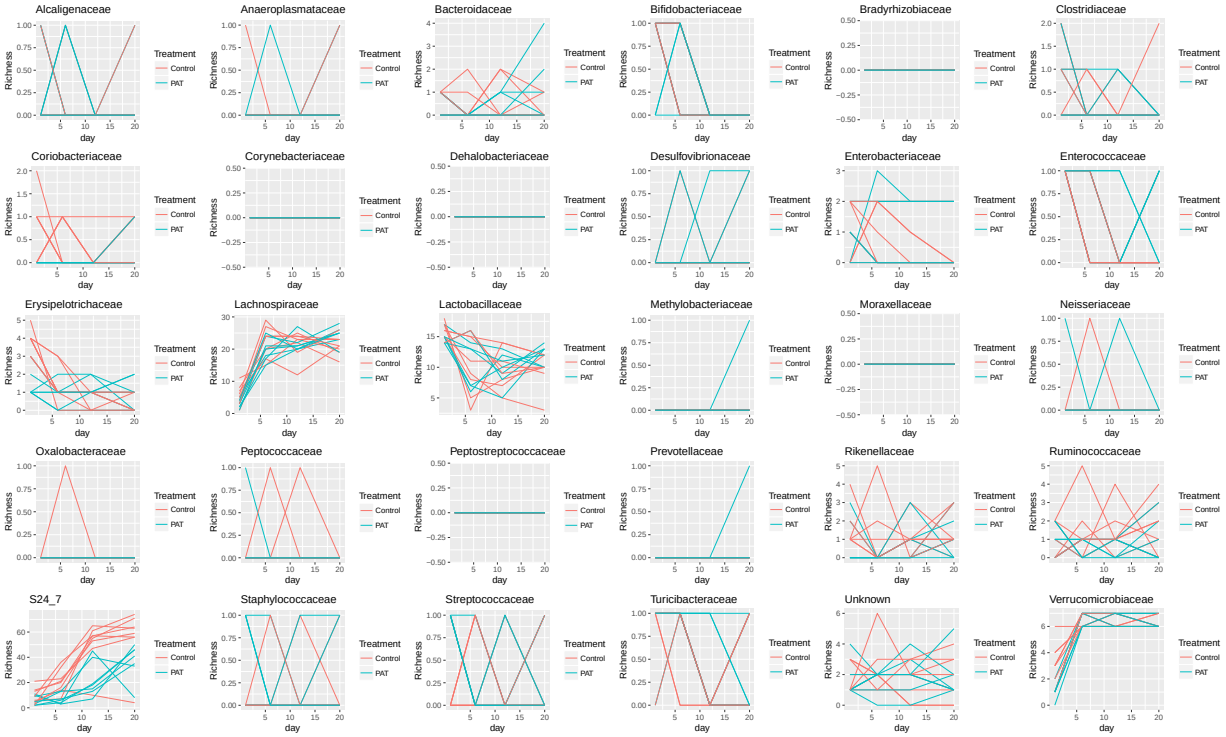

Figure 9: *Individual profile plots for all the Bacterial families*

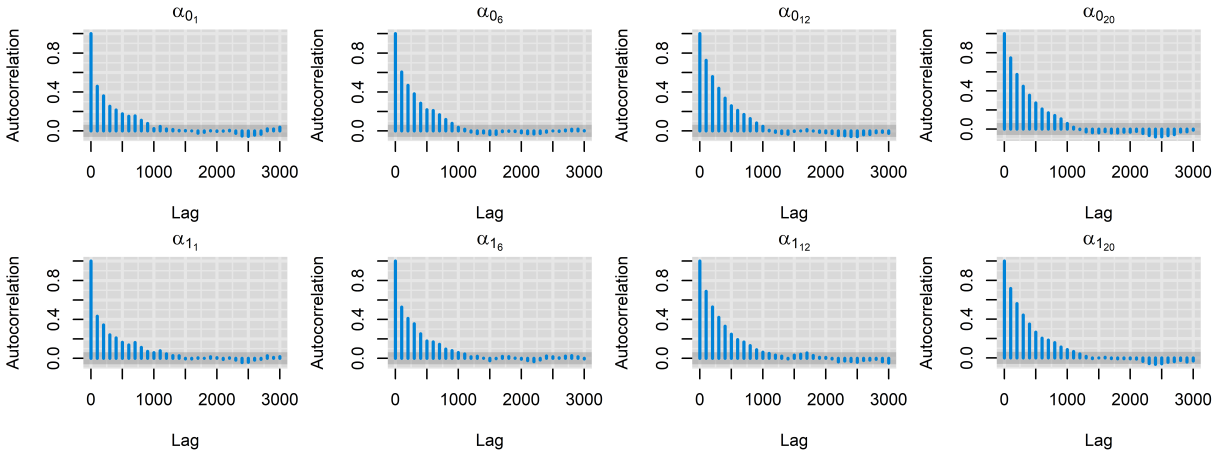

Figure 10: *Autocorrelation plot (ACF) for time specific intercept and treatment effects on family richness*

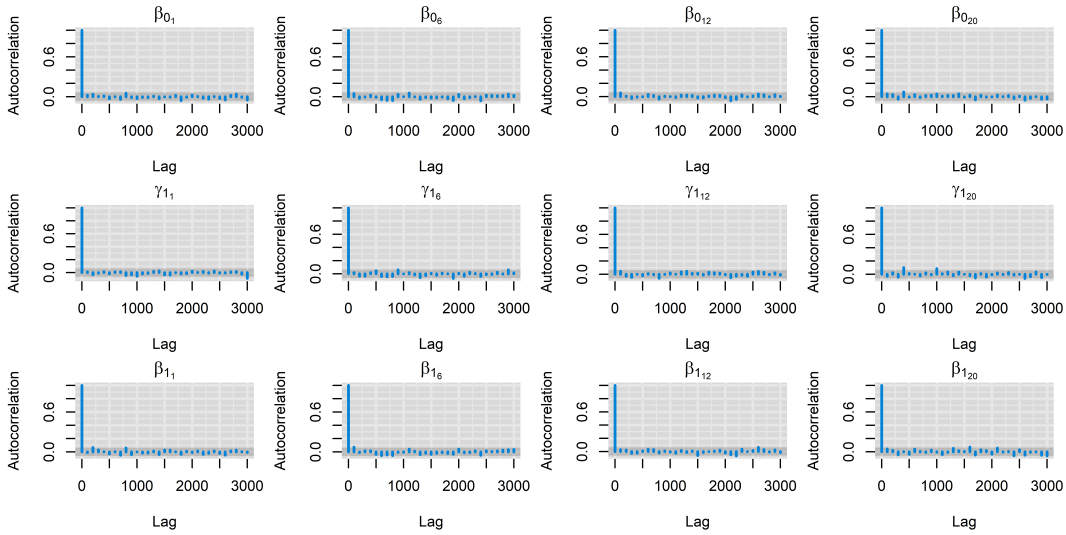

Figure 11: *ACF for time specific intercept, treatment effects and family richness effects on transformed IgA levels*

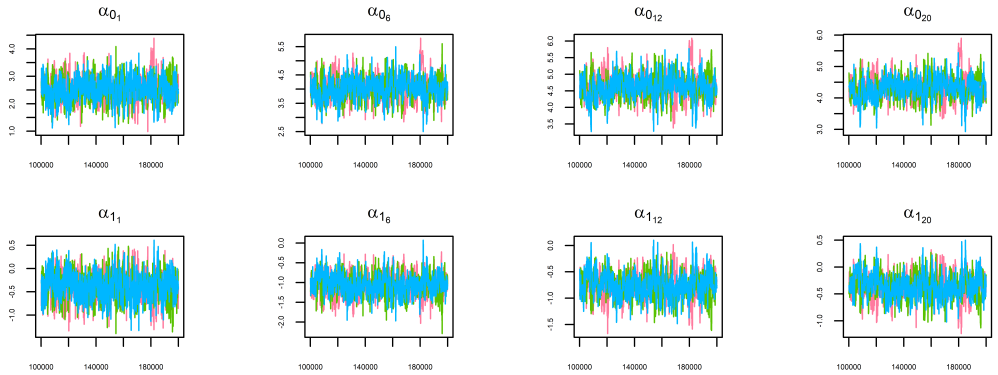

Figure 12: *Trace plot for time specific intercept and treatment effects on family richness*

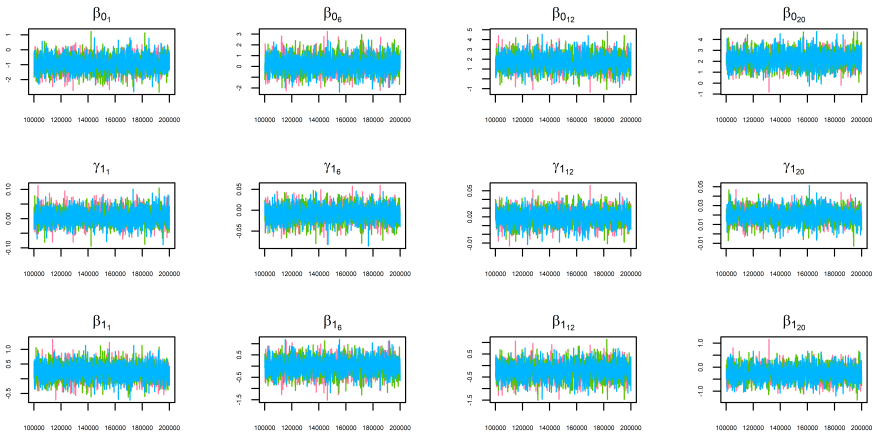

Figure 13: *Trace plot for time specific intercept,treatment effects and family richness effects on transformed IgA levels (M5)*

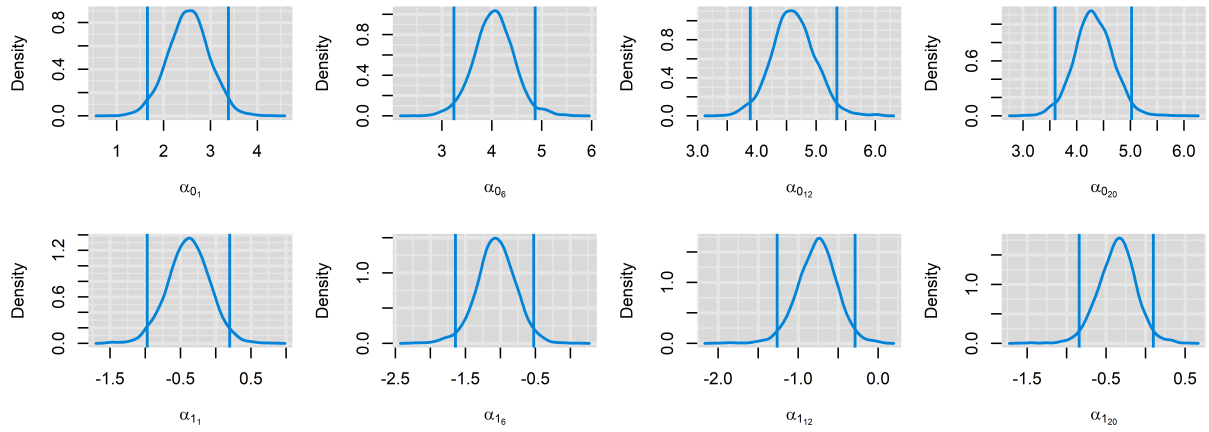

Figure 14: *Posterior density plot for time specific intercept and treatment effects on family richness (Model M5)*

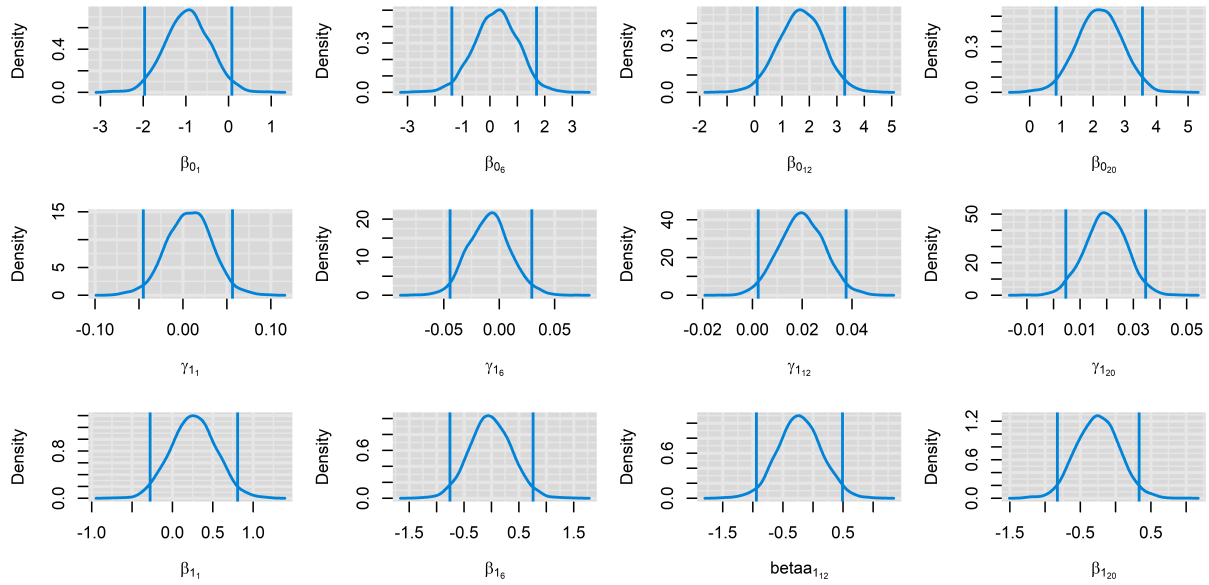

Figure 15: *Posterior density plot for time specific intercept (upper panel), family richness effects on IgA (middle Panel) and treatment effects on IgA (lower panel)*

9.2 Tables

Table 4: *Model Parameters*

	Model M1 ($\sigma_0^2 = \sigma_1^2$)			Model M2 ($\sigma_0^2 \neq \sigma_1^2$)			Model M3 ($\sigma_0^2 = \sigma_1^2$)		
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
$\alpha_{0(1)}$	2.655	2.098	3.194	2.218	1.989	2.446	2.513	1.557	3.464
$\alpha_{0(6)}$	4.160	3.746	4.574	3.047	2.894	3.196	4.016	3.120	4.886
$\alpha_{0(12)}$	4.739	4.493	4.992	3.924	3.829	4.022	4.597	3.787	5.428
$\alpha_{0(20)}$	4.435	4.214	4.658	4.022	3.927	4.112	4.291	3.477	5.103
$\alpha_{1(1)}$	-0.438	-0.820	-0.060	-0.439	-0.819	-0.054	-0.373	-1.000	0.260
$\alpha_{1(6)}$	-1.112	-1.433	-0.793	-1.111	-1.434	-0.792	-1.047	-1.655	-0.466
$\alpha_{1(12)}$	-0.816	-0.998	-0.630	-0.817	-1.001	-0.630	-0.751	-1.290	-0.220
$\alpha_{1(20)}$	-0.413	-0.568	-0.262	-0.413	-0.568	-0.261	-0.347	-0.868	0.184
$\beta_{0(1)}$	-0.935	-1.923	0.071	-0.770	-1.419	-0.149	-0.928	-2.157	0.261
$\beta_{0(6)}$	0.287	-1.200	1.762	0.299	-0.604	1.195	0.258	-1.459	1.975
$\beta_{0(12)}$	1.813	0.292	3.373	1.697	0.729	2.672	1.647	-0.115	3.485
$\beta_{0(20)}$	2.291	0.950	3.613	2.143	1.196	3.082	2.171	0.601	3.753
$\gamma_{1(1)}$	0.007	-0.043	0.056	0.016	-0.039	0.070	0.006	-0.047	0.063
$\gamma_{1(6)}$	-0.010	-0.046	0.025	-0.013	-0.051	0.027	-0.009	-0.050	0.029
$\gamma_{1(12)}$	0.019	0.001	0.035	0.016	-0.002	0.033	0.021	0.001	0.040
$\gamma_{1(20)}$	0.018	0.004	0.033	0.016	0.001	0.032	0.020	0.004	0.037
$\beta_{1(1)}$	0.251	-0.278	0.777	0.282	-0.261	0.811	0.249	-0.408	0.927
$\beta_{1(6)}$	-0.048	-0.781	0.641	-0.088	-0.824	0.660	-0.036	-0.887	0.815
$\beta_{1(12)}$	-0.265	-0.958	0.427	-0.349	-1.065	0.358	-0.206	-1.045	0.631
$\beta_{1(20)}$	-0.276	-0.850	0.291	-0.319	-0.901	0.269	-0.245	-0.940	0.487
σ_c^2	0.242	0.153	0.348	0.310	0.147	0.506	0.248	0.146	0.360
σ_p^2	0.242	0.153	0.348	0.195	0.088	0.328	0.248	0.146	0.360

Table 5: *Model Parameters*

	Model M4 ($\sigma_0^2 \neq \sigma_1^2$)			Model M5 ($\sigma_0^2 = \sigma_1^2$)				Model M6 ($\sigma_0^2 \neq \sigma_1^2$)		
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%	PSRF	Mean	2.5%	97.5%
$\alpha_{0(1)}$	2.143	1.721	2.549	2.523	1.655	3.410	1.000	2.145	1.768	2.516
$\alpha_{0(6)}$	2.972	2.596	3.351	4.027	3.232	4.833	1.000	2.975	2.628	3.302
$\alpha_{0(12)}$	3.848	3.500	4.215	4.609	3.865	5.324	1.000	3.850	3.527	4.164
$\alpha_{0(20)}$	3.947	3.590	4.305	4.305	3.575	5.024	1.000	3.949	3.623	4.254
$\alpha_{1(1)}$	-0.377	-1.018	0.243	-0.379	-0.945	0.218	1.000	-0.377	-0.977	0.185
$\alpha_{1(6)}$	-1.051	-1.645	-0.455	-1.053	-1.591	-0.505	1.000	-1.052	-1.585	-0.502
$\alpha_{1(12)}$	-0.754	-1.285	-0.208	-0.758	-1.228	-0.278	1.000	-0.755	-1.249	-0.293
$\alpha_{1(20)}$	-0.352	-0.877	0.182	-0.355	-0.833	0.103	1.000	-0.352	-0.822	0.114
$\beta_{0(1)}$	-0.765	-1.506	-0.047	-0.943	-1.968	0.082	1.000	-0.770	-1.419	-0.128
$\beta_{0(6)}$	0.282	-0.703	1.342	0.259	-1.279	1.763	1.000	0.287	-0.618	1.224
$\beta_{0(12)}$	1.624	0.512	2.742	1.720	0.094	3.326	1.000	1.661	0.626	2.661
$\beta_{0(20)}$	2.159	1.107	3.335	2.220	0.858	3.607	1.000	2.130	1.151	3.148
$\gamma_{1(1)}$	0.015	-0.043	0.076	0.007	-0.043	0.059	1.000	0.016	-0.039	0.071
$\gamma_{1(6)}$	-0.012	-0.056	0.029	-0.009	-0.046	0.028	1.000	-0.012	-0.052	0.026
$\gamma_{1(12)}$	0.017	-0.003	0.037	0.020	0.002	0.038	1.000	0.016	-0.002	0.035
$\gamma_{1(20)}$	0.016	-0.002	0.034	0.019	0.005	0.035	1.000	0.016	0.000	0.033
$\beta_{1(1)}$	0.279	-0.387	0.968	0.254	-0.279	0.814	1.000	0.282	-0.269	0.835
$\beta_{1(6)}$	-0.077	-0.973	0.796	-0.036	-0.776	0.688	1.000	-0.079	-0.838	0.678
$\beta_{1(12)}$	-0.308	-1.161	0.552	-0.232	-0.939	0.510	1.000	-0.329	-1.054	0.426
$\beta_{1(20)}$	-0.326	-1.042	0.421	-0.258	-0.850	0.322	1.000	-0.314	-0.928	0.294
σ_0^2	0.328	0.144	0.553	0.241	0.151	0.346	1.000	0.313	0.145	0.513
σ_1^2	0.189	0.076	0.332	0.241	0.151	0.346	1.000	0.191	0.085	0.329
d_{11}	0.241	0.084	0.462	0.182	0.046	0.380	1.000	0.182	0.047	0.378
d_{12}	-0.007	-0.146	0.137	-0.009	-0.084	0.048	1.000	-0.005	-0.076	0.062
d_{21}	-0.007	-0.146	0.137	-0.009	-0.084	0.048	1.000	-0.005	-0.076	0.062
d_{22}	0.153	0.050	0.303	0.016	0.000	0.062	1.004	0.017	0.000	0.065
$\hat{\rho}$	-0.034	-0.599	0.555	-0.155	-1.000	0.818	1.000	-0.095	-1.000	0.845

9.3 Rcodes

```
################################################################################
######## Bayesian Joint Fixed Effects Model M2: Unequal Residual Variances   ##
################################################################################
nburn<-100000    # Burn-in Period
samples<-100000  # Iteration for Posterior Inference
set.seed(2018)
model.microbiome<-"model{
for (i in 1:N){          # N is the Number of Observations, here N=15
for(t in 1:K){           # K is the Number of time periods
# Poisson Likelihood For Surrogate Endpoint (S24-7 Family Richness)
familyRichness[i,t] ~ dpois(lamda[i,t])
log(lamda[i,t]) <- beta1[t] + beta2[t]*trt[i] # Linear predictor (LP) with
                                              # trt=treatment
# Normal Likelihood For True Endpoint (log transformed IgA levels)
log_IgA[i,t] ~ dnorm(mu[i,t],taux[i]) # taux= precision for residual variance
mu[i,t] <-  beta3[t] +beta4[t]*familyRichness[i,t]+beta5[t]*trt[i] # LP
 }
# Precision for Residual Variance for Each Treatment Group
taux[i]<-(tau.c*(1-trt[i])+tau.p*(trt[i]))
}
# Prior Distribution for Residual Variance for Each Treatment Group
tau.c ~ dgamma(0.001,0.001) # Precision of Residual Variance for Control Group
tau.p ~ dgamma(0.001,0.001) # Precision of Residual Variance for PAT Group
sigma.c2<- 1 / tau.c        # Residual Variance for Control group
sigma.p2<-1 / tau.p         #Residual Variance for PAT group

# Prior Distribution for Regression Coefficients
for(j in 1:4){beta1[j] ~ dnorm(0,0.00001)}
for(j in 1:4){beta2[j] ~ dnorm(0,0.00001)}
for(j in 1:4){beta3[j] ~ dnorm(0,0.00001)};
for(j in 1:4){beta4[j] ~ dnorm(0,0.00001)}
for(j in 1:4){beta5[j] ~ dnorm(0,0.00001)} }"
```

```
################################################################################
########  Bayesian Hierarchical Joint Model (M5); Equal Residual Variances + ##
##########  Uniform Hyperpriors for Random Effects                            ##
################################################################################
model.microbiome5<-"model{
for (i in 1:N){          # N is the Number of Observations, here N=15
for(t in 1:K){           # K is the Number of time periods
# Poisson Likelihood (Generalized Linear Mixed Model)
familyRichness[i,t] ~ dpois(lamda[i,t])
log(lamda[i,t]) <- beta1[t] +b[i,1] + beta2[t]*trt[i] # Linear Predictor
# Posterior Predictive Ordinate (PPOi) for Checking Outliers in Family Richness
ppo1[i,t]<-exp(-lamda[i,t]+
familyRichness[i,t]*log(lamda[i,t])-logfact(familyRichness[i,t]))
# Normal Likelihood for IgA levels (Linear Mixed Model (LMM))
log_IgA[i,t] ~ dnorm(mu[i,t],tau.e)
mu[i,t] <-  beta3[t] +b[i,2]+beta4[t]*familyRichness[i,t]+beta5[t]*trt[i]
# Posterior Predictive Ordinate (PPOi) for Checking Outliers in IgA levels
ppo2[i,t]<- exp(-0.5*log(2*3.14)+0.5*log(tau.e)-0.5*tau.e*(
(log_IgA[i,t]-mu[i,t])*(log_IgA[i,t]-mu[i,t]))
}
# Distribution of Random effects
b[i,1:2]~dmnorm(zero,precision)
}
# Mean Vector of Random Effects
zero[1]<-0 ; zero[2]<-0
# Uniform Distribution for Standard Deviation of Covariance Matrix
d1  ~ dunif(0,20)  ; d2  ~ dunif(0,20) ; corr ~ dunif(-1,1); d12<-d1*d2*corr
# Covariance Matrix (D) of Random effects
cov[1,1]<-pow(d1,2)  # d11
cov[1,2]<-d12        # d12
cov[2,1]<-d12        # d21
cov[2,2]<-pow(d2,2)  # d22
```

```
precision[1:2,1:2]<-inverse(cov[,]) # D inverse
# Prior Distribution for residual variance
tau.e ~ dgamma(0.001,0.001) ;  sigma.e2<-1/tau.e
# Prior Distribution for Model Coefficients: Normal Distribution
for(t in 1:4){beta1[t] ~ dnorm(0,0.00001)}
for(t in 1:4){beta2[t] ~ dnorm(0,0.00001)}
for(t in 1:4){beta3[t] ~ dnorm(0,0.00001)}
for(t in 1:4){beta4[t] ~ dnorm(0,0.00001)}
for(t in 1:4){beta5[t] ~ dnorm(0,0.00001)}
 }"
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
High dimensional surrogacy in microbiome experiments: hierarchical Bayesian Approach

Richting: **Master of Statistics-Biostatistics**
Jaar: **2018**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Kipruto, Edwin

Datum: **15/06/2018**