Universiteit
Leiden
The Netherlands

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

Martin, I.

Citation

Martin, I. (2019, October 8). *Mixed models for correlated compositional data: applied to microbiome studies in Indonesia*. Retrieved from https://hdl.handle.net/1887/79254

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/79254

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/79254 holds various files of this Leiden University dissertation.

Author: Martin, I.
Title: Mixed models for correlated compositional data: applied to microbiome studies in Indonesia
Issue Date: 2019-10-08

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-

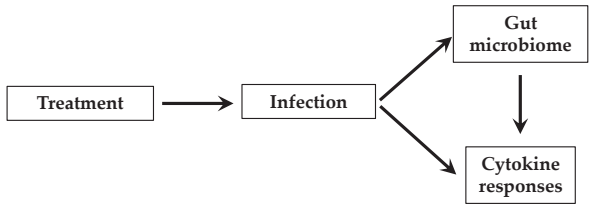

Figure 6.1: The hypothesized relationship based on the findings of our analyses.

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)].

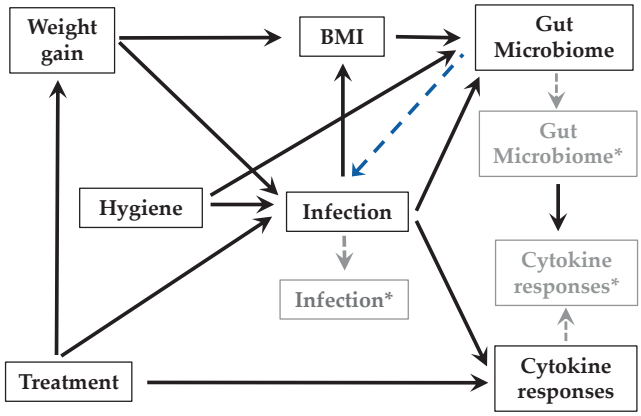

Figure 6.2: The DAG representing the relationship of all variables when measurement errors were included. The grey variables represent the observed variables with errors and blue line represents the possible causal direction.

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.

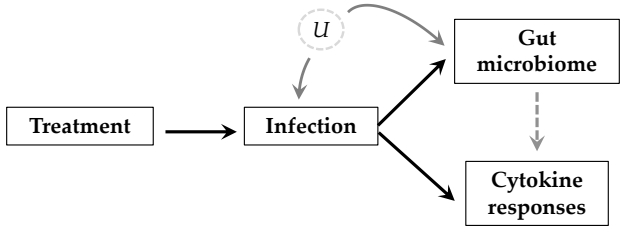

Figure 6.3: The concluded causal effect. The variable U represents latent variable to account for unobserved confounders.

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.