ADMIRE Project Proposal WP1

Olivier THAS

# Problem statement

*Describe concisely the state-of-the-art in the academic literature in the domains of the project. Indicate why the execution of the proposed research is essential.Position your project in relation to ongoing national and international research. Clarify how your project complements and overlaps with these developments.*

## Microbiome Data Analysis

Many statistical methods for microbiome data rely on distributional assumptions, e.g. the negative binomial (NB) or the zero-inflated NB (ZINB). However, recently we have demonstrated that these assumptions do not hold for the majority of the taxa (Hawinkel et al. 2020); this explains the poor performance of (ZI)NB-based methods w.r.t. their control of the false discovery rate (Hawinkel et al. 2019). Other methods rely on compositional data analysis (CoDa), which is a well established and mathematically supported class of methods for compositional data(Gloor et al. 2017)(Aitchison 1982). However, most CoDa methods start with log-transforms of (ratios of) counts, which are problematic with zero counts which are high highly abundant in typical microbiome datasets.

## Longitudinal Data Analysis

Most of the statistical methods for microbiome have been developed for independently distributed observations. However, complex experimental designs and longitudinal studies may result in observations that show a particular dependence structure for which other statistical methods are needed for correct statistical inference. The importance of longitudinal studies is stressed by several authors (e.g. (Schmidt, Raes, and Bork 2018), (Mai, Prosperi, and Yaghjyan 2016)), particularly because they are needed for studying the dynamics of microbial species and for discovering microbiome biomarkers and establishing causal relationships. Several methods for longitudinal data analysis of microbiome studies have been proposed, but these again rely on unrealistic assumptions. For example, (Chen and Li 2016) proposed a two-components model based on the zero-inflated beta distribution. This method is further improved by (Ho et al. 2019) by making it more flexible, but still relying on zero-inflated beta families. The methods of (Zhang et al. 2018) and (Zhang and Yi 2020) rely on the negative binomial (NB) and zero-inflated NB (ZINB) distribution, respectively. All these authors have performed simulation studies from which they conclude that their methods work well, but they all make the same fallacy: they simulated data from the same model as the one their method is based on. So this results in a self-fulfulling prophecy. For the evaluation of statistical tests for differential abundance in the simple 2-sample setting, simulation methods for generating realistically distributed count data have been proposed (Hawinkel et al. 2019; Assefa, Vandesompele, and Thas 2020), but these are not applicable yet for longitudinal study designs. (Williams et al. 2019) have developed a simulation framework for longitudinal microbiome data, but it is based on the multivariate normal distribution.

Other methods for the analysis of longitudinal microbiome data have been developed in the CoDa paradigm. (Xu and Chen 2020) proposed a two-parts model based on a zero-inflated multivariate normal distribution. The latter distribution should approximately hold for the log-ratio transformed counts. The between-taxa correlation structure is deduced from the taxonomic relationships and the serial correlation is assumed to have the exchangeable structure. The parameter are estimated as in the semiparametric GEE (generalised estimating equations) method, in which the correlation structure is part of the working correlation matrix. Valid asymptotic inference is thus guaranteed, even if the working correlation structure is misspecified. On the other hand, the parameter estimators are only semiparametric efficient if the correlation structure is correctly specified. The simulation study makes use of the multivariate normal distribution, which is consistent with their estimation equations and therefore they make the same fallacy as mentioned earlier.

A completely different approach has been taken by (Wang et al. 2020). Their method (MTA: Microbial Trend Analysis) combines splines for modelling the time trend and principal component analysis for dimension reduction. The method can be represented as a penalised matrix decomposition problem. In contrast to the methods we discussed earlier, MTA can be used for studying the time trend of a community instead of individual taxa. The bootstrap is used for p-value calculation and hence no distributional assumptions are imposed on the count data. The performance of the method is empirically evaluated in a simulation study in which the relative abundance data were generated from a Dirichlet distribution, which is not very realistic because it can only induce negatieve correlations between taxa. Also the methods of (Shields-Cutler et al. 2018) and (Luo, Ziebell, and An 2017) make use of splines for nonparametrically modelling the time trend. The former makes use of permutations for null distribution enumeration and can be used for comparing time trends between groups, and the latter can also identify time intervals in which the time trends differ between groups.  The nonparametric nature of these methods make that no continuous covariates can be included and that no clear interpretable parameter estimates are provided.

## Nonparametric and Semiparametric Methods

For the i.i.d. setting in which two groups are compared, several simulation studies have indicated that rank tests, such as the Wilcoxon rank sum test or SAMSeq(Li and Tibshirani 2013),  applied to appropriately normalised data work well in the sense that they control the FDR at the nominal level. These nonparametric tests, however, also suffer from a few disadvantages. The first is that they are not applicable to studies with a design that is more complicated than comparing two or more groups. For example, the inclusion of a continuous covariate or confounder is not possible. The second disadvantage is that these procedures are specifically developed for hypothesis testing and that they do not come with a corresponding estimate of a relevant effect size. Parametric methods, on the other hand, can be used for testing hypotheses that are explicitly formulated in terms of e.g. (log) fold changes, and these methods also provides estimates of this effect size. Finally, these rank tests are often much less sensitive than their parametric competitors. This is a price that these methods pay for being nonparametric.

**Probabilistic Index Models (PIM)**are a class of semiparametric models (Thas et al. 2012) that (1) generate the classical rank tests (De Neve and Thas 2015), (2) provide estimates of corresponding and informative parameters (effect sizes), and (3) possess the flexibility of regression models and hence allow for the inclusion of covariates. Although the original theory is asymptotic in nature, hypothesis tests can often be implemented as permutation tests or empirical likelihood methods can be used, with good results with sample sizes as small as 25 (Amorim et al. 2018). Thus, PIMs resolve the first two disadvantages that we listed for the classical nonparametric rank tests. Regarding the third disadvantage (small sensitivity), semiparametric methods may bring a solution. Upon using semiparametric efficiency theory (Tsiatis 2007), the estimation method can be designed so as to provide semiparametric efficient parameter estimators within a well-defined class of semiparametric models and restricted to asymptotically linear estimators. The latter can be considered as a compromise between the very wide class of nonparametric models and the very restrictive class of parametric models. For PIMs these semiparametric efficient estimators have been constructed and studied by (Vermeulen et al., n.d.). Their simulation study, however, indicated that, as compared to the original estimator of (Thas et al. 2012), only a small gain in precision could be obtained, and this came with a strongly increased computational cost.

Recent advances in semiparametric estimation theory and machine learning have resulted in algorithm-driven procedures for the construction of optimal statistical procedures. (Luedtke et al. 2020), for example, developed a deep adversarial learning method for constructing optimal minimax estimators, without the need of data and not restricted to asymptotically linear estimators. These methods are very useful in settings where a parametric model is known in advance but the optimal estimator is very hard to construct analytically. When data are available, adaptive semiparametric efficient estimators may be constructed as in (Bickel et al. 1993) or by the method of targeted maximum likelihood learning (Van Der Laan and Rubin 2006)(Van der Laan and Rose 2018). These methods adapt the estimator to the data that have to be analysed. A disadvantage that these procedures have in common is that they typically require a large sample size. In high-dimensional or large scale settings, as in microbiome studies, we are in a different situation: we have data on many similar features (taxa), but for each feature we only have a small number of observations. This largely prohibits the use of adaptive procedures to be applied to each individual feature, but on the other hand, information can be shared among the features, particularly because it is realistic to assume that all features have distributions that are not very dissimilar.

## Causal Mediation Analysis

### General Concepts

Unless applied to well designed (randomised) and executed studies, most statistical methods only allow for establishing associations but no causal effects. To translate microbiome research into solutions for crop production and treatment of diseases, causal relations need to be detected. Causal inference is well on its way to become an established theory and it is already frequently applied in e.g. clinical research, epidemiology, econometrics and sociology. Several authors have already cried out for the need for causal inference methods for microbiome studies (e.g. (Gilbert et al. 2016) (Maruvada et al. 2017)), and few methods have already been proposed for microbiome studies (see further). In microbiome research, it is often not known whether an intervention (e.g. change of diet) directly affects an outcome (e.g. immune response) or whether the outcome is caused by a change in the microbial communities, or both. Moreover, the microbiome can in turn also be affected by the intervention. This is illustrated in Figure ???. Mediation analysis may reveal the causal pathway of how a treatment or an intervention (A) affects an outcome (Y). This effect may take the direct path from A to Y (*direct effect*) and/or the path may go via a mediator (M) (*indirect effect*). In our context, this mediator is the (relative) abundance of a single taxon. Traditional methods for mediation analysis, as developed by (Judd and Kenny 1981) and (Baron and Kenny 1986),  were developed for a continuous outcome Y and rely on linear Structural Equation Models (SEM); see (MacKinnon 2012). In particular, the direct and indirect effects are estimated from two models: (1) an **outcome model** for , and (2) a **mediator model** for . The microbiome comprises more than a single taxon, but the general framework of (MacKinnon 2012) allows for mediation analysis with multiple mediators (see Figure ???).

The concepts of direct and indirect effects are properly defined in the causal inference literature, independently of the SEM formulation (Robins and Greenland 1992). We first introduction notation for **counterfactual outcomes**. Let  denote an outcome in treatment group  and with mediator M equal to , and let  denote a mediator in treatment group a. These outcomes and mediators may be counterfactual to what has been observed in the data. Another way of looking at it:  and  may be looked at as the two potential outcomes of a single subject; of course only one of the two can be realised. The (total natural) direct effect of the treatment on the outcome is defined as

and the (total natural) indirect effect is defined as

In the context where the causal effects are the target of statistical inference, they are referred to as the estimands. The next step is then to find good estimates of these estimands, which typically . estimands, which typically requires a connection between expectations of potential outcomes and conditional expectations as in the outcome model. To allow for this connection, a set of **identification assumptions** must hold  (see e.g. (VanderWeele 2015)). With the outcome and mediator  models of a SEM, these expectations can be expressed in terms of the model parameters, and upon plugging in the maximum likelihood estimators of the SEM, the (in)direct effects can be estimated. Despite the flexibility of this approach, it suffers from a very important drawback: correct estimates of the (in)direct effects can only be obtained if the two models in the linear SEM are correctly specified (Vanderweele and Vansteelandt 2009). This becomes particularly difficult with high-dimensional mediators, such as the taxa in the microbiome. Moreover, when the outcome is not continuous (e.g. binary) and nonlinear models are specified in a SEM, then there is no longer a theoretical justification for the use of these models for estimating the causal mediation effects; see e.g. (Pearl 2012) and reference therein. The causal mediation literature has also proposed other estimation methods that require less stringent model specifications, e.g. natural effect models and inverse probability weighting (Lange, Vansteelandt, and Bekaert 2012)(Vansteelandt, Bekaert, and Lange 2012), G-estimation and imputation. The former are models for  and hence they directly relate to the natural effects; these models reduce to marginal structural models when .

Bias in the estimators of the (in)direct effects may can from three sources. First, identification bias comes from violations of the identification assumptions. Model bias comes from misspecified statistical methods (e.g. outcome and mediator models), and, finally, estimation bias comes from inappropriate estimation procedures. (Díaz 2020) mentions only two types of bias, but types of bias; he merged two type first types.

### Multiple mediations and longitudinal studies

Two extensions of the classical mediation analysis will be important for this project: (1) multiple mediators and (2) longitudinal data. Both topics are the subject of intense ongoing research in the causal mediation analysis community. When multiple mediators are involved as in Figure ???, the problem cannot be simplified by studying one mediator at a time, unless under the unrealistic assumption that the mediators do not affect one another (VanderWeele and Vansteelandt 2014). Model-based methods have been developed (VanderWeele and Vansteelandt 2014), but correct inference requires correct model specification. With multiple mediators this may be hard, moreover because these models must also be compatible (i.e. not contradict one another). Probability weighting (VanderWeele and Vansteelandt 2014) is an estimation method that may partly resolve this issue. When multiple mediators are involved, several methods are only applicable when the mediators can be causally ordered (Daniel et al. 2015)(Steen et al. 2017), but this seems not feasible for the high-dimensional microbiome as mediator. (Vansteelandt and Daniel 2017) relaxed the structural dependence requirement, by shifting the attention to *interventional (in)direct effects*, which were first introduced by (VanderWeele, Vansteelandt, and Robins 2014) in the context of a single mediator. The estimation methods of (Loh et al. 2019), based on interventional effects models, still need the specification of the joint mediator distribution, but (Loh et al. 2020) further relaxed the conditions so that only the marginal mediator distributions need to be modelled, making the method more appropriate for the microbiome mediator.

When the microbiome is considered as a mediator, longitudinal studies should be set up so as to follow up the dynamic response of the microbial community. Several causal longitudinal mediation analysis methods have been proposed (Bind et al. 2016)(VanderWeele and Tchetgen Tchetgen 2017)(Zheng and van der Laan 2018). We only mention two papers. The recent work of (Mittinty and Vansteelandt 2019) extended the natural effects models to the longitudinal setting.  They use inverse probability weighting for parameter estimation, and for correct standard error calculation (accounting for the serial dependence), they propose a procedure that combines GEE with the parametric bootstrap. A second paper (VanderWeele and Tchetgen Tchetgen 2017) focuses on interventional effects, and use G-estimation for parameter estimation.  To our knowledge, there are still no methods for longitudinal mediation analysis with high-dimensional mediators.

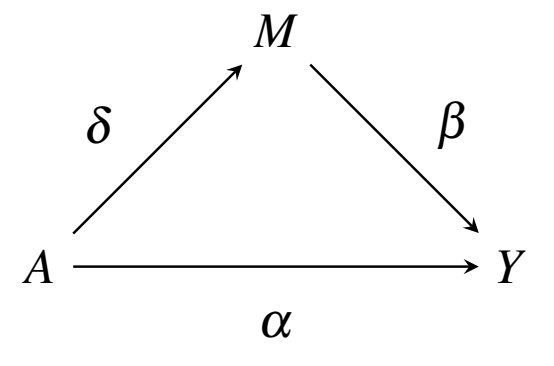
In this brief literature overview of mediation analysis, we so far avoided going into several other technical, though important issues: (1) identifiability assumptions, (2) confounding and covariates, and (3) decomposition. Regarding (2), most of the papers that we cited in the previous paragraphs, also consider the presence of confounders and covariates. In as far as they are observed, they can often be accounted for. Most problematic are the unmeasured confounders. Hardly any method can guarantee a correct causal conclusion in the presence of  unmeasured confounders, unless very restrictive assumptions are satisfied. This brings us needlessly to (1): all methods require identifiability conditions that need to be satisfied in order to make the (in)direct effects estimable. Throughout our literature overview, the definitions of (in)direct effects changed. Roughly summarised, it changed from effects, over natural effect to interventional effects. One of the motivations for changing definitions, is that it allows for relaxed identifiability assumptions, and with moving from a single mediator to multiple mediators and time-varying exposures and mediators.  Another motivation for changing definitions is related to (3): the decomposition of the total effect of the intervention into direct and indirect effects so that e.g. the indirect mediator effect can be expressed as a percentage of the total effect. This was the original crux of mediation analysis, but even without a decomposition, . decomposition, the estimates of the (in)direct effects are just as informative. When moving from a single to multiple mediators, and moving from cross sectional to longitudinal settings, the changes of the definitions of the effects were necessary to allow for such decompositions.

In previous paragraphs: better stress continuous versus generalised approaches.

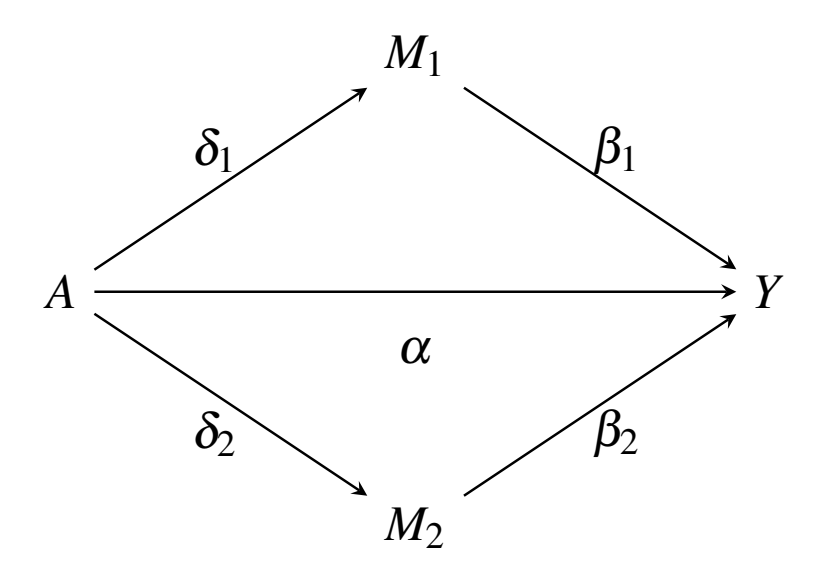
### Mediation Analysis for Microbiome Studies

(Zhang et al. 2019) proposed a method in the SEM framework (MacKinnon 2012) for microbiome data with a continuous outcome variable. Although their method allows for multiple mediators (taxa), their inference is targeted towards a single taxon. Their technical focus is on the high dimensionality and the compositionality of the mediator. For tackling the latter they transform the counts to their isometric log-ratio transform, and the high dimensionality is resolved by first fitting the outcome model with the lasso and then de-biasing the lasso estimator that corresponds to the targeted taxon (Zhang and Zhang 2014). Their method only focuses on testing the mediator effect on the outcome, and it does not directly address the distinction between the direct and indirect effects, nor do the authors consider the sparseness of typical microbiome data. (Sohn, Li, and others 2019) also focused on testing of the mediation effect. They also followed the path of the SEM and they also used a lasso-type of penalised estimation method for the outcome model to deal with the high-dimensionality; this method was particularly developed for high-dimensional compositional regressors (Shi et al. 2016) which also includes a de-biassing step.  The compositionality of the microbiome mediator is also accounted for in the mediator model, which is now modelling the full vector of additive log-ratio transformed counts by assuming that this vector is approximately normally distributed. The bootstrap is used for hypothesis testing. Yet another recent SEM-based mediation analysis method for microbiome studies was proposed by (Wang et al. 2020). Their outcome model includes the high-dimensional taxa data as additive log-ratios, but they also include the interaction effects between these taxa and the treatment. The model parameters are estimated using an penalised least squares method (Radchenko and James 2010) that involves two penalty terms to address both the hierarchy of main and interaction effects in the model and the high-dimensionality of the mediator. The model for the high-dimensional mediator makes use of the Dirichlet distribution; parameters are estimated by maximising the log-likelihood function with lasso-penalty terms. The latter helps in selecting the taxa that are most affected by the treatment. In contrast to the two previous methods, (Wang et al. 2020) does not use the model parameter estimates directly for statistical inference. They continue with defining the *average causal direct effect*of the treatment and the *average mediation effect* of the microbiome on the outcome, as in the counterfactual causal framework for mediation analysis (VanderWeele and Vansteelandt 2014). The parameter estimates of the two models are subsequently used for estimating these two effect sizes. For hypotheses testing for the mediation effect, a permutation scheme is used. The counterfactual paradigm has also been used by (Li et al. 2019) in a similar fashion. However, their models consider only a single mediator (and hence the mediation analysis has to be repeated for each taxon separately) and account for the many zero counts by separating zeroes from non zeroes. For the mediation model, they consider zero-inflated (ZI) distributions, such as ZI Poisson (ZIP, ZI Beta and ZI Lognormal. Maximum likelihood is used for parameter estimation, and these estimates are plugged into expression for the direct and indirect effects as defined in the counterfactual framework. Resampling methods are used for p-value calculation. (Zhang, Wei, and Chen 2018) also starts from a SEM, but instead of imposing a penalty in the estimation of the parameters in the high-dimensional models, they propose to perform a dimension reduction method on the microbiome data, and include the resulting low-dimensional constructs in the outcome model, and also use these constructs as outcome variables in the mediation model. The dimension reduction is accomplished by applying metric multidimensional scaling to an appropriate distance matrix. They consider distance metrics that are frequently used in ecology and microbiome research: Jaccard, Bray-Curtis, unweighted, weighted and generalised UniFrac. The latter three metrics account for the taxonomic relations between the taxa. The mediation effect is tested by relying on permutation testing.

All microbiome-specific methods from the previous paragraph may suffer from serious validity issues, as discussed in some more detail earlier in this project proposal. First, many methods rely on the SEM framework, which is, however, only valid if all models are correctly specified and only for continuous outcome variables (Vanderweele and Vansteelandt 2009). Correct model formulation becomes particularly hard when high-dimensional taxa data are involved. A taxon-by-taxon approach, as in (Li et al. 2019), has also been demonstrated to be invalid (Vanderweele and Vansteelandt 2009). In conclusion, most methods are model-based and do not rely on modern causal mediation analysis results.



Causal diagram



Causal diagram for two mediators

# Scientific Project Objectives

*Describe as explicitly as possible the scientific objectives that you wish to reach in this project proposal. What is the essence of the intended scientific progress from which the various concrete goals, criteria, activities and the desired results can be understood? List explicitly and concretely the results (or partial results) that will have to be achieved, such as specific knowledge, the solution to particular problems and the academic breakthroughs. Per sub-aspect, mention the major quantitative (preferably) and qualitative target values, criteria, requirements and norms that at the end of the project could determine to what extent the anticipated results were reached.*

**Overall Research Objective**: Development of robust and flexible data analysis methods for the analysis of microbiome studies, with a focus on longitudinal studies, visualisation, causal inference and biomarker discovery. A pilot microbiome study on the influence of the plant microbiome on the host microbiome, immune cell balance and disease is also included.

**Work package 1 (WP1)**: Development of flexible,  robust and efficient statistical methods for analysing microbiome data from clustered and longitudinal studies and for causal mediation analysis.

Supervisors: Prof. O. Thas (DSI), Prof. D. Valkenborg (DSI) and Prof. J. Vangronsveld (CMK)

**RO1**: We aim to develop a valid statistical modelling framework for analysing clustered and longitudinal microbiome data. The methods must be (1) flexible in the sense that they adapt to complex study designs (multiple factors, hierarchical dependence structure, correction for covariates/confounders, …)(RO1.1), (2) robust in the sense that they do not rely on unrealistic distributional assumptions and can cope with all microbiome-specific data characteristics (RO1.1) , and (3) efficient in the sense that the method provides precise estimates and powerful tests (RO1.2), despite its flexibility.

**RO2**: We aim to adapt counterfactual causal mediation methods to make them generally applicable to microbiome data (RO1. (RO2.1).

**RO3**: Part of the objective of WP1 is it to use machine learning methods for (1) improving the efficiency of the estimators (RO3.1+RO1.2), and (2) estimating the causal effects (RO3.2+RO2)

**Essence of intended scientific progress**: As in many empirical research disciplines, longitudinal studies are very common in microbiome research, but none of the many existing methods for longitudinal data analysis works well for microbiome data. There is thus a need for appropriate methods that are sufficiently flexible to be applicable to a wide variety of longitudinal study designs. If they work well and a convenient software implementation is available, they will be applied at large scale. Moreover, our envisaged methods will not only be applicable to microbiome data, but also to other data types that cause problems with existing methods.

Many longitudinal  microbiome studies with randomised treatment assignment aim at unravelling the causal effect of the treatment on an outcome, either a direct effect, an indirect effect via the microbiome, or a combination of both. Current mediation analysis methods, however, are inappropriate for the microbiome as high-dimensional mediator. We do not aim at contributing to the fundamental theory of causal inference, but rather to use state of the art mediation methods and adapt them to the very specific data characteristics of microbiome data by using the wide scope of data science methods (statistics, machine learning, visualisation) that is available at DSI.

Our approach to increase the efficiency of estimators (or power of hypothesis tests) is very novel and it can be ported to many other high dimensional statistical inference problems. When no distributional assumptions are made, efficiency can be gained by using the data itself for adapting the estimator to make it more efficient (adaptive estimators). Such methods usually rely on asymptotic theories and work only well on large datasets (large in the sense of number of observations). Our method improves the efficiency of an estimator related to one taxon by using the data from the other taxa. We anticipate that our approach will work well even when the number of observations for that one taxon is we use small. Although we will develop the method in the context of longitudinal data analysis, the general approach is more widely applicable to other high dimensional statistical inference problems.

# Research Methodology and Work Plan

*Describe how the research project will be executed and justify why this approach was chosen and why any specific strategic choices were made. Clarify how the scientific objectives will be reached, taking into account the proposed (sub)objectives, criteria and the capabilities of the partners.From the perspective of this general approach, describe the experimental set-up, the cohesion of the work packages as well as the milestones.Describe in the work plan: WHAT (divided into work packages), WHY and HOW (research approach, work method), WHEN (planning) and WHO (division of tasks, synergy and complementarity). Also state how interim decision points and general project risks are taken into account.Provide an overview of the deployment of the person power across the different work packages and its positioning in time for each of the work packages. Indicate whether it concerns scientific or technical personnel and their needed seniority. Personnel allocation should be given as estimated person-months (pm).Indicate how the coordination and monitoring of the project will be organized and how the cooperation among the various partners will be structured.*

**Overal Work Plan**

explain why we have 4 WPs

**Work package 1 (WP1): Development of flexible and robust statistical methods for analysing microbiome data from clustered and longitudinal studies and for causal statistical inference**

*Supervisors: Prof. O. Thas (DSI), Prof. D. Valkenborg (DSI) and Prof. J. Vangronsveld (CMK)*

**WP1.1 Semiparametric rank-based method for longitudinal analysis for a single taxon**

Starting from the observation that sequencing count data very often cannot be properly described by a (zero-inflated) negative binomial distribution (Hawinkel et al. 2020) and that almost all available methods for the analysis of longitudinal microbiome data rely on such distributional assumptions for the count data, we will develop a class of methods that does not rely on these assumptions. However, the methods must target informative parameters with unambiguous interpretation and they must allow for the flexible analysis of a wide variety of study designs, including one or more covariates or factors. For all these reasons, the method cannot be fully parametric and also not fully nonparametric. We therefore will develop a method based on a semiparamatric model. In particular, we **will build upon the Probabilistic Index Models (PIM)**(Thas et al. 2012), which form a class of methods that includes the classical Wilcoxon-Mann-Whitney test, among many other rank tests (De Neve and Thas 2015), but which also allows for inclusion of covariates and which provides parameters with clear and informative interpretations. PIMs have been used before for flexible and powerful analysis of data from high-throughput molecular technologies such as qPCR (De Neve et al. 2014)(De Neve et al. 2013)and scRNASeq (Assefa, Vandesompele, and Thas 2019), but these PIMs were not applicable to longitudinal or clustered data. In this project, we will use the PIM framework for modelling the count outcome of a single taxon (referred to as the **target taxon**). The method can of course be applied to all taxa separately as in conventional differential abundance testing.

In the PhD thesis of (Zhang 2012) a first attempt to extend PIMs to clustered data structures was explored. However, this path was not further elaborated because the proposed estimating equations observations only used pairs of observations within clusters and hence ignoring some information in the data. The model was based on the inclusion of random effects in an outcome model for the outcome Y, and integrating out the random effects distribution from the implied PIM. In this project we propose to walk along a different path. We will start from formulating two PIMs. The first PIM is the **within-subject model**,

where S, X and T are the subject, covariate and time of outcome Y, respectively (similarly for S\*, X\*, T\* and Y\*). The second PIM is the **between-subject model**,

where Z and Z\* are covariates corresponding to Y and Y\*, respectively. These covariates may overlap with X and X\*. For some applications, it may make sense to further restrict the second model to T=T\*.If we denote the outcome for a single taxon by  for subject i at time t, then parameter estimators will be defined as solutions to estimating equations of the form

where  is a matrix, which is referred to as the index matrix, that may depend on the covariates of observations (i,s) and (j,t), and  comes from PIM 1 or PIM 2, depending on whether i=j or not. The exact form of the matrix   will be inspired by (1) the index matrix of the ordinary PIM (Thas et al. 2012) and by (2) the working correlation matrices of GEE for longitudinal models. The variance-covariance matrix of the parameter estimator can be consistently estimated by a sandwich estimator. Asymptotic normality of the parameter estimators and consistency of the sandwich estimator will be demonstrated. The same theories as for ordinary PIMs can be used. In particular, as in (De Neve 2013), the theory of two-step estimators  (Newey and McFadden 1994) can be used (also referred to as V-estimators in the current context).

The theoretical properties of the estimators will be empirically evaluated in a simulation study and compared to competitor methods. For this purpose the flexible microbiome simulation method NPSimSeq (Assefa, Vandesompele, and Thas 2020) will be extended to allow for simulation of realistic longitudinal microbiome data.  NPSimSeq already includes a method for simulating correlated taxa, by combining estimating sparse  correlation networks (Kurtz et al. 2015)  and Gaussian copulas to backtransform to the correct marginal taxon-wise count distributions. This method can be extended to impose the serial correlations into all simulated taxon-wise longitudinal trajectories. However, we will make some working assumptions to make the method work when data with realistic sample sizes are available: (1) stationarity of the correlation network, and (2) taxon-wise serial dependence is modelled by an AR(1) process. This will allow for estimability of all parameters, and the Gaussian copula can again be used.

All methods will be implemented in R packages.

Application of method …

…

**WP1.2 Improving efficiency**

The estimation methods developed in WP1.1 are based on the naive estimation theory of (Thas et al. 2012), which does not result in (semiparametric) efficient estimators . For the simple i.i.d. case, (Vermeulen et al., n.d.) developed the semiparametric efficient estimator of the parameters in the PIM. However, their simulation study demonstrates that the efficiency gain is only marginal and does not outweigh the additional computation cost. In this project we take a different approach for improving estimation efficiency. We will take advantage of the high dimensionality of the microbiome data, which will allow us to construct data-adaptive estimators for which hardly a price has to be paid.

We will make use of the property that the distribution theory of (Thas et al. 2012) for the parameter estimator defined as the solution of equation ???, will remain valid as long as the index matrix ***A*** does not depend on the outcome data. In other words, a data-driven construction of the index matrix would complicate the distribution theory. However, in typical microbiome datasets, data on hundreds to thousands of taxa are available. Hence, under the (working) assumption that the data from the other taxa are independent of the data of the target taxon, we may construct the index matrix from data from the other taxa, without jeopardising the asymptotic distribution theory. We will develop an algorithm that aims at improving the efficiency of the parameter estimators by computing the index matrix  as the minimiser of some loss function computed over the the observed distributions of the other non-target taxa. In particular, let ${\cal{P}}=\{P\_j: \;\; j=1,\ldots, p\}$ denote the set of the distribution functions of the outcomes of the p other taxa (w.l.g. we ignore the conditioning on covariates) and suppose that we are interested in a parameter  that can be defined as a functional of Pj, i.e.  for some b(.). For each other taxon j the outcome distribution can be estimated, say , from which  serves as the true parameter. Repeatedly sampling from , say N times, gives for each repeated sample an estimate of  by solving the estimating equation ???. Based on these estimates and the true parameter , the MSE can be approximated, say . The performance of the index matrix ***A*** in estimating equation ??? can be quantified by means of a summary of the N MSEs, which is denoted by . We will use machine learning methods for finding the optimal matrix A in the sense of  . One particular example is . For this choice, will give a minimax estimator but with the restriction that the data-generating distribution is a member of ${\cal{P}}$. This shows resemblance with the procedure of (Luedtke et al. 2020), except that (1) we restrict the set of data-generating distributions to those observed in the other taxa, and (2) we consequently require other machine learning methods.

Although the new method is motivated by the longitudinal or clustered setting, we will also evaluate the method for the simpler situation of cross sectional studies studies with only a treatment indicator in the model (this is the traditional two-sample problem for testing for differential abundance). All evaluations will happen in simulations studies, using the SPSimSeq (Assefa, Vandesompele, and Thas 2020) simulation method (and its extension to longitudinal data, as developed in WP1.1).

All methods will be implemented in an R package.

Application …

**WP1.3: Mediation analysis**

Given that research on causal mediation analysis with high-dimensional mediators and for longitudinal studies has only started less than about five years ago, and is still a very active research topic, we anticipate that further improvements will have been published by the time this WP will commence. Developing new causal methodology, however, is no objective of this project; only the adaptation, evaluation and application of appropriate methods to realistic microbiome studies is part of this project.

From the current status of the literature, we anticipate that we will need to focus on interventional (in)direct effects, as in (Loh et al. 2020), for which the taxa do no have to be modelled jointly and for which a complete decomposition of the total effect may be feasible. This theory, however, does not encompass compositional mediators for which the dependence in the joint distribution is not structural (causal), but imposed by the compositionality. Log-ratio transformations could resolve this issue, but they do not deal properly with the many zeroes. We will therefore rely on a normalisation method that expresses the counts relative to the average count of a reference frame of taxa (Morton et al. 2019) that are stable between conditions (treatments) and over time. The selection of the taxa in the reference frame can proceed in a data-adaptive manner as in (Brill, Amir, and Heller 2019) without affecting the statistical inference. A further advantage of this approach is that it allows for inference on absolute abundances, under mild conditions. Our developments in this WP, will therefore also be applicable in the (near) future when new  technologies will allow for interrogating absolute abundances.

In this project we aim at estimating the outcome model  with machine learning methods. The flexibility of machine learning methods should reduce the risk of model bias. In high dimensional settings, however, regularisation is often required, which trades a reduction in variance for bias (estimation bias).  For this reason, we will adopt the double/debiased machine learning (DML) approach (Chernozhukov et al. 2018)(Farbmacher et al. 2020). The estimation of the (interventional) effects also involves models for the individual mediators (individual taxa). Instead of relying on resampling, as in (Loh et al. 2020), we will rely on the flexible semiparametric modelling framework of (Assefa, Vandesompele, and Thas 2020) for describing the mediator distributions so that parametric bootstrap samples can be generated from the fitted distributions. This will result in additional smoothing. Inference (e.g. hypothesis testing) will be based on the bootstrap procedure, as in many papers referred to in this proposal.

The procedure will eventually result in estimates for the indirect effects of all taxa. Depending on the taxonomic level that is targeted, the  number of simultaneous effect estimates and corresponding hypothesis tests may be large. We will examine different strategies, ranging from exploratory to formal hypothesis testing. For the latter, we will  aim at FDR-control using the taxonomic hierarchy information available in the phylogenetic tree, as in (Bogomolov et al. 2017) or (Barber and Ramdas 2017). As an exploratory tool, visualisations such as forest plots may be helpful. Forests plots are often used in meta-analysis for summarising effect size estimates from different studies (Lewis and Clarke 2001), but they could also be used for the visualisation of taxon-specific indirect effect estimates. We will further improve this visualisation by incorporating the phylogenetic tree structure.

The performance of the new methods will be assessed in a simulation study. As before, SPSimSeq will be used, but now the data simulation must proceed in two steps: (1) simulate all potential outcomes; (2) sample observed data from these simulated potential outcomes. See e.g. (Loh et al. 2019; Loh et al. 2020). If competitor methods are available, they will also be included in the simulation study.

Back-up plan in case no further methodological progress in mediation analysis: mediation analysis at each time point separately, or longitudinal mediation analysis after dimension reduction.

# References

Hawinkel, Stijn, JCW Rayner, Luc Bijnens, and Olivier Thas. 2020. “Sequence Count Data Are Poorly Fit by the Negative Binomial Distribution”. *PloS One* 15 (4): e0224909.

Hawinkel, Stijn, Federico Mattiello, Luc Bijnens, and Olivier Thas. 2019. “A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate”. *Briefings in Bioinformatics* 20 (1): 210–21.

Gloor, Gregory B, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. 2017. “Microbiome Datasets Are Compositional: and This Is Not Optional”. *Frontiers in Microbiology* 8: 2224.

Aitchison, John. 1982. “The Statistical Analysis of Compositional Data”. *Journal of the Royal Statistical Society: Series B (Methodological)* 44 (2): 139–60.

Schmidt, Thomas SB, Jeroen Raes, and Peer Bork. 2018. “The Human Gut Microbiome: from Association to Modulation”. *Cell* 172 (6): 1198–1215.

Mai, Volker, Mattia Prosperi, and Lusine Yaghjyan. 2016. “Moving Microbiota Research toward Establishing Causal Associations That Represent Viable Targets for Effective Public Health Interventions”. *Annals of Epidemiology* 26 (5): 306–10.

Chen, Eric Z, and Hongzhe Li. 2016. “A Two-Part Mixed-Effects Model for Analyzing Longitudinal Microbiome Compositional Data”. *Bioinformatics* 32 (17): 2611–17.

Ho, Nhan Thi, Fan Li, Shuang Wang, and Louise Kuhn. 2019. “MetamicrobiomeR: an R Package for Analysis of Microbiome Relative Abundance Data Using Zero-Inflated Beta GAMLSS and Meta-Analysis across Studies Using Random Effects Models”. *BMC Bioinformatics* 20 (1): 188.

Zhang, Xinyan, Yu-Fang Pei, Lei Zhang, Boyi Guo, Amanda H Pendegraft, Wenzhuo Zhuang, and Nengjun Yi. 2018. “Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data”. *Frontiers in Microbiology* 9: 1683.

Zhang, Xinyan, and Nengjun Yi. 2020. “Fast Zero-Inflated Negative Binomial Mixed Modeling Approach for Analyzing Longitudinal Metagenomics Data”. *Bioinformatics*.

Assefa, Alemu Takele, Jo Vandesompele, and Olivier Thas. 2020. “SPsimSeq: Semi-Parametric Simulation of Bulk and Single-Cell RNA-Sequencing Data”. *Bioinformatics* 36 (10): 3276–78.

Williams, Justin, Hector Corrada Bravo, Jennifer Tom, and Joseph Nathaniel Paulson. 2019. “MicrobiomeDASim: Simulating Longitudinal Differential Abundance for Microbiome Data”. *F1000Research* 8 (1769): 1769.

Xu, W, and B Chen. 2020. “Generalized Estimating Equation Modeling on Correlated Microbiome Sequencing Data with Longitudinal Measures”.

Wang, Chan, Jiyuan Hu, Martin J Blaser, and Huilin Li. 2020. “Microbial Trend Analysis for Common Dynamic Trend, Group Comparison and Classification in Longitudinal Microbiome Study”. *BioRxiv*.

Shields-Cutler, Robin R, Gabe A Al-Ghalith, Moran Yassour, and Dan Knights. 2018. “Splinectomer Enables Group Comparisons in Longitudinal Microbiome Studies”. *Frontiers in Microbiology* 9: 785.

Luo, Dan, Sara Ziebell, and Lingling An. 2017. “An Informative Approach on Differential Abundance Analysis for Time-Course Metagenomic Sequencing Data”. *Bioinformatics* 33 (9): 1286–92.

Li, Jun, and Robert Tibshirani. 2013. “Finding Consistent Patterns: a Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data”. *Statistical Methods in Medical Research* 22 (5): 519–36.

Thas, Olivier, Jan De Neve, Lieven Clement, and Jean-Pierre Ottoy. 2012. “Probabilistic Index Models”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (4): 623–71.

De Neve, Jan, and Olivier Thas. 2015. “A Regression Framework for Rank Tests Based on the Probabilistic Index Model”. *Journal of the American Statistical Association* 110 (511): 1276–83.

Amorim, Gustavo, Olivier Thas, Karel Vermeulen, Stijn Vansteelandt, and Jan De Neve. 2018. “Small Sample Inference for Probabilistic Index Models”. *Computational Statistics & Data Analysis* 121: 137–48.

Tsiatis, Anastasios. 2007. *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Vermeulen, Karel, Jan De Neve, Gustavo Amorim, Olivier Thas, and Stijn Vansteelandt. n.d. “Semiparametric Estimation Theory of Probabilistic Index Models: Efficiency and Bias”. *Scandinavian Journal of Statistics*.

Luedtke, Alex, Marco Carone, Noah Simon, and Oleg Sofrygin. 2020. “Learning to Learn from Data: Using Deep Adversarial Learning to Construct Optimal Statistical Procedures”. *Science Advances* 6 (9): eaaw2140.

Bickel, Peter J, Chris AJ Klaassen, Peter J Bickel, Ya’acov Ritov, J Klaassen, Jon A Wellner, and YA’Acov Ritov. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. Vol. 4. Johns Hopkins University Press Baltimore.

Van Der Laan, Mark J, and Daniel Rubin. 2006. “Targeted Maximum Likelihood Learning”. *The International Journal of Biostatistics* 2 (1).

Laan, Mark J Van der, and Sherri Rose. 2018. *Targeted Learning in Data Science*. Springer.

Gilbert, Jack A, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg, Janet K Jansson, Pieter C Dorrestein, and Rob Knight. 2016. “Microbiome-Wide Association Studies Link Dynamic Microbial Consortia to Disease”. *Nature* 535 (7610): 94–103.

Maruvada, Padma, Vanessa Leone, Lee M Kaplan, and Eugene B Chang. 2017. “The Human Microbiome and Obesity: Moving beyond Associations”. *Cell Host & Microbe* 22 (5): 589–99.

Judd, Charles M, and David A Kenny. 1981. “Process Analysis: Estimating Mediation in Treatment Evaluations”. *Evaluation Review* 5 (5): 602–19.

Baron, Reuben M, and David A Kenny. 1986. “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.”. *Journal of Personality and Social Psychology* 51 (6): 1173.

MacKinnon, David. 2012. *Introduction to Statistical Mediation Analysis*. Routledge.

Robins, James M, and Sander Greenland. 1992. “Identifiability and Exchangeability for Direct and Indirect Effects”. *Epidemiology*, 143–55.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

Vanderweele, Tyler J, and Stijn Vansteelandt. 2009. “Conceptual Issues Concerning Mediation, Interventions and Composition”. *Statistics and Its Interface* 2 (4): 457–68.

Pearl, Judea. 2012. “The Causal Mediation Formula—a Guide to the Assessment of Pathways and Mechanisms”. *Prevention Science* 13 (4): 426–36.

Lange, Theis, Stijn Vansteelandt, and Maarten Bekaert. 2012. “A Simple Unified Approach for Estimating Natural Direct and Indirect Effects”. *American Journal of Epidemiology* 176 (3): 190–95.

Vansteelandt, Stijn, Maarten Bekaert, and Theis Lange. 2012. “Imputation Strategies for the Estimation of Natural Direct and Indirect Effects”. *Epidemiologic Methods* 1 (1): 131–58.

Díaz, Iván. 2020. “Machine Learning in the Estimation of Causal Effects: Targeted Minimum Loss-Based Estimation and Double/Debiased Machine Learning”. *Biostatistics* 21 (2): 353–58.

VanderWeele, Tyler, and Stijn Vansteelandt. 2014. “Mediation Analysis with Multiple Mediators”. *Epidemiologic Methods* 2 (1): 95–115.

Daniel, RM, BL De Stavola, SN Cousens, and Stijn Vansteelandt. 2015. “Causal Mediation Analysis with Multiple Mediators”. *Biometrics* 71 (1): 1–14.

Steen, Johan, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. 2017. “Flexible Mediation Analysis with Multiple Mediators”. *American Journal of Epidemiology* 186 (2): 184–93.

Vansteelandt, Stijn, and Rhian M Daniel. 2017. “Interventional Effects for Mediation Analysis with Multiple Mediators”. *Epidemiology (Cambridge, Mass.)* 28 (2): 258.

VanderWeele, Tyler J, Stijn Vansteelandt, and James M Robins. 2014. “Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder”. *Epidemiology (Cambridge, Mass.)* 25 (2): 300.

Loh, Wen Wei, Beatrijs Moerkerke, Tom Loeys, and Stijn Vansteelandt. 2019. “Interventional Effect Models for Multiple Mediators”. *ArXiv Preprint ArXiv:1907.08415*.

———. 2020. “Non-Linear Mediation Analysis with High-Dimensional Mediators Whose Causal Structure Is Unknown”. *ArXiv Preprint ArXiv:2001.07147*.

Bind, M-AC, TJ Vanderweele, BA Coull, and JD Schwartz. 2016. “Causal Mediation Analysis for Longitudinal Data with Exogenous Exposure”. *Biostatistics* 17 (1): 122–34.

VanderWeele, Tyler J, and Eric J Tchetgen Tchetgen. 2017. “Mediation Analysis with Time Varying Exposures and Mediators”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (3): 917–38.

Zheng, Wenjing, and Mark J van der Laan. 2018. “Mediation Analysis with Time-Varying Mediators and Exposures”. In *Targeted Learning in Data Science*, 277–99. Springer.

Mittinty, Murthy N, and Stijn Vansteelandt. 2019. “Longitudinal Mediation Analysis Using Natural Effect Models”. *ArXiv Preprint ArXiv:1912.01200*.

Zhang, Haixiang, Jun Chen, Zhigang Li, and Lei Liu. 2019. “Testing for Mediation Effect with Application to Human Microbiome Data”. *Statistics in Biosciences*, 1–16.

Zhang, Cun-Hui, and Stephanie S Zhang. 2014. “Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 217–42.

Sohn, Michael B, Hongzhe Li, and others. 2019. “Compositional Mediation Analysis for Microbiome Studies”. *The Annals of Applied Statistics* 13 (1): 661–81.

Shi, Pixu, Anru Zhang, Hongzhe Li, and others. 2016. “Regression Analysis for Microbiome Compositional Data”. *The Annals of Applied Statistics* 10 (2): 1019–40.

Wang, Chan, Jiyuan Hu, Martin J Blaser, and Huilin Li. 2020. “Estimating and Testing the Microbial Causal Mediation Effect with High-Dimensional and Compositional Microbiome Data”. *Bioinformatics* 36 (2): 347–55.

Radchenko, Peter, and Gareth M James. 2010. “Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions”. *Journal of the American Statistical Association* 105 (492): 1541–53.

Li, Zhigang, Janaka SS Liyanage, A James O’Malley, Susmita Datta, Raad Z Gharaibeh, Christian Jobin, Modupe O Coker, et al. 2019. “Mediation Analysis for Zero-Inflated Mediators with Applications to Microbiome Data”. *ArXiv Preprint ArXiv:1906.09175*.

Zhang, Jie, Zhi Wei, and Jun Chen. 2018. “A Distance-Based Approach for Testing the Mediation Effect of the Human Microbiome”. *Bioinformatics* 34 (11): 1875–83.

De Neve, Jan, Joris Meys, Jean-Pierre Ottoy, Lieven Clement, and Olivier Thas. 2014. “UnifiedWMWqPCR: the Unified Wilcoxon-Mann-Whitney Test for Analyzing RT-QPCR Data in R.”. *Bioinformatics* 30 (17): 2494–95.

De Neve, Jan, Olivier Thas, Jean-Pierre Ottoy, and Lieven Clement. 2013. “An Extension of the Wilcoxon-Mann-Whitney Test for Analyzing RT-QPCR Data”. *Statistical Applications in Genetics and Molecular Biology* 12 (3): 333–46.

Assefa, Alemu Takele, Jo Vandesompele, and Olivier Thas. 2019. “Probabilistic Index Models for Testing Differential Expression in Single Cell RNA Sequencing Data”. *BioRxiv*, 718668.

Zhang, Fanghong. 2012. “From Parametric towards Nonparametric Mixed Modeling of Correlated Data”. PhD thesis, Ghent University.

De Neve, Jan. 2013. “Probabilistic Index Models”. PhD thesis, Ghent University.

Newey, KW, and D McFadden. 1994. “Large Sample Estimation and Hypothesis”. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. 2015. “Sparse and Compositionally Robust Inference of Microbial Ecological Networks”. *PLoS Computational Biology* 11 (5).

Vermeulen, Karel, Gustavo Amorim, Jan De Neve, Olivier Thas, and Stijn Vansteelandt. n.d. “Semiparametric Estimation of Probabilistic Index Models: Efficiency and Bias”. *Under Review*.

Morton, James T, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. 2019. “Establishing Microbial Composition Measurement Standards with Reference Frames”. *Nature Communications* 10 (1): 1–11.

Brill, Barak, Amnon Amir, and Ruth Heller. 2019. “Testing for Differential Abundance in Compositional Counts Data, with Application to Microbiome Studies”. *ArXiv Preprint ArXiv:1904.08937*.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. “Double/Debiased Machine Learning for Treatment and Structural Parameters”. Oxford University Press Oxford, UK.

Farbmacher, Helmut, Martin Huber, Henrika Langen, and Martin Spindler. 2020. “Causal Mediation Analysis with Double Machine Learning”. *ArXiv Preprint ArXiv:2002.12710*.

Bogomolov, Marina, Christine B Peterson, Yoav Benjamini, and Chiara Sabatti. 2017. “Testing Hypotheses on a Tree: New Error Rates and Controlling Strategies”. *ArXiv Preprint ArXiv:1705.07529*.

Barber, Rina Foygel, and Aaditya Ramdas. 2017. “The p-Filter: Multilayer False Discovery Rate Control for Grouped Hypotheses”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (4): 1247–68.

Lewis, Steff, and Mike Clarke. 2001. “Forest Plots: Trying to See the Wood and the Trees”. *Bmj* 322 (7300): 1479–80.