MedZIM: Mediation analysis for Zero-Inflated Mediators with applications to microbiome data

Zhigang Li*†[1], Janaka S.S. Liyanage*[1], A. James O'Malley[2], Susmita Datta[1], Raad Z. Gharaibeh[3], Christian Jobin[3], Modupe O. Coker[4], Anne G. Hoen[4], Brock C. Christensen[4], Juliette C. Madan[4] and Margaret R. Karagas[4]

[1]Department of Biostatistics, University of Florida, Gainesville, FL, USA, 32611
[2]The Dartmouth Institute, Geisel School of Medicine at Dartmouth, Hanover, NH, USA, 03755
[3]Department of Medicine, University of Florida, Gainesville, FL, USA, 32611
[4]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA, 03755

May 21, 2020

Abstract

The human microbiome can contribute to the pathogenesis of many complex diseases such as cancer and Alzheimer's disease by mediating disease-leading causal pathways. However, standard mediation analysis is not adequate in the context of microbiome data due to the excessive number of zero values in the data. Zero-valued sequencing reads, commonly observed in microbiome studies, arise for technical and/or biological reasons. Mediation analysis approaches for analyzing zero-inflated mediators are still lacking largely because of challenges raised by the zero-inflated data structure: (a) disentangling the mediation effect induced by the point mass at zero; and (b) identifying the observed zero-valued data points that are actually not zero (i.e., false zeros). We develop a novel mediation analysis method under the potential-outcomes framework to fill this gap. We show that the mediation effect of the microbiome can be decomposed into two components that are inherent to the two-part nature of zero-inflated distributions. The first component corresponds to the mediation effect attributable to a unit-change over the positive relative abundance and the second component corresponds to the mediation effect attributable to discrete binary change of the mediator from zero to a non-zero state. With probabilistic models to account for observing zeros, we also address the challenge

*Equal contribution
†Corresponding email: zhigang.li@ufl.edu.

with false zeros. A comprehensive simulation study and the applications in two real microbiome studies demonstrate that our approach outperforms existing mediation analysis approaches.

1 Introduction

Emerging evidence suggest that the human microbiome and the immune system are constantly shaping each other [4]. Thus the human microbiome can contribute to disease pathogeneses by mediating disease-leading causal pathways in complex diseases such as Alzheimer's disease [46] and cancer [21, 36]. To quantitatively study human microbiome, 16S ribosomal RNA gene sequencing and metagenomic shotgun sequencing have been popular methods to quantify microbiome composition in microbiome studies. An important feature of microbiome sequencing data is that it has excessive number of zeros [23]. Many microbiome data sets have more than 50% datapoints equal to 0, and it could be as high as 80% or more. These zeros are likely to be a mixture of structural zeros (i.e., true zeros) that represent true absence of microbial taxa and undersampling zeros (i.e., false zeros) that result from failure of detection. The zero-inflated data feature poses a challenge that needs to be addressed specifically in mediation analysis. Although there have been some exciting efforts to model microbiome as a high-dimensional mediator [34, 44, 48, 49], it remains a daunting task to address the zero-inflated data structure in mediation analyses.

Mediation analysis is an important tool to investigate the role of intermediate variables (i.e., mediators) in a causal pathway where the causal effect partially or completely relies on the mediators. For example, people with higher socioeconomic status tend to have longer life expectancy, but this causal pathway may be explained by many possible mediators including access to better health care, fewer stressors, better living environment and so forth. In a mediation analysis, the indirect effect (i.e., mediation effect) through one or more mediators can be estimated and tested along with the direct effect. This technique was first popularized in psychology and social sciences where traditional linear mediation analysis was proposed, and now is a common tool in many research areas such as epidemiology, environmental health sciences, medicine, randomized trials and psychiatry. Mediation analysis is a powerful approach largely because of its capability to estimate and test the mediation effect that cannot be achieved with typical regression approaches. There are two general types of mediation analysis approaches: potential-outcomes (PO) or counterfactual-outcomes methods [19, 40, 41] and traditional linear mediation analysis methods [3, 26]. The former approach stems from a counterfactual nonparametric function of a causal relationship without relying on linear assumptions and the latter is based on linear regression models. These approaches coincide with each other under linearity assumptions. PO approaches are more flexible because they can allow interaction effects of the exposure/treatment variable (referred to as exposure variable hereafter) with mediators as well as nonlinear effects to be modeled. It is worth noting that assumptions on unmeasured and measured confounders are required to draw causal inference regardless of which type of mediation analysis approach is used. Reviews of mediation analysis approaches and their assumptions can be found in the literature [22, 27, 42].

Although mediation modeling frameworks have been well established, to the best of our knowledge, there have been few studies which address the zero-inflated distributions for mediators. In a typical mediation analysis, the total effect of an exposure variable

can be decomposed into a mediation effect and a direct effect where the mediation effect measures the amount of the total causal effect attributable to change in the mediator caused by the exposure variable and the direct effect measures the causal effect due to change in the exposure variable while keeping the mediator variable constant. When the mediator has a zero-inflated distribution such as a zero-inflated Beta (ZIB) distribution or any other zero-inflated distributions, we show that its mediation effect can be further decomposed into two parts with one part being the mediation effect attributable to the amount of numeric change in the mediator and the other part being the mediation effect attributable to the binary change of the mediator from zero to a non-zero state. This phenomenon can be explained by the two-part nature of a zero-inflated distribution. For example, a ZIB distribution is essentially a two-component mixture distribution [13]: one component is a degenerate distribution with probability mass of one at zero, and the other component is a Beta distribution. The mediator changing from zero to a positive value results in the discrete jump from zero to a non-zero state as well as the change in the numerical metric of the mediator and thus the mediation effect can be divided accordingly. Both changes have important consequences as we can see in the real study examples later where the absence of a microbial taxon and the abundance level of a microbial taxon given its presence are considered. What makes it more complicated is that the observed zero-valued data points could be false zeros meaning that the true values are non-zero but observed as zero due to failure of detection. This is similar to a missing data problem and will be addressed here as well.

To fill the research gap in mediation modeling development, we propose a novel mediation analysis approach under the PO framework to model mediators with zero-inflated distributions. This approach can allow a mixture of truly zero-valued datapoints and false zeros. Our method is able to decompose the mediation effect into two components that are inherent to zero-inflated mediators: one component is the mediation effect attributable to the numeric change of the mediator on its continuum scale and the other component is the mediation effect attributable to the binary change of the mediator from zero to a non-zero state. So the mediation effect is actually the total mediation effect of the two components each of which can be estimated and tested. Although we focused on ZIB distribution for microbial taxa, our approach has a general framework to accommodate many zero-inflated distributions. A simulation study is conducted to evaluate our approach in comparison with a standard PO mediation analysis approach and two other existing approaches that can analyze microbiome composition as a mediator.

We introduce the model and its associated notations in Section 2. Estimation and inference procedures are provided in Section 3. A simulation study to assess the performance of our model in comparison with existing approaches is presented in Section 4, followed by an application of our model in two real studies in Section 5, and a discussion in Section 6. Additional details and derivations can be found in the Appendix.

2 Model and Notation

For simplicity, we suppress subject index in all notations in this section. Let Y, M and X denote the continuous outcome variable, the mediator variable and the independent variable respectively. We assume the outcome Y depend on M and X through the following regression equation:

$$Y = \beta_0 + \beta_1 M + \beta_2 1_{(M>0)} + \beta_3 X + \beta_4 X 1_{(M>0)} + \beta_5 XM + \epsilon \tag{1}$$

where $1_{(\cdot)}$ is an indicator function, the random error ϵ follows a normal distribution $N(0, \delta)$, and β_1, β_2, β_3, β_4 and β_5 are regression coefficients. The mediator M represents relative abundance (RA) of a microbial taxon for microbiome data. The advantage of using M instead of $\log(M)$, which has been used the literature [34, 44], is that we do not need to impute zero values of M with an positive number.

In order to construct a mediation model, we also need to model the association between M and X. For the zero-inflated mediator M, we write its distribution into a mixture distribution with the two-part density function given by:

$$f(m;\theta) = \begin{cases} \mathcal{G}(\theta), & m = 0 \\ (1-\mathcal{G}(\theta))G(m;\theta), & m > 0 \end{cases}$$

where θ is the K-dimensional parameter vector associated with the zero-inflated distribution, $\mathcal{G}(\cdot)$ is a $R^K \to R$ mapping with $0 < \mathcal{G}(\theta) < 1$ being the probability of M taking the value of 0 and $G(m;\theta)$ is the conditional probability density (or mass) function of M given that M is positive. We use a ZIB distribution [10, 11] for M when modeling relative abundance (RA) of microbial taxa as the mediator. This two-part model is fairly general and can model other parametric zero-inflated distributions as well, such as zero-inflated log-normal distribution and zero-inflated Poisson distribution. With this formulation, we can model the association beween M and X by assuming the parameter θ depends on X through the following equation:

$$T(\theta) = \nu_0 + \nu_1 X \tag{2}$$

where $T : R^K \to R^K$ is a known one-to-one (possibly nonlinear) transformation of the parameter vector θ, and ν_0 and ν_1 are two K-dimensional parameters where ν_0 can be interpreted as the intercept vector and ν_1 as the slope vector.

Equations (1) and (2) together form our full mediation model. Notice that X is a scalar here, but it is obvious that other covariates such as potential confounders can be included in the two equations. This model is fully compatible with allowing interactions between the exposure variable and mediators as the two interaction terms: $X1_{(M>0)}$ and XM are included in equation (1). In practice, investigators can also include only one or no interaction term depending on the hypothesis of interest. Notice that it will be challenging for traditional linear mediation methods to estimate the mediation effect here because of: 1) the interaction terms: $\beta_4 X1_{(M>0)}$ and $\beta_5 XM$; 2) the term $\beta_2 1_{(M>0)}$ in equation (1) which is a nonlinear function of the mediator M.

More specifically for a ZIB mediator, the two-part density function can be written as:

$$f(m;\theta) = \begin{cases} \mathcal{G}(\theta), & m = 0 \\ (1-\mathcal{G}(\theta))\frac{m^{\mu\phi-1}(1-m)^{(1-\mu)\phi-1}}{B\left(\mu\phi,(1-\mu)\phi\right)}, & m > 0 \end{cases}$$

where $\theta = (\mu, \phi, \Delta)^T$, $\mathcal{G}(\theta) = \Delta$, $B(\cdot,\cdot)$ is the Beta function and μ and ϕ are the mean and dispersion parameters respectively of the Beta distribution for the non-zero part [12, 16]. To model the association of the mediator with X, we use $T(\theta) = (\log(\mu/(1-\mu)), \log(\phi), \log(\Delta/(1-\Delta))^T$, $\nu_0 = (\alpha_0, \xi, \gamma_0)^T$ and $\nu_1 = (\alpha_1, 0, \gamma_1)^T$ and thus equation (2) becomes the following three equations:

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha_0 + \alpha_1 X, \tag{3}$$

$$\log(\phi) = \xi, \tag{4}$$

$$\log\left(\frac{\Delta}{1-\Delta}\right) = \gamma_0 + \gamma_1 X, \tag{5}$$

where the dispersion parameter ϕ is assumed not to depend on X which is why the right-hand side of equation (4) does not involve X. More details on formulation of the model are provided in the Appendix.

2.1 Mechanism for observing zeros of the mediator

For microbiome abundance data, observations below the limit of detection are set to be zero. Consequently, there are two types of zeros in the observed abundance data: true abundance of zero (i.e., absence) and abundance that is reported as zero as a consequence of the measurement procedure. We will use real microbiome studies to illustrate our method in a later section. Let M be the relative abundance of a microbial taxon and let M^* denote the observed value of M. When the observed value is positive (i.e., $M^* > 0$), we assume that $M^* = M$. But when $M^* = 0$, we don't know whether M is truly zero or M is positive but observed as zero. We consider the following mechanism for observing a zero of the microbial taxon abundance:

$$P(M^* = 0|M, L) = 1_{(ML<1)}, \tag{6}$$

where L is the library size (i.e., sequencing depth) and the product ML can be interpreted as the sample absolute abundance (SAA) of the taxon in a sample. Under this mechanism, all SAA below 1 have an observed value of zero. Here 1 can be considered as the Limit of Detection (LOD). We refer to this mechanism as "LOD mechanism" hereafter. This mechanism assumes that the probability of observing a zero only depends on SAA which implies that it is independent of the outcome Y conditional on SAA. Since SAA depends on both L and M, the LOD mechanism is not deterministic conditional on the library size. The probability of observing a zero conditional on L, the library size, is equal to $E(1_{(ML<1)}|L) = P(M < 1/L)$.

2.2 Mediation effect and direct effect

Under the potential-outcomes (PO) framework [42], we can define the natural indirect effect (NIE), natural direct effects (NDE) and controlled direct effect (CDE) where NIE is the mediation effect. The total effect of the exposure variable X is equal to the summation of NIE and NDE. Let M_x denote the value of M if X equals x. Let Y_{xm} denote the value of Y if $(X, M) = (x, m)$. The average NIE, NDE and CDE for X changing from x_1 to x_2 are defined as:

NIE $= E\big(Y_{x_2 M_{x_2}} - Y_{x_2 M_{x_1}}\big)$

NDE $= E\big(Y_{x_2 M_{x_1}} - Y_{x_1 M_{x_1}}\big)$

CDE $= E\big(Y_{x_2 m} - Y_{x_1 m}\big)$, for a fixed (i.e., controlled) value of M,

where $Y_{x_2 M_{x_1}}$ is a counterfactual outcome. By plugging the equations (1)-(2) into the above definitions and using Riemann-Stieljes integration [37], we can obtain the following formulas:

$$\begin{aligned}\text{NIE} &= E(Y_{x_2 M_{x_2}}) - E(Y_{x_2 M_{x_1}}) = E(E(Y_{x_2 M_{x_2}}|M_{x_2})) - E(E(Y_{x_2 M_{x_1}}|M_{x_1})) \\ &= E(\beta_0 + \beta_1 M_{x_2} + \beta_2 1_{(M_{x_2}>0)} + \beta_3 x_2 + \beta_4 x_2 1_{(M_{x_2}>0)} + \beta_5 x_2 M_{x_2})\end{aligned}$$

$$
\begin{aligned}
&\quad - E(\beta_0 + \beta_1 M_{x_1} + \beta_2 1_{(M_{x_1}>0)} + \beta_3 x_2 + \beta_4 x_2 1_{(M_{x_1}>0)} + \beta_5 x_2 M_{x_1}) \\
&= (\beta_1 + \beta_5 x_2)(E(M_{x_2}) - E(M_{x_1})) + (\beta_2 + \beta_4 x_2)(E(1_{(M_{x_2}>0)}) - E(1_{(M_{x_1}>0)})) \\
&= \mathrm{NIE}_1 + \mathrm{NIE}_2, \\
\mathrm{NIE}_1 &= (\beta_1 + \beta_5 x_2)(E(M_{x_2}) - E(M_{x_1})) \\
&= (\beta_1 + \beta_5 x_2)\left(\int_{m\in\Omega} m dF_{M_{x_2}}(m) - \int_{m\in\Omega} m dF_{M_{x_1}}(m) \right) \\
&= (\beta_1 + \beta_5 x_2)(1 - \mathcal{G}(\theta_{x_2})) \int_{m\in\Omega\setminus 0} m G(m;\theta_{x_2}) dm \\
&\quad - (\beta_1 + \beta_5 x_2)(1 - \mathcal{G}(\theta_{x_1})) \int_{m\in\Omega\setminus 0} m G(m;\theta_{x_1}) dm), \\
\mathrm{NIE}_2 &= (\beta_2 + \beta_4 x_2)(\mathcal{G}(\theta_{x_1}) - \mathcal{G}(\theta_{x_2})), \\
\mathrm{NDE} &= E\big(Y_{x_2 M_{x_1}} - Y_{x_1 M_{x_1}}\big) = E\big(Y_{x_2 M_{x_1}}\big) - E\big(Y_{x_1 M_{x_1}}\big) \\
&= \beta_3(x_2 - x_1) + \beta_4(x_2 - x_1)(1 - \mathcal{G}(\theta_{x_1})) + \beta_5(x_2 - x_1)E(M_{x_1}) \\
&= \beta_3(x_2 - x_1) + \beta_4(x_2 - x_1)(1 - \mathcal{G}(\theta_{x_1})) + \beta_5(x_2 - x_1) \int_{m\in\Omega} m dF_{M_{x_1}}(m) \\
&= \beta_3(x_2 - x_1) + \beta_4(x_2 - x_1)(1 - \mathcal{G}(\theta_{x_1})) \\
&\quad + \beta_5(x_2 - x_1)(1 - \mathcal{G}(\theta_{x_1})) \int_{m\in\Omega\setminus 0} m G(m;\theta_{x_1}) dm \\
\mathrm{CDE} &= \beta_3(x_2 - x_1) + \beta_4(x_2 - x_1)1_{(m>0)} + \beta_5 m(x_2 - x_1),
\end{aligned}
$$

where Ω denotes the domain of the mediator M, M_x denotes the value of M conditional on $X = x$, $F_{M_x}(m)$ denotes the CDF of M_x, $dF_{M_x}(m)$ denotes the stieltjes integration [37] with respect to $F_{M_x}(m)$, $\theta_x = T^{-1}(\nu_0 + \nu_1 x)$, $\Omega \setminus 0$ denotes the subset of Ω that does not contain 0, and $T^{-1}(\cdot)$ denotes the inverse function of $T(\cdot)$. NIE, NIE_1, NIE_2, NDE and CDE can be estimated by plugging the parameter estimates into the formulas. Confidence intervals (CI) can be obtained using the delta method. An alternative approach for finding standard errors to construct CI is bootstrapping [14]. More details of the calculation for mediation effect and direct effect and their CI for ZIB mediators can be found in the Appendix. NIE_1 can be interpreted as the mediation effect due to the change of the mediator on its numeric scale and NIE_2 can be interpreted as the mediation effect due to the discrete binary change of the mediator from zero to a non-zero status. This decomposition can be also seen in Figure 1 where there are two possible indirect causal pathways from X to Y through the mediator M. When the distribution of mediator is not zero-inflated, NIE_2 becomes 0 since $\mathcal{G}(\theta_x)$ reduces to 0, and thus the NIE reduces to a usual NIE that can be calculated by standard approaches [19, 40].

3 Parameter Estimation

Maximum likelihood estimation (MLE) will be used to estimate the parameters. The variables of observed data for each subject can be denoted by the vector (Y, R, M^*, L, X) where $R = 1_{(M^*>0)}$ and the subject index is suppressed. We are not considering other

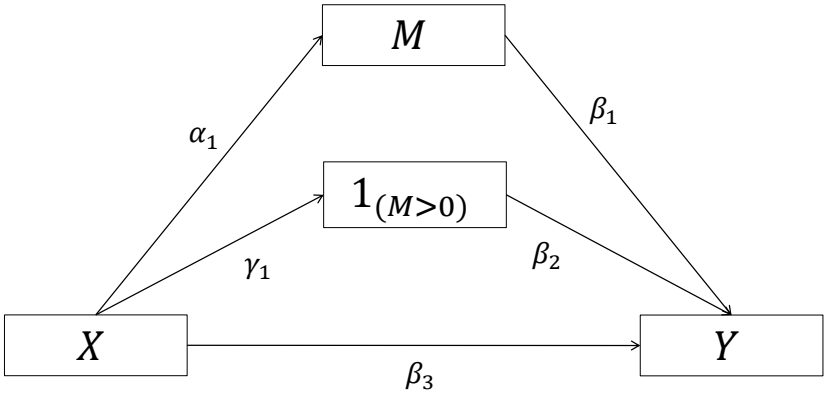

Figure 1: Potential causal mediation pathways of a zero-inflated mediator.

covariates here, but the method can be easily extended to include more covariates in the equations (1)-(2). The estimation challenge is that M is not always observable due to false zeros. The log-likelihood contribution from those subjects with false zeros cannot be directly calculated. However, given that we know the probability of observing a zero in equation (6), we can still obtain their log-likelihood contributions by integrating the joint density function over all possible values of M using Riemann–Stieltjes integration [37]. Let $(y_i, r_i, m_i^*, l_i, x_i)$ denote the observed data values for the ith subject in the study and m_i denote the true value of the mediator. We use i as subject index hereafter throughout the paper. The subjects are divided into two groups by whether M^* is non-zero. The first group consists of subjects whose observed value of m_i is non-zero (i.e., $m_i^* \neq 0$). In this group we have $m_i^* = m_i$ and the log-likelihood contribution from the ith subject can be calculated as:

$$\begin{aligned}
\ell_i^1 &= \log(f(y_i, r_i|m_i^*, x_i, l_i)f(m_i^*|x_i, l_i)) = \log(f(y_i|m_i^*, x_i, l_i)f(r_i|m_i^*, x_i, l_i)f(m_i^*|x_i, l_i)) \\
&= \log(f(y_i|m_i^*, x_i, l_i)) + \log(f(r_i|m_i^*, l_i)) + \log(f(m_i^*|x_i, l_i)) \\
&= -0.5\log(2\pi) - \log(\delta) - \frac{\left(y_i - \beta_0 - \beta_1 m_i^* - \beta_2 - (\beta_3 + \beta_4)x_i - \beta_5 x_i m_i^*\right)^2}{2\delta^2} \\
&\quad + \log(1 - P(M_i^* = 0|M_i = m_i^*, l_i)) + \log((1 - \mathcal{G}(\theta_{x_i}))G(m_i^*; \theta_{x_i})),
\end{aligned}$$

where $f(\cdot|m_i^*, x_i, l_i)$, $f(\cdot|m_i^*, x_i, l_i)$ and $f(\cdot|x_i, l_i)$ are the (conditional) density (or probability mass function) for Y, R and M respectively. Let $F(m|x)$ denote the (conditional) cumulative distribution function for M. The second group consists of subjects whose observed m_i are 0 (ie, $m_i^* = 0$) and their log-likelihood contribution can be calculated as:

$$\ell_i^2 = \log(f(y_i, r_i, m_i^*|x_i)) = \log\left(\int_{m\in\Omega} f(y_i|m, x_i)f(r_i|m)dF(m|x_i)\right)$$

$$= \log\left(\mathcal{G}(\theta_{x_i})\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(-\frac{\left(y_i-\beta_0-\beta_3 x_i\right)^2}{2\delta^2}\right)+\right.$$
$$\left.(1-\mathcal{G}(\theta_{x_i}))\int\limits_{m\in\Omega\setminus 0} f(y_i|m,x_i)P(M_i^*=0|M_i=m))G(m;\theta_{x_i})dm\right),$$

where

$$f(y_i|m,x_i)=\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(-\frac{\left(y_i-\beta_0-\beta_1 m-\beta_2-(\beta_3+\beta_4)x_i-\beta_5 x_i m\right)^2}{2\delta^2}\right).$$

Taken together, we have the complete log-likelihood function given by

$$\ell=\sum_{i\in\text{group 1}}\ell_i^1+\sum_{i\in\text{group 2}}\ell_i^2. \tag{7}$$

There is no general closed-form expression for ℓ due to the integration. More details of calculating ℓ for ZIB mediators are given in the Appendix. We obtain the MLE of the parameters by maximizing the above complete log-likelihood function. With the parameter estimates, we will be able to calculate NIE, NIE_1, NIE_2, NDE and CDE and their CI.

4 Simulation

Extensive simulations were carried out to demonstrate the performance of MedZIM in comparison with three existing approaches. First, we compared MedZIM with a current standard practice in causal mediation analyses developed by Imai, Keele and Tingley [19] (IKT approach hereafter) which is a PO approach and can be implemented in R using the package "mediation" [39]. The Marginal Structural Models developed in VanderWeele [40] is also a PO approach with a very similar definition of indirect effect. These standard causal mediation analysis approaches were not developed to analyze microbiome data, and thus could have poor performance when applied to microbiome data. Second, MedZIM was compared with two existing approaches: CCMM [34] and SparseMCMM [44] that were developed specifically to model microbiome composition as a mediator. In all simulation settings, the independent variable X was binary and generated using the Bernoulli distribution Ber(0.5) such that the number of subjects was balanced between the two groups. The LOD mechanism in equation (6) for observing zero-valued data points of the mediator was used to generate zeros for the mediator M. To mimic the real study data, the library size was generated by randomly picking the library size with replacement from a real study [2] where the library size ranges from 31,607 to 911,652. The RA data was generated in a way such that it mimicked the distribution of RA in real data. We generated 100 random datasets for each of the simulation settings.

4.1 Comparison with the IKT approach

In this comparison, the outcome Y was assumed to be a continuous variable and generated using equation (1) where β_5 is set to be 0 in the simulation and other true parameter values can be found in Table 1. Similar to simulation studies in the literature [10, 11]

where RA were generated individually, we generated an individual taxon RA with ZIB distributions based on equation (1) and equations (3-5). The sample size was 100 in each of the 100 random datasets. Two scenarios were considered for the taxon RA: low RA (Scenario 1: mean of positive RA is equal to 0.0025) and high RA (Scenario 2: mean of positive RA is equal to 0.5). About 20% of all sequencing reads were generated as true zeros (i.e., structured zeros) in both scenarios. Under the LOD mechanism in equation (6), about 30% sequencing reads were false zeros in Scenario 1 and there were no false zeros in Scenario 2 because the RA in Scenario 2 was high and thus SAA were greater than 1 for all truly non-zero RA. Model performance was evaluated by estimation bias, standard error, coverage probability (CP) of 95% CI of the estimators for parameters and the mediation effects in this comparison. For Scenario 1, the simulation results (Table 1) showed good performance for MedZIM in terms of bias and CP of the mediation effects and the parameter estimates. All the biases were small and the CP were around the desired level of 95%. The IKT approach, however, had a poor performance with a large bias (84.81%) and a small CP (9%). These poor performances were likely due to the false zeros not being appropriately accounted for by the IKT approach. Another disadvantage of IKT is that it cannot decompose the mediation effect into NIE1 and NIE2. For Scenario 2 with high RA where there were no false zeros, MedZIM showed good performance again in terms of the performance measures. IKT also showed satisfactory performance for the estimation of the NIE which is because there were no false zeros in the data under this scenario, but IKT cannot decompose the mediation effect according to the zero-inflated distribution of mediator.

Table 1: Simulation results for comparison between MedZIM and IKT with sample size of $n = 100$. Bias, percentage of the bias, the empirical standard errors, the the mean of estimated standard errors and the empirical coverage probability of the 95% CI for each estimator is respectively reported under the columns Bias, Bias %, SE, Mean SE and CP(%). Mediation effects from the IKT approach are provided at the bottom part of the table.

	Low relative abundance (mean=0.0025)							High relative abundance (mean=0.5)						
Parameter /Effect	True	Mean Estimate	Bias	Bias %	SE	Mean SE	CP(%)	True	Mean Estimate	Bias	Bias %	SE	Mean SE	CP(%)
							MedZIM							
NIE1	0.10	0.11	0.01	10.0	0.08	0.07	91	9.30	9.11	-0.18	-1.98	2.68	2.70	96
NIE2	0.55	0.52	-0.03	-5.67	0.55	0.56	97	0.55	0.50	-0.06	-10.15	0.62	0.56	94
NIE	0.65	0.63	-0.02	-3.31	0.58	0.58	96	9.85	9.61	-0.24	-2.44	3.25	3.20	95
β_0	-2.00	-2.05	-0.05	-2.45	0.32	0.33	96	-2.00	-1.92	0.07	3.82	0.32	0.29	94
β_1	100.00	101.89	1.89	1.89	18.04	19.04	97	100.00	99.96	-0.04	-0.04	1.89	1.74	91
β_2	4.00	4.05	0.05	1.37	0.38	0.36	94	4.00	3.93	-0.07	-1.73	0.58	0.57	91
β_3	5.00	5.08	0.08	1.53	0.53	0.51	94	5.00	4.97	-0.03	-0.62	0.46	0.46	99
β_4	3.00	2.93	-0.07	-2.40	0.58	0.55	92	3.00	3.02	0.02	0.55	0.53	0.54	99
δ	1.00	0.99	-0.01	-1.00	0.07	0.07	90	1.00	0.97	-0.03	-2.99	0.07	0.07	89
α_0	-6.20	-6.24	-0.04	-0.69	0.36	0.36	94	-1.00	-1.01	-0.01	-0.93	0.05	0.05	90
α_1	0.40	0.42	0.02	5.52	0.33	0.29	92	0.40	0.41	0.01	1.69	0.06	0.07	95
ξ	50.00	56.42	6.42	12.83	24.21	19.35	97	50.00	53.37	3.37	6.74	8.22	8.40	96
γ_0	-1.16	-1.23	-0.07	-5.75	0.35	0.36	99	-1.16	-1.20	-0.04	-3.18	0.37	0.34	95
γ_1	-0.50	-0.53	-0.03	-5.10	0.55	0.55	97	-0.50	-0.47	0.03	6.91	0.58	0.53	91
							IKT							
NIE	0.65	0.10	-0.55	-84.81	-	-	9	9.85	9.20	-0.65	-6.62	-	-	94

4.2 Comparison with CCMM and SparseMCMM

In this comparison, we generated microbiome RA data with Dirichlet distributions (which are essentially multivariate extensions of ZIB distributions) reflecting the real microbial composition in one of our real study examples. The number of taxa and sample size used in this simulation were similar to those in SparseMCMM [44]. Multiple testing was adjusted using the Benjamini-Hochberg Procedure [5] in this comparison such that the targeted FDR is 5%. The Dirichlet distribution is parameterized by the parameters $\mu_1, \mu_2, \ldots, \mu_{K+1}$ and ϕ where $\mu_1, \mu_2, \ldots, \mu_{K+1}$ are the mean parameters of RA and $\sum_{k=1}^{K+1} \mu_k = 1$, and ϕ is the dispersion parameter. In the data generation, the mean parameters were assumed to depend on the independent variable X through a typical multinomial logistic regression equations:

$$\mu_k = \frac{\exp\left(\alpha_0^k + \alpha_1^k X\right)}{1 + \sum_{k=1}^{K} \exp\left(\alpha_0^k + \alpha_1^k X\right)}, \quad k = 1, \ldots, K,$$
$$\mu_{K+1} = \frac{1}{1 + \sum_{k=1}^{K} \exp\left(\alpha_0^k + \alpha_1^k X\right)}.$$

These equations are essentially a multivariate extension of equation (3). Let $\alpha_0 = (\alpha_0^1, \alpha_0^2, \ldots,$
$\alpha_0^K)$ and $\alpha_1 = (\alpha_1^1, \alpha_1^2, \ldots, \alpha_1^K)$. We set $\alpha_0^1 = -3.8$ and all other elements of α_0 were generated from the uniform distribution $U(-2, -1)$ and $\alpha_1^1 = 4$ and all other elements of α_1 were generated from the uniform distribution $U(0, 0.6)$. The dispersion parameter ϕ was set to be 50 to mimic overdispersion in real data. Since both CCMM and SparseMM impute zero values with a positive number because they require all RA to be non-zero for their analyses, we generated zero-valued data points for only the first taxon (to minimize the imputation burden for CCMM and SparseMM) using equation (5) where $\gamma_0 = 0$ and $\gamma_1 = 2$. False zeros were also generated only for the first taxon with the LOD mechanism in equation (6) where library size was generated from the empirical distribution of library size in the VSL#3 study data [2].

We considered two scenarios, Scenario 3 and Scenario 4, for generating Y where the binary variable $1_{(M_1>0)}$ is the mediator in Scenario 3, and both $1_{(M_1>0)}$ and M_1 are mediators in Scenario 4. The outcome Y was generated using the following equation for Scenario 3:

$$Y = \beta_0 + \beta_2 1_{(M_1>0)} + \beta_3 X + \beta_4 X 1_{(M_1>0)} + \epsilon, \tag{8}$$

where M_1 denote the RA of the first taxon, $(\beta_0, \beta_2, \beta_3, \beta_4) = (-2, 6, 5, 4)$ and ϵ follows the standard normal distribution. Notice that there is only one taxon (i.e., the first taxon) mediating the effect X on Y under this model. The outcome Y only depends on the first taxon through the binary indicator variable $1_{(M_1>0)}$ which implies that NIE1 is zero in this scenario. In the data analysis step of the simulation, MedZIM analyzed each taxon as a mediator one by one whereas the other two approaches employed regularization methods to handle high dimensionality. When analyzing a taxon that did not have any zeros, MedZIM used equation (1) for the model of Y and assumed $\beta_2 = \beta_4 = 0$.

Six indices were used to evaluate the model performance: Recall, Precision, F1, bias of the estimate of mediation effect, bias percentage and coverage probability of a 95% CI. Recall, Precision and F1 were calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{F1} = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$$

where TP, FP, TN and FN denote true positive, false positive, true negative and false negative respectively. Recall is a measure of statistical power, the higher the better. Precision has an inverse relationship with false discovery rate (FDR) which is equal to (1-Precision), and thus the higher the Precision, the lower the FDR. When FP=0, Precision was set to be 1 regardless of whether or not TP=0. F1 is the Harmonic mean [30] of Recall and Precision that measures the overall performance in terms of Recall and Precision. The targeted FDR level is set to be 5% for all the three approaches in this comparison which means that the targeted Precision is 95%. As shown in Table 2, MedZIM had good Recall (>80%) and good F1 (>80%) for all cases except the case with sample size of 100 and 50 taxa. MedZIM either identified the only correct taxon or did not identify any taxon at all which led to FP=0 and was why the Precision was 100% and thus it achieved the targeted Precision of 95% for all cases. CCMM had low Recall and F1 throughout all cases, but CCMM had Precision around the targeted Precision of 95% for most cases. Notice that the Precision of CCMM was much higher than its Recall. This was because CCMM did not identify any taxon most of the times and that led to FP=0 which in turn generated high Precision. These three measures were not available for SparseMCMM because it does not provide p-values for mediation effects of individual taxa and thus FDR adjustment cannot be applied. The reason MedZIM had much better performance for identifying the mediation effect is because CCMM was not developed to identify the mediation effect of the binary change (i.e., $1_{(M_1>0)}$) whereas MedZIM can accommodate the mediation effect of $1_{(M_1>0)}$ induced by zero-inflated mediators. With respect to the estimate of mediation effect, Table 2 also shows that MedZIM can provide virtually unbiased estimates along with 95% CI that had coverage probability (CP) around the desired level of 95% under various settings. CCMM generated large biases for its estimates and thus the CP of 95% CI was 0%. SparseMCMM generated large biases as well. CP is not available with SparseMCMM because it does not provide CI for any individual taxon when its 95% CI contains 0. The observed suboptimal performance of CCMM and SparseMCMM is probably because they were not developed to estimate the effect mediated by the binary variable $1_{(M_1>0)}$ caused by the zero-inflated data structure of the mediator whereas MedZIM can handle zero-inflated mediators reasonably well.

For Scenario 4, the outcome Y was generated using the following equation:

$$Y = \beta_0 + \beta_1 M_1 + \beta_2 1_{(M_1>0)} + \beta_3 X + \beta_4 X 1_{(M_1>0)} + \epsilon. \tag{9}$$

Both the continuous M_1 and the binary variable $1_{(M_1>0)}$ were mediators in this scenario and thus both NIE_1 and NIE_2 are nonzero. Notice that the only difference between the above equation (9) and equation (8) is that $\beta_1 M_1$ is included in equation 9 which is the only difference between Scenario 4 and Scenario 3. The value of β_1 was set to be 100 in Scenario 4 and the values of all other parameters were set to be same as in Scenario 3. Notice that because of the compositional structure of RA data, model (9) is equivalent to the following model:

$$Y = (\beta_0 + \beta_1) M_1 + \beta_0 \sum_{k=2}^{K+1} M_k + \beta_2 1_{(M_1>0)} + \beta_3 X + \beta_4 X 1_{(M_1>0)} + \epsilon, \tag{10}$$

where all the taxa are included and thus the compositional structure is accounted for to some extent in model (9).

The simulation results (See Table 3) showed that MedZIM had a very good overall performance for identifying NIE_1 and NIE_2 in terms of Recall (>80%), Precision (=100%)

Table 2: Simulation results for the comparison of MedZIM with CCMM and SparseMCMM when the presence/absence of a taxon is the only mediator. Here n denotes the sample size and $K+1$ denotes the number of taxa. Recall, Precision, F1 and CP cannot be extracted from SparseMCMM because it does not provide a p value for testing individual taxa and it does not provide CI if the 95% CI contains zero.

		Recall (%)			Precision (%)			F1 (%)		
$K+1$	n	MedZIM (NIE2)	CCMM	SparseMCMM	MedZIM (NIE2)	CCMM	SparseMCMM	MedZIM (NIE2)	CCMM	SparseMCMM
10	100	80	57	–	100	99	–	80	57	–
	150	100	72	–	100	95	–	100	70	–
25	100	81	16	–	100	92	–	81	16	–
	150	96	24	–	100	96	–	96	23	–
50	100	54	2	–	100	83	–	54	2	–
	150	89	6	–	100	82	–	89	6	–
		Bias			Bias percentage (%)			Coverage probability (%)		
10	100	0.02	1.05	1.11	1.37	68.85	72.74	96	0	–
	150	-0.01	1.03	1.07	-0.67	67.76	70.56	96	0	–
25	100	-0.03	1.29	1.38	-2.25	84.94	90.39	98	0	–
	150	0.001	1.31	1.39	0.08	86.23	91.05	94	0	–
50	100	0.04	1.50	1.58	2.47	98.36	103.47	97	0	–
	150	-0.004	1.50	1.58	-0.27	98.76	104.02	97	0	–

and F1 (>80%) for most cases except for NIE_1 with sample size of 300 and 50 taxa and NIE_2 with sample size of 300 and 10 taxa. Again, MedZIM either identified the only correct taxon or did not identify any taxon at all which led to the Precision=100% and thus it achieved the targeted Precision of 95% for all cases. MedZIM also had a good performance for estimating NIE_1 and NIE_2 in terms of bias (<0.07), bias percentage (<4.1%) and CP (88-99%). CCMM had good performance in terms of Recall (84-98%) and F1 (82-88%) for the cases with 10 taxa, but not for other cases. It achieved the targeted Precision of 95% only for case with sample size of 300 and 10 taxa. The estimates of CCMM had large bias > 94% and low CP (0-74%) throughout all the cases in Table 3. Again, the Recall, Precision, F1 and CP were not available for SparseMCMM and the estimate of mediation effect for SparseMM had high bias (>116%). The reason CCMM and SparseMM generated high bias in their estimates was likely because (a) CCMM and SparseMCMM were proposed to model the RA on log-scale whereas equation (9) is on the original scale of RA, and (b) CCMM and SparseMCMM were not developed to incorporate the mediation effect of the binary variable $1_{(M_1>0)}$.

5 Real data applications

5.1 New Hampshire Birth Cohort Study (NHBCS)

The NHBCS is an NIH-funded ongoing prospective epidemiological study to investigate the health impacts of environmental exposures with a focus on arsenic exposure in pregnant women and their children [15]. Pregnant women were recruited at about 24 to 28 weeks of gestational age and both mothers and babies are followed up regularly after birth. We applied our approach in the NHBCS study to examine the mediation effect of gut microbiome in the causal pathway from maternal arsenic exposure to infant's health outcomes during the first year of life. In our analysis, the total *in-utero* arsenic level [31] was the exposure variable X, gut microbiome of infants at 6 weeks of age was the

Table 3: Simulation results for the comparison of MedZIM with CCMM and SparseMCMM when both the presence/absence of a taxon and its abundance level are mediators. Here n denotes the sample size and $K+1$ denotes the number of taxa. Recall, Precision, F1 and CP cannot be extracted from SparseMCMM because it does not provide a p value for testing individual taxa and it does not provide CI if the 95% CI contains zero.

		Recall (%)				Precision (%)				F1 (%)			
$K+1$	n	MedZIM (NIE1)	MedZIM (NIE2)	CCMM	SparseMCMM	MedZIM (NIE1)	MedZIM (NIE2)	CCMM	SparseMCMM	MedZIM (NIE1)	MedZIM (NIE2)	CCMM	SparseMCMM
10	300	90	71	84	–	100	100	97	–	90	71	82	–
	400	99	98	98	–	100	100	85	–	99	98	88	–
25	300	85	98	32	–	100	100	76	–	85	98	30	–
	400	100	99	59	–	100	100	81	–	100	99	54	–
50	300	64	99	5	–	100	100	82	–	64	99	4	–
	400	89	100	2	–	100	100	68	–	89	100	2	–
		Bias				Bias(%)				CP (%)			
10	300	-0.01	0.06	-3.41	-5.10	-0.20	4.09	-274.71	-299.23	91	94	0	–
	400	-0.02	0.004	-3.43	-5.09	-0.65	0.26	-275.86	-298.14	94	97	0	–
25	300	0.01	0.01	-0.53	-0.91	0.82	0.97	-383.50	-342.89	99	91	74	–
	400	0.01	0.02	-0.53	-0.93	0.91	1.43	-382.46	-350.06	96	88	65	–
50	300	-0.002	-0.002	0.76	0.67	-0.22	-0.13	94.45	125.20	97	95	29	–
	400	-0.01	-0.005	0.75	0.93	-1.76	-0.32	94.09	116.30	96	96	1	–

mediator M and the outcome Y is the total number of infections treated with a prescribed medicine between 4 and 12 months of age. Here X is a continuous variable and Y is treated as a continuous variable. The gut microbiome data was measured in DNA extracted from infant stool samples using 16S rRNA sequencing [24, 28]. After quality control and data cleaning, there were 195 subjects and 224 genera available in the data set. 85% of the microbiome data points were zero. Relative abundance (RA) of each genus was analyzed as a mediator variable using a ZIB distribution. We estimated all mediation effects (i.e., NIE_1, NIE_2, NIE) and their 95% CI for the exposure variable increasing from 0 to 1 meaning $x_1 = 0$ and $x_2 = 1$. Notice that x_1 and x_2 can take other values as needed depending on the interest of investigators. We used the BH approach [5] for multiple-testing adjustment (targeted FDR=20%) and the 95% CI were calculated before the adjustment. We found 1 genus *Fusobacterium* that was statistically significantly mediating the effect of *in-utero* arsenic on the infection outcome through NIE_1 which means that the RA level of *Fusobacterium* had a statistically significant mediation effect but the presence of *Fusobacterium* did not appear to have a significant mediation effect. The estimate for NIE_1 was -0.005 (95% CI: -0.007, -0.002). *Fusobacterium* has been well known to be associated with childhood infection in the literature[1, 8]. To give a full picture of the mediation effects, a heatmap was constructed (see Figure 2) to illustrate the NIE_1 effects of all genera. IKT and CCMM did not find any significant mediation effects of the microbial taxa. SparseMCMM is not applicable to this study because it requires the X variable to be a binary variable.

5.2 VSL#3 mouse model

VSL#3 is a commercially available probiotic cocktail (Sigma-Tau Pharmaceuticals, Inc.) of eight strains of lactic acid-producing bacteria: *Lactobacillus plantarum, Lactobacillus delbrueckii subsp. Bulgaricus, Lactobacillus paracasei, Lactobacillus acidophilus, Bifidobacterium breve, Bifidobacterium longum, Bifidobacterium infantis, and Streptococcus salivarius subsp.* Orally administered VSL#3 has shown success in ameliorating symptoms and reducing inflammation in human pouchitis [18] and ulcerative colitis [35]. Preventive VSL#3 administration can also attenuate colitis in Il10-/- mice [29] and ileitis

in SAMP1/YitFc mice [32]. When used as a preventative strategy, it has the potential capability to prevent inflammation and carcinogenesis. In a mouse model, Arthur et al. [2] studied the ability of a probiotic cocktail VSL#3 to alter the colonic microbiota and decrease inflammation-associated colorectal cancer when administered as interventional therapy after the onset of inflammation. The study duration was 24 weeks. In this study, there were 24 mice of which 10 were treated with VSL#3 and 14 served as control. Gut microbiome data were collected from stools at the end of the study with 16S rRNA sequencing. We obtained sequence data from Arthur et al. [2] and generated open reference OTUs using the Quantitative Insights into Microbial Ecology (QIIME) [9] version 1.9.1 at 97% similarity level using the Greengenes 97% reference dataset (release 13_8). Chimeric sequences were detected and removed using QIIME. OTUs that had 0.005% of the total number of sequences were excluded according to Bokulich and colleagues [6]. Taxonomic assignment was done using the RDP (ribosomal database project) classifier [45] through QIIME with confidence set to 50%. There were 362 OTUs in total in the data sets after quality control and data cleaning. 40% of the OTU RA data points were zero.

RA of each OTU was analyzed as a mediator variable using a ZIB distribution. The outcome variable in our analysis was dysplasia score (the higher the worse), a continuous variable measuring the abnormality of cell growth. The treatment variable is coded as 1/0 indicating VSL#3/control. Again, the FDR approach was used for adjusting for multiple testing such that the targeted FDR is 20% and the 95% CI were calculated before adjustment. Two OTUs were found to be significantly mediating the treatment effects. One of the two OTUs was assigned to the family S24-7 under order Bacteroidales and the other one was assigned to class Bacilli. All the significant mediation effects come from NIE_1 implying that only the mediation effects through the change in continuous abundance scale were statistically significant. The estimates of NIE_1 were 0.27 (95% CI: 0.1, 0.42) and -1.28 (95% CI: -2.06, -0.49) respectively. The family S24-7 and class Bacilli found by our approach have also been reported to be related with colorectal cancer in the literature [7, 33]. To give a full picture of the mediation effects in this data set, a heatmap was constructed (see Figure 3) to illustrate the NIE_1 effects of all OTUs. IKT and CCMM did not find any significant mediation effects of the OTUs. SparseMCMM identified 18 OTUs whose (bootstrapped) 95% CI's of their mediation effects did not contain zero, but the widths of identified CI's were all 0. The 18 OTUs were respectively assigned to species *distasonis*, *acidifaciens*, *farmeri*, *producta* and *muciniphila*, genera *Blautia*, *Clostridium* and *Klebsiella*, families S24-7 and Lachnospiraceae, Phylum Firmicutes and Kindom Bacteria.

6 Discussion

We developed an innovative mediation modeling approach under the PO framework to analyze mediators that have zero-inflated distributions such as the microbiome. We showed that the mediation effect for zero-inflated mediators can be decomposed into two components of which the first is due to the change in the mediator over its positive domain and the second is due to the discrete binary change from zero to a non-zero status. These two components have different interpretations and are equally important for investigating causal mechanisms. Although the derivation of the decomposition was done for continuous outcome variables, it can be easily extended to cases with other types of outcome variables such as binary outcomes by employing a generalized linear

model for Y. When the point mass $\mathcal{G}(\theta)$ is zero for the mediator (i.e., the distribution is not zero-inflated), the model reduces to a usual mediation analysis model. Therefore, this approach can be also used for data sets after zero values are imputed with a positive number or other normalization techniques are applied. Commonly used ZIB distributions were considered for microbiome RA in this paper even though this is a fairly general framework for mediators with zero-inflated distributions. This tool will be useful for researchers to evaluate the mediation effects of zero-inflated mediators and disentangle causal pathways that are scientifically important. Hence it can play an important role in translating research findings into medical practice. R scripts for implementing the method are available upon request.

This paper considered X as a univariate variable and did not include covariates as potential confounders in the models. It is straightforward to adjust for a set of covariates using our approach. Let C denote a vector of covariates or potential confounders. Then the NIE and NDE can be calculated at a specific value, c, of C as NIE $= E(Y_{x_2M_{x_2}} - Y_{x_2M_{x_1}}|C = c)$, NDE $= E(Y_{x_2M_{x_1}} - Y_{x_1M_{x_1}}|C = c)$ and CDE $= E(Y_{x_2m} - Y_{x_1m}|C = c)$. The value of c can be taken as the mean value of the covariates similar to how least squares mean is calculated in regression models [17]. CI can be obtained using the delta method or resampling methods. Decomposition of NIE follows the same procedure as shown in Section 2.2.

Several extensions of our approach in future research are worth noting. For high-dimensional mediators, our method can analyze the mediators one by one and employ the BH method [5] to adjust for multiple testing. Although this can analyze the RA by allowing a ZIB distribution for a single taxon at a time which can partially address the compositional structure as shown in equation (10), the correlation due to hierarchical structure of the phylogenetic tree is not utilized when the microbial taxa are analyzed one by one. A natural extension of our approach would be to include all the taxa in a more general model as follows:

$$Y = \sum_{k=1}^{K+1} \beta_k M_k + \beta^2 1_{(M_1>0)} + \beta^3 X + \beta^4 X 1_{(M_1>0)} + \epsilon$$

and use regularization approaches [38, 47] to select mediators for the model. Four assumptions on unmeasured and measured confounders [40] are required to make causal inference for mediation analysis. Sensitivity analysis [20, 43] can be useful to check the robustness of model performance with respect to validation of the assumptions. Existing sensitivity analysis procedures can be adapted in our setting for developing a sensitivity analysis method for our approach.

Misspecification of the mechanisms for observing zero-valued data points could have an impact on the model performance. This is similar to missing data where partial information is available on the missing data. It can be considered as missing not at random (MNAR) [25] because the probability of a data point being observed as zero depends on its true value. Besides the LOD mechanism in equation 6, another possible mechanism could be $P(M^* = 0|M, L) = \exp(-\eta ML)$ where $\eta > 0$ and thus it is a decreasing function of ML, the SAA, such that smaller values of ML are more likely to be observed as zero. Notice that the observed value M^* is equal to zero with probability of one when $M = 0$ which corresponds to the case that M is truly zero. Model selection approaches such BIC or AIC can be used to choose different mechanisms. Although these mechanisms may not be perfect to account for MNAR, it can, to a large extent, alleviate

the burden of not accounting for false zeros in the data at all. A future project has been planned to study the robustness of our model with respect to the mechanism for observing zeros using sensitivity analysis techniques.

7 Appendix

7.1 Mediation model for microbial relative abundance (RA)

Subject index i is suppressed in this section for simplicity. Let M denote the RA of a microbial taxon, we use a ZIB distribution for modeling M and its two-part density can be written as the following:

$$f(m;\theta) = \begin{cases} \mathcal{G}(\theta), & m = 0 \\ (1-\mathcal{G}(\theta))G(m;\theta), & m > 0 \end{cases}$$

where $\theta = (\mu, \phi, \Delta)$, $\mathcal{G}(\theta) = \Delta$ and

$$G(m;\theta) = \frac{m^{\mu\phi-1}(1-m)^{(1-\mu)\phi-1}}{B\big(\mu\phi,(1-\mu)\phi\big)},\ \ m \in (0,1), 1 > \mu > 0, \phi > 0.$$

Here $B(\cdot,\cdot)$ is the Beta function and we use the mean and dispersion parameterization for the Beta density function $G(m;\theta)$ [12, 16]. The transformation function and vectors in equation (2) are given by: $T(\theta) = (\log(\mu/(1-\mu)), \log(\phi), \log(\Delta/(1-\Delta))^T$, $\nu_0 = (\alpha_0, \xi, \gamma_0)^T$ and $\nu_1 = (\alpha_1, 0, \gamma_1)^T$. We use identity link for $g(\cdot)$ in equation (1) since Y is a continuous outcome, and thus the mediation model consists of the following equations:

$$\begin{aligned} &Y = \beta_0 + \beta_1 M + \beta_2 1_{(M>0)} + \beta_3 X + \beta_4 X 1_{(M>0)} + \beta_5 XM + \epsilon \\ &\log\left(\frac{\mu}{1-\mu}\right) = \alpha_0 + \alpha_1 X, \\ &\log(\phi) = \xi, \\ &\log\left(\frac{\Delta}{1-\Delta}\right) = \gamma_0 + \gamma_1 X. \end{aligned}$$

Let $\zeta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \delta, \alpha_0, \alpha_1, \xi, \gamma_0, \gamma_1)^T$. The formulas for NIE, NIE_1, NIE_2, NDE and CDE are given below where NIE_1, NIE_2, NDE and CDE can be considered as functions of the parameter vector ζ:

$$\begin{aligned} \text{NIE} &= E(Y_{x_2 M_{x_2}} - Y_{x_2 M_{x_1}}) \\ &= E((\beta_1 + x_2\beta_5)(M_{x_2} - M_{x_1}) + (\beta_2 + x_2\beta_4)(1_{(M_{x_2}>0)} - 1_{(M_{x_1}>0)})) \\ &= (\beta_1 + x_2\beta_5)(E(M_{x_2}) - E(M_{x_1})) + (\beta_2 + x_2\beta_4)(E(1_{(M_{x_2}>0)}) - E(1_{(M_{x_1}>0)})) \\ &= \text{NIE}_1 + \text{NIE}_2, \\ \text{NIE}_1 &= f_1(\zeta) \\ &= (\beta_1 + x_2\beta_5)\Big(\text{expit}(\alpha_0 + \alpha_1 x_2) - \text{expit}(\alpha_0 + \alpha_1 x_1)\Big) \\ &- (\beta_1 + x_2\beta_5)\Big(\text{expit}(\gamma_0 + \gamma_1 x_2)\text{expit}(\alpha_0 + \alpha_1 x_2) \\ &- \text{expit}(\gamma_0 + \gamma_1 x_1)\text{expit}(\alpha_0 + \alpha_1 x_1)\Big) \end{aligned}$$

$$
\begin{aligned}
\text{NIE}_2 &= f_2(\zeta) \\
&= (\beta_2 + x_2\beta_4)\big(\text{expit}(\gamma_0 + \gamma_1 x_1) - \text{expit}(\gamma_0 + \gamma_1 x_2)\big)
\end{aligned}
$$

$$
\begin{aligned}
\text{NDE} &= f_3(\zeta) \\
&= E\big(Y_{x_2 M_{x_1}} - Y_{x_1 M_{x_1}}\big) = (\beta_3 + \beta_4 E(1_{(M_{x_1}>0)}) + \beta_5 E(M_{x_1}))(x_2 - x_1) \\
&= \big(\beta_3 + \beta_4 \text{expit}(\gamma_0 + \gamma_1 x_1) + \beta_5(1 - \text{expit}(\gamma_0 + \gamma_1 x_1))\text{expit}(\alpha_0 + \alpha_1 x_1)\big)(x_2 - x_1) \\
\text{CDE} &= f_4(\zeta) \\
&= E\big(Y_{x_2 m} - Y_{x_1 m}\big) = \big(\beta_3 + \beta_4 1_{(m>0)} + \beta_5 m\big)(x_2 - x_1).
\end{aligned}
$$

7.2 Log-likelihood function under the LOD mechanism for observing zero values

Let $(y_i, r_i, m_i^*, x_i, l_i)$ denote the observed data for the ith subject where l_i is the observed library size. If the ith subject is in the first group, its log-likelihood contribution can be calculated as:

$$
\begin{aligned}
\ell_i^1 &= \log(f(y_i, r_i | m_i^*, x_i, \mathcal{L}_i) f(m_i^* | x_i, l_i)) = \log(f(y_i | m_i^*, x_i, l_i) f(r_i | m_i^*, x_i, l_i) f(m_i^* | x_i, l_i)) \\
&= \log(f(y_i | m_i^*, x_i, l_i)) + \log(f(r_i | m_i^*, l_i)) + \log(f(m_i^* | x_i, l_i)) \\
&= -\log(\delta) - \frac{\big(y_i - \beta_0 - \beta_1 m_i^* - \beta_2 - (\beta_3 + \beta_4) x_i - \beta_5 x_i m_i^*\big)^2}{2\delta^2} \\
&\quad + \log(1 - \Delta_i) - \log\Big(B\big(\mu_i\phi, (1-\mu_i)\phi\big)\Big) \\
&\quad + (\mu_i\phi - 1)\log(m_i^*) + \big((1-\mu_i)\phi - 1\big)\log(1 - m_i^*) + cons,
\end{aligned}
$$

where $\Delta_i = \text{expit}(\gamma_0 + \gamma_1 x_i)$, $\mu_i = \text{expit}(\alpha_0 + \alpha_1 x_i)$ and $\phi = \exp(\xi)$. If the ith subject is in the second group, its log-likelihood contribution can be calculated as:

$$
\begin{aligned}
\ell_i^2 &= \log(f(y_i, r_i, m_i^* | x_i)) = \log\Big(\int_0^{1/l_i} f(y_i, r_i | m, x_i) dF(m | x_i)\Big) \\
&= \log\Big(\int_0^1 f(y_i | m, x_i) f(r_i | m) dF(m | x_i)\Big) \\
&= \log\Bigg(\frac{\Delta_i}{\sqrt{2\pi\delta^2}} \exp\Big(-\frac{(y_i - \beta_0 - \beta_3 x_i)^2}{2\delta^2}\Big) \\
&\quad + \int_0^{1/l_i} f(y_i | m, x_i)(1 - \Delta_i) \frac{m^{\mu_i\phi - 1}(1-m)^{(1-\mu_i)\phi - 1} \exp(-\eta m l_i)}{B\big(\mu_i\phi, (1-\mu_i)\phi\big)} dm\Bigg) \\
&= -\log(\delta) + \log\Bigg(\Delta_i \exp\Big(-\frac{(y_i - \beta_0 - \beta_3 x_i)^2}{2\delta^2}\Big) + \frac{1 - \Delta_i}{B\big(\mu_i\phi, (1-\mu_i)\phi\big)} \int_0^{1/l_i} h_i(m) dm\Bigg) \\
&\quad + cons,
\end{aligned}
$$

where

$$
h_i(m) = m^{\mu_i\phi - 1}(1-m)^{(1-\mu_i)\phi - 1}
$$

$$\times \exp\left(- \frac{\big(y_i - \beta_0 - \beta_1 m - \beta_2 - (\beta_3 + \beta_4)x_i - \beta_5 x_i m\big)^2}{2\delta^2} \right).$$

Taken together, the complete log-likelihood function can be calculated as

$$\ell_{ZIB} = \sum_{i \in \text{group 1}} \ell_i^1 + \sum_{i \in \text{group 2}} \ell_i^2.$$

7.3 Delta method for obtaining 95% CI of NIE_1, NIE_2, NDE and CDE

Since NIE_1, NIE_2, NDE and CDE can be treated as functions of the parameter vector ζ as shown in Section 7.1, it suffices to derive the 95% CI for $f_1(\hat{\zeta})$, $f_2(\hat{\zeta})$, $f_3(\hat{\zeta})$ and $f_4(\hat{\zeta})$ where $\hat{\zeta}$ is the MLE of ζ. We first calculate the observed Fisher information matrix which can be calculated as $I_{obs} = -\frac{\partial^2 \ell_{ZIB}}{\partial \zeta \partial \zeta^T}|_{\zeta=\hat{\zeta}}$ where ℓ_{ZIB} is the log-likelihood function derived in Section 7.2. By using the multivariate Delta method, we can calculate the variance of the estimates as follows:

$$\begin{aligned}
\text{var}(\text{NIE}_1) = \text{var}(f_1(\hat{\zeta})) &= \left(\frac{\partial f_1(\zeta)}{\partial \zeta}|_{\zeta=\hat{\zeta}}\right)^T \text{var}(\hat{\zeta}) \left(\frac{\partial f_1(\zeta)}{\partial \zeta}|_{\zeta=\hat{\zeta}}\right) \\
&= \left(\frac{\partial f_1(\zeta)}{\partial \zeta}|_{\zeta=\hat{\zeta}}\right)^T I_{obs}^{-1} \left(\frac{\partial f_1(\zeta)}{\partial \zeta}|_{\zeta=\hat{\zeta}}\right),
\end{aligned}$$

where $\frac{\partial f_1(\zeta)}{\partial \zeta} = \left(\frac{\partial f_1(\zeta)}{\partial \beta_0}, \frac{\partial f_1(\zeta)}{\partial \beta_1}, \ldots, \frac{\partial f_1(\zeta)}{\partial \gamma_1}\right)^T$. Let $z_{0.025}$ denotes the 97.5th percentile of the standard normal distribution and the 95% CI of NIE_1 can calculated as $\left(f_1(\hat{\zeta}) - z_{0.025}\sqrt{\text{var}(f_1(\hat{\zeta}))}, f_1(\hat{\zeta}) + z_{0.025}\sqrt{\text{var}(f_1(\hat{\zeta}))}\right)$. The 95% CI for NIE_2, NDE and CDE can be calculated similarly.

Acknowledgements

The first author is supported in part by NIH Grants R01GM123014, UH3OD023275 and P20GM104416. The second and third authors are supported in part by NIH Grant R01GM123014. The last fours authors are supported in part by R01GM123014, UH3OD023275, P01ES022832 and P20GM104416, and U.S. EPA grant RD 83544201.

References

[1] Karen Arane and Ran D Goldman. Fusobacterium infections in children. *Canadian family physician Medecin de famille canadien*, 62:813–814, October 2016.

[2] Janelle C Arthur, Raad Z Gharaibeh, Joshua M Uronis, Ernesto Perez-Chanona, Wei Sha, Sarah Tomkovich, Marcus Mühlbauer, Anthony A Fodor, and Christian Jobin. Vsl# 3 probiotic modifies mucosal microbial composition but does not reduce colitis-associated colorectal cancer. *Scientific reports*, 3:2868, 2013.

[3] Reuben M. Baron and David A Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182, 1986.

[4] Yasmine Belkaid and Timothy W Hand. Role of the microbiota in immunity and inflammation. *Cell*, 157:121–141, March 2014.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A Practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.*, 57:289–300, 1995.

[6] Nicholas A Bokulich, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, and J Gregory Caporaso. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nature methods*, 10(1):57, 2013.

[7] Line Skute Bråten, Marianne Sødring, Jan Erik Paulsen, Lars Gustav Snipen, and Knut Rudi. Cecal microbiota association with tumor load in a colorectal cancer mouse model. *Microbial ecology in health and disease*, 28(1):1352433, 2017.

[8] Itzhak Brook. Fusobacterial head and neck infections in children. *International journal of pediatric otorhinolaryngology*, 79:953–958, July 2015.

[9] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335, 2010.

[10] Haitao Chai, Hongmei Jiang, Lu Lin, and Lei Liu. A marginalized two-part beta regression model for microbiome compositional data. *PLoS computational biology*, 14(7):e1006329, 2018.

[11] Eric Z Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics (Oxford, England)*, 32:2611–2617, September 2016.

[12] Francisco Cribari-Neto and Achim Zeileis. Beta regression in R. *Journal of Statistical Software*, 34:24848, 2010.

[13] M. L. Dalrymple, I. L. Hudson, and R. P. K. Ford. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41(3-4):491–504, 2003.

[14] B Efron and R Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

[15] S Farzan, S Korrick, Z Li, R Enelow, A Gandolfi, J Madan, K Nadeau, and M Karagas. In utero arsenic exposure and infant infection in a united states cohort: A prospective study. *Environmental Research*, 126:24–30, 2013.

[16] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31:799–815, 2004.

[17] D. Gianola. Least-squares means vs population marginal means. *American Statistician*, 36(1):65–66, 1982.

[18] Paolo Gionchetti, Fernando Rizzello, Alessandro Venturi, Patrizia Brigidi, Diego Matteuzzi, Gabriele Bazzocchi, Gilberto Poggioli, Mario Miglioli, and Massimo Campieri. Oral bacteriotherapy as maintenance treatment in patients with chronic pouchitis: a double-blind, placebo-controlled trial. *Gastroenterology*, 119(2):305–309, 2000.

[19] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15:309–334, 2010.

[20] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

[21] Chengcheng Jin, Georgia K Lagoudas, Chen Zhao, Susan Bullman, Arjun Bhutkar, Bo Hu, Samuel Ameh, Demi Sandel, Xu Sue Liang, Sarah Mazzilli, Mark T Whary, Matthew Meyerson, Ronald Germain, Paul C Blainey, James G Fox, and Tyler Jacks. Commensal microbiota promote lung cancer development via gammadelta t cells. *Cell*, 176:998–1013.e16, February 2019.

[22] Theis Lange, Kim Wadt Hansen, Rikke Sørensen, and Søren Galatius. Applied mediation analyses: a review and tutorial. *Epidemiology and health*, 39:e2017035, 2017.

[23] Hongzhe Li. Statistical and computational methods in microbiome and metagenomics. *Handbook in Statistical Genomics*, 2018.

[24] Zhigang Li, Katherine Lee, Margaret R. Karagas, Juliette C. Madan, Anne G. Hoen, A. James O' Malley, and Hongzhe Li. Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Stat Biosci*, pages 1–22, jul 2018.

[25] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[26] David P. MacKinnon. *Introduction to statistical mediation analysis*. New York: Erlbaum, 2008.

[27] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annual review of psychology*, 58:593–614, 2007.

[28] Juliette C Madan, Anne G Hoen, Sara N Lundgren, Shohreh F Farzan, Kathryn L Cottingham, Hilary G Morrison, Mitchell L Sogin, Hongzhe Li, Jason H Moore, and Margaret R Karagas. Association of cesarean delivery and formula supplementation with the intestinal microbiome of 6-week-old infants. *JAMA pediatrics*, 170:212–219, March 2016.

[29] Karen Madsen, Anthony Cornish, Paul Soper, Conor McKaigney, Humberto Jijon, Christine Yachimec, Jason Doyle, Lawrence Jewell, and Claudio De Simone. Probiotic bacteria enhance murine and human intestinal epithelial barrier function. *Gastroenterology*, 121(3):580–591, 2001.

[30] Marilyn N Martinez and Mary J Bartholomew. What does it "mean"? a review of interpreting and calculating different types of means and standard deviations. *Pharmaceutics*, 9(2), apr 2017.

[31] Kari C Nadeau, Zhigang Li, Shohreh Farzan, Devin Koestler, David Robbins, Dennis Liang Fei, Meena Malipatlolla, Holden Maecker, Richard Enelow, Susan Korrick, et al. In utero arsenic exposure and fetal immune repertoire in a us pregnancy cohort. *Clinical Immunology*, 155(2):188–197, 2014.

[32] Cristiano Pagnini, Rubina Saeed, Giorgos Bamias, Kristen O Arseneau, Theresa T Pizarro, and Fabio Cominelli. Probiotics promote gut health through stimulation of epithelial innate immunity. *Proceedings of the national academy of sciences*, 107(1):454–459, 2010.

[33] Brandilyn A Peters, Christine Dominianni, Jean A Shapiro, Timothy R Church, Jing Wu, George Miller, Elizabeth Yuen, Hal Freiman, Ian Lustbader, James Salik, et al. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome*, 4(1):69, 2016.

[34] Michael B. Sohn and Hongzhe Li. Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics*, 13(1):661–681, 2019.

[35] Ajit Sood, Vandana Midha, Govind K Makharia, Vineet Ahuja, Dinesh Singal, Pooja Goswami, and Rakesh K Tandon. The probiotic preparation, vsl# 3 induces remission in patients with mild-to-moderately active ulcerative colitis. *Clinical Gastroenterology and Hepatology*, 7(11):1202–1209, 2009.

[36] Takeshi Tanoue, Satoru Morita, Damian R Plichta, Ashwin N Skelly, Wataru Suda, Yuki Sugiura, Seiko Narushima, Hera Vlamakis, Iori Motoo, Kayoko Sugita, Atsushi Shiota, Kozue Takeshita, Keiko Yasuma-Mitobe, Dieter Riethmacher, Tsuneyasu Kaisho, Jason M Norman, Daniel Mucida, Makoto Suematsu, Tomonori Yaguchi, Vanni Bucci, Takashi Inoue, Yutaka Kawakami, Bernat Olle, Bruce Roberts, Masahira Hattori, Ramnik J Xavier, Koji Atarashi, and Kenya Honda. A defined commensal consortium elicits cd8 t cells and anti-cancer immunity. *Nature*, 565:600–605, January 2019.

[37] H. J. Terhorst. On stieltjes integration in euclidean-space. *Journal of Mathematical Analysis and Applications*, 114(1):57–74, 1986.

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(3):273–282, 2011.

[39] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. mediation: R package for causal mediation analysis. *https://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf*, 2017.

[40] Tyler J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20:18–26, 2009.

[41] Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford Univ. Press, 2015.

[42] Tyler J. VanderWeele. Mediation analysis: A practitioner's guide. *Annu Rev Public Health*, 37:17–32, 2016.

[43] Tyler J VanderWeele and Yasutaka Chiba. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, biostatistics, and public health*, 11, 2014.

[44] Chan Wang, Jiyuan Hu, Martin J Blaser, and Huilin Li. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*, 2019.

[45] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, 2007.

[46] Xinyi Wang, Guangqiang Sun, Teng Feng, Jing Zhang, Xun Huang, Tao Wang, Zuoquan Xie, Xingkun Chu, Jun Yang, Huan Wang, Shuaishuai Chang, Yanxue Gong, Lingfei Ruan, Guanqun Zhang, Siyuan Yan, Wen Lian, Chen Du, Dabing Yang, Qingli Zhang, Feifei Lin, Jia Liu, Haiyan Zhang, Changrong Ge, Shifu Xiao, Jian Ding, and Meiyu Geng. Sodium oligomannate therapeutically remodels gut microbiota and suppresses gut bacterial amino acids-shaped neuroinflammation to inhibit alzheimer's disease progression. *Cell research*, 29:787–803, October 2019.

[47] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[48] Haixiang Zhang, Jun Chen, Zhigang Li, and Lei Liu. Testing for mediation effect with application to human microbiome data. *Statistics in Biosciences*, In press, 2019.

[49] Jie Zhang, Zhi Wei, and Jun Chen. A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics (Oxford, England)*, 34:1875–1883, June 2018.

Figure 2: Heatmap of mediation strength based on NIE_1 in NHBCS study. The mediation strength is measured by (1-p) where p is the unadjusted p-value. Negative sign indicates negative NIE_1. Genera are labeled on the vertical axis and samples are labeled on the horizontal axis. Absence of a genus in a sample is coded as 0 in the heatmap.

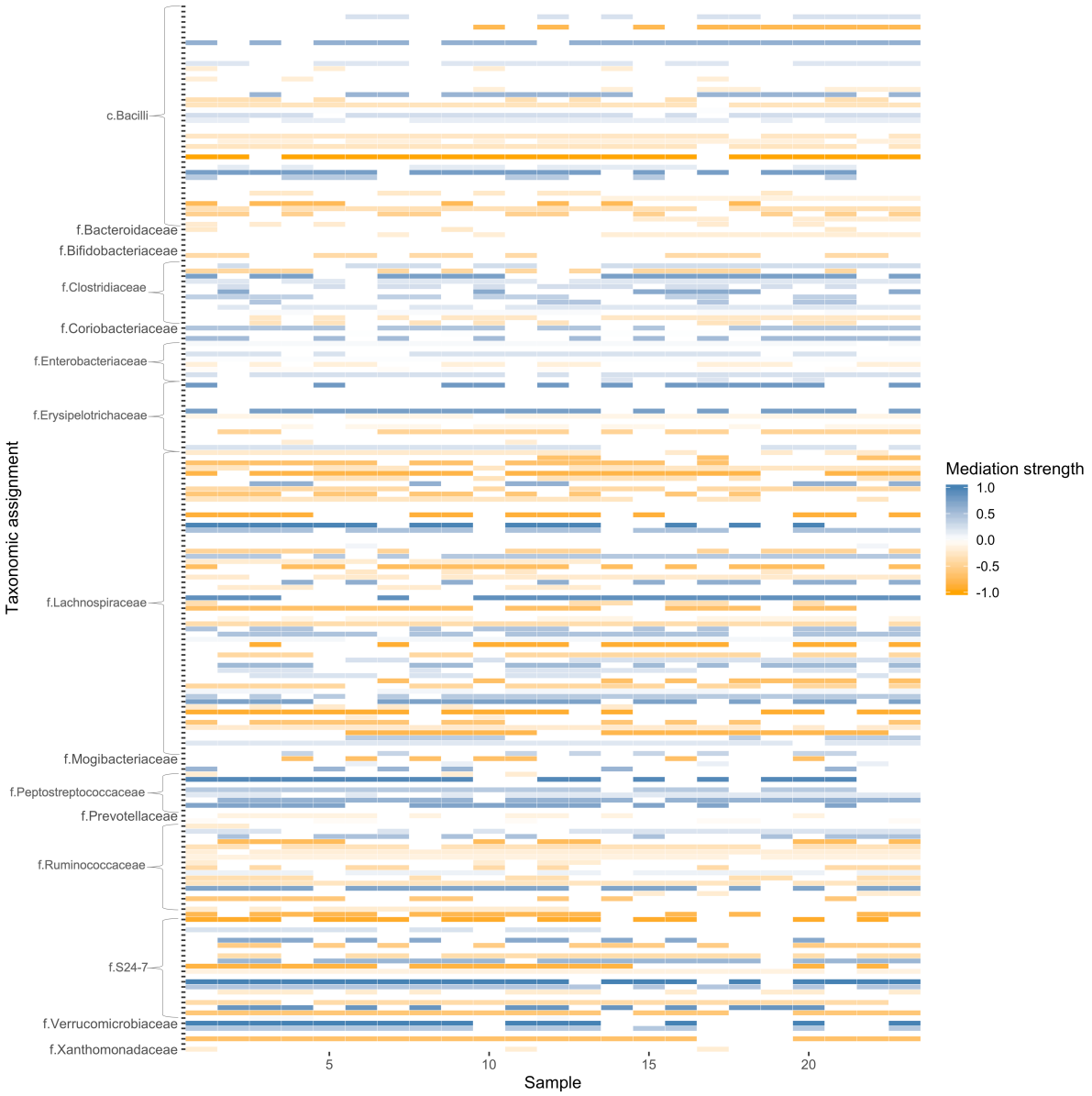

Figure 3: Heatmap of mediation strength based on NIE_1 in VSL#3 study. The mediation strength is measured by (1-p) where p is the unadjusted p-value. Negative sign indicates negative NIE_1. Taxonomic assignment is labeled on the vertical axis. Samples are labeled on the horizontal axis. Absence of an OTU in a sample is coded as 0 in the heatmap.