



## Article

# MarZIC: A Marginal Mediation Model for Zero-Inflated Compositional Mediators with Applications to Microbiome Data

Quran Wu <sup>1</sup>, James O'Malley <sup>2</sup>, Susmita Datta <sup>1</sup>, Raad Z. Gharaibeh <sup>3</sup>, Christian Jobin <sup>3</sup> , Margaret R. Karagas <sup>4</sup>, Modupe O. Coker <sup>4</sup> , Anne G. Hoen <sup>4</sup>, Brock C. Christensen <sup>4</sup>, Juliette C. Madan <sup>4</sup> and Zhigang Li <sup>1,\*</sup>

- <sup>1</sup> Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA; wuquran@ufl.edu (Q.W.); susmita.datta@ufl.edu (S.D.)
- <sup>2</sup> The Dartmouth Institute, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA; james.omalley@dartmouth.edu
- <sup>3</sup> Department of Medicine, University of Florida, Gainesville, FL 32611, USA; raad.gharaibeh@medicine.ufl.edu (R.Z.G.); christian.jobin@medicine.ufl.edu (C.J.)
- <sup>4</sup> Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA; margaret.r.karagas@dartmouth.edu (M.R.K.); modupe.o.coker@dartmouth.edu (M.O.C.); anne.g.hoen@dartmouth.edu (A.G.H.); brock.c.christensen@dartmouth.edu (B.C.C.); juliette.c.madan@hitchcock.org (J.C.M.)
- \* Correspondence: zhigang.li@ufl.edu



**Citation:** Wu, Q.; O'Malley, J.; Datta, S.; Gharaibeh, R.Z.; Jobin, C.; Karagas, M.R.; Coker, M.O.; Hoen, A.G.; Christensen, B.C.; Madan, J.C.; et al. MarZIC: A Marginal Mediation Model for Zero-Inflated Compositional Mediators with Applications to Microbiome Data. *Genes* **2022**, *13*, 1049. <https://doi.org/10.3390/genes13061049>

Academic Editor: Yi-Juan Hu

Received: 10 May 2022

Accepted: 7 June 2022

Published: 11 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Background: The human microbiome can contribute to pathogenesis of many complex diseases by mediating disease-leading causal pathways. However, standard mediation analysis methods are not adequate to analyze the microbiome as a mediator due to the excessive number of zero-valued sequencing reads in the data and that the relative abundances have to sum to one. The two main challenges raised by the zero-inflated data structure are: (a) disentangling the mediation effect induced by the point mass at zero; and (b) identifying the observed zero-valued data points that are not zero (i.e., false zeros). Methods: We develop a novel marginal mediation analysis method under the potential-outcomes framework to address the issues. We also show that the marginal model can account for the compositional structure of microbiome data. Results: The mediation effect can be decomposed into two components that are inherent to the two-part nature of zero-inflated distributions. With probabilistic models to account for observing zeros, we also address the challenge with false zeros. A comprehensive simulation study and the application in a real microbiome study showcase our approach in comparison with existing approaches. Conclusions: When analyzing the zero-inflated microbiome composition as the mediators, MarZIC approach has better performance than standard causal mediation analysis approaches and existing competing approach.

**Keywords:** mediation; microbiome; relative abundance; zero-inflated composition; sparse data

## 1. Introduction

Emerging evidence suggests that the human microbiome and the immune system are constantly shaping each other [1]. The human microbiome can contribute to disease pathogenesis by mediating disease-leading causal pathways in complex diseases such as Alzheimer's disease [2] and cancer [3,4]. To study the human microbiome, 16S ribosomal RNA gene sequencing and metagenomic shotgun sequencing have been popular methods to quantify microbiome composition in microbiome studies. A challenging feature of microbiome sequencing data is that it has excessive number of zeros [5]. Many microbiome data sets have more than 50% of the sequencing reads being 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 compounded

by a compositional structure poses a challenge that needs to be addressed specifically in mediation analyses. Although there have been some exciting efforts to model microbiome as a high-dimensional mediator [6–8], it remains a daunting task to address the zero-inflated data structure.

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 and it has become a common analysis tool in many research areas such as epidemiology, environmental health sciences, medicine, randomized trials and psychiatry. There are two general types of mediation analysis approaches: potential-outcomes (PO) or counterfactual-outcomes methods [9–11] and traditional linear mediation analysis methods [12,13]. The latter approach can be considered as a special case of the former approach that can allow for nonlinear associations and interactions between independent variables and mediators. PO approaches are more flexible because they can allow interaction effects of the independent variable with mediators as well as nonlinear effects. Reviews of mediation analysis approaches and their assumptions can be found in the literature [14–16].

Although mediation modeling frameworks have been well established, to the best of our knowledge, there have been few studies to address zero-inflated compositional mediators. In a typical mediation analysis, the total effect of an independent 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 independent variable and the direct effect measures the causal effect due to change in the independent variable while keeping the mediator variable constant. When the mediator has a marginal zero-inflated distribution such as a zero-inflated Beta (ZIB) distribution, 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 [17]: 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 decomposed accordingly. Both changes have important interpretations in microbiome research. 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 marginal mediation analysis approach under the PO framework to deal with zero-inflated compositional mediators. This approach can allow a mixture of truly zero-valued data points 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. An extensive simulation study is conducted to evaluate our approach MarZIC in comparison with a standard PO

mediation analysis approach [10] and another approach [6] 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 Section 5, and a discussion in Section 6. Additional details and derivations can be found in the Appendices A–C.

## 2. Model and Notation

For simplicity, we suppress the subject index in all notations in this section. Let  $Y$ ,  $M = (M_1, \dots, M_{K+1})$  and  $X$  denote the continuous outcome variable, the compositional mediator variable and the independent variable respectively. For example,  $M$  could be the vector of relative abundances (RA) of microbial taxa. Before constructing the model for zero-inflated data, we first describe the model for the special case where the mediator  $M$  have no zeros which could happen if investigators choose to impute zeros with a Pseudocount or a small positive number. The model for zero-inflated data will be provided after that.

### 2.1. Model for Data without Zeros

We first consider cases where there are no zeros for the mediator  $M$  in the data which is very rare, but it could happen if zeros are replaced by a Pseudocount or a small positive number. We will move to cases with  $M$  containing zeros in the next section. As Dirichlet distributions have been widely used for modeling the RA of microbiome taxa [18–23], let  $M$  follow a  $(K + 1)$ -dimensional Dirichlet distribution indexed by its mean parameters  $\mu_1, \dots, \mu_{K+1}$  with  $\sum_{k=1}^{K+1} \mu_k = 1$  and a dispersion parameter  $\phi$ . We assume the outcome  $Y$  depends on  $M$  and  $X$  through the following regression equation:

$$Y = \sum_{k=1}^{K+1} \beta^k M_k + \beta^X X + \sum_{k=1}^{K+1} \beta^{kk} X M_k + \epsilon \quad (1)$$

where the random error  $\epsilon$  follows a normal distribution with mean of 0 and a constant variance,  $\beta^k$ ,  $\beta^X$  and  $\beta^{kk}$  are regression coefficients, and  $X M_k$  is the interaction term between the independent variable  $X$  and the mediator  $M_k$ . An advantage of using  $M_k$  instead of  $\log(M_k)$  in the model is that it does not require imputing zeros with a positive number. All taxa and their interactions with  $X$  are included and thus the compositional structure is accounted for in this model. Later, we will show that a marginal model can also account for the compositional structure. Equation (1) implies that the marginal association between  $Y$  and any taxon  $M_j$ ,  $j = 1, \dots, K + 1$ , has the following form (derivation can be found in the Appendix A):

$$E_X(Y|M_j) = \beta_0^* + \beta_1^* M_j + \beta_2^* X + \beta_3^* X M_j, \quad (2)$$

where  $E_X(Y|M_j)$  is the mean of  $Y$  conditional on  $M_j$  given  $X$ , and

$$\beta_0^* = \frac{\sum_{k \neq j} \beta^k \mu_k}{\sum_{l \neq j} \mu_l}, \quad \beta_1^* = \beta^j - \beta_0^*, \quad \beta_2^* = \beta^X + \frac{\sum_{k \neq j} \beta^{kk} \mu_k}{\sum_{l \neq j} \mu_l}, \quad \beta_3^* = \beta^{jj} - \frac{\sum_{k \neq j} \beta^{kk} \mu_k}{\sum_{l \neq j} \mu_l}.$$

It is straightforward to see that the full model (1) uniquely determines the marginal association for each taxon. Therefore, without violating model (1), we can construct the following marginal regression model for the association between  $Y$  and  $M_j$  and  $X$  such that it is equivalent to model (1):

$$Y = \beta_0 + \beta_1 M_j + \beta_2 X + \beta_3 X M_j + \epsilon^*, \quad (3)$$

where the random error  $\epsilon^*$  has a normal distribution with mean of 0. An advantage of the above marginal model over model (1) is that it is straightforward to interpret the regression

coefficient  $\beta_1$  as a typical regression coefficient, whereas the corresponding regression coefficient  $\beta^j$  in Equation (1) does not have such a straightforward interpretation. That is because there has to be at least one  $M_k$ ,  $k \neq j$ , changing when  $M_j$  changes due to the compositional structure, and thus it is not possible to hold all  $M_k$ 's,  $k \neq j$ , constant while changing  $M_j$  to interpret  $\beta^j$  as a typical regression coefficient.

Another nice feature of the marginal model (3) is that the true values of its regression parameters ( $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ ) are functions of the parameters  $\mu_1, \dots, \mu_{K+1}$  of the Dirichlet distribution of  $M$  as shown in Equation (2); therefore, the marginal model accounts for the compositional structure.

It is also much more convenient to work on the marginal model (3) due to its simpler form. With that and the above advantages, we propose to use the marginal model (3) for constructing the mediation model. When the vector  $M$  has the Dirichlet distribution as previously assumed in this section,  $M_j$  has a Beta distribution with mean parameter  $\mu_j$  and scale parameter  $\phi$ . The following equation can be used to model the association between  $M_j$  and  $X$ :

$$\ln\left(\frac{\mu_j}{1-\mu_j}\right) = \alpha^0 + \alpha^1 X. \quad (4)$$

Equations (3) and (4) together form our marginal mediation model for the scenario without zeros for  $M$ .

## 2.2. Model for Data with Zeros

Now we consider scenarios where the data for  $M$  contain zeros. Given the advantages of a marginal model as demonstrated in the above subsection, we will again use a marginal model for the association between  $Y$  and any taxon  $M_j$  to form a mediation model. For any taxon  $M_j$ ,  $j = 1, \dots, K + 1$ , we construct the marginal model as follows:

$$Y = \beta_0 + \beta_1 M_j + \beta_2 1_{(M_j > 0)} + \beta_3 X + \beta_4 X 1_{(M_j > 0)} + \beta_5 X M_j + \epsilon \quad (5)$$

where  $1_{(\cdot)}$  is an indicator function indicating whether  $M_j$  is 0, the random error  $\epsilon$  follows a normal distribution  $N(0, \delta)$ , and  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  and  $\beta_5$  are regression coefficients. This model is fully compatible with allowing interactions between the independent variable and mediators as the two interaction terms:  $X 1_{(M_j > 0)}$  and  $X M_j$  are included in Equation (5). In practice, investigators can also include only one or no interaction term depending on the hypothesis of interest.

For the marginal distribution of  $M_j$ , it is natural to use a zero-inflated Beta (ZIB) distribution because the marginal of a Dirichlet distribution is a Beta distribution [18,19]. Its two-part density function is given as follows:

$$f(m) = \begin{cases} \Delta, & m = 0 \\ (1 - \Delta) \frac{m^{\mu_j \phi - 1} (1 - m)^{(1 - \mu_j) \phi - 1}}{B(\mu_j \phi, (1 - \mu_j) \phi)}, & m > 0 \end{cases}$$

where  $\Delta$  is the probability of being 0,  $B(\cdot, \cdot)$  is the Beta function and  $\mu_j$  and  $\phi$  are the mean and dispersion parameters respectively of the Beta distribution for the non-zero part [24,25]. To model the association of the mediator  $M_j$  with  $X$ , we use the following equations:

$$\ln\left(\frac{\mu_j}{1-\mu_j}\right) = \alpha_0 + \alpha_1 X, \quad (6)$$

$$\ln\left(\frac{\Delta}{1-\Delta}\right) = \gamma_0 + \gamma_1 X. \quad (7)$$

Equations (5)–(7) together form our mediation model. The parameter  $\alpha_1$  in Equation (6) measures the association between  $X$  and the RA level of the mediator and  $\gamma_1$  in Equation (7) measures the association between  $X$  and the binary presence of the mediator. Notice that  $X$

is a scalar here, but it is obvious that other covariates such as potential confounders can be included in the model equations.

### 2.3. Mechanism for Observing Zeros of the Mediator

For microbiome abundance data, observations that cannot be detected 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 failure. Let  $M_j^*$  denote the observed value of  $M_j$ . When the observed value is positive (i.e.,  $M_j^* > 0$ ), we assume that  $M_j^* = M_j$ . But when  $M_j^* = 0$ , we don't know whether  $M_j$  is truly zero or  $M_j$  is positive but observed as zero. We consider the following mechanism for the probability of observing a zero of the microbial taxon abundance:

$$P(M_j^* = 0 | M_j, L) = 1_{(M_j L < 1)}, \quad (8)$$

where  $L$  is the library size (i.e., sequencing depth) and the product  $M_j L$  can be interpreted as the sample absolute abundance (SAA) of the  $j$ th 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. Since SAA depends on both  $L$  and  $M_j$ , 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  $E(1_{(M_j L < 1)} | L) = P(M_j < 1/L)$ .

### 2.4. Marginal Mediation Effect and Direct Effect

Under the potential-outcomes (PO) framework [15], we can define the natural indirect effect (NIE), natural direct effects (NDE) and controlled direct effect (CDE) where NIE is the mediation effect. We refer to NIE as the marginal mediation effect because the proposed mediation models are based on marginal models as shown in Section 2. The total effect of  $X$  is equal to the summation of NIE and NDE. For any  $j, j = 1, \dots, K + 1$ , let  $M_j(x)$  denote the value of  $M_j$  if  $X$  equals  $x$ . Let  $Y_{xm}$  denote the value of  $Y$  if  $(X, M_j) = (x, m)$ . The average NIE, NDE and CDE for  $X$  changing from  $x_1$  to  $x_2$  are defined as:

$$\text{NIE} = E(Y_{x_2 M_j(x_2)} - Y_{x_2 M_j(x_1)})$$

$$\text{NDE} = E(Y_{x_2 M_j(x_1)} - Y_{x_1 M_j(x_1)})$$

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

where  $Y_{x_2 M_j(x_1)}$  is a counterfactual outcome. By plugging the Equations (5)–(7) into the above definitions and using Riemann-Stieljes integration [26], we can obtain the following formulas:

$$\begin{aligned} \text{NIE} &= E(Y_{x_2 M_j(x_2)}) - E(Y_{x_2 M_j(x_1)}) = E(E(Y_{x_2 M_j(x_2)} | M_j(x_2))) - E(E(Y_{x_2 M_j(x_1)} | M_j(x_1))) \\ &= E(\beta_0 + \beta_1 M_j(x_2) + \beta_2 1_{(M_j(x_2) > 0)} + \beta_3 x_2 + \beta_4 x_2 1_{(M_j(x_2) > 0)} + \beta_5 x_2 M_j(x_2)) \\ &\quad - E(\beta_0 + \beta_1 M_j(x_1) + \beta_2 1_{(M_j(x_1) > 0)} + \beta_3 x_2 + \beta_4 x_2 1_{(M_j(x_1) > 0)} + \beta_5 x_2 M_j(x_1)) \\ &= (\beta_1 + \beta_5 x_2)(E(M_j(x_2)) - E(M_j(x_1))) + (\beta_2 + \beta_4 x_2)(E(1_{(M_j(x_2) > 0)}) - E(1_{(M_j(x_1) > 0)})) \\ &= \text{NIE}_1 + \text{NIE}_2, \end{aligned}$$

$$\begin{aligned}
 \text{NIE}_1 &= (\beta_1 + \beta_5 x_2)(E(M_j(x_2)) - E(M_j(x_1))) \\
 &= (\beta_1 + \beta_5 x_2) \left( \int_{m \in [0,1]} m dF_{M_j(x_2)}(m) - \int_{m \in [0,1]} m dF_{M_j(x_1)}(m) \right) \\
 &= (\beta_1 + \beta_5 x_2) \left( \text{expit}(\alpha_0 + \alpha_1 x_2) - \text{expit}(\alpha_0 + \alpha_1 x_1) \right) \\
 &\quad - (\beta_1 + \beta_5 x_2) \left( \text{expit}(\gamma_0 + \gamma_1 x_2) \text{expit}(\alpha_0 + \alpha_1 x_2) \right. \\
 &\quad \left. - \text{expit}(\gamma_0 + \gamma_1 x_1) \text{expit}(\alpha_0 + \alpha_1 x_1) \right), \\
 \text{NIE}_2 &= (\beta_2 + \beta_4 x_2)(\text{expit}(\gamma_0 + \gamma_1 x_1) - \text{expit}(\gamma_0 + \gamma_1 x_2)),
 \end{aligned}$$

where  $\text{expit}(\cdot)$  is the inverse function of  $\text{logit}(\cdot)$ ,  $F_{M_j(x)}(m)$  denotes the CDF of  $M_j(x)$  and  $dF_{M_j(x)}(m)$  denotes the stieljes integration [26] with respect to  $F_{M_j(x)}(m)$ . So NIE,  $\text{NIE}_1$ ,  $\text{NIE}_2$ , NDE and CDE can be estimated by plugging the parameter estimates into the formulas. Confidence intervals (CI) are obtained using the multivariate delta method as outlined in the Appendix B. An alternative approach for finding standard errors to construct CI is bootstrapping [27].  $\text{NIE}_1$  can be interpreted as the marginal mediation effect due to the change of the mediator on its numeric scale and  $\text{NIE}_2$  can be interpreted as the marginal 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  $M_j$ .

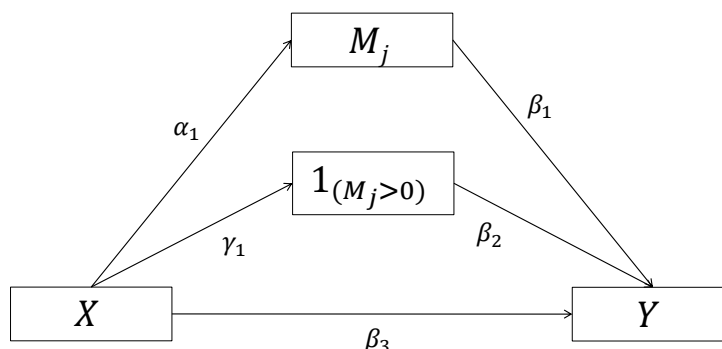


Figure 1. Potential causal mediation pathways of a zero-inflated mediator.

2.5. Sequential Ignorability Assumption

Mediation analyses require assumptions to make causal inference and there have been different forms of assumptions proposed in the literature [9,28–32]. The key of the assumptions is to identify the terms involving counterfactual outcomes so that they can be estimated with the observed data. One of the popular assumptions is the sequential ignorability assumption proposed in [28]. In the definition of NIE and NDE, the variable  $Y_{x_2 M_j(x_1)}$  is a counterfactual outcome because  $M_j(x_1)$  can not be observed when  $X$  takes the value of  $x_2$ . The sequential ignorability assumption [28] for identifying  $E(Y_{x_2 M_j(x_1)})$  can be written as follows in our setting:

$$\{Y_{x'm}, M_j(x)\} \perp\!\!\!\perp X|Z, \tag{9}$$

$$Y_{x'm} \perp\!\!\!\perp M_j(x)|X = x, Z, \tag{10}$$

where  $x'$  and  $x$  are any values in the support of  $X$ ,  $m$  is any value in the support of  $M_j$ , and  $Z$  is a vector of confounders (if any). The first assumption in Equation (9) says the outcome  $Y$  at any given value of the vector  $(X, M_j)$  and the mediator  $M_j$  at any given value of  $X$  should all be dependent of  $X$  conditional on confounders in  $Z$ . A randomized trial where  $X$  is the random assignment typically makes this assumption automatically satisfied. The second assumption in Equation (10) says the outcome  $Y$  at any given value

of the vector  $(X, M_j)$  is independent of the mediator  $M_j$  at  $X = x$  conditional on  $X = x$  and confounders in  $Z$ . The second assumption is essentially saying that the mediator  $M_j$  is effectively randomly assigned given  $X$  and  $Z$ . A straightforward interpretation for the first assumption is that there are no unmeasured confounders for the  $X - M_j$  association and the  $X - Y$  association. A straightforward interpretation for the second assumption is that there are no unmeasured confounders for the  $M_j - Y$  association. In our setting, the indicator variable  $1_{(M_j > 0)}$  is also considered as a mediator. Because it is completely determined by  $M_j$ , the above assumptions are enough to ensure the identifiability of  $E(Y_{x_2 M_j(x_1)})$  such that it can be estimated by the observed data.

### 3. Parameter Estimation

Maximum likelihood estimation (MLE) will be used to estimate the parameters. The data that is needed to estimate the marginal mediation effects for the  $j$ th taxon is  $(Y, R, M_j^*, L, X)$  where  $R = 1_{(M_j^* > 0)}$ . The estimation challenge is that  $M_j$  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 (8), we can still obtain their log-likelihood contributions by integrating the joint density function over all possible values of  $M_j$  using Riemann–Stieltjes integration [26]. Let  $(y_i, r_i, m_{ij}^*, l_i, x_i)$  denote the observed data values of  $(Y, R, M_j^*, L, X)$  for the  $i$ th subject in a study and  $m_{ij}$  denote the true value of the mediator  $M_j$  for the  $i$ th subject. We use  $i$  for subject index hereafter throughout the paper. The subjects can be divided into two groups by whether  $m_{ij}^*$  is non-zero and we derive the log-likelihood contribution for each group. The first group consists of subjects whose observed value of the mediator is non-zero (i.e.,  $m_{ij}^* > 0$ ). Based on the assumptions in the Equations (5)–(7) where  $\epsilon$  is assumed to have a normal distribution, the log-likelihood contribution from the  $i$ th subject (if it is in group 1) can be calculated as:

$$\begin{aligned} \ell_i^1 &= \ln(f(y_i, r_i | m_{ij}^*, x_i, l_i) f(m_{ij}^* | x_i, l_i)) = \ln(f(y_i | m_{ij}^*, x_i, l_i) p(r_i | m_{ij}^*, x_i, l_i) f(m_{ij}^* | x_i, l_i)) \\ &= \ln(f(y_i | m_{ij}^*, x_i, l_i)) + \ln(p(r_i | m_{ij}^*, l_i)) + \ln(f(m_{ij}^* | x_i, l_i)) \\ &= -0.5 \ln(2\pi) - \ln(\delta) - \frac{(y_i - \beta_0 - \beta_1 m_{ij}^* - \beta_2 - (\beta_3 + \beta_4)x_i - \beta_5 x_i m_{ij}^*)^2}{2\delta^2} \\ &\quad + \ln(1 - \Delta_i) - \ln\left(B(\mu_i \phi, (1 - \mu_i)\phi)\right) \\ &\quad + (\mu_i \phi - 1) \ln(m_{ij}^*) + ((1 - \mu_i)\phi - 1) \ln(1 - m_{ij}^*), \end{aligned}$$

where  $f(\cdot | m_{ij}^*, x_i, l_i)$ ,  $p(\cdot | m_{ij}^*, x_i, l_i)$  and  $f(\cdot | x_i, l_i)$  are the (conditional) density (or probability mass function) for  $Y$ ,  $R$  and  $M_j$  respectively,  $\Delta_i = \text{expit}(\gamma_0 + \gamma_1 x_i)$  and  $\mu_i = \text{expit}(\alpha_0 + \alpha_1 x_i)$ . Let  $F(m|x)$  denote the (conditional) cumulative distribution function for  $M_j$ . The second group consists of subjects with  $m_{ij}^* = 0$ . The log-likelihood contribution from the  $i$ th subject (if it is in group 2) can be calculated as:

$$\begin{aligned} \ell_i^2 &= \ln(f(y_i, r_i, m_{ij}^* | x_i)) = \ln\left(\int_{m \in [0,1]} f(y_i | m, x_i) p(r_i | m) dF(m | x_i)\right) \\ &= \ln\left(\frac{\Delta_i}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_3 x_i)^2}{2\delta^2}\right)\right. \\ &\quad \left.+ \int_0^{1/l_i} f(y_i | m, x_i) (1 - \Delta_i) \frac{m^{\mu_i \phi - 1} (1 - m)^{(1 - \mu_i)\phi - 1}}{B(\mu_i \phi, (1 - \mu_i)\phi)} dm\right) \\ &= -0.5 \ln(2\pi) - \ln(\delta) + \ln\left(\Delta_i \exp\left(-\frac{(y_i - \beta_0 - \beta_3 x_i)^2}{2\delta^2}\right) + \frac{1 - \Delta_i}{B(\mu_i \phi, (1 - \mu_i)\phi)} \int_0^{1/l_i} h_i(m) dm\right) \end{aligned}$$

where

$$h_i(m) = m^{\mu_i \phi - 1} (1 - m)^{(1 - \mu_i) \phi - 1} \times \exp \left( - \frac{(y_i - \beta_0 - \beta_1 m - \beta_2 - (\beta_3 + \beta_4)x_i - \beta_5 x_i m)^2}{2\delta^2} \right).$$

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

$$\ell = \sum_{i \in \text{group 1}} \ell_i^1 + \sum_{i \in \text{group 2}} \ell_i^2. \quad (11)$$

The MLE of the parameters can be obtained by maximizing the above complete log-likelihood function. With the parameter estimates and the observed Fisher information matrix, we will be able to calculate NIE, NIE<sub>1</sub>, NIE<sub>2</sub>, NDE and CDE and their CI's.

#### 4. Simulation

Extensive simulations were carried out to demonstrate the performance of our approach MarZIC in comparison with two existing approaches under two settings. In setting 1 where the mediator was generated by univariate ZIB distributions which is univariate version of Dirichlet distributions, we compared MarZIC with a current standard practice in causal mediation analyses developed by Imai, Keele and Tingley [10] (IKT approach hereafter) which is a PO approach and can be implemented in R using the package “mediation” [33]. The Marginal Structural Models [9] is also a standard PO approach with a very similar definition of indirect effect. These causal mediation analysis approaches were not developed to analyze microbiome data, and thus could have poor performance when applied to microbiome data. In setting 2 where the mediator was generated by multivariate zero-inflated Dirichlet-Multinomial distributions, MarZIC was compared with IKT and CCMM [6] which was 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  $\text{Ber}(0.5)$  such that the number of subjects was balanced between the two groups. To mimic the real study data, the library size was generated by randomly picking the library size with replacement from the real study data in Section 5 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 the real data. Multivariate delta method was used to derive confidence intervals in all settings.

##### 4.1. Simulation Setting 1: Univariate ZIB Distribution

In this setting, the outcome  $Y$  was assumed to be a continuous variable and generated using Equation (5) where  $\beta_5$  is set to be 0 and other true parameter values can be found in Table 1. Similar to simulation studies in the literature [18,19] where RA were generated individually, we generated individual taxon RA with ZIB distributions (i.e., univariate version of Dirichlet distributions) based on Equations (6) and (7). The LOD mechanism in Equation (8) for observing zero-valued data points of the mediator was used to generate false zeros for the mediator  $M_j$ . 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). We generated 100 random data sets for each scenario and the sample size was 200 for each data set. About 20% of all sequencing reads were generated as true zeros (i.e., structured zeros) in both scenarios. Under the LOD mechanism in Equation (8), 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 MarZIC 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 NIE<sub>1</sub> and NIE<sub>2</sub>. For Scenario 2 with high RA where there were no false zeros, MarZIC showed good performance again in terms of the performance measures. IKT also showed satisfactory performance for the estimation of the NIE 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 MarZIC and IKT with sample size of  $n = 200$ . 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.

Parameter /Effect	Low Relative Abundance (Mean = 0.0025)							High Relative Abundance (Mean = 0.5)						
	True	Mean	Bias	Bias	SE	Mean	CP(%)	True	Mean	Bias	Bias	SE	Mean	CP(%)
	Estimate		%		SE			Estimate		%		SE		
MarZIC														
NIE <sub>1</sub>	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
NIE <sub>2</sub>	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
$\beta_0$	−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
$\beta_1$	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
$\beta_2$	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
$\beta_3$	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
$\beta_4$	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
$\delta$	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
$\alpha_0$	−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
$\alpha_1$	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
$\zeta$	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
$\gamma_0$	−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
$\gamma_1$	−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. Simulation Setting 2: Multivariate Zero-Inflated Dirichlet-Multinomial Distribution

The subject index  $i$  is suppressed in this subsection for simplicity. The microbiome data was generated using a zero-inflated Dirichlet-multinomial model that can account for variability from both the Dirichlet distribution and the multinomial distribution. The microbiome data generation process can be found in Appendix C. As shown in Table 2, six different scenarios were considered, of which some had the number of taxa smaller than the sample size and the others had the number of taxa larger than the sample size. We generated 100 random data sets for each scenario and the sample size was 200 for each data set. The outcome  $Y$  was generated using the following equation:

$$Y = \beta_0 + \beta_{11}M_1 + \beta_{12}M_2 + \beta_2 1_{(M_2 > 0)} + \beta_3 X + \beta_4 X 1_{(M_2 > 0)} + \beta_5 X M_2 + \epsilon. \tag{12}$$

where  $M_1$  and  $M_2$  denote the RA of the first taxon and the second taxon respectively,  $(\beta_0, \beta_{11}, \beta_{12}, \beta_2, \beta_3, \beta_4, \beta_5) = (1, 80, 2, 3, 1, 1, 1)$  and  $\epsilon$  follows the standard normal distribution.

Notice that the data generation models are different from the analysis models in a few aspects: (a). The data generation model (12) involves both  $M_1$  and  $M_2$  which is different than the marginal model (5) where only one  $M_j$  is in the model; (b). The relationships between  $X$  and  $M_1$  and  $M_2$  in the data generation in Appendix C are different from the

data analysis model (6); (c). The zero mechanism for generating false zeros in the data generation as outlined in Appendix C is also different from the proposed mechanism in Section 2.3. Thus, to some extent, this simulation also demonstrated the robustness of our approach with respect to mis-specification of the model and the zero mechanism. Under the data generation model (12),  $Y$  has marginal associations with all taxa, but only the first two taxa marginally mediate the effect of  $X$  on  $Y$  because only their marginal mean values depend on  $X$  conditional on their presence according to the data generation in Appendix C. The indicator variable for the second taxon  $1_{(M_2>0)}$  also has a mediation effect because it has an impact on  $Y$  as shown in Equation (12) and the probability of presence of the second taxon depends on  $X$ . In summary,  $NIE_1$  should be significant for  $M_1$  and  $M_2$ , and  $NIE_2$  should be significant for  $M_2$  in the analysis results of this simulation. This setting also mimicked the real study case where there were only two OTU's with significant  $NIE_1$ .

Three indices were used to evaluate the model performance: Recall, Precision and F1 which 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  $TP = 0$ . F1 is the Harmonic mean [34] of Recall and Precision that measures the overall performance in terms of Recall and Precision. In the data analysis step of the simulation, MarZIC analyzed each taxon as a mediator one by one whereas CCMM employed  $\ell_1$  regularization to handle high dimensionality. Multiple testing was adjusted using the Benjamini-Hochberg Procedure [35] such that the targeted FDR is 20% for all approaches in this comparison which means that the targeted Precision should be around 80%.

The simulation results (See Table 2) showed that MarZIC had a very good overall performance for identifying  $NIE_1$  and  $NIE_2$  in terms of Recall (>77.5%), Precision (>87.2%) and F1 (>87.3%). MarZIC achieved the targeted Precision of 80% across all cases. CCMM had good performance in terms of Recall, but its Precision rates (38.8–52.4%) were much lower than the targeted Precision rate (80%) which resulted in low F1 values (55.3–66.1%). This suboptimal performance is likely due to: (a). CCMM was proposed to model the RA on log-scale whereas Equation (12) is on the original scale of RA, (b). CCMM was not developed to incorporate the mediation effect of the binary variable  $1_{(M_1>0)}$ , and (c). CCMM could not handle interactions between the independent variable and mediators such as  $X1_{(M_1>0)}$  in model (12). And CCMM could not generate any results for those scenarios with the number of taxa greater than or equal to 300 (See Table 2) due to computational issues whereas MarZIC can handle all cases very well. This is likely because CCMM is too computationally demanding for its  $\ell_1$  regularization algorithm which is not computationally capable of handling such high dimensionality. IKT had good Precision rates (>99.7%), but low recall rates (23.5–58.0%) compared to MarZIC, and thus also low F1 values.

In addition, we also considered cases with 5 taxa having significant  $NIE_1$  and one taxon having significant  $NIE_2$  and cases with 10 taxa having significant  $NIE_1$  and one taxon having significant  $NIE_2$ . The simulation results (See Table 3) also showed that MarZIC outperformed the other approaches. It had good recall rates for  $NIE_1$  (>85.3%) and  $NIE_2$  (>93%), and also achieved the target precision rate (80%) for both  $NIE_1$  and  $NIE_2$  except that it was 77.10%, slightly lower than 80%, for the case with 300 taxa of which 10 taxa had significant  $NIE_1$ . Its F1 values were also good for both  $NIE_1$  (>79.6%) and  $NIE_2$  (>86.6%). CCMM had fair recall (>66.0%), but much lower precision rate (19.0–66.2%) and therefore low F1 values (31.2–43.9%). IKT, on the other hand, achieved target precision rate for all cases (>99.1%), but low recall rate (29.3–66.2%), and thus low F1 values (44.3–78.2%).

**Table 2.** Simulation results for the comparison of MarZIC with CCMM and IKT. Here  $n$  denotes the sample size and  $K + 1$  denotes the number of taxa.

$K + 1$	$n$	Recall (%)				Precision (%)				F1 (%)			
		MarZIC (NIE <sub>1</sub> )	MarZIC (NIE <sub>2</sub> )	CCMM	IKT	MarZIC (NIE <sub>1</sub> )	MarZIC (NIE <sub>2</sub> )	CCMM	IKT	MarZIC (NIE <sub>1</sub> )	MarZIC (NIE <sub>2</sub> )	CCMM	IKT
10	200	99.00	100.00	100.00	58.00	97.70	98.00	38.80	99.70	97.90	98.60	55.30	68.10
25	200	99.50	100.00	96.00	39.50	98.20	99.50	52.40	100.00	98.50	99.60	66.10	48.30
50	200	97.50	100.00	97.00	44.00	100.00	100.00	46.40	100.00	98.30	100.00	60.60	54.70
100	200	96.00	98.90	100.00	32.50	95.50	100.00	42.80	100.00	94.50	98.90	58.00	41.30
300	200	86.00	97.80	-	25.00	90.80	99.50	-	100.00	85.80	97.50	-	31.30
500	200	77.50	94.70	-	23.50	97.80	87.20	-	99.00	83.00	87.30	-	30.00

Per the suggestion of a referee, we also did a simulation study with only 5 taxa (i.e.,  $K = 4$ ) in the data. The sample size was still 200 and the mean RA of the five taxa were approximately 0.196, 0.250, 0.220, 0.146 and 0.188 respectively. There were no false zeros because the five RA were large. The first two taxa had non-zero NIE<sub>1</sub> and the second taxon had non-zero NIE<sub>2</sub>. The simulation results from 100 random data sets showed good performance for both NIE<sub>1</sub> (Recall = 0.95, Precision = 0.96 and F1 = 0.94) and NIE<sub>2</sub> (Recall = 1, Precision = 0.97 and F1 = 0.98).

**Table 3.** Simulation results for the comparison of MarZIC with CCMM and IKT.

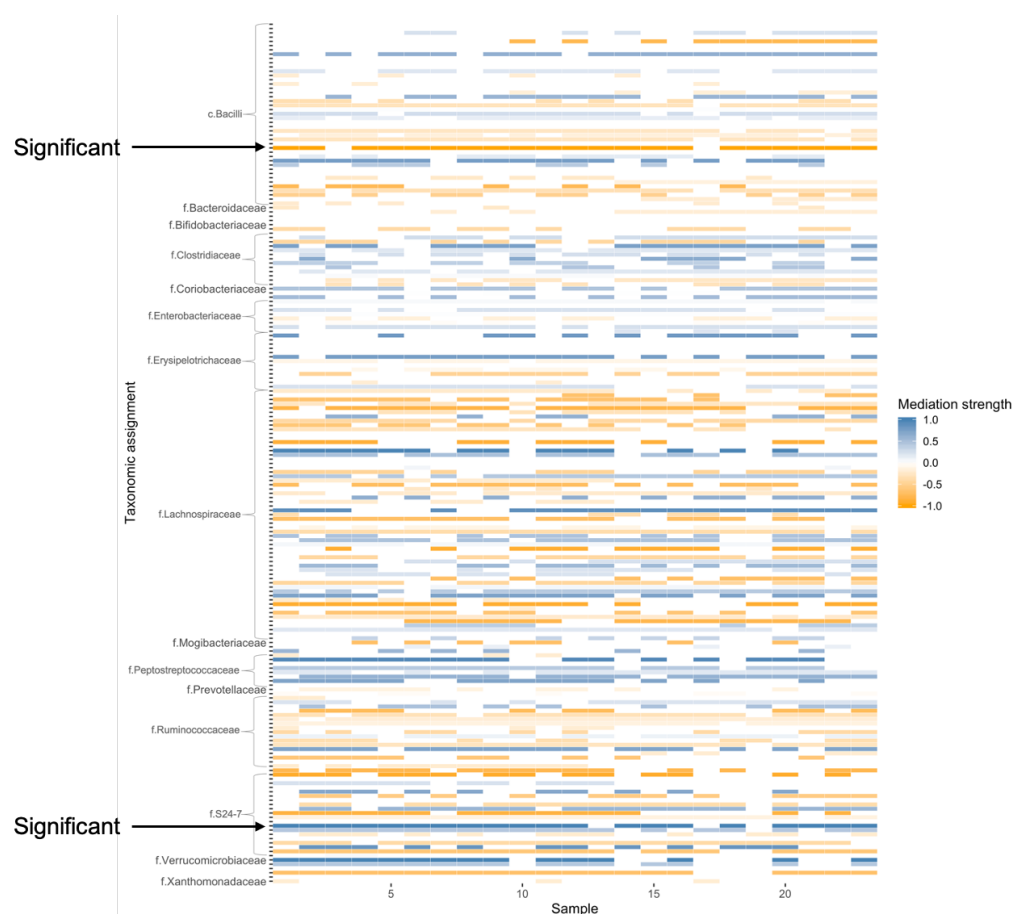
$K + 1$	Number of Taxa with Non-Zero NIE <sub>1</sub>	Recall (%)				Precision (%)				F1 (%)			
		MarZIC (NIE <sub>1</sub> )	MarZIC (NIE <sub>2</sub> )	CCMM	IKT	MarZIC (NIE <sub>1</sub> B)	MarZIC (NIE <sub>2</sub> )	CCMM	IKT	MarZIC (NIE <sub>1</sub> )	MarZIC (NIE <sub>2</sub> )	CCMM	IKT
50	5	95.00	100.00	89.00	66.20	99.00	98.50	27.90	99.60	96.60	99.00	42.20	78.20
50	10	95.70	92.00	66.00	62.40	98.80	91.80	33.20	99.60	97.10	86.20	43.90	75.70
100	5	96.60	99.00	89.40	60.60	92.70	98.30	19.00	99.10	94.10	97.80	31.20	73.30
100	10	92.10	91.00	80.10	46.00	93.70	97.80	27.20	100.00	92.50	89.50	40.40	61.20
300	5	94.20	96.00	-	56.10	80.50	97.00	-	99.70	85.20	94.00	-	69.90
300	10	85.30	93.00	-	29.30	77.10	91.00	-	99.60	79.60	86.60	-	43.40

## 5. Real Study Application

VSL#3 is a commercially available probiotic cocktail (Sigma-Tau Pharmaceuticals, Inc., Gaithersburg, MD, USA) 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 [36] and ulcerative colitis [37]. Preventive VSL#3 administration can also attenuate colitis in I10−/− mice [38] and ileitis in SAMP1/YitFc mice [39]. When used as a preventative strategy, it has the potential capability to prevent inflammation and carcinogenesis. In a mouse model, Arthur et al. [40] 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. [40] and generated open reference OTUs using the Quantitative Insights into Microbial Ecology (QIIME) [41] 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 [42]. Taxonomic assignment was done using the RDP (ribosomal database project) classifier [43] 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.

The relative abundance (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) which is an ordinal categorical variable measuring the abnormality of cell growth and it is treated as a continuous variable in the analysis because of its ordinal nature and its roughly bell-shaped density curve. 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.  $NIE_1$  of two OTUs were found to be statistically significant. The first OTU was assigned to the family S24-7 under order Bacteroidales and the second one was assigned to class Bacilli. 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 interpretation for the mediation effects are that the treatment had a marginal positive effect of 0.27 on the dysplasia score through changing the RA of the first OTU and it also had a marginal negative effect of  $-1.28$  on the dysplasia score through changing the RA of the second OTU. The family S24-7 and class Bacilli found by our approach have also been reported to be related with colorectal cancer in the literature [44,45]. To give a full picture of the mediation effects in this data set, a heatmap based on  $p$ -values was constructed (see Figure 2) to illustrate the  $NIE_1$  of all OTUs. CCMM and IKT did not find any significant mediation effects of the OTUs.



**Figure 2.** 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 left blank in the heatmap.

## 6. Discussion

We developed an innovative marginal mediation modeling approach under the PO framework to analyze zero-inflated compositional mediators such as microbiome. We showed that the marginal 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. The marginal model approach can also account for the compositional structure. When the point mass at zero (i.e.,  $\Delta$ ) is equal to zero for the mediator (i.e., the distribution is not zero-inflated), the model reduces to a marginal mediation model for data without zeros as described in Section 2.1. Therefore, this approach can be also used for data sets after zero-valued data points are imputed with a positive number such as a Pseudocount (or after other normalization techniques are applied). 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  $NIE = E(Y_{x_2 M_j(x_2)} - Y_{x_2 M_j(x_1)} | C = c)$ ,  $NDE = E(Y_{x_2 M_j(x_1)} - Y_{x_1 M_j(x_1)} | C = c)$  and  $CDE = E(Y_{x_2 m} - Y_{x_1 m} | C = c)$ . 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 [46]. CI can be obtained using the delta method or resampling methods. Decomposition of NIE follows the same procedure as shown in Section 2.4.

Misspecification of the mechanisms for observing zero-valued data points could have an impact on the model performance. This is similar to missing data issues where partial information is available on the missing data. It can be considered as missing not at random (MNAR) [47] because the probability of a data point being observed as zero depends on its true value. Besides the LOD mechanism in Equation (8), another possible mechanism could be  $P(M_j^* = 0 | M_j, L) = \exp(-\eta M_j / L)$  where  $\eta > 0$ . Model selection approaches such BIC or AIC can be used to choose the optimal mechanism among 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.

**Author Contributions:** Conceptualization, Q.W. and Z.L.; methodology, Q.W., J.O. and Z.L.; formal analysis, Q.W. and Z.L.; investigation, Q.W., J.O., S.D., M.R.K., M.O.C., A.G.H., B.C.C., J.C.M. and Z.L.; data curation, R.Z.G. and C.J.; writing—original draft preparation, Q.W., J.O., S.D., M.R.K., M.O.C., A.G.H., B.C.C., J.C.M. and Z.L.; supervision, Z.L.; funding acquisition, M.R.K. and Z.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by U.S. NIH Grants R01GM123014, UH3OD023275, P01ES022832 and P20GM104416, and U.S. EPA grant RD 83544201.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset analyzed in this paper is available upon request.

**Acknowledgments:** We thank Janaka Liyanage for helpful assistance.

**Conflicts of Interest:** The authors declare no conflict of interest.

### Appendix A. Marginal Association between $Y$ and $M_j$

Subject index  $i$  is again suppressed in this section for simplicity. To obtain the marginal association between  $Y$  and  $M_j$  under Equation (1), we derive the expression for the conditional expectation  $E_X(Y | M_j)$  which is the mean of  $Y$  conditional on  $M_j$  given  $X$ . By following basic principles of calculating conditional expectations, we have:

$$\begin{aligned}
 E_X(Y|M_j) &= E_X\left(\sum_{k=1}^{K+1} \beta^k M_k + \beta^X X + \sum_{k=1}^{K+1} \beta^{kk} X M_k + \epsilon \mid M_j\right) \\
 &= \sum_{k=1}^{K+1} \beta^k E_X(M_k|M_j) + \beta^X X + \sum_{k=1}^{K+1} \beta^{kk} X E_X(M_k|M_j) + E_X(\epsilon \mid M_j) \\
 &= \sum_{k=1}^{K+1} \beta^k E_X(M_k|M_j) + \beta^X X + \sum_{k=1}^{K+1} \beta^{kk} X E_X(M_k|M_j).
 \end{aligned}
 \tag{A1}$$

Next we need to derive the expression for  $E_X(M_k|M_j)$  for all  $k = 1, \dots, K + 1$  in the above equation. It is trivial to see that  $E_X(M_j|M_j) = M_j$ . Let  $M_{-j}$  denote the vector containing all but  $M_j$  and thus  $M_{-j} = (M_1, \dots, M_{j-1}, M_{j+1}, \dots, M_{K+1})^T$ . Since  $M$  has a Dirichlet distribution, the subcomposition  $\frac{M_{-j}}{1-M_j}$  conditional on  $M_j$  follows another Dirichlet distribution [48] with the mean parameters being  $\left(\frac{\mu_1}{\sum_{k \neq j} \mu_k}, \dots, \frac{\mu_{j-1}}{\sum_{k \neq j} \mu_k}, \frac{\mu_{j+1}}{\sum_{k \neq j} \mu_k}, \dots, \frac{\mu_{K+1}}{\sum_{k \neq j} \mu_k}\right)$  and the dispersion parameter being  $\phi \sum_{k \neq j} \mu_k$ . Thus, for any  $M_k$  in the subvector  $M_{-j}$ , we have

$$\begin{aligned}
 E_X(M_k|M_j) &= E_X\left((1 - M_j) \frac{M_k}{1 - M_j} \mid M_j\right) \\
 &= (1 - M_j) E_X\left(\frac{M_k}{1 - M_j} \mid M_j\right) \\
 &= (1 - M_j) \frac{\mu_k}{\sum_{l \neq j} \mu_l}.
 \end{aligned}$$

By plugging the above results into Equation (A1), we have

$$\begin{aligned}
 E_X(Y|M_j) &= \sum_{k=1}^{K+1} \beta^k E_X(M_k|M_j) + \beta^X X + \sum_{k=1}^{K+1} \beta^{kk} X E_X(M_k|M_j) \\
 &= \beta^j M_j + \sum_{k \neq j} \beta^k (1 - M_j) \frac{\mu_k}{\sum_{l \neq j} \mu_l} + \beta^X X + \beta^{jj} X M_j + \sum_{k \neq j} \beta^{kk} X (1 - M_j) \frac{\mu_k}{\sum_{l \neq j} \mu_l} \\
 &= \beta_0^* + \beta_1^* M_j + \beta_2^* X + \beta_3^* X M_j,
 \end{aligned}$$

where

$$\beta_0^* = \frac{\sum_{k \neq j} \beta^k \mu_k}{\sum_{l \neq j} \mu_l}, \quad \beta_1^* = \beta^j - \beta_0^*, \quad \beta_2^* = \beta^X + \frac{\sum_{k \neq j} \beta^{kk} \mu_k}{\sum_{l \neq j} \mu_l}, \quad \text{and} \quad \beta_3^* = \beta^{jj} - \frac{\sum_{k \neq j} \beta^{kk} \mu_k}{\sum_{l \neq j} \mu_l}.$$

**Appendix B. Multivariate Delta Method for Obtaining 95% CI of NIE<sub>1</sub>, NIE<sub>2</sub>, NDE and CDE**

Let  $\zeta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \delta, \alpha_0, \alpha_1, \gamma_0, \gamma_1)^\top$ . The formulas for NIE<sub>1</sub>, NIE<sub>2</sub>, NIE, NDE and CDE can be considered as functions of the full parameter vector  $\zeta$ . Let  $f_1(\zeta) = \text{NIE}_1$  as derived in Section 2.4 and thus  $f_1(\hat{\zeta})$  is the MLE of NIE<sub>1</sub> where  $\hat{\zeta}$  is the MLE of  $\zeta$ . We first calculate the observed Fisher information matrix which can be calculated as  $I_{obs} = -\frac{\partial^2 \ell}{\partial \zeta \partial \zeta^\top} \Big|_{\zeta = \hat{\zeta}}$  where  $\ell$  is the loglikelihood function in Equation (11). By using the multivariate Delta method, we can calculate the variance of the estimator as follows:

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

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

### Appendix C. Microbiome Data Generation Process for Simulation Setting 2

Let  $K + 1$  be the number of taxa, for the  $i$ th subject, the microbiome data generation steps are listed below:

**Step 1:** Generate true zeros for all taxa. The zeros for a taxon were generated using a Bernoulli distribution,  $Ber(\Delta)$ , with  $\Delta$  given in Equation (7). So the probability of the taxon being 0 is equal to  $\Delta$ . For taxon 1, we set  $\Delta = 0$  so that there's no zero in taxon 1. For taxon 2, we set  $\gamma_0 = 1$  and  $\gamma_1 = -3$  in Equation (7) for  $\Delta$ . For all the other taxa,  $\gamma_0$  were generated from  $U(1, 2)$  and  $\gamma_1 = 0$  in Equation (7) for  $\Delta$ . So only the absence (or presence) of taxon 2 was associated with  $X$ . The total percentage of zeros was between 68.8% and 81.6% with  $K + 1$  ranging from 10 to 500, which indicates high data sparsity.

**Step 2:** Generate RA for the non-zero taxa from a Dirichlet distribution. Assume we had  $P$  non-zero taxa (from Step 1) indexed by  $(t_1, t_2, \dots, t_p)$  in the ascending order meaning  $t_1 < \dots < t_p$ . Here  $t_1 = 1$  since the first taxon does not have any zeros from Step 1. The RA of those non-zero taxa was generated by the  $P$ -dimensional Dirichlet distribution with the dispersion parameter  $\phi$  and mean parameters  $(\mu_{t_1}, \dots, \mu_{t_p})$  that satisfies  $\sum_{p=1}^P \mu_{t_p} = 1$ . The dispersion parameter  $\phi$  was set to be 50 to mimic the over-dispersion in real data. The values of mean parameters were chosen in a way such that it has smaller values for taxa with larger  $\Delta$ 's in Step 1 so that taxa with lower abundance are more likely to have zeros. More specifically, the mean parameters were determined as follows:

If  $t_2 = 2$  (i.e., taxa 2 is one of the non-zero taxa):

$$\begin{aligned} \mu_{t_1} &= \frac{\exp(\alpha_0^1)}{\sum_{k=1}^{K+1} \exp(\alpha_0^k)} \times \frac{1}{1 + \exp(a_0 + a_1 X)}, \\ \mu_{t_2} &= \frac{\exp(\alpha_0^1)}{\sum_{k=1}^{K+1} \exp(\alpha_0^k)} \times \frac{\exp(a_0 + a_1 X)}{1 + \exp(a_0 + a_1 X)}, \\ \mu_{t_p} &= \left( 1 - \frac{\exp(\alpha_0^1)}{\sum_{k=1}^{K+1} \exp(\alpha_0^k)} \right) \times \frac{\exp(\alpha_0^{t_p})}{\sum_{p=3}^P \exp(\alpha_0^{t_p})}, \quad p \in \{3, \dots, P\}, \end{aligned}$$

where  $a_0 = -2$ ,  $a_1 = 5$ ,  $\alpha_0^1$  was set to be the value such that the false zeros for taxon 2 that will be generated in next step will be around 20%, and  $\alpha_0^k, k \in \{2, \dots, K + 1\}$  were generated from  $U(0, 1)$

If  $t_2 > 2$  (i.e., taxa 2 is not one of the non-zero taxa):

$$\begin{aligned} \mu_{t_1} &= \frac{\exp(\alpha_0^1)}{\sum_{k=1}^{K+1} \exp(\alpha_0^k)}, \\ \mu_{t_p} &= \left( 1 - \frac{\exp(\alpha_0^1)}{\sum_{k=1}^{K+1} \exp(\alpha_0^k)} \right) \times \frac{\exp(\alpha_0^{t_p})}{\sum_{p=2}^P \exp(\alpha_0^{t_p})}, \quad p \in \{2, \dots, P\}. \end{aligned}$$

Under this data generation, only the RA of taxa 1 and taxa 2 (conditional on presence) were dependent on  $X$ . The RA of taxa 3 to taxa  $K$  were independent of  $X$ .

**Step 3:** Generate sample absolute abundance (SAA) and false zeros from a multinomial distribution. Let  $(\mathcal{R}_{t_1}, \dots, \mathcal{R}_{t_p})$  denote the RA generated in Step 2 for the  $P$  non-zero taxa and thus  $\sum_{p=1}^P \mathcal{R}_{t_p} = 1$ . The  $P$ -dimensional multinomial distribution with the parameter vector  $(\mathcal{R}_{t_1}, \dots, \mathcal{R}_{t_p})$  and the library size (randomly selected from real data) was used to generate SAA for all the  $P$  non-zero taxa. Those taxa with SAA = 0 generated from the multinomial distribution are false zeros.

**Step 4:** Getting final RA for all non-zero taxa. After SAA were generated for all non-zero taxa in Step 3, the SAA were divided by the library size to get the final RA for all non-zero taxa.

**Step 5:** Repeat the above Steps 1–4 for each subject to get a full data set of microbiome data for 200 subjects.

## References

1. Belkaid, Y.; Hand, T.W. Role of the microbiota in immunity and inflammation. *Cell* **2014**, *157*, 121–141. [[CrossRef](#)] [[PubMed](#)]
2. Wang, X.; Sun, G.; Feng, T.; Zhang, J.; Huang, X.; Wang, T.; Xie, Z.; Chu, X.; Yang, J.; Wang, H.; et al. Sodium oligomannate therapeutically remodels gut microbiota and suppresses gut bacterial amino acids-shaped neuroinflammation to inhibit Alzheimer's disease progression. *Cell Res.* **2019**, *29*, 787–803. [[CrossRef](#)] [[PubMed](#)]
3. Jin, C.; Lagoudas, G.K.; Zhao, C.; Bullman, S.; Bhutkar, A.; Hu, B.; Ameh, S.; Sandel, D.; Liang, X.S.; Mazzilli, S.; et al. Commensal Microbiota Promote Lung Cancer Development via GammaDelta T Cells. *Cell* **2019**, *176*, 998–1013.e16. [[CrossRef](#)] [[PubMed](#)]
4. Tanoue, T.; Morita, S.; Plichta, D.R.; Skelly, A.N.; Suda, W.; Sugiura, Y.; Narushima, S.; Vlamakis, H.; Motoo, I.; Sugita, K.; et al. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **2019**, *565*, 600–605. [[CrossRef](#)]
5. Li, H. Statistical and Computational Methods in Microbiome and Metagenomics. *Handb. Stat. Genom.* **2018**.
6. Sohn, M.B.; Li, H. Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* **2019**, *13*, 661–681. [[CrossRef](#)]
7. Wang, C.; Hu, J.; Blaser, M.J.; Li, H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics* **2019**, *36*, 347–355. [[CrossRef](#)]
8. Zhang, H.; Chen, J.; Li, Z.; Liu, L. Testing for mediation effect with application to human microbiome data. *Stat. Biosci.* **2019**, *in press*.
9. VanderWeele, T.J. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **2009**, *20*, 18–26. [[CrossRef](#)]
10. Imai, K.; Keele, L.; Tingley, D. A General Approach to Causal Mediation Analysis. *Psychol. Methods* **2010**, *15*, 309–334. [[CrossRef](#)]
11. VanderWeele, T.J. *Explanation in Causal Inference: Methods for Mediation and Interaction*; Oxford University Press: New York, NY, USA, 2015.
12. Baron, R.M.; Kenny, D.A. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *J. Personal. Soc. Psychol.* **1986**, *51*, 1173–1182. [[CrossRef](#)]
13. MacKinnon, D.P. *Introduction to Statistical Mediation Analysis*; Erlbaum: New York, NY, USA, 2008.
14. MacKinnon, D.P.; Fairchild, A.J.; Fritz, M.S. Mediation analysis. *Annu. Rev. Psychol.* **2007**, *58*, 593–614. [[CrossRef](#)] [[PubMed](#)]
15. VanderWeele, T.J. Mediation Analysis: A Practitioner's Guide. *Annu. Rev. Public Health* **2016**, *37*, 17–32. [[CrossRef](#)] [[PubMed](#)]
16. Lange, T.; Hansen, K.W.; Sørensen, R.; Galatius, S. Applied mediation analyses: A review and tutorial. *Epidemiol. Health* **2017**, *39*, e2017035. [[CrossRef](#)] [[PubMed](#)]
17. Dalrymple, M.L.; Hudson, I.L.; Ford, R.P.K. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Comput. Stat. Data Anal.* **2003**, *41*, 491–504. [[CrossRef](#)]
18. Chai, H.; Jiang, H.; Lin, L.; Liu, L. A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput. Biol.* **2018**, *14*, e1006329. [[CrossRef](#)]
19. Chen, E.Z.; Li, H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **2016**, *32*, 2611–2617. [[CrossRef](#)]
20. Tang, Z.Z.; Chen, G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **2018**, *20*, 698–713. [[CrossRef](#)]
21. Peng, X.; Li, G.; Liu, Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *J. Comput. Biol.* **2016**, *23*, 102–110. [[CrossRef](#)]
22. Chen, J.; Li, H. Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *Ann. Appl. Stat.* **2013**, *7*, 418–442. doi: 10.1214/12-AOAS592. [[CrossRef](#)]
23. Martin, B.D.; Witten, D.; Willis, A.D. Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression. *Ann. Appl. Stat.* **2020**, *14*, 94–115. [[CrossRef](#)]
24. Ferrari, S.; Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [[CrossRef](#)]
25. Cribari-Neto, F.; Zeileis, A. Beta Regression in R. *J. Stat. Softw.* **2010**, *34*, 24848. [[CrossRef](#)]
26. Terhorst, H.J. On Stieltjes Integration in Euclidean-Space. *J. Math. Anal. Appl.* **1986**, *114*, 57–74. [[CrossRef](#)]
27. Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat. Sci.* **1986**, *1*, 54–75. [[CrossRef](#)]
28. Imai, K.; Keele, L.; Yamamoto, T. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Stat. Sci.* **2010**, *25*, 51–71. [[CrossRef](#)]
29. Pearl, J. Direct and indirect effects. In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, 26–29 August 2001; Breese, J., Koller, D., Eds.; Morgan Kaufman: San Francisco, CA, USA, 2001; pp. 411–420.
30. Robins, J. Semantics of causal DAG models and the identification of direct and indirect effects. In *Proceedings of the Highly Structured Stochastic Systems*; Green, P., Hjort, N., Richardson, S., Eds.; Oxford University Press: Oxford, UK, 2003; pp. 70–81.



31. Peterson, M.; Sinisi, S.; van der Laan, M. Estimation of Direct Causal Effects. *Epidemiology* **2006**, *17*, 276–284. [[CrossRef](#)]
32. Hafeman, D.M.; VanderWeele, T.J. Alternative Assumptions for the Identification of Direct and Indirect Effects. *Epidemiology* **2011**, *22*, 753–764. [[CrossRef](#)] [[PubMed](#)]
33. Tingley, D.; Yamamoto, T.; Hirose, K.; Keele, L.; Imai, K. mediation: R Package for Causal Mediation Analysis. 2017. Available online: <https://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf> (accessed on 6 June 2022).
34. Martinez, M.N.; Bartholomew, M.J. What does it “mean”? A review of interpreting and calculating different types of means and standard deviations. *Pharmaceutics* **2017**, *9*, 14. [[CrossRef](#)]
35. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A Practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **1995**, *57*, 289–300. [[CrossRef](#)]
36. Gionchetti, P.; Rizzello, F.; Venturi, A.; Brigidi, P.; Matteuzzi, D.; Bazzocchi, G.; Poggioli, G.; Miglioli, M.; Campieri, M. Oral bacteriotherapy as maintenance treatment in patients with chronic pouchitis: A double-blind, placebo-controlled trial. *Gastroenterology* **2000**, *119*, 305–309. [[CrossRef](#)]
37. Sood, A.; Midha, V.; Makharia, G.K.; Ahuja, V.; Singal, D.; Goswami, P.; Tandon, R.K. The probiotic preparation, VSL# 3 induces remission in patients with mild-to-moderately active ulcerative colitis. *Clin. Gastroenterol. Hepatol.* **2009**, *7*, 1202–1209. [[PubMed](#)]
38. Madsen, K.; Cornish, A.; Soper, P.; McKaigney, C.; Jijon, H.; Yachimec, C.; Doyle, J.; Jewell, L.; De Simone, C. Probiotic bacteria enhance murine and human intestinal epithelial barrier function. *Gastroenterology* **2001**, *121*, 580–591. [[CrossRef](#)] [[PubMed](#)]
39. Pagnini, C.; Saeed, R.; Bamias, G.; Arseneau, K.O.; Pizarro, T.T.; Cominelli, F. Probiotics promote gut health through stimulation of epithelial innate immunity. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 454–459. [[CrossRef](#)] [[PubMed](#)]
40. Arthur, J.C.; Gharaibeh, R.Z.; Uronis, J.M.; Perez-Chanona, E.; Sha, W.; Tomkovich, S.; Mühlbauer, M.; Fodor, A.A.; Jobin, C. VSL# 3 probiotic modifies mucosal microbial composition but does not reduce colitis-associated colorectal cancer. *Sci. Rep.* **2013**, *3*, 2868. [[PubMed](#)]
41. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.D.; Costello, E.K.; Fierer, N.; Pena, A.G.; Goodrich, J.K.; Gordon, J.I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335. [[CrossRef](#)] [[PubMed](#)]
42. Bokulich, N.A.; Subramanian, S.; Faith, J.J.; Gevers, D.; Gordon, J.I.; Knight, R.; Mills, D.A.; Caporaso, J.G. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **2013**, *10*, 57. [[CrossRef](#)] [[PubMed](#)]
43. Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–5267. [[CrossRef](#)]
44. Peters, B.A.; Dominianni, C.; Shapiro, J.A.; Church, T.R.; Wu, J.; Miller, G.; Yuen, E.; Freiman, H.; Lustbader, I.; Salik, J.; et al. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome* **2016**, *4*, 69. [[CrossRef](#)]
45. Bråten, L.S.; Sødning, M.; Paulsen, J.E.; Snipen, L.G.; Rudi, K. Cecal microbiota association with tumor load in a colorectal cancer mouse model. *Microb. Ecol. Health Dis.* **2017**, *28*, 1352433. [[CrossRef](#)]
46. Gianola, D. Least-Squares Means Vs Population Marginal Means. *Am. Stat.* **1982**, *36*, 65–66.
47. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2014; Volume 333.
48. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1982**, *44*, 139–177. [[CrossRef](#)]