mathematics

MDPI

Article

Estimation of Mediation Effect on Zero-Inflated Microbiome Mediators

Dongyang Yang [1] and Wei Xu [1,2,*]

[1] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada; dongyang.yang@mail.utoronto.ca
[2] Department of Biostatistics, Princess Margaret Cancer Centre, Toronto, ON M5G 2C1, Canada
[*] Correspondence: wei.xu@uhnresearch.ca

Abstract: The mediation analysis methodology of the cause-and-effect relationship through mediators has been increasingly popular over the past decades. The human microbiome can contribute to the pathogenesis of many complex diseases by mediating disease-leading causal pathways. However, standard mediation analysis is not adequate for microbiome data due to the excessive number of zero values and the over-dispersion in the sequencing reads, which arise for both biological and sampling reasons. To address these unique challenges brought by the zero-inflated mediator, we developed a novel mediation analysis algorithm under the potential-outcome framework to fill this gap. The proposed semiparametric model estimates the mediation effect of the microbiome by decomposing indirect effects into two components according to the zero-inflated distributions. The bootstrap algorithm is utilized to calculate the empirical confidence intervals of the causal effects. We conducted extensive simulation studies to investigate the performance of the proposed weighting-based approach and some model-based alternatives, and our proposed model showed robust performance. The proposed algorithm was implemented in a real human microbiome study of identifying whether some taxa mediate the relationship between LACTIN-V treatment and immune response.

Keywords: mediation model; microbiome data; zero-inflation; semiparamatric; direct effects; indirect effects

MSC: Primary 62H12; Secondary 62P10

check for updates

Citation: Yang, D.; Xu, W. Estimation of Mediation Effect on Zero-Inflated Microbiome Mediators. *Mathematics* **2023**, *11*, 2830. https://doi.org/10.3390/math11132830

Academic Editors: Zijiang Yang and Xin Gao

Received: 24 May 2023
Revised: 21 June 2023
Accepted: 22 June 2023
Published: 24 June 2023

1. Introduction

The methodology used for mediation analysis of the cause-and-effect relationship through intermediate variables, also known as mediators, has been increasingly popular over the past decades. Mediation considers how a third variable affects the explored variable and outcome variable in a causal pathway. In the mediation analysis framework, the relationship between the exposure and outcome is called a direct effect, and the causal effect through the mediator is named an indirect effect. The topic of causal inference, which particularly involves mediating factors, was initially established by social scientists and psychologists in their fields. In recent years, mediation methodologies have been expanded to epidemiology and healthcare research to better understand how treatment effects occur or which interventions can change the outcome of interest by targeting specific mediators. The mediators can be one or more, and both direct and indirect effects can be estimated and tested in a counterfactual approach by modeling covariance and correlation matrices [1–5].

The approaches to mediation analysis involve traditional methodologies such as the difference method and the product method using the structural equation modeling (SEM) method [6]. SEM is derived from path analysis, which allows latent variables and confounders to be incorporated into the analysis of mediating relationships [7]. Both SEM and path analysis require adequate linear model specification. As a result, the counterfactual framework was developed in the past decade to model mediator or outcome variable without the linearity assumptions [8]. Within the counterfactual framework, causal inference analysis models observed and potential outcomes in which the unobserved outcomes are treated as missing data. This approach is more flexible in terms of incorporating nonlinear models such as hazard survival and marginal structural models (MSMs) for estimation of the indirect effects [3,9–11]. In addition to the relaxation of linearity assumptions, the interaction effects of the exposure and the mediator are also allowed.

The field of microbes concentrates on the relationships among the microbiome and its host, showing profound connections between the microbiome with human health. Studies to understand the effects of the microbiome on health status have been conducted. Evidence has shown that the gut microbiota is associated with many chronic diseases such as obesity [12–15], cardiovascular disease [16], inflammatory bowel disease [17–19], diabetes [20,21], fatty liver disease [22], liver cirrhosis [23], and colorectal cancer [24]. The microbiome composition can be quantified using 16S rRNA [25,26] or shotgun sequencing technology [27,28] or quantification of microbial absolute abundance differences (QWDs) [29] as markers to categorize organisms into taxonomic groups with specific taxonomic identity, which are called operational taxonomic units (OTUs). Researchers have shown that OTUs are usually skewed and heavy-tailed with excessive zeros [30–32].

Although mediation frameworks have been well established, there is a lack of research on the applications of the microbiome as a mediator. Classical models such as linear regression and logistic regression models cannot accommodate the features of microbiome data without sacrificing the information on the zero parts for microbiome sequencing data and violate the normality and constant variance assumptions. Recent studies on the human microbiome have shown its potential mediation effects between risk factors and human diseases. For example, Sohn and Li [33] proposed a sparse compositional mediation framework to investigate the mediated effect of gut microbiome between fat intake and body mass index. They directly established a compositional mediation model in the simplex space and adopted the additive log-ratio transformation on the composition. Zhang [34] mainly used linear structural equation models to model the mediator and the outcome. Considering the compositional features, they used the isometric log-ratio transformed mediators to construct a mediation model in Euclidean space and applied methods that can be used in Euclidean space. They studied the relative mediation effect denoted by a specific composition in contrast to the rest of the compositions. These two studies are specific for relative-abundance microbiome data using a structural equation modeling framework. Other researchers have used counterfactual frameworks, such as Wu et al. [35], who established a mediation analysis approach with the consideration of a zero-inflated mediator using a model-based standardization technique. They assumed that the mediator followed a zero-inflated beta distribution for relative abundance. For the outcome model, they separately modeled the count part and the zero part using an indicator. Zhang et al. [36] used inverse probability weighing-based mediation analysis for a specific measure called the interventional indirect effect.

Regarding current microbiome research, there is a shortage of causal framework models to understand how the exposure affects the disease outcome through the microbiome. Although the aforementioned methods can estimate the mediation effect of microbiome features, they parametrically estimate the mediation effects, which strongly relies on model assumptions. In addition, the existing studies have all focused on the relative abundance of the microbiome rather than count data. Our objective was to develop a semiparametric causal modeling framework to formulate a zero-inflated mediation model for microbiome count data. We aimed to estimate the mediation effect of the microbiome by simultaneously modeling the zero part and the count part of the mediator using semiparametric estimators. This method can avoid the specification of an outcome model and exposure-covariate interactions and can more easily identify positivity violations.

This paper is structured as follows: We provide the background of the counterfactual framework and its notations and assumptions at the beginning of Section 2. Models for the microbiome such as zero-inflated models are introduced in Section 2.1, followed by the proposed model in Section 2.2. Statistical estimation procedures are provided in Section 2.3. Our simulation studies to assess the performance of the proposed model in comparison with the existing approaches are described in Sections 3.1 and 3.2, followed by a real-data application of the proposed model and the comparison models in Section 3.3, with a discussion in Section 4.

2. Materials and Methods

For mediation analysis, interest is usually focused on separating the indirect pathway from the total causal effect between exposure and outcome. Counterfactual variables [37] are introduced to better estimate the effects. Directed acyclic graphs (DAGs) are intuitive ways to describe a causal structure. An example is provided in Figure 1, where the variables defined in the graph are used below to define the models. Let A_i denote the exposure for subject i, $Y_i(a,m)$ denote the outcome achieved for subject i if the subject was exposed to level a and mediator level m, and $M_i(a)$ denote the mediator achieved for subject i if the exposure level is a. When the mediator is randomized, nested counterfactuals are defined as $Y(a,M(a^*))$. When $a=a^*$, the counterfactual variable is simply the observations.

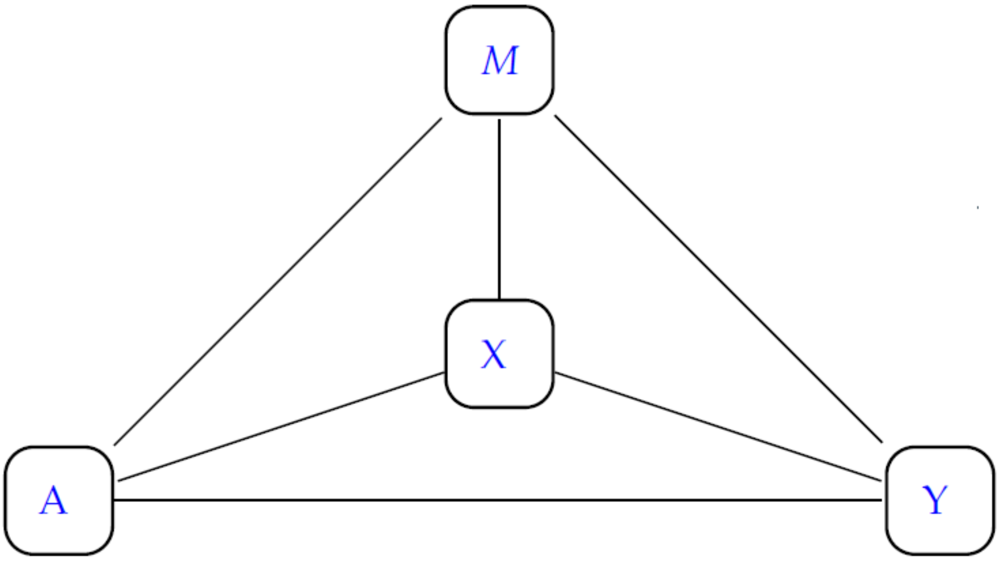

Figure 1. Generic directed acyclic graph for mediation analysis. X, confounder; A, exposure; M, mediator; Y, outcome.

We define the causal effects of interest using counterfactual variable notations. The effect that one would observe is referred to as the total effect (TE) of an exposure, and it is defined as the difference between $Y(1,M(1))$ and $Y(0,M(0))$ if the exposure is binary. More generally, we can define $TE(a,a^*) = E[Y(a) - Y(a^*)]$. When the mediator is set to 0 and the exposure level is changed, the difference between $Y(1,M(0))$ and $Y(0,M(0))$ is referred as a natural direct effect $NDE(a,a^*) = E[Y(a,M(a))] - E[Y(a^*,M(a))]$. Likewise, the natural indirect effect ocurrs when fixing the exposure level and changing the mediator, i.e., $Y(1,M(1)) - Y(1,M(0))$ or $NIE(a,a^*) = E[Y(a^*,M(a))] - E[Y(a^*,M(a^*))]$. The total effect is the sum of the natural direct effect (NDE) and natural indirect effect (NIE):

$$TE = NDE + NIE.$$

In any mediation analysis or causal inference, the causal mediation interpretation relies on a number of assumptions, and most of them are untestable.

Sequential Ignorability: There is no unmeasured confounding of the exposure–outcome relationship, the exposure–mediator relationship, or the mediator–outcome relationship. Mathematically, we have:

$$Y(a, M(a)) \perp A|X, \qquad M(a) \perp A|X, \qquad Y(a, m) \perp M|(A, X)$$

Consistency: The consistency assumption is that the counterfactuals take the observed values when the risk factor or treatment and mediator are actively set to the values they would have had. Mathematically, we have

$$P(Y(A, M) = Y) = 1 \quad and \quad P(M(A) = M) = 1$$

Positivity: All levels of exposure and mediator have a nonzero probability for any values of the confounders. Mathematically, we have

$$P(A = a|X = x) > 0 \quad \forall a, c \qquad P(M = m|X = x, A = a) > 0 \quad \forall a, x, m$$

Identification of Natural Effects: In order to identify natural effects, the potential outcome $Y(a, m)$ is assumed to be independent of the potential mediator $M(a^*)$ whenever a and a^* are different. In other words, this assumption ensures that there are no confounders of the mediator–outcome relationship that are affected by exposure, so that the direct and indirect effects are two distinct systems between exposure and outcome. Mathematically, we have

$$Y(a, m) \perp M(a^*)|X \quad for\ any \ \ m \ \ and \ \ a \neq a^*$$

Noted that regardless of which mediation analysis approach is used, no unmeasured confounding assumption is required [6].

If all the above assumptions hold, then

$$E[Y(a, M(a^*))] = E_X[E_{M|A=a^*,X}(E[Y|A = a, M = m, X = x])] \tag{1}$$

which can be estimated via:

1. Model-based standardization that model outcome, mediator, and exposure

$$E[Y(a, M(a^*))] = \frac{1}{n}\sum_{i=1}^{n}\Big[E[Y_i|A_i = a, M_i = m_i, X_i = x_i]\Big] \tag{2}$$

$$E[M(a)] = \frac{1}{n}\sum_{i=1}^{n}\Big[E[M_i|A_i = a, X_i = x_i]\Big] \tag{3}$$

2. Inverse probability weighting approach (IPW) [38,39] that weights the the observed outcomes via a combination of exposure and mediator given covariates

$$E[Y(a, M(a^*))] = E\Big[\frac{Y\mathbb{1}(A = a)}{f(a|X)} \cdot \frac{f(m|a^*, X)}{f(m|a, X)}\Big] \tag{4}$$

2.1. Zero-Inflated and Zero-Hurdle Models

Mixed models have been proposed to avoid the loss of information when data contain excess zeros. Zero-inflated (ZI) models were first introduced by Cohen [40] and widely accepted after Mullahy [41] and Lambert [42]. ZI models are mixtures of a discrete distribution and a zero point mass. Structure zeros are distinguished from count data, and the counts are usually assumed to follow a Poisson or negative binomial (NB) distribution. The density function for ZI models can be generally defined as

$$f_{ZI}(m_i) = \begin{cases} \phi_i, & \text{for } m_i = 0 \\ (1 - \phi_i)g(m_i), & \text{for } m_i > 0 \end{cases} \tag{5}$$

On the other hand, zero-hurdle (ZH) models [41,43,44] process the data in two stages to account for the excessive number of zeros. The first part is a binary model to determine whether the outcome is zero or a positive value. Logistic regression models are usually used for the first part to incorporate the effects of the covariates on the probability of an observation being zero. For the second part, distributions truncated at zero are used, conditioning on all the nonzero count outcomes. Zero-truncated regression models are then applied to incorporate the covariates effects on the nonzero distribution.

Xu et al. [31] analyzed several commonly used models on microbiome abundances including standard parametric and nonparametric models, zero-hurdle models, and zero-inflated models with varying degrees of zero inflation, with or without dispersion in the count component, and different directions of the covariate effect on both the structural zero and the count components. The simulation studies showed that the ZH and ZI models outperform the other models in terms of type I errors, power, goodness of fit measures, and they are more accurate and efficient in the parameter estimation. Additionally, ZH models are more stable when structural zeros are absent. Therefore, we developed the zero-hurdle negative binomial (ZHNB) model as the mediation model in the mediation framework.

For ZHNB models, the response variable $M(a)$ has the distribution

$$f_{ZHNB}(m_i) = \begin{cases} \phi_i, & \text{for } m_i = 0 \\ (1-\phi_i)\frac{h}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}}, & \text{for } m_i > 0 \end{cases} \tag{6}$$

where

$$h = h(m_i; \mu_i, \alpha) = \frac{\Gamma(m_i + \alpha^{-1})}{\Gamma(m_i+1)\Gamma(\alpha^{-1})}(1+\alpha\mu_i)^{-\alpha^{-1}-m_i}\alpha^{m_i}\mu_i^{m_i} \tag{7}$$

and $\alpha \geq 0$ is a dispersion parameter that is assumed not to depend on covariates and $0 < \phi_0 < 1$. We can estimate the expectation of the variable M_i given the exposure and confounders by

$$\begin{aligned} E[M_i|A_i, \vec{X}_i] &= P(M_i > 0|A_i, \vec{X}_i) \cdot E_{M_i>0|A_i,\vec{X}_i}[M_i|M_i > 0, A_i, \vec{X}_i] \\ &= \frac{(1-\phi_i)\mu_i}{1-\phi_0}, \quad \phi_0 = (\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}})^{1/\alpha} \end{aligned} \tag{8}$$

2.2. Inverse Probability Weighting Two-Part Model

To incorporate a variable with excessive zeros as the mediator, we decompose the mediation effect of the microbiome into two components that are inherent in the zero-inflated distributions: one attributed by a zero part $(M = 0)$ and one attributed by a count part $(M > 0)$ (Figure 2). The zero part, which includes the structural zeros and the sampling zeros, suggests the probability of whether an OTU is present in the data, while the count part explains the change in the outcome of interest resulting from a unit change in the OTU counts. To simultaneously model the zero part and the count part of the mediator, we developed a weighting-based approach in which the estimation of exposure–covariate interactions and a separate averaging step can be avoided. Following Lange et al.'s work [38], we propose a semiparametric approach for estimating the direct and indirect effects while avoiding specification of the outcome model. In this weighting-based approach, the estimation of exposure–covariate interactions and a separate averaging step are also avoided. Furthermore, the IPW approach is less computationally intensive and easier to formulate and implement. The weighting-based approach estimates the expectation of $E[Y(a, M(a^*))]$ as

$$E[Y(a, M(a^*))] = E\Big[\frac{Y \cdot I(A=a)}{f_{A|X}(a|x)} \cdot \frac{f_{M|A,X}(m|a^*, x)}{f_{M|A,X}(m|a, x)}\Big] \tag{9}$$

$$f_{M|A,X}(m|a, x) = h(m_+|m_0, a, x) \cdot g(m_0|a, x) \tag{10}$$

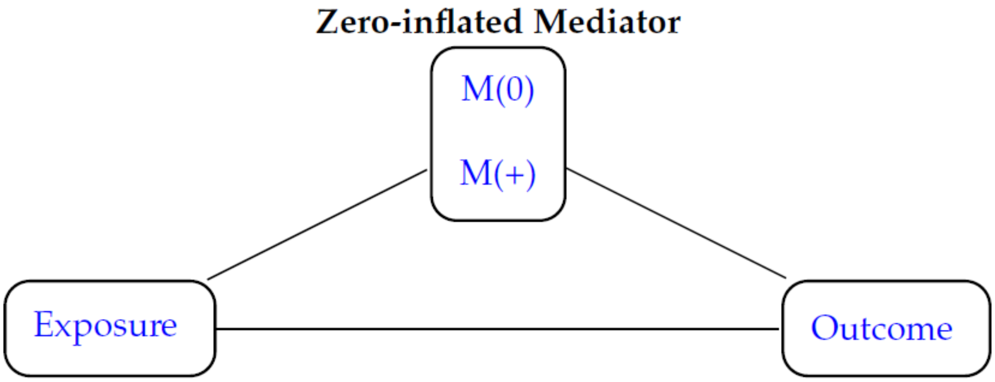

Figure 2. DAG describing the causal mechanism between a binary exposure, a binary outcome, and a zero-inflated mediator.

The IP weights remove confounding by creating a pseudo-population in which the confounders and the exposure are not associated. Subjects with a low probability of receiving the exposure level to which they indeed were exposed have high weights, which results in unstable estimators with large variance. Stabilized inverse probability weights can help the situation by timing the probability of the observed exposure value without any confounders of the original weights. We weight the outcomes by the stabilized inverse probability of each individual's exposure status and mediator levels to obtain the estimations.

As A and Y are binary variables, we model them both using a generalized linear model (GLM) with $g(\cdot)$ as a logit link function. Due to the characteristics of the mediator, we consider a ZH model to estimate the indirect effects. Specifically, for the exposure model with binary exposure, we have a model with a logit link:

$$logit(A|\vec{X}) = \theta_0 + \theta_x^T \vec{X} \tag{11}$$

and $P(A = a|\vec{X})$ can be written as

$$P(A = a|\vec{X}) = \frac{exp(\theta_0 + \theta_x^T \vec{X})}{1 + exp(\theta_0 + \theta_x^T \vec{X})} \tag{12}$$

The stabilized weighting function for the exposure can be written as

$$\begin{aligned} \widehat{sw_i^A} &= \frac{P(A_i = a_i; \hat{\theta}^*)}{P(A_i = a_i|\vec{X}; \hat{\theta}_a)} \\ &= \frac{1 + exp(-\hat{\theta}_0 - \hat{\theta}_x^T \vec{X})}{1 + exp(-\hat{\theta}_0^*)} \end{aligned} \tag{13}$$

For the mediator model, we consider a zero-hurdle negative binomial model with a logit link when the mediator is zero and a log link for the nonzero mediators:

$$logit(M_{m=0}|A = a, \vec{X}) = \beta_0 + \beta_a A + \beta_x^T \vec{X} \tag{14}$$

$$log(M_{m>0}|A = a, \vec{X}) = \gamma_0 + \gamma_a A + \gamma_x^T \vec{X} \tag{15}$$

Then, the probability of zero ϕ is defined as $\phi_i = \frac{1}{1+exp(-\beta_0-\beta_a a_i-\beta_x^T\vec{x}_i)}$, and the mean of counts μ_i is defined as $\mu_i = exp(\gamma_0 + \gamma_a a_i + \gamma_x^T \vec{x}_i)$. The stabilized weighting function for the mediator is

$$\widehat{sw_i^M} = \frac{P(M_i = m_i | A_i = a_i', \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)}{P(M_i = m_i | A_i = a_i, \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)}$$
$$= \begin{cases} \frac{1+exp(-\hat{\beta}_0-\hat{\beta}_a a_i-\hat{\beta}_x^T\vec{x}_i)}{1+exp(-\hat{\beta}_0-\hat{\beta}_a a_i'-\hat{\beta}_x^T\vec{x}_i)}, & \text{for } m_i = 0 \\ \frac{1+exp(\hat{\beta}_0+\hat{\beta}_a a_i+\hat{\beta}_x^T\vec{x}_i)}{1+exp(\hat{\beta}_0+\hat{\beta}_a a_i'+\hat{\beta}_x^T\vec{x}_i))} \left(\frac{1+\alpha exp(\hat{\gamma}_0+\hat{\gamma}_a a_i'+\hat{\gamma}_x^T\vec{x}_i)}{1+\alpha exp(\hat{\gamma}_0+\hat{\gamma}_a a_i+\hat{\gamma}_x^T\vec{x}_i)}\right)^{-\alpha^{-1}-m_i} & \\ \quad exp(\hat{\gamma}_a(a_i'-a_i))^{m_i} \frac{1-[1+\alpha exp(\hat{\gamma}_0+\hat{\gamma}_a a_i+\hat{\gamma}_x^T\vec{x}_i)]^{-\alpha^{-1}}}{1-[1+\alpha exp(\hat{\gamma}_0+\hat{\gamma}_a a_i'+\hat{\gamma}_x^T\vec{x}_i)]^{-\alpha^{-1}}}, & \text{for } m_i > 0 \end{cases} \tag{16}$$

Note that the models used for calculating the weights are not fit to the replicate dataset but to the original one.

2.3. Estimation of the Stabilized Direct and Indirect Effects

One of our major interests in mediation analysis lies in estimating the NDE, NIE, and TE of the counterfactual framework. Because we used the stabilized version of the weights, we named the targeted estimands as SNDE, SNIE, and STE.

The outcomes are weighted by the inverse probability of each individual's exposure status and mediator levels by fitting a weighted logistic regression model. The combined stabilized weights are the product of the exposure weights and the mediator weights. The natural direct and indirect effects can then be estimated accordingly:

$$SNDE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a)f_{A_i}(a)}{f_{A_i|X_i}(a|x)}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a')f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \cdot \frac{f_{M_i|A_i,X_i}(m|a,x)}{f_{M_i|A_i,X_i}(m|a',x)}\right] \tag{17}$$

$$SNIE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a')f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \cdot \frac{f_{M_i|A_i,X_i}(m|a,x)}{f_{M_i|A_i,X_i}(m|a',x)}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a')f_{A_i}(a')}{f_{A_i|X_i}(a'|x)}\right] \tag{18}$$

$$STE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a)f_{A_i}(a)}{f_{A_i|X_i}(a|x)}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\mathbb{I}(A_i = a')f_{A_i}(a')}{f_{A_i|X_i}(a'|x)}\right] \tag{19}$$

With the parameter estimates, we are able to calculate SNDE, SNIE, and STE and their empirical confidence intervals using bootstrapping. The detailed algorithm for the estimation of SNDE, SNIE, and STE can be found in Appendix A.

3. Results

A general framework was developed to accommodate all the zero-inflated distributions for microbiome mediators. This section presents our simulations and real-data implementation to assess the performance of the proposed weighting-based mediation framework.

3.1. Simulation Studies

This subsection presents extensive simulations that demonstrate the performance of the proposed IPW mediation framework in comparison with model-based standardization mediation analysis approaches [35]. Given that the data are absolute abundances, we adopted the zero-hurdle negative binomial model for the mediator model. We also incorporated other commonly used models such as GLM dealing with zero-inflated distributions for a mediator. We simulated 500 subjects with an independent binary exposure variable A, two confounders X_1 and X_2, and a binary outcome Y. Exposure A followed a binary distribution with an approximately 1:1 proportion of the two levels. The first confounder X_1 followed a uniform distribution with $U \sim [10, 80]$, and the second confounder X_2 followed a binary distribution with a 50% proportion.

Because a mediator can have effects on both te zero and count parts, the effects of the two components may act in the same or opposite direction [45]. When the effect of zeros decreases and the effect of positive parts increases, or vice versa, i.e., the two parts work in the same direction on the outcome overall mean, the phenomenon is called consonant effects [31]. In contrast, when the two parts work in the opposite direction with the coefficients having the same sign, this is defined as dissonant effects. A neutral effect is defined as an OTU that has an effect on the count component only. Multiple scenarios were designed in our simulation studies based on the possible effects of zero and count scenarios for the mediator and investigated the mediator influence on the effects of the exposure and outcome in the proposed model.

We considered a data-generating process based on the following equations:

$$logit(A|X_1, X_2) = 0.4 - 0.005X_1 + 0.05X_2 \tag{20}$$

$$logit(M_{m=0}|A, X_1, X_2) = -0.708A - 0.01X_1 + 0.5X_2 \tag{21}$$

$$log(M_{m>0}|A, X_1, X_2) = 0.5 + \gamma^T A - 0.01X_1 + 0.5X_2 \tag{22}$$

$$logit(Y|A, M, X_1, X_2) = -3 - 5A + M + 0.1X_1 + X_2 \tag{23}$$

Equation (20) is the exposure equation, in which the exposure A is a function of the observed variables X_1 and X_2. Using Equations (21) and (22), the mediator is divided into two parts: the zero part and the count part. The odds ratio for exposure and nonexposure groups was set to 1.5. We fixed the coefficient of the exposure for the zero part and varied the coefficient of A in Equation (22) to obtain consonant, neutral, and dissonant effects by setting $\gamma^T = [0.6,\ 0,\ -1]$, respectively. According to Equation (23), the observed outcome Y is a function of exposure A, mediator M, and the two confounders X_1 and X_2.

We ran 1000 replications with 500 subjects and estimated the effects of the weighting-based mediator with the ZHNB model (W-ZHNB). In addition, we considered model-based mediation models (M-ZHNB and M-GLM) to investigate the two counterfactual frameworks. With both the proposed weighted model and the model-based model, the generalized linear models within the NB family were used as the mediator model when we treated the microbiome as a continuous variable rather than as counts in comparison with the ZHNB model. The SNDE, SNIE, and STE for all the models were calculated.

3.2. Simulation Results

Table 1 presents the bias and bootstrap standard error (SE) of the various estimators including natural direct effects, natural indirect effects, and total effects of four different mediation models under consonant effects. For weighted models, when the mediator was modeled with ZHNB, the bias was the smallest for both SNDE and SNIE among all the models. The two weighting-based models had slightly smaller SEs than the model-based models, while the weighting-based GLM models (W-GLMs) had the smallest SE among all the models for SNDE and SNIE. Table A1 in Appendix B.1 presents the bias and bootstrap SEs of the SNDE, SNIE, and STE for the simulation models under neutral effects. We observed a similar pattern to that of the consonant effects. The bias for W-ZHNB was still the smallest among the four models for all three estimators. For SNDE, the two model-based models had a larger bias than both weighting-based models, with a slightly larger SE. However, for SNIE, W-GLM had the largest bias but smallest SE.

Table 1. Bias and bootstrap SEs of stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) for model W-ZHNB, W-GLM, M-ZHNB, and M-GLM under consonant effects.

Effects	**W-ZHNB**		**W-GLM**		**M-ZHNB**		**M-GLM**	
	Bias	**SE**	**Bias**	**SE**	**Bias**	**SE**	**Bias**	**SE**
SNDE	−0.001	0.027	0.026	0.022	−0.041	0.031	−0.038	0.035
SNIE	0.001	0.020	−0.026	0.014	0.043	0.028	0.042	0.029
STE	0.000	0.017	0.000	0.017	0.002	0.026	0.004	0.029

Table 2 provides the simulation results for dissonant effects. Similar patterns were observed for SNDE and SNIE with W-ZHNB. The two weighted models outperformed the model-based models for SNDE, while W-GLM, M-ZHNB, and M-GLM produced similar performance for the indirect effect under dissonant effects. The tables are visualized in Figure 3. In conclusion, our proposed W-ZHNB model performed well in the simulations in terms of the smallest bias and bootstrap SE under consonant, neutral, and dissonant effects.

Table 2. Bias and bootstrap SEs of stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) for models W-ZHNB, W-GLM, M-ZHNB, and M-GLM under dissonant effects.

Effects	**W-ZHNB**		**W-GLM**		**M-ZHNB**		**M-GLM**	
	Bias	**SE**	**Bias**	**SE**	**Bias**	**SE**	**Bias**	**SE**
SNDE	0.003	0.047	−0.012	0.047	−0.037	0.052	−0.040	0.052
SNIE	−0.002	0.043	0.012	0.043	−0.014	0.049	−0.010	0.049
STE	0.000	0.020	0.000	0.020	−0.051	0.026	−0.050	0.026

3.3. Real Data Implementation

We applied the proposed model and the other models to real microbiome data of infants' incidence of type 1 diabetes and allergies (DIABIMMUNE) study [46]. The cohort was recruited from three countries: 71 infants from Estonia, 71 infants from Finland, and 70 infants from Russia. For each infant in this study, 3 years of monthly stool samples; laboratory results; and questionnaires regarding breastfeeding, diet, allergies, family history, usage of drugs, and clinical examinations were collected. The composition of the gut microbiota was determined by sequencing the V4 region of the 16S rRNA gene from 1584 samples. The 16S data were processed using QIIME, and taxonomy was assigned according to the Greengenes taxonomy map of reference sequence OTUs to taxonomy. Because infants from Russia had a lower level of compliance than those in the other two countries, Vatanen et al. [46] sequenced all available Russian samples and used the sparse sampling of the Finnish and Estonian samples to achieve equal sample numbers. A mean sequence depth of 57,110 per sample was obtained, and samples with fewer than 3000 filtered sequences were excluded from the analysis. The metagenomic reads for taxonomic composition down to the genus level were obtained. The absolute abundances predicted when the total bacterial mass was estimated using universal 16S primers agreed with the interpolated absolute abundances. No multiple test adjustment was implemented.

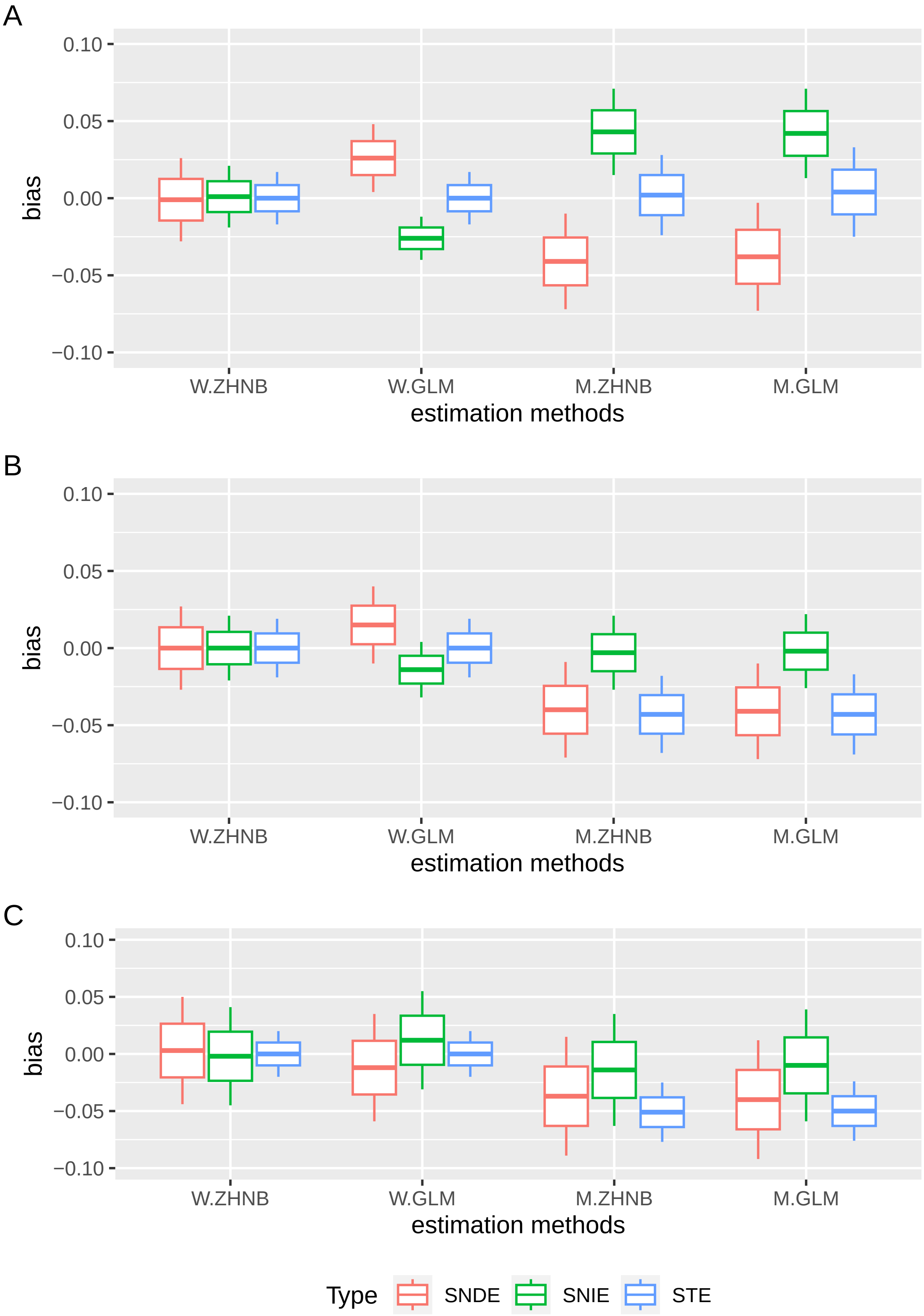

Figure 3. Boxplots for bias and corresponding bootstrap SEs of stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) for models W-ZHNB (W.ZHNB), W-GLM (W.GLM), M-ZHNB (M.ZHNB), and M-GLM (M.GLM) under consonant (**A**), neutral (**B**), and dissonant (**C**) effects.

Our aim was to explore how breastfeeding and baby formula impact allergies both directly and through specific OTUs (Figure 4). We combined various allergy outcomes for each subject as 'overall allergy', which contained allergy reactions of an infant to any of the following foods: milk, eggs, and peanuts. To illustrate the developed algorithm, we only selected two taxa in the analysis, which were the identified taxa that were significantly related to the study outcome: *Lactobacillus* related to breastfeeding and *Lactococcus* related to cow's milk. A total of 785 samples were available from 212 unique infants. Due to missing data, the subjects did not share common time points. For illustration of our proposed method, we kept the first observation for each subject with available allergy information, which resulted in a final sample size of 168.

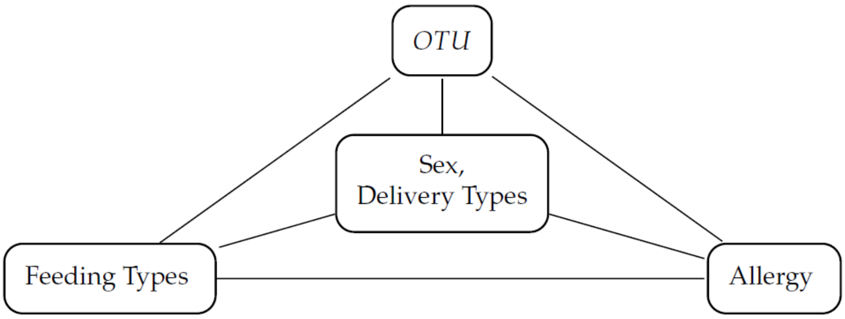

Figure 4. DAG for DIABIMMUNE.

In the study, the microbial taxa that were associated with breastfeeding and cow's milk were identified. The zero proportions were 47.6% for *Lactobacillus* and 42.3% for *Lactococcus*. The two microbial taxa were considered mediators after square root transformation in the causal mediation framework. In addition to the exposure feeding types and the outcome (overall allergy status), we treated sex and delivery types as confounders in the model. Table A2 in Appendix B.2 summarizes the variables of the DIABIMMUNE Cohort that were considered in our model.

We applied four models to assess the SNDE, SNIE, and STE of breastfeeding status on overall allergy through possible related OTUs. Two models used a weighting-based mediator with the ZHNB model (W-ZHNB) and GLM with NB distribution (W-GLM), and the other two were model-based mediation models for ZHNB (M-ZHNB) and GLM with NB distribution (M-GLM). The results of the effects are presented in Table 3 for *Lactobacillus* as the mediator and Table 4 for *Lactococcus*. Standard errors were estimated via 1000 bootstraps. None of the four models found any significant mediation effects of the two OTUs. Notably, M-ZHNB tended to produce larger estimates of SNIE (*Lactobacillus*: −0.0021, 95% CI: −0.0317, 0.0275; *Lactococcus*: 0.0073, 95% CI: −0.024, 0.0386) than W-ZHNB (*Lactobacillus*: −0.0002, 95% CI:−0.0017, 0.0013; *Lactococcus*: 0.0003, 95% CI: −0.0021, 0.0027).

Table 3. Stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) of breastfeeding status, the OTU that related to *Lactobacillus* (a breastfeeding-related OTU), and overall allergy using four causal mediation models after adjusting for sex and delivery type.

Effects	W-ZHNB		W-GLM		M-ZHNB		M-GLM	
	Est	SE	Est	SE	Est	SE	Est	SE
SNDE	−0.1518	0.0679	−0.1487	0.0716	−0.1468	0.1068	−0.1475	0.1085
SNIE	−0.0002	0.0015	−0.0034	0.0210	−0.0021	0.0296	−0.0082	0.0485
STE	−0.1520	0.0681	−0.1520	0.0681	−0.1489	0.1007	−0.1557	0.0763

Table 4. Stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) of breastfeeding status, the OTU related to *Lactococcus* (a cow-milk-related OTU), and overall allergy using four causal mediation models after adjusting for sex and delivery type.

Effects	W-ZHNB		W-GLM		M-ZHNB		M-GLM	
	Est	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**
SNDE	−0.1517	0.0601	−0.1502	0.0619	−0.1455	0.0759	−0.1456	0.0798
SNIE	0.0003	0.0024	−0.0018	0.0117	0.0073	0.0313	−0.0008	0.0268
STE	−0.1519	0.0602	−0.1520	0.0602	−0.1528	0.0797	−0.1465	0.0830

4. Discussion

In this study, we developed an innovative mediation counterfactual framework for the microbiome as a mediator to adopt the zero-inflation characteristics of the microbial mediator. In particular, we incorporated the inverse probability weighting method for parameter estimations and used zero-hurdle models for the zero-inflated mediator in the count data form. We showed that a zero-inflated mediator can be decomposed into two components of which the first part is for whether an OTU presents for the subject, and the second part is for a unit increase in outcome with the OTU's increase. We also constructed the bootstrap standard errors for the microbial variables in a real-data application and provided the corresponding empirical confidence intervals. Simulation studies demonstrated the robust performance of the proposed method in terms of the mediation effect estimation, including both direct and indirect effects. It was shown that the weighting-based approach has less bias than the model-based approach when consonant or dissonant effects are present in the microbiome data. If the true models are known, model-based mediation models should work well, theoretically. However, the true models of the relationship between exposure to the mediator, mediator to the outcome, and exposure to the outcome are usually unknown, especially when dealing with complex human microbiome data. The weighing-based mediation model can avoid the specification of the outcome model and exposure–covariate interactions, resulting a more robust estimation for the unknown true relationships. In addition, the simulation revealed that ZHNB models fit the microbiome mediator better than GLM, especially when excessive zeros were present. In the DIABIMMUNE dataset, model-based M-ZHNB models estimated larger SNIE than W-ZHNB for both taxa *Lactobacilllus* and *Lactococcus*.

Several extensions of our proposed approach could be explored in the future. We considered the situation where both exposure and outcome are binary. One natural extension is to accommodate multilevel or continuous exposure and/or outcomes. Although our framework can handle microbiome count data as mediators, it would be desirable to extend microbiome count data to relative abundance in the proposed weighting mediation framework in future research. The human microbiota is a dynamic ecosystem that can interact within itself and change over time. In this study, we focused on the mediation effect of a single time point of one OTU to address the knowledge gap of existing mediation frameworks being unable to handle the overdispersion and zero-inflation problems produced by OTU abundances. Multiple OTUs could result in interaction effects among the taxa as microbiome mediators, which would be our next step to explore. The dynamic change in the OTUs is interesting, so it would be important to explore the meaning of a mediation effect of OTUs. We will investigate longitudinal human microbiota data in the future.

Author Contributions: Conceptualization, D.Y. and W.X.; methodology, D.Y. and W.X.; software, D.Y.; validation, D.Y. and W.X.; formal analysis, D.Y.; investigation, D.Y.; resources, D.Y.; data curation, D.Y. and W.X.; writing—original draft preparation, D.Y.; writing—review and editing, D.Y. and W.X.; visualization, D.Y.; supervision, W.X.; project administration, D.Y.; funding acquisition, W.X. All authors have read and agreed to the published version of the manuscript.

Funding: D.Y. was supported by funding from Edwin S.H. Leong Scholarship. W.X. was funded by Natural Sciences and Engineering Research Council of Canada (NSERC Grant RGPIN-2017-06672).

Data Availability Statement: The data used in Section 4 real data application were obtained from a publicly archived dataset from the DIABIMMUNE study [46].

Acknowledgments: We thank for Olli Saarela for methodology discussion and simulation results.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SEM	Structural Equation modeling
MSM	Marginal Structural Models
OTU	Operational Taxonomic Units
DAG	Directed Acyclic Graph
TE	Total Effect
NDE	Natural Direct Effect
NIE	Natural Indirect Effect
IPW	Inverse Probability Weighting
ZI	Zero-Inflated
NB	Negative Binomial
ZH	Zero-Hurdle
ZHNB	Zero Hurdle Negative Binomial
GLM	Generalized Linear Model
STE	Stabilized Total Effect
SNDE	Stabilized Natural Direct Effect
SNIE	Stabilized Natural Indirect Effect
W-ZHNB	Weighting-based Mediator with Zero Hurdle Negative Binomial Model
M-ZHNB	Model-based Mediator with Zero Hurdle Negative Binomial Model
W-GLM	Weighting-based Mediator with GLM Model
M-GLM	Model-based Mediator with GLM Model
SE	Standard Error

Appendix A

Appendix A.1. Algorithm for Estimation of the Stabilized Direct and Indirect Effects

A summary of the estimation steps is provided below:

Step 1: Fit a logistic model including only an intercept term to estimate the probability $P(A_i = a_i; \hat{\theta}^*)$ exposed under treatment level a_i for each subject i.

Step 2: Fit a logistic model for the exposure given covariates to conditionally estimate the probability $P(A_i = a_i|\vec{X}; \hat{\theta}_a)$ exposed under treatment level a_i on the observed values of the covariate vector x_i for each subject i.

Step 3: For each subject i, calculate the stabilized exposure weight

$$sw_i^A = \frac{P(A_i = a_i; \hat{\theta}^*)}{P(A_i = a_i|\vec{X}; \hat{\theta}_a)} \tag{A1}$$

Step 4: Fit a model for the mediator given exposure and covariates using the hurdle negative binomial model

$$E[M|A, \vec{X}; \hat{\beta}, \hat{\gamma}]$$

Step 5: Construct an expanded dataset by repeating each observation of the original cohort four times and creating an additional variable A' that takes the exposure values a, a'.

Step 6: For each subject i, calculate the following zero-inflated function values

$$f_M(M = M_i|A_i = a_i, \vec{X}_i = x_i)$$

$$f_M(M = M_i | A_i = a_i', \vec{X}_i = x_i)$$

by applying the model fitted in Step 4 to the replicated data of Step 5.

Step 7: For each subject i, calculate the stabilized mediator weight

$$sw_i^M = \frac{P(M_i = m_i | A_i = a_i', \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)}{P(M_i = m_i | A_i = a_i, \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)} \tag{A2}$$

Step 8: Calculate the overall weights sw_i

$$sw_i = sw_i^A \times sw_i^M = \frac{P(A_i = a_i; \hat{\theta}^*)}{P(A_i = a_i | \vec{X}; \hat{\theta}_a)} \times \frac{P(M_i = m_i | A_i = a_i', \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)}{P(M_i = m_i | A_i = a_i, \vec{X}_i = \vec{x}_i; \hat{\beta}_m, \hat{\gamma}_m)} \tag{A3}$$

Step 9: Weight the outcomes using the inverse probability of each individual's exposure status and mediator levels, then fit a weighted logistic regression model

$$E[Y | A, M, \vec{X}; \hat{\delta}]$$

Step 10: Estimate the natural direct and indirect effects accordingly.

$$SNDE = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a) f_{A_i}(a)}{f_{A_i|X_i}(a|x)} \right] - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a') f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \cdot \frac{f_{M_i|A_i,X_i}(m|a,x)}{f_{M_i|A_i,X_i}(m|a',x)} \right] \tag{A4}$$

$$SNIE = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a') f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \cdot \frac{f_{M_i|A_i,X_i}(m|a,x)}{f_{M_i|A_i,X_i}(m|a',x)} \right] - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a') f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \right] \tag{A5}$$

$$STE = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a) f_{A_i}(a)}{f_{A_i|X_i}(a|x)} \right] - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i \mathbb{I}(A_i = a') f_{A_i}(a')}{f_{A_i|X_i}(a'|x)} \right] \tag{A6}$$

Appendix B

Appendix B.1. Table of Bias and Bootstrap SE of SNDE, SNIE, and STE under Neutral Effects

Table A1. Bias and bootstrap SEs of stabilized natural direct effects (SNDE), stabilized natural indirect effects (SNIE), and stabilized total effects (STE) for the W-ZHNB, W-GLM, M-ZHNB, and M-GLM models under neutral effects.

Effects	W-ZHNB		W-GLM		M-ZHNB		M-GLM	
	Bias	**SE**	**Bias**	**SE**	**Bias**	**SE**	**Bias**	**SE**
SNDE	0.000	0.027	0.015	0.025	−0.040	0.031	−0.041	0.031
SNIE	0.000	0.021	−0.014	0.018	−0.003	0.024	−0.002	0.024
STE	0.000	0.019	0.000	0.019	−0.043	0.025	−0.043	0.026

Appendix B.2. Table of Summary of Covariates by Overall Allergy Status for Infants

Table A2. Summary of covariates by overall allergy status for infants.

	Overall (N = 168)	**Allergy Yes (N = 62)**	**Allergy No (N = 106)**
Feeding Types			
Breastfeeding	56(33.3%)	15(24.2%)	41(38.7%)
Baby Formula	112(66.7%)	47(75.8%)	65(61.3%)
Sex			
Female	76(45.2%)	28(45.2%)	48(45.3%)
Male	92(54.8%)	34(54.8%)	58(54.7%)

Table A2. *Cont.*

	Overall (N = 168)	Allergy Yes (N = 62)	Allergy No (N = 106)
Delivery Type			
Vaginal	154(91.7%)	59(95.2%)	95(89.6)%
Cesarean section	14(8.3%)	3(4.8%)	11(10.4%)
g-Lactobacillis			
Mean (SD)	115(1140)	285(1870)	14.9(47.8)
Median[Min, Max]	1.00[0, 14700]	1.00[0, 14700]	1.00[0, 322]
g-Lactococcus			
Mean (SD)	93.4(390)	84.3(360)	98.7(408)
Median[Min, Max]	1.00[0, 3080]	1.00[0, 2520]	1.00[0, 3080]

References

1. Robins, J.M.; Greenland, S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **1992**, *3* , 143–155. [CrossRef] [PubMed]
2. Pearl, J. Direct and indirect effects. *arXiv* **2013**, arXiv:1301.2300.
3. VanderWeele, T.J. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **2009**, *20*, 18–26. [CrossRef] [PubMed]
4. Pearl, J. Principal stratification—A goal or a tool? *Int. J. Biostat.* **2011**, *7*, 20 . [CrossRef] [PubMed]
5. Imai, K.; Keele, L.; Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **2010**, *15*, 309. [CrossRef]
6. VanderWeele, T.J. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **2016**, *37*, 17–32. [CrossRef]
7. Muthén, B.; Asparouhov, T. Causal effects in mediation modeling: An introduction with applications to latent variables. *Struct. Equ. Model. Multidiscip. J.* **2015**, *22*, 12–23. [CrossRef]
8. Rubin, D. Estimating causal effects of treatments in experimental and observational studies. *ETS Res. Bull. Ser.* **1972**, *1972*, i-31. [CrossRef]
9. VanderWeele, T.J. Commentary: Causal mediation analysis with survival data. *Epidemiology* **2011**, *22*, 582–585. [CrossRef]
10. Tchetgen Tchetgen, E.J. On causal mediation analysis with a survival outcome. *Int. J. Biostat.* **2011**, *7*, 0000102202155746791351. [CrossRef]
11. Williamson, T.; Ravani, P. Marginal structural models in clinical research: When and how to use them? *Nephrol. Dial. Transplant.* **2017**, *32*, ii84–ii90. [CrossRef]
12. Ley, R.E.; Bäckhed, F.; Turnbaugh, P.; Lozupone, C.A.; Knight, R.D.; Gordon, J.I. Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11070–11075. [CrossRef]
13. Turnbaugh, P.J.; Ley, R.E.; Mahowald, M.A.; Magrini, V.; Mardis, E.R.; Gordon, J.I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **2006**, *444*, 1027–1031. [CrossRef]
14. Turnbaugh, P.J.; Hamady, M.; Yatsunenko, T.; Cantarel, B.L.; Duncan, A.; Ley, R.E.; Sogin, M.L.; Jones, W.J.; Roe, B.A.; Affourtit, J.P.; et al. A core gut microbiome in obese and lean twins. *Nature* **2009**, *457*, 480–484. [CrossRef] [PubMed]
15. Ley, R.E.; Turnbaugh, P.J.; Klein, S.; Gordon, J.I. Human gut microbes associated with obesity. *Nature* **2006**, *444*, 1022–1023. [CrossRef] [PubMed]
16. Koeth, R.A.; Wang, Z.; Levison, B.S.; Buffa, J.A.; Org, E.; Sheehy, B.T.; Britt, E.B.; Fu, X.; Wu, Y.; Li, L.; et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **2013**, *19*, 576–585. [CrossRef] [PubMed]
17. Lepage, P.; Häsler, R.; Spehlmann, M.E.; Rehman, A.; Zvirbliene, A.; Begun, A.; Ott, S.; Kupcinskas, L.; Doré, J.; Raedler, A.; et al. Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **2011**, *141*, 227–236. [CrossRef]
18. Garrett, W.S.; Gallini, C.A.; Yatsunenko, T.; Michaud, M.; DuBois, A.; Delaney, M.L.; Punit, S.; Karlsson, M.; Bry, L.; Glickman, J.N.; et al. Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host Microbe* **2010**, *8*, 292–300. [CrossRef]
19. Yang, D.; Xu, W. Clustering on Human Microbiome Sequencing Data: A Distance-Based Unsupervised Learning Model. *Microorganisms* **2020**, *8*, 1612. [CrossRef]
20. Wen, L.; Ley, R.E.; Volchkov, P.Y.; Stranges, P.B.; Avanesyan, L.; Stonebraker, A.C.; Hu, C.; Wong, F.S.; Szot, G.L.; Bluestone, J.A.; et al. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **2008**, *455*, 1109–1113. [CrossRef]
21. Qin, J.; Li, Y.; Cai, Z.; Li, S.; Zhu, J.; Zhang, F.; Liang, S.; Zhang, W.; Guan, Y.; Shen, D.; et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **2012**, *490*, 55–60. [CrossRef] [PubMed]
22. Yan, A.W.; E. Fouts, D.; Brandl, J.; Stärkel, P.; Torralba, M.; Schott, E.; Tsukamoto, H.; Nelson, K.E.; A. Brenner, D.; Schnabl, B. Enteric dysbiosis associated with a mouse model of alcoholic liver disease. *Hepatology* **2011**, *53*, 96–105. [CrossRef] [PubMed]
23. Qin, N.; Yang, F.; Li, A.; Prifti, E.; Chen, Y.; Shao, L.; Guo, J.; Le Chatelier, E.; Yao, J.; Wu, L.; et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature* **2014**, *513*, 59–64. [CrossRef]

24. Zackular, J.P.; Baxter, N.T.; Chen, G.Y.; Schloss, P.D. Manipulation of the gut microbiota reveals role in colon tumorigenesis. *MSphere* **2016**, *1*, e00001-15. [CrossRef] [PubMed]
25. Petti, C.; Polage, C.; Schreckenberger, P. The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. *J. Clin. Microbiol.* **2005**, *43*, 6123–6125. [CrossRef]
26. Janda, J.M.; Abbott, S.L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* **2007**, *45*, 2761–2764. [CrossRef]
27. Weber, J.L.; Myers, E.W. Human whole-genome shotgun sequencing. *Genome Res.* **1997**, *7*, 401–409. [CrossRef]
28. Chen, K.; Pachter, L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* **2005**, *1*, e24. [CrossRef]
29. Mi, K.; Xu, Y.; Li, Y.; Liu, X. QMD: A new method to quantify microbial absolute abundance differences between groups. *iMeta* **2023**, *2*, e78. [CrossRef]
30. Kaul, A.; Mandal, S.; Davidov, O.; Peddada, S.D. Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **2017**, *8*, 2114. [CrossRef]
31. Xu, L.; Paterson, A.D.; Turpin, W.; Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* **2015**, *10*, e0129606. [CrossRef]
32. Yang, D.; Xu, W. Statistical modeling on human microbiome sequencing data. *Big Data Inf. Anal.* **2019**, *4*, 1–12. [CrossRef]
33. Sohn, M.B.; Li, H. Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* **2019**, *13*, 661–681. [CrossRef]
34. Zhang, X.; Mallick, H.; Tang, Z.; Zhang, L.; Cui, X.; Benson, A.K.; Yi, N. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* **2017**, *18*, 4. [CrossRef] [PubMed]
35. 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. [CrossRef] [PubMed]
36. Zhang, Y.; Wang, J.; Shen, J.; Galloway-Pena, J.; Shelburne, S.; Wang, L.; Hu, J. Inverse Probability Weighting-based Mediation Analysis for Microbiome Data. *arXiv* **2021**, arXiv:2110.02440.
37. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
38. Lange, T.; Vansteelandt, S.; Bekaert, M. A simple unified approach for estimating natural direct and indirect effects. *Am. J. Epidemiol.* **2012**, *176*, 190–195. [CrossRef]
39. Huber, M. Identifying causal mechanisms (primarily) based on inverse probability weighting. *J. Appl. Econom.* **2014**, *29*, 920–943. [CrossRef]
40. Cohen, A.C. Estimation in Mixtures of Discrete Distributions. In *Proceedings of the International Symposium on Discrete Distributions, Montreal, QC, Canada*; Pergamon Press: New York, NY, USA, 1963; pp. 373–378.
41. Mullahy, J. Specification and testing of some modified count data models. *J. Econom.* **1986**, *33*, 341–365. [CrossRef]
42. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14. [CrossRef]
43. Cragg, J.G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econom. J. Econom. Soc.* **1971**, *39*, 829–844. [CrossRef]
44. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge University Press: New York, NY, USA, 2013; Volume 53.
45. Lachenbruch, P.A. Comparisons of two-part models with competitors. *Stat. Med.* **2001**, *20*, 1215–1234. [CrossRef] [PubMed]
46. Vatanen, T.; Kostic, A.D.; d'Hennezel, E.; Siljander, H.; Franzosa, E.A.; Yassour, M.; Kolde, R.; Vlamakis, H.; Arthur, T.D.; Hämäläinen, A.M.; et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **2016**, *165*, 842–853. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.