PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A joint modeling approach for longitudinal microbiome data improves ability to detect microbiome associations with disease

Pamela N. Luna[1,2], **Jonathan M. Mansbach**[3], **Chad A. Shaw**[1,2]*

1 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, **2** Department of Statistics, Rice University, Houston, Texas, United States of America, **3** Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

* cashaw@bcm.edu

Citation: Luna PN, Mansbach JM, Shaw CA (2020) A joint modeling approach for longitudinal microbiome data improves ability to detect microbiome associations with disease. PLoS Comput Biol 16(12): e1008473. https://doi.org/10.1371/journal.pcbi.1008473

Editor: Benjamin Muir Althouse, Institute for Disease Modeling, UNITED STATES

Received: December 20, 2019

Accepted: October 27, 2020

Published: December 14, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008473

Data Availability Statement: The pregnancy microbiome dataset published by Zhang, et al. (https://doi.org/10.3389/fmicb.2018.01683) is directly available at https://github.com/abbyyan3/

Abstract

Changes in the composition of the microbiome over time are associated with myriad human illnesses. Unfortunately, the lack of analytic techniques has hindered researchers' ability to quantify the association between longitudinal microbial composition and time-to-event outcomes. Prior methodological work developed the joint model for longitudinal and time-to-event data to incorporate time-dependent biomarker covariates into the hazard regression approach to disease outcomes. The original implementation of this joint modeling approach employed a linear mixed effects model to represent the time-dependent covariates. However, when the distribution of the time-dependent covariate is non-Gaussian, as is the case with microbial abundances, researchers require different statistical methodology. We present a joint modeling framework that uses a negative binomial mixed effects model to determine longitudinal taxon abundances. We incorporate these modeled microbial abundances into a hazard function with a parameterization that not only accounts for the proportional nature of microbiome data, but also generates biologically interpretable results. Herein we demonstrate the performance improvements of our approach over existing alternatives via simulation as well as a previously published longitudinal dataset studying the microbiome during pregnancy. The results demonstrate that our joint modeling framework for longitudinal microbiome count data provides a powerful methodology to uncover associations between changes in microbial abundances over time and the onset of disease. This method offers the potential to equip researchers with a deeper understanding of the associations between longitudinal microbial composition changes and disease outcomes. This new approach could potentially lead to new diagnostic biomarkers or inform clinical interventions to help prevent or treat disease.

Author summary

Evaluating how changes in the human microbiome influence the onset of disease could lead to the development of novel approaches for diagnosis and treatment. Although

NBZIMM-tutorial/tree/master/NBMM-longitudinal-temporal-data/. This methodology can be implemented using the development version of rstanarm on GitHub at https://github.com/stan-dev/rstanarm. A tutorial for using this approach is available online at https://pamelanluna.github.io/mbjm-tutorial/.

Funding: This work was funded by grant R01 AI108588 (JMM) from the National Institute of Allergy and Infectious Diseases (https://www.niaid.nih.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

various methods exist to determine significant differences in the microbial compositions between disease outcomes, no methods exist to measure how much changes in the microbiome affect disease onset. This deficiency in analytic methods can be attributed to the difficulty of determining associations between time-dependent covariates and time-to-event outcomes in conjunction with unique challenges of microbiome data analysis. Here we propose a new methodology capable of quantifying the effects of longitudinal microbiome data on time-to-event outcomes that overcomes these obstacles, demonstrating its performance and utility via simulation study and application to real data from a case-control study.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Multiple studies have found differences in microbial compositions among people with various illnesses, including depression, obesity, asthma, and autism spectrum disorder [1–7]. Importantly, the microbiome can fluctuate over time due to diet or other exposures [8–10]. Furthermore, longitudinal studies have shown that changes in the composition of the microbiome over time are associated with disease outcomes [11–14]. Understanding the complex trajectories of different microbes within a community and the relationship of these trajectories to the onset of human disease is important to uncovering the origins of dysbiosis. This enhanced understanding may eventually help researchers develop new methods for diagnosing and treating disease.

Many methods have been developed to find associations between changes in the microbiome and different outcomes [15]. First, cross-sectional analyses compare microbial compositions between phenotypic groups at a single time point and are extended to longitudinal data by contrasting the results across time points [16, 17]. Second, longitudinal regression models determine significant associations between data covariates and taxa abundances over time [18–20]. Third, multiple methods use smoothing splines to determine the time intervals in which microbial compositions significantly differ between phenotypic groups [21–23]. While all of these methods analyze associations between longitudinal microbiome data and an outcome, they do not account for how these changes affect time-to-event disease outcomes. Two methods for determining associations between microbial compositions and event times have been developed [24, 25], but they only examine the microbiome composition at a single time point.

Evaluating associations between time-dependent biomarkers, such as longitudinal microbiome data, and time-to-event outcomes requires a specialized analytic approach. Typically time-to-event models such as the Cox proportional hazards model are used to determine associations between covariates and event times. However, the inclusion of time-dependent biomarkers in a time-to-event model exposes parameter estimates to increased bias due to potential measurement error, imputation of data at event times, or correlation with other covariates and often violates proportional hazards assumptions [26, 27]. A joint modeling approach was developed to address these issues, allowing the incorporation of time-dependent biomarkers as covariates in a time-to-event model [28–31]. This joint modeling method simultaneously estimates a longitudinal submodel for the time-dependent biomarker and an event submodel for the time-to-event outcomes. The event submodel determines associations between the time-dependent biomarker and event times by including their estimated values

from the longitudinal submodel, rather than the observed values, as a covariate [31]. Given that longitudinal microbiome data are time-dependent biomarkers, this joint modeling approach could be used to determine associations between longitudinal microbiome data and time-to-event outcomes. However, microbiome data do not meet the Gaussian assumption of the longitudinal submodel.

Indeed, microbiome data analysis presents unique challenges. Typically, researchers use 16S rRNA gene sequencing or whole-genome shotgun sequencing as the basis for classifying microbes in a sample. This methodology results in a dataset containing counts of each taxon across all samples (microbiome count data). However, the total number of sequence read counts, or library size, varies across samples. This variation is generally recognized as an experimental artifact of the next-generation sequencing procedure and not biologically informative. To address these differences in library sizes, the data are often transformed into relative abundances (microbiome compositional data). This transformation, however, also has limitations. Indeed, microbiome compositional data cannot be analyzed using typical analytic techniques since the data are 1) non-Gaussian and 2) subject to a unit-sum constraint resulting in a simplex sample space [32, 33]. Other data normalization techniques (e.g., edgeR, DESeq2, cumulative sum scaling) have been developed to allow researchers to analyze microbiome data without transforming the data into relative abundances [34–36]. Unfortunately, these normalization methods hinder the interpretability of the resulting statistical models. Another approach to dealing with varying library sizes is to rarefy the data (i.e., subsample sequence read counts so the total number of read counts is consistent across all samples). While rarefying microbiome count data has become a common approach, this methodology reduces statistical power by discarding useful sample data and thus, results in less precise models [37]. More recent analytic approaches have turned away from the Gaussian distribution and instead directly model microbiome count data using discrete probability distributions, such as the Dirichlet multinomial distribution [38] or negative binomial distribution [18].

We hypothesize that a direct methodological extension of the joint model which accounts for the discrete nature of microbial abundances and the variation in library sizes will identify quantitative associations between longitudinal microbiome data and time-to-event outcomes. In turn, these methodological contributions lead to improved sensitivity and specificity in determining time-to-event outcomes influenced by microbial composition changes. To evaluate this hypothesis we develop a joint modeling framework with its longitudinal submodel formulated as a negative binomial mixed effects model that includes an offset term to adjust for library size. We additionally introduce a parameterization that represents the estimated longitudinal submodel values as scaled relative abundances in the event submodel to address the proportional nature of microbiome data [39] and to improve model interpretability. We then outline how to simulate event times associated with longitudinal microbiome data and apply our joint modeling approach to simulated datasets to illustrate its improved performance over existing alternative methods. Finally, we demonstrate the utility of this methodology by quantifying a previously detected association between longitudinal *Prevotella* abundances in the vaginal microbiome during pregnancy and earlier delivery times [40].

Methods

The joint model for longitudinal microbiome count data

The joint model for longitudinal and time-to-event data (joint model) determines associations between endogenous time-dependent covariates and event times [27]. The joint model accomplishes this goal by using a longitudinal submodel to model the time-dependent covariate and then incorporating those model values into the time-to-event model. We extended this joint

modeling approach to appropriately model unrarefied microbiome count data in the longitudinal submodel and incorporate the model values into the time-to-event submodel in a way that allows for interpretable results.

Longitudinal submodel. We modified the longitudinal submodel of the joint model to model subject-specific taxon abundances over time. We analyze taxon abundances in the form of unrarefied sequence read counts, which are non-Gaussian and overdispersed. Rarefying sequencing data essentially subsamples the counts so that the total number of sequence reads in each sample is the same. This throws away potentially useful data and decreases the power of analyses. Although it would be possible to use transformed relative abundances in the joint model assuming a Gaussian distribution, relative abundances often do not follow a Gaussian distribution even after performing common transformations.

To appropriately represent this overdispersed count data, we use a negative binomial distribution. For subject i with sample j, we assume the abundance of a single taxon in a sample y_{ij} follows a negative binomial distribution with probability mass function given in Eq 1. This parameterization of the negative binomial distribution has expected value $E[y_{ij}] = \mu_{ij}$ and variance $Var(y_{ij}) = \mu_{ij} + (\mu_{ij}^2 / \theta)$. The shape parameter $\theta > 0$ ensures that $Var(y_{ij}) > E[y_{ij}]$ and controls the amount of overdispersion in the distribution.

$$P(Y = y_{ij}) = \frac{\Gamma(y_{ij} + \theta)}{y_{ij}!\Gamma(\theta)} \cdot \left(\frac{\theta}{\mu_{ij} + \theta}\right)^{\theta} \cdot \left(\frac{\mu_{ij}}{\mu_{ij} + \theta}\right)^{y_{ij}} \quad (1)$$

We model the subject-specific taxon abundances over time using a negative binomial linear mixed effects model. The linear predictor with log link function for the j^{th} sample for subject i at time t (Eq 2) has fixed effects β for covariates $x_{ij}(t)$ and random effects $b_i \sim$ MVNormal(0, D) for covariates $z_{ij}(t)$. To account for the varying library sizes across samples, we introduced an offset variable into the linear model representing the log of the total number of sequence reads in a sample C_{ij}.

$$\eta_{ij}(t) = \log(\mu_{ij}(t)) = x_{ij}(t)^T\beta + z_{ij}(t)^T b_i + \log(C_{ij}) \quad (2)$$

To represent the subject-specific abundances, we include the subject as the random intercept covariate. We ensure the time variable is included as a fixed or random effect to analyze the abundances over time.

Event submodel. The original implementation of the joint model determines associations between the linear predictor values η_{ij} for the time-dependent covariate and the time-to-event. In the case of the negative binomial linear mixed effects model, the linear predictor is the log of the expected sequence read counts for a given sample. However, the event submodel also needs to account for the total number of sequence reads in a sample. Rearranging Eq 2 shows how we can easily determine the predicted relative abundances using the linear predictor.

$$\frac{\mu_{ij}(t)}{C_{ij}} = \exp\left(x_{ij}(t)^T\beta + z_{ij}(t)^T b_i\right) \quad (3)$$

We include these relative abundance values (Eq 3) in the hazard function for the event submodel (Eq 4) to determine the effect size α between the relative abundance of the taxon and the time-to-event. The hazard function has baseline hazard $h_0(t)$ and effect sizes γ for covariates w. The hazard function uses the entire longitudinal history up to time t, $\mathcal{M}_i(t)$.

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t)\exp\left(\gamma^T w_i + \alpha \cdot \phi \cdot \exp\left(x_i^T(t)\beta + z_i^T(t)b_i\right)\right) \quad (4)$$

The parameter α represents the increase in the expected log hazard of disease onset for each one unit increase in relative taxon abundance. However, a unit increase in relative abundance indicates going from 0% to 100% abundance of the taxon, which is uncommon. Therefore, the model may not be able to determine the effect size if the relative abundances are very small and do not show a unit increase. To improve the performance and interpretability of the model, we incorporated a scaling factor ϕ for these relative abundance terms. The scaling factor allows for flexibility in the model depending on the types of relative abundances and abundance changes in the data. Using $\phi = 10$ will make the unit a 10% change in abundance, and using $\phi = 100$ will make the unit a 1% change in abundance.

The microbiome joint model simultaneously estimates the two submodels with shared fixed effects β and random effects b_i parameters.

Software implementation. This methodology can be implemented using the `rstanarm` R package, which provides tools for Bayesian statistical inference of applied regression models, including the joint model for longitudinal and time-to-event data [41, 42]. We have extended the joint modeling `rstanarm` software to provide the functionality necessary to apply this approach. Code for replicating our analyses is included in supplemental file S1 Code. Furthermore, a tutorial for preprocessing and analyzing data using our methodology is available online and is also included here as S1 Appendix.

Simulation study

Longitudinal microbiome count data. We simulated the taxon abundances for multiple microbiome samples for each subject over time. We modeled the association between the microbial abundances and sample covariates using the model structure given in Eq 5. The fixed effects include a binary time-independent sample covariate X_1, the continuous time variable t_{ij}, and their interaction. To emulate the taxon abundance trajectories for each subject, we also included a subject-specific random intercept and a random slope based on the time variable.

$$Y \sim X_1 + Time + X_1 : Time + (Time \,|\, ID) + \text{offset}(\log(Counts)) \tag{5}$$

We assume the taxon abundances Y follow a negative binomial distribution. Using the log link function, Eq 6 gives the linear predictor for the j^{th} sample from subject i.

$$\eta_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 t_{ij} + \beta_3 X_{1ij} t_{ij} + b_{0i} + b_{1i} t_{ij} + \log(C_{ij}) \tag{6}$$

We set the parameter values for the fixed effects β_1, β_2, and β_3. We sampled the random effects $b_i \sim \text{Normal}_2(0, D)$, where D is the variance-covariance matrix with $d_{11} = \text{Var}(b_0) = 0.003$, $d_{22} = \text{Var}(b_1) = 0.001$, correlation parameter ρ, and covariance term $d_{12} = d_{21} = \rho \cdot d_{11} d_{22}$.

We then determined the model covariates by randomly sampling N values for the time-independent covariate $X_{1i} \sim \text{Bernoulli}(0.5)$ and $N \times K$ values for the time covariate $t_{ij} \sim \text{Uniform}(0, 8)$. We assumed the total number of sequence reads for each sample followed a normal distribution with $C_{ij} \sim \text{Normal}(10000, 1000)$.

Using these simulated covariates and parameter values, we then evaluated the linear predictor η_{ij}. However, because μ_{ij} should give values representing relative abundances, we must restrict its range to $\mu_{ij} \in [0, 1]$. Noting that the intercept term β_0 scales μ_{ij} multiplicatively (Eq 7), we initially set $\beta_0 = 0$ when calculating η_{ij}. We then set $\beta_0 = -\max(\eta_{ij} + \epsilon)$ and recalculated

η_{ij} and μ_{ij} using the new value for β_0.

$$\begin{aligned}\mu_{ij} &= \exp(\eta_{ij}) \\ &= \exp(\beta_0 + \beta_{-0}^T X_{ij} + b_i^T Z_{ij} + \log(C_{ij})) \\ &= \exp(\beta_0) \cdot \exp(\beta_{-0}{}^T X_{ij} + b_i^T Z_{ij} + \log(C_{ij}))\end{aligned} \tag{7}$$

Finally, the longitudinal abundances Y_{ij} were determined by taking $N^* K$ samples from NegativeBinomial(μ_{ij}, θ), where θ is the dispersion parameter.

Event times. Generally, event times can be simulated using the event function $S(t) = \exp(-H(t))$, where $H(t) = \int_0^t h(u)du$ is the cumulative hazard function for $h(t)$, by applying the probability inverse transform. The cumulative hazard function is determined by evaluating the integral of the hazard function from 0 to t. However, the integral over the hazard function for the joint model for microbiome count data (Eq 8) is intractable.

$$h(t|\mathcal{M}(t), w) = h_0(t) \exp(\gamma^T w + \alpha \cdot \phi \cdot \exp[x^T(t)\beta + z^T(t)b]). \tag{8}$$

Crowther and Lambert present a solution for generating event times in instances where the hazard function cannot be integrated analytically to determine a cumulative hazard function [43]. Briefly, the method derives an approximation for the cumulative hazard integral by using Gaussian quadratures. Once the cumulative hazard is calculated, a root finding procedure is then applied in order to solve for the event time t. To simulate microbiome joint model event times we apply this methodology, which is implemented in the `simsurv` R package [44].

For the event submodel, we extended the hazard function for a Cox proportional hazards model with covariates W_1 and W_2. The hazard function for this joint model (Eq 9) incorporates the model values μ_{ij} from the longitudinal submodel scaled by $\phi = 10$ with effect size α.

$$h_i(t) = \lambda \exp(\gamma_1 W_{1i} + \gamma_2 W_{2i} + \alpha \cdot \phi \cdot \exp(\beta_0 + \beta_1 X_{1ij} + \beta_2 t_{ij} + \beta_3 X_{1ij} t_{ij} + b_{0i} + b_{1i} t_{ij})) \tag{9}$$

We assumed an exponential baseline hazard $h_0 = \lambda = 0.1$ for the simulated event times. After setting the parameters γ_1 and γ_2, we sampled variables $W_1 \sim$ Bernoulli(0.5) and $W_2 \sim$ Bernoulli(0.3). The parameters and covariates for both the longitudinal and time-to-event submodels were then used by the `simsurv` R package, to simulate event times for the hazard function (Eq 9). Once the event times are simulated, all longitudinal observations after a subject's event time are removed from the dataset for time-to-event analysis. The simulated event times are right censored at $t_{\max} = 10$. Example R code for simulating event times associated with microbiome count data is included in supplemental file S1 Code.

Results

Model overview

We developed a joint modeling framework to determine associations between longitudinal microbiome count data and time-to-event outcomes that accommodates the distribution of microbiome data while still respecting the inherent proportional characteristic of the microbiome. The model, outlined in Fig 1, consists of a longitudinal and a time-to-event submodel.

The longitudinal component models taxon abundance over time using a negative binomial distribution. The longitudinal model structure addresses the issue of varying library sizes by incorporating an offset variable of the log number of sequence reads for each sample. The model values from the longitudinal submodel are incorporated into the time-to-event submodel in the form of scaled relative abundances. The scaling of the relative abundances improves detection and interpretability of the effect sizes.

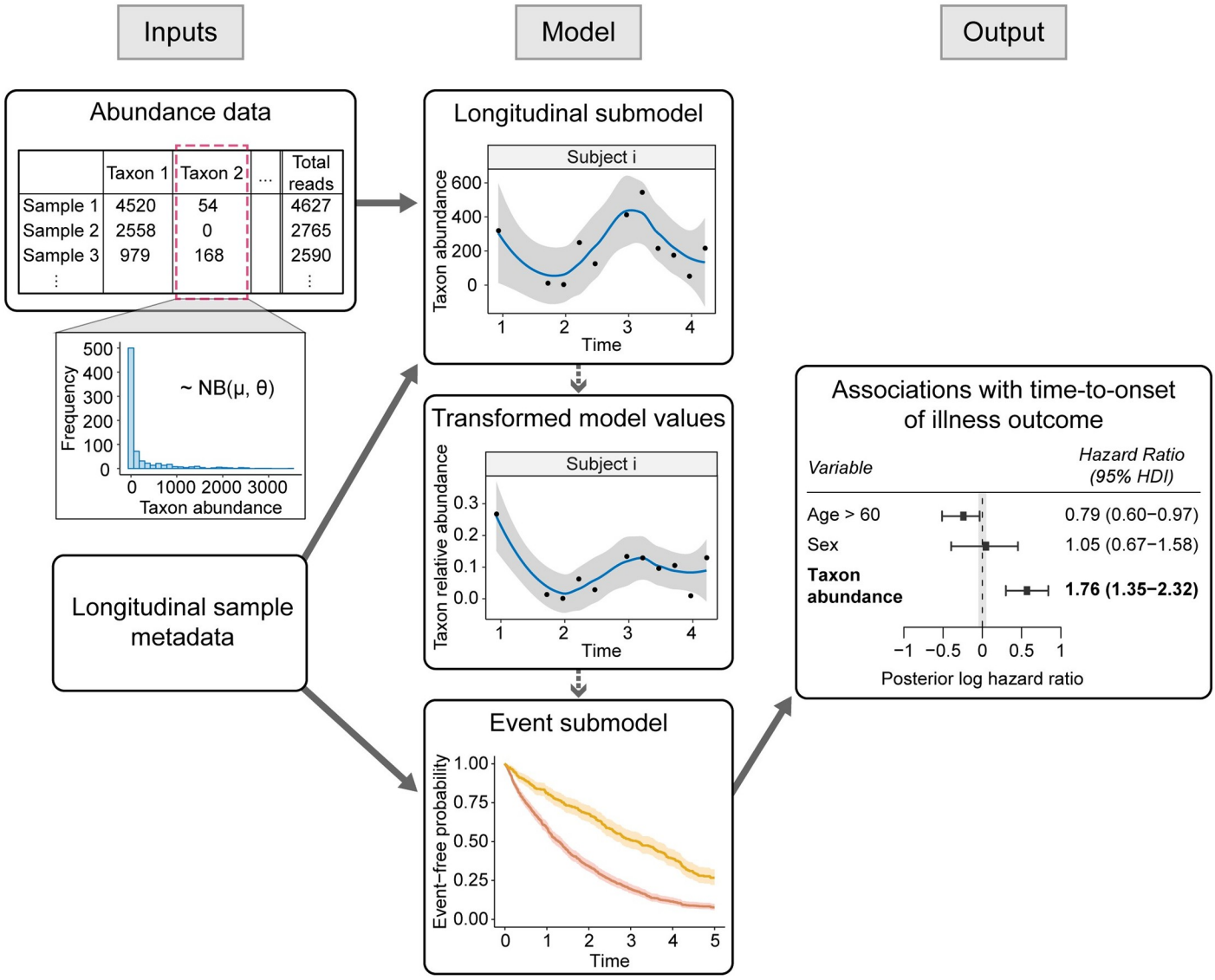

Fig 1. Overview of joint model for longitudinal microbiome count data. The inputs, model structure, and output of the joint model for longitudinal microbiome count data. (Inputs) The taxa abundance table contains the sequence read counts for all taxa across samples. The abundances for a single taxon following a negative binomial distribution and the total reads for each sample are passed to the longitudinal submodel. Additionally, metadata for the longitudinal microbiome samples are passed to both the longitudinal and time-to-event submodels. (Model) The longitudinal submodel analyzes subject-specific taxon abundances over time using a negative binomial mixed effects model. The model values for the taxon abundance are transformed to relative abundances before being included in the event submodel. The event submodel determines associations between longitudinal sample data, including the taxon abundances, and the time-to-event for an outcome. (Output) Parameter estimates from the joint model quantify the associations between the time-to-event and model covariates via hazard ratios.

https://doi.org/10.1371/journal.pcbi.1008473.g001

The parameter estimates from the joint model quantify the effect of microbial abundances on the time-to-onset of disease. These parameter estimates can then be used to determine posterior predictions for the longitudinal and time-to-event submodels. Additionally, the joint model's event-free probability predictions can be updated as more longitudinal data is included in the model.

Simulation study

To assess model performance on data with known parameter values, we analyzed simulated data for $N = 1000$ subjects over $K = 10$ time points. This simulation analysis illustrates that our

model accurately estimates parameter values. While no other methodology exists with the direct aim of quantifying the associations with microbiome data, we show that the joint model performs better than existing alternatives.

Model performance. We applied the joint model for longitudinal microbiome count data to a simulated dataset with taxon abundance effect size of $\alpha = 0.5$. The posterior high density intervals (HDIs) for the longitudinal and time-to-event parameter values are shown in Fig 2A, with true parameter values denoted by the vertical dotted line. For both the longitudinal and time-to-event submodels, the true parameter values fall within their respective 95% HDIs. In particular, the effect size of the taxon abundance parameter α is accurately detected by the joint model, falling in the 50% HDI.

To assess the predictive performance of the microbiome joint model, we predicted the posterior longitudinal trajectories and event-free probabilities using varying amounts of longitudinal data. Fig 2B shows plots of the predicted longitudinal abundances and event-free probabilities using longitudinal data up to $t = 1$ and $t = 4$ split by true event outcome. This figure shows that the model is able to detect the difference in longitudinal trajectories between event outcomes, particularly when predicting the longitudinal trajectory at the later time. Additionally, the model predicts lower event-free probabilities for subjects without the event. This separation becomes more apparent as more longitudinal data is included in the model predictions for event-free probabilities.

Comparison to alternative methods. Although no other methods exist to analyze associations between longitudinal microbiome data and time-to-event outcomes, we were able to compare the performance of our joint model for microbiome count data to existing analytic alternatives. Namely, we compare the performance of our model to the Cox proportional hazards model and the joint model using log-transformed relative abundances. The Cox model does not include the effect of microbial abundances on the time-to-event outcomes. The original formulation of the joint model is the only currently available event model method to include longitudinal microbiome data, but it expects a Gaussian distribution for its longitudinal submodel. To accommodate microbiome data, we normalize the count values to relative abundances and perform a log transformation to shift the data closer to a Gaussian distribution.

Using the same parameter values as above, we simulated event times using various α values and compared the results from these three approaches (Fig 3). The posterior distributions of the model parameters using the moderate $\alpha = 0.5$ taxon abundance effect size are shown in Fig 3A. The parameter estimates for our joint model for microbiome count data always fall in the 95% high density interval (HDI) of the posterior distributions for both the longitudinal and time-to-event submodels. Noting that the Cox model does not have a longitudinal submodel or taxon abundance parameter, we can see that its parameter estimates are close to their true values but that the model compensates for the taxon effect via its baseline hazard. The joint model with transformed relative abundances does have a longitudinal submodel, which has an inaccurately low intercept term. Because the longitudinal model values included in the event submodel are less accurate, the parameter estimates in the event submodel are affected as well. Specifically, the effect of the taxon abundance and the baseline hazard on the time-to-event are both overestimated relative to our method. While we do not expect these models to perform as well as our model since they are not explicitly suited for the simulated data, these results show how analyzing a dataset where this effect is present could skew analytic results.

We also compared the predictive performance of the different models based on the amount of longitudinal data and the true effect size α of taxon abundance. Fig 3B shows the receiver operating characteristic (ROC) curves comparing the ability of the predicted event-free probabilities to differentiate between event outcomes. For $\alpha = 0.1$, the models all have similar

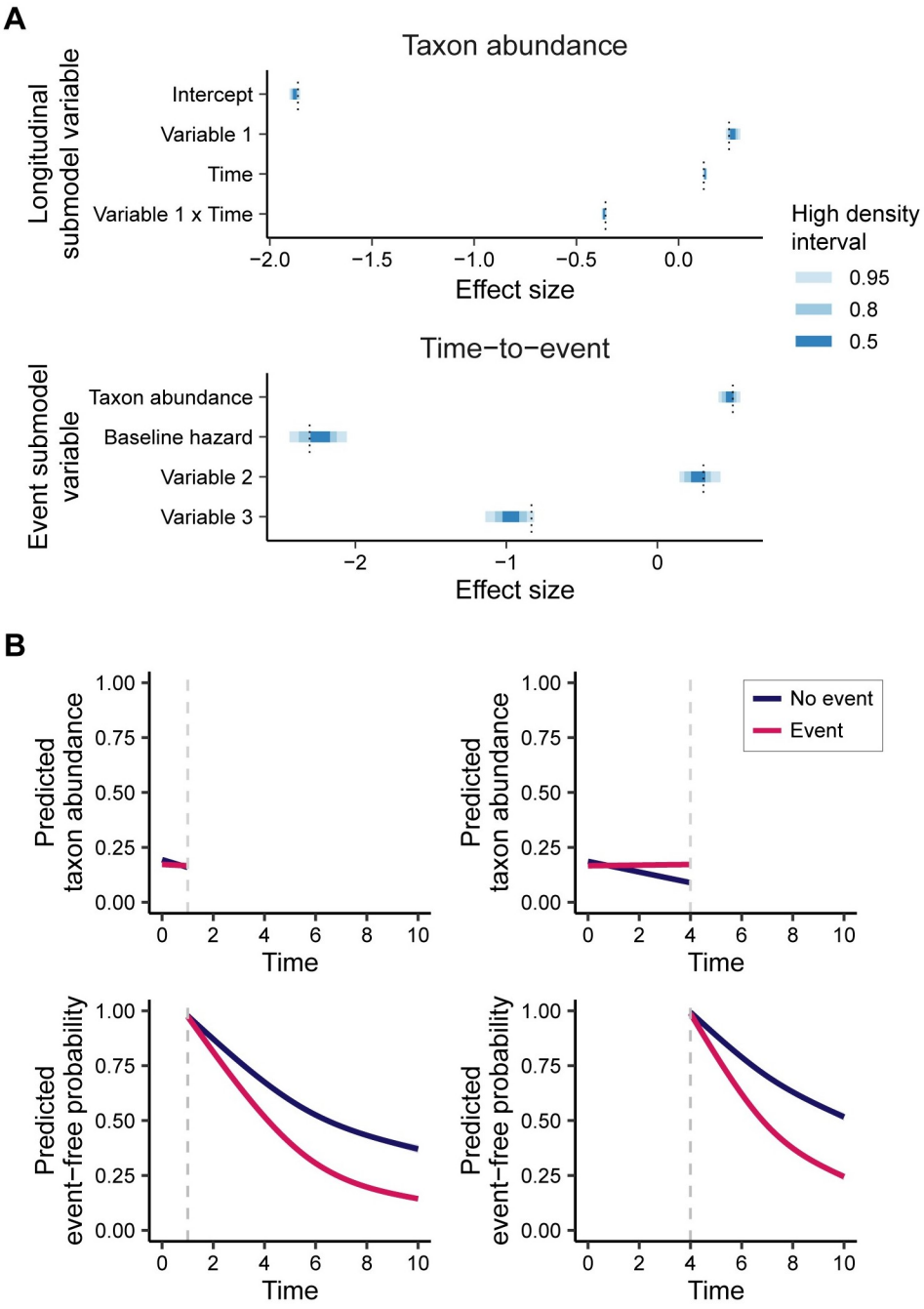

Fig 2. Model results for simulated data. Parameter estimates and predictive ability of the joint model for longitudinal microbiome count data on a simulated dataset. (A) Posterior high density intervals (HDIs) for parameters from the longitudinal and time-to-event submodels. Dotted lines show true parameter values. All parameter values fall within the 95% HDIs for the posterior distributions. (B) Marginal predicted longitudinal trajectories and event-free probabilities from the joint model using longitudinal data up to $t = 1$ (left) and $t = 4$ (right) split by true event outcome. As more longitudinal data is provided for the predictions, there is more separation between the marginal event-free probability predictions.

https://doi.org/10.1371/journal.pcbi.1008473.g002

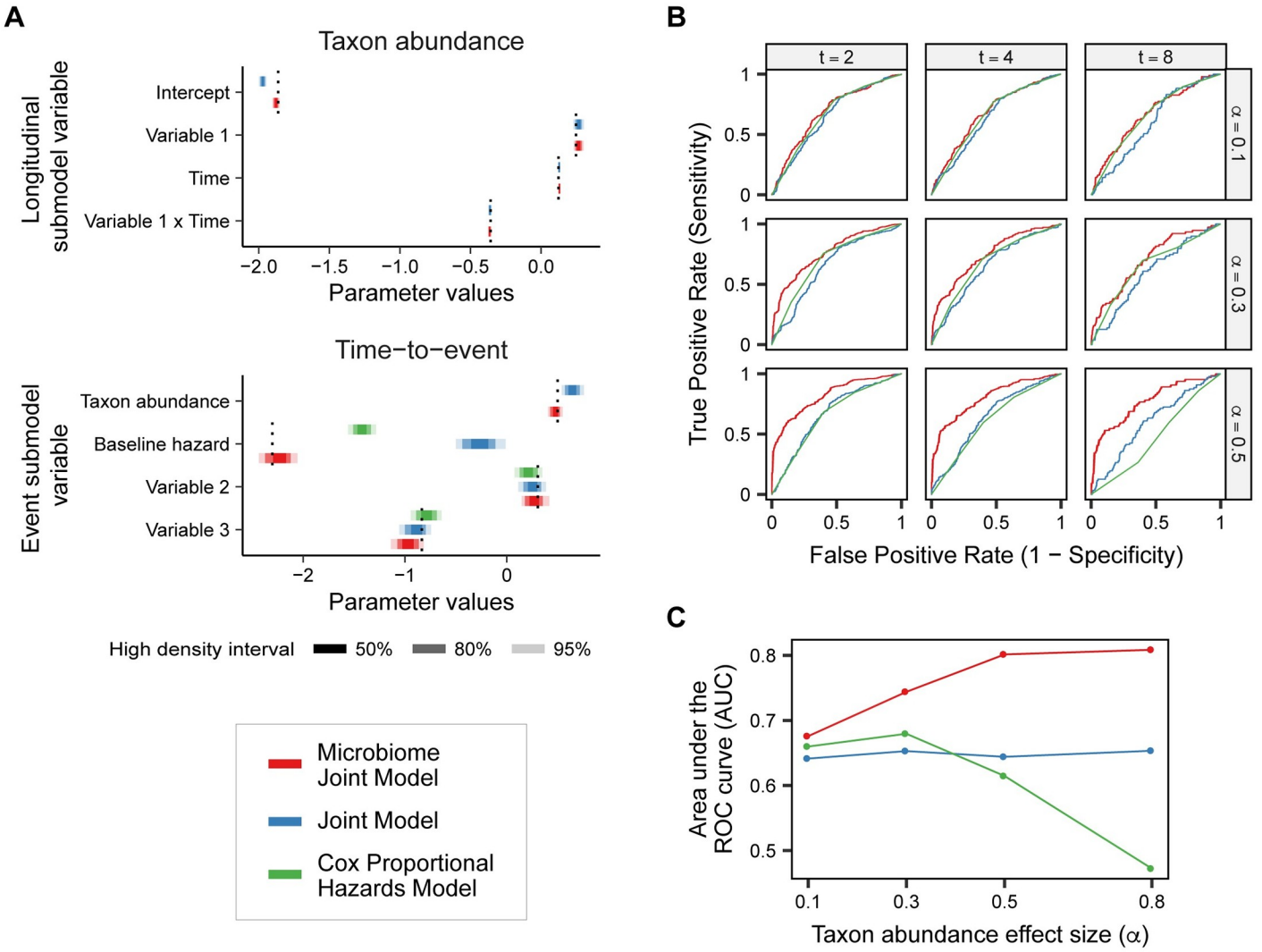

Fig 3. Model performance compared to alternative methods. Analysis of the simulated dataset using the joint model for longitudinal microbiome count data, the original joint model with transformed relative abundances, and the Cox proportional hazards model shows that our model best detects relationships within the data. (A) Posterior high density intervals (HDIs) of the parameter estimates for each of the models. The association with taxon abundance is over estimated in the original joint model. Both the Cox model and joint model poorly estimate the baseline hazard, likely accounting for the differences introduced by the effect of the taxon abundance. (B) Receiver operating characteristic (ROC) curves comparing the performance of the event-free probability predictions compared to the true event outcome for the three models. The panels across the x-axis vary the amount of longitudinal data included in the model, and the panels across the y-axis vary the true taxon abundance effect sizes. (C) Comparison of the area under the ROC curves (AUC) for increasing taxon abundance effect size across all three models. As the effect size increases, the performance of the microbiome joint model improves. The microbiome joint model always performs better than the alternative models.

https://doi.org/10.1371/journal.pcbi.1008473.g003

performance since the taxon abundance does not have a large effect on the time-to-event. However, for increasing alpha our model performs successively better than the other models. Additionally, the facets across the x-axis show the model performance using different amounts of longitudinal data to predict the event-free probabilities. As more longitudinal time points are included, the other models perform worse. In fact, when we include longitudinal data up to time $t = 8$ (all longitudinal data), the Cox model actually performs worse than random when the effect size of the taxon abundance is moderate $\alpha = 0.5$. Because the Cox model does not gain any new longitudinal information, its performance only changes with larger values of t since the predicted event-free probabilities are conditioned on not having the event by time t.

Looking only at the models with longitudinal data up to time $t = 4$, we compared the area under the ROC curves (AUC) across different taxon abundance effect sizes (Fig 3C). This comparison of the AUCs between the models illustrates that our model always performs better than the alternatives, even for lower values of α, and improves for larger values of α. The joint model with transformed relative abundances performs about the same regardless of the effect size, while the Cox model predictions deteriorate with larger effect sizes.

Sample size analysis. To understand how this methodology performs on datasets of varying size, we examined how the number of subjects N and number of longitudinal samples K affect the model's accuracy in estimating the taxon abundance effect size. For each combination of $N \in \{50, 100, 100\}$ and $K \in \{3, 5, 10\}$, we simulated 100 microbiome joint model data sets using randomly selected parameter values consistently across all combinations. We compared the taxon abundance effect sizes estimated using our methodology to the true parameter values. The results for this performance analysis, shown in S1 Fig, illustrate that the error rates for parameter estimates did not increase dramatically with fewer subjects or longitudinal samples.

Application to pregnancy dataset

To demonstrate the utility of our methodology, we applied this joint modeling technique to a pregnancy dataset published by Zhang, et al. [19]. The dataset originated from a case-control study on preterm birth outcomes by DiGiulio, et al. [40] which examined the microbiome of various anatomic sites in women throughout pregnancy. For our analysis, we focused on only vaginal swab samples collected from 40 women prior to birth.

DiGiulio, et al. found that women with microbiome profiles with high abundances of *Lactobacillus* were less likely to experience preterm births, defined as delivery before 37 gestational weeks. However, the *Lactobacillus* count abundances did not fit a negative binomial distribution, so *Lactobacillus* was not appropriate for our model. The study also determined a specific microbiome profile containing high amounts of *Prevotella* that had a higher occurrence of preterm births. Looking at the longitudinal *Prevotella* abundances based on preterm outcome (Fig 4A), we found that women who experienced preterm births had higher levels of *Prevotella* throughout pregnancy than those who did not experience preterm births. Therefore, we chose to examine the association between longitudinal *Prevotella* abundances, which follow a negative binomial distribution, and time to delivery outcomes.

Using microbiome samples combined at the genus level, we modeled the longitudinal abundance of *Prevotella* using a generalized linear mixed effects model adjusted for gestational week of collection, history of preterm births, preeclampsia, and race/ethnicity with random slope based on trimester and random intercept by subject. The longitudinal model was offset by the log library size for each sample. The time-to-event component modeled the association between the relative abundance of *Prevotella* and the time to delivery and was adjusted for preeclampsia, race/ethnicity, and income.

The resulting posterior predictions for the parameters of the longitudinal and time-to-event submodels (Fig 4B) show a positive association between longitudinal *Prevotella* abundance and the time to delivery (HR: 1.5; 80% HDI: 1.17-1.97). Because we used a scaling factor of $\phi = 10$, this result indicates that a 10% increase in *Prevotella* abundance is associated with a 1.5-fold increase in hazard of delivery.

Discussion

We present a discretized extension of the joint model for longitudinal and time-to-event data [30, 42] to evaluate associations between microbial abundances and the onset of disease. As

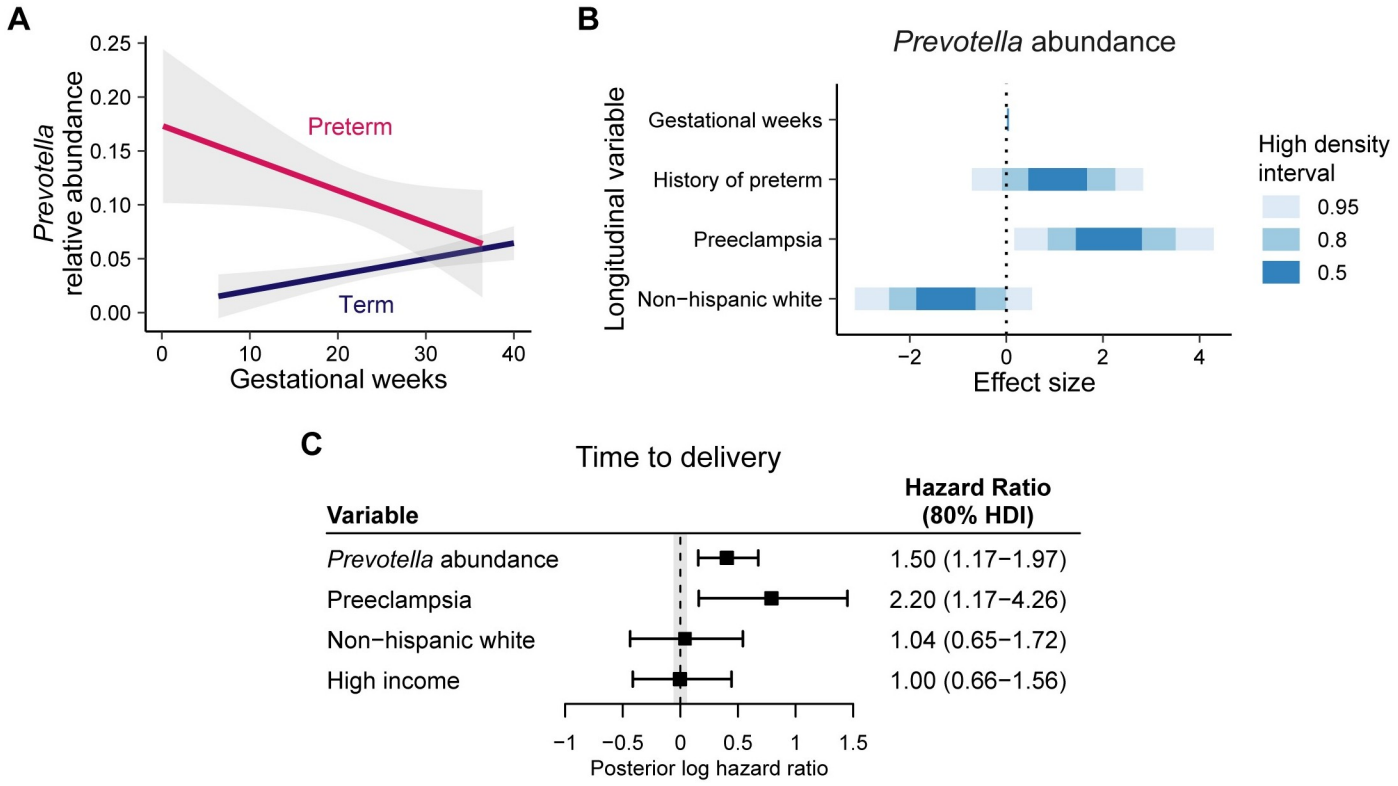

Fig 4. Analysis of longitudinal pregnancy microbiome dataset. Joint model for longitudinal microbiome count data analysis of a longitudinal pregnancy microbiome dataset. (A) Observed longitudinal relative abundances of *Prevotella* split by preterm outcome, defined as time to delivery less than 37 gestational weeks. Subjects with a preterm birth outcome initially have higher levels of *Prevotella* that decrease over time. (B) Posterior predictions for the effect sizes of the longitudinal submodel covariates. (C) Posterior predictions of the hazard ratios for the event submodel covariates. The *Prevotella* abundance parameter shows a positive association with the time to delivery outcome.

https://doi.org/10.1371/journal.pcbi.1008473.g004

hypothesized, our approach correctly quantifies associations between longitudinal microbiome data and time-to-event outcomes. Additionally, this joint modelling approach offers improved sensitivity and specificity relative to existing alternative methods in predicting subject-specific event-free probabilities.

In constructing this joint model, we acknowledged the underlying structure of microbiome data by tailoring an existing method to use statistical assumptions appropriate for the data, rather than transforming data to fit the model's assumptions. First, we modified the longitudinal submodel to reflect the characteristics of microbiome count data by using a negative binomial distribution. Second, we included an offset term in the longitudinal submodel that adjusts for the library size of each sample, avoiding the statistically undesirable process of rarefying microbiome data [37]. Third, we parameterized the event submodel to represent microbial abundances estimated by the longitudinal submodel as scaled relative abundances, which addresses the proportional nature of the microbiome [39] and provides interpretable model results.

Using a simulated dataset, we demonstrated that our method accurately models the effects of microbial abundances and other model covariates on time-to-event outcomes. We have also shown the beneficial predictive properties of this model which allow for improved event predictions with additional longitudinal data [31, 42]. Furthermore, we illustrate how this method could be applied in longitudinal microbiome studies via analysis of a pregnancy microbiome dataset [40]. Our results support an association between Prevotella abundance and preterm

birth detected in previous studies [45–49]. However, in addition to reinforcing this finding of Prevotella as a biomarker, we also determined that a 10% increase in the relative abundance of Prevotella indicates a 1.5-fold increase in the hazard of early delivery. This quantification of the relationship between Prevotella abundance during pregnancy and time to delivery is a new result that was unattainable using prior approaches.

Although our novel methodology solves a problem not previously addressed in the field of microbiome research, there remain opportunities for future research in this area. In its current implementation, our model examines the relationship of an individual microbiome taxon and additional covariates with a time-to-event outcome. This approach can be applied in parallel across individual taxa to perform a comprehensive analysis. We recommend using this parallel analysis approach on a methodically selected subset of individual taxa. The Bayesian hierarchical modeling approach utilized in the joint modeling software produces conservative model estimates that obviate the need for multiple testing corrections [50]. An alternative approach is to model the combined longitudinal dynamics and correlations of many taxa at once within the joint modeling framework. Although of interest, the actualization of joint modeling for many taxa is difficult due to the computational complexity of hierarchical Bayesian analyses, where the model complexity grows exponentially in the number of parameters considered [51]. The current implementation of the joint model provides functionality for a multivariate joint model that could model up to three taxa; however, this implementation could violate model assumptions due to the dependency issues intrinsic to microbiome and compositional data [32, 33]. In the future high dimensional Bayesian approaches may enable such model estimation [52].

We argue that the single taxon joint analysis is effective for two reasons. First, we note that clinicians and biologists focus their interest on the largest and most easily interpretable effects —such as the risk impact of individual taxa on outcomes that we demonstrated in the preterm birth application. Higher order effects of many taxa are less interpretable and therefore less actionable. Second, we note that commonly used analytic methods that consider the entire microbial community by clustering data often result in microbiome profiles dominated by a single taxon [40, 48, 53–56]. In these instances, a cluster that is driven by an individual taxon is used in downstream association analyses, which is analogous to our approach.

Another limitation of our model is the restriction of the longitudinal submodel to a negative binomial distribution, which precludes the analysis of taxa with bimodal distributions or excess zeroes. The number of taxa with these distributional complications is often low but can be highly dependent on the microbial diversity within the dataset. Taxa with an overabundance of zero counts might be better modeled using zero-inflated or hurdle models, but these solutions not currently implemented in existing joint model software. In these situations our negative binomial approach could still be applied but with a reduction in statistical power and performance due to lack of model fit. We advise performing preliminary analyses on microbiome data to determine taxa of interest and to ensure model assumptions are satisfied.

Despite its limitations, our methodology could be a powerful tool in understanding the relationship between changes in the human microbiome and disease. While we have discussed this approach in the context of the human microbiome, this joint model is also applicable to general microbiome studies. Furthermore, this analytic method could generally be applied to any dataset with a time-dependent endogenous covariate that follows a negative binomial distribution and that also has time-to-event outcomes.

Supporting information

S1 Code. Example R code for analysis and simulation.
(R)

S1 Appendix. Tutorial for preprocessing and analyzing data.
(PDF)

S1 Fig. Taxon abundance effect size errors using varying sample sizes. Application of the joint modeling methodology on simulated data sets with varying sizes for the number of subjects (N) and number of longitudinal samples (K) shows that the model retains accuracy with smaller sample sizes. The mean squared errors (MSEs) for the effect size predictions remain low with sample sizes as small as N = 100 with any number of longitudinal samples. The MSEs for effect size predictions are larger with sample size N = 50, but the MSEs are reduced with an increased number of longitudinal samples K.
(TIF)

Author Contributions

Conceptualization: Pamela N. Luna, Jonathan M. Mansbach.

Formal analysis: Pamela N. Luna.

Funding acquisition: Jonathan M. Mansbach.

Methodology: Pamela N. Luna.

Software: Pamela N. Luna.

Supervision: Chad A. Shaw.

Validation: Pamela N. Luna.

Visualization: Pamela N. Luna.

Writing – original draft: Pamela N. Luna.

Writing – review & editing: Jonathan M. Mansbach, Chad A. Shaw.

References

1. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nature reviews Genetics. 2012; 13(4):260–70. https://doi.org/10.1038/nrg3182
2. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016; 535(7610):94–103. https://doi.org/10.1038/nature18850 PMID: 27383984
3. Integrative HMPRNC. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014; 16(3):276–89. https://doi.org/10.1016/j.chom.2014.08.014
4. Naseribafrouei A, Hestad K, Avershina E, Sekelja M, Linløkken A, Wilson R, et al. Correlation between the human fecal microbiota and depression. Neurogastroenterology and Motility. 2014; 26(8):1155–1162. https://doi.org/10.1111/nmo.12378 PMID: 24888394
5. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: Human gut microbes associated with obesity. Nature. 2006; 444(7122):1022–1023. https://doi.org/10.1038/4441022a
6. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, et al. Disordered microbial communities in asthmatic airways. PloS one. 2010; 5(1):e8578. https://doi.org/10.1371/journal.pone.0008578 PMID: 20052417
7. Kang DW, Park JG, Ilhan ZE, Wallstrom G, LaBaer J, Adams JB, et al. Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. PLoS ONE. 2013; 8(7). https://doi.org/10.1371/journal.pone.0068322 PMID: 23844187
8. Biesbroek G, Tsivtsivadze E, Sanders EA, Montijn R, Veenhoven RH, Keijser BJ, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. Am J Respir Crit Care Med. 2014; 190(11):1283–92. https://doi.org/10.1164/rccm.201407-1240OC PMID: 25329446

9. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biology. 2011; 12(5):R50. https://doi.org/10.1186/gb-2011-12-5-r50 PMID: 21624126

10. Perez-Losada M, Alamri L, Crandall KA, Freishtat RJ. Nasopharyngeal Microbiome Diversity Changes over Time in Children with Asthma. PLoS One. 2017; 12(1):e0170543. https://doi.org/10.1371/journal.pone.0170543

11. Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Fofanova T, Nelson A, et al. Longitudinal development of the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls. Microbiome. 2017; 5(1):75. https://doi.org/10.1186/s40168-017-0295-1 PMID: 28701177

12. Zhou Y, Shan G, Sodergren E, Weinstock G, Walker WA, Gregory KE. Longitudinal Analysis of the Premature Infant Intestinal Microbiome Prior to Necrotizing Enterocolitis: A Case-Control Study. PLOS ONE. 2015; 10(3):e0118632. https://doi.org/10.1371/journal.pone.0118632

13. Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosh DW, et al. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome. 2013; 1(1):29. https://doi.org/10.1186/2049-2618-1-29 PMID: 24451163

14. Lambert JA, John S, Sobel JD, Akins RA. Longitudinal Analysis of Vaginal Microbiome Dynamics in Women with Recurrent Bacterial Vaginosis: Recognition of the Conversion Process. PLoS ONE. 2013; 8(12):e82599. https://doi.org/10.1371/journal.pone.0082599

15. Gerber GK. Longitudinal Microbiome Data Analysis. In: Metagenomics for Microbiology. Elsevier; 2015. p. 97–111. http://linkinghub.elsevier.com/retrieve/pii/B9780124104723000075.

16. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 2013; 8(4):e61217. https://doi.org/10.1371/journal.pone.0061217

17. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106

18. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, et al. Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics. 2017; 18(1):4. https://doi.org/10.1186/s12859-016-1441-7 PMID: 28049409

19. Zhang X, Pei YF, Zhang L, Guo B, Pendegraft AH, Zhuang W, et al. Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. Frontiers in Microbiology. 2018; 9:1683. https://doi.org/10.3389/fmicb.2018.01683 PMID: 30093893

20. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016; 32(17):2611–2617. https://doi.org/10.1093/bioinformatics/btw308

21. Shields-Cutler RR, Al-Ghalith GA, Yassour M, Knights D. SplinectomeR Enables Group Comparisons in Longitudinal Microbiome Studies. Frontiers in Microbiology. 2018; 9:785. https://doi.org/10.3389/fmicb.2018.00785

22. Metwally AA, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. Microbiome. 2018; 6(1):32. https://doi.org/10.1186/s40168-018-0402-y

23. Paulson JN, Talukder H, Bravo HC. Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. bioRxiv. 2017; p. 099457.

24. Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. MiRKAT-S: a community-level test of association between the microbiota and survival times. Microbiome. 2017; 5(1):17. https://doi.org/10.1186/s40168-017-0239-9

25. Koh H, Livanos AE, Blaser MJ, Li H. A highly adaptive microbiome-based association test for survival traits. BMC Genomics. 2018; 19(1):210. https://doi.org/10.1186/s12864-018-4599-8

26. Fisher LD, Lin DY. TIME-DEPENDENT COVARIATES IN THE COX PROPORTIONAL-HAZARDS REGRESSION MODEL. Annual Review of Public Health. 1999; 20(1):145–157. https://doi.org/10.1146/annurev.publhealth.20.1.145

27. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica. 2004; 14:809–834.

28. Tsiatis AA, Degruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and cd4 counts in patients with aids. Journal of the American Statistical Association. 1995; 90(429):27–37. https://doi.org/10.1080/01621459.1995.10476485

29. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. Statistics in Medicine. 1996; 15(15):1663–1685. https://doi.org/10.1002/(SICI)1097-0258(19960815)15:15%3C1663::AID-SIM294%3E3.0.CO;2-1

30. Wulfsohn MS, Tsiatis AA. A Joint Model for Survival and Longitudinal Data Measured with Error. Biometrics. 1997; 53(1):330. https://doi.org/10.2307/2533118

31. Rizopoulos D. Joint models for longitudinal and time-to-event data: with applications in R. CRC Press; 2012.

32. Aitchison J. The Statistical Analysis of Compositional Data; 1982. https://www.jstor.org/stable/2345821.

33. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology. 2016; 26(5):330–335. https://doi.org/10.1016/j.annepidem.2016.03.002

34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15(12):550. https://doi.org/10.1186/s13059-014-0550-8

35. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012; 40(10):4288–4297. https://doi.org/10.1093/nar/gks042

36. Paulson JN, Colin Stine O, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature Methods. 2013; 10(12):1200–1202. https://doi.org/10.1038/nmeth.2658

37. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Computational Biology. 2014; 10(4):e1003531. https://doi.org/10.1371/journal.pcbi.1003531

38. Holmes I, Harris K, Quince C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. PLoS ONE. 2012; 7(2):e30126. https://doi.org/10.1371/journal.pone.0030126

39. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is not optional; 2017.

40. DiGiulio DB, Callahan BJ, McMurdie PJ, Costello EK, Lyell DJ, Robaczewska A, et al. Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112(35):11060–5. https://doi.org/10.1073/pnas.1502875112 PMID: 26283357

41. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan.; 2018. http://mc-stan.org/.

42. Brilleman S, Crowther M, Moreno-Betancur M, Buros Novik J, Wolfe R. Joint longitudinal and time-to-event models via Stan.;. https://github.com/stan-dev/stancon_talks/.

43. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Statistics in Medicine. 2013; 32(23):4118–4134. https://doi.org/10.1002/sim.5823 PMID: 23613458

44. Brilleman S. simsurv: Simulate Survival Data; 2019. https://CRAN.R-project.org/package=simsurv.

45. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nature Medicine. 2019; 25(6):1012–1021. https://doi.org/10.1038/s41591-019-0450-2 PMID: 31142849

46. Doyle RM, Harris K, Kamiza S, Harjunmaa U, Ashorn U, Nkhoma M, et al. Bacterial communities found in placental tissues are associated with severe chorioamnionitis and adverse birth outcomes. PLOS ONE. 2017; 12(7):e0180167. https://doi.org/10.1371/journal.pone.0180167 PMID: 28700642

47. Tabatabaei N, Eren A, Barreiro L, Yotova V, Dumaine A, Allard C, et al. Vaginal microbiome in early pregnancy and subsequent risk of spontaneous preterm birth: a case-control study. BJOG: An International Journal of Obstetrics & Gynaecology. 2019; 126(3):349–358. https://doi.org/10.1111/1471-0528.15299 PMID: 29791775

48. Freitas AC, Bocking A, Hill JE, Money DM, VOGUE Research Group tVR. Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. Microbiome. 2018; 6(1):117. https://doi.org/10.1186/s40168-018-0502-8

49. Stafford GP, Parker JL, Amabebe E, Kistler J, Reynolds S, Stern V, et al. Spontaneous Preterm Birth Is Associated with Differential Expression of Vaginal Metabolites by Lactobacilli-Dominated Microflora. Frontiers in Physiology. 2017; 8:615. https://doi.org/10.3389/fphys.2017.00615 PMID: 28878691

50. Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness. 2012; 5(2):189–211. https://doi.org/10.1080/19345747.2011.618213

51. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo. 2017;.

52. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian Computing with INLA: A Review. Annual Review of Statistics and Its Application. 2016; 4(1):395–421.

53. Hasegawa K, Mansbach JM, Ajami NJ, Espinola JA, Henke DM, Petrosino JF, et al. Association of nasopharyngeal microbiota profiles with bronchiolitis severity in infants hospitalised for bronchiolitis. The European respiratory journal. 2016; 48(5):1329–1339. https://doi.org/10.1183/13993003.00152-2016 PMID: 27799386

54. Biesbroek G, Tsivtsivadze E, Sanders EAM, Montijn R, Veenhoven RH, Keijser BJF, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. American journal of respiratory and critical care medicine. 2014; 190(11):1283–92. https://doi.org/10.1164/rccm.201407-1240OC PMID: 25329446

55. Luna PN, Hasegawa K, Ajami NJ, Espinola JA, Henke DM, Petrosino JF, et al. The association between anterior nares and nasopharyngeal microbiota in infants hospitalized for bronchiolitis. Microbiome. 2018; 6(1):2. https://doi.org/10.1186/s40168-017-0385-0 PMID: 29298732

56. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011; 473(7346):174–180. https://doi.org/10.1038/nature09944 PMID: 21508958