Correlating High-dimensional longitudinal microbial features with time-varying outcomes with FLORAL

Teng Fei*[1], Victoria Donovan[1,2], Tyler Funnell[3], Mirae Baichoo[4], Nicholas R. Waters[4], Jenny Paredes[3], Anqi Dai[4], Francesca Castro[5], Jennifer Haber[6], Ana Gradissimo[6], Sandeep S. Raj[7], Alexander M. Lesokhin[5,8], Urvi A. Shah†[5,8], Marcel R. M. van den Brink†[3,9], and Jonathan U. Peled†[7,8]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[3]Department of Hematology and Hematopoietic Cell Transplantation, City of Hope National Medical Center

[4]Department of Medicine, Memorial Sloan Kettering Cancer Center

[5]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center

[6]Molecular Microbiology Facility, Department of Immunology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center

[7]Adult Bone Marrow Transplantation Service, Department of Medicine, Memorial Sloan Kettering Cancer Center

[8]Department of Medicine, Weill Cornell Medical College

[9]Hematologic Malignancies Research Institute, City of Hope National Medical Center

Abstract

Correlating time-dependent patient characteristics and matched microbiome samples can be helpful to identify biomarkers in longitudinal microbiome studies. Existing approaches typically repeat a pre-specified modeling approach for all taxonomic features, followed by a multiple testing adjustment step for false discovery rate (FDR) control. In this work, we develop an alternative strategy of using log-ratio penalized generalized estimating equations, which directly models the longitudinal patient characteristic of interest as the outcome variable and treats microbial features as high-dimensional compositional covariates. A cross validation procedure is developed for variable selection and model selection among different working correlation structures. In extensive simulations, the proposed method achieved superior sensitivity over the state-of-the-art methods with robustly controlled FDR. In the analyses of correlating longitudinal dietary intake and microbial features from matched samples of cancer patients, the proposed method effectively identified gut health indicators and clinically relevant microbial markers, showing robust utilities in real-world applications. The method is implemented under the open-source R package `FLORAL`, which is available at (https://vdblab.github.io/FLORAL/).

1 Introduction

Longitudinal patient data collection has become increasingly more prevalent in microbiome studies, where microbial samples are paired with longitudinal patient data, such as dietary intake [1–3], body mass index (BMI) [4], blood counts [5], immune cell measurements [6], and metabolite abundance [7]. Rich longitudinal clinical data offer valuable opportunities to explore microbial associations with various temporal variables from the clinical side, which further assists hypothesis generation in basic biological research which can be facilitated by mouse models derived based on the clinical observations. For example, it is of interest to investigate the associations between microbial taxa and the amount of fiber intake for patients undergoing allogeneic hematopoietic cell transplantation (allo-HCT), which will contribute to identifying strategies for dietary interventions.

Despite the availability of longitudinal microbiome data, the relevant literature in statistical and computational methods is limited. Feature selection methods have been

[*] Corresponding author, email: feit1@mskcc.org, lead contact

[†] These authors contributed equally to this work as co-last authors.

proposed to correlate longitudinal microbial features with a continuous, binary, or time-to-event disease outcome observed after the last time point of sample collection [8, 9], which is not generalizable to the paired longitudinal microbiome and longitudinal patient characteristic data scenario. A versatile dimension reduction method was developed to perform temporal tensor decomposition for the taxa trajectories with respect to taxa, individual, and temporal patterns [10], which demonstrated utility in correlating pre-specified patient groups and longitudinal microbial patterns as temporal loadings. Nevertheless, the method did not provide an explicit feature selection approach which incorporates dynamically changing patient characteristics. In contrast to the above methods, the mixed-effect model is a more widely applied class of methods that effectively incorporates longitudinal patient characteristics into the models of individual taxon trajectories [11–13]. Typically, the same mixed-effect model structure is repeatedly applied to model all taxonomic features, followed by a false discovery rate (FDR) control procedure across all models to identify significant associations between longitudinal patient characteristics and taxa. In practice, however, different taxa may have largely varying stability across repeated observations and highly variable patient-specific distributions, making it challenging to apply a single model configuration (for example, linear time effect and random intercept) for hundreds of taxa. Additionally, the sparse and compositional nature of microbiome data brings additional challenges in modeling the abundance trajectories of taxa, which motivated the use of complex modeling strategies to account for zero-inflation, over-dispersion and potentially non-linear associations [13]. As a common alternative to the mixed-effect model, generalized estimating equations (GEE) have also been applied to study microbial associations with longitudinal characteristics [14], yet the method focused on inferring on global microbial associations instead of taxa-level associations.

In this work, we propose a penalized log-ratio GEE model to select longitudinal microbial features associated with a longitudinal patient outcome. Here, the *outcome* variable refers to any patient characteristics collected at roughly the same time as each microbiome sample, such as dietary intake, BMI, CD4 T-cell count from flow cytometry, or the concentration of a certain short-chain fatty acid from a metabolic assay (**Fig.1A**). Unlike the widely applied mixed-effect models, we treat the longitudinal patient characteristic as the outcome variable and the microbial taxa as covariates in a multivariable regression framework. As shown in **Fig.1B**, the proposed GEE model accounts for the within-subject dependency of the outcome variable with common working correlation structures such as independence, compound symmetry, and autoregressive (AR)-1 structures. Similar to the "fitting a log-ratio lasso" (`FLORAL`) regression framework we previously developed [9], we assume a sparse set of taxa are associated with the longitudinal outcome, where the penalized estimating procedure is extended from the standard approaches [15, 16] by adding the zero-sum constraint to account for the compositional nature of the covariates. We develop a model and variable selection procedure based on the cross-validated deviance residual [17] with two-step feature filtering to further control the false discovery rate (FDR) [9, 18]. The method is publicly available as a new module within the R package `FLORAL`.

Compared to the mixed-effect models, the proposed penalized log-ratio GEE model addresses several challenges of modeling individual taxon trajectories by flipping the roles of longitudinal taxa and longitudinal patient characteristics in the regression framework. Instead of modeling the highly sparse, volatile, and compositional taxa features, we focus on modeling the more tractable and stable patient characteristics which can be conveniently depicted by Gaussian or binomial link functions. In addition, the log-ratio covariate space effectively transforms zero-inflated quantities into a continuous variable space, which mitigates the computational burdens caused by complex models. Moreover, the proposed approach focuses on modeling the marginal expectation of one fixed outcome variable, where the non-linear associations between the outcome variable and time can be easily captured by using splines [19] without specifying the forms of subject-specific random effects (e.g. random intercepts and random slopes). Finally, the cross-validated variable selection process of the proposed method is more data-driven than the existing methods which are based on a pre-specified threshold of significance. We demonstrate by extensive simulations that the proposed method requires smaller number of patients or samples to achieve similar variable selection performances as the mixed-effect models while controlling for FDR. In real-data analyses, the proposed method identifies meaningful associations between fiber intake and taxa abundance from two studies conducted at Memorial Sloan Kettering Cancer Center (MSK) with different patient populations [1, 2], showing strong practical utilities in detecting clinically relevant microbial markers.

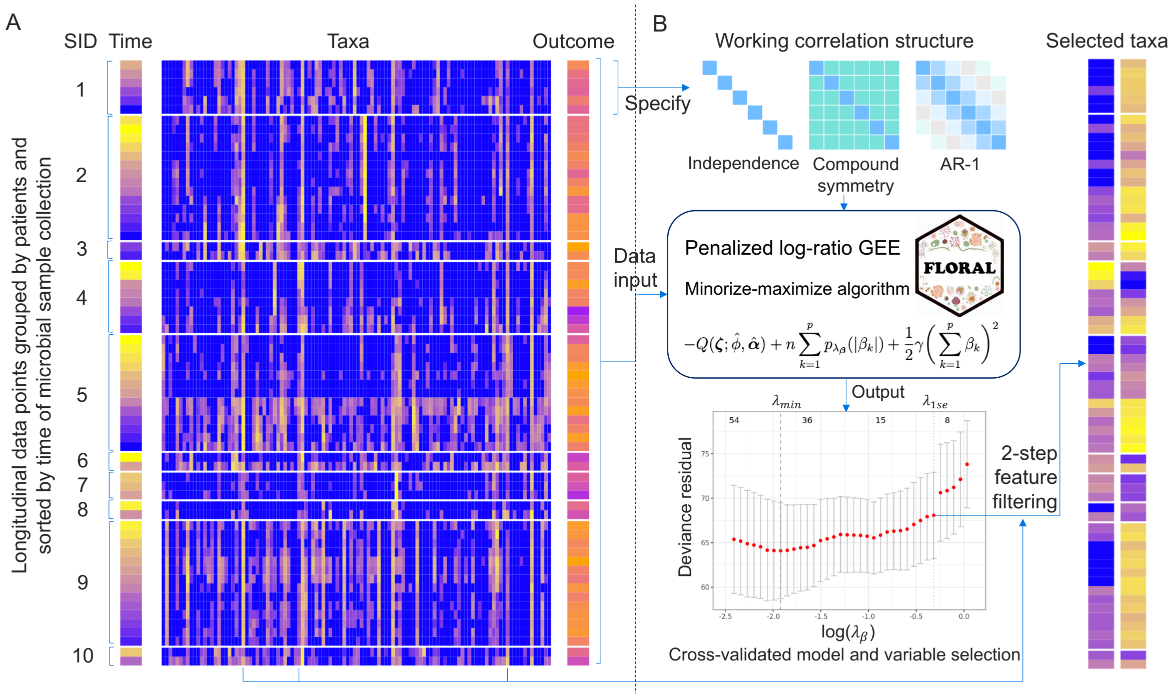
A
SID
Time
Taxa
Outcome
Longitudinal data points grouped by patients and sorted by time of microbial sample collection
1
2
3
4
5
6
7
8
9
10
B
Working correlation structure
Selected taxa
Specify
Independence
Compound symmetry
AR-1
Data input
Penalized log-ratio GEE
FLORAL
Minorize-maximize algorithm
$-Q(\boldsymbol{\zeta};\hat{\phi},\hat{\boldsymbol{\alpha}})+n\sum_{k=1}^{p}p_{\lambda_\beta}(|\beta_k|)+\frac{1}{2}\gamma\left(\sum_{k=1}^{p}\beta_k\right)^2$
Output
λ_{min}
λ_{1se}
54
36
15
8
Deviance residual
75
70
65
60
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
$\log(\lambda_\beta)$
Cross-validated model and variable selection
2-step feature filtering

Figure 1: Flow chart of using the proposed penalized log-ratio generalized estimating equation (GEE) approach. **A.** Heatmaps of the input data based on 10 randomly chosen patients from the MSK allo-HCT cohort, including longitudinal taxa features, the corresponding longitudinal outcome variable (fiber intake), and time of sample collection, where brighter colors represent larger numerical values. Data points were grouped by artificially assigned subject IDs (SID) and sorted by time of sample collection. **B.** The pipeline of the proposed method. First, a user-specified working correlation structure is required for the GEE model. Then the penalized log-ratio GEEs are solved by a minorize-maximize algorithm with a zero-sum constraint. Finally, cross validations are used to determine penalty parameters ($\lambda_{\text{min}}, \lambda_{\text{1se}}$) based on deviance residual, where the selected features will be further screened by an additional ratio-based procedure (2-step filtering). The heatmaps of two selected taxa were shown on the right.

2 Methods

2.1 Notations and model formulations

Let Y_{ij} denote the jth observation of the longitudinal outcome for the ith subject, $i = 1, \ldots, n, j = 1, \ldots, m_i$. Let $\boldsymbol{X}_{ij}$ denote the associated $p \times 1$ microbial count vector and $\boldsymbol{W}_{ij}$ denote the $L \times 1$ confounder feature vector. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_i})$, and $\boldsymbol{W}_i = (\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{im_i})$. While the number of observations m_i varies across different subjects in practice, we assume $m_i = m < \infty$ without loss of generality in the following description of the proposed model.

In the proposed generalized estimating equation (GEE) model for compositional covariates, we adapt the framework of log-contrast framework [9, 20] to model the mean and the variance of the longitudinal outcome Y_{ij}:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &\equiv \mu_{ij}(\boldsymbol{\zeta}) = g(\sum_{k=1}^{p} \beta_k \log X_{ij,k} + \sum_{l=1}^{L} \omega_l W_{ij,l}) \\ \mathrm{Var}(Y_{ij}) &\equiv \phi v_{ij}(\boldsymbol{\zeta}) = \phi g'(\sum_{k=1}^{p} \beta_k \log X_{ij,k} + \sum_{l=1}^{L} \omega_l W_{ij,l}) \\ &\text{subject to} \sum_{k=1}^{p} \beta_k = 0, \end{aligned} \tag{1}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\omega}^T)^T$ is the vector of unknown regression coefficients, including $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ for log-transformed compositional features and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)^T$ for non-compositional features. In addition, $g(\cdot)$ is a differentiable link function, and ϕ is a scaling factor which can be assumed as fixed or to be estimated. We impose a zero-sum constraint $\sum_{k=1}^{p} \beta_k = 0$ for the unknown coefficients $\boldsymbol{\beta}$ associated with the log-count features $\log(\boldsymbol{X}_{ij})$, which makes the model equivalent to an unconstrained linear model of all possible log-ratios of the compositional features [9, 20], thus accounting for the compositional nature of the features. Similar to the GEE model [21], we also consider a working correlation structure to depict the correlations within the repeated measurements $\boldsymbol{Y}_i$ from the same individual or cluster, where the working correlation matrix of $\boldsymbol{Y}_i$ is denoted by $\boldsymbol{R}(\boldsymbol{\alpha})$. Popular choices of $\boldsymbol{R}(\boldsymbol{\alpha})$ include independence, compound symmetry (or exchangeable), or autocorrelation (AR)-1, where $\boldsymbol{\alpha}$ follows different configurations. It then follows that $\boldsymbol{V}_i(\boldsymbol{\zeta}, \phi, \boldsymbol{\alpha}) = \phi \boldsymbol{A}_i^{1/2}(\boldsymbol{\zeta}) \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}(\boldsymbol{\zeta})$ is the variance-covariance matrix of $\boldsymbol{Y}_i$, where $\boldsymbol{A}_i(\boldsymbol{\zeta}) = \mathrm{diag}\{v_{i1}(\boldsymbol{\zeta}), \ldots, v_{im}(\boldsymbol{\zeta})\}$. To investigate the association between longitudinal compositional features and the corresponding outcomes, the main parameter of interest is the effect size vector $\boldsymbol{\beta}$, while ϕ and $\boldsymbol{\alpha}$ are treated as nuisance parameters in the estimation procedure. In practice, the number of compositional features p can be larger than the number of subjects n and the number of samples $n \times m$. We assume that only a sparse set of the features are associated with the outcome, which means the majority of the elements in $\boldsymbol{\beta}$ are zeros.

2.2 Estimation procedure

Given the above formulation of the first two moments of $\boldsymbol{Y}$, we obtain an unbiased constrained estimating function for $\boldsymbol{\zeta}$ given ϕ and $\boldsymbol{\alpha}$

$$\boldsymbol{S}(\boldsymbol{\zeta};\phi,\boldsymbol{\alpha}) \equiv \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}}\right)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta},\phi,\boldsymbol{\alpha})\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\zeta})\} = 0, \quad \text{subject to} \sum_{k=1}^{p} \beta_k = 0, \quad (2)$$

which follows the same form as the classic GEE [21] with an additional zero-sum constraint. To impose sparsity of $\boldsymbol{\beta}$, one natural approach is to consider a regularized regression framework such as lasso [22]. However, unlike the standard lasso regression which minimizes a penalized negative log-likelihood function, our model assumption (1) does not have an explicit likelihood function to construct an optimization problem. Instead, the proposed estimating equation (2) is a constrained zero point finding problem. Therefore, we adapt alternative strategies for penalized estimating equations [15] to fulfill the regularization of the coefficients. Specifically, we extended the penalized generalized estimating equations (PGEE) [16] framework, where we incorporated the zero-sum constraint into the PGEE, established a more systematic model and feature selection mechanism via cross-validation, and developed easily accessible software for wide applications.

The PGEE function with zero-sum constraint is formulated as

$$\boldsymbol{U}(\boldsymbol{\zeta}) = \boldsymbol{S}(\boldsymbol{\zeta};\hat{\phi},\hat{\boldsymbol{\alpha}}) - n \begin{pmatrix} \boldsymbol{q}_{\lambda_{\boldsymbol{\beta}}}(|\boldsymbol{\beta}|) \odot \operatorname{sign}(\boldsymbol{\beta}) - \mathbf{1}_p(\gamma \sum_{k=1}^{p} \beta_k) \\ \boldsymbol{q}_{\lambda_{\boldsymbol{\omega}}}(|\boldsymbol{\omega}|) \odot \operatorname{sign}(\boldsymbol{\omega}) \end{pmatrix}, \quad (3)$$

where $\boldsymbol{S}(\boldsymbol{\zeta};\hat{\phi},\hat{\boldsymbol{\alpha}})$ is the estimating function as defined in (2), where $\hat{\phi}$ and $\hat{\boldsymbol{\alpha}}$ are estimates of ϕ and $\boldsymbol{\alpha}$ to be updated at each iteration of the algorithm. Here, $\boldsymbol{q}_{\lambda_{\boldsymbol{\beta}}}(|\boldsymbol{\beta}|) = \{q_{\lambda_{\boldsymbol{\beta}}}(|\beta_1|), \ldots, q_{\lambda_{\boldsymbol{\beta}}}(|\beta_p|)\}^T$ and $\boldsymbol{q}_{\lambda_{\boldsymbol{\omega}}}(|\boldsymbol{\omega}|) = \{q_{\lambda_{\boldsymbol{\omega}}}(|\omega_1|), \ldots, q_{\lambda_{\boldsymbol{\omega}}}(|\omega_L|)\}^T$ are penalty functions corresponding to each element of $\boldsymbol{\zeta}$, where the non-negative penalty parameters $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\omega}}$ are separately specified for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ to flexibly impose penalties to compositional features while adjusted for covariates. In addition, $\odot$ denotes the element-wise multiplication operator, $\text{sign}(\boldsymbol{\beta}) = \{\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p)\}^T$ and $\text{sign}(\boldsymbol{\omega}) = \{\text{sign}(\omega_1), \ldots, \text{sign}(\omega_L)\}^T$, where $\text{sign}(x) = I(x > 0) - I(x < 0)$. The term $\mathbf{1}_p(\gamma \sum_{k=1}^p \beta_k)$ corresponds to the zero-sum constraint, where $\mathbf{1}_p$ is a p-vector with all elements equal to one and γ is the penalty parameter which governs the strength of the constraint $\sum_{k=1}^p \beta_k = 0$. The proposed estimator $\hat{\boldsymbol{\zeta}}$ satisfies $\boldsymbol{U}(\hat{\boldsymbol{\zeta}}) = \mathbf{0}$.

The formulation of (3) can be derived from the following target function of an optimization problem

$$-Q(\boldsymbol{\zeta}; \hat{\phi}, \hat{\boldsymbol{\alpha}}) + n \sum_{k=1}^{p} p_{\lambda_{\boldsymbol{\beta}}}(|\beta_k|) + n \sum_{l=1}^{L} p_{\lambda_{\boldsymbol{\omega}}}(|\omega_l|) + \frac{1}{2}\gamma \left(\sum_{k=1}^{p} \beta_k \right)^2.$$

Here $Q(\boldsymbol{\zeta}; \phi, \boldsymbol{\alpha})$ is the quasi log-likelihood corresponding to $\boldsymbol{S}(\boldsymbol{\zeta}; \phi, \boldsymbol{\alpha})$, which satisfies $\frac{\partial}{\partial \boldsymbol{\zeta}} Q(\boldsymbol{\zeta}; \phi, \boldsymbol{\alpha}) = \boldsymbol{S}(\boldsymbol{\zeta}; \phi, \boldsymbol{\alpha})$. Functions $p_{\lambda_{\boldsymbol{\beta}}}(\cdot)$ and $p_{\lambda_{\boldsymbol{\omega}}}(\cdot)$ determine the form of the penalties, such as L_1 or L_2 penalty terms, which satisfy $p'_{\lambda_{\boldsymbol{\beta}}}(x) = q_{\lambda_{\boldsymbol{\beta}}}(x)$ and $p'_{\lambda_{\boldsymbol{\omega}}}(x) = q_{\lambda_{\boldsymbol{\omega}}}(x)$. For example, $p_\lambda(|x|) = \lambda |x|$ and $q_\lambda(|x|) = \lambda$ for lasso penalty [22], while $q_\lambda(|x|) = \lambda\{I(|x| < \lambda) + \frac{(a\lambda - |x|)_+}{(a-1)\lambda} I(|x| \geq \lambda)\}$ for SCAD penalty [23] with $a > 2$. The term $\frac{1}{2}\gamma(\sum_{k=1}^p \beta_k)^2$ utilizes the penalty method to penalize the zero-sum constraint with the penalty parameter γ [24]. It can be shown that (3) approximates the negative differentiation of the above target function with respect to $\boldsymbol{\zeta}$, such that we transform the optimization problem to the estimating equation solving problem of $\boldsymbol{U}(\boldsymbol{\zeta}) = \mathbf{0}$. In practice, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are set to be penalized by the same type of penalty functions. In addition, $\lambda_{\boldsymbol{\omega}}$ is set as a fraction of $\lambda_{\boldsymbol{\beta}}$ with $\lambda_{\boldsymbol{\omega}} = r\lambda_{\boldsymbol{\beta}}, r \in [0, 1]$. The penalty parameter γ is set as 10^5 to numerically enforce the zero-sum constraint.

As described in [15], a minorize-maximize (MM) algorithm with local quadratic approximations for $\boldsymbol{q}_\lambda(|\cdot|)\text{sign}(\cdot)$ is used to obtain the estimates $\hat{\boldsymbol{\zeta}}$. In the following, we use subscript $\{i\}$ to denote the index of pathwise solutions with respect to a series of $\lambda_{\boldsymbol{\beta}}$, and use superscript (j) to denote the index of iterations of the MM algorithm with a fixed $\lambda_{\boldsymbol{\beta}}$. Like lasso regression, we obtain estimates $\hat{\boldsymbol{\zeta}}_{\{i\}}, i = 1, \ldots, b$ along a decreasing path of $\lambda_{\boldsymbol{\beta}\{1\}}, \ldots, \lambda_{\boldsymbol{\beta}\{b\}}$, where the largest parameter $\lambda_{\boldsymbol{\beta}\{1\}}$ is determined by a standard formula used for lasso regression while treating all observations independent of each other [9], while the value of the smallest parameter $\lambda_{\boldsymbol{\beta}\{b\}}$ is by default set as $0.01\lambda_{\boldsymbol{\beta}\{1\}}$. The initial value $\hat{\boldsymbol{\zeta}}^{(0)}_{\{1\}}$ associated with $\lambda_{\boldsymbol{\beta}\{1\}}$ is set as zeros, while the initial value $\hat{\boldsymbol{\zeta}}^{(0)}_{\{i\}}$ associated with $\lambda_{\boldsymbol{\beta}\{i\}}$ can be set as the "warm-start" estimates $\hat{\boldsymbol{\zeta}}_{\{i-1\}}$ associated with $\lambda_{\boldsymbol{\beta}\{i-1\}}$. Given $\lambda_{\boldsymbol{\beta}}$

and estimate $\hat{\boldsymbol{\zeta}}^{(j)}$ after jth iteration, the parameters are updated in the $(j+1)$th iteration as following. First, let $\boldsymbol{\varepsilon}_i^{(j)} = \{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\zeta}}^{(j)})\} \odot \text{vecdiag}\{\boldsymbol{A}_i(\hat{\boldsymbol{\zeta}}^{(j)})\}^{-\frac{1}{2}}$ be the Pearson residual of the ith subject, the we update $\hat{\phi}^{(j+1)} = \sum_{i=1}^n \boldsymbol{\varepsilon}_i^{(j)T}\boldsymbol{\varepsilon}_i^{(j)}/(nm)$ and $\hat{\boldsymbol{\alpha}}^{(j+1)}$ as described in Table 45.13 of [25]. Then $\hat{\boldsymbol{\zeta}}$ is updated as

$$\begin{aligned}
&\hat{\boldsymbol{\zeta}}^{(j+1)} = \hat{\boldsymbol{\zeta}}^{(j)} + \mathcal{A}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}})^{-1}\mathcal{B}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}}), \text{where}\\
&\mathcal{A}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}}) = \boldsymbol{H}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}}) + n\boldsymbol{E}(\hat{\boldsymbol{\zeta}}^{(j)}) + \gamma\boldsymbol{J};\\
&\mathcal{B}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}}) = \boldsymbol{S}(\hat{\boldsymbol{\zeta}}^{(j)}, \hat{\phi}, \hat{\boldsymbol{\alpha}}) - n\boldsymbol{E}(\hat{\boldsymbol{\zeta}}^{(j)})\hat{\boldsymbol{\zeta}}^{(j)} - \begin{pmatrix} \mathbf{1}_p(\gamma\sum_{k=1}^p \hat{\beta}_k^{(j)}). \\ \mathbf{0}_L \end{pmatrix}
\end{aligned} \tag{4}$$

Here, $\boldsymbol{J}$ is a $(p+L) \times (p+L)$ matrix with the upper left $p \times p$ entries equal to one and zero elsewhere, $\mathbf{0}_L$ is a L-vector with all entries equal to zero,

$$\boldsymbol{H}(\boldsymbol{\zeta}, \phi, \boldsymbol{\alpha}) = \sum_{i=1}^n \phi^{-1}\boldsymbol{X}_i^T\boldsymbol{A}_i^{-\frac{1}{2}}(\boldsymbol{\zeta})R(\boldsymbol{\alpha})^{-1}\boldsymbol{A}_i^{-\frac{1}{2}}(\boldsymbol{\zeta})\boldsymbol{X}_i$$

and

$$\boldsymbol{E}(\boldsymbol{\zeta}) = \text{diag}\left\{\frac{q_{\lambda_{\boldsymbol{\beta}}}(|\beta_1|)}{\epsilon + |\beta_1|}, \ldots, \frac{q_{\lambda_{\boldsymbol{\beta}}}(|\beta_p|)}{\epsilon + |\beta_p|}, \frac{q_{\lambda_{\boldsymbol{\omega}}}(|\omega_1|)}{\epsilon + |\omega_1|}, \ldots, \frac{q_{\lambda_{\boldsymbol{\omega}}}(|\omega_L|)}{\epsilon + |\omega_L|}\right\}$$

is a $(p+L) \times (p+L)$ diagonal matrix such that $\boldsymbol{E}(\boldsymbol{\zeta})\boldsymbol{\zeta}$ is a quadratic approximation of $\{\boldsymbol{q}_{\lambda_{\boldsymbol{\beta}}}(|\boldsymbol{\beta}|) \odot \text{sign}(\boldsymbol{\beta})^T, \boldsymbol{q}_{\lambda_{\boldsymbol{\omega}}}(|\boldsymbol{\omega}|) \odot \text{sign}(\boldsymbol{\omega})^T\}^T$. It can be shown that that $\mathcal{A}(\boldsymbol{\zeta}, \hat{\phi}, \hat{\boldsymbol{\alpha}})$ is the derivative matrix of $\mathcal{B}(\boldsymbol{\zeta}, \hat{\phi}, \hat{\boldsymbol{\alpha}})$ with respect to $\boldsymbol{\zeta}$, forming a close connection to the Newton-Raphson algorithm. In our implementation, we set $\epsilon = 10^{-6}$ and set the convergence criterion as $\|\boldsymbol{\zeta}^{(j+1)} - \boldsymbol{\zeta}^{(j)}\|_\infty < 10^{-3}$. After reaching convergence or exceeding 100 iterations for a given $\lambda_{\boldsymbol{\beta}}$, parameter estimates with absolute values smaller than 10^{-3} are set as zeros as suggested by [16, 26]. The algorithm described above is summarized as Algorithm 1, which is implemented in R package `FLORAL` with `RcppArmadillo` [27].

2.3 Variable selection via cross validation and 2-step filtering

We utilize K-fold cross validation to determine the values of the penalty parameters. For each choice of the penalty parameter $\lambda_{\boldsymbol{\beta}}$, we split the data into K folds by individual identifiers, such that all observations from a certain individual will be assigned to the same fold. Then for $k = 1, \ldots, K$, the proposed model is fitted using all except the kth fold, then we calculate the deviance residual [17] according to the distribution family used by the GEE for all observations from the kth fold. The cross-validated average deviance residual is then used for model selection, where the penalty parameter achieving the

Algorithm 1 Iterative optimization algorithm for solving $\boldsymbol{U}(\hat{\boldsymbol{\zeta}}) = \mathbf{0}$ with given $\lambda_{\boldsymbol{\beta}}$ and γ. Note that the following algorithm assumes no intercept term. The algorithm with intercept term can be derived similarly. $\odot$ denotes element-wise multiplication. Given a general square matrix $\mathbf{M}$, the vecdiag$(\mathbf{M})$ operator creates a vector whose elements are on the diagonal of the matrix.

```
Input: Initial value of $\hat{\boldsymbol{\zeta}} = \tilde{\boldsymbol{\zeta}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\omega}}^T)^T$; $n \times (p+L)$ matrix $\boldsymbol{Z} = \{\log(\boldsymbol{X})^T, \boldsymbol{W}^T\}^T$; penalty parameters $\lambda_{\boldsymbol{\beta}}, \gamma$; penalty ratio $r$; tolerance parameter $\delta$; maximum iteration number $v$; Pre-specified working correlation structure (independence, exchangeable, or AR-1); Penalty function $p(\cdot)$ and $q(\cdot)$
Set $\hat{\boldsymbol{\zeta}}^{(0)} = \tilde{\boldsymbol{\zeta}}$, $j = 0$, $d_{\boldsymbol{\zeta}} = 1$, $\lambda_{\boldsymbol{\omega}} = r\lambda_{\boldsymbol{\beta}}$
while $d_{\boldsymbol{\zeta}} > \delta$ and $j \leq v$ do
    Set $j = j + 1$
    Update $\boldsymbol{\varepsilon}_i^{(j-1)} = \{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\zeta}}^{(j-1)})\} \odot \text{vecdiag}\{\boldsymbol{A}_i(\hat{\boldsymbol{\zeta}}^{(j-1)})\}^{-\frac{1}{2}}$
    Update $\hat{\phi}^{(j)} = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^{(j-1)T} \boldsymbol{\varepsilon}_i^{(j-1)}/(nm)$ and $\hat{\boldsymbol{\alpha}}^{(j)}$ according to the working correlation structure, as described in Table 45.13 of [25], where the average is taken without subtracting the degree of freedom.
    Update $\hat{\boldsymbol{\zeta}}^{(j)}$ by (4), with $\hat{\phi}^{(j)}$ and $\hat{\boldsymbol{\alpha}}^{(j)}$ plugged in.
    Set $d_{\boldsymbol{\zeta}} = \|\hat{\boldsymbol{\zeta}}^{(j)} - \hat{\boldsymbol{\zeta}}^{(j-1)}\|_{\infty}$
end while
Set $\{\boldsymbol{\zeta}_k : |\boldsymbol{\zeta}_k| < 10^{-3}\}$ as zeros
Output: $\hat{\boldsymbol{\zeta}}^{(j)}$
```

smallest cross-validated deviance residual (λ_{min}) and the largest penalty parameter with its deviance residual within one standard-error of the smallest deviance residual (λ_{1se}) are widely used choices [9, 22]. Subsequently, we report the features with non-zero regression coefficients at λ_{min} and λ_{1se}. The same cross-validation procedure is also used for model selection across different choices of working correlation structures, where the working correlation structure achieving the smallest cross-validated deviance residual is treated as the best option.

Following the cross-validated feature selection, we further implement the step-2 feature selection procedure [9, 18]. Specifically, all pairs of log-ratios based on selected features at λ_{min} (and λ_{1se}) are refitted by the PGEE method without the zero-sum constraint to achieve a higher sparsity of feature selection and ratio-based model interpretation. Similar to the other regularized log-ratio regression methods implemented in the `FLORAL` package, the PGEE model also allows users to repeat the cross validation steps for multiple times with random fold splits, then summarize the frequency of variable selection out of all the repeats.

2.4 Method assessment and benchmarking

We conducted extensive simulations and real-data analysis to study and benchmark the performance of the proposed constrained PGEE method under various scenarios. Due to the scarcity of tools developed for feature selection for the longitudinal microbiome sample - longitudinal outcome data structure, we mainly focused on comparing feature selection performances within the scope of PGEE. For log-transformed data, we compared the zero-sum constrained PGEE model implemented by `FLORAL` and the standard unconstrained PGEE model. Then we also investigated the performance of PGEEs with relative abundance data and CLR-transformed data. We also applied the popular `MaAsLin2` and `MaAsLin3` packages, which use mixed-effect regression models of microbial abundance over longitudinal covariates, to better understand the pros and cons of the proposed clinical outcome-oriented modeling strategy and the popular taxa-oriented modeling strategy.

2.4.1 Simulations

Let n be the number of individuals, m be the number of samples per individual, and p be the number of features. Longitudinal microbiome samples were simulated following a similar approach as described in [9]. We assume that the samples were observed at time $t = 0, \ldots, m-1$ for each individual. First, we simulated the true count of longitudinal microbiome data $\boldsymbol{C}$ based on a logistic-normal model [20], then the log-ratios consisting of the first ten features were utilized to generate the longitudinal outcome $\boldsymbol{Y}$ with a given correlation structure within each individual. Finally, the observable count data $\boldsymbol{X}$ was generated based on the true count data $\boldsymbol{C}$ and a randomly simulated sequencing depth for each sample. We considered scenarios with continuous outcomes and binary outcomes.

For the ith individual at time t, we first generate a vector $\boldsymbol{x}_i(t) = \{\boldsymbol{x}_{i1}(t), \ldots, \boldsymbol{x}_{ip}(t)\}$ from a p-variate normal distribution $N_p\{\boldsymbol{\xi}(t), \boldsymbol{\Sigma}(t)\}$ with $\boldsymbol{\xi}(t) = \{\xi_1(t), \ldots, \xi_p(t)\}^T$. We let $\xi_k(0) = \log p$ for $k = 1, 2, 3, 5, 6, 8$ and otherwise $\xi_k(0) = 0$, such that there are six features with higher abundance at time 0. In addition, we assume

$$\xi_k(t) = \begin{cases} \xi_k(0) + 0.5t & \text{if } k \in \{2, 4, 6, 8, 10\}, \\ \xi_k(0) & \text{otherwise,} \end{cases}$$

where 0.5 is the slope with respect to time for five pre-specified features. Regarding the covariance parameter $\mathbf{\Sigma}(t)$, we set the variances $\sigma_k(t)^2 = \sqrt{\log p/2}$ for $k = 1, 2, 3, 5, 6, 8$ and otherwise 1, such that the features with higher baseline abundances also have higher variations throughout the follow-up. We also set the covariance $\mathbf{\Sigma}_{j,k}(t) = \rho^{|j-k|}, \rho \in [0, 1)$ between features j and k. For the same individual, we also impose a sample-wise correlation of 0.4 for the first ten features. Additionally, we specify a sparsity level of 0.8 and randomly let 80% of the entries in $\boldsymbol{x}_i$ to be $-\infty$ to create zeros in compositions. After the above steps, we obtain the unobservable underlying time-dependent composition vector $\boldsymbol{c}_i(t)$ for $t = 0, \ldots, m-1$, where the kth entry satisfies $c_{ik}(t) = \exp\{x_{ik}(t)\}/\sum_d \exp\{x_{id}(t)\}$. With the true composition $\boldsymbol{c}_i(t)$, we assume that the total count is 10^6 for each sample and generate the true count vector $\boldsymbol{C}_i(t)$ from a multinomial distribution with 10^6 counts and probability vector $\boldsymbol{c}_i(t)$.

With true count vector $\boldsymbol{C}_i(t)$, we calculate the underlying true linear predictor $\boldsymbol{l}_i = \{l_i(0), \ldots, l_i(m-1)\}^T$ which consists of five log-ratios from the first ten features

$$l_i(t) = 0.5u\left\{\log\frac{C_{i1}(t)+1}{C_{i2}(t)+1} + \log\frac{C_{i3}(t)+1}{C_{i4}(t)+1}\right\} + u\left\{\log\frac{C_{i5}(t)+1}{C_{i6}(t)+1} + \log\frac{C_{i7}(t)+1}{C_{i8}(t)+1} + \log\frac{C_{i9}(t)+1}{C_{i10}(t)+1}\right\} + \kappa t.$$

Here, a pseudo value 1 is added to all counts to make sure the log-transformations are well defined, κ is the constant time effect on the longitudinal outcome, and effect size u governs the strength between compositional features and the associated longitudinal outcome $Y_i(t)$. Given $l_i(t)$, the continuous outcome vector $\boldsymbol{Y}_i = \{Y_i(0), \ldots, Y_i(m-1)\}^T$ is generated from $\boldsymbol{Y}_i = \boldsymbol{l}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ follows a zero-mean normal distribution with variance one and an exchangeable correlation 0.8 across repeated measurements. For binary outcomes, we adapt the approach from [28] to generate correlated binary outcomes from the probability vector $\{1+\exp\{-l_i(t)\}\}^{-1}, t = 0, \ldots, m-1$ with an exchangeable correlation 0.8 within the same individual. After generating the outcome $\boldsymbol{Y}_i$ by the underlying true count $\boldsymbol{C}_i(t)$, the observable count vector $\boldsymbol{X}_i(t)$ for each individual i and time t is simulated from multinomial distribution with probability vector $\boldsymbol{c}_i(t)$ and a randomly simulated sequencing depth as the largest integer smaller than a $Unif(5000, 50000)$ random variable. For the ith individual, longitudinal outcomes $Y_i(t)$, observable compositional features $\boldsymbol{X}_i(t)$ and the time vector $t, (t = 0, \ldots, m-1)$ are used in model fitting.

We considered multiple scenarios which focused on different aspects of simulated data characteristics. For simulations with continuous outcomes, we set the reference scenario with $n = 50, p = 200, m = 3, u = 0.15, \rho = 0$ and $\kappa = 2.5$, which was based on the empirical observation of strong time effect on longitudinal outcome variables (such as BMI or dietary intake). Then we performed simulations with $n = 10, 20, 50, 100, 200$, $p = 100, 200, 500$, $m = 2, 3, 4, 6, 8$, $u = 0.1, 0.15, 0.25, 0.5$, $\rho = 0, 0.4, 0.8$ and $\kappa = 0, 1.5, 2.5$ while fixing other parameters as specified in the reference scenario. For simulations with binary outcomes, we set the reference scenario with $n = 60, p = 200, m = 3, u = 0.3, \rho = 0$ and $\kappa = 0$, where the sample size and effect size were higher than the reference scenario used for simulating the continuous outcome. This was because models for binary outcomes requires higher sample size or effect size to reach similar level of power as models for continuous models. We let $\kappa = 0$ in the reference scenario for binary outcome simulations because a strong time effect would result in very small variations of a longitudinal binary variable (i.e. constantly equal to zero or one) as time increases. Centered at the reference scenario, we conducted simulations with $n = 20, 30, 60, 100, 200$, $p = 100, 200, 500$, $m = 2, 3, 4, 6, 8$, $u = 0.15, 0.3, 0.45, 0.75$, $\rho = 0, 0.4, 0.8$ and $\kappa = 0, 1.5, 2.5$.

2.4.2 Real data examples

Longitudinal diet and microbiome data of the NUTRIVENTION study Elevated BMI and diets lacking plant foods are significant risk factors for multiple myeloma, which led to the development of a high fiber dietary intervention strategy. The MSK NUTRIVENTION study (NCT04920084) was a prospective trial investigating the efficacy of a high-fiber dietary intervention on weight loss and also whether it may delay progression from monoclonal gammopathy or smoldering myeloma to multiple myeloma [2]. The study recruited 20 evaluable patients who received 12 weeks of high fiber plant-based meals and 24 weeks of nutrition coaching with the meals and were followed for a year. Various patient characteristics, including but not limited to BMI, dietary intake, and 16S rRNA sequencing data from stool samples, were collected at 5 planned time points (baseline, 1 month, 3 months, 6 months, and 1 year) across a whole year of intervention. It has been shown that the intestinal alpha diversity was significantly increased from baseline to 3 months after study, where the longitudinal intestinal alpha diversity also had a significantly negative association with BMI [2]. The study showed that a high fiber dietary intervention could improve BMI and reshape the gut microbiome.

We applied the proposed method and other existing methods to correlate the longitudinal microbial taxa abundance with the longitudinal fiber intake collected from the 20 precursor plasma cell disorder patients receiving the high-fiber food intervention. 65 matched pairs of stool samples and fiber intake data points between baseline and 6 months after intervention were identified for the analysis, where each patient contributed 2-4 matched data points with a median of 3 per patient. The distribution of the available fiber intake values had a slightly heavy tail (**Fig.2A**), where the fiber intake was the highest 1 month after the start of intervention (**Fig.2C**). Such trajectories of fiber intake reflected patients' adherence to the intervention protocol, where the adherence was the highest shortly after the study was initiated. Based on the above observations, we conducted natural logarithm transformation to the grams of fiber intake for modeling (**Fig.2B**), while using natural cubic spline terms to capture the non-linear time effect. **Fig.2D** displays the distribution of taxa prevalence, where taxa with prevalence below 10% are excluded. It can be seen that there are many taxa with high prevalence across samples.

Longitudinal diet and microbiome data of MSK allo-HCT cohort To investigate the associations between dietary intake and the change of gut microbiota during bone marrow transplantation, the investigators at MSK collected longitudinal diet and 16S rRNA microbiome data for allo-HCT patients during the period of inpatient stay [8]. Specifically, food intake was categorized as five macronutrients (sugar, fiber, fat, protein, and other carbohydrates) in grams based on receipts from cafeteria and records from the care team to reflect food items and the amount of actual intake of each food item during each recorded meal. Accordingly, longitudinal stool samples were also collected for 16S rRNA sequencing. Based on a Procrustes analysis between microbial and macronutrient compositions, the alignment was the highest between between dietary records and stool samples collected 2 days later, as compared to other choices of gap days [8], which makes it possible to pair each dietary record with a stool sample based on the availability.

To demonstrate the utility of the proposed log-ratio PGEE method, we treated fiber

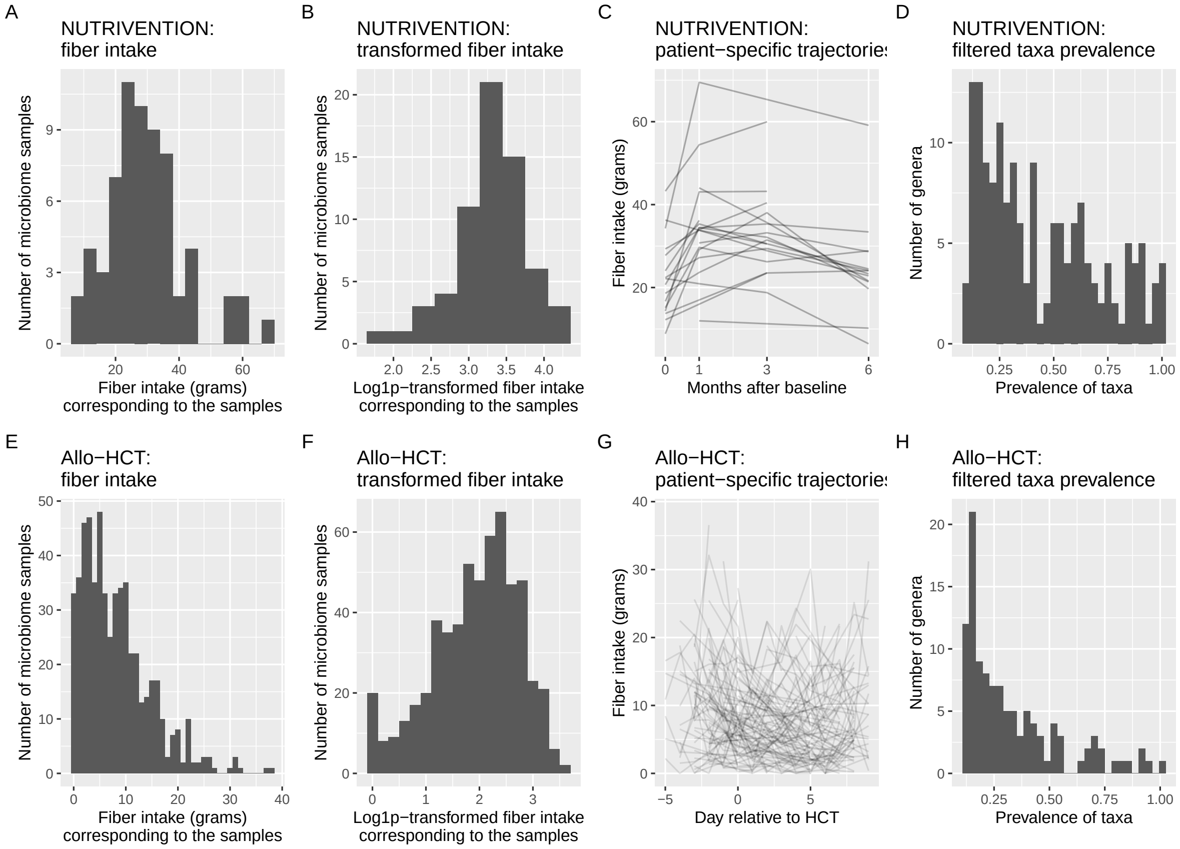

Figure 2: Data characteristics for the distribution of fiber intake values, fiber intake trajectories, and prevalence of filtered genera features. **A-D.** Histogram of fiber intake in grams (**A**), histogram of log-transformed fiber intake (**B**), spaghetti plot of patient specific fiber intake trajectories (**C**), and histogram of filtered taxa's prevalences (**D**) for patients in the NUTRIVENTION study. **E-H.** Histogram of fiber intake in grams (**E**), histogram of log$(\cdot+1)$-transformed fiber intake (**F**), spaghetti plot of patient specific fiber intake trajectories (**G**), and histogram of filtered taxa's prevalences (**H**) for patients in the allo-HCT cohort.

intake in grams as the outcome variable in the GEE model, while using the genera counts observed in the corresponding stool sample collected two days later as covariates, aiming to identify microbial markers associated with fiber intake. The analysis focused on the diet records collected between 7 days prior to transplantation (day -7) and 7 days after transplantation (day 7), which consisted of 505 patient-days of paired fiber-microbiome longitudinal samples from 137 unique patients collected between day -5 and day 9. The number of available paired samples varied across patients, ranging from a single pair to 10 pairs with the median of 3 pairs and the interquartile range between 2 and 5 pairs. Due to the conditioning chemotherapy given before the HCT, overall dietary intake, including fiber, declined rapidly from day -7 to day 0 followed by a slow recovery trajectory after day 0 (**Fig.2G**). The heavy shift of dietary pattern also caused highly skewed distribution of fiber intake with a long tail (**Fig.2E**). Therefore, we conducted natural logarithm(+1) transformation to the fiber intake values (**Fig.2F**) in the models and adjusted for the non-linear time effect by including the cubic natural spline of time as covariates. Due to the conditioning therapy prior to the transplantation, patients' gut microbiota were heavily shifted during the time window, where the prevalence was low for most taxa (**Fig.2H**).

2.4.3 Method configurations and assessment

Simulations We applied the proposed log-ratio PGEE model with two-step variable selection and other existing methods to benchmark the variable selection performance in simulations and real-data analysis. In simulation studies, log-ratio PGEE lasso models were fitted by the proposed method `FLORAL` with the correct compound symmetry working correlation structure (`FLORAL,cp`), and the incorrect independent (`FLORAL,ind`) and AR-1 (`FLORAL,ar1`) working correlation structures. In addition, we fitted standard PGEE models without the zero-sum constraint but with the correct compound symmetry working correlation structure using log-transformed count data (`PGEE,log`), relative abundance data (`PGEE,rel`), and centered log-ratio (CLR) transformed data (`PGEE,clr`). For the above penalized regression methods, a linear time effect was included in the GEE models without penalties ($r = 0$). Variable selection was performed by identifying the non-zero regression coefficients at λ_{min} and λ_{1se} based on a five-fold cross validation, where the fold split was set as identical for all methods within the same run. In addition, the implementation of the PGEE method without the zero-sum constraint was fulfilled by running the `FLORAL` function with $\gamma = 0$ to ensure identical cross-validation and variable selection procedures. As illustrated in the Introduction section, only a few existing methods can be applied to study associations between longitudinal microbial features and longitudinal outcomes [11–13]. We implemented the mixed-effect model `MaAsLin2` [11] and `MaAsLin3` [12] with CLR transformation (`normalization=‘CLR’, transform=‘NONE’`) or with total sum scaling (TSS) normalization and log-transformation (`normalization=‘TSS’,transform=‘LOG’`). For both methods, the mixed-effect models considered a linear time effect, a linear outcome effect from the simulated longitudinal outcome, and a random intercept for each simulated individual. Feature selection was based on FDR-adjusted p-values via the Benjamini-Hochberg approach [29] with significance level of 0.1. For `MaAsLin3`, we report features selected by the prevalence model and the abundance model separately.

In terms of method assessment, 100 runs were performed under each simulation scenario for performance evaluations. Four commonly applied metrics were calculated for each method, namely the F_1 score, number of false positive features, number of false negative features, and false discovery rate (FDR). The F_1 score is defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity), which is used to indicate the overall variable selection performance after balancing sensitivity and FDR. An F_1 score of 1 indicates perfect performance, while an F_1 score of 0 implies that no true features were selected. The cross-validated deviance residuals were also compared across the `FLORAL` models with different working correlation structures to evaluate its effectiveness of selecting models with the most appropriate working correlation structure. In addition, we also recorded the time (in seconds) used for each evaluated method to complete each simulation run, the number of iterations `FLORAL` took before convergence or reaching the maximum number of iterations, and the convergence criterion $\|\hat{\boldsymbol{\zeta}}^{(j)} - \hat{\boldsymbol{\zeta}}^{(j-1)}\|_\infty$ when `FLORAL`'s algorithm (**Algorithm 1**) stopped at each simulation run.

Real-data analysis As discussed in Section 2.4.2, we performed natural log-transformation to normalize the fiber intake data. Treating the transformed fiber intake as the outcome variable, we fitted the proposed log-ratio PGEE model with independent (`FLORAL,ind`), compound symmetry (`FLORAL,cs`), and AR-1 (`FLORAL,AR1`) working correlation structures. In addition, we also applied PGEE without zero-sum constraints with log-transformed (`PGEE,log`), relative abundance (`PGEE,rel`), and CLR-transformed (`PGEE,clr`) microbiome data. Cubic natural spline terms were included for both datasets as covariates without penalization ($r = 0$), where the knots were selected as the 10th, 50th and 90th quantiles of the time points of sample collection. Variable selection procedure follows the same procedure as described for the simulation studies with a 5-fold cross validation. Additionally, we conducted model fitting using the above PGEE methods for 100 times with random fold splits, then summarized the number of times for taxa being selected across 100 times as probabilities of being selected. Taxa with high frequency of selection will be interpreted as more likely to be associated with the fiber intake. We also applied mixed-effect models `MaAsLin2` and `MaAsLin3` by treating normalized or transformed fiber intake as a covariate. Similar to the penalized regression models, we also included the same cubic natural spline terms to capture non-linear time effects. For the MaAsLin packages, we applied the two taxa normalization-transformation configurations as used in simulations. Selected features are defined as features with FDR-adjusted p-values < 0.1. We also report both prevalence and abundance models for `MaAsLin3`.

We evaluated the methods based on their capabilities of detecting signals and the clinical relevance of selected microbial features. We also compared the features with the strongest signals from different methods, as ranked by the selection probabilities for the PGEE models and the p-values for the mixed-effect models.

3 Results

3.1 FLORAL achieves superior variable selection performances in simulations

We performed extensive simulations to assess the variable selection performance of the proposed log-ratio PGEE method `FLORAL`, the standard `PGEE` models with log-transformed, relative abundance, and CLR-transformed features, and mixed-effect models `MaAsLin2` and `MaAsLin3`. For both continuous and binary outcome simulations, we specified the compound symmetry (`cs`) working correlation structure for data generation. In model fitting, we tested using the correct structure (`cs`), the independence structure (`ind`) and the AR-1 structure (`ar1`) with `FLORAL`, while the model fitting with other PGEE methods were conducted with the correct correlation structure (`cs`). Details of data generation and performance assessment can be found in Section 2.4. **Fig.3** summarizes the median F_1 scores obtained by the PGEEs and mixed-effect models across 100 simulations for each scenario. Overall, the task of variable selection is more challenging for longitudinal binary outcomes as compared to longitudinal continuous outcomes, such that the simulations with continuous outcomes attained better performance than those with binary outcomes even with smaller sample sizes or effect sizes (Figs.3A-C). The above observation justifies our simulation strategies with different reference scenario for continuous and binary outcomes.

Comparing across the methods, most PGEE methods conducted better variable selection than the mixed-effect models under small numbers of individuals (n), repeated observations (m), and effect sizes (u), while the mixed-effect models achieved comparable performances as $n \geq 100$ (**Fig.3**,**Figs.S1-S3**, panel A). Specifically, the PGEE methods showed higher sensitivity in selecting the true features than the mixed-effect models under smaller sample sizes and effect sizes (**Figs.S1-S3**, panels C). Moreover, the FDR and false-positive control of the PGEE methods gradually improved with a higher sample size, while we observed an inflated FDR of `MaAsLin2` and `MaAsLin3` as n and m increased (**Figs.S1-S2**, panels D). Additional simulations also implied that an increasing number of features (p) corresponded to a decline in sensitivity or an inflation in FDR for all methods, resulting in a decreasing F_1 score (**Fig.S4**). In addition, feature-wise correlation level (ρ) and the strength of the linear time effect (κ) appeared not to heavily affect the variable selection performance (**Figs.S5-S6**).

Among the PGEE methods applying the underlying correct compound symmetry working correlation structure, `FLORAL` achieved a consistently better balance between sensitivity and FDR control while keep both in reasonably effective levels, resulting in a better overall F_1 score. Due to the log-ratio model used for data generation, `FLORAL` and the standard PGEE model with log-transformed data (`PGEE-log`) achieved comparably high level of sensitivity than the PGEE models using other data transformation schemes. Nevertheless, the `FLORAL` model obtained slightly higher sensitivity and consistently better FDR control than `PGEE-log` in most scenarios (**Figs.S1-S6**, panels C-D), where the improved sensitivity can be attributed to the implementation of zero-sum constraint to

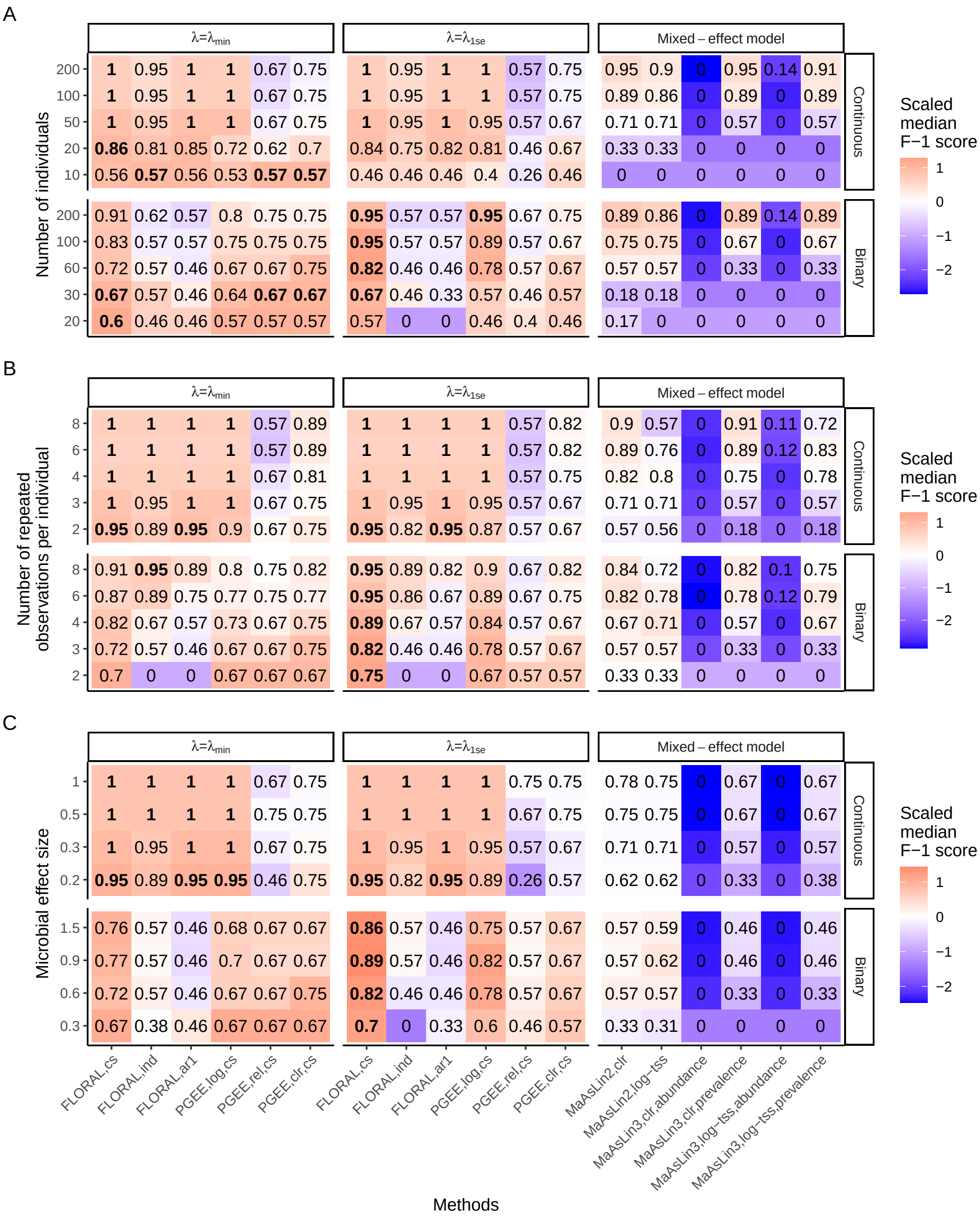

Figure 3: Median F_1 scores out of 100 simulations obtained by PGEE methods with $\lambda = \lambda_{\text{min}}$, $\lambda = \lambda_{\text{1se}}$, and mixed-effect models (`MaAsLin`) under different simulation scenarios with continuous or binary longitudinal outcome variables with different **A.** number of individuals (n), **B.** number of repeated observations per individual (m) and **C.** microbial feature effect sizes (u). As described in the Method section, the reference scenario is $n = 50, p = 200, m = 3, u = 0.15, \rho = 0, \kappa = 2.5$ for continuous outcome and $n = 60, p = 200, m = 3, u = 0.3, \rho = 0, \kappa = 0$ for binary outcome. For each scenario, the color scheme represents scaled median F_1 scores, where red color represents better performance. The highest F_1 score per scenario was shown in bold fonts. For the PGEE methods, we considered compound sysmmetry (cs), independence (ind) and AR-1 (ar1) working correlation structures.

account for compositionality, and the better FDR control is due to the additional two-step feature screen strategy. The performances of `FLORAL` with λ_{min} and λ_{1se} penalty parameters were generally similar, where λ_{min} tended to select more truly associated features at the cost of inflated false positive findings, while λ_{1se} tended to be more conservative with a well-controlled FDR. Overall, `FLORAL` achieved effective FDR control for continuous outcomes when the sample size satisfies $n \geq 20$, irrespective of the choice of λ. However, we did observe a more severely inflated FDR associated with λ_{min} for binary outcomes in most simulation scenarios, whereas `FLORAL` with λ_{1se} still maintained a decent level of FDR with binary outcomes (**Figs.S1-S6**, panel D).

`FLORAL` achieved robust variable selection performance when using different working correlation structures in modeling continuous outcomes, while the performance for binary outcomes depended more heavily on the specification of correct working correlation structure. As shown in **Fig.3** and **Figs.S1-S6** panel A, the performance of variable selection in the continuous outcome models was not strongly affected by the specification of working correlation, where `FLORAL,cs` and `FLORAL,ar1` reached slightly higher F_1 scores than `FLORAL,ind` due to better sensitivity (**Figs.S1-S6**, panel C). In simulations with binary outcomes, on the other hand, we observed a large performance gap between PGEE models with correctly specified working correlation and `FLORAL` with incorrectly specified correlation structures. Although the overall performance of `FLORAL,ind` and `FLORAL,ar1` improved with larger sample sizes, the sensitivity of variable selection was consistently lower than other PGEE methods with correctly specified working correlation structure (**Figs.S1-S6**, panel C). The above observations align with the comparisons of cross-validated deviance residuals across the three working correlation structures, where all three structures achieved comparable cross-validated deviance residuals with continuous outcome, while the compound symmetry structure obtained much smaller deviance residuals than the other two structures (**Fig.S7-S12**, panel B). Such alignment demonstrated the utility of cross-validated deviance residuals for model selection.

The two mixed-effect models, `MaAsLin2` and `MaAsLin3`, required a larger sample size to achieve comparable performances as compared to `FLORAL`. Due to the data generation mechanism based on a log-ratio model, mixed-effect models with CLR-transformation (`clr`) performed better than models with log-transformed TSS (`log-tss`) in most scenarios (**Figs.S1-S6**, panel A). In addition, the high sparsity with 80% zeros of simulated microbial features resulted in more informative variable selection results from `MaAsLin3`'s prevalence model as compared to the abundance model (**Fig.3**). Moreover, `MaAsLin2` and both models from `MaAsLin3` tended to have inflated FDRs under simulations with large n and large m (**Figs.S1-S2**, panels B,D), which was also described by the preprint of `MaAsLin3` as "precision loss with high power" [12, Figs.S4 and S7].

In terms of computational time, `FLORAL` generally took a longer time than `MaAsLin2` but a shorter time than `MaAsLin3` (**Figs.S7-S12**, panel A). As observed, `FLORAL` requires more computational time under simulations with larger number of features (**Fig.S10A**). Moreover, `FLORAL` typically took less than 50 iterations to converge under $\lambda = \lambda_{\text{min}}$ and $\lambda = \lambda_{\text{1se}}$, where models for binary outcomes may took more iterations than models for continuous outcomes (**Figs.S7-S12**, panel C). In very rare cases, `FLORAL` did not reach convergence after 100 iterations, while the convergence criterion was not distant from the pre-specified threshold of 0.001 (**Figs.S7-S12**, panel D).

3.2 FLORAL identifies meaningful taxonomic markers associated with the fiber intake of cancer patients

We correlated longitudinal fiber intake records and longitudinal microbial genera from two cancer studies. The NUTRIVENTION study [2] is a pilot trial with 20 patients and less frequent sample collections, while the MSK allo-HCT cohort [8] is a larger cohort with more than 100 patients and more frequent sample collections. To benchmark the feature selection performance of various methods under different data characteristics, we identified matched fiber intake and microbiome data points for the NUTRIVENTION study (65 samples from 20 patients with 161 genera) and the MSK allo-HCT cohort (505 samples from 137 patients with 112 genera), where the included genera were detected in more than 10% of all samples. Similar to the simulations, we applied `FLORAL` with three working correlation structures (`cs`, `ind` and `ar1`), the standard PGEE models with `cs` correlation structure with log-transformed, relative abundance, and CLR-transformed data, and mixed-effect models (`MaAsLin2` and `MaAsLin3`). The `FLORAL` and PGEE models were run for 100 times with random fold split to reflect a robust pattern of feature selection by the 5-fold cross-validation, where the more frequently selected taxa indicate stronger signals. We used the threshold of 0.1 of the adjusted p-values for feature selection from the mixed-effect models. Detailed information about the two studies and method configurations can be found in Section 2.4.

FLORAL identifies gut health indicating genera from NUTRIVENTION data

Fig.4 displays the variable selection results of FLORAL for the NUTRIVENTION data, where features were only identified using $\lambda = \lambda_{\min}$ due to the small sample size and limited statistical power. Cross-validated deviance residual implies similar model fitting performances by the three working correlation structures (**Fig.4A**), which is confirmed by the correspondingly similar variable selection results (**Fig.4B-D**). Genera *Coprococcus* and *Longicatena* were selected in more than 70% of cross-validated runs, where *Coprococcus* is a well-studied butyrate producer that secretes beneficial short chain fatty acids [30] and *Longicatena* is associated with gut dysbiosis and inflammatory bowel disease [31]. As expected, the abundance of *Coprococcus* was positively associated with fiber intake while the abundance of *Longicatena* was negatively associated with fiber intake, indicating the fiber-oriented dietary intervention was effective in boosting beneficial bacteria and controlling potential pathogens. Genera *Longibaculum* was also identified as positively associated with fiber intake in around 40% of runs, which has been shown to improve oral glucose tolerance in a mouse study [32].

Out of the standard PGEE models with different data transformation schemes, only the model using log-transformed count data identified several markers with $\lambda = \lambda_{\min}$ (**Fig.S13A**). This observation is consistent with our simulation studies where PGEE models with relative abundance and CLR-transformed data showed poorer sensitivities compared to FLORAL and PGEE with log-transformed count data (**Fig.S1-S6**, panel C). Similar to FLORAL, PGEE with log-transformed data also selected *Coprococcus* and *Longicatena*. However, the PGEE model did not adjust for varying sequencing depths across samples for the log-transformed taxa counts due to the lack of zero-sum constraint, which further caused the under-selection of *Coprococcus* and over-selection of *Longicatena*. In terms of the mixed-effect models, only MaAsLin2 with CLR-transformed data selected four taxa associated with fiber intake at the FDR threshold of 0.1 (**Fig.S13B**), where fiber intake was positively associated with *Anaerofilum* (q=0.07) and *Coprococcus* (q=0.08) and negatively associated with *Longicatena* (q=0.09) and *Dehalobacter* (q=0.10). While the clinical interpretation for Coprococcus and Longicatena is expected, the clinical interpretation for Anaerofilum is unclear given there is evidence for its association with

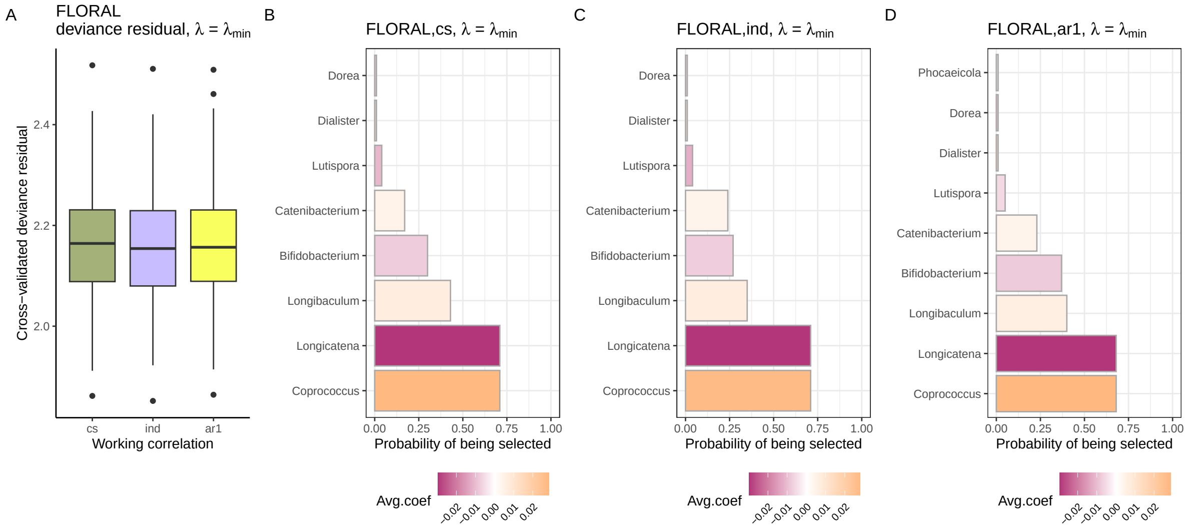

Figure 4: Model fitting and variable selection results for the NUTRIVENTION longitudinal fiber intake and microbiome data using FLORAL with $\lambda = \lambda_{\text{min}}$. **A.** Cross-validated deviance residual obtained by FLORAL using compound symmetry (cs), independence (ind) and AR-1 (ar1) correlation structures out of 100 runs of 5-fold cross-validation with random fold splits. **B-D.** Proportions of taxa being selected by FLORAL out of 100 runs of 5-fold cross-validation with random fold splits using **B.** compound symmetry, **C.** indpendence and **D.** AR-1 working correlation structures. Colors represent the average feature coefficient out of 100 runs, where a positive coefficient implies a positive association between fiber intake and the microbial feature. Results with $\lambda = \lambda_{\text{1se}}$ were omitted as no features were selected.

obesity [33].

FLORAL identifies clinically relevant genera from MSK allo-HCT data

Fig.5 shows model fitting and variable selection results by FLORAL for the MSK allo-HCT cohort. Compared to the NUTRIVENTION data, the allo-HCT cohort consists of substantially more samples and patients, where features were selected using both $\lambda = \lambda_{\text{min}}$ and $\lambda = \lambda_{\text{1se}}$ configurations. Similar to the simulations for continuous outcomes and the NUTRIVENTION analysis, different working correlation structures again reached a similar level of cross-validated deviance residual (**Fig.5A,E**), showing similar model fitting and variable selection performances (**Fig.5B-D,F-H**). Combining the results from $\lambda = \lambda_{\text{min}}$ and $\lambda = \lambda_{\text{1se}}$ across different working correlation structures, an increasing fiber intake were most strongly associated with an increasing abundance of *Blautia* and decreasing abundances of *Enterococcus*, *Alistipes* and *Veillonella*. Out of the above four markers with strongest associations, *Blautia* and *Enterococcus* have been extensively studied in allo-HCT literature as taxa associated with good and poor clinical outcomes, respectively [34–36]. Moreover, *Veillonella* is an oral bacteria and an indicator of gut microbiota depletion in allo-HCT patients [37], while *Alistipes* has been shown to have both protective and harmful effects on gut health [38], also offering reasonable interpretations of their associations with fiber intake.

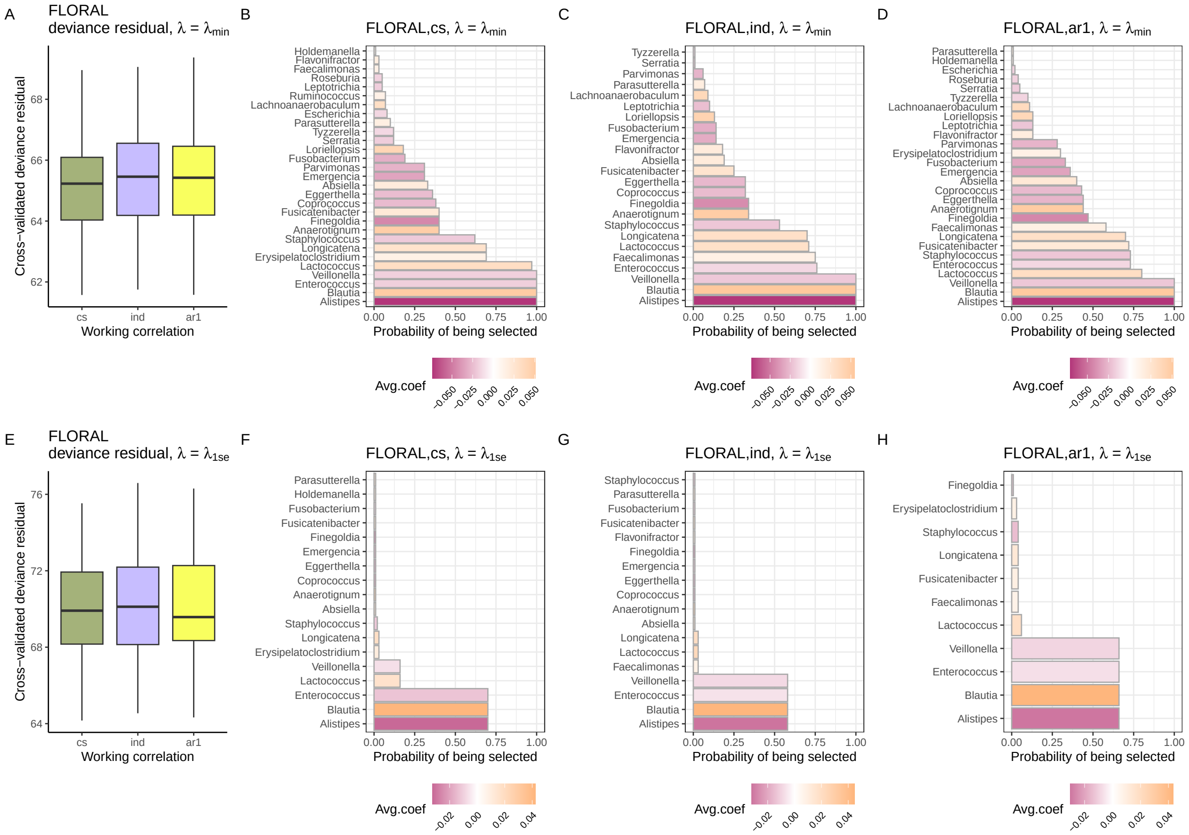

Figure 5: Model fitting and variable selection results for the allo-HCT longitudinal fiber intake and microbiome data using FLORAL with $\lambda = \lambda_{\text{min}}$ (**A-D**) and $\lambda = \lambda_{\text{1se}}$ (**E-H**). **A,E.** Cross-validated deviance residual obtained by `FLORAL` using compound symmetry (cs), independence (ind) and AR-1 (ar1) correlation structures out of 100 runs of 5-fold cross-validation with random fold splits. **B-D,F-H.** Proportions of taxa being selected by `FLORAL` out of 100 runs of 5-fold cross-validation with random fold splits using **B,F.** compound symmetry, **C,G.** indpendence and **D,H.** AR-1 working correlation structures. Colors represent the average feature coefficient out of 100 runs, where a positive coefficient implies a positive association between fiber intake and the microbial feature.

Similar to the NUTRIVENTION analysis results, the PGEE model with log-transformed count data identified similar features as selected by `FLORAL` with different ranks of selection frequencies (**Figs.S14A,D**). Additionally, PGEEs with relative abundance and CLR-transformation also identified *Enterococcus* or *Blautia* in most runs with $\lambda = \lambda_{\text{min}}$ (**Figs.S14B-C**), while still showing low sensitivities at $\lambda = \lambda_{\text{1se}}$ (**Figs.S14E-F**). In terms of the mixed-effect models, both `MaAsLin2` and `MaAsLin3` identified multiple genera significantly associated with fiber intake at the FDR level of 0.1 (**Figs.S15-S16**), where feature selection was mainly driven by the models instead of the data transformation and normalization strategies. Interestingly, *Enterococcus* was not identified by either of the mixed-effect models, while *Blautia* was only identified by `MaAsLin2`, with the 3rd strongest association (q=0.015) per the `MaAsLin2,clr` model (**Fig.S16D**) and only the 14th strongest association (q=0.09) in the `MaAsLin2,log-tss` model (**Fig.S16A**). While obtaining numerous significant associations, most of them had no established relevance with allo-HCT patient outcomes, making it challenging for generating new hypotheses.

Comparing across the tested methods in real-data analyses, `FLORAL` achieved desirable feature-selection performance not only by successfully identifying easily interpretable taxonomic markers, but also by effectively ranking the most plausible and relevant associations as the strongest signals. This can be an advantage in exploratory analysis with an aim of hypothesis generation, where the mixed-effect methods rely on subjective p-value thresholds and may not identify the most biologically meaningful markers as having the most significant associations.

4 Discussion

In this work we introduced the log-ratio penalized generalized estimating equation (PGEE) method to our recently described `FLORAL` package as a new approach to analyzing longitudinal associations between microbial features and patient characteristics. We account for the compositionality of microbial features by an added zero-sum constraint [20] to the standard penalized estimating equation framework [15, 16] and impose a two-step variable selection procedure to better control the FDR [18]. Our simulations demonstrated superior sensitivity with reasonable FDR control of the proposed method over standard PGEE methods and mixed-effect methods under our model assumptions. Real-data analyses further validated the utility of our method in reliably identifying clinically relevant and reported gut health indicating taxonomic markers. Unlike the mixed-effect models where the selected features may contain numerous noises, FLORAL's log-ratio PGEE more robustly ordered the highly relevant taxa as the strongest signals, showing high potentials in exploratory analysis for longitudinal microbiome studies of various scales.

The proposed log-ratio PGEE method is naturally implemented in the publicly available R package FLORAL, which was first introduced for penalized log-ratio generalized linear models and Cox proportional hazards regressions [9]. We provide a user-friendly interface with a standard format of visualizations of variable selection results. Our simulations also validated the stability of the proposed minorize-maximize algorithm with a zero-sum constraint by showing high convergence rates and fast computational speed. Moreover, we propose a built-in model selection criterion for different working correlation structures based on cross-validated deviance residual, which performs robustly in simulations.

Unlike the popularly applied mixed-effect models, we treat the longitudinal patient characteristic of interest as the outcome variable in a GEE model, which translated the research question into fitting a single model rather than hundreds of taxon-specific models. We believe this strategy is technically simpler and performance-wise more robust in real-data applications than the taxon-specific modeling approach, as the microbial trajectories are usually highly heterogeneous and are challenging to be explained by a unified set of model configurations. Under our setting, an analyst has the bandwidth to focus more carefully on modeling the single trajectory of the "outcome" variable of interest, such as dietary intake, which usually follows a more regular distribution than the sparse and skewed microbial abundance, and is easier for fine tuning the non-linear associations with respect to the time effect. Additionally, modeling individual taxon usually involves multiple factors such as dietary intake, antibiotics, and other medications, where the colinearity between factors can easily mask the signals. Under our setting, on the other hand, we are positioned to concentrate more on the associations between the outcome variable of interest and microbial features, where the confounding factors to the trajectory of the outcome variable can still be adjusted. Furthermore, the proposed PGEE method can also be applied to model the trajectory of the outcome variable conditioned on a single baseline microbiome sample, which is another widely available data structure, especially in mouse studies.

Like other penalized regression methods, the proposed log-ratio PGEE model has the following limitations. First, the proposed method tends to have an inflated FDR with $\lambda = \lambda_{\text{min}}$ when the number of individuals is small (**Fig.S1D**) and the number of feature is large (**Fig.S4D**) especially for binary outcomes. If strict FDR control is desired, we recommend using $\lambda = \lambda_{\text{1se}}$ for better FDR control at the cost of getting lower sensitivity. More systematic FDR control procedures, like knockoff [39], can be considered as a direction for future development. Second, the log-ratio regression framework cannot be easily extended to account for non-linear associations between microbial features and the outcome variable, which could be better captured by quantile-based methods [13] or dimension reduction methods for microbial trajectories [10]. Third, our implementation of the log-ratio PGEE model does not incorporate statistical inference of the regression coefficients as discussed in the original PGEE paper [16], where the construction of variance estimators and inference procedures can be further studied for log-ratio-based regression models.

Data and Code Availability

Open-source R package `FLORAL` can be accessed via GitHub (`https://vdblab.github.io/FLORAL`) or CRAN (`https://cran.r-project.org/package=FLORAL`). 16S rRNA sequencing data sets will be made available on FigShare by the time of publication.

Author Contributions

Teng.F. conceived of the project, developed the methodology and wrote the manuscript. Teng.F. and V.D. performed computational analysis. Teng.F., V.D., Tyler.F., M.B., N.R.W., A.D., S.S.R, U.A.S. and J.U.P analyzed and interpreted the analysis results. Tyler.F., M.B., N.R.W., J.P., A.D., F.C., J.H., A.G, S.S.R., U.A.S. assisted with microbiome and clinical data harmonization. J.P., F.C., J.H., A.G., S.S.R, A.M.L., U.A.S., M.R.M.v.d.B, and J.U.P. coordinated clinical data and sample collection and sequencing management. Teng.F., U.A.S., M.R.M.v.d.B. and J.U.P. co-supervised the study.

Authors' Disclosures

A.M.Lesokhin reports a grant from Novartis, during the conduct of the study; grants from BMS; personal fees from Trillium Therapeutics; grants, personal fees, and nonfinancial support from Pfizer; grants and personal fees from Janssen, outside the submitted work; and has a patent US20150037346A1, with royalties paid. U.A.Shah reports MSK Paul Calabresi Career Development Award for Clinical Oncology K12CA184746, Paula and Rodger Riney Foundation, Allen Foundation Inc, Parker Institute for Cancer Immunotherapy at MSK, International Myeloma Society, HealthTree Foundation and Willow Foundation as well as nonfinancial support from American Society of Hematology Clinical Research Training Institute, Transdisciplinary Research in Energetics and Cancer training workshop R25CA203650; research funding support from Celgene/BMS and Janssen to the institution, nonfinancial research support from Sabinsa pharmaceuticals, and M&M Labs to the institution; personal fees from Janssen Biotech, Sanofi, BMS, and i3Health outside the submitted work. M.R.M. van den Brink has received research support and stock options from Seres Therapeutics and stock options from Notch Therapeutics and Pluto Therapeutics; he has received royalties from Wolters Kluwer; has consulted, received honorarium from or participated in advisory boards for Seres Therapeutics, Vor Biopharma, Rheos Medicines, Frazier Healthcare Partners, Nektar Therapeutics, Notch Therapeutics, Ceramedix, Lygenesis, Pluto Therapeutics, GlaskoSmithKline, Da Volterra, Thymofox, Garuda, Novartis (Spouse), Synthekine (Spouse), Beigene (Spouse), Kite (Spouse); he has IP Licensing with Seres Therapeutics and Juno Therapeutics; and holds a fiduciary role on the Foundation Board of DKMS (a nonprofit organization). J.U. Peled reports funding from NHLBI NIH Award K08HL143189 and the V Foundation; he reports research funding, intellectual property fees, and travel reimbursement from Seres Therapeutics, and consulting fees from DaVolterra, CSL Behring, Crestone Inc, and from MaaT Pharma. He serves on an Advisory board of and holds equity in Postbiotics Plus Research; He has filed intellectual property applications related to the microbiome (reference numbers #62/843,849, #62/977,908, and #15/756,845).

Acknowledgments

This research was supported by National Cancer Institute award numbers P30-CA008748 MSK Cancer Center Support Grant/Core Grant, R25-CA272282, and R01-CA228308. Additional support was received from National Cancer Institute award numbers R01-CA228358, K12-CA184746 and R25-CA203650, National Heart, Lung, and Blood Institute (NHLBI) award numbers R01-HL123340, R01-HL147584 and K08-HL143189, National Institute on Aging award number P01-AG052359, The Allen Foundation, Inc, American Society of Hematology Scholar Award, Cycle for Survival, The Lymphoma Foundation, Parker Institute for Cancer Immunotherapy, Paula and Rodger Riney Multiple Myeloma Research Initiative, Seres Therapeutics, The Solomon Microbiome Nutrition and Cancer Program, Starr Cancer Consortium, The Susan and Peter Solomon Family Fund, Tri-Institutional Stem Cell Initiative, and V Foundation. The NUTRIVENTION trial was supported by Plantable that provided subsidized meals and participant nutrition coaching. We thank the MSK Society for the support of the Genomics Experience for Master's Students (GEMS) program.

References

1. Dai, A. *et al.* Sugar-rich foods exacerbate antibiotic-induced microbiome injury. *bioRxiv,* 2024–10 (2024).

2. Shah, U. A. *et al.* A High-Fiber Dietary Intervention (NUTRIVENTION) in Precursor Plasma Cell Disorders Improves Biomarkers of Disease and May Delay Progression to Myeloma. *Blood* **144,** 671–671 (2024).

3. Paredes, J. *et al.* Increased Fiber Intake Results in Better Overall Survival and Lower GI-aGVHD in Allo-HCT Recipients and Pre-Clinical Gvhd Models. *Blood* **144,** 259 (2024).

4. Shah, U. *et al.* A High-Fiber Dietary Intervention (NUTRIVENTION) in Precursor Plasma Cell Disorders Improves Disease Biomarkers and Delays Progression to Myeloma. *SSRN Electronic Journal.* https://ssrn.com/abstract=4850456 (2024).

5. Schluter, J. *et al.* The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588,** 303–307 (2020).

6. Miltiadous, O. *et al.* Early intestinal microbial features are associated with CD4 T-cell recovery after allogeneic hematopoietic transplant. *Blood, The Journal of the American Society of Hematology* **139,** 2758–2769 (2022).

7. Lindner, S. *et al.* Altered microbial bile acid metabolism exacerbates T cell-driven inflammation during graft-versus-host disease. *Nature Microbiology* **9,** 614–630 (2024).

8. Dai, Y. *et al.* Longitudinal Microbiome-based Interpretable Machine Learning for Identification of Time-Varying Biomarkers in Early Prediction of Disease Outcomes. *bioRxiv,* 2024–10 (2024).

9. Fei, T. *et al.* Scalable log-ratio lasso regression for enhanced microbial feature selection with FLORAL. *Cell Reports Methods* **4(11),** 100899 (2024).

10. Shi, P. *et al.* TEMPTED: time-informed dimensionality reduction for longitudinal microbiome studies. *Genome Biology* **25,** 317 (2024).

11. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology* **17,** e1009442 (2021).

12. Nickols, W. A. *et al.* MaAsLin 3: Refining and extending generalized multivariable linear models for meta-omic association discovery. *bioRxiv,* 2024–12 (2024).

13. Li, S., Li, R., Lee, J. R., Zhao, N. & Ling, W. ZINQ-L: A Zero-Inflated Quantile Approach for Differential Abundance Analysis of Longitudinal Microbiome Data. *Frontiers in Genetics* **15,** 1494401 (2024).

14. Sun, H. *et al.* Detecting sparse microbial association signals adaptively from longitudinal microbiome data based on generalized estimating equations. *Briefings in Bioinformatics* **23,** bbac149 (2022).

15. Johnson, B. A., Lin, D. & Zeng, D. Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103,** 672–680 (2008).

16. Wang, L., Zhou, J. & Qu, A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68,** 353–360 (2012).

17. Agresti, A. *Categorical data analysis* (John Wiley & Sons, 2012).

18. Bates, S. & Tibshirani, R. Log-ratio lasso: scalable, sparse estimation for log-ratio models. *Biometrics* **75,** 613–624 (2019).

19. Harrell, F. E. *et al. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (Springer, 2005).

20. Aitchison, J. *The Statistical Analysis of Compositional Data* ISBN: 0-412-28060-4 (Chapman and Hall, London, 1986).

21. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22 (1986).

22. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations* (CRC press, 2015).

23. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96,** 1348–1360 (2001).

24. Boyd, S. & Vandenberghe, L. *Convex optimization* (Cambridge university press, 2004).

25. Inc, S. SAS/STAT 14.3 user's guide. *Cary, NC: SAS Institute Inc* (2017).

26. Inan, G. & Wang, L. PGEE: an R package for analysis of longitudinal data with high-dimensional covariates. *R J.* **9,** 393 (2017).

27. Eddelbuettel, D. & Sanderson, C. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis* **71,** 1054–1063 (2014).

28. Qaqish, B. F. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90,** 455–463 (2003).

29. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300 (1995).

30. Notting, F., Pirovano, W., Sybesma, W. & Kort, R. The butyrate-producing and spore-forming bacterial genus Coprococcus as a potential biomarker for neurological disorders. *Gut Microbiome* **4,** e16 (2023).

31. Dahal, R. H., Kim, S., Kim, Y. K., Kim, E. S. & Kim, J. Insight into gut dysbiosis of patients with inflammatory bowel disease and ischemic colitis. *Frontiers in Microbiology* **14,** 1174832 (2023).

32. Hankir, M. K. *et al.* Gut Microbiota Contribution to Weight-Independent Glycemic Improvements after Gastric Bypass Surgery. *Microbiology Spectrum* **11,** e05109–22 (2023).

33. Niu, J. *et al.* Cottonseed meal fermented by Candida tropical reduces the fat deposition in white-feather broilers through cecum bacteria-host metabolic cross-talk. *Applied microbiology and biotechnology* **104,** 4345–4357 (2020).

34. Nguyen, C. L. *et al.* High-resolution analyses of associations between medications, microbiome, and mortality in cancer patients. *Cell* **186,** 2705–2718 (2023).

35. Peled, J. U. *et al.* Microbiota as predictor of mortality in allogeneic hematopoietic-cell transplantation. *New England Journal of Medicine* **382,** 822–834 (2020).

36. Taur, Y. *et al.* Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clinical infectious diseases* **55,** 905–914 (2012).

37. Liao, C. *et al.* Oral bacteria relative abundance in faeces increases due to gut microbiota depletion and is linked with patient outcomes. *Nature microbiology,* 1–11 (2024).

38. Parker, B. J., Wearsch, P. A., Veloo, A. C. & Rodriguez-Palacios, A. The genus Alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health. *Frontiers in immunology* **11,** 906 (2020).

39. Barber, R. F. & Candès, E. J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43,** 2055–2085 (2015).