The idea if personalized diet is to make decisions on what food type one should take in order to improve his/her health conditions. There is a rich statistical literature on personalized treatment models, including (53–55). The key statistical question behind personalized medicine is the heterogeneous responses across different individuals, which makes prediction more challenging. For example, in a linear model, the effect of food is the same (represented by the same beta coefficients) for all individuals, however, this may not be true [can we give a citation here in nutrition literature?]. Hence, we are interested in identifying such heterogeneity using a flexible and robust machine learning model. This makes random forest a perfect candidate. In addition, the difficult of our study lies in the potential of multiple outcomes (such as metabolites or clinical measurements) and the compositional property of the microbiome features. Furthermore, our data consists of measurements at both before and end of a period, which can be utilized differently than existing studies. To address these issues, we propose a new random forest based model (42,44,56–58) with linear combination splits that explores the canonical correlation between multiple covariates and multiple outcomes, and further construct splitting rules that differentiate these canonical covariates across different treatments (food labels) This allows us to model and identify potential markers that indicates if the subject may respond to the particular food. It should be noted that different from previous approaches, this new model would allow us to make such predictions before the subject actually take the given food. Hence it is possible to use this model as a screening tool for selecting subjects in future studies.

[add some further justification why this is interesting.]

For simplicity, we consider the case when there are two food types, denoted as . This could represent a food versus the control. In this case, the output of this model may suggest if the subject is going to respond to the food differently than control. The method can be easily generalized to multiple food types with slight modification. We modify previous notations slightly to facilitate the personalized model. At the beginning of each period, a participant’s microbiomes (and/or genes) are measured, denoted as , the measurements at the end is denoted as , and the difference between the two is . Here, could represent the covariate from either treatment. An addition -dimensional outcome vectors and are also observed before and after the period, respectively, with the difference .

The mechanism of the underlying data generating mechanism are visually presented in Figure 3. We assume a very general framework that there are two unobserved functions, the link function for the relationship between microbiome and outcomes, and the treatment effect function that alters the microbiome values due to the food. More precisely, the function is defined as

While the link function between the microbiome and the response is expressed as

The goal of our study is to model the difference of the multivariate outcome between the two treatments. This is of great interest because we may be able to make the treatment decision based on the predicted difference. Based on our model assumption, the difference between the multivariate outcome of interest caused by the treatment is

Which is the expected outcome difference starting from the same initial microbiome value . However, this quantity can be difficult to estimate.

Since the quantity we are interested in is a local estimation, we use random forest model to construct the neighborhood and perform a local averaging. However, the challenging part is to construct the splitting rule. The splitting rule is the core of a random forest model as it guides the model fitting process. It quantifies the reduction of impurity or variance [cite] associated with a split and find the best split that maximize the change of impurity. This means that we want to construct splitting rules on the initial covariate such that the terminal nodes contain mostly homogeneous subjects, with similar treatment effects. Consider that at a local region, the treatment effect function can be approximated by

Where is a matrix of coefficients for either . However, at an internal node we may not be interested in modeling all of them, instead, we could explore the most significant/drastic direction by utilizing the canonical correlation analysis (CCA (59)) for each treatment value separately:

and

Similarly, if we express as a linear combination of the covariate, i.e.

where is a matrix of coefficients. Then the difference of outcome and the difference of microbiome has the relationship

We can again use the CCA analysis for each treatment, but tie the direction to the directions specified previously for . This leads to solving

and

Furthermore, since is highly correlated with both and , we can jointly solve, for

And similarly for

Note that the entire derivation above is to locally approximate the relationship between different types of measurements of the two different treatments. However, we still need to develop the splitting rule that differentiate the treatment effect.

First, the split should be performed on the covariate

Note that this is a univariate variable after the linear combination. Then to perform a split on this variable, we need to calculate the impurity reduction on the (difference) of outcome of interest. To do this, we first calculate the treatment difference, which is simply

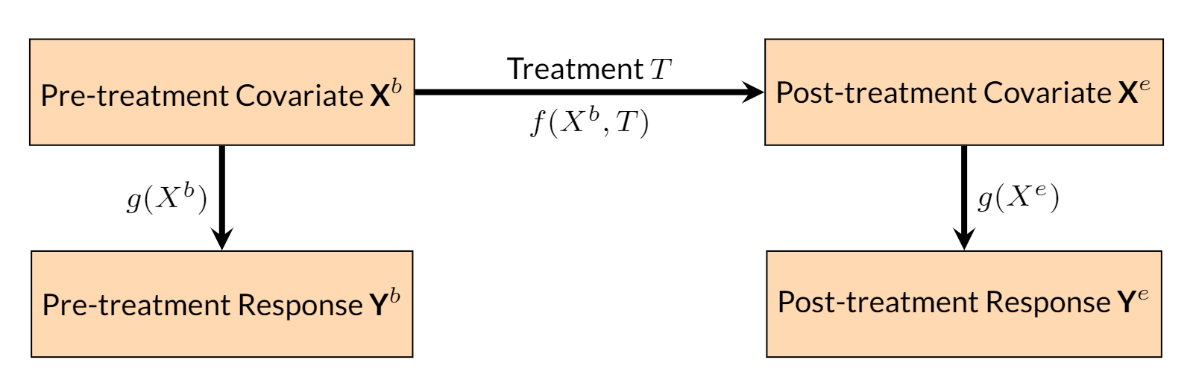
I think there is a problem here. For the people who received the treatment T=1, how to interpret , and similarly . Theoretically, this is what we want right?

We want to perform a split such that after the splitting, the treatment difference is difference between the left child node and the right child node. In a random forests regression model, this is essentially evaluating the variance of this quantity on the left and right node, and calculate the overall pooled variance

Current Implementation:

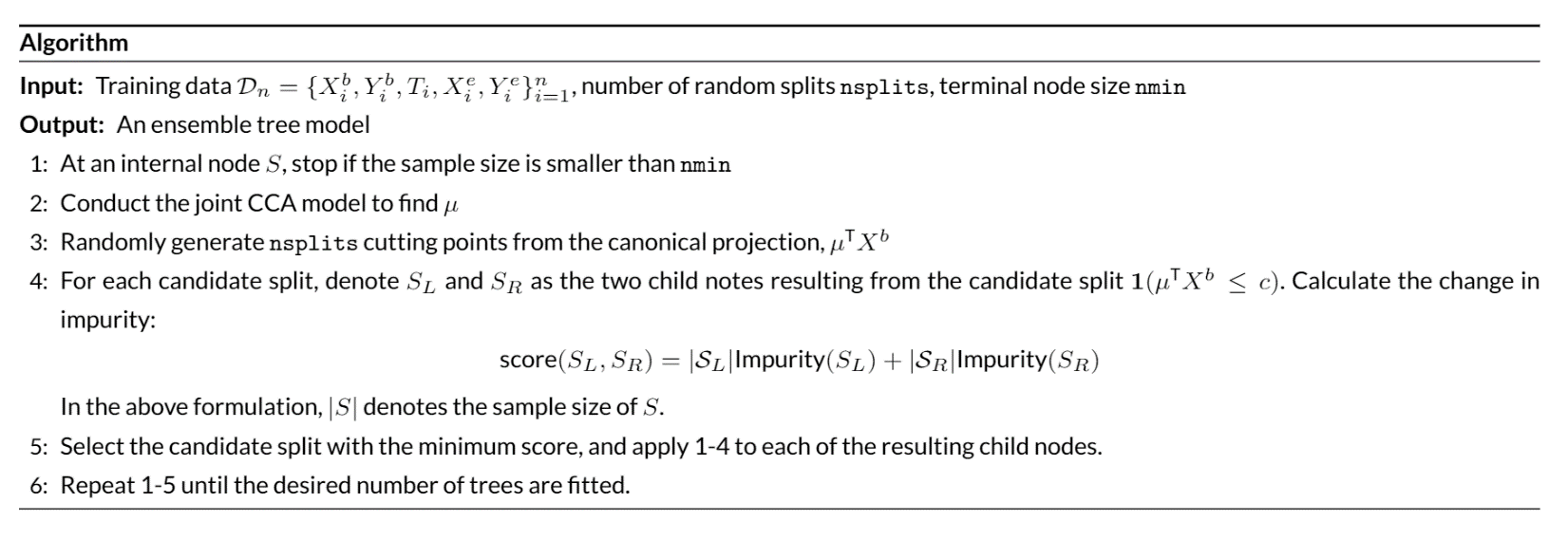
Will Try:

Where L and R denotes the left and right nodes, respectively, and denotes their sample size. And the split on that yields the largest splitting score is used. An outline of the algorithm is given in the following Table 1. Note that a random forest model usually involves several tuning parameters, such as the number of splits, terminal nodes size (60). We also incorporate them in our proposed model.

****

**Figure 3.** [Model assumption of the personalized treatment random forest model with multiple outcomes in two stages. ]

**Table 1.** [ Pseudo algorithm for the proposed random forest model]



Modeling Heterogeneity in Treatment Difference for Precision Medicine Using Multivariate Random Forest