University of Cincinnati

Date: 11/3/2022

I, Timothy M Stone, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Biostatistics (Environmental Health).

It is entitled:

Threshold Parameter Optimization in Weighted Quantile Sum Regression

Student's name: **Timothy M Stone**

This work and its defense approved by:

Committee chair: Roman Jandarov, Ph.D.

Committee member: Ashley Merianos, Ph.D.

Committee member: Marepalli Rao, Ph.D.

Committee member: Tiina Reponen, Ph.D.

44204

University of Cincinnati College of Medicine

Threshold Parameter Optimization in Weighted Quantile Sum Regression

A Dissertation in
Biostatistics
by
Timothy M. Stone

Submitted in Partial Fulfillment
of the Requirements
for the degree of

Doctor of Philosophy

3 November 2022

Abstract

Connecting and predicting health outcomes based on environmental exposures is critical for disease prevention. The ensemble approach, weighted quantile sum regression (WQS) is often used for that purpose. In this work, a novel application of WQS to analyzing microbiome data, an original method for determining the WQS selection threshold parameter τ, and the validation of this method with real world chemical exposure biomarker data are used. WQS provided useful estimates of overall effect of exposures to microbial mixtures on the observed presence of respiratory health conditions in children. Additionally, the proposed novel method for determining threshold selection parameter τ improved characterization of important predictors within the model. Validation of the τ selection procedure expanded the generalizability of ROC optimization for identifying groups of important predictors for relative comparisons.

Acknowledgements

I want to express my sincerest gratitude to my mentor, Dr. Roman Jandarov, for his guidance and support of my research, as well as his example of a statistician. His encouragement and advice were invaluable throughout my time at University of Cincinnati. Besides my advisor, I would like to thank the rest of my dissertation committee. Dr. Rao, you are a pillar of the biostatistics division. Dr. Reponen, my experience working with you was a large part of the motivation for my dissertation. Dr. Merianos, thank you for your insightful comments and support of my research.

Kotya, Scout, and Baloo, thank you for your diligence in attending every single conference call despite never being invited. To my parents, I am grateful for your love and support of my pursuits. Dad, thank you for providing the foundation for my academic curiosity, despite my left-handedness. Mom, thank you for always being in my corner, and for all the sacrifices you made that have allowed to get where I am today. I also want to thank my very first teacher, my grandmother. I think you would agree that I have progressed a bit beyond the days of counting numbers on the kitchen fridge.

Most importantly, I would like to thank my wife Kate. Thank you for listening to me ramble about work at home whenever I would get excited about an idea. Thank you for looking over my papers and enduring my presentations. Thank you for being so patient through the difficult times and all the long, stressful nights. Thank you for supporting me and believing in me every step of the way. This endeavor would not have been possible without your love.

Contents

1 An Introduction to Microbiome Analysis **6**
1.1 Characterizing Microbiome Data 7
1.2 Common Statistical Approaches and Applications 8
1.2.1 In-sample Diversity Analysis Using Classical Statistical Tools . . 9
1.2.2 Community Level Analysis of Microbiome 11
1.2.3 Regression Approaches in Differential Abundance Analysis . . . 18

2 Weighted Quantile Sum Regression **27**
2.1 A General Framework for Weighted Quantile Sum Regression 28
2.2 WQS Regression: Variations 30
2.2.1 Grouped WQS 30
2.2.2 Bayesian WQS 31
2.2.3 Lagged WQS Regression 33
2.3 Identifying Meaningful Predictors via Selection Threshold Parameter Tau 34
2.4 Simulation Study 35
2.4.1 Methods 35
2.4.2 Results 39
2.4.3 Discussion 44

3 ROC Optimization for Tau: Using Chemical Biomarker Profiles **49**
3.1 Tobacco Smoke Exposure and Classification 49
3.2 Random Forests 50
3.3 Description of the Data 52
3.4 Model Comparison and Implementation of Tau Optimization 55
3.4.1 Random Forest Model Results 56
3.4.2 Weighted Quantile Sum Results 57
3.4.3 Discussion 62

References **65**

List of Tables

1.1 Example of microbial species distribution (counts) 7
1.2 P-values for non-parametric alpha diversity comparisons 11
1.3 Community composition comparisons using PERMANOVA 15
1.4 Log 2-fold changes in differentially abundant bacterial and fungal species 20
1.5 Associations between unadjusted WQS mixtures and the presence or absence of health outcomes of asthma, wheeze, aeroallergen positivity, and rhinitis 25
2.1 Comparisons of AUC for varying experimental conditions ROC. Dimension comparisons for 30 to 80 and 30 to 59 utilize data structures 1 and 2, respectively. 40
2.2 Mean percent differences between τ_{ROC} and τ_h for each set of compared experimental conditions 43
2.3 Linear regression for effect of varying simulation conditions on mean percent difference in selection threshold parameter 45
3.1 Univariate logistic regression models for each chemical biomarker or ratio of interest 56
3.2 Random forest variable importance scores 57
3.3 Random forest model confusion matrix 57
3.4 Final WQS regression model 59
3.5 Component weights from final WQS model 60
3.6 WQS model confusion matrix 61

List of Figures

1.1 *Clustering of bacterial taxa using Bray's dissimilarity and Ward's method* 17
1.2 *Correlation matrix consisting of fungal taxa calculated using Spearman's* ρ 23
2.1 *Flowchart to describe the various sets of experimental conditions used for simulations.* 37
2.2 *ROC curves for models fit with differing sets of experimental conditions. Considerations for panel (A): Correlation Structures 1 and 2; (B): 30 dimensions and 80 dimensions; (C): 30 dimensions and 59 dimensions; (D): 5, 10, 15, and 20 signals; (E): Damping factor 0.5 or 1; (F): Signal strength ranges 0.05-0.15, 0.10-0.20, and 0.15-0.25.* 41
2.3 *Probability density curves for models fit with differing sets of experimental conditions. Considerations for panel (A): Correlation Structures 1 and 2; (B): 30 dimensions and 80 dimensions; (C): 30 dimensions and 59 dimensions; (D): 5, 10, 15, and 20 signals; (E): Damping factor 0.5 or 1; (F): Signal strength ranges 0.05-0.15, 0.10-0.20, and 0.15-0.25.* 44
3.1 *Chemical biomarker correlation matrix calculated using Spearman's* ρ . 55
3.2 *ROC curve from WQS simulations used to select threshold parameter* τ*. The black line indicates a random classifier. The magenta segments indicate sensitivity and specificity of the spot along the curve which minimizes the length of the green line.* 59
3.3 *Weight estimates for predictors in final WQS model* 61

1 An Introduction to Microbiome Analysis

Since the widespread acceptance of the germ theory, the deleterious roles of individual microorganisms in human health have been extensively studied. However, only more recently has it been shown that communities of microbial populations can affect human health. These populations, residing inside and on the surface of the human body, were termed the "human microbiome". The organisms that constitute the human microbiome are estimated to outnumber human cells by a factor of ten[1]. Many studies have established important links between the human microbiome and an individual's health, such as its influence on metabolism and its role in regulation of the immune system[2,3]. To this day, researchers continue to discover novel niches of microbiota in human organs and tissues, all with their unique compositions and roles, an imbalance of which can lead to deleterious consequences[4]. It is also known that not only is an individual's health determined by internal factors, but it is also affected by external exposures. In recent studies, dynamic interactions between environmental microbiota communities and humans have been investigated, revealing that they play a role in bronchopulmonary and inflammatory disorders[5,6]. This work focuses on investigating causative effects of the exposures on human disease.

1.1 Characterizing Microbiome Data

Microbiome data is most commonly generated through 16S rRNA sequencing and shotgun metagenomic sequencing. The sequences are then mapped to an existing phylogenetic tree or clustered, and then hierarchically assigned to a tree based on similarity[7]. The data produced after taxa assignments is compiled into tables consisting of read counts corresponding to nodes of a branch on a taxonomic tree[8]. Microbiome count data is characteristically high dimensional, sparse, and over-dispersed[9–11]. Table 1.1 illustrates characteristics commonly found in microbiome data. This table is extracted from data sets published in Cox et al[12].

Table 1.1: Example of microbial species distribution (counts)

Species	Zero Percentage	Median	Mean	Variance
A. lwoffii	16.55%	180	1032.91	7 171 935.46
A. muciniphila	71.03%	0	29.69	7924.17
A. illinoisensis	71.72%	0	32.32	15 664.50
A. phyllosphaerae	50.34%	0	137.99	163 575.72
D. aquatilis	68.28%	0	25.85	5239.49
F. prausnitzii	72.41%	0	52.43	26 075.08
F. periodonticum	43.45%	5	130.31	62 215.09
G. vaginalis	65.52%	0	266.48	1 603 716.53
K. rhamnosa	70.34%	0	75.03	73 685.01
M. vaginatus	35.17%	12	155.57	203 710.89
N. plantarum	75.17%	0	40.03	15 785.06
N. suwonense	13.10%	94	253.07	162 475.93
P. pasteri	28.28%	15	130.34	57 215.38
R. mucilaginosa	23.45%	87	716.80	1 985 095.09
S. yunnanensis	53.10%	0	331.83	1 238 812.21
S. maltophilia	45.52%	5	271.41	779 652.47
T. sanguinis	71.03%	0	23.49	3859.79
V. paradoxus	61.38%	0	120.09	87 864.21

The data contained 170 samples and 80 different species after processing (removing

species which were only present in fewer than 20% of the samples), but for ease of display only a cross section of the data is shown. Many of the species shown above have counts of 0 across more than 50% of samples in the corresponding data set, and approximately 52% of all data points were 0s. The example provided is high dimensional and sparse, which is typical of microbiome data. For each species, the variance is much larger than its mean, indicating over-dispersion in the count data.

In general, the goal of studies that produce this type of data is to explore differences in microbiome composition between experimental groups or to investigate the impact of external factors on microbiome composition[13,14].

1.2 Common Statistical Approaches and Applications

In the previous section, the typical objectives of microbiome studies and common characteristics of microbiome data sets were discussed. These sparse, high dimensional, over-dispersed data present numerous challenges for traditional statistical approaches. The presence of many zeroes limits the ability for parametric models to make accurate estimates of variance for meaningful inference[15]. Non parametric methods lack the power to perform inference on taxa with low counts. This section contains discussion on methods that are commonly used to provide insight on the relationship between the microbiome and human health. In addition, it also includes work that demonstrates application of these methods.

1.2.1 In-sample Diversity Analysis Using Classical Statistical Tools

The properties characteristic of microbiome data present challenges for the classical toolbox of statistical models, such as t-tests, ANOVA, or their corresponding non-parametric equivalents. However, these standard statistical tools can still be applied to analyze the changes in microbiome between groups, just not by directly using taxon counts. The most common approach uses taxa counts from each sample to calculate various diversity measures and then compares those calculated values between groups[16–18]. One such measure is α-diversity.

In the context of the microbiome, α-diversity measures the variation of microbes within a single sample. There are many differing estimates of α-diversity. Some common measures are species richness, Shannon's diversity index, or Simpson's diversity index. Species richness simply is a count of the number of unique species observed in a sample[19]. Shannon diversity index describes how evenly the microbes are distributed within a sample. Denoted as H, the formula for the index is given in equation 1.1, where p is the proportion of counts of a specific taxon and s is the number of species[20].

$$H = -\sum_{i=1}^{s} p_i \; ln(p_i) \tag{1.1}$$

Larger values are given by the presence of many species evenly distributed within a sample. Whereas the Shannon index measures the evenness of taxa within a sample, the Simpson index considers taxa dominance by giving more weight to commonly observed

species. Denoted by D, the calculation of the Simpson index is provided below in equation 1.2, where p is the proportion of counts of a given taxon, and s is the number of species[21].

$$D = \frac{1}{\sum_{i=1}^{s} {p_i}^2} \tag{1.2}$$

Values of D are between 0 and 1, where values approaching 1 indicate higher levels of diversity.

In work analyzing the associations between residential microbiomes and childhood respiratory health, Mann-Whitney and Kruskal-Wallis tests were used to compare different alpha diversity measures between groups of interest[12]. The results of those tests are shown in Table 1.2.

All the reported p values are unadjusted. Additionally, none of the p values were less than 0.05, thereby suggesting that there was not enough evidence to claim that there were meaningful associations between alpha diversity measures and experimental groups based on presence or absence of respiratory outcomes. However, other research groups have published work suggesting links between microbiome diversity and respiratory health[6,22]. The possibility was considered such that these classical statistical tools were not powerful enough to detect potential signals in this data set.

Table 1.2: P-values for non-parametric alpha diversity comparisons

Kingdom	Age	Outcome	Observed Richness	Shannon	Simpson
Bacteria	Year 7	Asthma	0.7107	0.4968	0.3061
		Wheeze	0.8359	1.0000	0.7185
		Aeroallergen+	0.0872	0.3308	0.5140
		Rhinitis	0.5867	0.4360	0.3771
		Wheeze Type	0.8816	0.9805	0.9214
	Year 12	Asthma	0.8532	0.2952	0.1781
		Wheeze	0.9241	0.5904	0.5753
		Aeroallergen+	0.6648	0.9071	0.4218
		Rhinitis	0.3639	0.1666	0.1310
		Wheeze Type	0.4990	0.6195	0.5572
Fungi	Year 7	Asthma	0.7107	0.4968	0.5825
		Wheeze	0.2102	0.3330	0.2576
		Aeroallergen+	0.0872	0.3308	0.8158
		Rhinitis	0.5867	0.4360	0.8762
		Wheeze Type	0.5102	0.9805	0.7252
	Year 12	Asthma	0.8532	0.2952	0.2655
		Wheeze	0.4017	0.4945	0.6358
		Aeroallergen+	0.6648	0.9071	0.7750
		Rhinitis	0.3639	0.1666	0.9311
		Wheeze Type	0.4990	0.6195	0.8811

1.2.2 Community Level Analysis of Microbiome

Another major objective of microbiome studies is to determine whether communities can be classified based on their composition[23]. β-diversity is a measurement used to assess the heterogeneity of taxa composition along experimental gradients[24]. It is important to emphasize the differences between α-diversity and β-diversity. α-diversity is akin to a point estimate for each sample, and is used for comparisons between samples. β-diversity is used for comparisons between groups. For microbiome count data, β-diversity measures are usually expressed as distance coefficients. One popular distance estimate is the Bray-Curtis dissimilarity, which is a statistic used to quantify compositional

differences between two samples, based on taxa counts[25]. Given as BC, the formula for Bray-Curtis dissimilarity is shown below in 1.3, where X_{ij}, X_{ik} are the number of individuals in species i in each sample (j, k) and n is the total number of species in samples.

$$BC = \frac{\sum_{i=1}^{n} |X_{ij} - X_{ik}|}{\sum_{i=1}^{n} (X_{ij} + X_{ik})} \tag{1.3}$$

The Bray-Curtis measure ignores cases in which a taxon is absent in both community samples. However, it is dominated by the prevalent taxa, so that rare taxa contribute minimally to the value of the coefficient. Identical communities have a value of "0", whereas entirely different communities have a value of "1"[26]. While simply calculating the Bray-Curtis dissimilarity coefficients between groups provides some concept of the levels of heterogeneity between groups, many researchers are interested in testing for meaningful associations between overall taxa composition and experimental groups. Typically, a multivariate analysis of variance (MANOVA) test would be used to make this comparison. However, since microbiome data is sparse and highly-skewed, it does not meet the assumptions of MANOVA, namely multivariate normality. To overcome this limitation, permutational analysis of variance is used instead (PERMANOVA). PERMANOVA follows the same framework as MANOVA, where multivariate comparisons are performed between experimental groups. However, rather than comparing multivariate means as in MANOVA, the comparison in a PERMANOVA test compares the centroids of the experimental groups. PERMANOVA is formulated using any dis-

tance measure, which permits the use of the Bray-Curtis dissimilarity matrix to perform statistical inference using this method.

To perform a PERMANOVA test, consider matrix $\mathbf{Y}$ with N rows representing samples and p columns representing variables. Define $\mathbf{D} = d_{ij}, i = 1, ..., N; j = 1, ..., N$ as the dissimilarity or distance between every pair of observations (i, j). The total sum of squares of dissimilarities is given below.

$$SS_T = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{d_{ij}^2}{N} \tag{1.4}$$

Total sum of squares adds up the squares of the distances in the upper or lower triangle of the distance matrix (not including the diagonal) and divides by N number of observations. SS_T is used as the average distance among all samples. The within-group or residual sum of squares is calculated below.

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij} \tag{1.5}$$

ϵ_{ij} is an indicator function that takes the value 1 if observation i and observation j belong to the same group. Otherwise, its value is 0. SS_W is the squares of all the distances between observations that occur in the same group divided by n, the number of observations in each respective group. The sum of squares within groups is the average distance between samples within groups. The average distance among groups, given as SS_A is obtained by subtracting SS_W from SS_T.

$$SS_A = SS_T - SS_W \tag{1.6}$$

A pseudo F-statistic, F^* is then calculated using the previously formulated SS_A, SS_W, p, and N, where p is the number of groups being compared, and N is number of samples.

$$F^* = \frac{\frac{SS_A}{p-1}}{\frac{SS_W}{N-p}} \tag{1.7}$$

Distribution free significance of the pseudo F-statistic is determined by performing permutations of the data, and calculating separate F statistics for each permutation of the data. The resulting p value is determined as a ratio of the number of permutated F statistics greater than F^* plus 1 and the total number of permutations plus 1.

$$P = \frac{(count\ F^p \geq F^*) + 1}{(total\ count\ F^p) + 1} \tag{1.8}$$

The rationale for PERMANOVA is that, under the null hypothesis, there should be no difference between the groups. Consequently, the multivariate observations would be exchangeable between the different groups. This random shuffling can be performed for all possible re-orderings of the sample labels, and then compared to see whether the randomized re-ordering results in greater separation among groups. If less than 5 of the permuted F-statistics are greater than or equal than F^*, the p value is less than 0.05, and the centroids between groups are significantly different[27].

PERMANOVA was used to analyze differences in community composition between samples with or without respiratory health conditions for both fungi and bacteria[12]. The results of that analysis are shown in Table 1.3.

Table 1.3: Community composition comparisons using PERMANOVA

Kingdom	**Age**	**Outcome**	**P value**
Bacteria	Year 7	Asthma	0.264
		Wheeze	0.532
		Aeroallergen+	0.184
		Rhinitis	0.292
	Year 12	Asthma	0.801
		Wheeze	0.394
		Aeroallergen+	0.512
		Rhinitis	0.314
Fungi	Year 7	Asthma	0.443
		Wheeze	0.528
		Aeroallergen+	0.548
		Rhinitis	0.822
	Year 12	Asthma	0.437
		Wheeze	0.939
		Aeroallergen+	0.336
		Rhinitis	0.815

The p values reported in Table 1.3 are all unadjusted. None of the values were less than 0.05, so there was not enough evidence to suggest that the centroids between groups were different. Although no signals were detected using this method, it is still a useful tool in other work comparing microbiome compositions between experimental groups[28,29]. However, statistical inference using PERMANOVA only provides insight as whether group compositions are different[27]. It does not describe how the groups are different (if at all), nor what those differences between the groups mean in a health context.

Another community level approach for analyzing microbiome data is clustering. The objective of clustering is to put samples into groups based on the similarity of their observations to those of other samples. Although no hypothesis testing is performed using clustering, this exploratory method provides community level insight by comparing groups of clustered samples as determined by their observed taxa to their designated experimental groups. Clustering analysis relies on a distance measure to form groups[30,31]. Bray-Curtis dissimilarity is a common distance measure used for microbiome data, and was discussed earlier. After calculating the Bray-Curtis dissimilarity for each pair of samples, clusters of samples are formed using the desired clustering method. Illustrated below (Figure 1.1) is a commonly used clustering method for microbiome analysis, Ward's Minimum Variance Clustering[32,33].

This method considers clustering as an analysis of variance problem, and aims to minimize the within-cluster sums of squared distances between samples. The initial state of the algorithm is all n samples belong to clusters of size 1 each. In the first step, $n-1$ clusters are formed, one with size 2 and the others with size $n-1$. The pair of samples that yield the smallest residual sum of squares (SS_R) becomes the first cluster. In the next step, $n-2$ clusters are formed from the remaining $n-1$ defined in the previous step. Again, the samples are grouped such that SS_R is minimized. This process continues iteratively until all samples are combined into a single cluster of size n[34]. In addition to PERMANOVA, clustering was used to perform bacterial community level analysis. This clustering was accompanied by a heatmap, to further examine how select species influence the clusters from the resulting analysis[12].

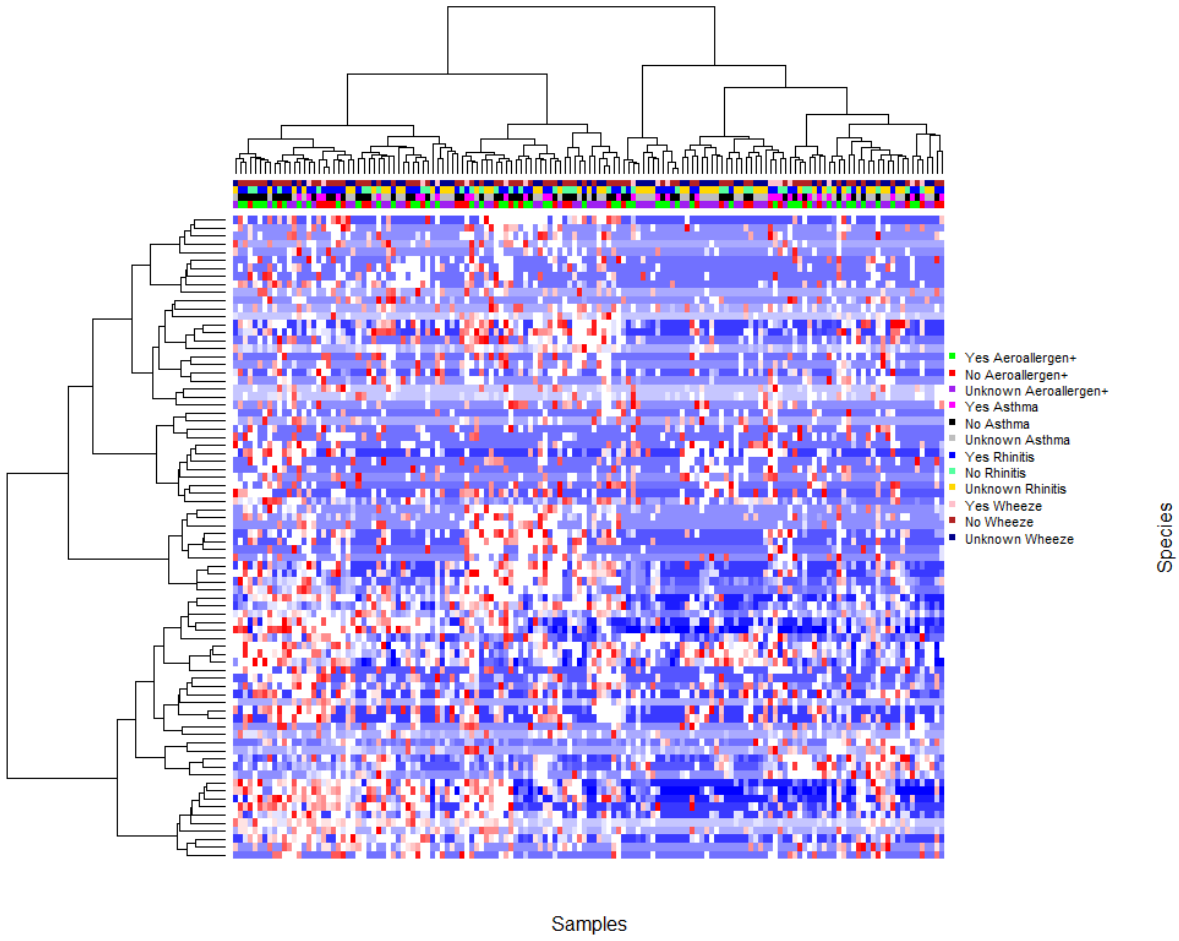

Figure 1.1: *Clustering of bacterial taxa using Bray's dissimilarity and Ward's method*

The experimental groups corresponding to each sample indicated by the color bars are underneath the dendrogram displayed above the heatmap along the x-axis (Figure 1.1). Based on sample labels, the use of clustering analysis did not achieve group separation for any outcomes of interest. This type of exploratory analysis can be useful to understand how the similarities between samples relate to community classification, but interpretation of patterns is up to the researcher's discretion.

1.2.3 Regression Approaches in Differential Abundance Analysis

Another point of interest for researchers is identifying taxa that are differentially abundant between groups defined by experimental conditions[35], which is similar to differential expression analysis with ribonucleic acid sequencing (RNA-seq) data. As with gene expression, a taxon is differentially abundant if its mean proportion is significantly different between experimental groups. RNA-seq data and microbiome data a fundamentally similar, they both consist of over-dispersed counts, where for RNA-seq those counts represent genes and for microbiome data they represent taxa[36]. Since microbiome data is similar to RNA-seq data and is generated identically as well, methods that were developed for differential gene expression analysis can also be applied to microbiome differential abundance analysis. One popular method that has been co-opted for microbiome analysis is DESeq2[11,12,35]. DESeq2 uses shrinkage estimation to deal with count dispersion and focuses on fold changes between groups for increased interpretability of estimates. It has been shown that a negative binomial distribution is a good fit for read count data, including microbiome data[36,37, 38]. Consider a matrix of taxa counts, one row for each taxon i and one column for each sample j. Let Y_{ij} represent the read counts in sample j for taxon i, then

$$Y_{ij} \sim NB(\mu_{ij}, \phi_i \mu_{ij}^2) \tag{1.9}$$

where μ_{ij} is considered as value q_{ij}, proportional to the number of read counts in a

sample scaled by a normalization factor s_{ij} to account for differences in sequencing depth between samples. The DESeq2 pipeline estimates size factors s_{ij} using a median-of-ratios method. After calculating dispersion factors and scaling value q_{ij}, differentially abundant taxa are identified by fitting negative binomially distributed generalized linear models (GLM) using logarithmic link function

$$log_2(q_{ij}) = x_{j.}\beta_i \tag{1.10}$$

where $x_{j.}$ is the model matrix column for sample j and β_i is the log-fold change for taxon i. After GLMs are fit univariately for each taxon, the estimate of logarithmic fold change (differential abundance) between experimental groups β_i is tested for significance using a Wald test[38]. The Wald test compares the estimate β_i divided by its estimated standard error to a standard normal distribution. The resulting p values are then adjusted for multiple testing using the Benjamini-Hochberg correction method. Work published by Cox et al utilized the DESeq2 analytic pipeline[12].

Table 1.4: Log 2-fold changes in differentially abundant bacterial and fungal species

Outcome	Species	Asthma		Wheeze		Aeroallergen+		Rhinitis	
		Age 7	Age 12	Age 7	Age 12	Age 7	Age 12	Age 7	Age 12
Positive	*A. tenebrio*	-2.69	-2.04	-3.28	-	-	-	-	-
	A. cibarius	-	-2.76	-2.15	-2.46	-	-	-	-
	B. maydis	-	-2.45	-	-	-	-	-	-
	C. lunata	-	-2.16	-	-2.73	-	-	-	-
	D. strelitziicola	-2.22	-2.55	-3.71	-	-	-	-	-
	M. tassiana	-	-2.46	-	-2.03	-	-	-	-
	N. oryzae	-	-2.45	-	-2.70	-	-	-	-
	P. aurea	-2.03	-	-	-	-	-	-	-
	P. podocarpi	-	-2.45	-	-	-	-	-	-
	S. flava	-	-2.07	-	-	-	-	-	-
	V. carnescens	-	-	-	-2.11	-	-	-	-
	V. victoriae	-2.20	-2.60	-	-2.34	-	-	-	-
Positive and Negative	*C. parapsilosis*	-	-	-2.98	-	-	-	-	3.39
	C. apollinis	-3.31	-3.89	-4.35	-3.41	-2.57	-	2.16	2.57
	C. americana	-2.56	-3.10	-	-2.62	-	2.71	-	2.78
	E. xenobiotica	-2.61	-	-2.38	-2.31	-	2.12	-	-
	F. oeirense	-	-	-2.91	-	-	-	-	2.90
	G. intricans	-3.27	-2.78	-2.95	-2.19	-	-	-	2.14
	R. mucilaginosa	-3.06	-	-2.79	-	-	3.69	-	2.28
	R. taiwanensis	-	-	-	-	-2.17	2.06	-	2.75
	T. irritans	-3.63	-5.83	-3.26	-5.29	-3.32	-3.42	-3.13	4.11
Negative	*A. sydowii*	-	-	-	-	-	2.83	-	3.15
	C. tropicalis	-	-	-	-	-	2.92	-	2.00
	C. cygneicollum	-	-	-	2.23	-	-	-	-
	D. catenulata	-	-	2.39	-	-	-	-	-
	F. acutatum	-	-	-	-	-	2.20	-	-
	N. albida	-	-	-	-	-	2.14	-	-

Many differentially abundant taxa across multiple health outcomes of interest were identified. The abundance of all species shown in the Table 1.4 above was significantly different among the groups after the adjustment for multiple testing, with log-fold changes in the magnitudes of 2 or greater (magnitude log-fold change of 2 being a biologically meaningful threshold). However, one limitation of the DESeq2 analysis pipeline is that all GLMs are fit univariately, and thus do not consider the entirety of the microbiome simultaneously. The effect of taxon's association with the outcome is estimated independently of the other taxa. The organisms that constitute the microbiome exist as a community and do not reside only in isolation. The effects of the microbiome on a sample are the result of a mixture of those exposures.

Standard approaches to quantify the effect of multiple predictors on a response are logistic regression or multiple linear regression, although usually for health related studies the response is categorical as determined by the experimental design, (thereby indicating the need for logistic regression). However, both of these methods struggle with handling some of the features typical of microbiome data, namely its high dimensionality. Fitting these types of regression models with large numbers of p predictors often provides a "better" fit of the observed data, however these estimates lead to many problems. The inclusion of too many variables can inflate the variance of the estimates, leading to a decrease in the model's predictive ability and interpretability, in addition to presenting false effects. To perform regression analysis considering the entirety of the microbiome, variable selection is essential.

To solve the problem of feature selection in regression models, least absolute shrinkage and selection operator (LASSO) regression and Elastic-Net Regression are used. Like the DESeq2 pipeline, LASSO regression also performs shrinkage. Consider a linear regression model with predictors x_{ij} and response values yi for samples $i = 1, 2, ..., N$ and predictors $j = 1, 2, ...p$. LASSO regression solves the l_1-penalized regression of finding $\beta = \{\beta_j\}$ that minimizes equation 1.11 below.

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j}^{p-1} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.11}$$

Some β_j are shrunk to zero, thereby selecting for predictors x_j, resulting in a more interpretable model. However, it has been demonstrated in cases with high correlations between predictors LASSO regression randomly selects are predictor from the set of correlated ones, where the implication is that those not selected have no association with the outcome[39]. With microbiome data analysis, this leads to problems when taxa are highly dependent on one another. An example of fungal microbiome correlation structure using data from previous work[12] is shown (Figure 1.2). Note, that Spearman's correlation coefficient is used.

Few fungal species were negatively correlated, but some groups of somewhat positively correlated species were identified (Figure 1.2). Consequently, it was decided that LASSO regression was unsuitable. To overcome the limitations of LASSO regression in selecting amongst correlated predictors, elastic net regression applies both l_1 (as in LASSO) and

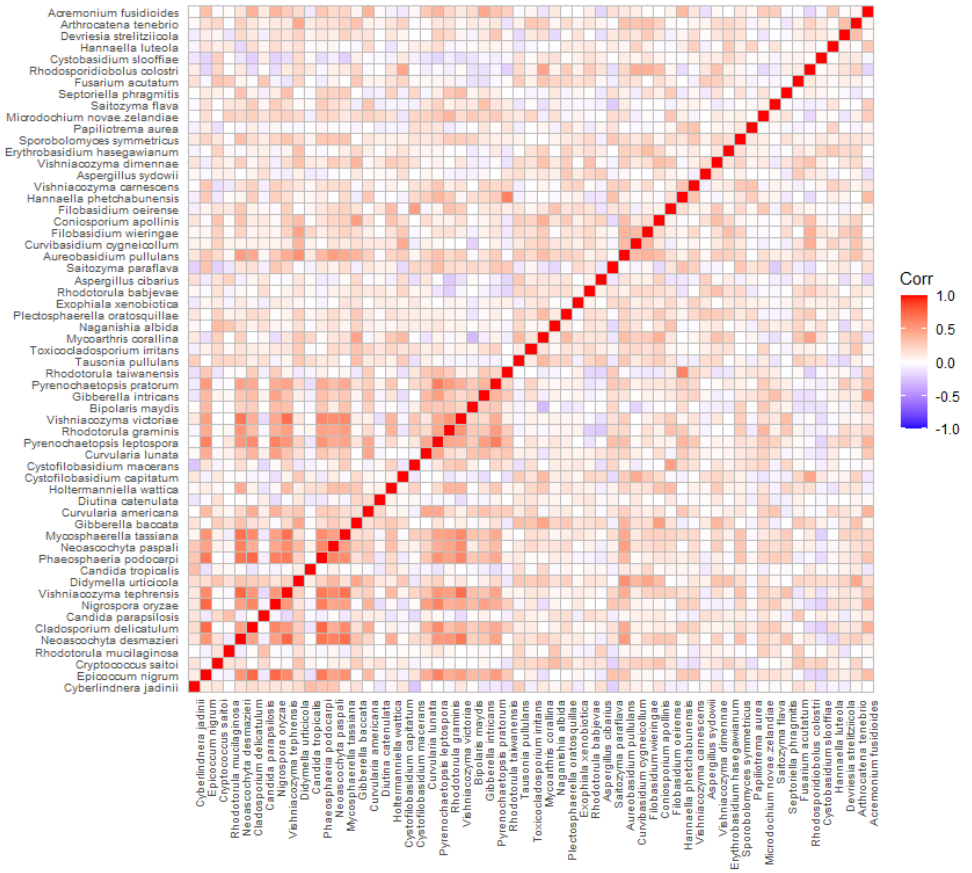

Figure 1.2: *Correlation matrix consisting of fungal taxa calculated using Spearman's* ρ

l_2 (as in Ridge Regression) penalties to the predictor coefficients, and uses a tuning parameter to determine the strength of the respective penalties that minimize equation 1.12.

$$\hat{\beta}_{elastic\ net} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}(\alpha|\beta_j| + (1-\alpha)\beta_j^2) \tag{1.12}$$

This encourages a grouping effect for correlated predictors whose coefficients are either all eliminated from the model or all selected[40]. While the grouping effect from elastic net (specifically the l_2 regularization penalty) results in more stable estimates than

in LASSO regression, this can be problematic for mixtures where correlations among predictors are due to exposure or behavioral patterns that are not necessarily associated with outcomes of interest.

Weighted Quantile Sum (WQS) regression is a method that was developed to overcome the previously discussed limitations of LASSO regression and elastic net. WQS regression empirically constructs a single weighted index of predictors for use in a regression model. The WQS method condenses tests of association for many predictors into a model using one predictor, which consists of weighted combinations of all the components of interest[41]. WQS was applied to the data to perform feature selection and quantify the effect of the microbiome on response variables of interest[42].

Table 1.5: Associations between unadjusted WQS mixtures and the presence or absence of health outcomes of asthma, wheeze, aeroallergen positivity, and rhinitis

Age			**Positive Fungi**				**Negative Fungi**		
		OR	Lower Bound	Upper Bound	p-value	OR	Lower Bound	Upper Bound	p-value
Age 7	Asthma	0.82	0.71	0.93	<0.01	1.20	1.07	1.34	<0.01
	Wheeze	0.68	0.55	0.85	<0.01	1.27	1.13	1.43	<0.001
	Aeroallergen+	0.91	0.83	1.01	0.09	1.22	1.11	1.35	<0.001
	Rhinitis	0.87	0.78	0.96	<0.05	1.13	1.03	1.24	<0.05
Age 12	Asthma	0.84	0.75	0.94	<0.01	1.24	1.09	1.41	<0.01
	Wheeze	0.81	0.68	0.97	<0.05	1.27	1.09	1.47	<0.01
	Aeroallergen+	0.92	0.83	1.02	0.12	1.32	1.13	1.54	<0.01
	Rhinitis	0.95	0.87	1.05	0.33	1.18	1.05	1.33	<0.01
			Positive Bacteria				**Negative Bacteria**		
		OR	Lower Bound	Upper Bound	p-value	OR	Lower Bound	Upper Bound	p-value
Age 7	Asthma	0.76	0.63	0.90	<0.05	1.06	0.98	1.13	0.13
	Wheeze	0.83	0.73	0.95	<0.05	1.09	1.01	1.19	0.06
	Aeroallergen+	0.93	0.87	1.00	0.06	1.11	1.04	1.18	<0.01
	Rhinitis	0.87	0.80	0.96	<0.05	1.09	1.00	1.19	0.06
Age 12	Asthma	0.89	0.80	0.99	0.06	1.16	1.05	1.27	<0.01
	Wheeze	0.83	0.68	1.01	0.07	1.16	1.05	1.28	<0.01
	Aeroallergen+	0.93	0.85	1.01	0.09	1.15	1.04	1.27	<0.05
	Rhinitis	0.86	0.76	0.96	<0.05	1.09	1.00	1.19	0.06

With this novel application of WQS to microbiome data, significant associations between weighted mixture indices of microbiota counts and childhood respiratory health outcomes were identified. Due to the conditions imposed on index term in WQS regression, the models were fit to be either positively or negatively associated with outcome[41]. This nuance and other aspects concerning WQS regression will be discussed in more detail in the next chapter.

2 Weighted Quantile Sum Regression

Analyzing correlated predictors presents many challenges for researchers, especially in the context of environmental exposures or biological response -omics type research. The standard for quantifying the effect of predictors on a response is linear regression through use of one or multiple predictor variables. Previously, the limitations of univariate linear regression were discussed alongside the importance of simultaneous consideration of all predictors. LASSO and elastic net regression were presented as methods that consider all predictors in the model while also performing variable selection. The use of LASSO regression for analyzing microbiome data is limited due to high correlation between sets of taxa[40], and elastic net regression is unsuitable in cases where correlations between predictors are due to similar sources of exposure and not necessarily due to their shared relationship with the outcome. In the previous chapter, weighted quantile sum (WQS) regression was briefly introduced as a method that addresses the limitations of LASSO and elastic net regression in analyzing microbiome data. WQS regression constructs a single weighted index of predictors for use in a regression model. The model condenses tests of association for many predictors into one using only the weighted index[41]. In this section, WQS regression and its variants are discussed in further detail, in addition to a novel procedure for selecting a value for selection threshold parameter τ. Simulation studies are performed to evaluate how different data structures influence the accuracy of the WQS model and optimal values of τ.

2.1 A General Framework for Weighted Quantile Sum Regression

Weighted Quantile Sum regression constructs a weighted index of predictors for use in a singular model. The model condenses tests of association for many predictors into one, using only the weighted index[41]. Fitting a WQS model involves multiple steps. First, values for each predictor are scored into quantiles for every component. The values of the quantiled predictors are then combined into an index, referred to as the WQS index. The WQS index is then used to fit a linear regression model, where the weights of the components and the regression coefficients are estimated simultaneously. The basic WQS model is shown in equation 2.1.

$$g(\cdot) = \beta_0 + \beta_1(\sum_{j=1}^{c} w_j q_{ji}) + z_i'\phi \tag{2.1}$$

In 2.1, the weighted index is given by $\sum_{j=1}^{c} w_j q_{ji}$, where q_{ji} is the quantile of predictor j for the ith sample, and w_j is the weight of each $j_{1,\ldots,c}$ predictors included in the index. Two conditions are imposed upon the weights:

$$\begin{aligned} 0 < w_j < 1 \\ 1 = \sum_{j=1}^{c} w_j \end{aligned} \tag{2.2}$$

By constraining the weights to sum to 1 and fall within a range of 0 and 1, WQS regression reduces dimensionality through near-zero weights and diminishes potential issues with

collinearity. Interpretation of the model follows naturally, where the individual weight w_j indicates the relative importance of that component in the mixture's association with outcome. The intercept is given by β_0, similar to a generalized linear model, and the effect of the mixture is summarized by parameter β_1. Both parameters are related to the outcome of interest using any monotonic, differentiable link function $g(\cdot)$. The effects of covariates not included in the index term and their corresponding associations with outcome are represented by z_i' and ϕ respectively. WQS makes a critical assumption of unidirectional association between components comprising the index with respect to the outcome. In other words, a WQS model is fit to identify mixtures of predictors to be either positively or negatively associated with outcome. By limiting the direction of association, the model avoids the reversal paradox. Without this assumption, the inclusion of correlated predictors with opposite signs in a single index could cancel out their respective associations with the outcome. Consequently, it is necessary to fit two WQS models separately to assess the association between predictors and outcome in both positive and negative directions.

To improve estimate stability, B number of bootstrap samples are generated and are used to fit a corresponding number of models described by 2.1. The estimated weights from each model that have a statistically significant β_1 parameter are then averaged to obtain the final WQS index shown in 2.3.

$$WQS_{final} = \sum_{j=1}^{c} \bar{w}_j q_{ji} \tag{2.3}$$

The final index can then be used to test for associations between the mixture and outcome in a generalized linear model, similar in structure to 2.1. Recent work proposed a repeated holdout validation procedure, combining cross-validation and bootstrapping steps to estimate parameters in the model[43]. Data are randomly partitioned (with replacement) 100 times, and each partitioned data set is used to fit a WQS model. The model fit for each data partition also incorporates the bootstrap step to ensure weight stability. The final component weights and beta coefficients equal the average across each partition. This method provides approximately normal distributions of beta coefficients and component weights, which allows for characterization of component weight uncertainty.

2.2 WQS Regression: Variations

It is important to discuss other variations of the WQS regression approach. These extensions build upon the general framework described in the previous section, in that they all utilize the strategy of constructing a quantile index with empirically determined weights, the WQS term, to model the effect of a mixture of correlated components on an outcome. Conditions for the application of a specific WQS extension vary with the characteristics of the data being analyzed.

2.2.1 Grouped WQS

Grouped WQS considers sets of predictors partitioned into multiple WQS indices, or groups, within a singular generalized linear model. Different magnitudes of effect and

directions of association for each group are considered simultaneously within the model. Groups are constructed based on the similarities between predictors[44]. Note, each predictor may only belong to one group.

$$g(\cdot) = \beta_0 + \sum_{j=1}^{K}[\beta_j(\sum_{i=1}^{c_j} w_{ji}q_{ji})] + z_i'\phi \tag{2.4}$$

In 2.4, observe K number of WQS terms and non-intercept coefficients. Each term is independently subjected to the same constraints described in 2.2. Estimation of the component weights and β coefficients involves bootstrap and nonlinear optimization steps to maximize the log likelihood as with basic WQS[44]. Threshold selection parameter τ is used to identify meaningfully associated components within each group.

2.2.2 Bayesian WQS

The variations of WQS discussed earlier rely on a random split of the data into training and validation subsets. The training set is used to estimate predictor weights averaged across bootstrap samples. Using the validation subset, the coefficients of the weighted mixture are estimated[41,44]. This internal splitting of the data can reduce the statistical power and may lead to unstable estimates, especially with small sample sizes. A Bayesian extension of WQS was developed to overcome these limitations. Bayesian WQS regression uses the GLM framework described in equation 2.1 to model association

between the mixture and outcome Y.

$$g(\cdot) = \beta_0 + \beta_1[\sum_{j=1}^{C} w_j q_{ji})] + z_i'\phi \tag{2.5}$$

The values for correlated mixture components C are scored into quantiles q_{ji} for each $j = 1, ..., C$ predictor and $i = 1, ..., N$ samples. As before, $g(\cdot)$ is a monotonic differentiable link function, β_0 is the intercept, β_1 is the effect of the weighted index, given by $\sum_{j=1}^{c} w_j q_{ji}$, and z_i' is a vector of covariates with their corresponding effects in ϕ. The index weights $w_1, ..., w_C$ are assigned a Dirichlet prior with parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_c)$. The Dirichlet prior assures that the weights are compliant with the usual constraints imposed on weights in other WQS approaches, $w_j \in (0, 1)$ and $\sum_{j=1}^{C} w_j = 1$. The intercept, index regression coefficients, and covariate regression coefficients are assigned uninformative normal priors with $\mu = 0$ and large variance[45].

BWQS requires prior probability distributions on all parameters in the model. Typically, an uninformative prior is assumed, defined as a normal distribution centered around 0 with large variance. This allows for the estimate of the mixture effect to assume all real numbers, without selecting *a priori* the direction of association the mixture has with the outcome. Informative priors can also be defined and embedded into the model when more information regarding the effect of the mixture on outcome is known. Markov Chain Monte Carlo (MCMC) procedures are used to obtain estimates for the model parameters, β_0,β_1,$w_1, ...w_c$. The Bayesian approach can also be generalized for grouped

WQS discussed earlier. Here, the priors for parameters in grouped WQS are similar to the BWQS regression model. A Dirichlet prior for the weights of each group index is assumed and the mixture effect for each group uses an uninformative prior. That is, a normal distribution centered around 0 with large variance. Important predictors are identified by comparing their weight estimates to *a prior* selected threshold τ[45].

2.2.3 Lagged WQS Regression

Lagged WQS regression (LWQS) is an approach that evaluates associations between fixed response data and mixed effects of multiple time varying predictors. LWQS applies the logic of using an empirically constructed index consisting of weighted quantiled predictors to a reverse distributed lag model (DLM). In a reverse DLM, the roles of outcomes and exposures are interchanged, and their associations are evaluated using a functional spline model with time varying coefficients[46].

$$X(t) = \beta_0(t) + \beta_1(t)Y + \gamma z + u + \epsilon \tag{2.6}$$

The equation of an inverse DLM is shown above in 2.1, where $X(t)$ are standardized exposure concentrations over time, and Y is the standardized outcome variable, γ are covariates and z are their corresponding effects. The focus of the inference is the time-varying correlation between X and outcome Y, represented by $\beta_1(t)$. The random effects term u allows for the assumption of a compound correlation pattern for intra-subject observations. Values of $\beta_1 > 0$ indicate associations between higher mean outcome

values and higher concentrations of exposures[45].

In LWQS, the use of an empirically weighted index of predictors is used to extend the reverse DLM for analysis of mixtures. Fitting an LWQS model requires an ensemble approach. First, the average weight of each component within the mixture is estimated for each time unit and used to calculate a weighted index per sample, given by $WQS_i(t)$. The time-varying association across these indices are evaluated using a reverse DLM model.

$$WQS(t) = \beta_0(t) + \beta_1(t)Y + \gamma z + u + \epsilon \tag{2.7}$$

LWQS allows for inferences in both directions of association with outcome. When the estimate of $\beta_1(t)$ for $WQS(t)$ is positive, the mixture and corresponding weights are positively associated with outcome. When fit in the negative direction, the estimate of $\beta_1(t)$ is negative and the resulting weighted mixture is negatively correlated with outcome. As with other WQS approaches, the index is subject to the same constraints, $w_j \in (0, 1)$ and $\sum_{j=1}^{C} w_j = 1$. Similar to other applications of WQS, variable selection in LWQS compares final weight estimates for the predictors in each mixture index to an a priori selected threshold τ.

2.3 Identifying Meaningful Predictors via Selection Threshold Parameter Tau

WQS regression and its extensions are used to analyze data sets consisting of high dimensional and highly correlated mixtures[47–51]. Although the weight estimates indicate

the relative importance of each component in the mixture's association with outcome, meaningfully associated components are identified by comparing estimated weights with *a priori* selection of threshold parameter τ. Components with weights greater than τ are considered to be meaningfully associated with the outcome of interest. The τ threshold is used to identify meaningful index contributors across all WQS variations. A low value for τ increases the number of incorrectly selected components. Conversely, an excessively high value for τ leads to a decrease in correctly selected components. Therefore, selection of τ is a critical step in WQS regression. The predominant technique for selecting τ uses formula $\frac{1}{c}$, where c is the number of components being assessed in the mixture[43,44]. Here, it is referred to as the heuristic method. Although useful, this approach assumes that no proportion of the total weight is assigned to components erroneously, and that each component contributes the same amount to the mixture. Simulation studies were performed to examine different data conditions which influence the optimal value for the selection threshold parameter τ. In addition, a novel alternative to the heuristic method for selecting a value of τ was proposed.

2.4 Simulation Study

2.4.1 Methods

To assess the factors that affect optimal value for threshold parameter τ, simulation studies with several varying data conditions were performed. Although WQS regression was developed with the intent of analyzing chemical mixture exposures, some researchers

have recently extended this application to microbiome analysis[42,52]. To evaluate the accuracy of expanded use cases for WQS, simulations using two different data structures were performed, both based on separate bacterial and fungal taxonomic count data from a previous study[42]. The two data sets were referred as structure 1 and structure 2. In simulations with structure 1, the correlations between the predictors were similar to the correlations observed in the bacterial count data; the correlations in structure 2 were similar to the correlations in the fungal taxonomic count data. The impact of dimension number on optimal value for τ was also assessed because the current standard for selecting τ is entirely reliant on the number of predictors in the data. Two cases of data dimensions were considered: either the full number of predictors corresponding to the respective correlation structure were used, or a reduced number, 30. In complete cases, data structures 1 and 2 had 80 and 59 dimensions, respectively. The number of true signals in each simulation was also varied. Each simulation had either 5,10,15 or 20 signals. Naturally, the effect of signal strength on optimal τ was also assessed. The correlation between predictor and response were considered as measures of signal strength. Individual microbiota are often only weakly associated with an outcome of interest. More commonly, they form bundles of dense signals, where many taxa are weakly associated with the outcome, but together have a strong joint effect[53,54]. Therefore, $\rho = 0.25$ was used as the upper limit for signal strength in the simulations. Three signal strength scenarios were used: a hard scenario, with each ρ between 0.05 and 0.15, a medium scenario, with each ρ between 0.10 and 0.20, and an easy scenario, with each ρ between 0.15 and 0.25. Associations between other predictors and the outcome

were set to 0. Furthermore, different levels of intercorrelation between predictors were also considered and referred to as damping factors. Non-diagonal entries in correlation matrix derived from structure 1 or structure 2 were multiplied by one of two damping factor levels, 1 or 0.5. See Figure 2.1 for a summary of all conditions used in this study.

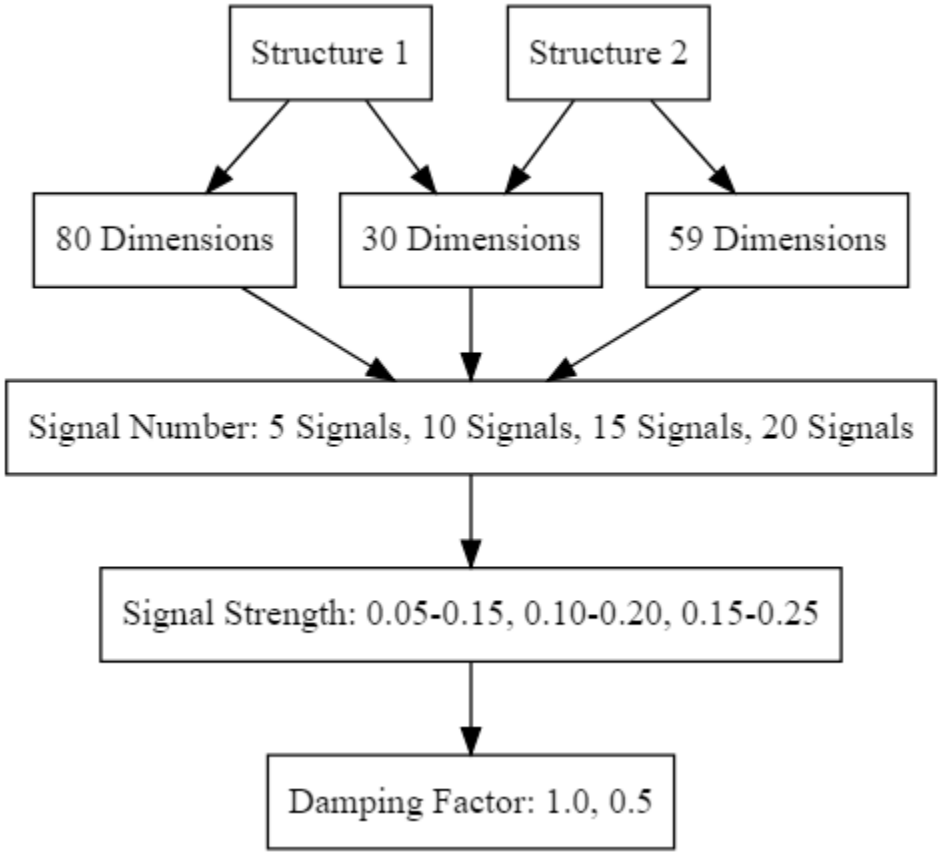

Figure 2.1: *Flowchart to describe the various sets of experimental conditions used for simulations.*

Simulations were performed 100 times for each of the 96 unique sets of conditions. For each iteration, correlation matrix D was defined for every possible pair of selected c components. A number of values equal to the number of specified signals were sampled from a uniform distribution, with parameters a and b set according to the corresponding

signal strength ranges from Figure 2.1. All other associations between $x_{1,...,c}$ and y were set to 0. The resulting vector was defined as Y_ρ. Two-hundred values were sampled from a multivariate normal distribution with parameters $\boldsymbol{\mu} = [0, ..., 0]^T$ and

$$\Sigma = \begin{bmatrix} D & Y_\rho \\ Y_\rho^T & 1 \end{bmatrix} \tag{2.8}$$

Prior to sampling, Σ was made to be a positive definite matrix[55]. Note that microbiome taxon count data, on which these simulations are based, is often modeled using a negative-binomial distribution[37,56,57]. However, since the WQS model quantiles predictors to estimate their respective effects, the actual simulated values are negligible so long as the relationships among predictors are preserved. Similarly, Σ consists of the correlation matrix between predictors, rather than the covariance. Other work describing multiple imputation using WQS also utilizes a similar logic[58]. Selection threshold parameter τ_h was calculated using the heuristic (or $\frac{1}{c}$) method. The newly proposed ROC optimized selection threshold parameter τ_{ROC} was calculated by identifying the point along a model's ROC curve that minimized the distance to the point $(0, 1)$, indicating a perfect classifier. Threshold parameter τ_{ROC} was then selected corresponding to the sensitivity and specificity of the model at that point. Measured as area-under-the-curve, model accuracy between varying simulation conditions were compared using two-sample t-tests. Similarly, mean percent change between τ_{ROC} and τ_h was also evaluated for each set simulation conditions using one-sample t-tests. The effects of changing simulation

parameters on values of τ_{ROC} were evaluated using linear regression and ANOVA. All simulations and statistical analysis were done using R v4.0.5.

2.4.2 Results

First, WQS models were fit using two correlation structures for simulated data and assessed how varying experimental conditions impacted accuracy of the models and optimal selection of selection threshold parameter τ (Figure 2.1). Simulations were designed with varying degrees of difficulty. It was posited that scenarios with high signal strength, few dimensions, and low intercorrelation among predictors would have higher overall accuracy than models fit using data sets with lower signal strength, high dimensions and greater intercorrelation among predictors. Table 2.1 shows AUC comparisons between simulations with varying experimental conditions. To allow for balanced design, two separate comparisons for data dimensions were performed. Differences were evaluated between other scenarios using simulation conditions that used 30 dimensions only. Note that the comparison between 30 and 80 dimensions only considered the data structure 1 simulations, while the comparison between 30 and 59 dimensions only utilized the data structure 2 simulations.

Mean area under the curve (AUC) for simulations using data structure 1 was greater than that for data structure 2 (Table 2.1). As shown in Table 2.1, the 95% confidence interval for AUC estimates was wider for structure 2 scenarios than for structure 1 scenarios. However, the differences in accuracy between structures 1 and 2 were not statistically significant (Table 2.1). Results in Table 2.1 also demonstrate that, for

Table 2.1: Comparisons of AUC for varying experimental conditions ROC. Dimension comparisons for 30 to 80 and 30 to 59 utilize data structures 1 and 2, respectively.

	Conditions	Mean	CI Lower	CI Upper	P Value
Structure					0.4254
	1	0.8439	0.8173	0.8704	
	2	0.8288	0.7647	0.8930	
Dimensions					0.1631
	30	0.8169	0.7517	0.8822	
	80	0.8439	0.8168	0.8709	
Dimensions					0.3558
	30	0.8122	0.7509	0.8735	
	59	0.8288	0.8034	0.8543	
Signals					0.5950
	5	0.8562	0.7647	0.9478	
	10	0.8392	0.8012	0.8771	
	15	0.8279	0.7363	0.9195	
	20	0.8221	0.7306	0.9137	
Damping Factor					0.6963
	0.5	0.8400	0.8133	0.8668	
	1	0.8327	0.7682	0.8972	
Signal Strength					<0.0001
	0.05-0.15	0.7577	0.7466	0.7689	
	0.10-0.20	0.8480	0.8210	0.8749	
	0.15-0.25	0.9034	0.8765	0.9303	

structure 1 scenarios with 30 dimensions and 80 dimensions, the mean AUC was greater than that for the 30-dimension data, but no statistical difference was detected. Similarly, varying data dimensions for structure 2 scenarios resulted in a higher mean AUC for the higher dimension count (59) than when the same data structure but with 30 dimensions was used. Nonetheless, these differences were not statistically significant (Table 2.1). Interestingly, using data with fewer signals resulted in greater mean AUCs for scenarios with 5, 10, 15, and 20 signals (Table 2.1). The mean AUC for simulations with a damping factor of 0.5 was greater than the mean AUC for simulations with a damping

factor of 1. However, this modest increase in accuracy was not statistically significant (Table 2.1). The only experimental factor that resulted in significant differences in model accuracy was signal strength ($p < 0.0001$), where increases in signal strength resulted in higher model accuracy (Table 2.1).

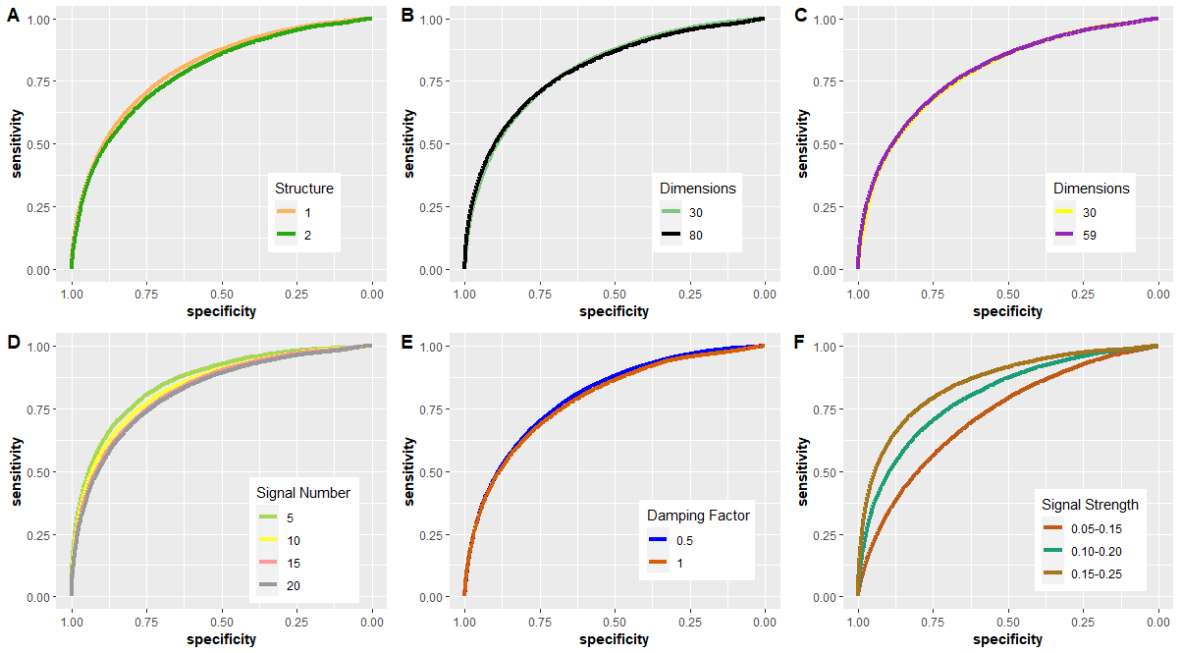

Figure 2.2: *ROC curves for models fit with differing sets of experimental conditions. Considerations for panel (A): Correlation Structures 1 and 2; (B): 30 dimensions and 80 dimensions; (C): 30 dimensions and 59 dimensions; (D): 5, 10, 15, and 20 signals; (E): Damping factor 0.5 or 1; (F): Signal strength ranges 0.05-0.15, 0.10-0.20, and 0.15-0.25.*

ROC curves for model accuracy estimation are shown in Figure 2.2, with each panel corresponding to each of the comparisons displayed in Table 2.1. Note that the ROC curves in Figure 2.2 consist of all results for a specific set of experimental conditions unlike the comparisons from Table 2.1 which calculate AUC for each unique set of conditions. Despite this nuance, results similar to the comparisons shown in Table 2.1 were observed. The ROC curves comparing model accuracy for varying data structure,

dimension number, number of signals, and damping factor are not different from each other reflecting the results in Table 2.1 (Figure 2.2). The changes in the corresponding conditions did not significantly affect the accuracy of the model. The ROC curves exhibit the greatest degree of separation in Figure 2.2F, further demonstrating that changes in signal strength have a profound impact on model accuracy.

After evaluating the effect of varying experimental conditions on the accuracy of WQS regression, the ROC optimization procedure was implemented to determine selection threshold parameter τ_{ROC} for each set of conditions. Selection threshold parameters τ_h were calculated using the heuristic method. For each dimension total, 30, 59, and 80, τ_h was 0.0333, 0.0169, and 0.0125, respectively.

The difference between τ_{ROC} and τ_h was less than 0 for all simulation conditions except when 5 signals were used (Table 2.2). In other words, τ_h overestimates the optimal value of the selection threshold parameter in most cases. Density plots of percent change for each comparison reported in Table 2.2 are shown in Figure 2.3. For every set of experimental conditions except simulations with 5 signals, most of the density for percent change between τ_{ROC} and τ_h was below 0, further demonstrating that in most cases, the heuristic method for determining τ diminishes the model's ability to correctly identify predictors truly associated with the outcome.

The effects of changing simulation conditions on mean percent difference between τ_{ROC} and τ_h were further quantified using linear regression and ANOVA. For each comparison, the reference variable is the first listed simulation condition. Mean percent change

Table 2.2: Mean percent differences between τ_{ROC} and τ_h for each set of compared experimental conditions

	Condition	Mean Change	LB	UB	P-Value
Structure					
	1	−43.57	−55.95	−31.19	<0.0001
	2	−36.84	−49.66	−24.03	<0.0001
Dimensions					
	30	−43.57	−55.95	−31.19	<0.0001
	80	−21.28	−40.67	−1.90	0.0328
Dimensions					
	30	−36.84	−49.66	−24.03	<0.0001
	59	−26.87	−37.22	−16.51	<0.0001
Signals					
	5	0.42	−11.72	12.56	0.9403
	10	−36.43	−44.21	−28.66	<0.0001
	15	−56.62	−64.54	−48.71	<0.0001
	20	−68.19	−75.62	−60.76	<0.0001
Damping Factor					
	0.5	−31.29	−43.26	−19.32	<0.0001
	1	−49.12	−61.32	−36.92	<0.0001
Signal Strength					
	0.05-0.15	−40.49	−55.06	−29.93	<0.0001
	0.10-0.20	−39.88	−55.16	−24.60	0.0001
	0.15-0.25	−40.24	−61.22	−19.27	0.0010

between τ_{roc} and τ_h varied significantly among changing number of signals and damping factor (Table 2.3). However, for both sets of dimension comparisons, the magnitude of mean percent difference between threshold selection parameters decreased, as the number of dimensions increased. Interestingly, varying signal strength had minimal impact on mean percent difference between τ_{ROC} and τ_h.

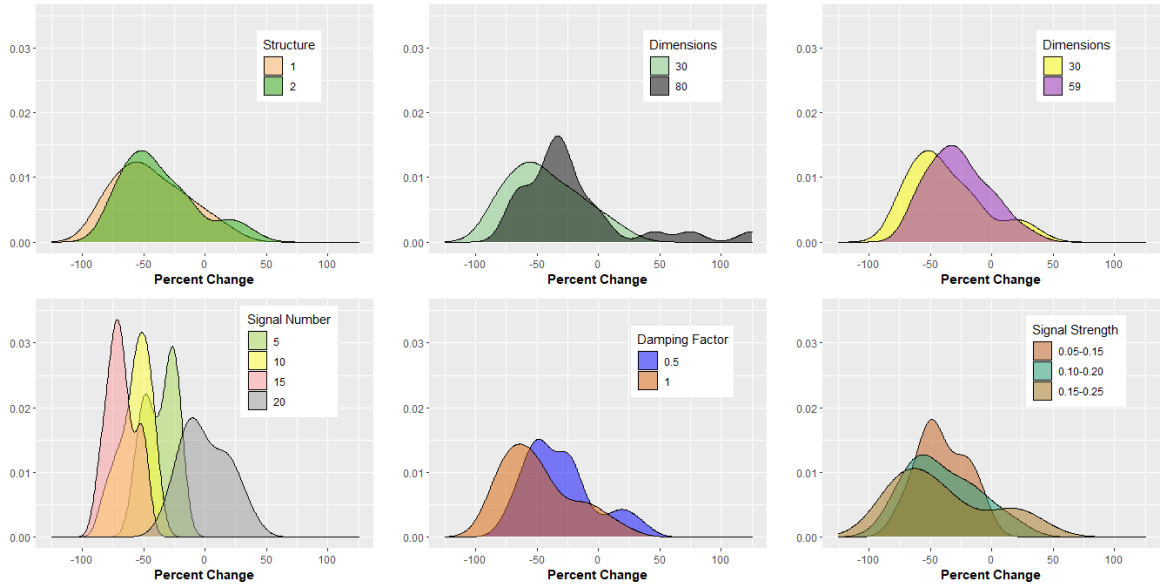

Figure 2.3: *Probability density curves for models fit with differing sets of experimental conditions. Considerations for panel (A): Correlation Structures 1 and 2; (B): 30 dimensions and 80 dimensions; (C): 30 dimensions and 59 dimensions; (D): 5, 10, 15, and 20 signals; (E): Damping factor 0.5 or 1; (F): Signal strength ranges 0.05-0.15, 0.10-0.20, and 0.15-0.25.*

2.4.3 Discussion

WQS regression is a useful and versatile tool to quantify the effects of mixtures of components on an outcome of interest. By combining the components into a weighted index, researchers can identify meaningfully associated components by comparing their weight within the index to threshold selection parameter τ. Selecting τ is a key part of fitting a WQS model, as a value too low leads to an increased number of incorrectly selected variables, where a value too high leads to an decreased number of correctly selected variables. The current standard for determining τ only considered dimension number, therefore it was necessary to investigate the effects of other conditions on such a critical aspect of WQS regression.

Table 2.3: Linear regression for effect of varying simulation conditions on mean percent difference in selection threshold parameter

	Conditions	Beta	LB CI	UB CI	P-Value
Structure					0.4389
	1	−43.57	−55.83	−31.31	
	2	6.72	−10.61	24.06	
Dimensions					0.0510
	30	−43.57	−55.39	−27.74	
	80	22.28	−0.10	44.66	
Dimensions					0.2166
	30	−36.84	−48.18	−25.51	
	59	9.98	−6.05	26.01	
Signals					<0.0001
	5	0.42	−7.84	8.68	
	10	−36.86	−48.54	−25.17	
	15	−57.05	−68.73	−45.36	
	20	−68.61	−80.29	−56.93	
Damping Factor					0.0361
	0.5	−31.29	−43.05	−19.53	
	1	−17.83	−34.46	−1.20	
Signal Strength					0.9983
	0.05-0.15	−40.49	−55.78	−25.21	
	0.10-0.20	0.62	−21.00	22.24	
	0.15-0.25	0.25	−21.37	21.87	

This simulation study demonstrated the effects of varying data conditions such as correlation structure, dimension number, signal number, correlation among predictors, and signal strength on the performance of WQS models. Interestingly, changes in data structure did not affect the accuracy of the WQS model, supporting further use cases for WQS beyond quantifying the effects of chemical mixture exposures, or more recently, the microbiome. Surprisingly, increasing the number of dimensions resulted in a higher overall accuracy for both sets of dimension comparisons. However, this trend is likely explained by an increase in the number of correctly identified unrelated predictors,

rather than an improvement in the model's ability to identify ones truly associated with the outcome. Increasing the number of signals within the data slightly reduced the accuracy of the model. The constraints applied to the WQS term in the model require a proportion of weight to be assigned to predictors truly associated with the outcome (2.2). On the other hand, a portion of the total weight in the index is assigned erroneously to predictors with no relation to the outcome, thereby reducing the proportion of weight available to be assigned to predictors truly associated with outcome. In other words, more signals lead to lower accuracy in the WQS model because the shared proportion of weight that is assigned to them becomes dispersed as more signals are added to the WQS index. As expected, predictors that had stronger associations with outcome were more readily identified by the model. Damping factor did not significantly impact model accuracy. However, the trend suggested that reduced correlation among predictors facilitates slightly more accurate classification of a predictor's association with outcome.

This work demonstrated that under most simulated conditions, the heuristic method[41] resulted in a much higher selection threshold parameter than the one selected via ROC optimization. However, the approach of using ROC curves to select a value for the selection threshold parameter in a WQS model is contingent upon knowing how many true signals between components and outcome exist. With simulated data, determining the true number of signals in the data is trivial, since the true associations between the predictors and response are defined as such. This results in clear labels of either signal or noise for the resulting predictors, that determine the true positive rate and

false positive rate as the weight threshold is varied. This is critical for construction of the ROC curve and the optimization step performed afterwards. With real data the true number of signals is unknown.

One potential strategy of obtaining a signal number estimate is univariately testing the associations of each predictor with the response and using the number of significant associations as an estimate for the true number of signals for simulations. Such analytic approaches are often used in risk analysis studies to provide preliminary insights[59,60]. The information gleaned from these initial steps can be used to estimate signal number for simulations and ROC optimization to select an ideal τ parameter. However, the main limitation of this approach is that it relies on the power of the method being used to evaluate the univariate associations between response and predictors, and consequently, may underestimate the number of signals present within the data.

Performing simulations and ROC optimization is computationally inefficient, especially for high dimensional data. While the procedure was relatively quick (15 minutes) for the scenarios with 30 dimensions, the simulations which used 80 dimensions took many hours to complete. This is impractical for analysis of non-simulated high dimensional data. These simulations demonstrated that optimal τ was affected by signal number, damping factor, and potentially dimension number. To efficiently and practically implement the ROC optimization procedure to select τ, additional simulations with more variants of the aforementioned conditions should be performed to fully characterize their effect on τ via an equation.

In summary, these simulations provide validation for the application of WQS to analyze high dimensional, correlated data beyond chemical mixture exposures and provide insight into characteristics which impact the current standard used by researchers to identify components of interest within the WQS index. Although the proposed method of ROC optimization to determine selection threshold parameter τ has some limitations, its application can be readily adapted by researchers who wish to utilize WQS as part of their analysis to identify meaningful risk factors within a mixture of predictors while still considering the entirety of the mixture's effect.

3 ROC Optimization for Tau: Using Chemical Biomarker Profiles

The previous chapter described the general framework of the WQS regression approach and also highlighted its original applications in chemical mixture analysis[41]. Simulation studies demonstrated the effectiveness of WQS in analyzing microbiome data and also validated a novel procedure for determining selection threshold parameter τ. The work in this chapter pivots back to the analytic roots of WQS regression in chemical mixture assessment. Based on other work, the similarities between the ensemble modeling approaches of random forests and WQS regression are closely examined. Additionally, WQS and the ROC optimization procedure are used to distinguish chemical biomarker profiles between different levels of tobacco smoke exposure in children.

3.1 Tobacco Smoke Exposure and Classification

The adverse health effects of secondhand tobacco smoke exposure have been well documented since they were considered in the 1972 United States Surgeon General's report, *Health Consequences of Smoking*[61–65]. More recently, there has been increased interest in quantifying the health effects from thirdhand smoke exposure[66–68]. That is, residual tobacco smoke contamination that remains after the cigarette has been extinguished[69]. The thirdhand smoke residue mixture contains tobacco-specific pollutants such as nicotine, nitrosamines, and nicotelline, along with tobacco non-specific pollutants such as polycyclic aromatic hydrocarbons and volatile organic compounds known to be harmful

to humans[70,71]. Nicotine intake from exposure to secondhand and thirdhand smoke can be measured using several metabolites, including cotinine, which is converted to other metabolites such as trans-3'-hydroxycotinine[72]. However, measurement of nicotine intake via cotinine and its metabolites underestimate thirdhand exposure to other toxicants present in tobacco smoke, such as potent carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, which is rapidly metabolized to urinary 4-(methynitrosamino)-1-(3-pyridyl)-1-butanol (NNAL), another potent carcinogen[73]. In addition to these tobacco specific compounds, non-tobacco specific biomarkers 3-hydroxyfluorene, 2-hydroxyfluorene, and N-acetyl-S-(2-cyanoethyl)-L-cysteine also provide insight into levels of tobacco smoke exposure[74,75]. In a recently work[76], random forest models were utilized to distinguish between self-reported thirdhand smoke exposure and mixed secondhand and thirdhand smoke exposure categories using chemical biomarker profiles.

3.2 Random Forests

Random forests are ensemble models consisting of a collection of classification and regression trees (CARTs). Given a set of input variables, each tree calculates a single response, and the aggregate of those responses is the final prediction by the random forest model. For classification problems, the final response is the class that receives the most votes by the trees. For regression problems, the mean of all trees' estimates is the final prediction[77]. Note that CARTs are unable to estimate values outside the range of the response used for training. Although this limitation can be exploited under certain circumstances[76,78], decision trees and random forest models are often more suitable for

classification problems rather than regression. The basic structure of a tree consists of three parts, root nodes, internal nodes, and leaf nodes. The root node is the first node of the tree, which consists of the entire training data. Internal nodes are in the middle of the tree, holding the resulting bins of data after a split has been performed. Leaf nodes are terminal, as the final partition of the training data based on previous splits of the data. Splitting of the root and internal nodes is based on the criterion that minimizes the *Gini Impurity* of the resulting nodes.

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.1}$$

In 3.1, p is the proportion of items in class i, where i is determined by the binary splitting criterion. Node splitting criteria can also be determined using other metrics, most often some measure of variance[79].

Another key feature of random forest models is the bagging method (also called bootstrapping) used to construct each decision tree[77]. Bagging is a resampling technique that improves stability. A random sample of size N is drawn with replacement from the data and used to train a model. The average (or the majority) of the predictions across the bootstrapped samples is used as the final estimate[80]. Additionally, each tree in a random forest model is constructed using a random subset of predictors. This ensures low correlation among trees and limits the influence of especially strong predictors[77], while improving the overall accuracy of the model compared to a single tree. However,

the gains in accuracy provided by using a random forest model instead of a single decision tree come with a loss of general interpretability. The rule-based splitting criteria in a decision tree is easy to follow, whereas understanding the paths of hundreds of trees is considerably more challenging.

Interpretation of a random forest is limited to measures of variable importance within the model. Each tree in the model has its own out-of-bag data that was not used for fitting. Suppose there are C number of predictors used to fit a tree. For each out-of-bag data set corresponding to a tree, the values for one predictor are permuted and the resulting data is run through the corresponding tree. The classification for each out-of-bag sample is saved. This process is repeated for predictors $c = 1, 2..., C$. The decrease in accuracy on the shuffled data is measured by comparing the predictions using permuted data to the data using true labels[77]. Intuitively, permuting the values of a variable that has little predictive power has minimal impact on model accuracy. Conversely, permuting the values of a highly predictive variable reduces model accuracy.

3.3 Description of the Data

Some of the advantages of random forest models are similar to the key benefits of WQS regression described in the previous chapter. Through use of the bootstrapping step, both models have an increased estimate stability and disrupt the correlation structure among predictors. Furthermore, both types of models explain the importance of variables via relative comparisons of one component's predictive ability to another. For random

forest models, this is done by internal variable importance scores[77]. For WQS regression, predictor importance is explained by the component weights within the WQS index[41]. The analysis data set used in Merianos et al[76] contains 16 predictors - some chemical biomarkers measurements, some chemical biomarker ratios, and some categorical questionnaire responses. The response variable, TSE group, was defined using self-reported (or parental proxy) questionnaire data to identify three types of smoking exposure among subjects: no reported thirdhand smoke exposure (NEG), reported thirdhand smoke exposure (TEG), and both reported secondhand plus thirdhand smoke exposure (MEG). Each sample was weighted such that it represents a corresponding number of people within the United States population. To leverage all information within the data, weighted random forest imputation was performed to account for missing values and ensure that similar subjects had similarly imputed values.

The sample weighting allows for generalization of model results to the sample population. However, the objective in comparing WQS and random forest models was to use chemical biomarkers to distinguish between different groups of smoking exposure and not necessarily to quantify the effects of biomarkers in a way that is reflective of the national population. Work by Merianos et al.[76] demonstrated that the subjects belonging to the NEG group were most easily identified by random forest model across all sets of predictors. Correctly identifying subjects belonging to the TEG and MEG groups was more challenging. Here, the comparison between WQS and random forest models uses subjects belonging to either the TEG or MEG classes. The survey sample weights were used for the imputation step, but for model fitting, each sample was inversely weighted

according to the frequency of its respective TSE group. For example, the smaller MEG group was weighted more heavily than the TSE group. This inverse frequency weighting was performed to ensure that the models did not favor prediction of one outcome class more than the other.

Recall that one of the assumptions for the WQS model is that predictors included in WQS the index must be quantiled, and therefore continuous. The survey questions contained in the original data are not. Therefore, only the chemical biomarker data was used in the analysis comparing random forest models and WQS regression. After omitting all NEG subjects and categorical predictors, the resulting data set contained 1116 samples and 12 continuous predictors that can quantiled and reasonably combined into an index.

The correlation pattern (Figure 3.1) among the predictors stems from two sources: correlation as a result of a shared source of exposure by the manner of tobacco smoke residue or correlation as a result of consideration within a ratio. Due to the bootstrapping steps in both random forest models and WQS regression, the challenges encountered when performing analysis with correlated predictors is mitigated.

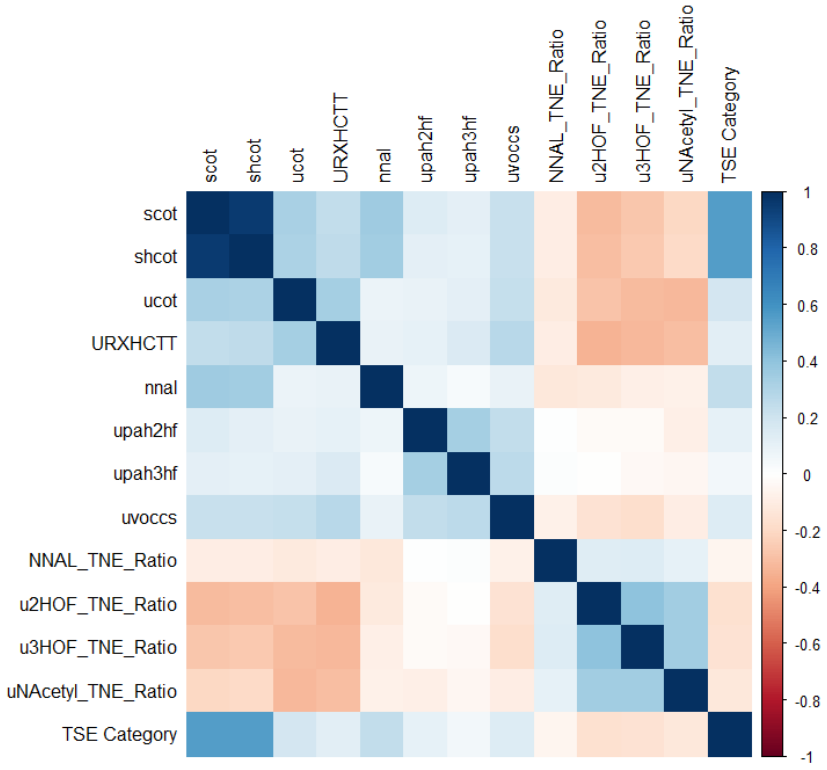

Figure 3.1: *Chemical biomarker correlation matrix calculated using Spearman's ρ*

3.4 Model Comparison and Implementation of Tau Optimization

To implement the τ optimization procedure, it is first necessary to estimate the number of signals, or predictors significantly associated with the outcome. A weighted logistic regression model was fit between each predictor and the TSE exposure group response using TEG as a baseline.

It is necessary to reiterate that the goal of these models is not to assess the effect of the biomarkers individually, but simply to estimate how many signals are required for the τ optimization procedure. Nonetheless, β-estimates and 95% confidence intervals are

Table 3.1: Univariate logistic regression models for each chemical biomarker or ratio of interest

	OR	**CI LB**	**CI UB**	**p-value**
Serum Cotinine	5.14	4.03	6.56	<0.0001
Serum Hydroxycotinine	6.00	4.62	7.79	<0.0001
Urinary Cotinine	1.37	1.18	1.59	<0.0001
Urinary Hydroxycotinine	1.37	1.18	1.58	<0.0001
NNAL	2.25	1.83	2.75	<0.0001
2-Hydroxyfluorene	1.54	1.07	2.22	0.0189
3-Hydroxyfluorene	1.53	1.08	2.18	0.0179
N-acetyl-S-(2-cyanoethyl)-L-cysteine	2.06	1.52	2.79	<0.0001
NNAL\Total Nicotine Equivalent	1.03	0.96	1.10	0.4598
2-Hydroxyfluorene\Total Nicotine Equivalent	1.00	1.00	1.00	0.2476
3-Hydroxyfluorene\Total Nicotine Equivalent	1.00	1.00	1.00	0.9036
NNAL\Total Nicotine Equivalent	1.00	1.00	1.00	0.5386

shown alongside p-values (Table 3.1) to show the direction of association for each of the predictors. Eight chemical biomarkers were positively associated with MEG, indicating that higher levels of these chemicals increases the likelihood of being a member of the MEG exposure group.

3.4.1 Random Forest Model Results

A class weighted random forest model consisting of 1000 trees was fit using the chemical biomarkers and ratios shown in Figure 3.1 with TSE category as the response. Mean decrease in Gini impurity score was used to measure variable importance within the model, where the greater the magnitude of the mean decrease, the greater the importance of that particular predictor.

The random forest model had two groups of predictors. Serum cotinine (scot) and serum hydroxycotinine (shcot) formed the higher tier, with importance greater than

Table 3.2: Random forest variable importance scores

Predictor	Importance Score
Serum Cotinine	134.5575
Serum Hydroxycotinine	104.1869
Urinary Cotinine	36.7037
Urinary Hydroxycotinine	36.0207
NNAL	35.1059
2-Hydroxyfluorene\Total Nicotine Equivalent	34.8293
3-Hydroxyfluorene\Total Nicotine Equivalent	33.6325
N-acetyl-S-(2-cyanoethyl)-L-cysteine\Total Nicotine Equivalent	31.8090
NNAL\Total Nicotine Equivalent	30.0032
2-Hydroxyfluorene	28.3551
3-Hydroxyfluorene	27.4283
N-acetyl-S-(2-cyanoethyl)-L-cysteine	24.8631

100. The remainder formed the lower tier, with variable importance scores in range $[24.8631, 36.7037]$ (Table 3.2). The accuracy of the random forest model was also assessed using an 80% training and 20% split. The model had a 78.48% accuracy rate but did not favor one type of misclassification over another (Table 3.3). In other words, the model struggled equally with properly classifying the TEG subjects as it did with the MEG subjects.

Table 3.3: Random forest model confusion matrix

		Predicted	
		TEG	MEG
Observed	TEG	115	26
	MEG	22	60

3.4.2 Weighted Quantile Sum Results

A weighted quantile sum regression model was fit using the chemical biomarkers and biomarker ratios to predict TSE group membership. Meaningfully associated predictors

were identified using the τ selection threshold procedure described in Chapter 2. Recall the steps in the ROC τ optimization procedure.

First, $p+1$ x $p+1$ correlation matrix D was defined using $1, ...p$ number of predictors and 1 response. All associations among predictors and between predictors and response were calculated using Spearman's rank correlation (Figure 3.1). Data was simulated from a multivariate normal distribution with $\boldsymbol{\mu} = [0, ..., 0]^T$ and variance D. Note, that because the predictors are quantiled, the use of the correlation matrix rather than the covariance matrix to describe the associations among components within the model is preferred, so long as relationships between variables are conserved. The response was also simulated from the multivariate normal distribution. However, the resulting values were converted into a binary response based on quantile. For example, approximately 63% of subjects belonged to the TEG group. Therefore, all simulated responses within the 63rd percentile were set to TEG, the remainder were set to MEG. Although this transformation does not perfectly capture the associations between the predictor and response, it is reasonably close and has been used in other studies[81].

Next, a WQS regression model was fit using the simulated multivariate normal data. The process of simulating data based on correlation structure D and subsequently fitting a WQS model was repeated 500 times. The aggregated final component weights from the 500 WQS models were used to construct an ROC curve.

The optimization procedure identifies a point along the ROC curve which minimizes the euclidean distance to the point (1,1), indicative of a perfect classifier. A value of τ is

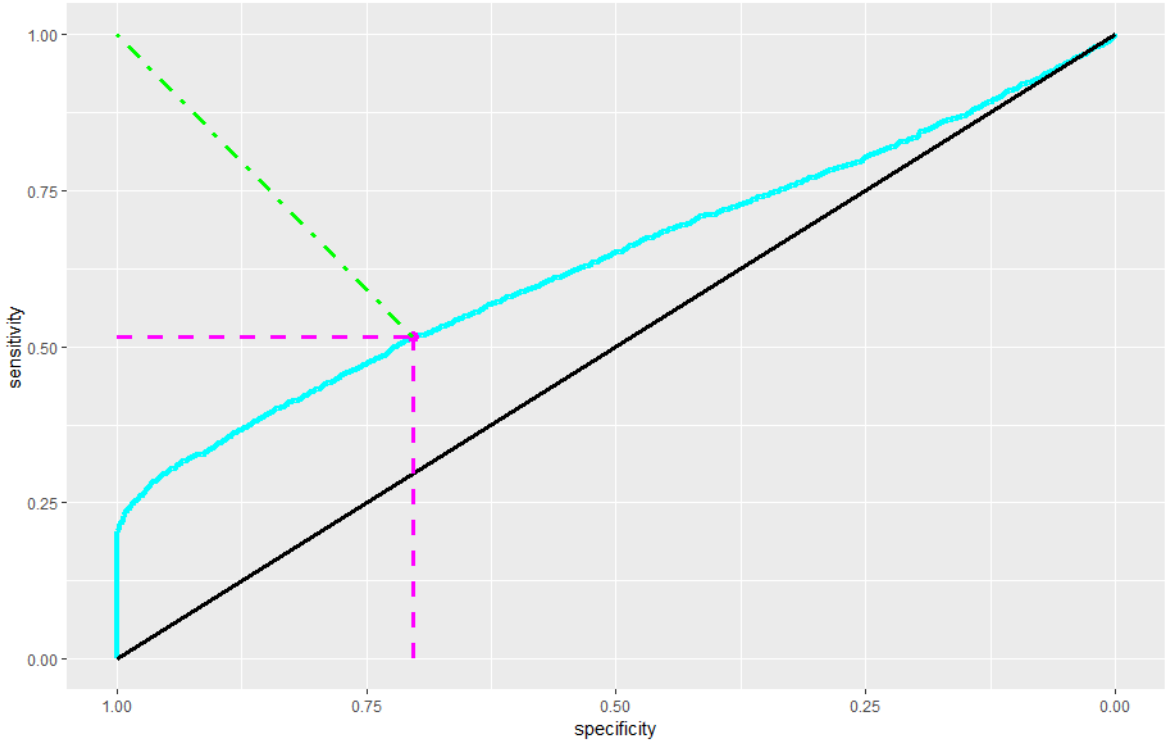

Figure 3.2: *ROC curve from WQS simulations used to select threshold parameter τ. The black line indicates a random classifier. The magenta segments indicate sensitivity and specificity of the spot along the curve which minimizes the length of the green line.*

then selected corresponding to that point along the ROC curve, which maximizes the specificity and sensitivity of the model (Figure 3.2). The optimal value for τ was 0.04661, which yields a sensitivity of 0.5150 and specificity of 0.7035 to identify meaningfully associated components in final WQS model using real data.

Using selection threshold parameter $\tau = 0.04661$, the final WQS model was fit using 1000 bootstrapped samples of the actual chemical biomarker and ratio data.

Table 3.4: Final WQS regression model

	Odds Ratio	**CI Lower**	**CI Upper**	**P Value**
Intercept	0.0450	0.0275	0.0738	<0.0001
WQS Index	2.2188	1.9763	2.4910	<0.0001

The odds of belonging to the MEG group increase by a factor of 2.2188 as the value of the WQS index increases one unit. However, further interpretability of the model is provided by component weights.

Table 3.5: Component weights from final WQS model

Predictor	**Mean Index Weight**
Serum Hydroxycotinine	0.4354
Serum Cotinine	0.2416
NNAL	0.0760
Urinary Cotinine	0.0639
N-acetyl-S-(2-cyanoethyl)-L-cysteine	0.0559
Urinary Hydroxycotinine	0.0237
u3HOF\Total Nicotine Equivalent	0.0233
NNAL\Total Nicotine Equivalent	0.0209
NNAL\Total Nicotine Equivalent	0.0172
2-Hydroxyfluorene\Total Nicotine Equivalent	0.0161
2-Hydroxyfluorene	0.0136
3-Hydroxyfluorene	0.0122

Together, serum hydroxycotinine (shcot) and serum cotinine (scot) accounted for more than 60% of the effect of the mixture. Individually, the other components included in the WQS term each contributed to less than 8% of the mixture's effect (Table 3.5). Although the majority of the mixture's effect is explained by two components, that does not necessarily suggest that the other predictors are not meaningfully associated with the outcome. Important components within the mixture can be identified by weights greater than selection threshold parameter $\tau = 0.04661$.

Five components within the WQS index had weights greater than τ (Figure 3.3), despite the initial impression that only two of the components within the mixture were meaningfully associated with the outcome. The five components identified by the WQS

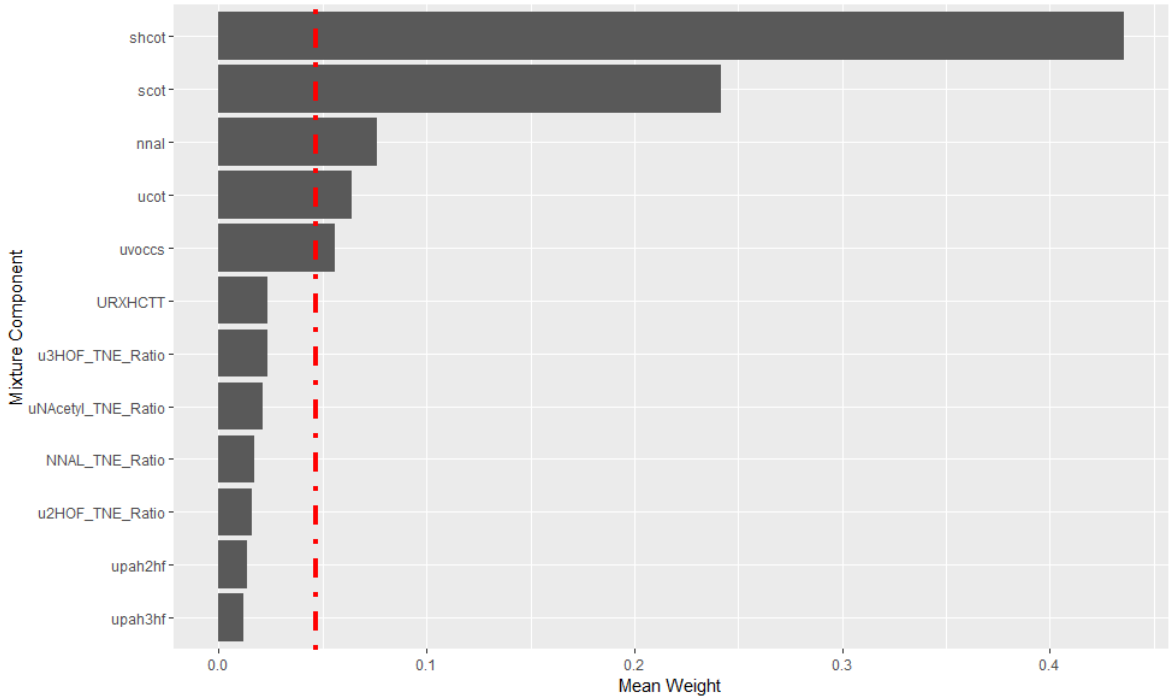

Figure 3.3: *Weight estimates for predictors in final WQS model*

model were also identified by the univariate logistic regression models (Table 3.1). The accuracy of the WQS model was measured using an 80% training and 20% testing split. Selection threshold parameter τ was ignored, as it is only relevant for identifying important mixture components, and has no impact on the overall predictive ability of the model. Samples were weighted inversely according to the frequency of the response to address TSE class imbalance.

Table 3.6: WQS model confusion matrix

		Predicted	
		TEG	MEG
Observed	TEG	104	37
	MEG	18	64

The WQS model was able to predict belonging to TSE group with an accuracy of 75.34%.

However, the model had a greater bias towards misclassifying TEG subjects as MEG than misclassifying MEG subjects as TEG (Table 3.6).

3.4.3 Discussion

Random forest models are useful for situations where accurately distinguishing between different categorical responses is of the utmost importance. The ensemble bootstrapping step and the use of random subsets of predictors for tree fitting disrupt the correlation structure of the data and avoid the pitfalls of modeling with many correlated predictors. However, due to how the random forest captures the relationships between predictors and response, the gains in accuracy come at the cost of model interpretability. In a random forest model, interpretability is limited to measures of relative variable importance. In exposure science, model interpretability is crucial for assessing and quantifying the biological impacts of external exposures. Weighted Quantile Sum regression is another ensemble approach that uses bootstrapping to disrupt the correlation patterns among predictors. By quantiling and then combining variables of interest into a single weighted index, the model estimates the effect of the mixture on the response, while retaining the form and interpretability of a generalized linear model. As with random forest models, a variable's importance within the WQS model is relative to the others considered in the index. However, with many predictors, it becomes increasingly difficult to distinguish which components in the mixture are most important. Predictors in the index that are meaningfully associated with the response can be identified through a comparison with *a priori* selected threshold parameter τ.

Variables with weights greater than τ are considered important. In this work, τ was selected using the ROC optimization procedure described in an earlier section.

The random forest and WQS regression models exhibited similar levels of accuracy in distinguishing between TEG and MEG group membership. However, WQS regression was approximately 3% less accurate. This decrease in accuracy is negligible, especially considering the estimated effect of the mixture provided by WQS regression. Despite weighting the training samples such that each outcome class would be considered equally, the WQS model favored misclassification of TEG as MEG. This bias likely results from loss of information due to quantiling of the data. TEG subjects whose values for important predictors were similar to those of MEG subjects would be assigned to the same quantile, thereby decreasing the model's ability to identify which TSE group those subjects belong to.

The ROC optimization procedure for selection of threshold parameter τ was effective for identifying important predictors within the model. The component weights in the WQS provided a similar impression to the variable importance measures from the random forest. For both models, there were two tiers of variables- a high tier, consisting of two variables that were critical in explaining outcome; and a low tier, consisting of the remaining variables that still provided some information but were not very influential in predicting TSE group (Table 3.2, Table 3.5). Exploratory univariate logistic regression models identified 8 biomarkers that were significantly associated with differences in TSE group membership (Table 3.1). Using the ROC optimization procedure to select threshold parameter τ, the WQS model identified 5 of the 8 biomarkers that were

significant in the univariate models, but considered the effects of all variables of interest simultaneously.

The ROC optimization procedure could also conceivably be extended for use with a random forest model to formally identify which variables are meaningful. Random forest variable importance is relative and by constraining the sum of the importance scores to 1, they become something akin to component weights as in WQS regression. From there, one could implement the same ROC optimization procedure to select a value for threshold parameter τ, where predictors in the random forest model with importance scores (or weights) greater than τ are deemed meaningful. Such an approach could be useful to an investigator who requires the slightly higher predictive accuracy of a random forest but also desires more model interpretability. In summary, this work validates the use of the ROC optimization procedure for selection of threshold parameter τ for non-simulated data. Had the heuristic $\frac{1}{c}$ method been used to determine τ, a reduced number of significantly associated variables would have been identified by the WQS model, and consequently ignored when making conclusions about critical exposures.

References

1. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449,** 804–810 (2007).
2. Martin, A. M., Sun, E. W., Rogers, G. B. & Keating, D. J. The influence of the gut microbiome on host metabolism through the regulation of gut hormone release. *Frontiers in Physiology* **10,** 428 (2019).
3. Wu, H.-J. & Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut microbes* **3,** 4–14 (2012).
4. Chen, W., Liu, F., Ling, Z., Tong, X. & Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS one* **7,** e39743–e39743 (2012).
5. I., A. R. *et al.* Fungal Signature of Moisture Damage in Buildings: Identification by Targeted and Untargeted Approaches with Mycobiome Data. *Applied and Environmental Microbiology* **86,** e01047–20 (2021).
6. Karvonen, A. M. *et al.* Indoor bacterial microbiota and development of asthma by 10.5 years of age. *Journal of Allergy and Clinical Immunology* **144,** 1402–1410 (2019).
7. Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nature reviews. Genetics* **13,** 47–58 (2011).
8. Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics (Oxford, England)* **14,** 244–258 (2013).
9. Shankar, J. *et al.* A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC Bioinformatics* **16,** 1–18 (2015).
10. Pan, A. Y. Statistical analysis of microbiome data: The challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research* **19,** 35–40 (2021).
11. Xia, Y. & Sun, J. Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases* **4,** 138–148 (2017).
12. Cox, J. *et al.* Associations of observed home dampness and mold with the fungal and bacterial dust microbiomes. *Environmental Science: Processes and Impacts* **23,** 491–500 (2021).
13. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology 2011 9:4* **9,** 279–290 (2011).
14. Virgin, H. W. & Todd, J. A. Metagenomics and Personalized Medicine. *Cell* **147,** 44–56 (2011).
15. Martin, T. G. *et al.* Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters* **8,** 1235–1246 (2005).

16. Menni, C. *et al.* Serum metabolites reflecting gut microbiome alpha diversity predict type 2 diabetes. *Gut Microbes* **11,** 1632–1642 (2020).

17. Mendes, L. W. *et al.* Soil-Borne Microbiome: Linking Diversity to Function. *Microbial Ecology* **70,** 255–265 (2015).

18. Prehn-Kristensen, A. *et al.* Reduced microbiome alpha diversity in young patients with ADHD. *PLOS ONE* **13,** e0200728 (2018).

19. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *The Journal of Animal Ecology* **12,** 42 (1943).

20. Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal* **27,** 623–656

21. Simpson, E. H. Measurement of Diversity. *Nature 1949 163:4148* **163,** 688–688 (1949).

22. Sokolowska, M., Frei, R., Lunjani, N., Akdis, C. A. & O'Mahony, L. Microbiome and asthma. *Asthma research and practice* **4,** (2018).

23. Goodrich, J. K. *et al.* Conducting a Microbiome Study. *Cell* **158,** 250 (2014).

24. Whittaker, R. H. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30,** 279–338 (1960).

25. Gilbert, J. A. & Lynch, S. V. Community ecology as a framework for human microbiome research. *Nature medicine* **25,** 884–889 (2019).

26. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27,** 325–349 (1957).

27. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26,** 32–46 (2001).

28. Cope, E. K., Goldberg, A. N., Pletcher, S. D. & Lynch, S. V. Compositionally and functionally distinct sinus microbiota in chronic rhinosinusitis patients have immunological and clinically divergent consequences. *Microbiome* **5,** (2017).

29. Ramakrishnan, V. R. *et al.* Sinus microbiota varies among chronic rhinosinusitis phenotypes and predicts surgical outcome. *Journal of Allergy and Clinical Immunology* **136,** 334–342.e1 (2015).

30. Subedi, S., Neish, D., Bak, S. & Feng, Z. Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69,** 1163–1187 (2020).

31. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174 (2011).

32. Yang, D. & Xu, W. Clustering on Human Microbiome Sequencing Data: A Distance-Based Unsupervised Learning Model. *Microorganisms 2020, Vol. 8, Page 1612* **8,** 1612 (2020).

33. Clooney, A. G. *et al.* A comparison of the gut microbiome between long-term users and non-users of proton pump inhibitors. *Alimentary Pharmacology & Therapeutics* **43,** 974–984 (2016).

34. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58,** 236–244 (1963).

35. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 1–18 (2017).

36. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* **14,** 1–13 (2013).

37. Zhang, X. *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* **18,** 4 (2017).

38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).

39. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1996).

40. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67,** 301–320 (2005).

41. Carrico, C., Gennings, C., Wheeler, D. C. & Factor-Litvak, P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. *Journal of Agricultural, Biological, and Environmental Statistics* **20,** 100–120 (2015).

42. Cox, J. *et al.* Residential bacteria and fungi identified by high-throughput sequencing and childhood respiratory health. *Environmental Research* **204,** 112377 (2022).

43. Tanner, E. M., Bornehag, C. G. & Gennings, C. Repeated holdout validation for weighted quantile sum regression. *MethodsX* **6,** 2855–2860 (2019).

44. Wheeler, D. C. *et al.* Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. *International Journal of Environmental Research and Public Health* **18,** 1–20 (2021).

45. Colicino, E., Pedretti, N. F., Busgang, S. A. & Gennings, C. Per- and polyfluoroalkyl substances and bone mineral density: Results from the Bayesian weighted quantile sum regression. *Environmental Epidemiology* **4,** (2020).

46. Chen, Y.-H., Ferguson, K. K., Meeker, J. D., McElrath, T. F. & Mukherjee, B. Statistical methods for modeling repeated measures of maternal environmental exposure biomarkers during pregnancy in association with preterm birth. *Environmental Health* **14,** 9 (2015).

47. Batzella, E. *et al.* Perfluoroalkyl substance mixtures and cardio-metabolic outcomes in highly exposed male workers in the Veneto Region: A mixture-based approach. *Environmental Research* **212,** 113225 (2022).

48. Zhang, Y. *et al.* Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: Comparison of three statistical models. *Environment International* **123,** 325–336 (2019).

49. Luo, K. *et al.* Associations between organophosphate esters and sex hormones among 6–19-year old children and adolescents in NHANES 2013–2014. *Environment International* **136,** 105461 (2020).

50. Eguchi, A. *et al.* Association between Total and Individual PCB Congener Levels in Maternal Serum and Birth Weight of Newborns: Results from the Chiba Study of Mother and Child Health Using Weighted Quantile Sum Regression. *International Journal of Environmental Research and Public Health* **19,** (2022).

51. Daniel, S. *et al.* Perinatal phthalates exposure decreases fine-motor functions in 11-year-old girls: Results from weighted Quantile sum regression. *Environment International* **136,** (2020).

52. Eggers, S., Bixby, M., Renzetti, S., Curtin, P. & Gennings, C. Human Microbiome Mixture Analysis using Weighted Quantile Sum Regression. (2022). doi:10.1101/2022.07.11.22277512

53. Faust, K. & Raes, J. Microbial interactions: From networks to models. *Nature Reviews Microbiology 2012 10:8* **10,** 538–550 (2012).

54. Xiao, J. *et al.* Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model. *Frontiers in Microbiology* **9,** (2018).

55. Higham, N. J. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications* **103,** 103–118 (1988).

56. Revers, A., Zhang, X. & Zwinderman, A. H. A Bayesian Negative Binomial Hierarchical Model for Identifying Diet–Gut Microbiome Associations. *Frontiers in Microbiology* **12,** 711861 (2021).

57. Ye, P., Qiao, X., Tang, W., Wang, C. & He, H. Testing latent class of subjects with structural zeros in negative binomial models with applications to gut microbiome data. *Statistical Methods in Medical Research* 09622802221115881 (2022). doi:10.1177/09622802221115881

58. Hargarten, P. M. & Wheeler, D. C. Accounting for the uncertainty due to chemicals below the detection limit in mixture analysis. *Environmental Research* **186,** 109466 (2020).

59. Malin, A. J. *et al.* Fluoride exposure and sleep patterns among older adolescents in the United States: A cross-sectional study of NHANES 2015–2016. *Environmental Health* **18,** 106 (2019).

60. Slattery, M. L., Pellatt, D. F., Mullany, L. E. & Wolff, R. K. Differential Gene Expression in Colon Tissue Associated With Diet, Lifestyle, and Related Oxidative Stress. *PLoS ONE* **10,** e0134406 (2015).

61. United States Surgeon General. *The Health Consequences of Smoking – 50 Years of progress: A Report of the Surgeon General: (510072014-001).* (American Psychological Association, 2014). doi:10.1037/e510072014-001

62. Gillis, C. R., Hole, D. J., Hawthorne, V. M. & Boyle, P. The effect of environmental tobacco smoke in two urban communities in the west of Scotland. *European Journal of Respiratory Diseases. Supplement* **133,** 121–126 (1984).

63. Garland, C., Barrett-Connor, E., Suarez, L., Criqui, M. H. & Wingard, D. L. Effects of passive smoking on ischemic heart disease mortality of nonsmokers. A prospective study. *American Journal of Epidemiology* **121,** 645–650 (1985).

64. Judson Wells, A. An estimate of adult mortality in the United States from passive smoking. *Environment International* **14,** 249–265 (1988).

65. Glantz, S. A. & Parmley, W. W. Passive smoking and heart disease. Epidemiology, physiology, and biochemistry. *Circulation* **83,** 1–12 (1991).

66. Jacob, P. *et al.* Thirdhand Smoke: New Evidence, Challenges, and Future Directions. *Chemical Research in Toxicology* **30,** 270–294 (2017).

67. Merianos, A. L., Mahabee-Gittens, E. M. & Choi, K. Tobacco Smoke Exposure and Inadequate Sleep among U.S. School-Aged Children. *Sleep medicine* **86,** 99–105 (2021).

68. Burton, A. Does the Smoke Ever Really Clear? Thirdhand Smoke Exposure Raises New Concerns. *Environmental Health Perspectives* **119,** A70–A74 (2011).

69. Winickoff, J. P. *et al.* Beliefs About the Health Effects of 'Thirdhand' Smoke and Home Smoking Bans. *Pediatrics* **123,** e74–e79 (2009).

70. Matt, G. E. *et al.* Thirdhand Tobacco Smoke: Emerging Evidence and Arguments for a Multidisciplinary Research Agenda. *Environmental Health Perspectives* **119,** 1218–1226 (2011).

71. Schick, S. F. *et al.* Thirdhand cigarette smoke in an experimental chamber: Evidence of surface deposition of nicotine, nitrosamines and polycyclic aromatic hydrocarbons and de novo formation of NNK. *Tobacco Control* **23,** 152–159 (2014).

72. Benowitz, N. L., Hukkanen, J. & Jacob, P. Nicotine Chemistry, Metabolism, Kinetics and Biomarkers. *Handbook of experimental pharmacology* 29–60 (2009). doi:10.1007/978-3-540-69248-5_2

73. Benowitz, N. *et al.* Urine cotinine underestimates exposure to the tobacco-derived lung carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) in passive compared to active smokers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **19,** 2795–2800 (2010).

74. St. Helen, G. *et al.* Exposure and Kinetics of Polycyclic Aromatic Hydrocarbons (PAHs) in Cigarette Smokers. *Chemical Research in Toxicology* **25,** 952–964 (2012).

75. St.Helen, G. *et al.* Intake of Toxic and Carcinogenic Volatile Organic Compounds from Secondhand Smoke in Motor Vehicles. *Cancer Epidemiology, Biomarkers & Prevention* **23,** 2774–2782 (2014).

76. Merianos, A. L. *et al.* Distinguishing Perceived Exposure to Secondhand and Thirdhand Tobacco Smoke among U.S. Children based on Questionnaire and Biomarker Profiles using Machine Learning: NHANES 2013-2016. (2022).

77. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

78. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

79. Breiman, L. *Classification And Regression Trees.* (Routledge, 2017). doi:10.1201/9781315139470

80. Breiman, L. Bagging predictors. *Machine Learning* **24,** 123–140 (1996).

81. Czarnota, J., Gennings, C. & Wheeler, D. C. Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. *Cancer Informatics* **14s2,** CIN.S17295 (2015).