# Exploring the Interplay Between Fibre Intake, Exercise, and Gut Microbiota in Modulating Cardiovascular Health in a Westernization Context

Rui Xiang Yu, Houria Afshar Moghaddam, Brooke Macleod, Quinlan Torstensen

## Table of contents

## 1 Introduction

This report contains the methods and results generated throughout the making of this project. Our research question is:

*In a westernization context, how do dietary fibre intake and exercise affect cardiometabolic health, and does this associate with unique microbiome profiles?*

Our aims are:

1. Investigate the optimal predictors for cardiovascular health, such as dietary fibre and exercise.
2. Compare diversity metrics of groups with adequate and inadequate dietary fibre intake, and high and low exercise levels in relation to cardiometabolic health.
3. Identify taxa that are differentially abundant and co-occur among the different fibre intake and exercise groups, as well as cardiovascular condition.

## 2 Methods

### 2.1 Data processing

The dataset was sourced from de la Cuesta-Zuluaga et al. 2018's paper, whose raw DNA FASTQ files can be found at the SRA-NCBI under BioProject PRJNA417579 (Cuesta-Zuluaga et al. (2018)). The dataset comprises the sequences corresponding to the V4 hypervariable region of the 16S rRNA gene, collected in 2014 (Cuesta-Zuluaga et al. (2018)). The primers used were F515 and R806 and sequencing was done with Illumina MiSeq (Cuesta-Zuluaga et al. (2018)). There are 441 samples of men and women, across a wide range of ages and body mass indices, that lived in different Colombian cities (Bogotá, Medellín, Cali, Barranquilla, Bucaramanga). None of the participants were underweight (Body Mass Index <18.5 kg/m$^2$), pregnant, consumed antibiotics or antiparasitics three months before sample collection, with cancer, with gastrointestinal diseases, or with neurodegenerative diseases (Cuesta-Zuluaga et al. (2018)).

The accompanying metadata for each patient consists of multiple demographic parameters (age, biological sex, city of residence), anthropometric measures (body mass index, body fat percentage, waist circumference), lipid profiles in blood (HDL, LDL, adiponectin), macronutrient consumption, blood pressure, glucose metabolism profiles, and stool consistencies. A full list of available metadata with each variable's range or factors can be found in Table **??**.

Table 1: Table of the multiple metadata information available for each patient. For categorical variables, it contains all factors. For numerical variables, it contains the numerical range of available values.

| Variable | Range_or_Categories |
|---|---|
| adiponectin | 0 – 28.21 |
| age_years | 18 – 62 |
| age_range | 18_40, 41_62 |
| BMI | 18.6 – 47.4 |
| BMI_class | Lean, Obese, Overweight |
| Body_Fat_Percentage | 18.7 – 48.7 |
| Calorie_intake | 634 – 4034 |
| Cardiometabolic_status | Abnormal, Healthy |
| city | Barranquilla, Bogota, Bucaramanga, Cali, Medellin |
| country | Colombia |
| diastolic_bp | 50 – 126 |
| fiber | 7 – 44 |
| glucose | 64 – 335 |
| Hemoglobin_a1c | 4.6 – 10.77 |
| CRP | 0.12 – 44.3 |
| insulin | 1.95 – 57.07 |
| latitude | 3.42 – 10.96 |
| Total_Cholesterol | 67 – 302 |
| HDL | 11 – 134 |
| LDL | 30 – 219 |
| VLDL | 6 – 218 |
| Triglycerides | 28 – 1090 |

Table 1: Table of the multiple metadata information available for each patient. For categorical variables, it contains all factors. For numerical variables, it contains the numerical range of available values.

| Variable | Range_or_Categories |
|---|---|
| medication | No, Yes |
| per_carbohydrates | 45.89 – 64.87 |
| per_total_protein | 11.95 – 21.11 |
| per_total_fat | 21.51 – 36.83 |
| per_animal_protein | 39.41 – 74.48 |
| per_monoinsaturated_fat | 6.85 – 12.55 |
| per_polyunsaturated_fat | 3.48 – 7.86 |
| per_saturated_fat | 7.46 – 16.06 |
| sex | female, male |
| smoker | No, Yes |
| stool_consistency | Diarrheic, Hard, Normal, soft |
| systolic_bp | 76 – 204 |
| MET_mins_per_week | 0 – 45204.6 |
| waist_circumference | 65.2 – 131.3 |

The data was processed with QIIME2 version 2025.4.0 (Bolyen et al. (2019)). The script to import the data and demultiplex the sequences can be found in `bin/01-qiime2_data_processing.sh`. The maximum read length was 251 nucleotides (nts). Based on a random subsample of 10,000 reads, 98% of reads have a read length of 251 nts, while 2% have a read length of 250 nts. The maximum number of reads in a sample was 117,562, whereas the minimum number was 4,305. The mean number of reads was 40,657.89.

The script for filtering and downstream QIIME2 processes are in `bin/02-qiime2_data_filtering.sh`. The denoising tool chosen was DADA2 (Callahan et al. (n.d.)). The Phred quality score threshold chosen was 30. The base pair at position 251's median quality was 29. A truncation length of 250 was chosen. After denoising, all 441 samples were still retained. The maximum number of reads in a sample changed to 106,116, whereas the minimum number changed to 3,665.

With QIIME2's taxonomic classifier, amplicon sequence variants (ASVs) were classified with a pre-trained Naïve Bayes model built on the SILVA version 138 99% OTUs, trained for the primer pair 515F/806R, which targets the V4 region of the 16S rRNA gene. A phylogenetic tree was subsequently generated by aligning multiple sequences with MAFFT, masking hypervariable regions, constructing an approximately maximum-likelihood tree with FastTree2, and applying midpoint rooting to obtain a rooted phylogeny for downstream analyses.

Afterwards, sequences corresponding to mitochondria or chloroplasts, as well as non-bacterial sequences were filtered out. Based on the subsequently generated rarefaction curve (Figure **??**), ASVs saturate at around ~10,000 reads for most samples. A cutoff of 24,406 reads was chosen as this retained 79.49% of the original number of features (8,705 out of 10,951) and 333 samples. Thus, 108 samples were discarded. No batch correction was applied to the data as de la Cuesta-Zuluaga et al. found no significant difference across runs and an internal control was also present (Cuesta-Zuluaga et al. (2018)).

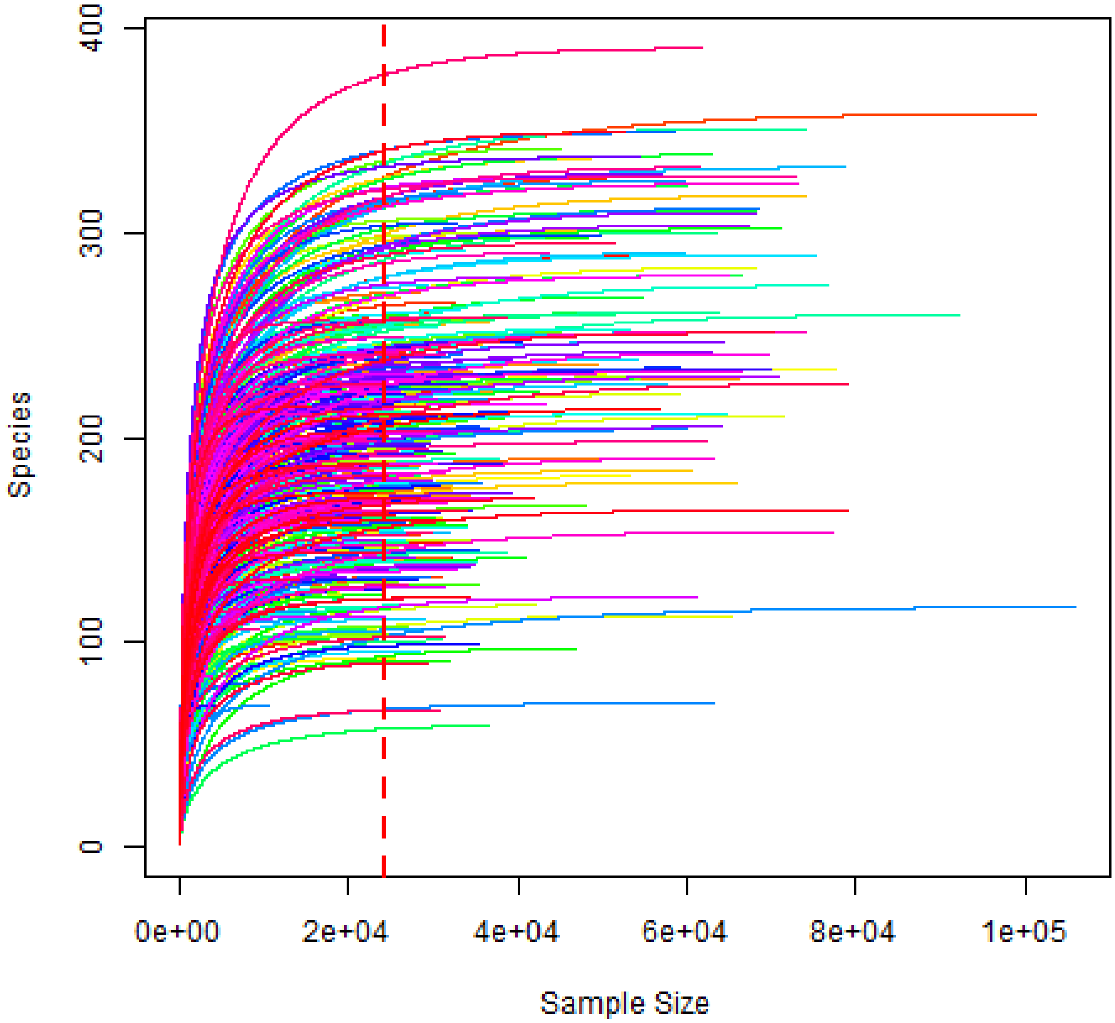

Figure 1: Alpha rarefaction plot where each line represents a sample. The dotted red line represents a sequencing depth of 24,406 reads.

## 2.2 Exploratory Data Analyses and Preliminary Analyses

Prior to rarefying, using all 441 samples, a redundancy analysis (RDA) was performed. The OTU table was aggregated to the phylum level and a Hellinger transformation was applied. The transformed and aggregated OTU table was used as the response variable and the entire metadata was used as the predictors. An ANOVA Permutation test was run on the model to determine significance of the covariates. Results can be found in Table **??**. We can see that fibre appears to be significantly associated to the OTU table.

Table 2: Redundancy Analysis results with hellinger-transformed phylum counts as a response matrix. Contains for each variable its degrees of freedom (Df), variance explained by the variable (Variance), its F-statistic (F), and the p-value from ANOVA-like permutation tests (Pr(>F)).

| variable | Df | Variance | F | Pr(>F) |
|---|---|---|---|---|
| adiponectin | 1 | 0.0019950 | 5.0756737 | 0.001 |
| age_years | 1 | 0.0004970 | 1.2643629 | 0.265 |
| age_range | 1 | 0.0004762 | 1.2114433 | 0.302 |
| BMI | 1 | 0.0014115 | 3.5910102 | 0.007 |
| BMI_class | 2 | 0.0015760 | 2.0048381 | 0.035 |
| Body_Fat_Percentage | 1 | 0.0022959 | 5.8410633 | 0.001 |
| Calorie_intake | 1 | 0.0008806 | 2.2403008 | 0.061 |
| Cardiometabolic_status | 1 | 0.0004781 | 1.2162441 | 0.273 |
| city | 4 | 0.0168991 | 10.7484504 | 0.001 |
| diastolic_bp | 1 | 0.0010521 | 2.6767860 | 0.026 |
| fiber | 1 | 0.0014159 | 3.6022940 | 0.005 |
| glucose | 1 | 0.0003692 | 0.9394260 | 0.442 |
| Hemoglobin_a1c | 1 | 0.0002473 | 0.6290934 | 0.656 |
| CRP | 1 | 0.0000917 | 0.2332970 | 0.959 |
| insulin | 1 | 0.0002153 | 0.5478133 | 0.727 |
| Total_Cholesterol | 1 | 0.0003405 | 0.8663439 | 0.465 |
| HDL | 1 | 0.0001065 | 0.2709362 | 0.940 |
| LDL | 1 | 0.0003832 | 0.9748675 | 0.404 |
| VLDL | 1 | 0.0006079 | 1.5465834 | 0.183 |
| Triglycerides | 1 | 0.0001867 | 0.4751187 | 0.763 |
| medication | 1 | 0.0005710 | 1.4525957 | 0.185 |
| per_carbohydrates | 1 | 0.0007468 | 1.8999085 | 0.097 |
| per_total_protein | 1 | 0.0002285 | 0.5813990 | 0.702 |
| per_total_fat | 1 | 0.0001462 | 0.3720172 | 0.885 |
| per_animal_protein | 1 | 0.0007834 | 1.9930137 | 0.087 |
| per_monoinsaturated_fat | 1 | 0.0001688 | 0.4293863 | 0.829 |
| per_polyunsaturated_fat | 1 | 0.0002719 | 0.6918226 | 0.607 |

Table 2: Redundancy Analysis results with hellinger-transformed phylum counts as a response matrix. Contains for each variable its degrees of freedom (Df), variance explained by the variable (Variance), its F-statistic (F), and the p-value from ANOVA-like permutation tests (Pr(>F)).

| variable | Df | Variance | F | Pr(>F) |
|---|---|---|---|---|
| per_saturated_fat | 1 | 0.0004138 | 1.0526674 | 0.360 |
| sex | 1 | 0.0005928 | 1.5082614 | 0.164 |
| smoker | 1 | 0.0005700 | 1.4501757 | 0.227 |
| stool_consistency | 3 | 0.0035411 | 3.0030428 | 0.001 |
| systolic_bp | 1 | 0.0006574 | 1.6726463 | 0.133 |
| MET_mins_per_week | 1 | 0.0001677 | 0.4266723 | 0.807 |
| waist_circumference | 1 | 0.0007175 | 1.8254798 | 0.096 |
| Residual | 399 | 0.1568304 | NA | NA |

Furthermore, a logistic regression model was fit onto the metadata, where cardiovascular status was the response variable. Model results are in Table ??. We can see that MET units are significantly associated with cardiovascular status.

Table 3: Logistic regression results with cardiometabolic status as the response variable. Contains for each variable the log-odds coefficient (Estimate), the standard error of the coefficient (Std. Error), its z-statistic (z), and its p-value (Pr(>|z|)).

| Estimate | Std. Error | z value | Pr(>\|z\|) | Variable |
|---|---|---|---|---|
| -9.9122263 | 77.4539894 | -0.1279757 | 0.8981682 | (Intercept) |
| -0.0602070 | 0.0648303 | -0.9286861 | 0.3530518 | adiponectin |
| 0.0656814 | 0.0441824 | 1.4865949 | 0.1371218 | age_years |
| -0.3639803 | 0.8645281 | -0.4210161 | 0.6737433 | age_range41_62 |
| 0.0936441 | 0.1754121 | 0.5338521 | 0.5934439 | BMI |
| -0.4970514 | 1.4419124 | -0.3447168 | 0.7303073 | BMI_classObese |
| -0.4844568 | 0.8015177 | -0.6044244 | 0.5455615 | BMI_classOverweight |
| -0.0052189 | 0.0811479 | -0.0643130 | 0.9487210 | Body_Fat_Percentage |
| -0.0003404 | 0.0007326 | -0.4646448 | 0.6421858 | Calorie_intake |
| -1.9390412 | 1.3234127 | -1.4651825 | 0.1428711 | cityBogota |
| -0.3654200 | 1.2887662 | -0.2835425 | 0.7767610 | cityBucaramanga |
| 0.2606411 | 1.0634218 | 0.2450966 | 0.8063816 | cityCali |
| -0.1763447 | 1.0270975 | -0.1716922 | 0.8636795 | cityMedellin |
| 0.0339491 | 0.0339727 | 0.9993058 | 0.3176466 | diastolic_bp |
| -0.0285924 | 0.0631854 | -0.4525165 | 0.6508969 | fiber |
| 0.0955358 | 0.0324401 | 2.9449870 | 0.0032297 | glucose |
| -0.2352485 | 0.5294736 | -0.4443064 | 0.6568211 | Hemoglobin_a1c |

Table 3: Logistic regression results with cardiometabolic status as the response variable. Contains for each variable the log-odds coefficient (Estimate), the standard error of the coefficient (Std. Error), its z-statistic (z), and its p-value (Pr(>|z|)).

| Estimate | Std. Error | z value | Pr(>\|z\|) | Variable |
|---:|---:|---:|---:|:---|
| 0.1768625 | 0.0453103 | 3.9033602 | 0.0000949 | CRP |
| 0.2162195 | 0.0508586 | 4.2513881 | 0.0000212 | insulin |
| -0.0525445 | 0.0539141 | -0.9745964 | 0.3297605 | Total_Cholesterol |
| -0.1135320 | 0.0632763 | -1.7942248 | 0.0727773 | HDL |
| 0.0593667 | 0.0535188 | 1.1092685 | 0.2673144 | LDL |
| -0.1594781 | 0.1546671 | -1.0311058 | 0.3024912 | VLDL |
| 0.0587170 | 0.0305979 | 1.9189893 | 0.0549857 | Triglycerides |
| 0.8986201 | 0.4667142 | 1.9254185 | 0.0541770 | medicationYes |
| -0.0338268 | 0.7495101 | -0.0451318 | 0.9640022 | per_carbohydrates |
| -0.1803468 | 0.7899163 | -0.2283113 | 0.8194043 | per_total_protein |
| 0.1254492 | 0.8568957 | 0.1463997 | 0.8836059 | per_total_fat |
| -0.0357073 | 0.0661648 | -0.5396719 | 0.5894233 | per_animal_protein |
| -0.4547436 | 0.3918152 | -1.1606076 | 0.2458015 | per_monoinsaturated_fat |
| -0.4089035 | 0.3731494 | -1.0958171 | 0.2731588 | per_polyunsaturated_fat |
| 0.2903317 | 0.2742740 | 1.0585463 | 0.2898065 | per_saturated_fat |
| -1.3558918 | 0.8241822 | -1.6451361 | 0.0999418 | sexmale |
| 0.4601585 | 0.6921807 | 0.6647953 | 0.5061814 | smokerYes |
| -1.4746828 | 1.2264204 | -1.2024285 | 0.2291975 | stool_consistencyHard |
| 0.3962243 | 1.0608879 | 0.3734837 | 0.7087885 | stool_consistencyNormal |
| -0.8717309 | 1.5102800 | -0.5771982 | 0.5638056 | stool_consistencysoft |
| 0.0479254 | 0.0220354 | 2.1749271 | 0.0296356 | systolic_bp |
| -0.0000771 | 0.0000433 | -1.7807330 | 0.0749561 | MET_mins_per_week |
| -0.0153338 | 0.0485372 | -0.3159191 | 0.7520639 | waist_circumference |

The distributions of the variables of interest did not change considerably after rarefying. The distribution metabolic equivalent of task (MET) was heavily right-skewed (Figure **??**). The distribution of daily fibre intake was bell-shaped (Figure **??**). The number of female and male participants, as well as participants with a healthy or abnormal cardiometabolic status, are balanced throughout (Figure **??**). The number of participants with healthy cardiometabolic status went from 269 to 204, while those with an abnormal status went from 172 to 129. The number of female participants went from 229 to 178, while male participants went from 212 to 155. The distribution of age in years was also explored and it was roughly bell-shaped, almost uniform-like (Figure **??**). The distributions of these variables prior to filtering (all 441 initial samples portrayed) can be found in Figure **??**.

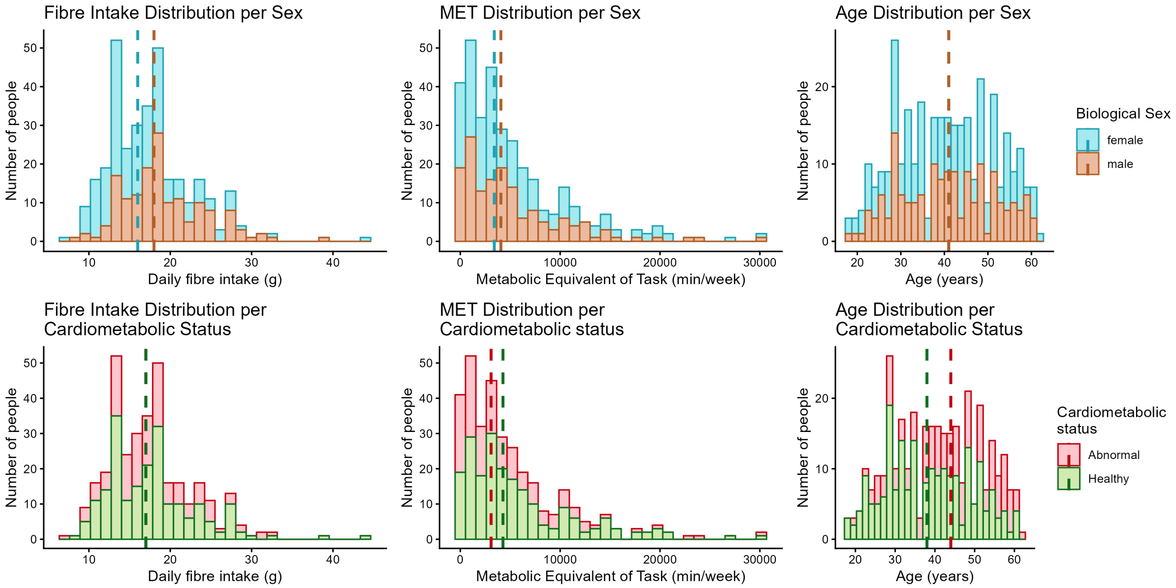

Figure 2: Histogram distributions of the 333 participants' fibre intake (left), metabolic equivalent of tasks (middle), and age (right). Distributions are colored by biological sex (top) and cardiometabolic status (bottom). Dotted vertical lines represent the median for each sub-group.

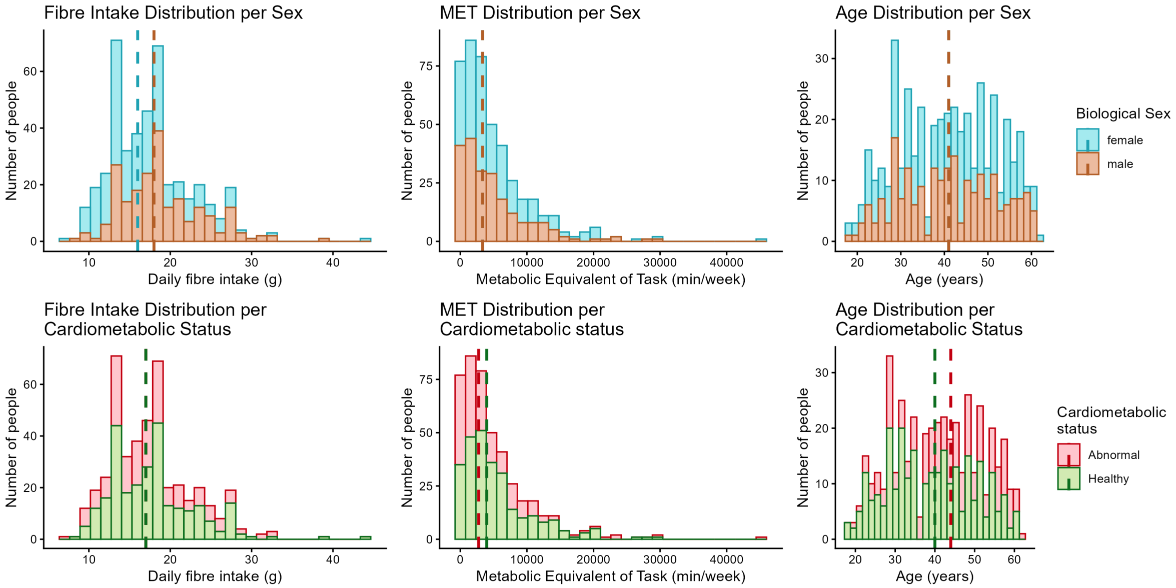

Figure 3: Histogram distributions of the 441 participants' fibre intake (left), metabolic equivalent of tasks (middle), and age (right). Distributions are colored by biological sex (top) and cardiometabolic status (bottom). Dotted vertical lines represent the median for each sub-group.

As mentioned previously, participants will be divided into two fibre intake and exercise level groups, while considering cardiovascular status as well. High fibre intake is considered as 20 grams of fibre or above. High exercise level is considered as more than 1000 metabolic equivalent of task minutes per week. These cut-offs were considered based in the distributions observed as well as guidelines on the adequate number of MET units and fibre intake. The sample sizes of each sub-group are reported in Table **??**. The smallest groups, representing participants with high fibre intake and low exercise, included 8 with a healthy and 9 with an abnormal cardiovascular status. Given these small sample sizes, all 333 samples were retained for downstream analyses, with no further filtering performed. Based on the observed distributions and lack of clear outliers, this approach was considered valid.

Table 4: Distribution of Participants by Cardiometabolic Health Status, Fibre Intake, and Exercise Level. High fibre intake is 20 grams or more of daily fibre. High exercise level is more than 1000 MET minutes per week.

| Cardiometabolic_status | fibre_group | exercise_group | n |
|---|---|---|---|
| Abnormal | high | high | 28 |
| Abnormal | high | low | 9 |
| Abnormal | low | high | 66 |
| Abnormal | low | low | 26 |

Table 4: Distribution of Participants by Cardiometabolic Health Status, Fibre Intake, and Exercise Level. High fibre intake is 20 grams or more of daily fibre. High exercise level is more than 1000 MET minutes per week.

| Cardiometabolic_status | fibre_group | exercise_group | n |
|---|---|---|---|
| Healthy | high | high | 51 |
| Healthy | high | low | 8 |
| Healthy | low | high | 120 |
| Healthy | low | low | 25 |

### 2.3 Aim 1: Multiple Linear Regression

## References

Bolyen, Evan, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57.

Callahan, BJ, PJ McMurdie, MJ Rosen, AW Han, AJA Johnson, and SP Holmes Dada. n.d. "High-Resolution Sample Inference from Illumina Amplicon Data., 2016, 13." *DOI: Https://Doi. Org/10.1038/Nmeth* 3869: 581–83.

Cuesta-Zuluaga, Jacobo de la, Vanessa Corrales-Agudelo, Eliana P Velásquez-Mejía, Jenny A Carmona, José M Abad, and Juan S Escobar. 2018. "Gut Microbiota Is Associated with Obesity and Cardiometabolic Disease in a Population in the Midst of Westernization." *Scientific Reports* 8 (1): 11356.