Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8033806, 16 pages
https://doi.org/10.1155/2022/8033806

Research Article

A Copula Type-Model for Examining the Role of Microbiome as a Potential Tool in Diagnosis

Enrique Calderín–Ojeda,[1] **Guillermo López–Campos**,[2] **and Emilio Gómez–Déniz**[3]

[1]*Centre for Actuarial Studies, Department of Economics, University of Melbourne, Melbourne, Australia*
[2]*Wellcome-Wolfson Institute for Experimental Medicine, Medical School, Queen's University Belfast, Belfast, UK*
[3]*Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain*

Correspondence should be addressed to Enrique Calderín–Ojeda; ecalderin@unimelb.edu.au

Received 24 January 2022; Revised 14 March 2022; Accepted 2 April 2022; Published 6 June 2022

Academic Editor: Xiaofeng Li

Continuous advancements in biotechnology are generating new knowledge and data sources that might be of interest for the insurance industry. A paradigmatic example of these advancements is genetic information which can reliably notify about future appearance of certain diseases making it an element of great interest for insurers. However, this information is considered by regulators in the highest confidentiality level and protected from disclosure. Recent investigations have shown that the microbiome can be correlated with several health conditions. In this paper, we examine the potential use of microbiome information as a potential tool for cardiovascular diagnosis. By using a recent dataset, we analyze the relation of some variables associated to coronary illnesses and several components of the microbiome in the organism by using a new copula-based multivariate regression model for compositional data in the predictor. Our findings show that the coabundance group associated to *Ruminococcaceae-Bifidobacteriaceae* has a negative impact on the age for nonsedentary individuals. However, one should be cautious with this conclusion since environmental conditions also influence the baseline microbiome.

1. Introduction

In recent years, the advances in biomedical sciences and biotechnology have enabled an unprecedented leap forward in the amounts and variety of data and information available for research and other purposes. This has translated in new applications and the development of new approaches such as a personalised or precision medicine where the aim is to use these new data and information sources (mostly related with genetic/genomic information) for diagnostic and therapeutic purposes and tailoring them to individuals or groups of patients (see Ginsburg and Phillips [1]). However, genomic information is considered in the highest level of confidentiality and protected from disclosure. For these reasons, the use of genomic data sparked a debate around the ethics and limits associated with the use of this knowledge and information in different sectors how it could be eventually regulated. In an attempt to overcome this potential limitation, in this paper, we propose to explore new avenues and information sources and the use of microbiome as a potential tool for disease diagnosis.

The microbiome is defined as the set of microorganisms that live inside or on the organism and its analysis has attracted a great interest in the biomedical domain, particularly since the development of new technologies that have facilitated and reduced the costs of accessing this information. Microbiome is an extremely dynamic element, and it changes with time and environmental conditions and other external factors, such as diet, geographical location, or physical activity and even with interaction between microbes and microbes and the host. The advances in biotechnology allow researchers to measure dynamic behaviors of the microbiota at a large scale (see [2]). Recent studies have shown that differences in the microbiome composition are correlated with an increasing number of conditions ranging from cardiovascular diseases, autoimmune diseases, metabolic diseases, or neurological disorders and mental health aspects [3–7]. Another interesting reference linking a well-established cohort in the biomedical domain (the Framingham Study) with changes in the microbiome for multiple relevant health parameters such as cardiovascular risk, metabolic syndrome, and diabetes is Walker et al. [8]. BMI and physical activity have been also studied in the context of the microbiome finding relationships with different microbiome compositions [9–12]. Quite often, these relationships have been studied under the umbrella of either different ages or other health conditions. For a recent review highlighting that physical activity has an impact in gut microbiota and that physical exercise could be used to control obesity and health (see [13]). Both BMI and physical activity are important factors considered in insurance underwriting. For example, a higher BMI was predominantly related to blood pressure and lipids, which is consistent with results found in the literature (see [14] or [15]). There also exists an association between obesity and higher BMI with all-cause mortality (see [16]). Besides, BMI is connected to increased cancer risk as was recently described by Bhaskaran et al. [17] in a recent paper. On the other hand, microbiome data present a singular challenge due to its inherently high-dimensional and sparse structure. To handle the high dimensionality and compositional nature of the data, Wang et al. [18] proposed a sparse microbial causal mediation model specifically; also, Zhang et al. [19] used an isometric log-ratio transformation of the relative abundances as the mediator variables between treatment and outcome. A statistical approach that enables the inclusion of all daily activity behaviors, based on the principles of compositional data analysis was described by Dumuid et al. [20].

In our cross-sectional analysis, using a sample of eligible individuals with unique microorganisms across the Indian microbiome population, due to the large number of operational taxonomic units available in the gut microbiome across the sample, a clustering analysis to reduce the dimensionality of the dataset was initially carried out. Then, the resulting proportions of each of the groups of bacterial coabundance are combined in compositional data predictors that will be used to jointly explain the relationship of age, BMI, and level of physical activity by using a new bivariate regression model for compositional data in the predictor. In this paper, the margins are a beta regression and mixture of logistic regression models. As the age in years of the individuals is restricted to the interval 18–65, a beta regression model indexed by mean and dispersion parameters is considered. This regression family is useful in situations where the dependent variable is continuous and restricted to a bounded interval. On the other hand, regardless of the gender and physical activity level, the empirical distribution of BMI in humans is bimodal. Therefore, choosing suitable parametric models that can capture this feature is crucial; for that reason, a mixture of logistic regression model has been chosen due to its flexibility and simplicity. These margins are linked via a t-copula. This is an elliptical copula that is particularly well suited for this purpose as they not only allow for separate modeling of the univariate marginal distributions from the dependency structure but also for covariate adjustment in the margins and uncertainty quantification of their dependence estimates. In addition, they can specify different levels of correlation between the marginals. The compositional data included in the linear predictor are rewritten as logarithms of ratios. Then, we perform estimation via inference for margins method to explain the age, BMI, and level of physical activity using as margins a beta regression and mixture of regression models. Although copula models have been widely applied to model the joint distributions with mixed margins, copula models with the margins proposed in this work with compositional data in the predictor have not been extensively studied in the literature.

The rest of the paper is structured as follows: in Section 2, an examination of a human microbiome dataset is carried out. Here, a cluster analysis to classify the proportion of the most significant bacterial coabundance groups in the sample is completed. Furthermore, an approach to deal with the implementation of microbiome data as compositional data in the predictor is presented. The relationship of age, body mass index, and physical activity level with microbiome is analyzed in Section 3. Here, we firstly consider the marginal relation of age given a level of physical activity with microbiome through a beta regression model. Next, we examine the connection of BMI with the microbiome given the level of physical activeness by using a mixture of logistics regression model. Later, the joint relationship of these variables is examined by using a t-copula. Finally, discussion and extensions conclude the paper.

2. Analysis of a Human Microbiome Dataset

In our analyses we use a dataset available in Dubey et al.'s [21] *LogMPIE* study. This dataset is freely accessible, and it may be downloaded from the *European Nucleotide Archive* (ENA) portal of the *European Bioinformatics Institute* (https://www.ebi.ac.uk/ena/data/view/PRJEB25642). In this study, as it was portrayed in the original description of the dataset, they identify and map the Indian gut microbiome. It was carried out in fourteen geographical locations. Individuals were uniformly selected across geographical regions and some variables associated with changes in the structure of microbiome such as BMI, age in years of the individual, restricted to the interval 18–65 and level of physical activity (sedentary-nonsedentary) and gender (male-female) were also considered in their study design. In addition, a subject is classified as an obese if his/her BMI is greater than 30. This study recorded data from 1004 eligible individuals and reported 993 unique microorganisms across the Indian microbiome population. Unfortunately, in this dataset neither a longitudinal analysis across time of individuals nor changes in the composition of microbiome in old subjects are available.

2.1. Cluster Analysis. In general, microbiome empirical distribution includes a high proportion of zero observation and a truncation point mass to account for high values that are too sparse to model; for that reason, models that gives an accurate estimates of the true proportion of zeros have been considered in the literature (see [22] and [23]). In addition, given the dynamic character of the microbiome other techniques such as functional response regression on correlated longitudinal microbiome sequencing data has been recently considered in the literature [24]. In this work, in order to facilitate further analyses and reduce the dimensionality of our data set, we started carrying out a cluster analysis. A main task of exploratory data mining, to group a set of bacterial coabundance collections in such a way that objects in the same group or cluster are more similar to each other than to those in other clusters. We performed our clustering based on the coabundance of genus-like groups at a taxonomic level of species within a sample of 1004 subjects. A total of 993 bacterial genera were identified. The core microbiota analysis was completed by using the *Hierarchical Ordered Partitioning and Collapsing Hybrid* (HOPACH) package in **R** that can be downloaded from the *Bioconductor* website http://www.bioconductor.org/. This package includes the HOPACH clustering algorithm that assembles a hierarchical tree of clusters by recursively portioning the whole dataset while ordering and collapsing clusters at each level. In our analysis, we have discarded redgenus that contain at least a minimum relative abundance of 30%, i.e., 70% of zeros in the sample of 1004 individuals. The algorithm uses the MSS (Mean/Median Split Silhouette) criteria to identify the level of the tree with maximally homogeneous clusters. The correlation distance (cor) was the metric selected for clustering the microbiome species by calculating dissimilarities between variables. We have also used a nonparametric bootstrap to estimate the probability that each species belongs to each cluster and to better understand the variability of each cluster. For that reason, we employed the "boothopach" function by taking 1000 bootstrap resample datasets to obtain a suitable balance between precision and speed. As a result of this, we were able to group the microbiome in five groups containing different numbers of genera (see supplementary tables in Table 1). The five different group of bacteria (classes) identified from the cluster analyses could be associated with different taxonomic groups according to the most abundant or representative genus for each of the identified clusters. Groups 1 and 4 are the two largest groups in terms of number of taxonomic elements. Also, as in Group 1, a majority of members comes from the *Bacteroidales-Bacteroidaceae* group that represents almost 2/3 of the species contained in this cluster (17 out of 27 members), it could be related to *Bacteroidales-Bacteroidaceae* cluster. Group 4 is associated with *Lachnospiraceae* which represent almost 1/3 of the total in this group (7 out of 23 members). The other three groups (2, 3, and 5) were assigned to the *Ruminococcaceae-Bifidobacteriaceae* group (5 out of 19 members), *Negativicutes* group (4 out of 19 members), and *Pasteurellaceae* group (3 out of 15 members), respectively. The results and relationships between the different elements on each of the clusters are presented in Figure 1. Here, species close to each other in the tree are shown in a similar way. The ordered distance matrix shows the clustering structure. Similar clusters appear as blocks on the diagonal of this heatmap. Darker colours represent small distances whereas the lighter colours represent large distances. The identified clusters have different sizes and compositions, with two large coabundance clusters, grouping the majority of the genus analyzed. It is important to note that we have combined under the name Group 0 all the discarded operational taxonomic units, that is, all the species with more that 70% of zeroes in the sample.

Table 2 displays the mean, median, and standard deviation for each one of the coabundance groups. It is noticeable that the proportion of bacteria that belongs to Group 1 is higher in average than the proportion in the other groups. The variability is also larger for the first coabundance group.

In Figure 2, some ternary plots for different combinations of the bacterial groups are displayed. In particular we have compared the coabundance Group 1, with Group 2 (top left), Group 3 (top right), Group 4 (bottom left), and Group 5 (bottom right). In order to ensure that the total sum is one, we have combined the coabundance proportion for the rest of the groups in each graph as *Others*. Group 1 is always located at the top of each triangle. The proportion of coabundance of Group 1 is measured in terms of the horizontal lines, i.e., 0% of coabundance is measured in terms of base of the triangle (farthest from the vertex Group 1). In the lower left apex of each triangle is represented the groups compared to Group 1. The right side of the triangle now becomes the baseline for the percentage of the groups located in this vertex. Finally, the combined groups are located at the lower right apex of the triangle.

The rate of coabundance for the combined groups is calculated from the left side of the triangle (0% abundance) to the lower right corner (100% abundance). It is observable that the data lie from a high amount of coabundance of Group 1 and Group 2 with a low coabundance of third, fourth, and fifth groups (top left graph). From the rest of the graphs, it can be inferred that when Group 1 is compared to the other groups, the coabundance of these groups is lower than in the former graph. Also, as Group 2 has been included in the lot *Others*, the corresponding coabundance of the combined group is higher than in the top left graph.

2.1.1. Compositional Data Predictor. Compositional data can be defined as arrays of strictly positive numbers for which ratios between them are important without any further requirement [25]. Microbiome data are compositional, that is, the distance between component values is only meaningful proportionally (see [26]). The elements of the composition are non-negative and sum to unity. An important issue in microbiome data is the large presence of zeros; however, the issue of zero values in some components is not addressed in most papers and especially in the task of regression. In general, in compositional research problems, most of the basic statistical analysis tools are incorrect unless the variables are rewritten in terms of logarithms of ratios as proposed in the log-ratio methodology for compositional data. After computing these log-ratios, standard regression methods can be used since the relative character of the

Table 1: Composition of the five clusters identified in the analysis of the Indian dataset (Dubey et al., [21]).

Group 1
Bacteroidetes-Bacteroidia-Bacteroidales-Rikenellaceae-Alistipes-onderdonkii
Bacteroidetes-Bacteroidia-Bacteroidales-Rikenellaceae-Alistipes-putredinis
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-coprocola
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-coprophilus
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-dorei
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-fragilis
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-intestinalis
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-plebeius
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-thetaiotaomicron
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-uniformis
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-vulgatus
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-xylanisolvens
Bacteroidetes-Bacteroidia-Bacteroidales-Porphyromonadaceae-Barnesiella-intestinihominis
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-biforme
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-ventriosum
Firmicutes-Clostridia-Clostridiales-Clostridiales-Flavonifractor-plautii
Proteobacteria-Gammaproteobacteria-Enterobacteriales-Enterobacteriaceae-Klebsiella-variicola
Firmicutes-Bacilli-Lactobacillales-Lactobacillaceae-Lactobacillus-rogosae
Firmicutes-Negativicutes-Veillonellales-Veillonellaceae-Megasphaera-sp.
Bacteroidetes-Bacteroidia-Bacteroidales-Porphyromonadaceae-Odoribacter-splanchnicus
Bacteroidetes-Bacteroidia-Bacteroidales-Porphyromonadaceae-Parabacteroides-distasonis
Bacteroidetes-Bacteroidia-Bacteroidales-Porphyromonadaceae-Parabacteroides-merdae
Bacteroidetes-Bacteroidia-Bacteroidales-Prevotellaceae-Prevotella-copri
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Roseburia-sp.
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-bromii
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-callidus
Proteobacteria-Betaproteobacteria-Burkholderiales-Sutterellaceae-Sutterella-wadsworthensis
Group 2
Actinobacteria-Actinomycetales-Actinomycineae-Actinomycetaceae-Actinomyces-odontolyticus
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-massiliensis
Actinobacteria-Actinobacteria-Bifidobacteriales-Bifidobacteriaceae-Bifidobacterium-adolescentis
Actinobacteria-Actinobacteria-Bifidobacteriales-Bifidobacteriaceae-Bifidobacterium-bifidum
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Blautia-luti
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Blautia-wexlerae
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Butyrivibrio-crossotus
Actinobacteria-Coriobacteriia-Coriobacteriales-Coriobacteriaceae-Collinsella-aerofaciens
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-eligens
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Faecalibacterium-prausnitzii
Firmicutes-Clostridia-Clostridiales-Clostridiales-Howardella-ureilytica
Proteobacteria-Gammaproteobacteria-Enterobacteriales-Enterobacteriaceae-Klebsiella-pneumoniae
Firmicutes-Negativicutes-Veillonellales-Veillonellaceae-Megasphaera-micronuciformis
Proteobacteria-Betaproteobacteria-Burkholderiales-Sutterellaceae-Parasutterella-excrementihominis
Firmicutes-Clostridia-Clostridiales-Peptostreptococcaceae-Peptostreptococcus-stomatis
Firmicutes-Negativicutes-Acidaminococcales-Acidaminococcaceae-Phascolarctobacterium-faecium
Spirochaetes-Spirochaetes-Spirochaetales-Spirochaetaceae-Treponema-succinifaciens
Firmicutes-Erysipelotrichia-Erysipelotrichales-Erysipelotrichaceae-Turicibacter-sanguinis
Lentisphaerae-Lentisphaeria-Victivallales-Victivallaceae-Victivallis-vadensis
Group 3
Proteobacteria-Alphaproteobacteria-Rhodospirillales-Acetobacteraceae-Acidiphilium-sp.
Verrucomicrobia-Verrucomicrobiae-Verrucomicrobiales-Akkermansiaceae-Akkermansia-muciniphila
Proteobacteria-Gammaproteobacteria-Aeromonadales-Succinivibrionaceae-Anaerobiospirillum-succiniciproducens
Actinobacteria-Actinobacteria-Bifidobacteriales-Bifidobacteriaceae-Bifidobacterium-longum
Proteobacteria-Deltaproteobacteria-Desulfovibrionales-Desulfovibrionaceae-Bilophila-wadsworthia
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Blautia-obeum
Firmicutes-Erysipelotrichia-Erysipelotrichales-Erysipelotrichaceae-Bulleidia-p-1630-c5
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Coprococcus-catus
Proteobacteria-Deltaproteobacteria-Desulfovibrionales-Desulfovibrionaceae-Desulfovibrio-piger
Firmicutes-Negativicutes-Veillonellales-Veillonellaceae-Dialister-succinatiphilus
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-siraeum
Firmicutes-Negativicutes-Veillonellales-Veillonellaceae-Megasphaera-elsdenii

Table 1: Continued.

Firmicutes-Negativicutes-Selenomonadales-Selenomonadaceae-Mitsuokella-jalaludinii
Firmicutes-Negativicutes-Selenomonadales-Selenomonadaceae-Mitsuokella-multacida
Bacteroidetes-Bacteroidia-Bacteroidales-Prevotellaceae-Prevotella-stercorea
Firmicutes-Clostridia-Clostridiales-Clostridiaceae-Romboutsia-ilealis
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminiclostridium-siraeum
Actinobacteria-Coriobacteriia-Eggerthellales-Eggerthellaceae-Slackia-isoflavoniconvertens
Proteobacteria-Betaproteobacteria-Burkholderiales-Sutterellaceae-Sutterella-sp.
Group 4
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-caccae
Bacteroidetes-Bacteroidia-Bacteroidales-Bacteroidaceae-Bacteroides-ovatus
Firmicutes-Erysipelotrichia-Erysipelotrichales-Erysipelotrichaceae-Catenibacterium-mitsuokai
Firmicutes-Clostridia-Clostridiales-Clostridiaceae-Clostridium-bartlettii
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Coprococcus-comes
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Coprococcus-eutactus
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Dorea-formicigenerans
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Dorea-longicatena
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-hadrum
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-hallii
Firmicutes-Clostridia-Clostridiales-Eubacteriaceae-Eubacterium-ramulus
Proteobacteria-Alphaproteobacteria-Rhizobiales-Hyphomicrobiaceae-Gemmiger-formicilis
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Lachnoclostridium-clostridioforme
Firmicutes-Clostridia-Clostridiales-Oscillospiraceae-Oscillibacter-sp.
Firmicutes-Negativicutes-Acidaminococcales-Acidaminococcaceae-Phascolarctobacterium-succinatutens
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Roseburia-faecis
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Roseburia-inulinivorans
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-faecis
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-gnavus
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-sp.
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-torques
Firmicutes-Negativicutes-Selenomonadales-Selenomonadaceae-Selenomonas-bovis
Proteobacteria-Betaproteobacteria-Burkholderiales-Sutterellaceae-Sutterella-stercoricanis
Group 5
Proteobacteria-Gammaproteobacteria-Pasteurellales-Pasteurellaceae-Actinobacillus-minor
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Blautia-faecis
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Blautia-producta
Firmicutes-Clostridia-Clostridiales-Clostridiaceae-Clostridium-disporicum
Firmicutes-Clostridia-Clostridiales-Clostridiaceae-Clostridium-perfringens
Firmicutes-Clostridia-Clostridiales-Clostridiaceae-Clostridium-sp.
Proteobacteria-Deltaproteobacteria-Desulfovibrionales-Desulfovibrionaceae-Desulfovibrio-D168
Proteobacteria-Gammaproteobacteria-Enterobacteriales-Enterobacteriaceae-Escherichia-coli
Proteobacteria-Gammaproteobacteria-Pasteurellales-Pasteurellaceae-Haemophilus-parainfluenzae
Proteobacteria-Gammaproteobacteria-Pasteurellales-Pasteurellaceae-Haemophilus-pittmaniae
Firmicutes-Bacilli-Lactobacillales-Lactobacillaceae-Lactobacillus-ruminis
Proteobacteria-Gammaproteobacteria-Pseudomonadales-Pseudomonadaceae-Pseudomonas-lini
Firmicutes-Clostridia-Clostridiales-Lachnospiraceae-Roseburia-intestinalis
Firmicutes-Clostridia-Clostridiales-Ruminococcaceae-Ruminococcus-gauvreauii
Firmicutes-Negativicutes-Veillonellales-Veillonellaceae-Veillonella-dispar

information is considered when analyzing the results, as one group or variable can only increase in relative terms if some other group or groups reduce. In this work we focus on the case of compositional data being included in the predictor variables. The effect of increasing one of the variables in relative terms in the predictor therefore depends on which other variables are decreased when this occur. In log-ratio parlance, the effect of increasing one log-ratio is interpreted while keeping all other log-ratios constant as the same log-ratio may have different meaning depending on the way that the other log-ratios in the model are assembled. Thus, the interpretation of log-ratios as explanatory variables is usually different from other approaches. Several different approaches of building and interpreting the log-ratios have been considered in the literature, often leading to the same predictions and residuals [27]. Among the different parametrizations, in this work, we have chosen centred log-ratios [28]. In our analysis, we consider a vector of 6-dimensional real space that carries information on the relative importance of its components,

$$\begin{aligned} \mathbf{x}_i &= \left(x_{i0}, x_{i1}, \ldots, x_{i5}\right) \in \mathbb{R}_{+}^{6}, \text{with } x_{ij} > 0, \\ j &= 0, 1, 2, \ldots, 5, \sum_{j=0}^{5} x_{ij} = 1, \end{aligned} \tag{1}$$

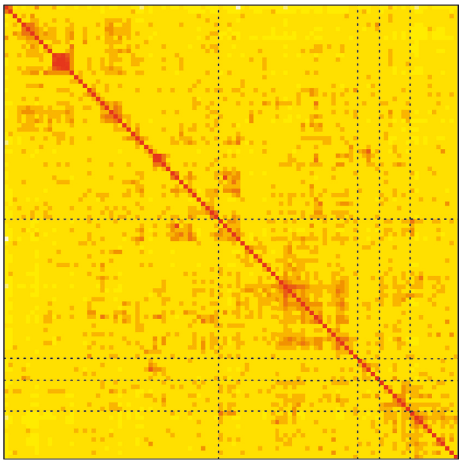


Figure 1: Heatmap of coabundance groups. This figure represents the distance among the abundances of the different genus of bacteria characterised in the sequencing analyses. The dashed lines depict the boundaries between the five different clusters. Darker colours represent closer distances (coabundance) between genus whereas lighter colours displays larger distances.

Table 2: Mean, median, and standard deviation for each coabundance group.

Coabundance group	Mean	Median	Standard deviation
Group 1	0.5548	0.5669	0.1542
Group 2	0.1899	0.1670	0.1024
Group 3	0.0816	0.0624	0.0688
Group 4	0.0899	0.0813	0.0517
Group 5	0.0603	0.0239	0.0859
Group 0	0.0235	0.0120	0.0632

where $i = 1, \dots, 1004$. Note that the explanatory variables x_{ij} represent the proportion of the bacterial coabundance proportion of Group j in individual i. Centred log-ratios are calculated by using a quotient between each variable and the geometric mean of all components (see [28]),

$$\log_2\left(\frac{x_j}{\sqrt[6]{\prod_{j=0}^{5} x_j}}\right), \quad \text{with } j = 0, 1, 2, \dots, 5. \tag{2}$$

The fact that we are using logarithms to base 2 means that a unit increase in this logarithm leads to a double increase in the original magnitude. In order to avoid perfect collinearity one centred log-ratio must be deleted from the regression equation. Since all six centred log-ratios add-up to zero, by increasing a fixed centred log-ratio while keeping the other four remaining log-ratios in the regression equation (with regressors β_j with $j = 0, 1, 2, \dots, 5$) constant implies increasing the given centred log-ratio whilst reducing the omitted centred log-ratio by the same amount. In this regard, a positive statistically significant regression β_j coefficient indicates an increasing value of the covariate x_j at the expense of decreasing the amount of the omitted component has a significant positive effect on expected value of the response variable. This is equivalent to say that in terms of logarithm of base 2, β_j is interpreted as the expected change in the response variable when the ratio between x_j and the omitted explanatory variable is multiplied by six. Finally, in order to obtain the estimates and their corresponding p values for all possible pair combinations, the model needs to be repeated six different times by ignoring each time a different centred log-ratio.

3. Relation of Age, BMI, and Level of Physical Activity with Microbiome

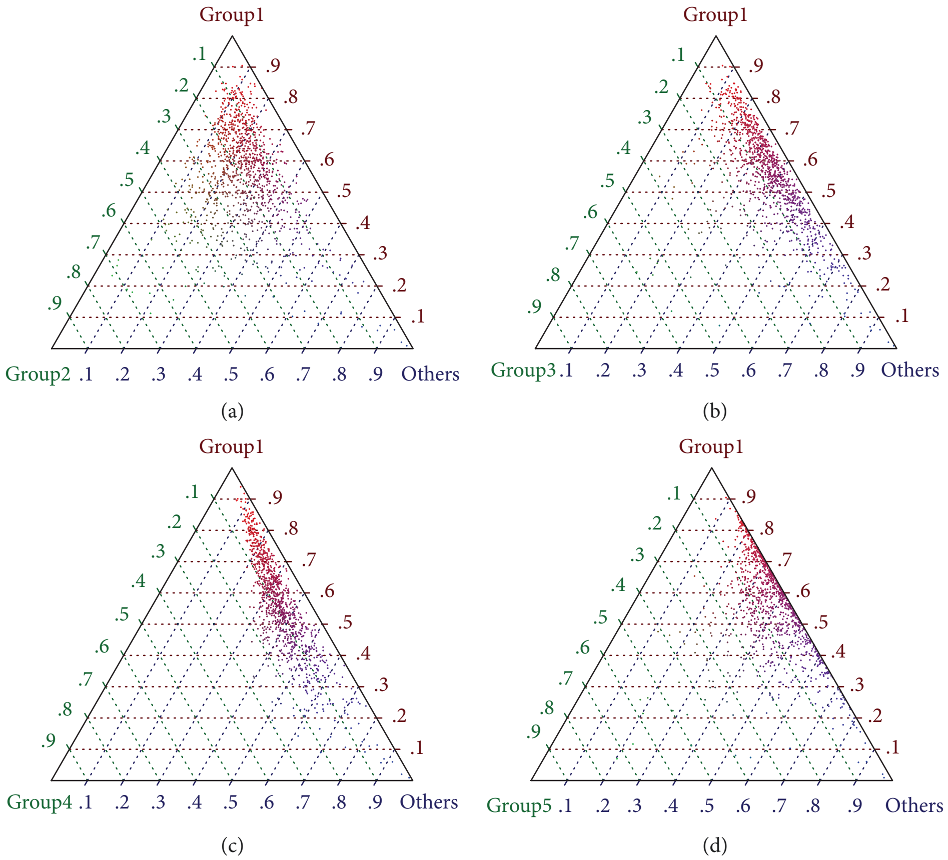
In this section, we firstly consider the marginal relation of age given a level of physical activity with microbiome through a beta regression model. Next, we examine the connection of BMI with the microbiome given the level of physical activeness by using a mixture of logistics regression model. Finally, the joint relationship of these variables is examined via a t-copula.

3.1. Relation of Age and Level of Physical Activity with Microbiome. It is our interest to model the relationship between the age of the subject and the proportion of each coabundance genus-like groups at taxonomic level of species via a beta regression model (see Ferrari and Cribari-Neto [29]). This model assumes that the response variable is beta distributed using a parametrization of the beta law that is indexed by mean and dispersion parameters. This regression family is useful for modeling rates and proportions, that is, in situations where the dependent variable of interest is continuous and restricted to a bounded interval (a, b) where a and b are known scalars with $a < b$. This model is related to other variables through a regression structure. Our goal is to explain a continuous response variable Y_1 with $a < y_1 < b$. The density of Y_1 is defined as follows:

$$f_1(y_{1i} | \omega_i, \phi) = \frac{\Gamma(\phi)(b-a)^{1-\phi}}{\Gamma(\omega_i \phi)\Gamma((1-\omega_i)\phi)} \left(\frac{y_{1i}-a}{b-a}\right)^{\omega_i \phi - 1} \left(\frac{b-y_{1i}}{b-a}\right)^{(1-\omega_i)\phi - 1}, \tag{3}$$

with $0 < \omega_i < 1$ and $\phi > 0$ with $i = 1, \dots, n$. This parametrization allows us to obtain a regression structure for the mean of the response along with a dispersion parameter ϕ. Here, n is the sample size and $E(Y_{1i}) = \omega_i(b-a) + a$. The variance of the response variable can be easily explained in terms of its mean by the following expression $\mathrm{Var}(Y_{1i}) = (b-a)^2 \omega_i(1-\omega_i)/1+\phi$. The variance decreases with the value of the dispersion parameter.

Let us now consider that a random variable Y_{1i} denoting age of the individual i in the sample is related to a compositional data predictor related to each one the coabundance groups, $\mathbf{x}_i = (1, u_{i1}, \dots, u_{i5})^\top$ where $(u_{i1}, \dots, u_{i5})$ are chosen among all combinations without repetition from the vector $(x_{i0}/\sqrt[6]{\prod_{j=0}^{5} x_j}, \dots, x_{i5}/\sqrt[6]{\prod_{j=0}^{5} x_j})^\top$ taking 5 components at a time. Then, by using the logit link (i.e., $h(\omega_i) = \log\omega_1/1-\omega_i$), we have that $\omega_i = \exp(\mathbf{x}_\mathbf{i}^\top \underline{\beta})/1 + \exp(\mathbf{x}_\mathbf{i}^\top \underline{\beta})$, where $\underline{\beta} = (\beta_0, \beta_1, \dots, \beta_5)^\top$ is



(a)

(b)

(c)

(d)

Figure 2: Ternary plots associated to different coabundance groups. (a) Group1/Group 2/Other groups. (b) Group 1/Group 3/Other groups. (c) Group 1/Group 4/Other groups. (d) Group 1/Group 5/Other groups.

Table 3: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as sedentary. The response variable is age.

	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Group 1	−0.0341	—	0.1369	0.0047	−0.0050	0.0921
	0.6026	—	0.1388	0.9593	0.9473	0.1262
Group 2	−0.1708	−0.1368	—	−0.1313	−0.1500	−0.0448
	0.0117	0.1388	—	0.0861	0.0686	0.4267
Group 3	−0.0404	−0.0064	0.1305	—	−0.0076	0.0857
	0.49404	0.9439	0.0880	—	0.7957	0.0956
Group 4	−0.0448	−0.0112	0.1253	−0.0046	—	0.0809
	0.4992	0.9017	0.1923	0.9515	—	0.2157
Group 5	−0.1260	−0.0921	0.0446	−0.0863	−0.1023	—
	0.0044	0.1261	0.4286	0.0931	0.0284	—
Group 0	—	0.0341	0.1708	0.0397	0.0145	0.1261
	—	0.6025	0.0117	0.5010	0.6334	0.0044
Intercept	−0.1209	−0.1210	−0.1214	−0.1178	−0.1088	−0.1210
	0.5175	0.5173	0.5157	0.5285	0.5659	0.5173
ϕ	2.9797	2.9795	2.9796	2.9789	2.9807	2.9795
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
AIC	3481.07	3481.07	3481.07	3481.07	3480.86	3481.07

a vector of regressors. Other choices for the link function link functions for the response model are feasible.

We have fitted this beta regression model to this dataset to explain the response variable *Age* by considering two levels of physical activity: sedentary and nonsedentary by assuming $a = 17.5$ and $b = 65.5$. Below in Table 3, the estimates and p values associated with the six predictors for individuals classified as sedentary for each microbiome coabundance group's proportion obtained under the regression model (1). Similarly, in Table 4 estimates and p

Table 4: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as nonsedentary. The response variable is age.

	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Group 1	−0.0465	—	0.1209	−0.0105	−0.0745	0.0159
	0.4022	—	0.0283	0.8851	0.3583	0.7459
Group 2	−0.1673	−0.1209	—	−0.1328	−0.1954	−0.1052
	0.0063	0.1151	—	0.0461	0.0344	0.0459
Group 3	−0.0358	0.0108	0.1318	—	−0.0637	0.0265
	0.4957	0.8814	0.0477	—	0.4014	0.5497
Group 4	0.0277	0.0745	0.1954	0.0641	—	0.0899
	0.6623	0.3583	0.0344	0.3983	—	0.1771
Group 5	−0.0623	−0.0157	0.1052	−0.0263	−0.0902	—
	0.1196	0.7484	0.0458	0.5522	0.1753	—
Group 0	—	0.0465	0.1674	0.0349	−0.0280	0.0622
	—	0.4021	0.0063	0.5069	0.6589	0.1206
Intercept	−0.1641	−0.1642	−0.1641	−0.1667	−0.1643	−0.1649
	0.2890	0.2888	0.2892	0.2820	0.2887	0.2869
ϕ	2.9002	2.8999	2.9000	2.8981	2.8999	2.8999
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
AIC	4155.08	4155.08	4155.08	4155.08	4155.08	4155.08

values results for each predictor are shown for nonsedentary subjects. From these tables, it is discernible that for the first predictor the regressor associated to *Group 5* for sedentary individuals is statistically significant at the 5% level whereas it is not for nonsedentary subjects. Its value, −0.1260, is interpreted as the decrease in the covariate x_5 (i.e., *Pasteurellaceae*) at the expense of increasing the amount of x_0 has a significant negative effect on $h(\omega_i)$ while keeping the remaining four log-ratios in the equation constant. In a similar fashion for the third predictor for the nonsedentary subjects the explanatory variables, x_1, x_3, x_4, and x_5 are significant at the same level while they are not for the sedentary individuals. For all these covariates, the sign of their regressors is positive; therefore, an increase in these regressors at the expense of decreasing the value of x_2 has a significant effect on the transformation of the expected value of the mean of the model. For the fourth predictor, the regression coefficient associated to the second group is only significant for the nonsedentary party. Similarly for the fifth predictor, x_5 is significant for the sedentary individuals whereas the regressor associated to *Group 2*, i.e., *Ruminococcaceae-Bifidobacteriaceae* is significant for the nonsedentary individuals. Similar situation is also verified for the sixth predictor. On the other hand, for the sedentary subjects, the variable x_0 is a positive significant variable.

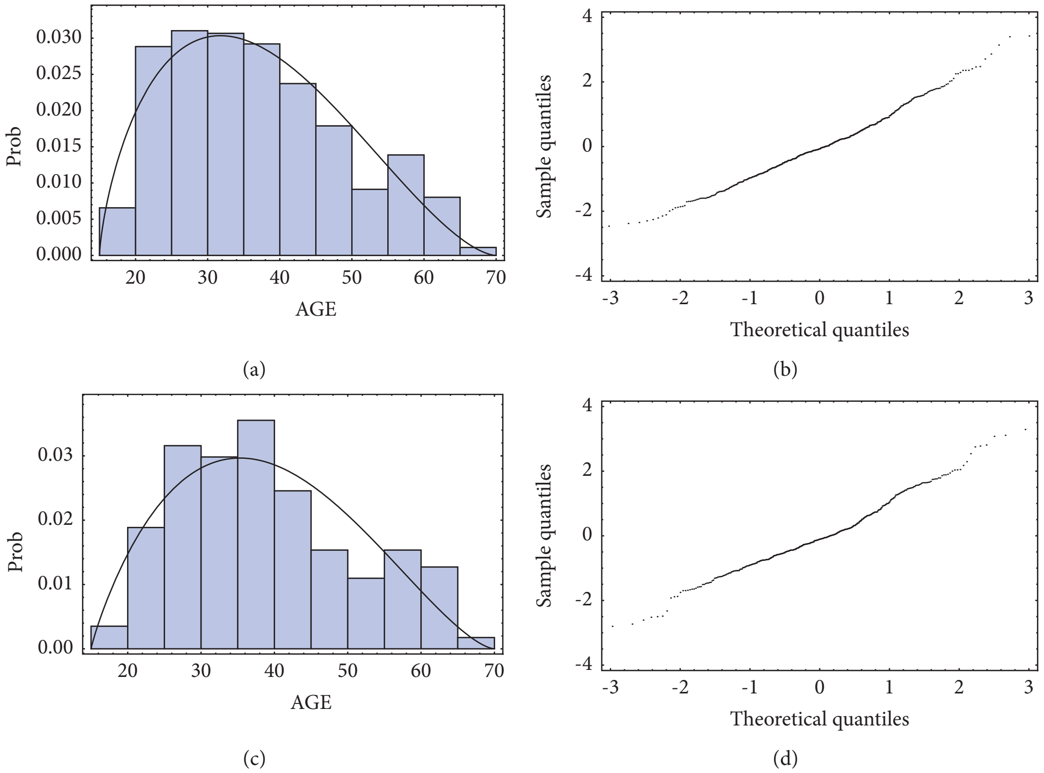
In Figure 3, we have plotted the histograms of the empirical distribution of the response variable *Age* for the nonsedentary (top left panel) and sedentary (bottom left panel) subjects. For both histograms, we have superimposed the probability density function of the beta distribution. It is observable that this distribution provides a better fit to empirical data for the nonsedentary group than for the sedentary party. Furthermore, we have performed a diagnostic analysis to check the goodness-of-fit of the estimated model by providing a global measure of explained variation and graphical tools based on QQ-plots, to detect departures from the given model and influential observations. Residuals are used to check the appropriateness of a chosen model and to identify outliers. For that reason, randomized quantile residuals (Dunn and Smyth [30]) are used since other type of residuals, i.e., Pearson's and deviance residuals are far from normality when the parameters of the model are known and they fail to provide useful information of the inadequacy of the model. The ith randomized quantile residuals for a discrete response variable is defined as $r_{q\,i} = \Phi^{-1}(F(y_{1i}; \widehat{\omega}_i, \widehat{\phi}))$ where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution and $F(y_{1i}; \widehat{\omega}_i, \widehat{\phi})$ is the cumulative distribution function associated to the beta regression model evaluated at the estimated parameters for $i = 1, \ldots, n$. In the right panels of Figure 3, the QQ-plots of the randomized quantile residuals of the beta regression models when the predictor 1 is considered for the nonsedentary (top right) and sedentary (bottom right) subjects. Each dot on the plots represents an empirical residual. A perfect alignment with the 45° line implies that the residuals are normally distributed. In general, it is observable that the residuals for the nonsedentary group adhere closer to the line in the whole distribution.

3.2. Relation of BMI and Level of Physical Activity with Microbiome.

Regardless of the gender, the empirical distribution of BMI in humans is bimodal. Then, finding appropriate statistical models that have the capacity to explain bimodal datasets is an issue of vital importance. In this work, we use a mixture of two logistic distributions with different locations and scale parameters. We have chosen this family for its flexibility and simplicity. It is now our interest to explain the BMI in the population in terms of a random variable $Y_2 \in \mathbb{R}$. The probability density function of this random variable is

$$f_2(y_{2i} | w_i, \mu_1, \sigma_1, \mu_2, \sigma_2) = \frac{w_i}{4\sigma_1} \operatorname{sech}^2\left(\frac{y_{2i} - \mu_1}{2\sigma_1}\right), \quad (4)$$

$$+ \frac{1 - w_i}{4\sigma_2} \operatorname{sech}^2\left(\frac{y_{2i} - \mu_2}{2\sigma_2}\right), \quad (5)$$



(a)

(b)

(c)

(d)

Figure 3: Histograms of the *Age* distributions and the beta density superimposed and QQplots of the randomized quantile residuals (RQRs) for each regression model by using Predictor 1. (a) Histogram of age and nonsedentary. (b) RQR of age and nonsedentary against Predictor 1 regression model. (c) Histogram of age and sedentary. (d) RQR of age and sedentary against Predictor 1 regression model.

where $\mu_1, \mu_2 \in \mathbb{R}$ are location parameters and $\sigma_1, \sigma_2 > 0$ are scale parameters. In this model, it will also be assumed that the weight parameter for each individual in the sample is again expressed as a function of the same group of covariates, $w_i = \exp(\mathbf{x}_\mathbf{i}^\top \underline{\alpha})/1 + \exp(\mathbf{x}_\mathbf{i}^\top \underline{\alpha})$, where $\underline{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_5)^\top$ is a vector of regressors. Other choices for the link function link functions for the response model are also possible. This parametrization enable us to obtain a regression structure for the mean of the response in the following way, $E(Y_{2i}) = w_i\mu_1 + (1 - w_i)\mu_2$. The variance of the response variable is written in terms of the following linear combination of the scale parameters:

$$\operatorname{Var}(Y_{2i}) = \frac{\pi^2}{3}\left\{(w_i\sigma_1)^2 + ((1 - w_i)\sigma_2)^2\right\}. \qquad (6)$$

We have now fitted the mixture of logistics regression model given by (5) to this dataset to explain the dependent variable *BMI* by again considering two levels of physical activity: sedentary and nonsedentary. In Table 5, the estimates and p values associated with the six predictors for individuals classified as sedentary for each microbiome coabundance groups proportion obtained under this mixture of logistics regression model (5). In a similar way, in Table 6, estimates and p values results for each predictor are shown for nonsedentary subjects. From these tables, it is apparent that for the first predictor the regressor associated to *Group 1* for sedentary individuals is statistically significant at the 5% level whereas it is not for non-sedentary subjects. The estimated value is −0.3631, that it is interpreted as the decrease in the covariate x_1 (i.e., *Bacteroidales-Bacteroidaceae*) at the expense of increasing the amount of x_0 has a significant negative effect on $h(\omega_i)$ while keeping the remaining four log-ratios in the equation of the predictor constant. Similarly, for the second predictor and the sedentary subjects, the explanatory variables, x_0 is statistically significant at the 10% significance level while it is not for the nonsedentary individuals. The sign of this regression coefficient is positive; therefore, an increase in this regressor at the expense of decreasing the value of x_1, while keeping the other four log-ratios in the equation constant has a significant effect on the transformation of the expected value of the mean of the model. Finally, for predictor 3, the regression coefficient associated to the first group is only significant at the 10% significance level for the sedentary individuals. The sign of this regressor is negative.

In Figure 4, we have plotted the histograms of the empirical distribution of the response variable *BMI* for the nonsedentary (top left panel) and sedentary (bottom left panel) individuals. For both histograms, we have superimposed the density function of the mixture of logistics distributions. It can be seen that this distribution is able to reproduce the two modes of the empirical distribution for both cohorts. Note that for the group of sedentary individuals, the second modal value located around the *BMI* value of 31 is clearly more predominant. Once again, we have plotted the QQ-plots of the randomized quantile residuals of this mixture of logistics regression when the first predictor 1

Table 5: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as sedentary. The response variable is BMI.

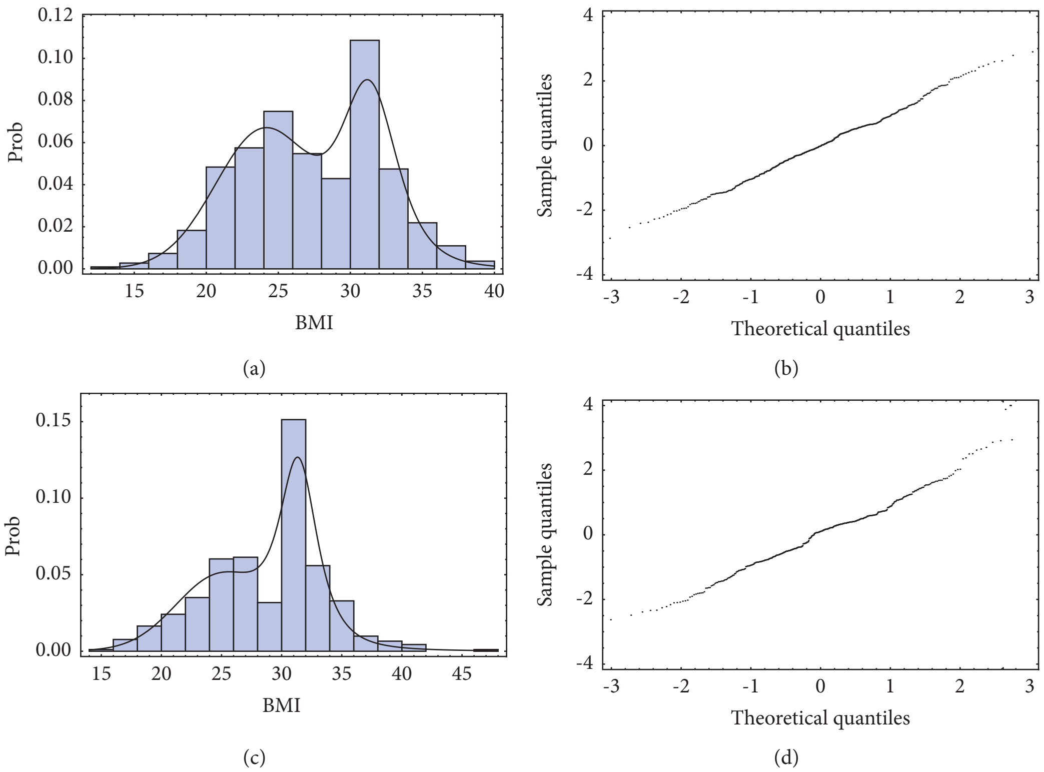
	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Group 1	−0.3631	—	−0.5055	−0.2915	−0.1853	−0.1829
	0.0492	—	0.0534	0.2393	0.4711	0.2635
Group 2	0.2200	0.3914	—	0.2154	0.3215	0.1421
	0.2429	0.1217	—	0.3117	0.2173	0.3455
Group 3	−0.0267	0.2575	−0.2111	—	0.1062	−0.0071
	0.8716	0.2966	0.3209	—	0.6199	0.9604
Group 4	−0.1777	0.0996	−0.3113	−0.1064	—	−0.1145
	0.3450	0.6939	0.2281	0.6195	—	0.5254
Group 5	−0.0527	0.2175	−0.2280	−0.0181	0.0881	—
	0.6648	0.1641	0.1379	0.9004	0.6246	—
Group 0	—	0.3143	−0.1292	0.0849	0.1911	0.1057
	—	0.0819	0.4838	0.6061	0.3049	0.3826
Intercept	0.3101	0.3894	0.3626	0.3633	0.3634	0.3374
	0.5688	0.4873	0.4983	0.4989	0.4989	0.5406
μ_1	24.4923	24.70000	24.4250	24.4537	24.4532	24.6451
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
μ_2	31.3471	31.4683	31.3505	31.3616	31.3613	31.4405
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
s_1	2.1269	2.1761	2.0905	2.0984	2.0984	2.1679
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
s_2	1.6660	1.6123	1.6664	1.6643	1.6644	1.6371
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
AIC	2717.15	2717.25	2716.83	2716.83	2716.83	2717.54

Table 6: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as nonsedentary. The response variable is BMI.

	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Group 1	0.0419	—	0.1453	0.0264	0.2168	0.0526
	0.6145	—	0.4500	0.8847	0.2696	0.6742
Group 2	−0.1095	−0.1466	—	−0.1185	0.0662	−0.1277
	0.4624	0.4459	—	0.4661	0.7654	0.3291
Group 3	0.0164	−0.0282	0.1186	—	0.1793	0.0202
	0.8980	0.8769	0.4660	—	0.3258	0.8540
Group 4	−0.1438	−0.1902	−0.0425	−0.1618	—	−0.1669
	0.3464	0.3322	0.8481	0.3746	—	0.2921
Group 5	0.0084	−0.0252	0.1204	0.0022	0.1827	—
	0.9295	0.8366	0.3558	0.9843	0.2488	—
Group 0	—	−0.0328	0.1126	−0.0056	0.1725	−0.0027
	—	0.8068	0.4497	0.9650	0.2594	0.9771
Intercept	0.2075	0.2282	0.2308	0.2307	0.1902	0.1731
	0.6145	0.5798	0.5754	0.5754	0.6438	0.6438
μ_1	23.9065	23.8923	23.8953	23.8921	23.8906	23.8869
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
μ_2	31.2608	31.2582	31.2606	31.2576	31.2563	31.2567
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
s_1	1.9478	1.9444	1.9434	1.9417	1.9408	1.9418
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
s_2	1.4769	1.4785	1.4779	1.4788	1.4790	1.4803
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
AIC	3264.49	3264.45	3264.45	3264.45	3264.47	3264.58

is considered for the nonsedentary (top right) and sedentary (bottom right) individuals. In general, it is observable that the residuals for the nonsedentary group adhere closer to the line in the whole distribution but it underestimates the top part of the distribution of residuals.

3.3. Joint Relation of Age, BMI, and Level of Physical Activity with Microbiome. The degree of association between the two variables age and BMI in the sample for the different levels of physical activity can be summarized in terms of some measures of correlation for bivariate data. In Table 7,



(a)

(b)

(c)

(d)

Figure 4: Histograms of the BMI distributions and the mixture of logistic density superimposed and QQplots of the randomized quantile residuals (RQR) for each regression model by using Predictor 1. (a) Histogram of BMI and nonsedentary. (b) RQR of BMI and nonsedentary against Predictor 1 regression model. (c) Histogram of BMI and sedentary. (d) RQR of BMI and sedentary against Predictor 1 regression model.

Table 7: Pearson's, Spearman's, and Kendall's measures of correlation for the variables age and BMI and different levels of physical activity.

Measure of correlation	Physical activeness level	
	Sedentary	Nonsedentary
Pearson's	0.2238	0.1738
Spearman's	0.2060	0.1828
Kendall's	0.1421	0.1242

Pearson's, Spearman's, and Kendall's measures of correlation for these continuous random variables are displayed. It is noticeable that there exists weak positive correlation between these two variables. The degree of association is less intense for the nonsedentary individuals.

We model the joint dependence of age and BMI for different level of physical activity and their relationship with the proportion of each coabundance genus-like groups at taxonomic level of species via a t-copula with degrees of freedom (df) parameter $\nu > 1$ with marginal distributions given by the beta regression model given in (1) and the mixture of logistics distributions provided in (5). The density of this of this multivariate distribution is defined as

$$g\left(\underline{y}_i|\Theta_1,\Theta_2,\Theta_3\right) = \frac{\Gamma(\nu+1/2)\Gamma(\nu/2)\left(1+\underline{z}\,\Sigma^{-1}\underline{z}^{\top}/\nu\right)^{-\nu+2/2}}{\Gamma(\nu+1/2)^2|\Sigma|^{1/2}\prod_{j=1}^{2}\left(1+z_j^2/\nu\right)^{-\nu+1/2}} \times f_1(y_{i1}|\Theta_1)\times f_2(y_{i2}|\Theta_2), \tag{7}$$

where $\underline{y}_i\colon = (y_{i1}, y_{i2})^{\top}$, $\Theta_1 = (\beta,\phi)$, $\Theta_2 = (\underline{\alpha},\mu_1,\sigma_1,\mu_2,\sigma_2)$, and $\Theta_3 = (\nu,\Sigma)$. Here, $\Sigma\colon = (\rho_{ij})_{1\le i,j\le 2}$ is a symmetric and positive definite scatter matrix with dimension 2×2 with unit diagonal entries and $-1<\rho_{ij}<1$, $|\cdot|$ denotes the determinant of a matrix, and $\Gamma(\cdot)$ is the complete gamma function. Also, $\underline{z}\colon = (z_1,z_2)^{\top}$ with $z_j = t_\nu^{-1}(F_j(y_j|\Theta_j))$ with $j=1,2$, where $t_\nu^{-1}(\cdot)$ is the quantile function of univariate t-distribution with ν df and $F_j(y_{ij}|\Theta_j)$ is the cdf associated to the regression models presented above with $i = 1, \ldots, n$.

The corresponding log-likelihood function, given a sample $\mathbf{y}\colon = \underline{y}_1, \ldots, \underline{y}_n$ is provided by

$$\begin{aligned}\ell(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3|\mathbf{y}) &= \ell_C(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3|\mathbf{y}) + \ell_M(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2|\mathbf{y}) \\ &= n \log \Gamma\left(\frac{\nu+1}{2}\right) + n \log \Gamma\left(\frac{\nu}{2}\right) - 2n \log \Gamma\left(\frac{\nu+1}{2}\right) \\ &\quad - \frac{n}{2}\log|\Sigma| - \frac{\nu+2}{2}\sum_{i=1}^{n}\log\left(1 + \frac{\underline{z}_i\Sigma^{-1}\underline{z}_i^{\top}}{\nu}\right) \\ &\quad + \frac{\nu+1}{2}\sum_{i=1}^{n}\sum_{j=1}^{2}\log\left(1 + \frac{z_{ij}^2}{\nu}\right) + n \log \Gamma(\phi) + n(1-\phi)\log(b-a) \\ &\quad - \sum_{i=1}^{n}(\log \Gamma(\omega_i\phi) + \log \Gamma((1-\omega_i)\phi) + \sum_{i=1}^{n}(\omega_i\phi - 1)\log\left(\frac{y_i - a}{b-a}\right) \\ &\quad + \sum_{i=1}^{n}((1-\omega_i)\phi - 1)\log\left(\frac{b - y_i}{b-a}\right) \\ &\quad + \sum_{i=1}^{n}\log\left(\frac{w_i}{4\sigma_1}\text{sech}^2\left(\frac{y_{2i}-\mu_1}{2\sigma_1}\right) + \frac{1-w_i}{4\sigma_2}\text{sech}^2\left(\frac{y_{2i}-\mu_2}{2\sigma_2}\right)\right),\end{aligned} \tag{8}$$

where $\ell_C(\cdot)$ and $\ell_M(\cdot)$ are the log-likelihood functions of the copula and marginal model, respectively. Maximum likelihood estimation can be used to estimate the parameters of expression (8) via an adaptive maximization by parts (MBP) algorithm as described in [31], by using initial estimates $(\widehat{\Theta}_1^{(0)}, \widehat{\Theta}_2^{(0)}, \widehat{\Theta}_3^{(0)})$ generated by inference for margins algorithm. In the step k of this algorithm for $k = 1, 2, \ldots$, we find,

$$\begin{aligned}\left(\widehat{\boldsymbol{\Theta}}_1^{(k)}, \widehat{\boldsymbol{\Theta}}_2^{(k)}\right) &= \arg\max\left\{\ell_M(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2|\mathbf{y}) + \ell_C\left(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2|\mathbf{y}, \widehat{\boldsymbol{\Theta}}_3^{(k-1)}\right)\right\}, \\ \boldsymbol{\Theta}_3^{(k)} &= \arg\max \ell_C\left(\boldsymbol{\Theta}_3|\mathbf{y}, \widehat{\boldsymbol{\Theta}}_1^{(k)}, \widehat{\boldsymbol{\Theta}}_2^{(k)}\right).\end{aligned} \tag{9}$$

The algorithm stops when a terminating condition between two consecutive iterations is reached, i.e.,

$$\left\|\ell\left(\widehat{\boldsymbol{\Theta}}_1^{(k)}, \widehat{\boldsymbol{\Theta}}_2^{(k)}, \widehat{\boldsymbol{\Theta}}_3^{(k)}|\mathbf{y}\right) - \ell\left(\widehat{\boldsymbol{\Theta}}_1^{(k-1)}, \widehat{\boldsymbol{\Theta}}_2^{(k-1)}, \widehat{\boldsymbol{\Theta}}_3^{(k-1)}|\mathbf{y}\right)\right\|_1 < 10^{-3}. \tag{10}$$

Finally, we have fitted the bivariate distribution given in (5) to the bivariate data set. Once again, the two levels of physical activeness have been considered. Results are shown in Table 8 for the sedentary case and Table 9 for the nonsedentary situation. When using the first predictor, i.e., the omitted covariate is x_0, the variable x_5 and x_1 are statistically significant at the 5% level of significance for the variables age and BMI, respectively, for sedentary individuals while they are not for the nonsedentary group. Also, the regressor associated with the covariates x_2 and x_0 for the *BMI* are significant at the same level of significance for the sedentary individuals whereas they are not for the nonsedentary group. The sign of these regression coefficients is negative. Conversely, for the third predictor and the variable *Age*, the regressors for the variables x_3 and x_5 are positive significant only for the nonsedentary subjects. In addition, the variable x_3 is positive significant for the variable BMI for the same level of physical activeness. Regression coefficients associated to x_2 in the fourth and fifth predictors (omitted variables x_3 and x_4, respectively) are negative significant for the response variable age. Finally, when the explanatory variable x_5 is deleted, the regressor linked to x_0 is positive significant at the 5% level for the sedentary subjects and response variable *Age*. Similarly, for this sixth predictor, the regressor associated to the variable x_2 for the same response variable is negative significant for individuals classified as nonsedentary.

4. Discussion and Extensions

Although genetic information can reliably inform about the future appearance of certain diseases and it is an element of great interest for different stakeholders, this genomic information is considered in the highest level of confidentiality and protected from disclosure. In this sense, in the insurance industry, a particularity of genomic information in this context is that it does not only provide information and knowledge about the individual taking the insurance but also in respect to their ancestors and descendants. As a consequence of these limitations, international regulators, e.g., the Council of Europe encourages insurers to update their actuarial bases according to relevant and new scientific knowledge and this may open the gates to explore new avenues and data types and information sources. As part of

Table 8: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as sedentary (bivariate regression model).

		Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Age	Group 1	−0.0051	—	0.0603	−0.0000	0.0346	0.0857
		0.9316	—	0.5081	0.9997	0.6975	0.1475
	Group 2	−0.1580	−0.1524	—	−0.1233	−0.0983	−0.0437
		0.0178	0.0944	—	0.1011	0.2985	0.4303
	Group 3	−0.0462	−0.0293	0.1203	—	0.0167	0.0824
		0.4265	0.7436	0.1103	—	0.8232	0.1031
	Group 4	−0.0430	−0.0371	0.0880	−0.0065	—	0.0765
		0.5101	0.6761	0.0116	0.0745	—	0.2339
	Group 5	−0.1125	−0.1045	0.0161	−0.0825	−0.0603	—
		0.0099	0.0779	0.7711	0.1027	0.3485	—
	Group 0	—	0.0126	0.1463	0.0440	0.0645	0.1257
		—	0.8448	0.0282	0.4480	0.3238	0.0039
	Intercept	−0.2183	−0.1924	−0.0095	−0.1094	−0.1543	−0.1174
		0.2361	0.2958	0.9587	0.5517	0.4014	0.5226
	ϕ	3.1241	3.1261	3.1143	3.1354	3.1258	3.1411
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
BMI	Group 1	−0.3887	—	−0.4573	−0.2960	−0.1711	−0.2530
		0.0292	—	0.0642	0.2154	0.4850	0.1132
	Group 2	0.1077	0.4876	—	0.1827	0.3142	0.2227
		0.5429	0.0500	—	0.3680	0.2120	0.1314
	Group 3	−0.0647	0.3060	−0.1777	—	0.1355	0.0421
		0.6835	0.2018	0.3813	—	0.5090	0.7622
	Group 4	−0.2022	0.1727	−0.3104	−0.1363	—	−0.0920
		0.2598	0.4813	0.2169	0.5057	—	0.5961
	Group 5	−0.1169	0.2616	−0.2148	−0.0417	0.0902	—
		0.3182	0.1023	0.1444	0.7640	0.6033	—
	Group 0	—	0.3713	−0.1111	0.0640	0.1979	0.1079
		—	0.0362	0.5286	0.6857	0.2685	0.3546
	Intercept	0.4546	0.3999	0.3452	0.3700	0.3962	0.3764
		0.3794	0.4365	0.4995	0.4692	0.4403	0.4631
	μ_1	24.5370	24.5337	24.5521	24.5309	24.5345	24.5355
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	μ_2	31.4511	31.4477	31.4557	31.4509	31.4486	31.4552
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	s_1	1.9793	1.9764	1.9786	1.9708	1.9752	1.9791
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	s_2	1.5562	1.5576	1.5530	1.5526	1.5572	1.5599
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	ρ_{12}	0.2554	0.2550	0.2571	0.2557	0.2558	0.2557
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	ν	17.1151	17.1042	16.9295	16.8143	17.0653	16.9971
		0.0050	0.0050	0.0046	0.0043	0.0049	0.0047

this new vision, we have examined the potential use of microbiome information in some variables associated with the insurance underwriting. Recent investigations have shown that changes in the gut microbiome are associated to certain risk of pathologies could be a potential proximal predictor of disease onset.

Recently, in an unpublished work by using text mining techniques in life insurance literature and microbiome research, a significant overlap between certain diseases and health conditions and other elements that are considered in insurance underwriting. One of these elements is the body mass index (BMI). This is one of the variables considered in the standard health declaration. Traditionally, this declaration is the first step in the risk assessment in health insurance underwriting practice. Certainly, depending on the level of insured capital and age of the policyholder, extra medical examination will be the obligatory required guarantee regardless of the outcome of the standard health declaration. However, medical examinations are expensive, disturbing for the applicant and time-consuming in the underwriting process (see [32]). The importance of BMI is linked to obesity that will lead to large number of chronic diseases, and consequently increase health expenditures and claims costs. Therefore, an early detection of obesity is crucial to safeguard the financial structure of the health insurance provider. Similar conclusions can be drawn about the early detection of cardiovascular, mental metabolic or immune diseases. Then, it is extremely important for the private health insurers to monitor their policyholders' health status in order to reduce future claims costs.

Table 9: Parameter estimates (first row) and p values (second row) for the regressors associated with the six predictors for individuals classified as nonsedentary (bivariate regression model).

		Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5	Predictor 6
Age	Group 1	−0.0256	—	0.1122	−0.0160	−0.0783	0.0123
		0.6443	—	0.1419	0.8252	0.3330	0.8009
	Group 2	−0.1632	−0.1279	—	−0.1313	−0.1980	−0.1032
		0.0076	0.0952	—	0.0478	0.0317	0.0493
	Group 3	−0.0324	0.0075	0.1344	—	−0.0650	0.0283
		0.5360	0.9174	0.0425	—	0.3906	0.5205
	Group 4	0.0353	0.0749	0.1917	0.0617	—	0.0913
		0.5767	0.3550	0.0371	0.4146	—	0.1690
	Group 5	−0.0546	−0.0205	0.1022	−0.0299	−0.0950	—
		0.1714	0.6748	0.0513	0.4980	0.1527	—
	Group 0	—	0.0420	0.1640	0.0323	−0.0328	0.0622
		—	0.4480	0.0072	0.5383	0.6037	0.1197
	Intercept	−0.2207	−0.1789	−0.1543	−0.1667	−0.1726	−0.1595
		0.1532	0.2471	0.3170	0.2802	0.2636	0.3015
	ϕ	2.9231	2.9215	2.9357	2.9269	2.9234	2.9294
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
BMI	Group 1	0.0183	—	0.1122	0.0182	0.1886	0.0139
		0.8906	—	0.1419	0.9198	0.3338	0.9090
	Group 2	−0.1063	−0.1419	—	−0.1120	0.0545	−0.1103
		0.4730	0.4582	—	0.4883	0.8052	0.3944
	Group 3	0.0070	−0.0286	0.1344	—	0.1682	0.0003
		0.9561	0.8746	0.0426	—	0.3537	0.9933
	Group 4	−0.1635	−0.1953	−0.0549	−0.1703	—	−0.1644
		0.2825	0.3168	0.8036	0.3474	—	0.2959
	Group 5	0.0037	−0.0281	0.1133	−0.0023	0.1680	—
		0.9687	0.8168	0.3815	0.9831	0.2856	—
	Group 0	—	−0.0344	0.1074	−0.0067	0.1623	−0.0054
		—	0.7963	0.4682	0.9576	0.2860	0.9546
	Intercept	0.2518	0.2081	0.2263	0.2352	0.2327	0.2471
		0.5380	0.6100	0.5790	0.5641	0.5689	0.5455
	μ_1	23.8809	23.8742	23.8818	23.8866	23.8793	23.8795
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	μ_2	31.2662	31.2593	31.2647	31.2716	31.2667	31.2640
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	s_1	1.9226	1.9203	0.9188	1.9169	1.9224	1.9238
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	s_2	1.4653	1.4656	1.4624	1.4634	1.4665	1.4655
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	ρ_{12}	0.1970	0.1968	0.1965	0.1968	0.1970	0.1967
		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	ν	114.9900	110.6990	98.4523	103.0250	110.6760	106.2180
		0.5978	0.5840	0.5393	0.5879	0.5816	0.5676

The main findings of our analysis show that the second bacterial coabundance group associated to *Ruminococcaceae-Bifidobacteriaceae* has a significant negative effect on the expected value of the response variable *Age* for the nonsedentary individuals. This issue is verified not only for the marginal model associated with the response variable age but also for the joint regression model. Concerning the fifth coabundance group related to *Pasteurellaceae*, it was observed a positive impact on the expected value of age for the sedentary individuals in the marginal model; in contrast, for some predictors, a negative impact in the mean of the response variable age is noticed under the bivariate regression model. This fact is somehow consistent with the recent work of Jollet et al. [33], where it is described a randomized clinical trial that shows in two of the bacterial groups of the *Pasteurellaceae* group are related to the level of physical activity. However, the degree that this conclusion is valid is arguable since, in general it appears that the majority of people have a unique baseline microbiome that is influenced by environmental conditions. The standard microbiome composition is affected not only by the level of physical activity but also for other factors such as age, diet, and medication. In addition, we have only analyzed a single observation of the gut microbiome per individual limited to the age interval 18–65. It should be highly recommended to perform a longitudinal analysis to relate microbiome information to risk mortality factors and probability of developing certain pathologies. For example, to deal with the high proportion of zeroes in the operational taxonomic units observed, a two part zero-inflated regression model with random effects could be used [26]. In this regard, the *American Gut Project* (a citizen science project containing more than 10,000 samples not only from the USA but also from several other countries around the world including Australia, UK, or Spain) represents an opportunity to access microbiome data for a variety of age group. This source of information together with the *Human Mortality Database* available in https://www.mortality.org could be used to analyze how changes in the gut microbiome are related to human longevity in different countries.

Data Availability

In our analyses, we use a dataset available in Dubey et al.'s [21] *LogMPIE* study. This dataset is freely accessible, and it may be downloaded from the European Nucleotide Archive (ENA) portal of the European Bioinformatics Institute (https://www.ebi.ac.uk/ena/data/view/PRJEB25642). Datasets are also included in the submission (Abundance.txt and Metadata.txt).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

E.C.O and G.L.C acknowledge Fundación MAPFRE for "Ignacio H. de Larramendi 2017" research grant that partially funded this work.

Supplementary Materials

Metadata: demographic information of the individuals. Table S1: microbiome abundance for each individual. (*Supplementary Materials*)

References

[1] G. S. Ginsburg and K. A. Phillips, "Precision medicine: from science to value," *Health Affairs*, vol. 37, no. 5, pp. 694–701, 2018.

[2] B. Chen and W. Xu, "Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures," *PLoS Computational Biology*, vol. 16, no. 9, Article ID c1008108, 2020.

[3] D. MacFabe, "Autism: metabolism, mitochondria, and the microbiome," *Global Advances in Health and Medicine*, vol. 2, no. 6, pp. 52–66, 2013.

[4] Y. Sanz, M. Olivares, A. Moya-Perez, and C. Agostoni, "Understanding the role of gut microbiome in metabolic disease risk," *Pediatric Research*, vol. 77, no. 1-2, pp. 236–244, 2015.

[5] B. Singh, N. Qin, and G. Reid, "Microbiome regulation of autoimmune, gut and liver associated diseases," *Inflammation and Allergy - Drug Targets*, vol. 14, no. 2, pp. 84–93, 2016.

[6] G. B. Stefano, R. Ptacek, J. Raboch, and R. M. Kream, "Microbiome: a potential component in the origin of mental disorders," *Medical Science Monitor*, vol. 23, pp. 3039–3043, 2017.

[7] H. Tilg and A. Kaser, "Gut microbiome, obesity, and metabolic dysfunction," *Journal of Clinical Investigation*, vol. 121, no. 6, pp. 2126–2132, 2011.

[8] R. L. Walker, H. Vlamakis, J. W. J. Lee et al., "Population study of the gut microbiome: associations with diet, lifestyle, and cardiometabolic disease," *Genome Medicine*, vol. 13, no. 1, p. 188, 2021.

[9] A. S. Meijnikman, O. Aydin, A. Prodan et al., "Distinct differences in gut microbial composition and functional potential from lean to morbidly obese subjects," *Journal of Internal Medicine*, vol. 288, no. 6, pp. 699–710, 2020.

[10] J. Bai, Y. Hu, and D. W. Bruner, "Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7-18 years old children from the American Gut Project," *Pediatric obesity*, vol. 14, no. 4, Article ID e12480, 2019.

[11] F. Valeriani, F. Gallé, M. S. Cattaruzza et al., "Are nutrition and physical activity associated with gut microbiota? a pilot study on a sample of healthy young adults," *Annali di Igiene: Medicina Preventiva e di Comunita*, vol. 32, no. 5, pp. 521–527, 2020.

[12] M. A. Stanislawski, D. Dabelea, L. A. Lange, B. D. Wagner, and C. A. Lozupone, "Gut microbiota phenotypes of obesity," *Npj-Biofilms and Microbiomes*, vol. 5, no. 1, p. 18, 2019.

[13] J. Aragón-Vela, P. Solis-Urra, F. J. Ruiz-Ojeda, A. I. Álvarez-Mercado, J. Olivares-Arancibia, and J. Plaza-Diaz, "Impact of exercise on gut microbiota in obesity," *Nutrients*, vol. 13, no. 11, p. 3999, 2021.

[14] S. Dua, M. Bhuker, P. Sharma, M. Dhall, and S. Kapoor, "Body mass index relates to blood pressure among adults," *North American Journal of Medical Sciences*, vol. 6, no. 2, pp. 89–95, 2014.

[15] L. Shamai, E. Lurix, M. Shen et al., "Association of Body Mass Index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese," *Obesity Surgery*, vol. 21, no. 1, pp. 42–47, 2011.

[16] K. M. Flegal, B. I. Graubard, D. F. Williamson, and M. H. Gail, "Cause-specific excess deaths associated with underweight, overweight, and obesity," *JAMA*, vol. 298, no. 17, p. 2028, 2007.

[17] K. Bhaskaran, I. Douglas, H. Forbes, I. dos Santos-Silva, D. A. Leon, and L. Smeeth, "Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults," *The Lancet*, vol. 384, no. 9945, pp. 755–765, 2014.

[18] C. Wang, J. Hu, M. J. Blaser, and H. Li, "Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data," *Bioinformatics*, vol. 36, no. 2, pp. 347–355, 2020.

[19] H. Zhang, J. Chen, Y. Feng, C. Wang, H. Li, and L. Liu, "Mediation effect selection in high-dimensional and compositional microbiome data," *Statistics in Medicine*, vol. 40, no. 4, pp. 885–896, 2021.

[20] D. Dumuid, T. E. Stanford, J. A. Martin-Fernández et al., "Compositional data analysis for physical activity, sedentary time and sleep research," *Statistical Methods in Medical Research*, vol. 27, no. 12, pp. 3726–3738, 2018.

[21] A. K. Dubey, N. Uppadhyaya, P. Nilawe et al., "LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing," *Scientific Data*, vol. 5, no. 1, 2018.

[22] K. Shestopaloff, M. D. Escobar, B. Graubard, and W. Xu, "Analyzing differences between microbiome communities using mixture distributions," *Statistics in Medicine*, vol. 37, no. 27, pp. 4036–4053, 2018.

[23] V. Jonsson, T. Osterlund, O. Nerman, and E. Kristiansson, "Modelling of zero-inflation improves inference of metagenomic gene count data," *Statistical Methods in Medical Research*, vol. 28, no. 12, pp. 3712–3728, 2019.

[24] B. Chen and W. Xu, "Functional response regression model on correlated longitudinal microbiome sequencing data," *Statistical Methods in Medical Research*, vol. 31, no. 2, pp. 361–371, 2022.

[25] J. J. Egozcue and V. Pawlowsky-Glahn, "Compositional data: the sample space and its structure," *Test*, vol. 28, no. 3, pp. 599–638, 2019.

[26] X. Yingling, J. Sun, and D. Chen, *Statistical Analysis Of Microbiome Data With R. ICSA Book Series in Statistics*, Springer Nature, Singapore, 2018.

[27] J. Aitchison, "The statistical analysis of compositional data," *Monographs on Statistics and Applied Probability*, Chapman and Hal, London, UK, 1986.

[28] J. Aitchison, "Principal component analysis of compositional data," *Biometrika*, vol. 70, no. 1, pp. 57–65, 1983.

[29] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[30] P. K. Dunn and G. K. Smyth, "Randomized quantile residuals," *Journal of Computational & Graphical Statistics*, vol. 5, no. 3, pp. 236–244, 1996.

[31] R. Zhang, C. Czado, and A. Min, "Efficient maximum likelihood estimation of copula based meta *t*-distributions," *Computational Statistics & Data Analysis*, vol. 55, pp. 1196–1214, 2011.

[32] M. Rothstein, *Genetics and Life Insurance: Medical Underwriting and Social Policy*, MIT Press, Boston, 2004.

[33] M. Jollet, K. Nay, A. Chopard et al., "Does physical inactivity induce significant changes in human gut microbiota? New answers using the dry immersion hypoactivity model," *Nutrients*, vol. 13, no. 11, p. 3865, 2021.