Relative Bacteria Abundance In Armpit
Project Principles of Statistical Data Analysis

Group 9: Jan Alexander[1], Paul Morbée [1], Steven Wallaert [1] and Joren Verbeke[1]

[1] *1/4 of effort for data analysis and reporting*

January 4, 2020

Contents

1 Introduction 1

2 Methods 1
- 2.1 Data intake 1
- 2.2 Relative abundance of genera 2
- 2.3 Relative abundance of species 2

3 Results 2
- 3.1 Results initial data analysis 2
- 3.2 Relative abundance of genera 3
- 3.3 Relative abundance of species 3

4 Conclusions and discussions 4

Abbreviations

Abbreviation	
BMI	Body mass index
CI	Confidence interval
IQR	Interquartile range
KW test	Kruskal-Wallis test
WMW test	Wilcoxon-Mann-Withney test
OR	Odds Ratio
RGA	Relative genus abundance
RSA	Relative species abundance
y.o.	year old

1 Introduction

Different bacteria can be isolated from an armpit. In this study, the association between abundance of malodor causing bacteria and a persons age was investigated. Additionally, potential associations between abundance of individual species were studied. A historical data set with relative abundance of 8 species in 40 volunteers was available. First, the data was read and explored. Secondly, a statistical protocol was written taking the objectives and data characteristics into account. Finally, the protocol was executed and results were discussed.

2 Methods

2.1 Data intake

2.1.1 Data cleaning

Data intake: The datafile `armpit.txt` was read.

Badly encoded genders: The gender was badly encoded for some observations due to trailing spaces, e.g. *'F '* was re-encoded to *'F'*.

Missing age values: One observation did not include a value for age (nor gender). This observation was omitted, as the objective of the analysis was to investigate an association between age and relative bacteria abundance.

Synthesising bacteria genera: The abundances of species within genera were summed to obtain the abundance for both of the genera.

Make age categories: The age was categorized by making two groups: '$age \leq 40y.o.$' and '$age > 40y.o.$'. The age of 40 years was chosen because the mean age of a Belgian is 40.8 years [1].

Species detection: Relative species abundance was also studied as a binary variable (detected or not). '$RSA = 0$' was considered as not detected whereas '$RSA > 0$' was considered as detection.

Data anomalies: It was verified that the relative quantities of both genera sum up to 100%.

After these operations, the final data-set used for the analysis in this document was obtained.

2.1.2 Initial data analysis

The team performing the analysis was not involved in the trial process, nor did we have any prior knowledge of microbiology. Before formulating a data analysis protocol, the presented dataset was inspected to gain an insight in its specificity. Basic statistics (mean, median, IQR, standard deviation, minimum and maximum) were calculated for relative genus and species abundance. Additionally, boxplots were used to inspect their distribution. The number of species detected per subject was plotted in a histogram. A histogram was made to investigate

[1] Census België, 2011 - `https://www.census2011.be/data/fresult/age-avg_nl.html`

the distribution of the age of the subjects. A frequency table was made to inspect the number of subjects per age, gender and BMI category.

2.2 Relative abundance of genera

The association between the patients age and the relative abundance of the malodor causing *'Corynebacterium'* genus in the armpit microbiome was investigated both with age as a categorized variable and with age as a continuous variable [2].

2.2.1 RGA as a of function of categorized age

Because the data deviate strongly from normality (see further) and the sample size was small, a WMW-test (rather than a t-test) was used to investigate a possible difference in *'Corynebacterium'* RGA distribution between the two age groups.

2.2.2 RGA as a function of age as a continuous variable

The association between RGA and age as a continuous variable was studied with the Kendall correlation coefficient. The Kendall correlation coefficient was calculated because the data contained many ties and the distribution showed a strong deviation from normality. The significance was estimated by comparing the observed Kendall τ with the exact permutation null distribution. Latter was obtained by calculating the Kendall τ of 10,000 random simulation runs with permuted age vector. A confidence interval was obtained through a bootstrapping procedure with 2000 replicates.

2.2.3 Influence by BMI or gender

Potential confounding by BMI or gender was investigated with KW-test. To investigate potential interactions, the WMW probability index, Kendall-τ correlation coefficient conditional on the BMI categories or the gender categories and their 95% confidence interval were estimated as described higher.

2.3 Relative abundance of species

The relative abundances of all 8 species[3] were investigated. Relative species abundance was studied both as a binary categorical variable (detected or not) and a continuous variable.

Binary approach: Odds ratios and Fisher exact tests were performed for each pair of species to investigate if the probability of detecting one species was independent from the detection of another species (H_0). The Fisher exact test was used instead of a χ^2-test, because expected cell counts in some 2x2 contingency tables were lower than 5 [4].

Quantitative approach: Kendall correlation coefficients were calculated and tested as described higher to explore if relative abundance of one species was correlated with abundance of another species.

Statistically significant was interpreted as $p \leq 0.05$ in this report.

The threshold level for statistical relevance was **not** corrected for the high number (28) of interaction tests. This choice was made to avoid inflation of test power, i.e. avoid increasing type II error probability. Due to the increased probability of a type I error, the results of these tests should be treated as a starting point for further testing.

3 Results

3.1 Results initial data analysis

BMI	Gender	$\leq$ 40 y.o.	>40 y.o.	Total
BMI $\leq$ 25	F	**14**	**4**	18
BMI $\leq$ 25	M	**6**	**2**	8
BMI > 25	F	**5**	**1**	6
BMI > 25	M	**2**	**5**	7
Total		27	12	39

Table 1: *Number of patients in each combination of categories*

The table in figure 6 contains the basic statistics of the relative distributions of the different bacteria species. The RGA shows a strongly skewed distribution. The relative Corynebacteria abundance is low ($\leq$ 3.2%) for at least half of the patients. Nevertheless, the armpit microbiome relative concentration of Corynebacteria can be as high as 97%.

Within the same genus, the skewness of the data is clear from the second part of the table. The *Staphylococcus* species 2, 3 and 4 are completely absent in at least half the samples. Nevertheless, the RSA of *Staphylococcus.2* can reach up to 89.7%.

Figure 1 shows that for no participant, all 8 species were detected.

Figure 2, shows the heavy skewness of the patient's age distribution. There are few older participants in the data set.

The distribution of the observations over the age, gender and BMI categories is shown in table 1.

[2] Since there are only two genera, the relative abundance of *'Staphylococcus'* genus is fully characterized with the relative abundance of *'Corynebacterium'* genus.

[3] 4 Corynebacterium species and 4 Staphylococcus species

[4] Introduction to categorical data analysis; - in Ysebaert M., Introductory Statistics, ICES Ghent University 2010-2011

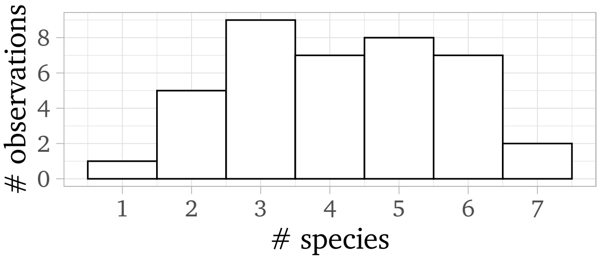
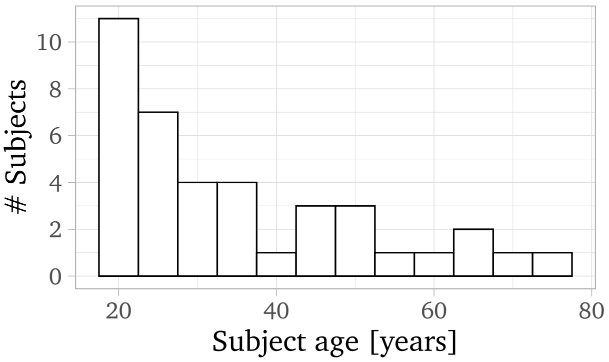


Figure 1: *Histogram of number of detected bacteria species in each test participant*



min	Q1	median	mean	Q3	max
19	22	30	35	43	76

Figure 2: *Histogram of age distribution*

3.2 Relative abundance of genera

3.2.1 Age as a categorical variable

The difference in RGA of corynebacteria can be observed in figure 3. For most people of at most 40 years old in this data set, a lower RGA (<10%) of corynebacteria is observed (with 4 exceptions). The distribution of RGA of corynebacteria for people over 40 years old shows a higher dispersion (IQR). The center of this distribution (both the median and the mean) is higher. The (two-sided) WMW-test indicates a probabilistic index for the Corynebacterium genus $P(RGA^C_{\leq 40} \leq RGA^C_{>40}) = 0.799$, $(CI_{95\%} = [0.605, 0.907])$ (p < 0.01), indicating that $H_0 : F_{\leq 40y.o.} = F_{>40y.o.}$ should be rejected.

3.2.2 Age as a continuous variable

Relative abundance of corynebacteria was positively correlated with higher age. A Kendall correlation coefficient $\tau = 0.26$ with a statistically significant p-value of $p < 0.025$ and a confidence interval: $CI_{95\%}\tau = [0.05, 0.48]$ was obtained.

3.2.3 Influence of BMI and Gender

When investigating the association between RGA of corynebacteria and age conditionally on gender or BMI, the WMW probability index and Kendall-τ correlation coefficient remained numerically higher than $1/2$ and 0, respectively (table 2). However, the CI cover a wider range and some were non significant. With the given sample size, interactions between age and BMI and between age and Gender can neither be proved nor disproved.

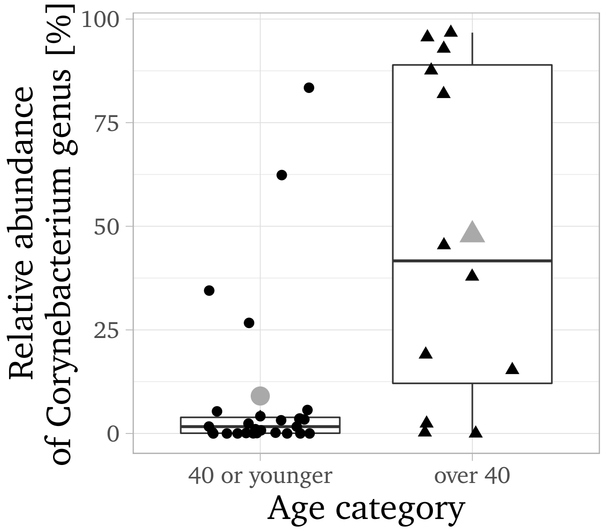


Figure 3: *Boxplot of corynebacteria abundance per age category. Black dots indicate individual observations. Grey markers indicate the mean for both groups.*

3.3 Relative abundance of species

In this part, the association between the abundance of the individual species was investigated using a binary and quantitative approach.

Binary approach: A bacterium species is detected if its measured abundance is > 0.0%. Odds ratios of detection between pairs of species are presented in figure 4. *Staphylococcus* 1 was not included in this analysis because it was detected in all participants [5]. Results indicate that detection of *Corynebacterium* 1, 2 and 3 was positively associated. Detection of any of these species was more likely when either were detected.

Quantitative approach: Kendall correlation coefficients between pairs of species are presented in figure 5. Results indicate that relative abundance of *Corynebacterium* 1, 2 and 3 were positively correlated with each other. The relative abundance of the same species were negatively correlated with relative abundance of *Staphylococcus* 1. The RSA of *Staphylococcus* 2 and 3 were negatively correlated and those of *Staphylococcus* 3 and 4 were positively correlated, though the correlations were not strong ($|\tau| \leq 0.26$).

As was discussed in 2.3, the colours of the ORs and correlations in figure 4 and figure 5 indicate whether $p < 0.05$ for each test or not, without correction for multiple testing.

[5] protocol amendment

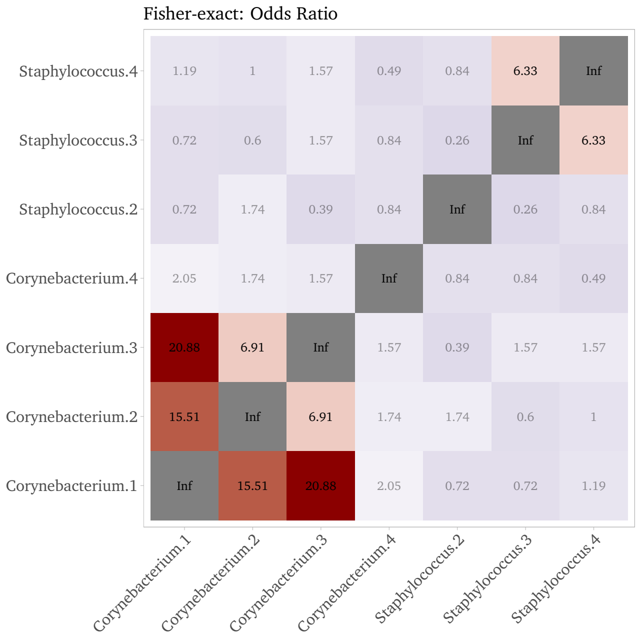


Figure 4: *Odds ratios, grey numbers were not significant according to a Fisher exact test*

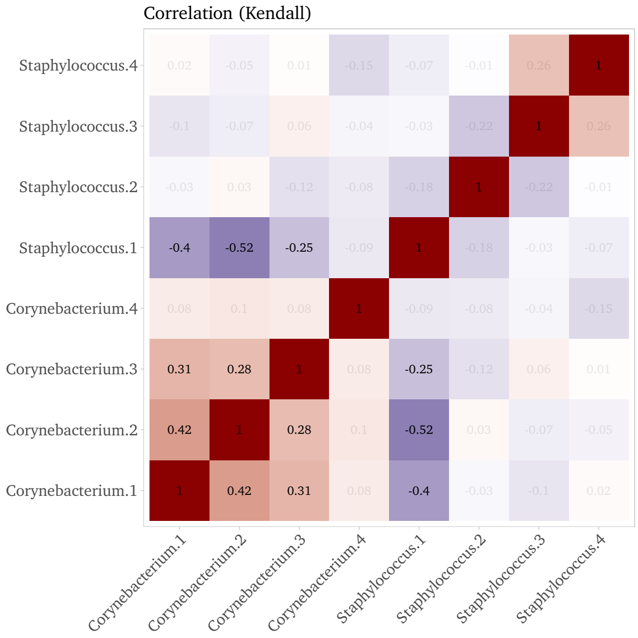


Figure 5: *Correlation plot, grey numbers indicate non-significant correlations*

4 Conclusions and discussions

In this study, the association between relative bacteria abundance in the armpit microbiome and age was studied. This study provides an indication that the relative abundance of the Corynebacterium genus tends to increase with age. No evidence for influence of BMI and gender on the relative genus abundance in the armpit microbiome could be found in the data set.

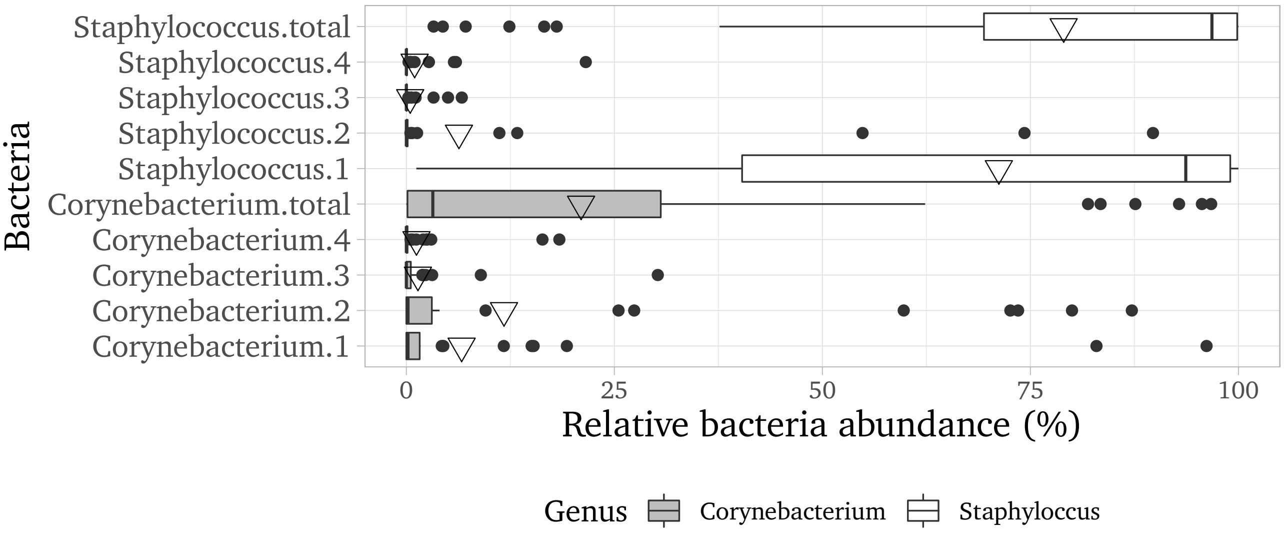
For the first research question, two approaches (age as a categorical vs. age as a continuous variable) were employed and yielded similar results. Yet, we want to remark that in the categorical case one is forced to make a choice on where to set the boundary between the different age groups whereas this extraneous intervention is not necessary in the continuous case. One also needs to make a choice on how many age groups to consider. Arguably, a method where arbitrary choices, with each choice carrying a potential impact on the results, don't need to be made is to be preferred over a method where these choices do need to be made. This is not to say there could be no reason to employ a categorical method, but the motivation to do this should be driven by theory, not by the data.

The data set which was highly biased towards younger individuals. Although biological insight can be gained, the results should not be generalized to the total population.

Associations between relative abundance of species were studied using Kendall correlation coefficient and Fisher exact test. Findings were highly similar indicating that correlations had limited added value compared to a dichotomous approach (detected or not) in this study. This could be explained by the fact that most species were only present in a limited proportion of subject (zero inflated data). Furthermore, correlations might be misleading as the abundance is expressed as relative values summed up to 100 procent. By default, abundance of the most prevalent species staphylococcus 1 is negatively correlated with abundance of the other species.

To maintain the test power, the p-values were not corrected for multiple testing. The obtained results should be subjected to further investigation, for example by conducting a new study focusing only on the currently found associations. Because of the focus in that study (limited number of tests) it will be more feasible to control the family wise error rate to 0.05, and thus would it give more insight into the validity of our results.

In conclusion, RGA of corynebacteria was associated with higher age and RSA of *Corynebacterium* 1, 2 and 3 were associated with each other in the study population.



Genus	mean [%]	median [%]	sd [%]	min [%]	max [%]	IQR [%]
Corynebacterium.total	21.0	3.2	33.0	0.0	96.7	30.4
Staphylococcus.total	79.0	96.8	33.0	3.3	100.0	30.4

Species	mean [%]	median [%]	sd [%]	min [%]	max [%]	IQR [%]
Corynebacterium.1	6.6	0.2	20.1	0.0	96.2	1.6
Corynebacterium.2	11.7	0.2	25.4	0.0	87.2	3.1
Corynebacterium.3	1.4	0.0	5	0.0	30.2	0.5
Corynebacterium.4	1.2	0.0	3.9	0.0	18.4	0.1
Staphylococcus.1	71.2	93.7	36.2	1.2	100	58.7
Staphylococcus.2	6.3	0.0	20.1	0.0	89.7	0.2
Staphylococcus.3	0.5	0.0	1.4	0.0	6.7	0.1
Staphylococcus.4	1.0	0.0	3.6	0.0	21.6	0.1

Figure 6: *Overview of base statistics of relative bacteria abundance*

$CI_{95\%}$	$P(RGA^C_{\leq 40} < RGA^C_{>40} \mid ...)$	$\tau(Age, RGA^C \mid ...)$
BMI >25	$[0.391, 0.931]$	$[-0.338, 0.539]$
BMI $\leq$ 25	$[0.611, 0.966]$	$[0.097, 0.565]$
M	$[0.493, 0.942]$	$[0.115, 0.767]$
F	$[0.495, 0.930]$	$[-0.178, 0.491]$
unconditional	[0.605, 0.907]	[0.05, 0.48]

Table 2: *Conditional $CI_{95\%}$ for both WMW probability index (categorical age) and Kendall τ correlation coefficient for RGA Corynebacterium and age as a continuous variable.*

Appendix

Influence of BMI and Gender

In 3.2.1, a difference between the relative genera abundance between the age groups was observed. This difference was not caused by confounding in the data set between age and another parameter such as BMI and gender. Based on a KW test, we found no evidence to reject H_0, that there is no difference in location between the 4 interaction groups of BMI and Gender (indicated in tabel 1):

1. BMI $\leq$ 25 & F: 18 observations.
2. BMI $\leq$ 25 & M: 8 observations.
3. BMI $>$ 25 & F: 6 observations.
4. BMI $>$ 25 & M: 7 observations.

Figure 7 shows box plots representing the difference in distribution of RGA for Corrynebacterium for both age groups where each age group is sub-devided by either BMI or gender. The similarity with figure 3 is clear.

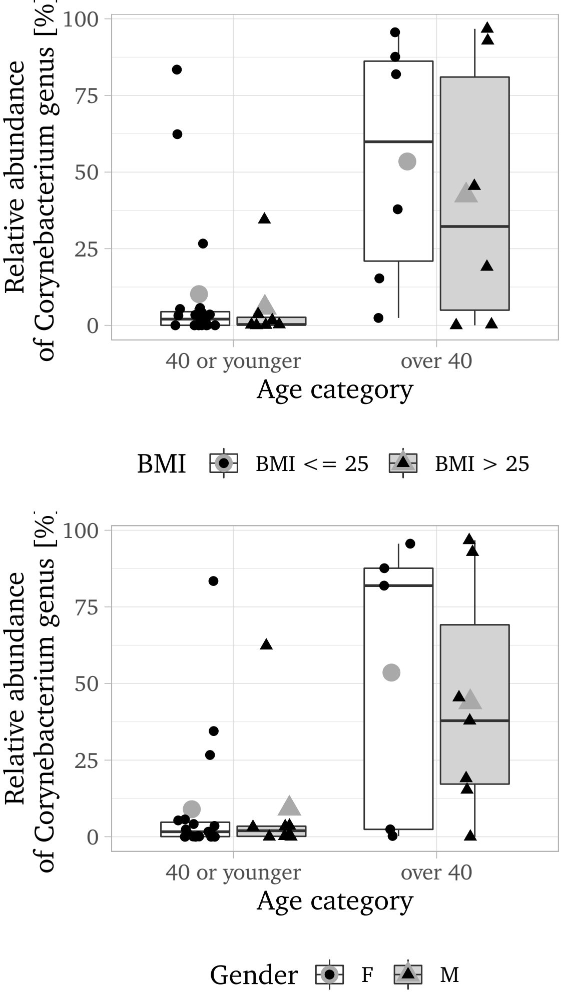


Figure 7: *Boxplot comparing BMI and age*

Fisher-Test odds ratio

To clarify, the ORs that were used in figure 4 were constructed in the following way: For each combination of 2 bacteria species, a table was constructed.

	B2	¬B2
B1	a	b
¬ B1	c	d

In this table, Bx indicates that bacterium x was observed. $\neg Bx$ indicates that is was not. a indicates the number of observations for which both bacteria *B1* and *B2* were detected, etc. The *odds ratio* is then:

$$OR = \frac{a/b}{c/d}$$

The statistical significance of these OR's was investigated with the Fisher-exact test.

As a side note, we would like to point out that `R` calculates the Maximum Likelihood Estimation of the odds ratio. This value, shown in figure 4, can differ slightly from the simple calculation given above.

Acknowledgement

Tables in this document where generated with the help of: `Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer`