

**Project J: Antibiotics associated diarrhea I**

**By:**

Name: Shahd Ayman Abdullah ID: 202002096

Name:Toka Mohamed Radwan ID: 20200228

Name: Laura Atef ID: 202001736

Name: Mohamed Yousry ID: 19104952

Under Supervision by Dr. Mohamed Maysara

**Table of Contents**

[Introduction 3](#_Toc19382)

[I Reading the Dataset 3](#_Toc19383)

[II Descriptive statistics 4](#_Toc19384)

[III Graphics 4](#_Toc19385)

[IV Outlier detection 8](#_Toc19386)

[V Testing for normality/ homoscedasticity 8](#_Toc19387)

[VI Statistical Inference 10](#_Toc19388)

[VII Hypothesis testing 11](#_Toc19389)

[VIII Linear model 12](#_Toc19390)

[IX Conclusion 14](#_Toc19391)

[X Contribution 15](#_Toc19392)

# Introduction

Antibiotics associated Diarrhea is a worldwide economically important disease in human. The disease is Caused by disturbance in the commensal bacteria that is followed by either colonization of pathogenic bacteria named Clostridioides difficile, referenced to as CDI (C. difficile infection) or diarrhea caused by the disturbance itself, referenced to as AAD (Antibiotics associated Diarrhea), or experienced no Diarrhea (ND).

To ad dress this a prospective clinical trial across 6 European countries (France, Netherland, Spain, Romania, Germany, Greece) was conducted and ~ 1000 patients were enrolled. The inclusion criteria were the fact they are CDI negative, haven’t taken Antibiotics prior to the enrolment. Patients received on of a three classes of Antibiotics and were sampled on Day1 (just before the antibiotics) and again on Day6, then they were monitored for 90 days and their stool samples (in case of diarrhea) were collected.

* A: PBL: penicillin + beta-lactamase inhibitor.
* OBL: other beta-lactamase antibiotics.
* FQN: fluoroquinolones.

The samples were sequenced using 16S microbial profiling and their microbial richness (Chao) and evenness (Shannon) and beta-diversity distance (between D1 and D6, Jclass) was assessed.

We are interested in the following outcomes:

* Changes in between the baseline samples across various countries, gender and age-stratas
* Changes over time (D1, D6 and Stool) for each antibiotic
* Changes across antibiotics over time

# Reading the Dataset

After reading the data by loading it in R we analyze it and found that the data contains eight columns which are:

1. **Patient.ID**
2. **Antibiotic.class (contains 3 Antibiotics "OBL ,FQN ,PBL” )**
3. **D1.Shannon.diversity**
4. **D6.Shannon.diversity**
5. **D1.Chao1.diversity**
6. **D6.Chao1.diversity**
7. **D1.D6.Jaccard.distance**
8. **Outcome**

# Descriptive statistics

1. Calculate the following: mean, median, minimum, maximum, first and third quartile

(for each variable), so we get them for continuous variable such as; D1.Shannon.diversity, D6.Shannon.diversity,D1.Chao.diversity,D6.Chao.diversity, D1.D6.Jaccard.distance**.**

1. For the categorical variable existing, calculate a frequency table; there is two categorical variables which is Antibiotic.Class ,Outcome.
2. Calculate the correlation coefficient (D1.Shannon.diversity and D6.Shannon.diversity) and (D1.Chao.diversity and D6.Chao.diversity), there is a positive relationship between (D1.Shannon.diversity and D6.Shannon.diversity) and (D1.Chao.diversity and D6.Chao.diversity) .

# Graphics

1. Generate a bar chart of a categorical variable for the outcome (AAD, CDI ,ND).

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Fig 1: we found that our data contains 25 AAD, 10 CDI and 300 ND from 335 patient.

1. Generate a bar chart graph with mean Outcome in BPL, FQN, OBL

A picture containing text, screenshot, diagram, design

Description automatically generated

Fig 2: mean outcome in FQN, OBL and PBL.

1. Make a histogram of a continuous variable: “D1 Chao” as well as “D6 Chao”.

A picture containing text, diagram, screenshot, technical drawing

Description automatically generated

Fig 3: the histogram visualization of D1 Chao shows that the distribution of data seems to be not normal (right skewed) but we will check normality using Shapiro test to check whether the data is really not normally distributed or normally distributed.

A picture containing diagram, text, screenshot, technical drawing

Description automatically generated

Fig 4: the histogram visualization of D1 Chao shows that the distribution of data seems to be not normal (right skewed) but we will check normality using Shapiro test to check whether the data is really not normally distributed or normally distributed.

1. Make a scatterplot of 2 continuous variables D1 Chao and D6 Chao, and add the regression lines for each antibiotic.

A picture containing screenshot, diagram, line, text

Description automatically generated

Fig 5: regression of OBL (green), PBL (blue) and FQN (red), the distribution of the points suggests a positive relationship D1 Chao and D6 Chao for each antibiotic.

1. Make a boxplot of Jaccard distance and a separate boxplots per Antibiotics (as factors).

A picture containing diagram, screenshot, text, rectangle

Description automatically generated

Fig 6: we found that there is variability between the Jaccard distance and Antibiotics.

# Outlier detection

**What do you think:**

After exploring the data for any existing outliers, we have outliers in D1.Shannon.diversity: 12 outliers, D6.Shannon.diversity: 26 outliers, D1.Chao.diversity: 9 outliers, D6.Chao.diversity: 3 outliers.

* Outliers are values in a dataset that are very different from the others. They could be due to measurement errors or unusual events. While outliers could affect an analysis negatively, it's not always necessary to remove them. Sometimes, outliers contain valuable information. Before deciding to remove outliers, it's important to evaluate their nature and impact on the analysis. Different statistical methods exist to deal with outliers. Ultimately, the decision to remove or keep outliers should be based on a good understanding of the data and the analysis objectives.

# Testing for normality/ homoscedasticity

**1) We check the normality using:**

**a - Shapiro test:**

* **D1.Shannon.diversity**: p-value = 4.134e-13
* **D6.Shannon.diversity**: p-value = 8.795e-13
* **D1.Chao1.diversity:** p-value = 2.489e-09
* **D6.Chao1.diversity:** p-value = 1.776e-06
* **D1.D6.Jaccard.distance:** p-value = 8.022e-05
* As the p-value is small, we can reject the null hypothesis (normally distributed) and accept the alternative hypothesis (not normally distributed) with a high level of confidence.

**b- Histogram:**

A picture containing text, diagram, screenshot, plot

Description automatically generatedA picture containing text, diagram, screenshot, line

Description automatically generatedA picture containing text, diagram, screenshot, line

Description automatically generatedA picture containing text, diagram, screenshot, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generated

**Second, we check the homoscedasticity:**

We checked the homoscedasticity using var test, box plot and bartlett test.

We found that all features are heteroscedastic as p-value is smaller than 0.05.

A picture containing text, diagram, screenshot, plan

Description automatically generatedA picture containing text, diagram, screenshot, technical drawing

Description automatically generatedA picture containing text, diagram, screenshot, line

Description automatically generated

# **Statistical Inference**

• **Calculate the 90%, 95%, 99% confidence interval for the means of ADWG0021per each gender.**

We calculate the confidence interval by this formula:

A picture containing font, white, number, symbol

Description automatically generated

* 90% confidence interval means that we are 90 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: FQN =0.6432289, OBL=0.6623997, PBL=0.6303106

Upper Interval: FQN =0.6646397, OBL=0.6821078, PBL=0.6537171

* 95% confidence interval means that we are 95 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: FQN =0.6401832, OBL=0.6595961, PBL=0.6269809

Upper Interval: FQN =0.6676855, OBL=0.6849114, PBL=0.6570468

* 99% confidence interval means that we are 99 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: FQN =0.6344462, OBL=0.6543154, PBL=0.6207092

Upper Interval: FQN =0.6734224, OBL=0.6901921, PBL=0.6633186

• **How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence (i.e. 99% C.I.)?**

When the confidence interval increases the width increases, so 99% confidence interval is narrow compared to 90% and 95% confidence interval.

# Hypothesis Testing

1. We hypothesis that Chao/Shannon at day 6 different from day1 for CDI. Assuming normality and homoscedasticity, **p-value > 0.05 (not significant). Chao diversity is not different in day 6 from day 1 for CDI outcome.**

**Testing normality:**

* 'CDI' "D1.Shannon.diversity: p-value 0.7488 (not significant; normally distributed)
* 'CDI' "D6.Shannon.diversity: p-value 0.1407 (not significant; normally distributed)
* 'CDI' "D1.Chao.diversity: p-value 0.636 (not significant; normally distributed)
* 'CDI' "D6.Chao.diversity: p-value 0.04323 (significant; NOT normally distributed)

**Homogeneity of variance test:**

* 'CDI' "D1.Shannon.diversity D6.Shannon.diversity: p-value = 0.05039 (not significant; variances are not different)
* 'CDI' "D1.Chao.diversity D6.Shannon.diversity: p-value = 0.7106 (not significant; variances are not different)

1. We hypothesis that Chao “different” in the group receiving BPL Antibiotics compared to the FQ antibiotics B. Assuming heteroscedasticity, **p-value = 0.228 (not significant). The two antibiotics are not different from one another on the level of chao diversity.**
2. We hypothesis that Chao is different between the different Antibiotics overtime. cComparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity).

**Chao for all antibiotics together: p-value = 2.306e-05 (significant)**

**Chao for FQN: p-value = 0.002091 (significant)**

**Chao for OBL: p-value = 0.004721 (significant)**

**Chao for PBL: p-value = 0.05336 (not significant)**

# Linear Model

1. Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above), the model predicts the outcome variable D6.Chao1.diversity from the predictor variables D1.Chao1.diversity and Antibiotic.class, the linear regression is the relationship between D1.Chao1.diversity, Antibiotic.class, and D6.Chao1.diversity. This is because the hypothesis tests suggest that there is a significant difference in Chao diversity between day 1 and day 6, and the linear regression is investigating whether the initial Chao diversity and the antibiotic class predict the change in Chao diversity from day 1 to day 6,

Residual standard error: 71.84 on 331 degrees of freedom.

Multiple R-squared: 0.1018, Adjusted R-squared: 0.09366

F-statistic: 12.5 on 3 and 331 DF, p-value: 9.148e-08

A picture containing text, screenshot, diagram

Description automatically generatedA picture containing text, screenshot, diagram, pattern

Description automatically generatedA picture containing text, diagram, plot, line

Description automatically generated

1. Calculate and interpret a 95% confidence interval of the regression slope,

The 95% confidence interval for the intercept is (82.2119038, 136.4873629), which means that we can be 95% confident that the true population intercept lies within this range.

The 95% confidence interval for the regression coefficient of D1.Chao1.diversity is (0.1816088, 0.3647027).

The 95% confidence interval for the regression coefficient of Antibiotic.classOBL is (-19.1627214, 27.2403198), which means that we cannot be 95% confident that the true population regression coefficient for Antibiotic.classPBL is different from zero,

the 95% confidence interval for the regression coefficient of Antibiotic.classPBL is (-4.1490992, 39.5413542).

1. Estimating the average CHAO change for with changing the Antibiotics, the difference in the predicted value of D6.Chao1.diversity between patients who received the PBL antibiotic and patients who received the OBL antibiotic is 13.65733, indicates that, on average, patients who received the PBL antibiotic had a higher predicted value of D6.Chao1.diversity compared to patients who received the OBL antibiotic, after controlling for the effect of D1.Chao1.diversity..
2. Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above), taking into account repeated measures!

While Holding D1.Chao1.diversity and the outcome variable constant, patients who received the antibiotic class OBL on average had a slightly lower Chao1 diversity value at day 6 (-0.468) than patients who received the reference antibiotic class,

patients who received the PBL antibiotic class had, on average, a greater Chao1 diversity value at day 6 (12.591) compared to patients who received the reference antibiotic class.

Patients who had the outcome variable CDI had, on average, a higher Chao1 diversity value at day 6 (20.612) compared to patients who had the reference outcome variable, while holding D1.Chao1.diversity and the antibiotic class constant, While holding D1.Chao1.diversity and the antibiotic class constant, patients with the outcome variable ND had, on average, an even greater Chao1 diversity value at day 6 (52.064) compared to patients with the reference outcome variable.

# Conclusion

In conclusion, this study aimed to investigate the impact of three different classes of antibiotics on the gut microbiota of patients. Different statistical techniques, such as descriptive statistics, normality and homoscedasticity tests, hypothesis testing, and linear regression, were used to analyses the data. The results showed that there were outliers in some of the variables, and the data was heteroscedastic. The findings of the hypothesis testing showed that there was no significant change in Chao diversity between the two antibiotics, BPL and FQ, or between day 6 and day 1 for CDI outcome. The three drugs' Chao diversity, however, changed dramatically over time. The linear regression model predicted the change in Chao diversity from day 1 to day 6 based on the initial Chao diversity and the antibiotic class. The study also found that the confidence interval width increased with higher confidence levels.

# Contribution

|  |  |  |
| --- | --- | --- |
| Name |  | Contribute |
| Toka Mohamed | 0. | Data reading |
|  | 1. | Descriptive statistics |
|  | 2. | Graphics |
|  | 3. | Outlier detection |
| Shahd Ayman | 4. | Testing for normality/ homoscedasticity  Statistical inference |
| Mohamed Yousry | 6. | Hypothesis testing |
| Laura Mostafa | 7. | Linear Model |