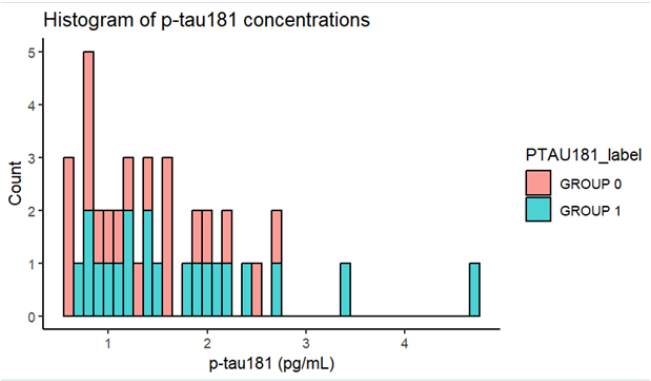
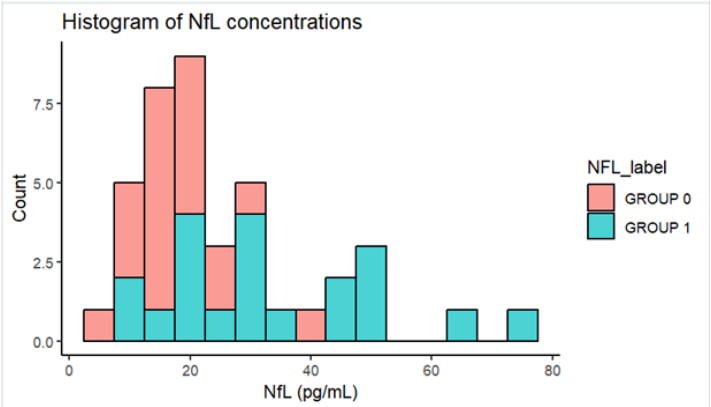
Task 1 Seminar 1 (Group A3)

Q1:

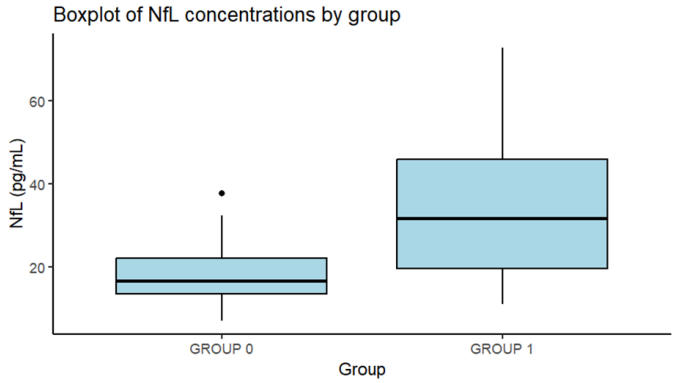
Explore the dataset visually and by carrying out appropriate statistical analyses.

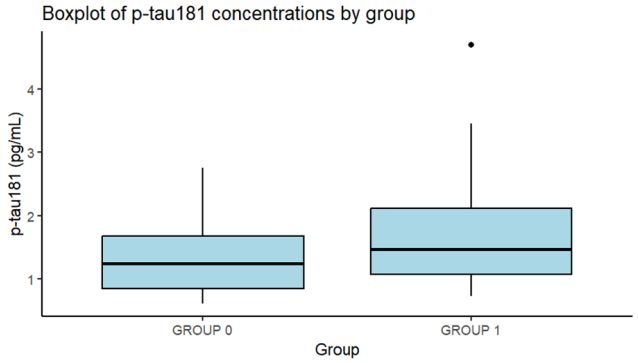
First we want to get a feel of the data and explore it visually. Since the sample size is low the visual representation is important because statistical tests lose power and p-values become less reliable.



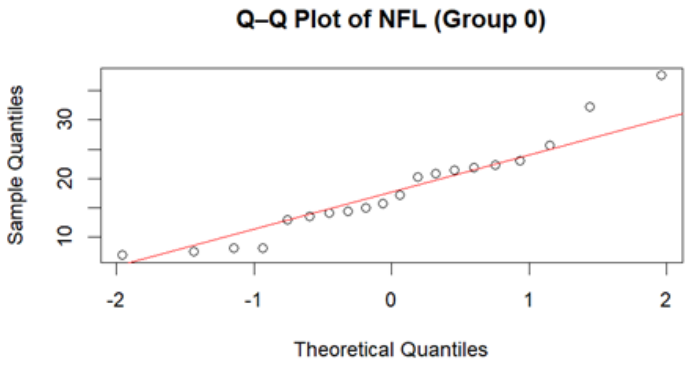


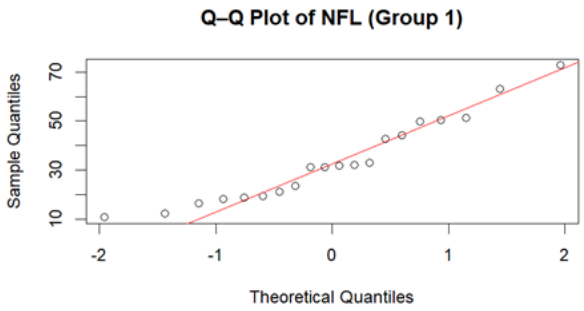
We first plot the biomarkers in a histogram. Here we see that the data looks right skewed and has some outliers.



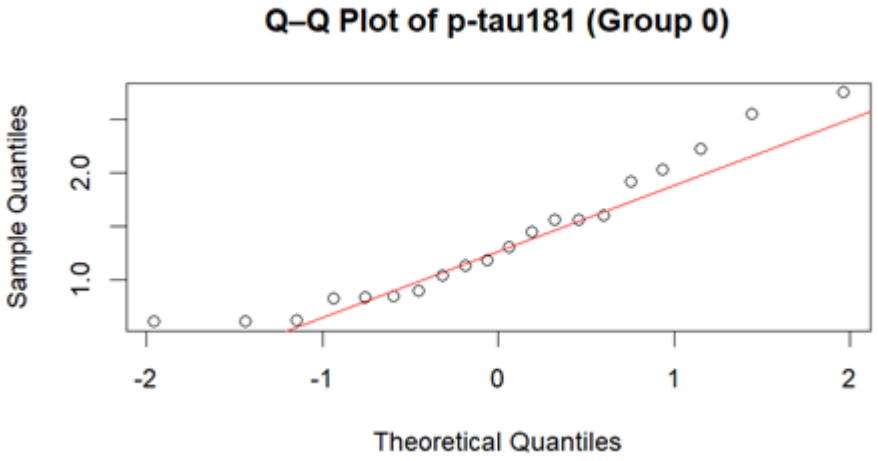


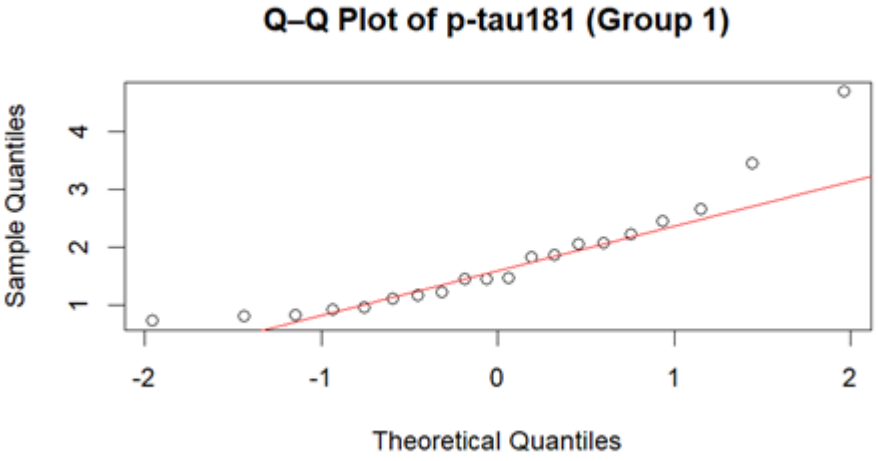
We then continue with plotting the data in boxplots. Which gives a representation of the data and how the values are distributed with the mean as a reference. This gives us an idea of the data. The p-tau181 groups look more asymmetric than NfL.





We go on to plot the data in QQ plots. We see that for the NfL data the data follows the line good, and we assume an approximate normality.





P-tau181 deviates substantially from the line, suggesting non-normality.

```
> shapiro_test_PTAU181_0 <- shapiro.test(df_plot[df_plot$GROUP == 0, ]$PTAU181)
> print(shapiro_test_PTAU181_0)

        Shapiro-Wilk normality test

data:  df_plot[df_plot$GROUP == 0, ]$PTAU181
W = 0.924, p-value = 0.1183


> shapiro_test_PTAU181_1 <- shapiro.test(df_plot[df_plot$GROUP == 1, ]$PTAU181)
> print(shapiro_test_PTAU181_1)

        Shapiro-Wilk normality test

data:  df_plot[df_plot$GROUP == 1, ]$PTAU181
W = 0.85896, p-value = 0.00756


> shapiro_test_NFL_0 <- shapiro.test(df_plot[df_plot$GROUP == 0, ]$NFL)
> print(shapiro_test_NFL_0)

        Shapiro-Wilk normality test

data:  df_plot[df_plot$GROUP == 0, ]$NFL
W = 0.93826, p-value = 0.2222


> shapiro_test_NFL_1 <- shapiro.test(df_plot[df_plot$GROUP == 1, ]$NFL)
> print(shapiro_test_NFL_1)

        Shapiro-Wilk normality test

data:  df_plot[df_plot$GROUP == 1, ]$NFL
W = 0.93541, p-value = 0.1961
```

To confirm, we performe the Shapiro–Wilk normality test:

- NfL (Group 0 & Group 1) P-tau181 (Group 0) : $p > 0.05$ → fail to reject normality assumption.

- P-tau181 (Group 1): $p < 0.05$ → reject normality.

Thus, NfL can reasonably be analyzed with a **t-test**, while P-tau181 requires either a non-parametric test (Mann–Whitney) or a log-transformation.

We have now looked at the data. NFL group 0, NFL group 1 are now assumed to be approximately normally distributed. Given the visualisation and the statistical test.
The p-tau181 biomarker has less of a normal distribution. Group 0 and Group 1 both follow the qqline poorly and Group 1 has a p-value lower than 0.05 in the shapiro-wilk normality test.

Moving on, we want to compare group 0 with group 1 for both biomarkers:

NFL analysis

For the NFL data we use the t-test:

```
> t_test_result <- t.test(df_plot[df_plot$GROUP == 0, ]$NFL, df_plot[df_plot$GROU
P == 1, ]$NFL)
> # Print the results
> print(t_test_result)

        Welch Two Sample t-test

data:  df_plot[df_plot$GROUP == 0, ]$NFL and df_plot[df_plot$GROUP == 1, ]$NFL
t = -3.721, df = 27.013, p-value = 0.0009209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -24.553374  -7.099626
sample estimates:
mean of x mean of y
  17.9620   33.7885
```

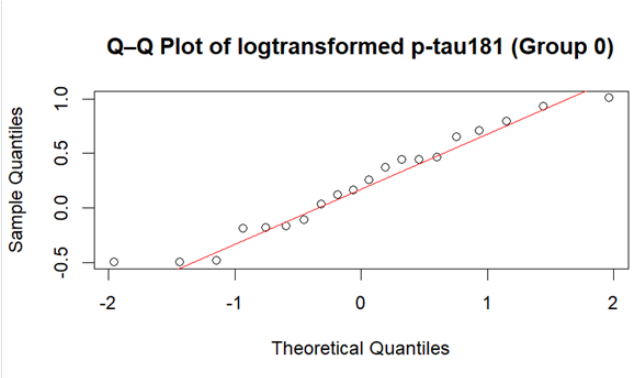
We get a p-value which is lower than 0.05 indicating that we can reject the null hypothesis. Meaning that the two groups have differing means.

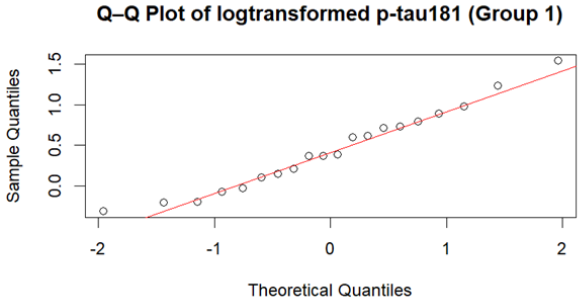
P-tau181 analysis

For the p-tau181 data we have two different things to do. Either we can use the willcoxon man whitney test, since it is non parametric meaning it does not assume a certain distribution. Or we can try to logtransform the data and see if it becomes approximately normally distributed making it possible to use the t-test.

Both tests have pros and cons. For this application we decide to use the t-test on the log data. We do this because the t-test is a more powerful test.

We start with logtransforming the data and do a qq-plot to confirm normality.





We can see that the data follows the line much better and we assume normality. For confirmation we do the shapiro test:

```
> shapiro.test(log_PTAU181_0)

        Shapiro-Wilk normality test

data:  log_PTAU181_0
W = 0.95544, p-value = 0.4573
```

```
> shapiro.test(log_PTAU181_1)

        Shapiro-Wilk normality test

data:  log_PTAU181_1
W = 0.97003, p-value = 0.7555
```

The p-values indicate that the data is normally distributed. We can now perform the t-test on the log data.

```
> t_test_result_PTAU181 <- t.test(log_data_PTAU181_0, log_PTAU181_1)
> # Print the results
> print(t_test_result_PTAU181)

        Welch Two Sample t-test

data:  log_data_PTAU181_0 and log_PTAU181_1
t = -1.4751, df = 37.808, p-value = 0.1485
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.54072530  0.08491306
sample estimates:
mean of x mean of y
0.2165314 0.4444375
```

The p-value is higher than 0.05 and we can not reject the null hypothesis. Indicating that the means of the data is similar and we can not distinguish between control and AD patients.

Q2:

What conclusions can you draw?

NfL:

- Two-sample t-test shows $p < 0.05$ → reject null hypothesis.
- Conclusion: mean NfL concentration is significantly higher in AD patients compared to controls.

P-tau181 (log-transformed):

- Log-transformation improved normality (confirmed by Q–Q plot and Shapiro–Wilk).
- Two-sample t-test on log(P-tau181): $p > 0.05$ → fail to reject null hypothesis.
- Conclusion: no significant difference in P-tau181 between groups, at least within this sample.

Q3:

How would you describe the statistical approach you are taking to solve this task? Discuss strengths and weaknesses of your approach.

Approach Taken

1. Exploratory visualization (histograms, boxplots, scatterplots, Q–Q plots).
2. Normality assessment (Shapiro–Wilk test, Q–Q plots).
3. Hypothesis testing:
    - Parametric t-test for NfL (normal).
    - Log-transform + t-test for P-tau181 (skewed).

Strengths

- Systematic workflow (EDA → normality check → hypothesis test).
- Transformation handled skewness and enabled use of a more powerful test.
- Results are straightforward to interpret for NfL.

Weaknesses

- Sample size is small (n=20/group), reducing statistical power and increasing risk of Type II error.
- Biomarkers considered separately, no combined model.
- Interpretability of log-transformed data: results need back-transformation to express differences on the original scale (e.g., as fold-change).
- A non-parametric approach (Mann–Whitney U) could have been a robustness check for P-tau181.

Conclusion

- NfL: significantly higher in AD patients, and thus shows promise as a biomarker.
- P-tau181: no significant group difference in this dataset. May require larger sample size, better measurement precision, or complementary biomarkers.