GSK/DigData Challenge 1

# Reading the data

library(tidyverse)

library(ggplot2)

# Import excel data and store in variables

clinical <- read\_csv("clinical-study.csv")

protein <- read\_csv("protein-levels.csv")

# Cleaning

When dealing with NA values, I decided to populate them with the mean. By doing so, it prevents the incumbent mean from being influenced or changed, which would occur if values other than the mean were to be inserted. Additionally, any missing weight values have been populated with an average which means that the original value were similar to the average value.

clinical <-  rename(

  clinical,

  participant\_id = subject\_id,

  treatment = trt\_grp,

  response = RESPONSE

)

protein <- rename(

  protein,

  concentration = protein\_concentration

)

clinical <- filter(clinical, age>=18)

clinical <- distinct(clinical, participant\_id, .keep\_all = TRUE)

# Replace with mean so that it does not introduce bias to incumbent mean.

clinical <- replace\_na(clinical, list(weight=91.39))

protein <- replace\_na(protein, list(concentration=121.7))

clinical <- mutate(clinical, clinical, response = as\_factor(response))

clinical <- mutate(clinical, clinical, BMI = weight/height\*\*2)

clinical <- left\_join(clinical, protein, by = "participant\_id")

# Aggregating

When comparing disparities among the drug and control treatment groups, I noticed that the average weight of the control group was higher, at a 1.1kg or 1.13% increase. Although it may seem insignificant, the disparity may still contribute to incorrect assumptions. On the other hand, the average age was almost identical, at a 0.1 difference in mean age.

On average, responders were 3.3kg heavier than non-responders which could suggest that Miraculon-A and the control drug are more effective among patients with a higher body weight. Both responders and non-responders had similar average ages, with a 0.1 difference, suggesting that it does not affect drug effectiveness.

In the drug group there were 210 responders whereas the control group had 124 responders, that is a 69.35% difference. This strongly suggests that Miraculon-A is over twice as effective as the control drug.

The protein concentration in responders was 105ug/L and 22% lower compared to non-responders. The data indicates that both drugs may be most effective when the protein concentration is around ~105ug/L.

# Checking patient age/weight disparity between the DRUG and CONTROL groups.

cmpr\_demo <- clinical %>%

  group\_by(treatment)

cmpr\_demo %>% summarise(average\_weight = mean(weight), average\_age = mean(age))

# Comparing mean weight and age in responders vs non responders

cmpr\_demo\_response <- clinical %>%

  group\_by(response)

cmpr\_demo\_response %>%

  summarise(average\_weight = mean(weight), average\_age = mean(age))

# Comparing response in treatment groups

cmpr\_responses <- clinical %>%

  filter(response=="Y")%>%

  group\_by(treatment)

cmpr\_responses %>%  summarise(no\_of\_responses = n())

# Comparing protein concentration between responders and non responders

cmpr\_protein <- clinical %>%

  group\_by(response) %>%

  summarise(avg\_concentration = mean(concentration))

# Visualizing

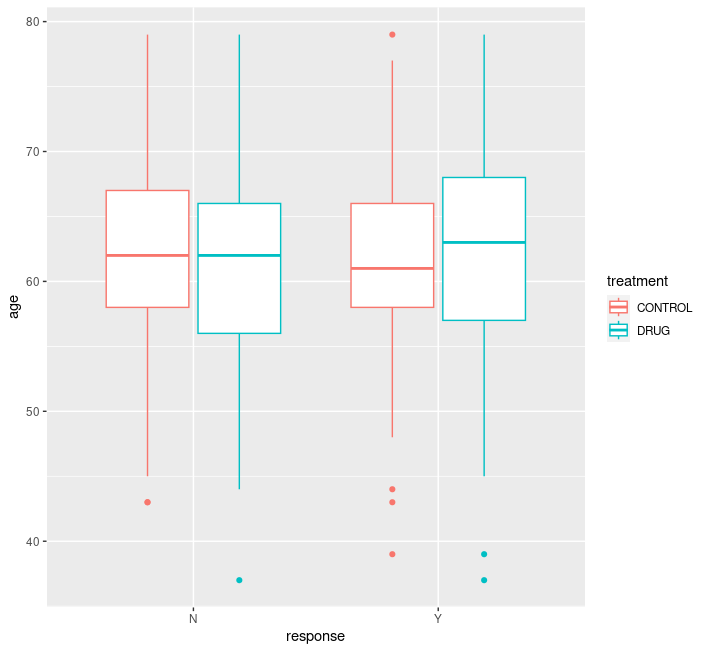
The box plot below shows the age difference between responders and non-responders and is coloured by its treatment group. The boxplot shows that non-responders had a similar median age suggesting that the efficacy of both drugs are not linked to age. However, the median age for a responder using Miraculon-A drug is noticeably higher compared to the control group’s median. This may suggest that GSK’s drug may be more effective among older patients.

Figure 1

ggplot(cmpr\_demo, aes(x=response, y=age, colour=treatment)) +

  geom\_boxplot() +

  geom\_smooth(method = "lm", se = FALSE)

Figure 2 shows the box plots of BMI for responders and non-responders. Non-responders had similar box plots. On average, responders in the control group had a slightly higher BMI than responders in the GSK drug group. However, the difference was very slight, and it is therefore reasonable to assume that BMI does not affect efficacy.

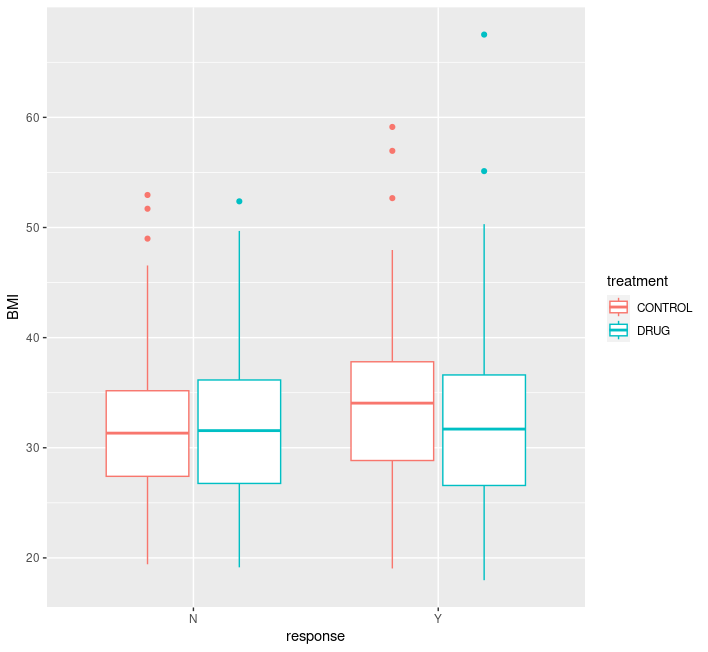


Figure 2

ggplot(cmpr\_demo, aes(x=response, y=BMI, colour=treatment)) +

  geom\_boxplot() +

  geom\_smooth(method = "lm", se = FALSE)

Figure 3 shows the protein concentration disparity among responders and non-responders: there is a significant difference between the two groups. On average, responders had a protein concentration of ~105 ug/L. As mentioned previously, protein concentration seems to be an important factor determining the efficacy of the drugs.

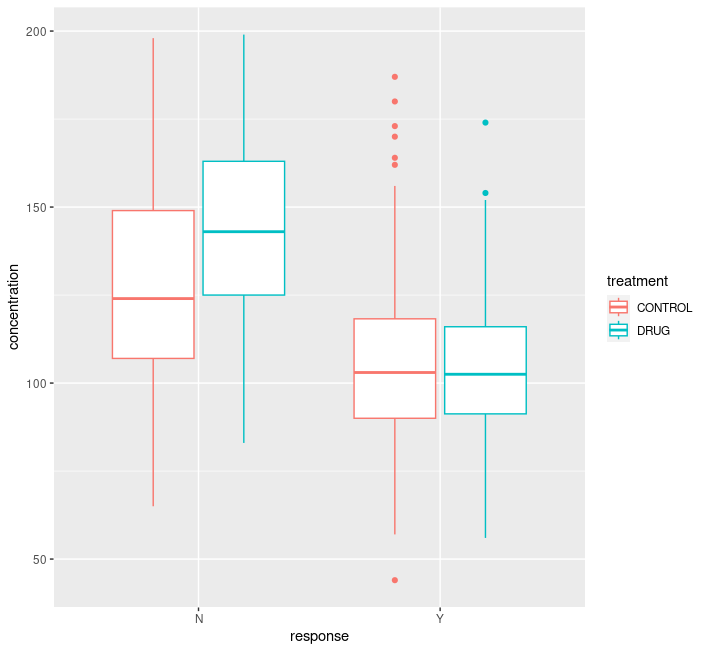


Figure 3

ggplot(cmpr\_demo, aes(x=response, y=concentration, colour=treatment)) +

  geom\_boxplot() +

  geom\_smooth(method = "lm", se = FALSE)

# Conclusion

In conclusion, a protein concentration of 105ug/L may improve the efficacy of the drug. Additionally, GSK’s drug Miraculon-A had over 60% more responses compared to the control group suggesting that the drug may be more effective than the control drug.