Metadata Figures

Elizabeth Anderson

7/22/2020

## Loading and Reviewing Metadata Info

To start off, we need to access the excel metadata that is found in this folder. It’s called ‘combinedMetadata.xlsx’

### Creating HMP Data Info

suppressMessages(library(httr))  
sample\_data <- read\_tsv('https://storage.googleapis.com/gbsc-gcp-project-ipop\_public/HMP/clinical\_tests/clinical\_tests.txt')  
GET('https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-019-1236-x/MediaObjects/41586\_2019\_1236\_MOESM3\_ESM.xlsx', write\_disk(tf <- tempfile(fileext = ".xlsx")))  
patient\_data <- read\_excel(tf, 1L)

### Summarizing Metadata

metadata\_df <- read\_excel('combinedMetadata.xlsx', sheet = 1)  
# Add summary statistics for the data (across cohorts and separately)  
skim(metadata\_df)

Data summary

|  |  |
| --- | --- |
| Name | metadata\_df |
| Number of rows | 616 |
| Number of columns | 47 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 47 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| dataset\_name | 0 | 1.00 | 3 | 8 | 0 | 3 | 0 |
| patientID | 103 | 0.83 | 9 | 19 | 0 | 513 | 0 |
| sampleID | 0 | 1.00 | 3 | 9 | 0 | 616 | 0 |
| Other Sample ID | 0 | 1.00 | 2 | 2 | 0 | 1 | 0 |
| body\_site | 0 | 1.00 | 2 | 5 | 0 | 2 | 0 |
| antibiotics\_current\_use | 0 | 1.00 | 2 | 2 | 0 | 2 | 0 |
| antibiotics\_family | 0 | 1.00 | 2 | 24 | 0 | 2 | 0 |
| study\_condition | 0 | 1.00 | 1 | 2 | 0 | 3 | 0 |
| disease | 0 | 1.00 | 1 | 9 | 0 | 7 | 0 |
| age | 0 | 1.00 | 2 | 18 | 0 | 161 | 0 |
| age\_category | 0 | 1.00 | 2 | 6 | 0 | 2 | 0 |
| gender | 0 | 1.00 | 2 | 6 | 0 | 3 | 0 |
| country | 0 | 1.00 | 3 | 3 | 0 | 12 | 0 |
| non\_westernized | 0 | 1.00 | 2 | 2 | 0 | 2 | 0 |
| number\_reads | 0 | 1.00 | 2 | 9 | 0 | 146 | 0 |
| number\_bases | 0 | 1.00 | 2 | 11 | 0 | 146 | 0 |
| minimum\_read\_length | 0 | 1.00 | 2 | 3 | 0 | 3 | 0 |
| median\_read\_length | 0 | 1.00 | 2 | 3 | 0 | 3 | 0 |
| hdl | 5 | 0.99 | 1 | 18 | 0 | 222 | 0 |
| triglycerides | 0 | 1.00 | 1 | 19 | 0 | 298 | 0 |
| hba1c | 0 | 1.00 | 1 | 18 | 0 | 105 | 0 |
| ldl | 3 | 1.00 | 1 | 19 | 0 | 269 | 0 |
| c\_peptide | 0 | 1.00 | 1 | 19 | 0 | 130 | 0 |
| cholesterol | 0 | 1.00 | 1 | 18 | 0 | 346 | 0 |
| glucose | 0 | 1.00 | 1 | 18 | 0 | 317 | 0 |
| adiponectin | 0 | 1.00 | 1 | 18 | 0 | 117 | 0 |
| insulin\_cat | 0 | 1.00 | 1 | 18 | 0 | 190 | 0 |
| fgf\_19 | 0 | 1.00 | 2 | 18 | 0 | 131 | 0 |
| hscrp | 0 | 1.00 | 1 | 19 | 0 | 127 | 0 |
| leptin | 0 | 1.00 | 2 | 18 | 0 | 129 | 0 |
| glutamate\_decarboxylase\_2\_antibody | 0 | 1.00 | 1 | 18 | 0 | 26 | 0 |
| creatinine | 0 | 1.00 | 2 | 3 | 0 | 48 | 0 |
| il\_1 | 0 | 1.00 | 1 | 19 | 0 | 33 | 0 |
| cd163 | 0 | 1.00 | 2 | 18 | 0 | 131 | 0 |
| glp\_1 | 0 | 1.00 | 1 | 18 | 0 | 44 | 0 |
| Curator | 0 | 1.00 | 2 | 2 | 0 | 1 | 0 |
| BMI | 0 | 1.00 | 2 | 18 | 0 | 433 | 0 |
| dyastolic\_p | 0 | 1.00 | 2 | 3 | 0 | 42 | 0 |
| systolic\_p | 0 | 1.00 | 2 | 3 | 0 | 58 | 0 |
| consented | 0 | 1.00 | 2 | 2 | 0 | 1 | 0 |
| IRIS | 0 | 1.00 | 2 | 7 | 0 | 4 | 0 |
| Steady-state.Plasma.Glucose(SSPG) | 0 | 1.00 | 2 | 6 | 0 | 60 | 0 |
| FPG | 0 | 1.00 | 2 | 6 | 0 | 44 | 0 |
| SSPG.Date | 0 | 1.00 | 2 | 10 | 0 | 57 | 0 |
| Ethnicity | 0 | 1.00 | 1 | 7 | 0 | 17 | 0 |
| Num\_All\_Visits | 0 | 1.00 | 1 | 3 | 0 | 27 | 0 |
| Study | 0 | 1.00 | 2 | 13 | 0 | 6 | 0 |

skim(metadata\_df, age)

Data summary

|  |  |
| --- | --- |
| Name | metadata\_df |
| Number of rows | 616 |
| Number of columns | 47 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

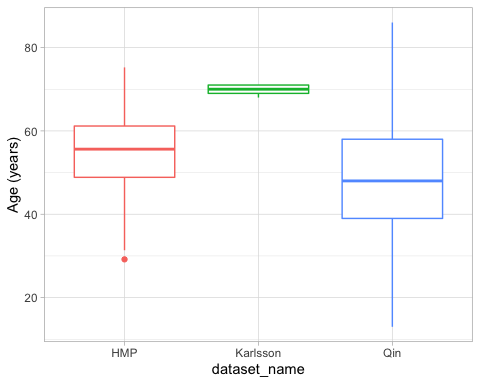
**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| age | 0 | 1 | 2 | 18 | 0 | 161 | 0 |

## Visualizing the Metadata

### Age

tibbleData = as\_tibble(metadata\_df)  
tibbleData <- tibbleData %>% filter(age != "NA")  
tibbleData$age <- as.numeric(tibbleData$age)  
  
ggplot(tibbleData, aes(x=dataset\_name, y=age)) +  
 geom\_boxplot(aes(color=dataset\_name)) +   
 theme\_light() +   
 ylab('Age (years)') +   
 theme(legend.position = "none")

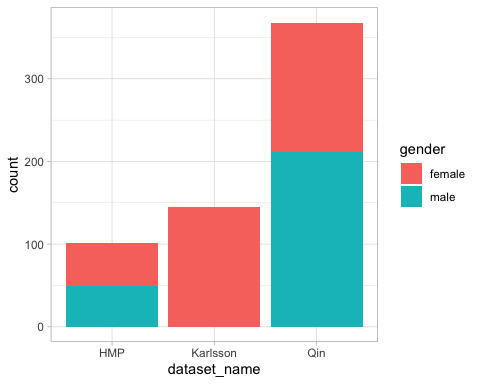


ggsave("age\_boxplot.png")

## Saving 5 x 4 in image

### Gender

tibbleData = as\_tibble(metadata\_df)  
tibbleData <- tibbleData %>% filter(gender != "NA")  
  
ggplot(tibbleData, aes(x=dataset\_name, fill = gender)) +  
 geom\_bar() +   
 theme\_light()

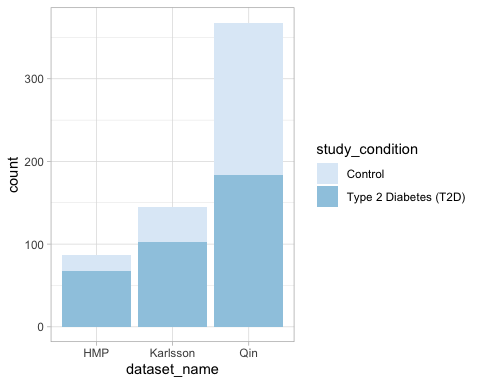


ggsave("gender\_barplot.png")

## Saving 5 x 4 in image

### Disease Condition

tibbleData = as\_tibble(metadata\_df)  
tibbleData <- tibbleData %>% filter(study\_condition != "NA")  
  
ggplot(tibbleData, aes(x=dataset\_name, fill = study\_condition)) +  
 geom\_bar() +   
 scale\_fill\_brewer(palette = "Blues",  
 labels = c("Control", "Type 2 Diabetes (T2D)")) +  
 theme\_light()

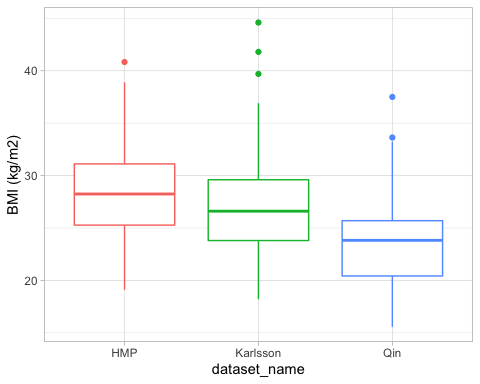


ggsave("disease\_barplot.png")

## Saving 5 x 4 in image

### Body-Mass Index (BMI)

tibbleData = as\_tibble(metadata\_df)  
tibbleData <- tibbleData %>% filter(BMI != "NA")  
tibbleData$BMI <- as.numeric(tibbleData$BMI)  
  
ggplot(tibbleData, aes(x=dataset\_name, y=BMI)) +  
 geom\_boxplot(aes(color=dataset\_name)) +   
 theme\_light() +   
 ylab('BMI (kg/m2)') +   
 theme(legend.position = "none")

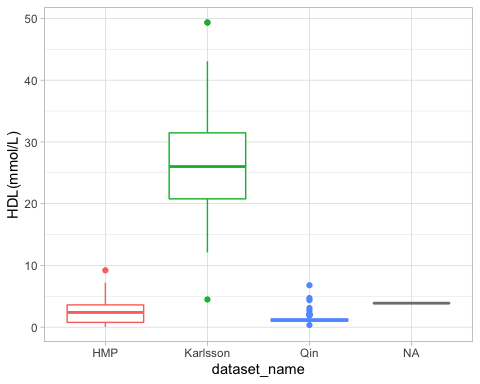


ggsave("BMI\_boxplot.png")

## Saving 5 x 4 in image

### HDL

metadata\_df <- read\_excel('combinedMetadata.xlsx', sheet = 2)  
tibbleData = as\_tibble(metadata\_df)  
tibbleData <- tibbleData %>% filter(hdl != "NA")  
tibbleData$hdl <- as.numeric(tibbleData$hdl)  
  
ggplot(tibbleData, aes(x=dataset\_name, y=hdl)) +  
 geom\_boxplot(aes(color=dataset\_name)) +   
 theme\_light() +   
 ylab('HDL(mmol/L)') +   
 theme(legend.position = "none")



ggsave("HDL\_boxplot.png")

## Saving 5 x 4 in image