Statistical Analysis of CYP-GUIDES Clinical Data

## Hello!

I wanted to showcase my R statistical analysis skills using a Kaggle clinical trial dataset. This is a **work in progress!** Please see my work below :)

# Data Description:

Source: [Kaggle](https://www.kaggle.com/datasets/shashwatwork/clinical-dataset-of-the-cypguides-trial/data)

### Context

CYP-GUIDES (Cytochrome Psychotropic Genotyping Under Investigation for Decision Support) is a randomized controlled trial (RCT) comparing 2 outcomes in hospitalized patients with severe depressive disorders treated according to the patient’s CYP2D6 genotype and functional status versus standard psychotropic therapy. The primary outcome was hospital Length of Stay (LOS) and the secondary outcome was the Re-Admission Rate (RAR) 30 days after discharge.

The trial setting was the Institute of Living at Hartford Hospital. CYP2D6 genotyping was implemented to characterize the functional status of the CYP2D6 enzyme with sub-normal, normal, or supra-normal function. The electronic medical record (EMR) was utilized to transmit clinically actionable drug prescribing guidance based on the patient’s CYP2D6 function to the physician.

### Content

The RCT recruited 1500 patients, genotyped CYP2D6 in 1459, and randomized 477 to standard therapy (Group S), for whom treatment-as-usual guidance was delivered without consideration of patient CYP2D6 genotype, and 982 to genetically-guided therapy (Group G) where CYP2D6-based treatment recommendations were provided via EMR to physicians. For inpatients in Group G whose CYP2D6 function was sub- or supra-normal, medications primarily metabolized by the CYP2D6 enzyme were proscribed.

The RCT developed a database of potential benefits to the field. The pharmacologic, clinical course, and pharmacogenetic therapeutic guidance is being published in a related article. These data should enable various investigators to assess effeeffects of clinical decision support on resource utilization and psychotropic therapy during psychiatric hospitalizations.

## Acknowledgements

Tortora, Joseph; Robinson, Saskia; Baker, Seth; Ruaño, Gualberto (2020), “Clinical Dataset of the CYP-GUIDES Trial”, Mendeley Data, V1, doi: 10.17632/25yjwbphn4.1 Source

### Summary Statistics of Data:

Table 1. Summary Statistics of the demographics of the CYP-GUIDES study

| **Variable** | **Category** | **Count** |
| --- | --- | --- |
| Sex | Female | 769 |
| Sex | Male | 731 |
| Age | Young Adults (18-24) | 379 |
| Age | Early Mid-Adulthood (25-39) | 414 |
| Age | Late Adulthood (40-64) | 641 |
| Age | Elderly Adults (65+) | 66 |
| Race | White | 864 |
| Race | Black | 182 |
| Race | Hispanic/Latino | 382 |
| Race | Other/Unknown | 72 |

## Is the LOS Significantly Different Between the Two Treatments?

Below is R code that performs statistical analyses on the LOS between the two groups, patients who were genotyped and treated accordingly, and patients who were not genotyped and given standard treatment. A descriptive statistics table of the mean LOS of the two groups is produced.

df <- read\_csv("Dataset.csv") # Read in dataset  
  
desc.stats <- df %>%  
 filter(!is.na(Assignment)) %>% # Filter all Assignments that are NA  
 mutate(Assignment = case\_when(  
 Assignment == "G" ~ "Genetically-guided therapy",  
 Assignment == "S" ~ ("Standard Therapy")  
 )) %>%  
 group\_by(Assignment) %>%  
 summarise(avg\_LOS = mean(LOS), sd\_LOS = sd(LOS), n = n()) %>% # Find mean LOS, SD LOS, and count of each group type  
 mutate(se = sd\_LOS / sqrt(n)) %>% # Find standard error between mean of each group  
 mutate(lower\_CI = avg\_LOS - 1.96 \* se,  
 upper\_CI = avg\_LOS + 1.96 \* se) %>% # Calculate upper and lower bounds of 95% confidence interval of mean  
 select(Assignment, avg\_LOS, lower\_CI, upper\_CI) %>%  
 rename(`Mean LOS` = avg\_LOS,  
 `Lower CI` = lower\_CI,  
 `Upper CI` = upper\_CI  
 )  
  
caption <- "Table 2. Descriptive statistics of the CYP-GUIDES study"  
tbl <- format\_output(desc.stats,caption)  
FitFlextableToPage(tbl)

Table 2. Descriptive statistics of the CYP-GUIDES study

| **Assignment** | **Mean LOS** | **Lower CI** | **Upper CI** |
| --- | --- | --- | --- |
| Genetically-guided therapy | 178.5112 | 166.8220 | 190.2004 |
| Standard Therapy | 172.6080 | 159.1926 | 186.0233 |

With this table, we can write our Hypotheses:

Null Hypothesis: There is no difference between the length of stay between the two therapies.

Alternative Hypothesis: There is a difference between the length of stay between two therapies.

We will now conduct a Welch’s T-Test to determine if there is a significant difference between the mean LOS between the two therapies.

p\_value <- df %>%  
 mutate(p\_value = t.test(LOS ~ Assignment, var.equal = FALSE)$p.value) %>%  
 distinct(p\_value)  
print(paste0(p\_value))

## [1] "0.515656957537065"

We have found that the p-value of the T-test is 0.515657, and can accept that the null hypothesis is true. There is no significant difference between the LOS in the two groups.

## ANOVA Testing for LOS between Therapeutic Guidances

There are three types of Therapeutic Guidances. I’m wondering if the descriptive statistics will be the similar to the descriptive statistics between the two types of therapies.

df <- read\_csv("Dataset.csv") # Read in dataset  
  
desc.stats2 <- df %>%  
 rename(therp.guidance = `Therapeutic Guidances`) %>%  
 filter(!is.na(therp.guidance)) %>%  
 group\_by(therp.guidance) %>%  
 summarise(avg\_LOS = mean(LOS), sd\_LOS = sd(LOS), n = n()) %>% # Find mean LOS, SD LOS, and count of each group type  
 mutate(se = sd\_LOS / sqrt(n)) %>% # Find standard error between mean of each group  
 mutate(lower\_CI = avg\_LOS - 1.96 \* se,  
 upper\_CI = avg\_LOS + 1.96 \* se) %>% # Calculate upper and lower bounds of 95% confidence interval of mean  
 select(therp.guidance, avg\_LOS, lower\_CI, upper\_CI) %>%  
 rename(`Mean LOS` = avg\_LOS,  
 `Lower CI` = lower\_CI,  
 `Upper CI` = upper\_CI,  
 `Therapy Guidance` = therp.guidance  
 )  
  
caption <- "Table 3. Descriptive statistics of the CYP-GUIDES study"  
tbl <- format\_output(desc.stats2,caption)  
FitFlextableToPage(tbl)

Table 3. Descriptive statistics of the CYP-GUIDES study

| **Therapy Guidance** | **Mean LOS** | **Lower CI** | **Upper CI** |
| --- | --- | --- | --- |
| CT | 179.3257 | 163.6169 | 195.0345 |
| EG | 176.5100 | 165.1102 | 187.9098 |
| TL | 161.3182 | 128.4926 | 194.1438 |

Again, with this table, we can write our hypotheses:

Null Hypothesis: There is no difference between the length of stay between the different Therapy Guidances.

Alternative Hypothesis: There is a difference between the length of stay between different Therapy Guidances.

Let’s see the variance between the three therapeutic groups:

df <- read\_csv("Dataset.csv") # Read in dataset  
  
aov.test <- df %>%  
 filter(!is.na(`Therapeutic Guidances`))  
  
result <- aov(LOS ~ `Therapeutic Guidances`, data = aov.test)  
summary(result)

## Df Sum Sq Mean Sq F value Pr(>F)  
## `Therapeutic Guidances` 2 18341 9170 0.297 0.743  
## Residuals 1456 44893897 30834

The F-value between the Therapeutic Guidances is 0.297, which suggests that the differences in the LOS between the types of therapeutic guidances are small compared to the variability within the groups. The p-value is 0.743, which indicates that the differences in LOS are not statistically significant. Based on this analysis, there is no evidence to suggest that the therapeutic guidance statistically affects the length of stay for patients.

## Power Analysis

Something to consider in further studies is a Power Analysis. The Power Analysis will help us see the sample size needed to know that regardless of the p-value we get, we used enough data to make a good decision. I’m thinking of making another .Rmd for it though.

## Patient-Readmittance Based on Treatment and Therapeutic Guidances

To be continued…

df <- read\_csv("Dataset.csv") # Read in dataset  
  
desc.stats2 <- df %>%  
 filter(!is.na(Assignment)) %>% # Filter all Assignments that are NA  
 mutate(Assignment = case\_when(  
 Assignment == "G" ~ "Genetically-guided therapy",  
 Assignment == "S" ~ ("Standard Therapy")  
 ),  
 RAR = case\_when(  
 RAR == 0 ~ ("Not Re-Admitted"),  
 RAR == 1 ~ ("Re-Admitted")  
 )) %>%  
 group\_by(Assignment,RAR) %>%  
 summarize(Count = n()) %>%  
 ungroup() %>%  
 mutate(Percent = (Count/sum(Count) \* 100))  
   
  
caption <- "Table 2. Summary statistics of the CYP-GUIDES study"  
tbl <- format\_output(desc.stats2,caption)  
FitFlextableToPage(tbl)

Table 2. Summary statistics of the CYP-GUIDES study

| **Assignment** | **RAR** | **Count** | **Percent** |
| --- | --- | --- | --- |
| Genetically-guided therapy | Not Re-Admitted | 883 | 60.520905 |
| Genetically-guided therapy | Re-Admitted | 99 | 6.785469 |
| Standard Therapy | Not Re-Admitted | 434 | 29.746402 |
| Standard Therapy | Re-Admitted | 43 | 2.947224 |