*Thank you for taking the time to complete our biostatistics skills assessment. Please note you are* ***not*** *expected to be an expert in all of these questions, but please complete to the best of your ability. To complete the assessment, please follow these guidelines:*

* *Use any resources, such as books, mentors, and google (In practice, these resources plus our team and I are available for guidance and support)*
* *Feel free to record any comments or questions – it may be helpful to write your thought process*
* *Definitions of the variables in the simulated data sets are listed at the end of this document*
* *Use all the time you need within the 48-hour timeframe to complete the assessment*

1. Using the *Encounter\_Dates* data, create a new variable called AGE to specify the patient’s age at each encounter
   1. Copy and paste code

enc\_data <- read\_csv("Encounters\_data.csv", show\_col\_types = FALSE)

head(enc\_data, n=10)

```{r}

#mdy data format

enc\_data$BIRTH\_DATE <- mdy(enc\_data$BIRTH\_DATE)

enc\_data$ENC\_DATE <- mdy(enc\_data$ENC\_DATE)

#create age at each encounter date

enc\_data$AGE <- round(as.numeric(difftime(enc\_data$ENC\_DATE, enc\_data$BIRTH\_DATE, units = "days"))/365.25)

head(enc\_data)

* 1. Specify the 5-point summary of the AGE variable (i.e. min, max; 25th, 50th, and 75th percentiles)

Min. 1st Qu. Median Mean 3rd Qu. Max.

30.00 38.00 43.00 44.08 51.00 58.00

1. Using the *Encounter\_Dates* data, create a new variable called RACE\_ETH. The RACE\_ETH variable should be coded as:

‘Non-Hispanic Black’ if RACE is Black and ETHNICITY is Non-Hispanic;

‘Non-Hispanic White’ if RACE is White and ETHNICITY is Non-Hispanic;

‘Non-Hispanic Other’ if RACE is Other and ETHNICITY is Non-Hispanic;

‘Hispanic’ if ETHNICITY is Hispanic.

* 1. Copy and paste code

table(enc\_data$RACE)

table(enc\_data$ETHNICITY)

encounter <- enc\_data %>%

mutate(RACE\_ETH = case\_when(

RACE=="Black" & ETHNICITY == "Non-Hispanic"~0,

RACE=="White" & ETHNICITY == "Non-Hispanic"~1,

RACE=="Other" & ETHNICITY == "Non-Hispanic"~2,

ETHNICITY == "Hispanic" ~ 3))

encounter$RACE\_ETH <- factor(encounter$RACE\_ETH,levels = 0:3,

labels = c("Non-Hispanic Black", "Non-Hispanic White",

"Non-Hispanic Other", "Hispanic"))

table(encounter$RACE\_ETH)

* 1. State the frequency of each Race/Ethnicity (RACE\_ETH) out of the 100 encounters

Non-Hispanic Black Non-Hispanic White Non-Hispanic Other Hispanic

22 35 15 28

1. Using the *Encounter\_Dates* data, create a new variable called Depression, which documents if depression is present at the time of encounter.
   1. Copy and paste code

encounter$Depression <- if\_else(grepl("depression",encounter$COMORBID, ignore.case = TRUE),"Yes","No")

table(encounter$Depression)

* 1. State the frequency of patients who had Depression at their first encounter

No = 72

Yes = 28

1. Choose **one** of the following (a or b) below:
   1. There are 7 unique patients in the *Encounter\_Dates* data. Calculate (via coding) the count of encounters by patient
      1. Copy and paste code

encounter\_counts <- encounter %>% group\_by(ID) %>% summarise(encounter\_count = n())

print(encounter\_counts)

* + 1. Paste/submit your table of counts

| **ID**  <chr> | **encounter\_count**  <int> |
| --- | --- |
| PAT 1 | 16 |
| PAT 2 | 13 |
| PAT 3 | 12 |
| PAT 4 | 10 |
| PAT 5 | 15 |
| PAT 6 | 22 |
| PAT 7 | 12 |

* 1. There are 7 unique patients in the *Encounter\_Dates* data. Create a variable called READMISSION indicating whether each encounter was within 10 days from the patient’s first encounter
     1. Copy and paste code
     2. Specify the distribution (e.g. number of yes’s and no’s) of READMISSION

1. Suppose a physician requests an analysis to compare systolic blood pressure in adult patients (18 years and older) prescribed Ketamine versus those prescribed Etomidate from January 1, 2018 to December 31, 2019 at the hospital. She hypothesizes those prescribed Ketamine will have lower systolic blood pressure compared to Etomidate.

Note that systolic blood pressure is known to range from 70 to 180 mmHg in this population. Use the *Ketamine\_SBP\_2018\_2019 and the Ketamine\_Covariate* data to answer the following questions. Both datasets correspond to the same patient cohort.

* 1. Write the PICO/ PECO (T) for this research project (as best you can with the information provided)

PECO

Population = Adult patients 18 years and older

Exposure = Ketamien Prescription

Comparison = Etomidate Prescription

Outcome = Systolic Blood Pressure

Time = January 1, 2018 to December 31, 2019

* 1. Using the data provided and the inclusion/exclusion criteria specified above, compute an unadjusted effect estimate to address the physician’s request. Please include all pre-processing steps (assumptions, missing data, etc). (Note: you could use a regression model to find an unadjusted estimate or other methods)
     1. Copy and paste code

#Load the datasets

```{r}

ketaminesbp <- read.csv("Ketamine\_SBP\_2018\_2019.csv")

ketaminecovariate <- read.csv("Ketamine\_Covariate.csv")

```

#merge datasets

```{r}

ket\_data <- merge(ketaminesbp, ketaminecovariate, by = "ID")

```

#Summary statistics

```{r}

summary(ket\_data)

```

#Missing Values

```{r}

#Total MV

sum(is.na(ket\_data))

#Columns missing values

colSums(is.na(ket\_data))

# complete case data

naniar::pct\_complete\_case(ket\_data)

```

#Removing missing cases

```{r}

ket\_complete <- ket\_data %>% drop\_na()

#recheck missingness

sum(is.na(ket\_complete))

dim(ket\_complete)

```

#Factoring Variables

```{r}

ket\_complete$MEDICATION <- as.factor(ket\_complete$MEDICATION)

ket\_complete$COMORB\_A <- ifelse(grepl("N", ket\_complete$COMORB\_A),"No","Yes")

ket\_complete$COMORB\_A <- as.factor(ket\_complete$COMORB\_A)

ket\_complete$COMORB\_B <- as.factor(ket\_complete$COMORB\_B)

ket\_complete$COMORB\_C <- as.factor(ket\_complete$COMORB\_C)

ket\_complete$HYPOTENSION <- ifelse(ket\_complete$HYPOTENSION=="0", "No", "Yes")

ket\_complete$HYPOTENSION <- as.factor(ket\_complete$HYPOTENSION)

ket\_complete$PAYER <- as.factor(ket\_complete$PAYER)

ket\_complete$SEX <- as.factor(ket\_complete$SEX)

ket\_complete$RACE\_ETH <- as.factor(ket\_complete$RACE\_ETH)

summary(ket\_complete)

```

#Removing Outliers in SBP and Age

In Minimum is -3.37 and maximum is 999. So as in Age, minimum is 17, this dataset is for adults 18 years and older.

```{r SBP outliers & Age}

#Under 70 mmHg

under70 <- ket\_complete %>% filter(SBP <70) %>% (n=nrow)

under70

#Over 180 mmHg

over180 <- ket\_complete %>% filter(SBP >180) %>% (n=nrow)

over180

#Filtering those with < 70 and > 180

ket\_complete <- ket\_complete %>% filter(SBP >= 70 & SBP <= 180 )

dim (ket\_complete)

#Under 18

under18 <- ket\_complete %>% filter (AGE < 18) %>% (n=nrow)

under18

#Filtering only 18 and above

ket\_complete <- ket\_complete %>% filter(AGE >= 18)

dim(ket\_complete)

```

#Assessing normality with Histogram

```{r histogram sbp}

#Histogram of SBP

sbp\_his <- ggplot(ket\_complete, aes(x = SBP)) +

geom\_histogram(binwidth =2.5, color = "black", fill = "white")+

geom\_vline(aes(xintercept = mean(SBP)), color = "red", linewidth = 0.8)+

geom\_vline(aes(xintercept = median(SBP)), color = "blue", linewidth = 0.8, linetype = "dashed")+

xlab("Systolic Blood Pressure of Patients") +

ylab("Count") +

ggtitle ("Histogram of Systoli Blood Pressure of Patients")

sbp\_his

```

# normality and variance test

```{r}

nortest::ad.test(ket\_complete$SBP)

ggqqplot(ket\_complete$SBP, main = "Q-Q Plot of SBP with 95% Confidence Interval")

car::leveneTest(SBP ~ MEDICATION, data = ket\_complete)

```

#T-test

```{r}

t.test(SBP ~ MEDICATION, data = ket\_complete)

```

* + 1. Describe data pre-processing

First, the two datasets pertaining to ketamine are imported into R and subsequently merged into a unified dataframe. The nature of the data is explored using the summary function, enabling identification of missing values, outliers, and the classification of variables. Subsequently, missing values are examined and records with missing data are eliminated, justified by the assertion that the dataset is predominantly complete with only 2% missingness. Variables are then transformed into categorical factors. Outliers in systolic blood pressure (SBP) and age are addressed and removed. Following these preprocessing steps, the data is deemed ready for analysis.

To assess the normality assumption, histograms, Anderson-Darling tests, and Q-Q plots are employed. Homogeneity of variance is evaluated using Levene’s Test. Results indicate that both normality and equal variance assumptions are met. Subsequently, a t-test is conducted to ascertain the mean difference in SBP between patients administered Ketamine and those given Etomidate.

* + 1. State and interpret the results.

Welch Two Sample t-test

data: SBP by MEDICATION

t = 9.9534, df = 555.55, p-value < 2.2e-16

alternative hypothesis: true difference in means between group Etomidate and group Ketamine is not equal to 0

95 percent confidence interval:

8.139712 12.142224

sample estimates:

mean in group Etomidate mean in group Ketamine

108.39921 98.25824

There is a statistically significant difference between patients who was prescribed Ketamine and Etomidate. The mean systolic blood pressure of Ketamine patients is lower than Etomidate patients. The mean SBP of group Ketamine is 98.2582 mmHg whereas the mean SBP of Etomidate group is 108.2992 with t = 9.9534 and p-value <0.05.

* 1. Patients’ sex is thought to confound the relation between Ketamine and systolic blood pressure. Repeat the analysis while accounting for sex as a confounding variable.
     1. Copy and paste code

#Fitting regression models

```{r unadjusted model}

m0 <- lm(SBP ~ MEDICATION, data = ket\_complete)

summary(m0)

```

#Sex adjusted model

```{r}

m1 <- lm(SBP ~ MEDICATION + SEX ,data = ket\_complete)

summary(m1)

```

* + 1. State and interpret the results

Call:

lm(formula = SBP ~ MEDICATION + SEX, data = ket\_complete)

Residuals:

Min 1Q Median 3Q Max

-31.059 -7.883 -0.595 7.804 64.141

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 111.0491 0.8821 125.896 < 2e-16 \*\*\*

MEDICATIONKetamine -10.2663 0.9978 -10.289 < 2e-16 \*\*\*

SEXMale -4.9954 0.9985 -5.003 7.58e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.78 on 555 degrees of freedom

Multiple R-squared: 0.1879, Adjusted R-squared: 0.1849

F-statistic: 64.19 on 2 and 555 DF, p-value: < 2.2e-16

The mean SBP of female patients who used Etomidate is 111.0491 mmHg. When others are held constant, patients who received Ketamine have on average 10.2663 mmHg lower SBP compared to patients who received Etomidate. When others are held constant, male patients on average have 4.9954 mmHg lower SBP than female patients. These results are all statistically significant with p-vale less than 2.2e-16. Overall, the model suggests that both medication and sex are significant predictors of SBP after adjusting for each other, with male and patients with Ketamine having lower average SBP.

* 1. Suppose that the physician presents a new outcome to measure. Instead of the numeric blood pressure outcome, the outcome is binary indicating whether or not the patient experiences a readmission. A logistic regression model was used to assess the unadjusted effect of Ketamine on readmission. Please interpret the results based on the model information below. Specifically, how do you explain the results (Table 1) to a non-statistician?

Logistic Regression Model:

, where

, and

for readmission; 0 otherwise, and

for Ketamine; 0 otherwise.

|  |  |  |
| --- | --- | --- |
| **Table 1: Model Results for** | | |
| Coefficient (1) | 95% Confidence Interval | p-value |
| 0.47 | (0.32, 0.56) | <0.001 |

The logistic regression model result from Table 1 provides information about the relationship between receiving ketamine medication and the likelihood of readmission to the hospital. Simply, patients who received ketamine medication may have a higher likelihood of being readmitted to the hospital compared to patients who received other medications. The coefficient shows 0.47 increase in the log odds of experiencing readmissions and it is also statistically significant with p-value less than 0.001.

1. Choose one of the following (a or b). Based on the Ketamine analysis from 5 above:
   1. Suppose the physician requests an additional analysis. She would like to know which independent variables are predictors of hypotension. Propose a statistical method to address this aim. (Include any thoughts, limitations, and justification)

In this case, the outcome variable is hypotension which is binary variable. Therefore, a suitable statistical method for this analysis would be logistic regression and the predictors could be either continuous or categorical variables like age, weight, sex and race-ethnicity. I would use forward stepwise regression with hypotension being the outcome variable and putting each of the rest of the variables in the regression model. Then, the model’s goodness of fit would be assessed by likelihood ratio test, ROC curves. However, logistic regression has its own assumptions such as independence of observations, linearity of logit, little or no multicollinearity, large enough sample size, binary outcome variable, and no outliers. Some of these assumptions can be assessed and handled using variance inflation factors (VIF) and cook’s distance.

* 1. Suppose the physician returns later with new data. Now, each patient’s systolic blood pressure is measured 1-3 times, but they remain within the same medication group (e.g. if the patient is on Ketamine at baseline, they remain on Ketamine throughout the study period). She again aims to evaluate systolic blood pressure in patients prescribed Ketamine compared to those prescribed Etomidate. Propose an analysis which accounts for the repeated measures. (Include any thoughts, limitations, and justification)

**Variable Definitions**

|  |  |
| --- | --- |
| **Table 1: Definitions of variables in *Encounter\_Dates* data** | |
| **Variable Name** | **Variable Definition** |
| **ID** | **Unique patient identifier** |
| **BIRTH\_DATE** | **Patient’s date of birth (mm/dd/yyyy)** |
| **RACE** | **Patient’s race** |
| **ETHNICITY** | **Patient’s ethnicity** |
| **ENC\_DATE** | **Patient’s date of hospital encounter (mm/dd/yyyy)** |
| **COMORBID** | **Patient’s comorbidities at time of encounter** |
| **Any missing data is denoted by “NA”** | |

|  |  |  |
| --- | --- | --- |
| **Table 2: Definitions of variables in *Ketamine\_SBP\_2018\_2019* data** | |  |
| **ID** | **Unique patient identifier** | **Character/Factor** |
| **MEDICATION** | **Medication prescribed to patient** | **Factor with 2 levels** |
| **WEIGHT** | **Patient’s weight in pounds** | **Numeric (pounds)** |
| **COMORB\_A** | **Receipt of unspecified comorbidity** | **Binary (yes/no)** |
| **COMORB\_B** | **Receipt of unspecified comorbidity** | **Binary (yes/no)** |
| **COMORB\_C** | **Receipt of unspecified comorbidity** | **Binary (yes/no)** |
| **SBP** | **Systolic blood pressure** | **Numeric (mmHg)** |
| **HYPOTENSION** | **Receipt of hypotension (SBP<90)** | **Binary (yes/no)** |
| **Any missing data is denoted by “NA” or 999.** | | |

|  |  |  |
| --- | --- | --- |
| Table 3: Definitions of variables in *Ketamine\_Covariate* data | |  |
| ID | Unique patient identifier | Character/Factor |
| PAYER | Payer/insurance | Factor with 3 levels |
| SEX | Patient’s sex | Factor with 2 levels |
| RACE\_ETH | Patient’s race/ethnicity | Factor with 4 levels |
| AGE | Patient’s Age | Numeric (17—99) |
| Any missing data is denoted by “NA” or 999. | | |