Statistical testing potential covariates

Annika Jorgensen

2023-12-11

PURPOSE: This document is to conduct statistical testing to see if any variables in the metadata are significantly different between males and females. For variables deemed significant further investigation will be done via MDS analysis and survival analysis. These variables may be incorportated as covariates in our DEA.

METHODS: Using R functions, statistical testing will be done on variables found in the metadata. The following variables will be testing

– Age – Tumor Stage – Nodes – Metadata – Edmondson Grade – Tumor stage in mm – Liver Fibrosis – Alcohol Intake – Smoking – Prognosis – Survival Month

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

Here the full metadata is read in as a dataframe titled "metadata" and a curated metadata set is titled "metadatasmall"

```
#reading in Full meta dataset
metadata <- read.table("~/Desktop/ResearchProjects/LiverCancer/Metadata/Full_RIKEN_clinical_data.csv", 
```

```
#reading in meta dataset
metadataSmall <- read.csv("~/Desktop/ResearchProjects/LiverCancer/DEA_removed_samples/MedicalCovariateA
```

Subsetting metadata such that analyses can be conducted. The curated metadata set only has the samples in the analysis that we want to analyze. However, the curated metadata doesn't contain the full metadata. So, the full metadata set was subsetted by matching the sample names in the curated dataset for only the samples that we want to study.

```
#subsetting full metadata for samples used in analysis
metadataAnalysis <- metadata[rownames(metadata) %in% metadataSmall$ID,]

#subsetting male samples
Male <- metadataAnalysis[which(metadataAnalysis$Gender== "M"),]

#subsetting female samples
Female <- metadataAnalysis[which(metadataAnalysis$Gender== "F"),]
```

F and T test age of males and females

Age is a numerical variable so a two-sample t test will be used to identify if the variable is statistically significant. To figure out the type of t test we want to use an f test is employed to determine whether or not the variances between the two datasets are equal or not. A significance level of 0.05 is used.

```
#subsetting age of male samples
MaleAge<- Male$Age

#subsetting age of female samples
FemaleAge<- Female$Age

#f test alpha=0.05
var.test(MaleAge, FemaleAge) #Fail to reject null no evidence variance different
```

```
##
##  F test to compare two variances
##
## data:  MaleAge and FemaleAge
## F = 0.72044, num df = 107, denom df = 41, p-value = 0.1851
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4179384 1.1700975
## sample estimates:
## ratio of variances
##          0.7204416
```

```
#variances equal
```

The p value is 0.1851 so therefore we fail to reject the null hypothesis. We will conduct a t test with equal variance.

```
#t test alpha=0.05
t.test(MaleAge,FemaleAge,var.equal=TRUE) #Fail to reject null mean age of males and females not signifi
```

```
##
##  Two Sample t-test
##
## data:  MaleAge and FemaleAge
## t = -1.556, df = 148, p-value = 0.1218
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.5037621  0.7736034
## sample estimates:
## mean of x mean of y
##  65.61111  68.47619
```

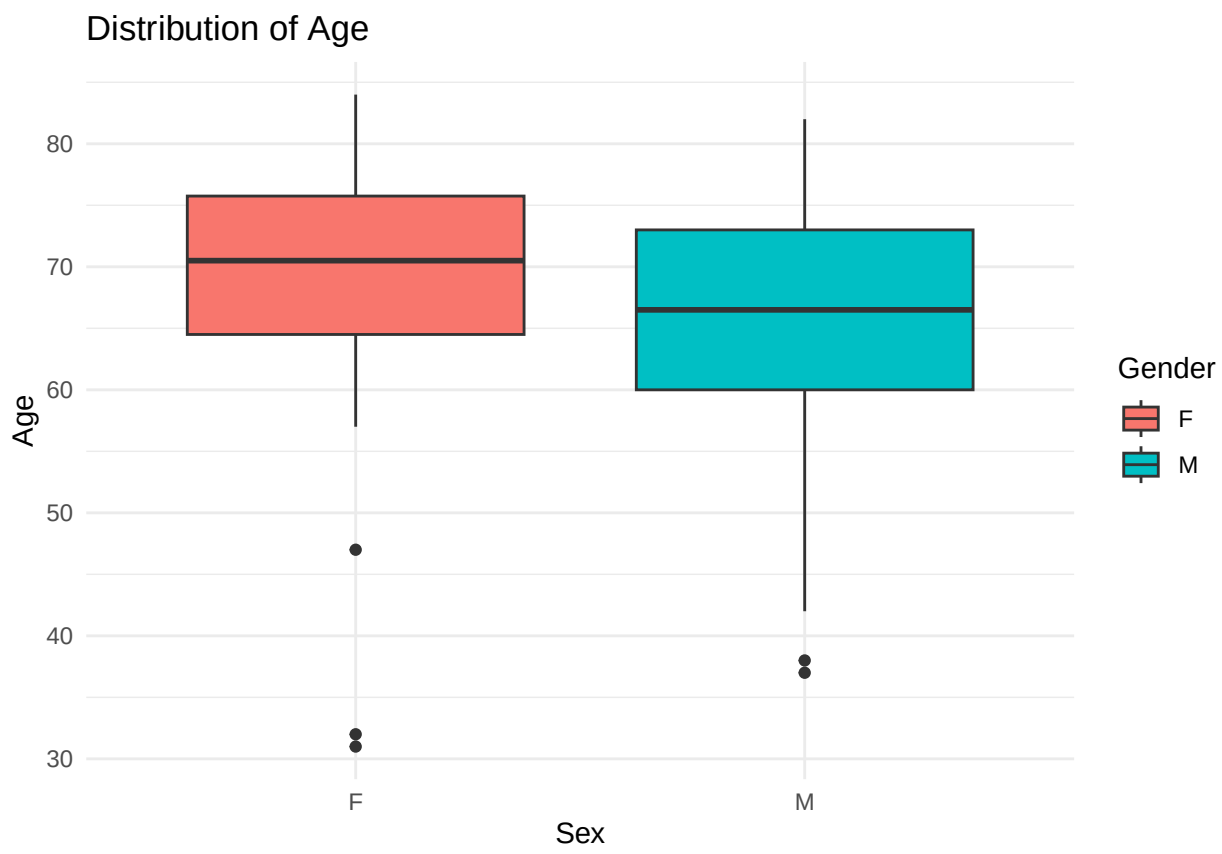
The mean age of males is 65.61 years and the mean age of females is 68.47. The p value is 0.1218 so we fail to reject the null hypothesis.

CONCLUSION: We are 95% confident that the mean age of male patients compared to female patients is not significantly different. This variable will not be further analyzed.

Here we are creating two side by side box and whisker plots showing the distribution of male and female ages

```
Age <- data.frame(Gender = metadataAnalysis$Gender, Age = metadataAnalysis$Age)

ggplot(Age, aes(x = Gender, y = Age, fill = Gender)) +
  geom_boxplot(position = position_dodge(width = 0.75)) +
  labs(title = "Distribution of Age",
       x = "Sex",
       y = "Age") +
  theme_minimal()
```



showing boxplot statistics for male age

```
#getting boxplot statistics for male age
boxplot.stats(MaleAge)
```

```
## $stats
## [1] 42.0 60.0 66.5 73.0 82.0
##
## $n
## [1] 108
##
## $conf
## [1] 64.52354 68.47646
##
## $out
## [1] 38 37
```

Showing boxplot statistics for female age distribution

```
boxplot.stats(FemaleAge)
```

```
## $stats
## [1] 47.0 64.0 70.5 76.0 84.0
##
## $n
## [1] 42
##
## $conf
## [1] 67.57441 73.42559
##
## $out
## [1] 32 31
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
Age <- data.frame(Gender = metadataAnalysis$Gender, Age = metadataAnalysis$Age)

ggplot(Age, aes(x = Gender, y = Age, fill = Gender)) +
  geom_boxplot(position = position_dodge(width = 0.75)) +
  labs(title = "Distribution of Age",
       x = "Sex",
       y = "Age") +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

Tumor Stage CHI-square

Since Tumor Stage is a categorical variable a CHI square homogenity of proportions test was conducted using the chisq.test function. A significance level of 0.05 is used

```
#male stage separated by severity
Male1st<- nrow(Male[Male$Ta== "1", ])
Male2st<- nrow(Male[Male$Ta== "2", ])
Male3st<- nrow(Male[Male$Ta== "3", ])
Male4st<- nrow(Male[Male$Ta== "4", ])

#female stage separated by severity
Female1st<- nrow(Female[Female$Ta== "1", ])
Female2st<- nrow(Female[Female$Ta== "2", ])
Female3st<- nrow(Female[Female$Ta== "3", ])
Female4st<- nrow(Female[Female$Ta== "4", ])

#counts of tumor stage
TumorStage<- data.frame(Males= c(Male1st,Male2st,Male3st,Male4st), Females= c(Female1st, Female2st, Fema
rownames(TumorStage)<- c(1,2,3,4)

#conducting chi square alpha 0.05
chisq.test(TumorStage)
```

```
## Warning in chisq.test(TumorStage): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  TumorStage
## X-squared = 6.177, df = 3, p-value = 0.1033
```

```
#using different integrator because of warning
chisq.test(TumorStage, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  TumorStage
## X-squared = 6.177, df = NA, p-value = 0.1059
```

CONCLUSION: We are 95 percent confident that the proportion of males and females at each Tumor Stage is the same. This variable does not need further analysis.

Here is the table of counts for the patients at each Tumor stage. The row names are the stage.

```
#sums of rows in TumorStage
TumorStageCounts<-c(27,72,42,9)

#add sums of rows to TumorSTage
TumorStage<- cbind(TumorStage,TumorStageCounts)

#sums of columns in TumorStage
StageTotalCounts<-c(108,42,150)

#add sums of column to TumorStage
TumorStage<- rbind(TumorStage,StageTotalCounts)
rownames(TumorStage)<- c("1","2","3","4","Totals")
```

```
TumorStage
```

```
##        Males Females TumorStageCounts
## 1         17      10               27
## 2         48      24               72
## 3         36       6               42
## 4          7       2                9
## Totals   108      42              150
```

Here is a paired percentage bar graph showing the percent of male and females samples in each tumor stage.

```
# Sample data with sex, stage, and counts
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 4),
  Category = rep(c("Stage 1", "Stage 2", "Stage 3", "Stage 4"), times = 2),
  Percentage = c(17, 48, 36, 7, 10, 24, 6, 2)
```

```
)

#colors for graph
custom_colors <- c("#40B0A6", "#FFC20A")

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

# Create a percentage paired bar graph
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Tumor Stage",
       x = "Tumor Stage",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

Investigating nodes and metastasis

We wanted to see if there were any number of patients with cancer in the lymph nodes or metastasized cancer.

```
#counting number of samples without cancer in nodes
MaleNodes<- nrow(Male[Male$N=="1",])
FemaleNodes<- nrow(Female[Female$N=="1",])
print(MaleNodes)
```

```
## [1] 0
```

```
print(FemaleNodes)
```

```
## [1] 0
```

```
#counting number of samples without metastasis
MaleMetastasis<- nrow(Male[Male$M=="1",])
FemaleMetastasis<- nrow(Female[Female$M=="1",])
print(MaleMetastasis)
```

```
## [1] 0
```

```
print(FemaleMetastasis)
```

```
## [1] 0
```

CONCLUSION: There are no patients with metastastized cancer or cancer in their lymph nodes. This variable does not need further analysis.

CHI square Edmondson Grade

Since Edmondson Grade is a categorical variable a CHI square homogenity of proportions test was conducted using the chisq.test function.

```
na_count_male <- sum(is.na(Male$Edmondson.grade))
na_count_female <- sum(is.na(Male$Edmondson.grade))
```

```
#replacing NAs with Fives
#Male[is.na(Male)]
#Female[is.na(Female)]

#male grade separated by severity
MaleEg0<- nrow(Male[Male$Edmondson.grade=="0", ])
MaleEg1<- nrow(Male[Male$Edmondson.grade=="1", ])
MaleEg1.5<- nrow(Male[Male$Edmondson.grade=="1~2", ])
MaleEg2<- nrow(Male[Male$Edmondson.grade=="2", ])
MaleEg2.5<- nrow(Male[Male$Edmondson.grade=="2~3", ])
MaleEg3<- nrow(Male[Male$Edmondson.grade=="3", ])
MaleEg4<- nrow(Male[Male$Edmondson.grade=="4", ])

#female grade separated by severity
FemaleEg0<- nrow(Female[Female$Edmondson.grade=="0", ])
FemaleEg1<- nrow(Female[Female$Edmondson.grade=="1", ])
FemaleEg1.5<- nrow(Female[Female$Edmondson.grade=="1~2", ])
FemaleEg2<- nrow(Female[Female$Edmondson.grade=="2", ])
FemaleEg2.5<- nrow(Female[Female$Edmondson.grade=="2~3", ])
FemaleEg3<- nrow(Female[Female$Edmondson.grade=="3", ])
FemaleEg4<- nrow(Female[Female$Edmondson.grade=="4", ])

#creating dataframe for edmondson grade
EdmondsonGrade<- data.frame(Males= c(MaleEg0,MaleEg1,MaleEg1.5,MaleEg2,MaleEg2.5,MaleEg3,MaleEg4), Fema
rownames(EdmondsonGrade)<- c(0,1,1.5,2,2.5,3,4)

#chi square test alpha 0.05
chisq.test(EdmondsonGrade)
```

```
## Warning in chisq.test(EdmondsonGrade): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  EdmondsonGrade
## X-squared = 3.1505, df = 6, p-value = 0.7897
```

```
#ran using different integrator because of warning
chisq.test(EdmondsonGrade, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
```

```
##
## data:  EdmondsonGrade
## X-squared = 3.1505, df = NA, p-value = 0.8186
```

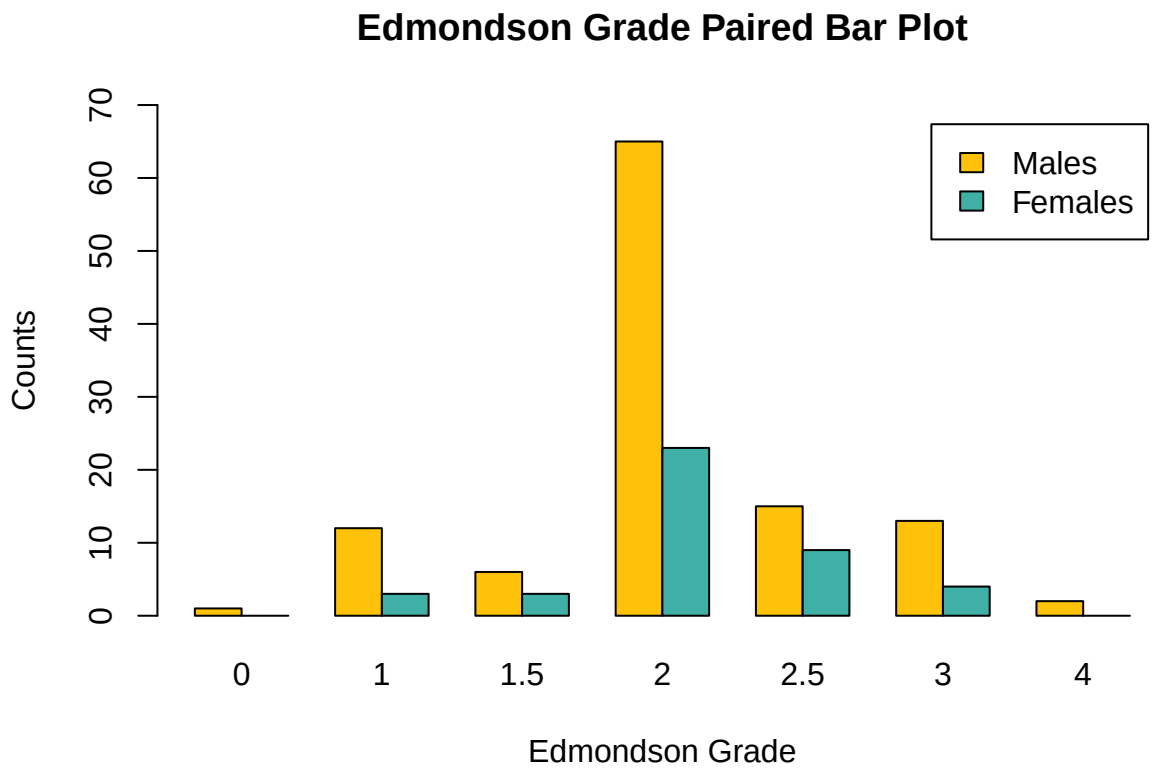
```
#cannot reject null
```

CONCLUSION: We are 95 percent confident that the proportion of males and females at each Edmondson Grade is the same. This variable does not need further analysis.

Below is a paired bar plot of raw counts of male and female samples at each Edmondson Grade

```
EdmondsonHist <- t(EdmondsonGrade)
EdmondsonHist <- as.matrix(EdmondsonHist)
```

```
barplot(EdmondsonHist, beside = TRUE, ylim = c(0,70), col = c("#FFC20A", "#40B0A6"), names.arg = c(0,1,
       legend.text = c("Males", "Females"), xlab = "Edmondson Grade", ylab = "Counts",
       main = "Edmondson Grade Paired Bar Plot")
```




```
dev.off()
```

```
## null device
##           1
```

Listed below is the table show the counts of males and females with each Edmondson grade The row names are the levels of Edmondson Grade

```r
#sums of rows in EdmondsonGrade
EdmondsonGradeCounts<-c(1,14,8,87,23,16,1)

#add sums of rows to EdmondsonGrade
EdmondsonGrade<- cbind(EdmondsonGrade,EdmondsonGradeCounts)

#sums of columns in EdmondsonGrade
GradeTotalCounts<-c(108,42,150)

#add sums of column to EdmondsonGrade
EdmondsonGrade<- rbind(EdmondsonGrade,GradeTotalCounts)
rownames(EdmondsonGrade)<- c("0","1","1.5","2","2.5","3","4","Totals")
```

```r
EdmondsonGrade
```

```
##        Males Females EdmondsonGradeCounts
## 0          1       0                    1
## 1         12       3                   14
## 1.5        6       3                    8
## 2         65      23                   87
## 2.5       15       9                   23
## 3         13       4                   16
## 4          2       0                    1
## Totals   108      42                  150
```

Below is a percentage paired bar plot showing the percentage of male and female samples at each Edmondson Grade.

```r
FemaleEg5<- nrow(Male[Male$Edmondson.grade=="5", ])
print(FemaleEg5)
```

```
## [1] 1
```

```r
FemaleEg5<- nrow(Male[Male$Edmondson.grade=="5", ])
print(FemaleEg5)
```

```
## [1] 1
```

```r
# Sample data with sex grade and count data
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 7),
  Category = rep(c("Grade 0", "Grade 1", "Grade 1.5", "Grade 2", "Grade 2.5", "Grade 3", "Grade 4"), ti
  Percentage = c(1, 11, 5, 64, 14, 12, 1, 0, 3, 3, 23, 9, 4, 0)
)

custom_colors <- c("#40B0A6", "#FFC20A")

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
```

```r
# Create a percentage paired bar graph
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Edmondson Grade",
       x = "Edmondson Grade",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

F and T test tumor size in mm

Tumor Size is a numerical variable so a two-sample t test will be used to identify if the variable is statistically significant. To figure out the type of t test we want to use an f test is used to determine whether or not the variances between the two datasets are equal or not. A significance level of 0.05 is used.

```r
#subsetting by tumor size in female samples
FemaleSize <- Female$Tumor.size..mm.

#subsetting by tumor size in male samples
MaleSize<- Male$Tumor.size..mm.

#f test alpha=0.05
var.test(MaleSize, FemaleSize) #reject null variances significantly different
```

```
##
##  F test to compare two variances
##
## data:  MaleSize and FemaleSize
## F = 3.1061, num df = 107, denom df = 41, p-value = 9.361e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.801914 5.044798
## sample estimates:
## ratio of variances
##           3.106136
```

```r
#variances unequal
```

The p value is 9.361e-05 so therefore we reject the null hypothesis. So, we will conduct a t test with unequal variances.

```r
#t test tumor size alpha=0.05
t.test(MaleSize,FemaleSize) #Do not reject null no evidence tumor size is different
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  MaleSize and FemaleSize
## t = 1.5319, df = 128.2, p-value = 0.128
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.271914 17.853924
## sample estimates:
## mean of x mean of y
##  41.31481  33.52381
```

```
#no significant difference
```

The mean male tumor size is 41.314 and the mean female tumor size is 33.52. The p value is 0.128 so we fail to reject null hypothesis.

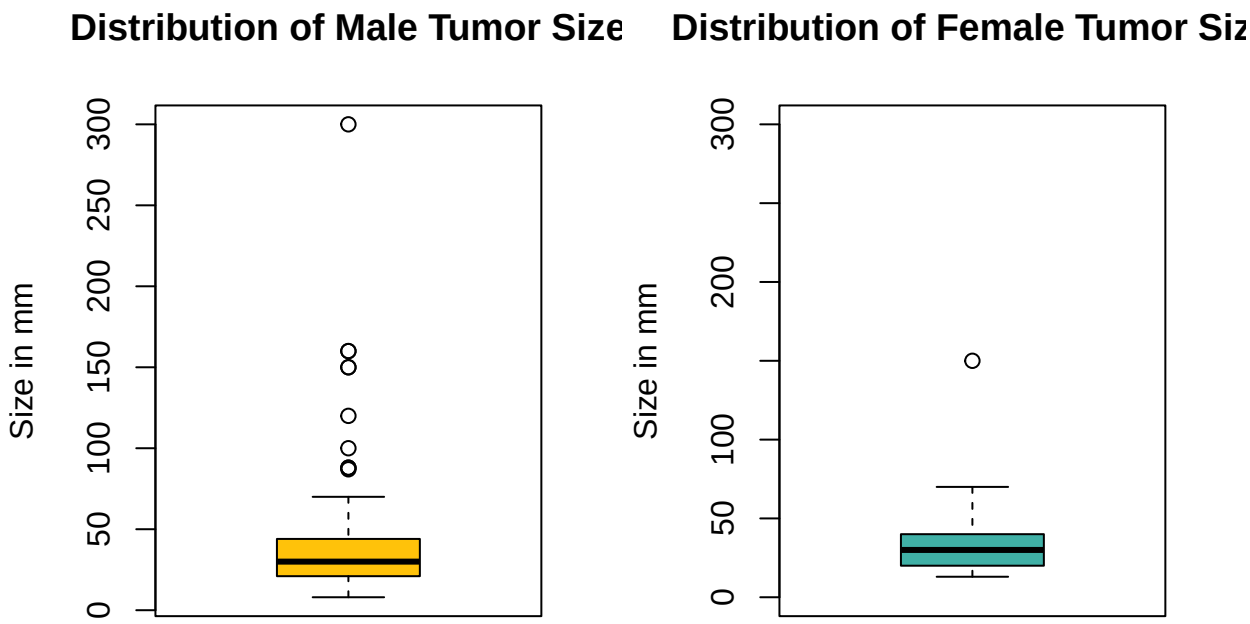
CONCLUSION:We are 95% confident that the mean tumor size (in mm) is not significantly different between males and females. We will not conduct further analysis on this variable.

Here two box and whisker plots are created side by side showing the male and female distribution of tumor size.

```
# Create side-by-side boxplots
par(mfrow = c(1, 2))  # Set up a 1x2 layout for two plots

# Boxplot for MaleAge
boxplot(MaleSize, main = "Distribution of Male Tumor Size", ylab = "Size in mm", col = "#FFC20A")

# Boxplot for FemaleAge
boxplot(FemaleSize, main = "Distribution of Female Tumor Size", ylab = "Size in mm", col = "#40B0A6", yl
```




```
# Reset the layout to default (1x1) after creating the plots
par(mfrow = c(1, 1))
```

```
TumorSize <- data.frame(Gender = metadataAnalysis$Gender, TumorSize = metadataAnalysis$Tumor.size..mm.)

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(TumorSize, aes(x = Gender, y = TumorSize, fill = Gender)) +
  geom_boxplot(position = position_dodge(width = 0.75)) +
  labs(title = "Distribution of Tumor Size in mm",
       x = "Sex",
       y = "Millimeters") +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

Liver Fibrosis CHI-square

Since Liver Fibrosis is a categorical variable a CHI square homogenity of proportions test was conducted using the chisq.test function.

```r
#male fibrosis separated by severity
Male0F<- nrow(Male[Male$Liver.fibrosisc== "0", ])
Male1F<- nrow(Male[Male$Liver.fibrosisc== "1", ])
Male2F<- nrow(Male[Male$Liver.fibrosisc== "2", ])
Male3F<- nrow(Male[Male$Liver.fibrosisc== "3", ])
Male4F<- nrow(Male[Male$Liver.fibrosisc== "4", ])

#female fibrosis
Female0F<- nrow(Female[Female$Liver.fibrosisc== "0", ])
Female1F<- nrow(Female[Female$Liver.fibrosisc== "1", ])
Female2F<- nrow(Female[Female$Liver.fibrosisc== "2", ])
Female3F<- nrow(Female[Female$Liver.fibrosisc== "3", ])
Female4F<- nrow(Female[Female$Liver.fibrosisc== "4", ])

#counts of liver fibrosis

LiverFibrosis<- data.frame(Males= c(Male0F,Male1F,Male2F,Male3F,Male4F), Females= c(Female0F, Female1F,

#chi square test alpha 0.05
chisq.test(LiverFibrosis)
```

```
## Warning in chisq.test(LiverFibrosis): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  LiverFibrosis
## X-squared = 6.5253, df = 4, p-value = 0.1632
```

```r
#using different integrator because of warning
chisq.test(LiverFibrosis, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  LiverFibrosis
## X-squared = 6.5253, df = NA, p-value = 0.1636
```

CONCLUSION: We are 95 percent confident that the proportion of males and females with liver fibrosis is the same. This variable does not need further analysis.

Listed below is the table show the counts of males and females with each level of liver fibrosis. The row names are the severity of liver fibrosis

```r
#sums of rows in LiverFibrosis
LiverFibrosisCounts<-c(5,16,30,40,59)

#add sums of rows to LiverFibrosis
LiverFibrosis<- cbind(LiverFibrosis,LiverFibrosisCounts)
```

```r
#sums of columns in LiverFibrosis
LiverTotalCounts<-c(108,42,150)

#add sums of column to LiverFibrosis
LiverFibrosis<- rbind(LiverFibrosis,LiverTotalCounts)
rownames(LiverFibrosis)<- c("0","1","2","3","4","Totals")
```

```r
LiverFibrosis
```

```
##        Males Females LiverFibrosisCounts
## 0          5       0                   5
## 1         13       3                  16
## 2         25       5                  30
## 3         26      14                  40
## 4         39      20                  59
## Totals   108      42                 150
```

Below is a percentage paired bar plot showing the percentage of male and female samples at each level of Liver Fibrosis.

```r
# Sample data with sex level of fibrosis and count data
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 5),
  Category = rep(c("Severity 0", "Severity 1", "Severity 2", "Severity 3", "Severity 4"), times = 2),
  Percentage = c(5, 13, 25, 26, 39, 0, 3, 5, 14, 20)
)

custom_colors <- c("#40B0A6", "#FFC20A") # colors denoting the sex

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

# Create a percentage paired bar graph

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Liver Fibrosis",
       x = "Severity",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

CHI-squared Alcohol

Since Alcohol Intake is a categorical variable a Chi square test will be conducted using chisq.test function.

```r
#male alcohol intake separated by severity
Male0<- nrow(Male[Male$Alcohol.intakee== "0", ])
Male1<- nrow(Male[Male$Alcohol.intakee== "1", ])
Male2<- nrow(Male[Male$Alcohol.intakee== "2", ])
Male3<- nrow(Male[Male$Alcohol.intakee== "3", ])
MaleAlcohol<- as.data.frame(cbind(Male0,Male1,Male2,Male3))

#female alcohol
Female0<- nrow(Female[Female$Alcohol.intakee== "0", ])
Female1<- nrow(Female[Female$Alcohol.intakee== "1", ])
Female2<- nrow(Female[Female$Alcohol.intakee== "2", ])
Female3<- nrow(Female[Female$Alcohol.intakee== "3", ])
FemaleAlcohol<- as.data.frame(cbind(Female0,Female1,Female2,Female3))

#counts of alcohol intake

AlcoholIntake<- as.data.frame(cbind(Males= c(Male0,Male1,Male2,Male3), Females= c(Female0, Female1, Fema
rownames(AlcoholIntake)<- c(0,1,2,3)

#chi square test alpha of 0.05
chisq.test(AlcoholIntake)
```

```
##
##  Pearson's Chi-squared test
##
## data:  AlcoholIntake
## X-squared = 30.204, df = 3, p-value = 1.25e-06
```

```r
# reject null hypothesis
```

CONCLUSION: We are 95% confident that the proportion of males and females who intake alcohol is different. This variable needs further analysis.

Here is a paired barplot showing the raw counts of males and females at each level of alcohol intake.

```r
#transposing data for paired bar plot
AlcoholIntake <- t(AlcoholIntake)
AlcoholIntake <- as.matrix(AlcoholIntake)

#paired bar plot
barplot(AlcoholIntake, beside = TRUE, ylim = c(0,40), col = c("#FFC20A", "#40B0A6"), names.arg = c(0,1,
       legend.text = c("Males", "Females"), xlab = "Level of Severity", ylab = "Counts",
       main = "Alcohol Intake Paired Bar Plot")
```


```
#transposing data
AlcoholIntake <- as.data.frame(AlcoholIntake)
AlcoholIntake <- t(AlcoholIntake)
```

Here is a sample table for males and female patients alcohol intake. The level of severity is the row name.

```
#sums of rows in Alcoholintake
AlcoholTotalCounts<-c(63,25,42,20)

#add sums of rows to AlcoholIntake
AlcoholIntake<- cbind(AlcoholIntake,AlcoholTotalCounts)

#sums of columns in AlcoholIntake
AlcoholTotalCounts<-c(108,42,150)

#add sums of column to AlcoholIntake
AlcoholIntake<- rbind(AlcoholIntake,AlcoholTotalCounts)
rownames(AlcoholIntake)<- c("0","1","2","3","Totals")
```

```
AlcoholIntake
```

```
##        Males Females AlcoholTotalCounts
## 0         31      32                63
## 1         20       5                25
## 2         37       5                42
```

```
## 3         20       0                 20
## Totals   108      42                150
```

Below is a percentage paired bar plot showing the percentage of male and female sample at each level of alcohol intake.

```
# Sample data sex level of alcohol intake and sample count
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 4),
  Category = rep(c("Level 0", "Level 1", "Level 2", "Level 3"), times = 2),
  Percentage = c(31, 20, 37, 20, 32, 5, 5, 0)
)

custom_colors <- c("#40B0A6", "#FFC20A")

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

# Create a percentage paired bar graph
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Alcohol Intake",
       x = "Level of Intake",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

Test of Proportions Smoking Males and Females

Smoking is a dichotomous variable so we will conduct a test of proportions

```
#number of male smokers
MaleSmokers<-sum(Male$Smoking)
MaleSmokers
```

```
## [1] 83
```

```
#number of female smokers
FemaleSmokers<- sum(Female$Smoking)
FemaleSmokers
```

```
## [1] 4
```

```r
# alpha=0.05
prop.test(x=c(MaleSmokers,FemaleSmokers), n= c(108,42), p = NULL, alternative = "two.sided", correct = T
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(MaleSmokers, FemaleSmokers) out of c(108, 42)
## X-squared = 53.543, df = 1, p-value = 2.53e-13
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.5375452 0.8090157
## sample estimates:
##    prop 1    prop 2
## 0.7685185 0.0952381
```

```r
#the number of men and women that smoke is significantly different
```

CONCLUSION: We are 95% confident that the proportion of males and females who smoke are different. Since 77% of males smoke and 9% of females smoke it is clear that more males that females smoke.

Here a paired bar plot showing the percentages of males and females who do and do not smoke is created.

```r
#creating data frame with sex, smoking, and sample counts
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 2),
  Category = rep(c("Does Smoke", "Does Not Smoke"), times = 2),
  Percentage = c(83,25,4,38))

custom_colors <- c("#40B0A6", "#FFC20A")

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

# Create a percentage paired bar graph
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Smoking",
       x = "Smoking",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

CHI square Prognosis

Sincle prognosis is a categorical variable a chi square homogenity of proportions will be conducted.

```r
#male prognosis separated by severity?
MaleProg0<- nrow(Male[Male$Prognosisf=="0",])
MaleProg1<- nrow(Male[Male$Prognosisf=="1",])
MaleProg2<- nrow(Male[Male$Prognosisf=="2",])
MaleProg3<- nrow(Male[Male$Prognosisf=="3",])

#female prognosis separated by severity?
FemaleProg0<- nrow(Female[Female$Prognosisf=="0",])
FemaleProg1<- nrow(Female[Female$Prognosisf=="1",])
FemaleProg2<- nrow(Female[Female$Prognosisf=="2",])
FemaleProg3<- nrow(Female[Female$Prognosisf=="3",])

#counts of Prognosis
Prognosis<- as.data.frame(cbind(Males= c(MaleProg0,MaleProg1,MaleProg2,MaleProg3), Females= c(FemaleProg

chisq.test(Prognosis)
```

```
## Warning in chisq.test(Prognosis): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Prognosis
## X-squared = 1.88, df = 3, p-value = 0.5977
```

```r
#using different integrator because of warning
chisq.test(Prognosis, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  Prognosis
## X-squared = 1.88, df = NA, p-value = 0.6798
```

CONCLUSION: We are 95% confident that the proportion of males and females at each level of prognosis is the same

Here is a sample table with number of patients at each level of prognosis. The level of severity is the row name.

```r
#sums of rows in Prognosis
PrognosisCounts<-c(124,21,3,2)

#add sums of rows to Prognosis
Prognosis<- cbind(Prognosis,PrognosisCounts)

#sums of columns in Prognosis
PrognosisTotalCounts<-c(108,42,150)

#add sums of column to Prognosis
Prognosis<- rbind(Prognosis,PrognosisTotalCounts)
rownames(Prognosis)<- c("0","1","2","3","Totals")
```

```
Prognosis
```

```
##         Males Females PrognosisCounts
## 0          92      32             124
## 1          13       8              21
## 2           2       1               3
## 3           1       1               2
## Totals    108      42             150
```

Here a paired bar plot with is created showing the percentage of male and female samples and each level of prognosis

```
# Sample data with sex level of prognosis and sample count
data <- data.frame(
  Sex = rep(c("Males", "Females"), each = 4),
  Category = rep(c("Level 0", "Level 1", "Level 2", "Level 3"), times = 2),
  Percentage = c(92, 13, 2, 1, 32, 8, 1, 1)
)

custom_colors <- c("#40B0A6", "#FFC20A")

# Calculate the percentages within each group
data <- transform(data, Percentage = Percentage / tapply(Percentage, Sex, sum)[Sex] * 100)

# Create a percentage paired bar graph
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/Statistical_testing_potential_covariates/figu
ggplot(data, aes(x = Category, y = Percentage, fill = Sex, group = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage Bar Graph Prognosis",
       x = "Level of Severity",
       y = "Percentage") +geom_text(aes(label = sprintf("%.1f%%", Percentage)),
            position = position_dodge(width = 0.9),    # Adjust the width as needed
            vjust = -0.5, hjust = 0.5, size = 3) +  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal()
dev.off()
```

```
## pdf
##   2
```

F and T test overall survival month

Since survival month is a numerical variable a two sample t test will be conducted. To see what type of t test shall be used a f test is conducted first

```
#subsetting overall survival month from male sample
MaleSurvivalMonth<- Male$Overall.survival..month.

#subsetting overall survival month from female sample
FemaleSurvivalMonth<- Female$Overall.survival..month.

#f test overall survival month alpha=0.05
var.test(MaleSurvivalMonth, FemaleSurvivalMonth) #fail to reject null no evidence variances are differe
```

```
##
##  F test to compare two variances
##
## data:  MaleSurvivalMonth and FemaleSurvivalMonth
## F = 1.3051, num df = 107, denom df = 41, p-value = 0.3362
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7571205 2.1197021
## sample estimates:
## ratio of variances
##           1.305123
```

```
#variances are equal
```

The p value is 0.3362 so we fail to reject the null hypothesis. A two sample t test of equal variance will be conducted.

```
#t test overall survival month alpha=0.05
t.test(MaleSurvivalMonth, FemaleSurvivalMonth, var.equal = TRUE) #Reject null survival month significan
```

```
##
##  Two Sample t-test
##
## data:  MaleSurvivalMonth and FemaleSurvivalMonth
## t = 2.1201, df = 148, p-value = 0.03567
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.3597716 10.2381120
## sample estimates:
## mean of x mean of y
##  29.20370  23.90476
```

```
#significant difference
```

The mean overall survival month of males is 29.2 months and the mean overall survival month of females is 23.9 months. Note that the survival month for females is shorter than males which is atypical.

CONCLUSION: We are 95% confident that the mean survival month of males is different that the mean survival month of females.