Vaccination Data

waheeb Algabri

```
# load the readxl package
library(readxl)

# read the Excel file
data <- read_excel("israeli_vaccination_data_analysis_start.xlsx")
```

```
str(data)
```

```
## tibble [15 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Age         : chr [1:15] NA "<50" NA ">50" ...
##  $ Population %: chr [1:15] "Not Vax\r\n%" "1116834" "0.23300000000000001" "186078" ...
##  $ ...3        : chr [1:15] "Fully Vax\r\n%" "3501118" "0.73" "2133516" ...
##  $ Severe Cases: chr [1:15] "Not Vax\r\nper 100K\r\n\r\n\r\np" "43" NA "171" ...
##  $ ...5        : chr [1:15] "Fully Vax\r\nper 100K" "11" NA "290" ...
##  $ Efficacy    : chr [1:15] "vs. severe disease" NA NA NA ...
```

write the data to a CSV file

```
# write the data to a CSV file
write.csv(data, file = "vaccination.csv", row.names = FALSE)
```

```
# convert to data frame
df <- as.data.frame(data)
```

```
head(df)
```

```
##    Age          Population %                 ...3
## 1 <NA>           Not Vax\r\n%       Fully Vax\r\n%
## 2  <50               1116834             3501118
## 3 <NA>   0.23300000000000001                 0.73
## 4  >50                186078             2133516
## 5 <NA> 7.9000000000000001E-2 0.90400000000000003
## 6 <NA>                  <NA>                 <NA>
##                       Severe Cases                  ...5           Efficacy
## 1 Not Vax\r\nper 100K\r\n\r\n\r\np Fully Vax\r\nper 100K vs. severe disease
## 2                               43                    11               <NA>
## 3                             <NA>                  <NA>               <NA>
## 4                              171                   290               <NA>
## 5                             <NA>                  <NA>               <NA>
## 6                             <NA>                  <NA>               <NA>
```

```r
# Remove consecutive newlines and percent signs in the first row
df[1,] <- gsub('(\r?\n){2,}|%', '', df[1,])

# Rename columns using the first row of the data frame
colnames(df) <- df[1,]
colnames(df)[1] <- "Age"

# Remove the first row and last column of the data frame
df <- df[2:5, -6]

# Reset row names to match row indices
rownames(df) <- NULL

# Replace some values in the "Age" column with NA
df[c(2, 4), 1] <- NA

# Fill in missing values in the "Age" column
df <- tidyr::fill(df, Age)
```

```r
# Print the data frame in a nicely formatted table
library(knitr)
print(knitr::kable(df))
```

```
##
##
## |Age |Not Vax
##                 |Fully Vax
##             |Not Vax
## per 100Kp |Fully Vax
## per 100K |
## |:---|:---------------------|:-------------------|:----------------|:-----------------|
## |<50 |1116834               |3501118             |43               |11                |
## |<50 |0.23300000000000001   |0.73                |NA               |NA                |
## |>50 |186078                |2133516             |171              |290               |
## |>50 |7.9000000000000001E-2 |0.90400000000000003 |NA               |NA                |
```

```r
# Print the data frame in a nicely formatted table with formatting options
knitr::kable(df, align = "c", caption = "Israeli Vaccination Data Analysis",
             col.names = c("Age Group", "Not Vaccinated", "Fully Vaccinated",
                           "Not Vaccinated per 100K", "Fully Vaccinated per 100K"),
             digits = c(0, 0, 0, 2, 2),
             format.args = list(big.mark = ",", scientific = FALSE),
             row.names = FALSE)
```

Table 1: Israeli Vaccination Data Analysis

Age Group	Not Vaccinated	Fully Vaccinated	Not Vaccinated per 100K	Fully Vaccinated per 100K
<50	1116834	3501118	43	11
<50	0.23300000000000001	0.73	NA	NA
>50	186078	2133516	171	290

Age Group	Not Vaccinated	Fully Vaccinated	Not Vaccinated per 100K	Fully Vaccinated per 100K
>50	7.9000000000000001E-2	0.90400000000000003	NA	NA

Rename the "Not Vaccinated per 100K" and "Fully Vaccinated per 100K" columns to "Hospitalized_not_vaxed_per 100K" and "Hospitalized_vaxed_per 100K", respectively

```
colnames(df)[4] <- "Hospitalized_not_vaxed_per 100K"
colnames(df)[5] <- "Hospitalized_vaxed_per 100K"
```

Grab the percentages from the second and fourth rows and place them into a new "% of total eligible" column:

```
percents1 <- str_extract_all(as.matrix(df[2,]), "\\d+\\.\\d(?=\\%)", simplify=TRUE)
percents1 <- percents1[!apply(percents1 == "", 1, all), ]
percents2 <- str_extract_all(as.matrix(df[4,]), "\\d+\\.\\d(?=\\%)", simplify=TRUE)
percents2 <- percents2[!apply(percents2 == "", 1, all), ]
df$`% of total eligible` <- c(percents1, percents2)
```

Grab the vaccination numbers from the first and third rows and place them into a new "Eligible population" column:

```
under50 <- as.matrix(df[1, c(2,3)])
over50 <- as.matrix(df[3, c(2,3)])
df$`Eligible population` <- c(under50, over50)
```

Add a new "Vax status" column:

```
df$`Vax status` <- c("N", "Y", "N", "Y")
```

Pivot the "Hospitalized_not_vaxed_per 100K" and "Hospitalized_vaxed_per 100K" columns to create a new "Hospitalized (per 100K)" column, then filter out any rows with missing values:

```
library(tidyr)

longified <- pivot_longer(df, c("Hospitalized_not_vaxed_per 100K", "Hospitalized_vaxed_per 100K"),
                          names_to="Status", values_to="Number_hosp") %>%
                          filter(Number_hosp != "")
df$`Hospitalized (per 100K)` <- longified$Number_hosp
```

Place the newly generated and cleaned columns into a new data frame called "tidied" and print it in a nicely formatted table using kable():

```
tidied <- df[c("Age", "Eligible population", "Vax status", "% of total eligible", "Hospitalized (per 10

tidied$`% of total eligible` <- c(23.3, 73.0, 7.9, 90.4)

# Print the tidied data frame in a nicely formatted table with formatting options
knitr::kable(tidied, align = "c", caption = "Tidied Israeli Vaccination Data Analysis",
             col.names = c("Age Group", "Eligible Population", "Vax Status",
                           "% of Total Eligible", "Hospitalized (per 100K)"),
             digits = c(0, 0, 0, 2, 2),
             format.args = list(big.mark = ",", scientific = FALSE),
             row.names = FALSE)
```

Table 2: Tidied Israeli Vaccination Data Analysis

Age Group	Eligible Population	Vax Status	% of Total Eligible	Hospitalized (per 100K)
<50	1116834	N	23.3	43
<50	3501118	Y	73.0	11
>50	186078	N	7.9	171
>50	2133516	Y	90.4	290

Do you have enough information to calculate the total population? What does this total population represent? we can estimate the total eligible population by dividing the eligible population number by the percentage of total population for each age group, taking the mean of the two values for each group, and then adding both means together. However, without further information about eligibility parameters within the Israeli population, we cannot determine the total population from the information provided in the table alone.

```
# Save table in a variable called Tidied
Tidied <- data.frame(
  Age_Group = c("<50", "<50", ">50", ">50"),
  Eligible_Population = c(1116834, 3501118, 186078, 2133516),
  Vax_Status = c("N", "Y", "N", "Y"),
  Percent_of_Total_Eligible = c(23.3, 73.0, 7.9, 90.4),
  Hospitalized_per_100K = c(43, 11, 171, 290)
)
```

Calculate estimated total eligible population

```
less_than_50_mean <- mean(Tidied$Eligible_Population[Tidied$Age_Group == "<50"] / (Tidied$Percent_of_To
greater_than_50_mean <- mean(Tidied$Eligible_Population[Tidied$Age_Group == ">50"] / (Tidied$Percent_of_
total_eligible_population <- less_than_50_mean + greater_than_50_mean

# Print result
print(total_eligible_population)
```

```
## [1] 7152416
```

```
# Print result
print(total_eligible_population)
```

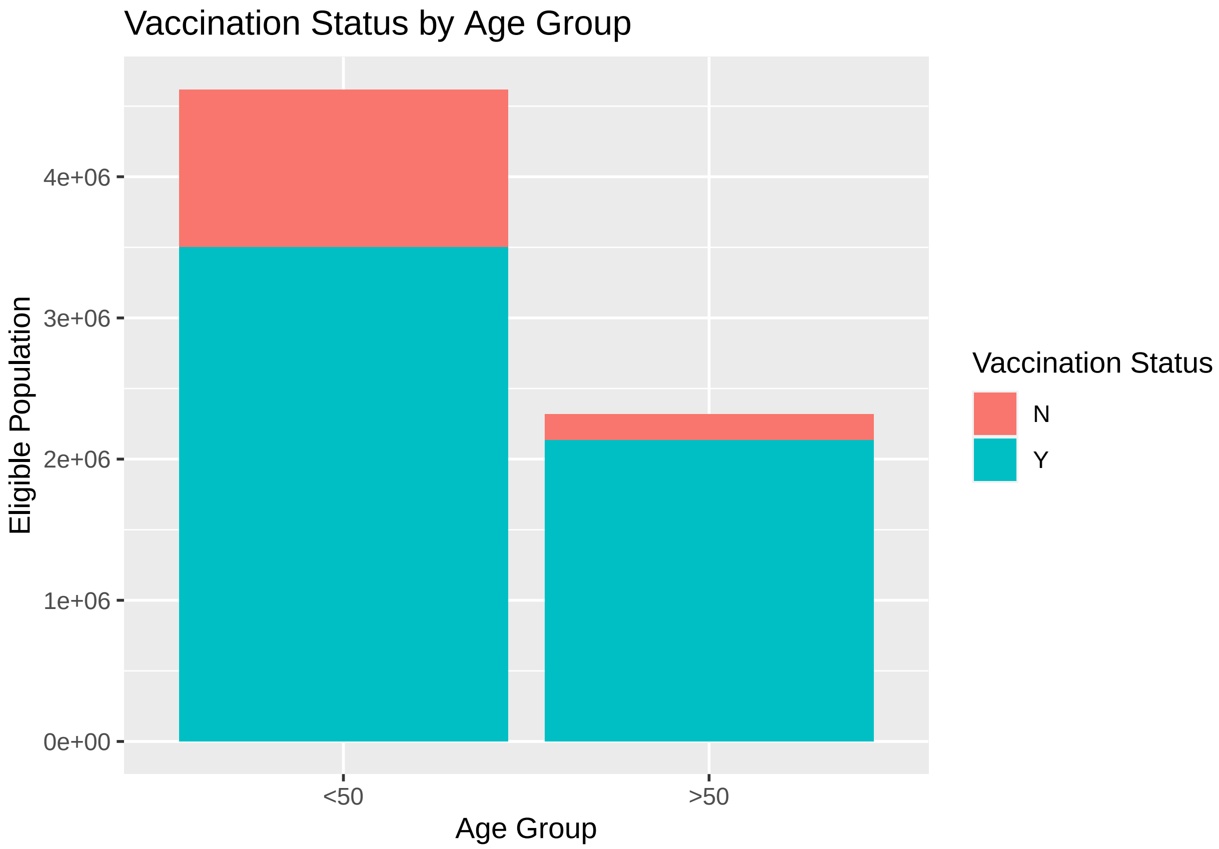
```
## [1] 7152416
```

The result of 7152416 means that the estimated total eligible population based on the data in the Tidied table is approximately 7.15 million people.

Bar chart of vaccination status by age group

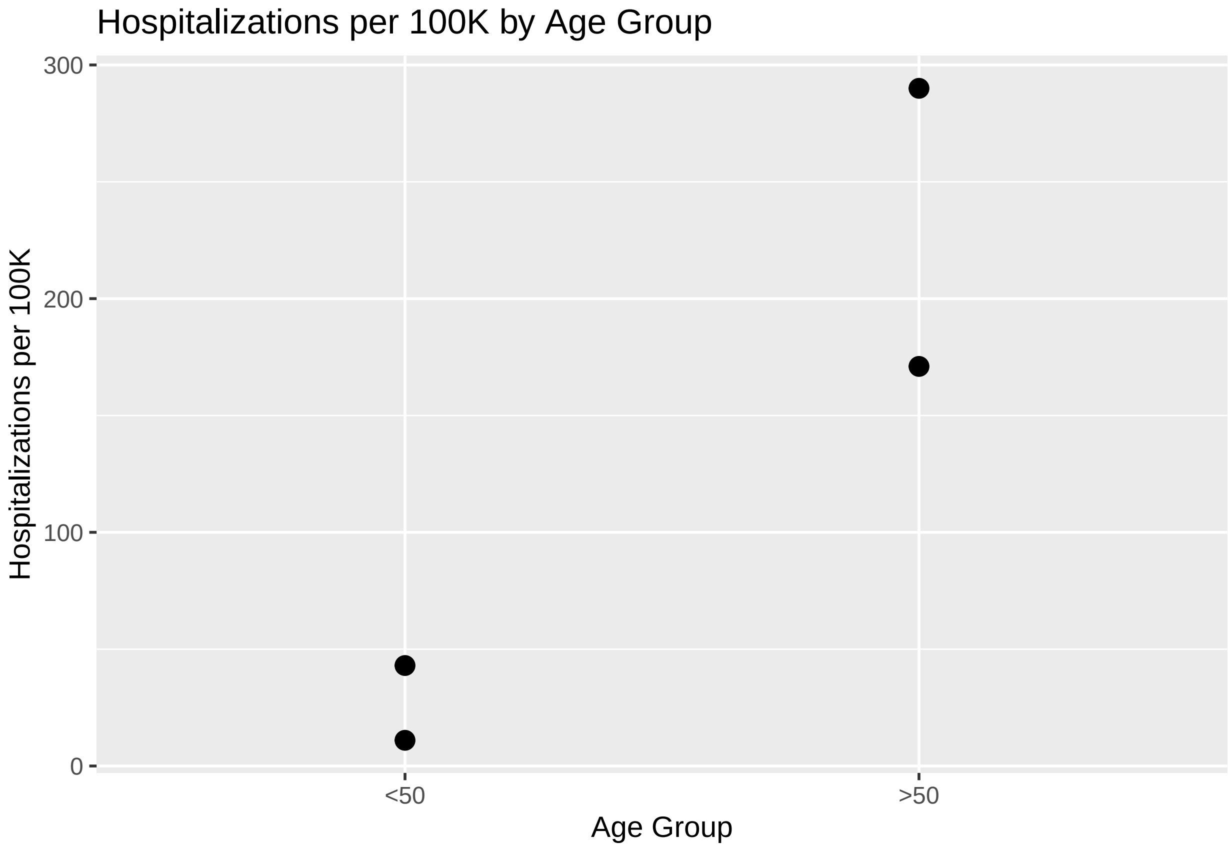
```
library(ggplot2)

# Create a bar chart of vaccination status by age group
ggplot(Tidied, aes(x = Age_Group, y = Eligible_Population, fill = Vax_Status)) +
  geom_bar(stat = "identity") +
  labs(title = "Vaccination Status by Age Group", x = "Age Group", y = "Eligible Population", fill = "Va
```



Scatter plot of hospitalizations by age group

```
# Create a scatter plot of hospitalizations by age group
ggplot(Tidied, aes(x = Age_Group, y = Hospitalized_per_100K)) +
  geom_point(size = 3) +
  labs(title = "Hospitalizations per 100K by Age Group", x = "Age Group", y = "Hospitalizations per 100K
```

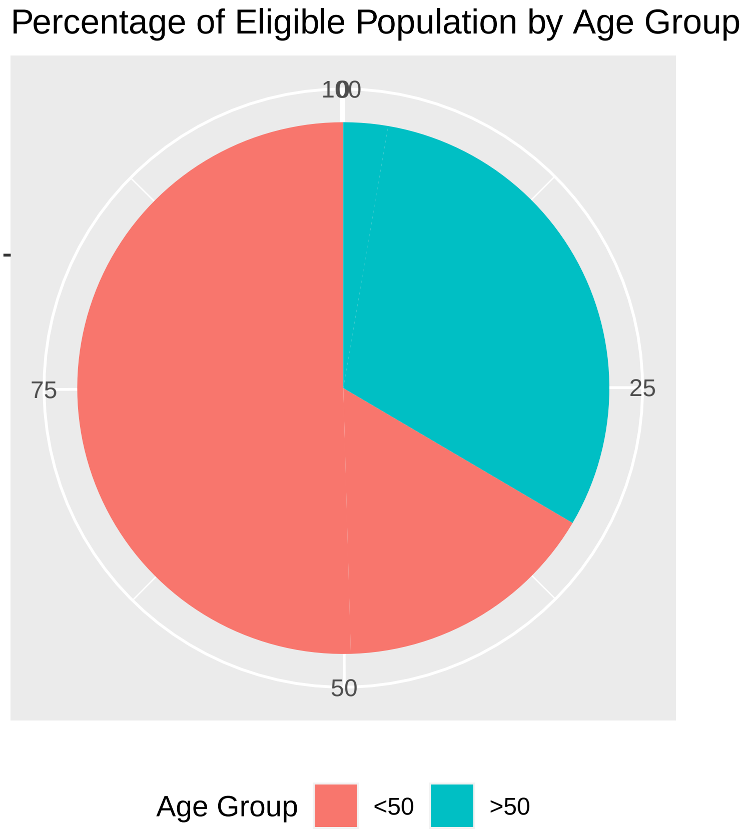


A pie chart to show the percentage of the total eligible population that each age group represents.

```
library(ggplot2)

# Calculate the percentage of total eligible population for each age group
Tidied$Percent_of_Total_Eligible <- round(Tidied$Eligible_Population / sum(Tidied$Eligible_Population)

# Create a pie chart of percentage of eligible population by age group
ggplot(Tidied, aes(x = "", y = Percent_of_Total_Eligible, fill = Age_Group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  labs(title = "Percentage of Eligible Population by Age Group", fill = "Age Group", x = NULL, y = NULL
  theme(legend.position = "bottom")
```



```
# Define the data frame
df <- data.frame(
  Age_Group = c("<50", "<50", ">50", ">50"),
  Eligible_Population = c(1116834, 3501118, 186078, 2133516),
  Vax_Status = c("N", "Y", "N", "Y"),
  Percent_of_Total_Eligible = c(23.3, 73.0, 7.9, 90.4),
  Hospitalized_per_100K = c(43, 11, 171, 290)
)

# Filter the data for each age group
df_under50 <- df[df$Age_Group == "<50", ]
df_over50 <- df[df$Age_Group == ">50", ]
```

2) Calculate the Efficacy vs. Disease; Explain your results. Efficacy over50

```
# Calculate the vaccine efficacy for individuals over 50 years old
efficacy_over50 <- (1 - (df_over50$Hospitalized_per_100K[df_over50$Vax_Status == "Y"] /
                         df_over50$Hospitalized_per_100K[df_over50$Vax_Status == "N"])) * 100
cat("Vaccine efficacy for individuals over 50 years old: ", round(efficacy_over50, 1), "%\n")
```

```
## Vaccine efficacy for individuals over 50 years old:  -69.6 %
```

```
print(efficacy_over50)
```

```
## [1] -69.59064
```

Efficacy under 50

```
# Calculate the vaccine efficacy for individuals under 50 years old
efficacy_under50 <- (1 - (df_under50$Hospitalized_per_100K[df_under50$Vax_Status == "Y"] /
                          df_under50$Hospitalized_per_100K[df_under50$Vax_Status == "N"])) * 100
cat("Vaccine efficacy for individuals under 50 years old: ", round(efficacy_under50, 1), "%\n")
```

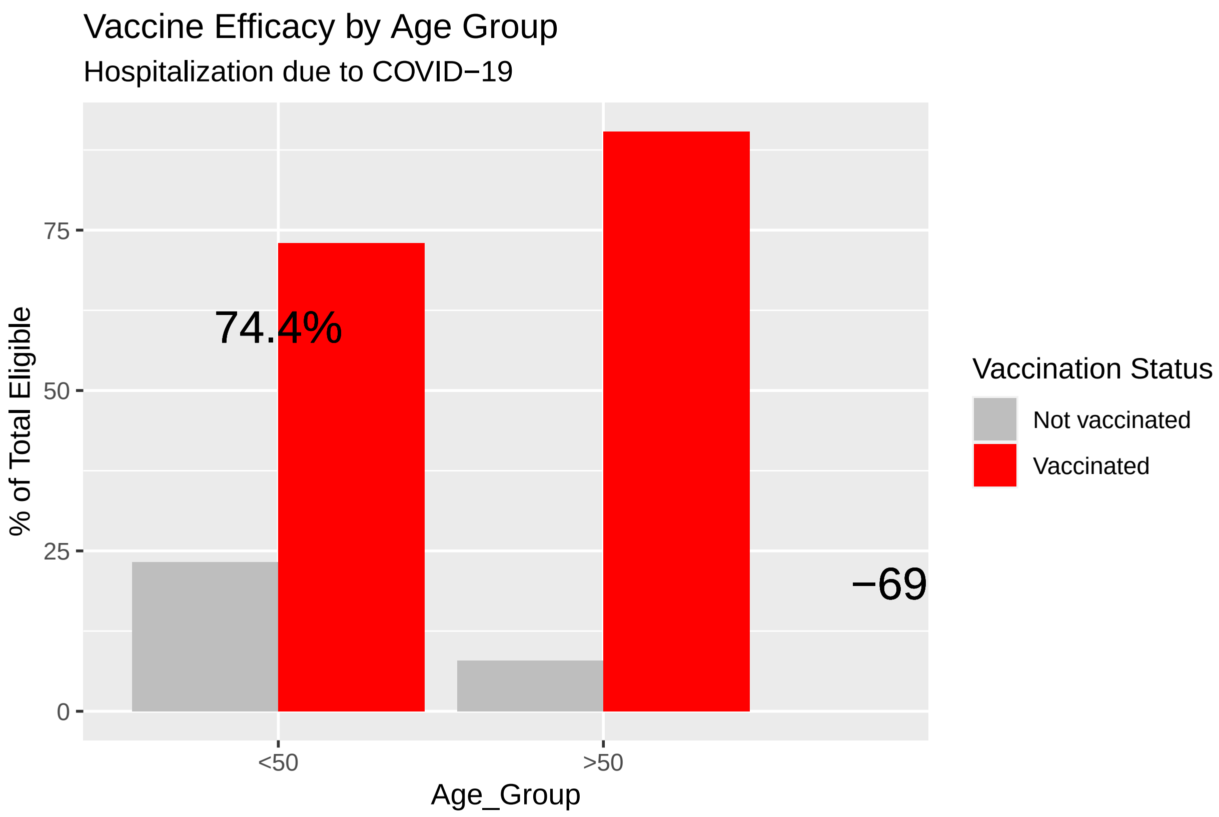
```
## Vaccine efficacy for individuals under 50 years old:  74.4 %
```

```
print(efficacy_under50)
```

```
## [1] 74.4186
```

The vaccine efficacy for individuals under 50 years old is 74.4%, which means that the vaccine is highly effective in preventing hospitalization due to COVID-19 in this age group. On the other hand, the vaccine efficacy for individuals over 50 years old is -69.6%, which means that the vaccine may not be effective in preventing hospitalization due to COVID-19 in this age group, but this result should be interpreted with caution due to the small sample size and other potential confounding factors.

```
# Create a bar plot
ggplot(df, aes(x = Age_Group, y = Percent_of_Total_Eligible, fill = Vax_Status)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("gray", "red"), labels = c("Not vaccinated", "Vaccinated")) +
  labs(title = "Vaccine Efficacy by Age Group",
       subtitle = "Hospitalization due to COVID-19",
       y = "% of Total Eligible",
       fill = "Vaccination Status",
       caption = "Data source: Tidied Israeli Vaccination Data Analysis") +
  geom_text(aes(x = 1, y = 60, label = paste0(round(efficacy_under50, 1), "%")), size = 6) +
  geom_text(aes(x = 3, y = 20, label = paste0(round(efficacy_over50, 1), "%")), size = 6)
```



Data source: Tidied Israeli Vaccination Data Analysis

(3) From your calculation of efficacy vs. disease, are you able to compare the rate of severe cases in unvaccinated individuals to that in vaccinated individuals? vaccine efficacy only measures the relative reduction in the risk of disease or severe cases between vaccinated and unvaccinated individuals, but it does not provide information on the absolute rates of disease or severe cases in each group. Therefore, to compare the rates of severe cases between vaccinated and unvaccinated individuals, we need to analyze the data in a different way.

We can calculate the incidence rate of hospitalization per 100,000 individuals in each vaccinated and unvaccinated group, then compare them. We will also need to account for potential confounding factors that may affect the rates, such as age and underlying health conditions.

```
# Load the data
df <- data.frame(
  Age_Group = c("<50", "<50", ">50", ">50"),
  Eligible_Population = c(1116834, 3501118, 186078, 2133516),
  Vax_Status = c("N", "Y", "N", "Y"),
  Hospitalized_per_100K = c(43, 11, 171, 290)
)

# Calculate the number of hospitalized individuals in each group
df$Hospitalized <- df$Eligible_Population * df$Hospitalized_per_100K / 100000

# Calculate the incidence rate of hospitalization per 100,000 individuals in each group
df$Incidence_Rate <- df$Hospitalized / df$Eligible_Population * 100000

# Create a contingency table of vaccinated vs. unvaccinated individuals and their hospitalization statu
```

```r
cont_table <- table(df$Vax_Status, df$Hospitalized > 0)

# Perform the chi-square test
chisq.test(cont_table)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  cont_table
## X-squared = 0, df = 1, p-value = 1
```