Analysis of Deaths Involving COVID-19 by Vaccination Status

Rahiq Raees, Saba Sheykh Hassani,and Parsa Moghaddamcharkari

2024-03-27

Description of the Variables and Data {#1}

```
##       _id            date                          age_group
##  Min.   :   1   Min.   :2021-03-01 00:00:00.00   Length:7812
##  1st Qu.:1954   1st Qu.:2021-12-04 18:00:00.00   Class :character
##  Median :3906   Median :2022-09-09 12:00:00.00   Mode  :character
##  Mean   :3906   Mean   :2022-09-10 03:07:27.93
##  3rd Qu.:5859   3rd Qu.:2023-06-15 06:00:00.00
##  Max.   :7812   Max.   :2024-03-27 00:00:00.00
##  deaths_boost_vac_rate_7ma deaths_full_vac_rate_7ma
##  Min.   :0.00000           Min.   :0.00000
##  1st Qu.:0.00000           1st Qu.:0.00000
##  Median :0.00000           Median :0.00000
##  Mean   :0.02642           Mean   :0.02664
##  3rd Qu.:0.01000           3rd Qu.:0.01000
##  Max.   :0.81000           Max.   :1.97000
##  deaths_not_full_vac_rate_7ma
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.2931
##  3rd Qu.: 0.0400
##  Max.   :18.3300
```

This dataset reports the daily reported number of the 7-day moving average rates of Deaths involving COVID-19 by vaccination status and by age group from March 1st, 2021 to March 20th, 2024.

_id: Row number of data

date: Date on which the death occurred

age_group: Age group with levels 0-4yrs, 5-11yrs, 12-17yrs 18-39yrs, 40-59yrs, 60+, and ALL

deaths_boost_vac_rate_7ma: 7-day moving average of the last seven days of the death rate per 100,000 for those vaccinated with at least one booster

deaths_full_vac_rate_7ma: 7-day moving average of the last seven days of the death rate per 100,000 for those fully vaccinated

deaths_not_full_vac_rate_7ma: 7-day moving average of the last seven days of the death rate per 100,000 for those not fully vaccinated

"Not fully vaccinated" category includes people with no vaccine and one dose of double-dose vaccine.

Background of the Data

https://data.ontario.ca/dataset/deaths-involving-covid-19-by-vaccination-status

The data is collected by Public Health Units across Ontario. A public health unit is a government organization under the supervision of a local board of health. These PHUs are under the direction of a Medical Officer of Health, who is appointed by the supervising board of health. PHUs continually clean up COVID-19 data and enter them into CCM. CCM is a dynamic disease reporting system which allows ongoing update to data previously entered. As a result, data extracted from CCM represents a snapshot at the time of extraction and may differ from previous or subsequent results.

Public health units collect vaccination data for the following reasons:

- **Monitoring Effectiveness:** This includes assessing how well vaccines prevent infection, transmission, as well as how long their protection lasts.
- **Safety Monitering:** Gathering data on adverse reactions or side effects.
- **Public Confidence:** Accurate information about vaccine efficacy builds trust with communities which promotes an increase in vaccinations among citizens.

Research Question

How does COVID-19 mortality vary among vaccinated and not fully vaccinated individuals in Ontario, considering factors such as age?

Tables {#2}

year	death_rate_boost_vac	death_rate_full_vac	death_rate_not_full_vac	total_death_rate
2021	6.93	6.93	223.16	237.02
2022	96.22	96.22	1686.49	1878.93
2023	34.57	34.57	266.78	335.92
2024	6.49	6.49	39.34	52.32

The above table displays the total death rate per 100,000 by vaccination type for years 2021 to 2024. Note that there are many repeat deaths since each obervation in our data set is the 7-day moving average of the last seven days of the death rate per 100,000.

We see that the covid death rate peaked in 2022 and has drastically decreased since then.

age_group	death_rate_boost_vac	death_rate_full_vac	death_rate_not_full_vac	total_death_rate
60+	139.09	139.09	2156.79	2434.97
40-59yrs	4.41	4.41	49.29	58.11
18-39yrs	0.71	0.71	6.96	8.38
0-4yrs	0.00	0.00	1.26	1.26
12-17yrs	0.00	0.00	1.19	1.19
5-11yrs	0.00	0.00	0.28	0.28

The table displays the total death rate per 100,000 by vaccination type for years each age group.

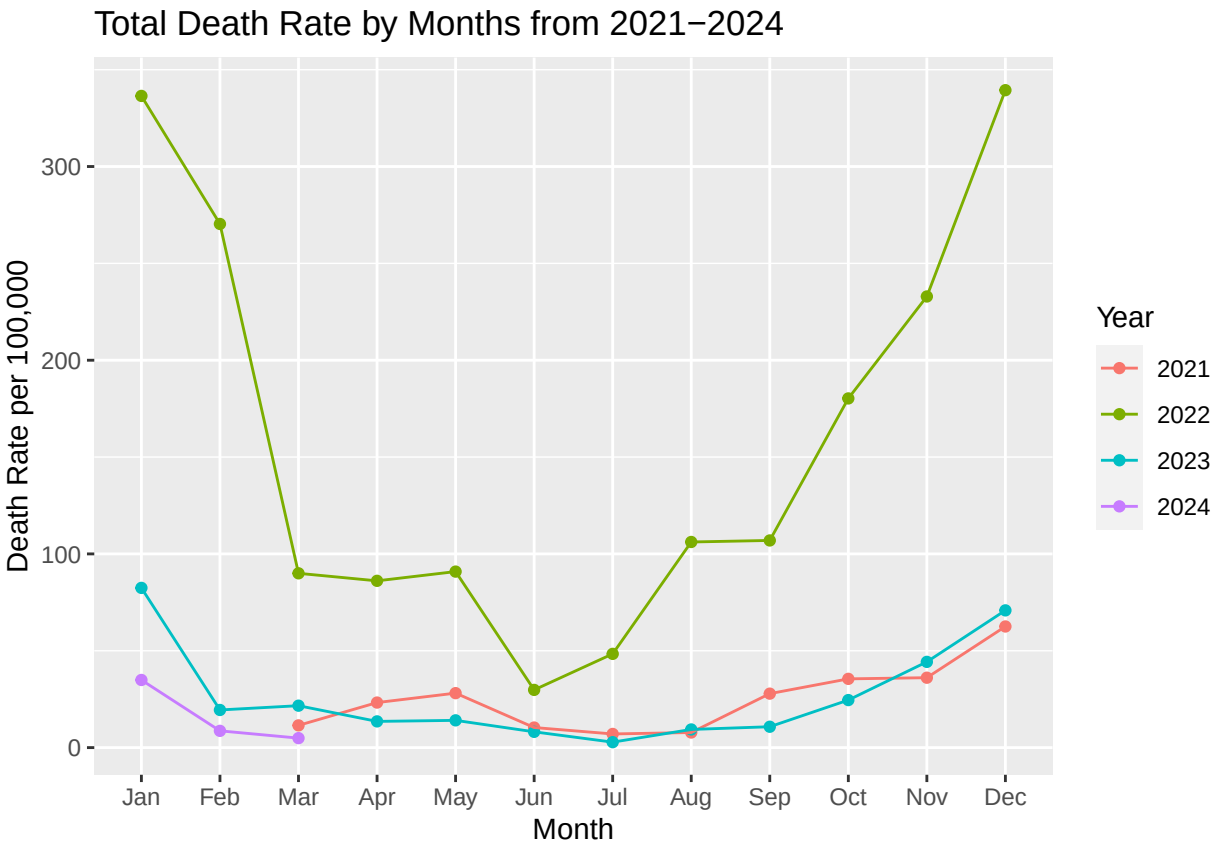
We see that the covid death rate is notably higher for 60+ year olds compared to all other age groups. The general trend is that younger people suffer less deaths than older people.

year	total (not_full - boost) vac diff	total (not_full - full) vac diff	total (full - boost) vac diff
2021	216.23	203.23	13.00
2022	1590.27	1551.55	38.72
2023	232.21	246.25	-14.04
2024	32.85	36.72	-3.87

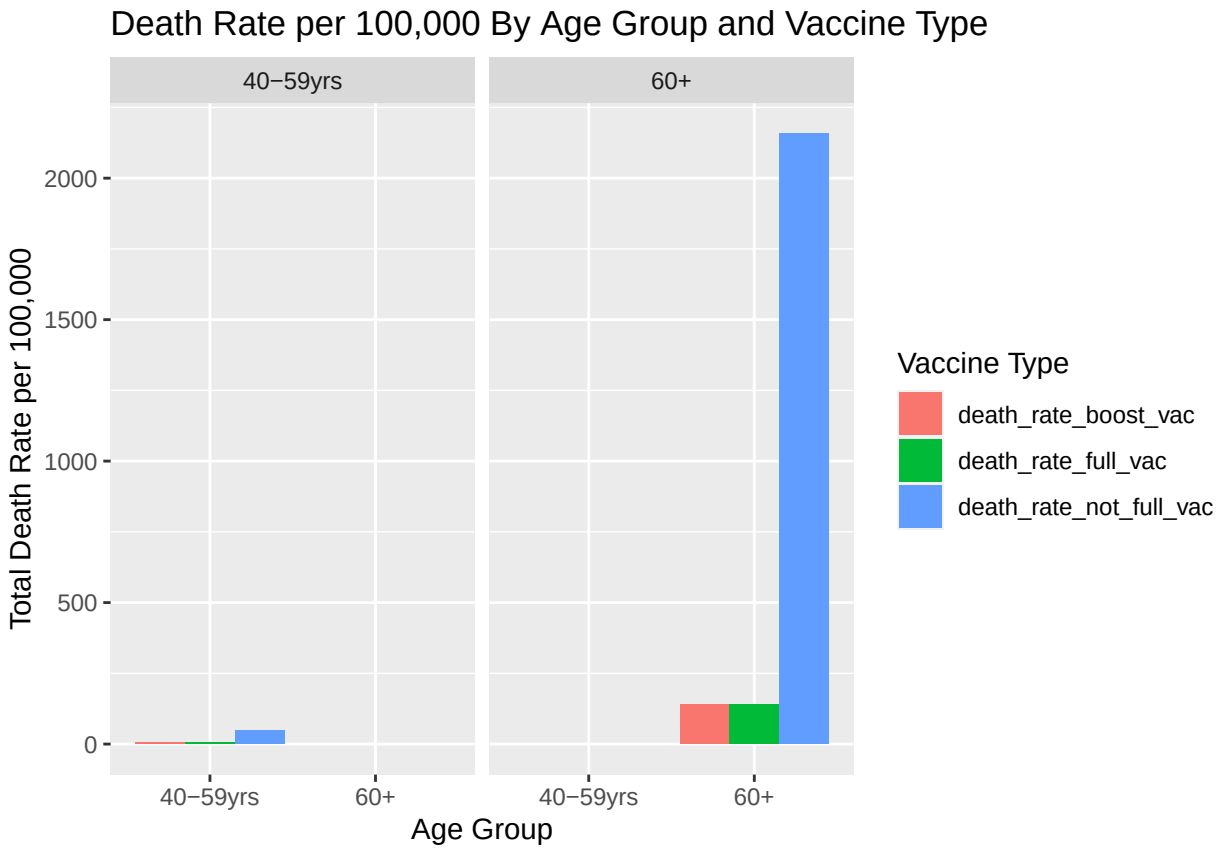
The table displays the difference in death rates for each type of vaccine among each age group.

The differences are almost all positive which shows the effectiveness of the vaccine in reducing the death rate. The first two columns have large positive differences which shows the dangers of not being fully vaccinated. The "total (not_full - boost) vac diff" column and "total (not_full - full) vac diff" are similar on a yearly basis which suggests that the getting the booster vaccine potentially may not drastically change your chance of survival if you're already fully vaccinated. This is also supported by the "total (full - boost) vac diff" column having much smaller values than the other two. Furthermore, a surprising result is that this column has negative values for 2023 and 2024 which discredits the effectiveness of the booster vaccine.

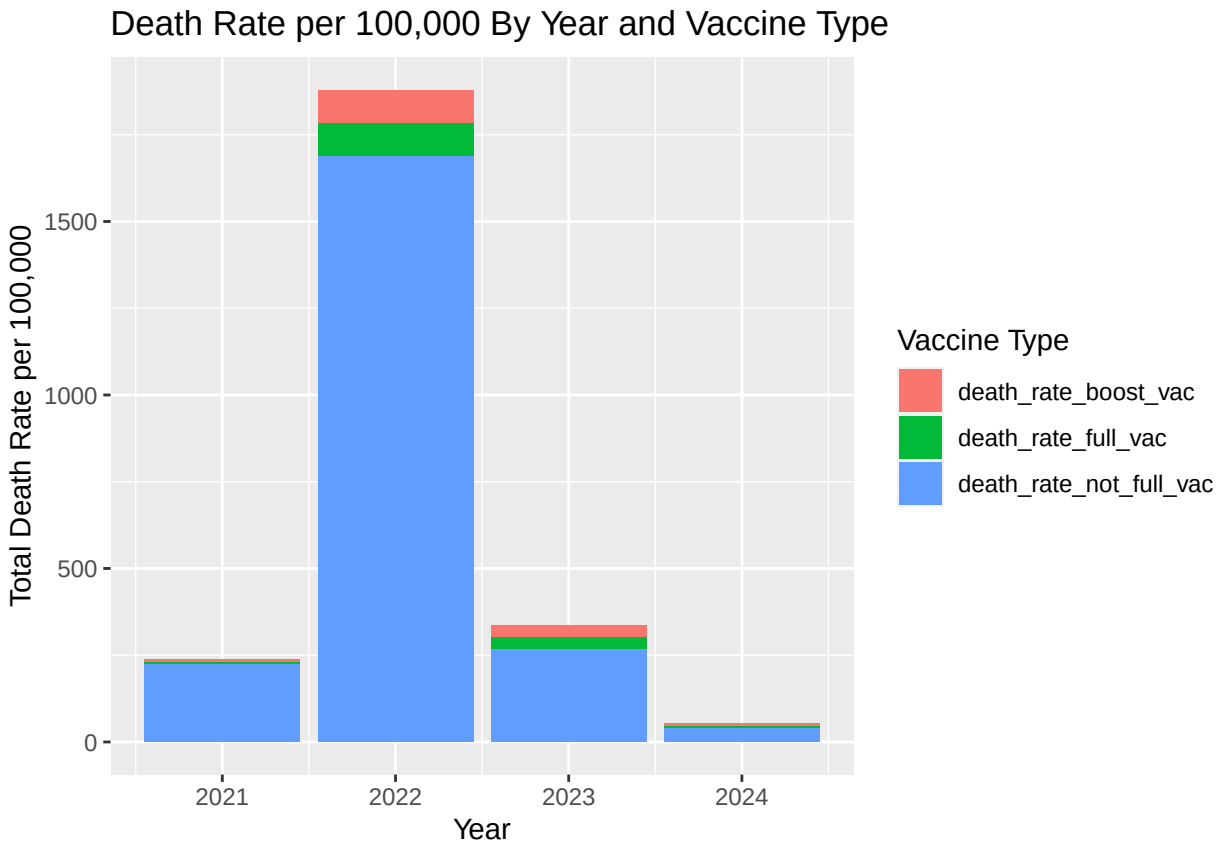
Graphs {#3}



The line chart shows that 2022 has been the worst year in terms of covid deaths. 2021 and 2023 are quite similar, and fortunately 2024 is off to a good start. The general pattern that we see is that the death rate is associated with temperature. The warmer months experience a dip in the death rate while the colder months experience a surge. This makes sense since cold weather encourages people to spend more time indoors in close proximity to others, and some research shows that our immune systems are weakened in cold temperatures.



The bar plot shows the dangers of not being fully vaccinated since the blue bars are much higher than the rest. We also learn about how vulnerable the elderly population since the bars for 40-59 year olds look almost negligable compared to the bars for 60+ year olds.



From the bar plot, we find that a vast majority of the deaths come from people who were not fully vaccinated. We also see that the total death rate in 2022 far exceeds the death rates in 2021 and 2023. 2021 and 2023 death rates were similar.

Hypothesis Testing and Confidence Intervals {#4}

Let μ_b be the average death rate for persons who received the booster vaccine, μ_f be the average death rate for fully vaccinated individuals, and μ_n be the average death rate for not-fully vaccinated individuals

We want to test the following:

- $H_0 : \mu_f - \mu_b = 0$ vs. $H_1 : \mu_f - \mu_b > 0$
- $H_0 : \mu_n - \mu_b = 0$ vs. $H_1 : \mu_n - \mu_b > 0$
- $H_0 : \mu_n - \mu_f = 0$ vs. $H_1 : \mu_n - \mu_f > 0$

Testing $H_0 : \mu_f - \mu_b = 0$ vs. $H_1 : \mu_f - \mu_b > 0$:

```
##
##  Welch Two Sample t-test
##
## data:  d$deaths_full_vac_rate_7ma and d$deaths_boost_vac_rate_7ma
## t = 2.8936, df = 10708, p-value = 0.001908
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.00218111        Inf
## sample estimates:
##  mean of x  mean of y
## 0.02661384 0.02155928
```

Our p-value is $0.001829 < 0.05$, thus we reject the null hypothesis and conclude that the average death rate for persons who are fully vaccinated is greater than the death rate for persons who received the booster.

Testing $H_0 : \mu_n - \mu_b = 0$ vs. $H_1 : \mu_n - \mu_b > 0$:

```
##
##  Welch Two Sample t-test
##
## data:  d$deaths_not_full_vac_rate_7ma and d$deaths_boost_vac_rate_7ma
## t = 18.126, df = 6723.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2815893       Inf
## sample estimates:
##  mean of x  mean of y
## 0.33125579 0.02155928
```

Our 95% confidence interval is $(0.283378, \infty)$, hence we reject the null hypothesis since the interval does not contain 0. We conclude that the average death rate for persons who are not fully vaccinated is greater than the death rate for persons who received the booster.

Testing $H_0 : \mu_n - \mu_f = 0$ vs. $H_1 : \mu_n - \mu_f > 0$:

```
##
##  Welch Two Sample t-test
##
## data:  d$deaths_not_full_vac_rate_7ma and d$deaths_full_vac_rate_7ma
## t = 17.784, df = 6793.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2764616       Inf
## sample estimates:
##  mean of x  mean of y
## 0.33125579 0.02661384
```

Our 95% confidence interval is $(0.2781953, \infty)$, hence we reject the null hypothesis since the interval does not contain 0. We conclude that the average death rate for persons who are not fully vaccinated is greater than the death rate for persons who are fully vaccinated.

Logistic Regression {#5}

We created a new column called total_death_rate that takes the value 1 if the sum of the death rates for each vaccine is greater than 0, and 0 otherwise. This will be our outcome variable in our logistic regression.

```
##
## Call:
## glm(formula = total_death_rate ~ age_group, family = binomial,
##     data = d)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.52734  -0.34138  -0.19518   0.00013   2.81790
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.8134     0.1297 -21.690  < 2e-16 ***
## age_group12-17yrs  -1.1378     0.2557  -4.451 8.56e-06 ***
```

```
## age_group18-39yrs   2.0035     0.1450  13.813  < 2e-16 ***
## age_group40-59yrs   3.6066     0.1450  24.881  < 2e-16 ***
## age_group5-11yrs   -1.1378     0.2557  -4.451 8.56e-06 ***
## age_group60+       21.3795   194.7274   0.110    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8656.4  on 6688  degrees of freedom
## Residual deviance: 3658.5  on 6683  degrees of freedom
## AIC: 3670.5
##
## Number of Fisher Scoring iterations: 17
```

From the R output, we see that the p-values for age_group5-11yrs, age_group12-17yrs, age_group18-39yrs, and age_group40-59yrs are all very close to zero, hence we conclude that these are significant predictors. From analyzing the p-values we see that age_group is a significant predictor for the most part.

Interpreting regression parameters:

- The log odds of having a total death rate greater than 0 when the age group measurement is age0-4yrs is -2.8067
- The log odds of having a total death rate greater than 0 when the age group measurement is age12-17yrs is $(-2.8067) + (-1.1381) = -3.9448$
- The log odds of having a total death rate greater than 0 when the age group measurement is age18-39yrs is $(-2.8067) + (2.0059) = -0.8008$
- The log odds of having a total death rate greater than 0 when the age group measurement is age40-59yrs is $(-2.8067) + (3.6033) = 0.7966$
- The log odds of having a total death rate greater than 0 when the age group measurement is age5-11yrs is $(-2.8067) + (-1.1381) = 0.7966$
- The log odds of having a total death rate greater than 0 when the age group measurement is age60+ is $(-2.8067) + (21.3728) = 18.5661$

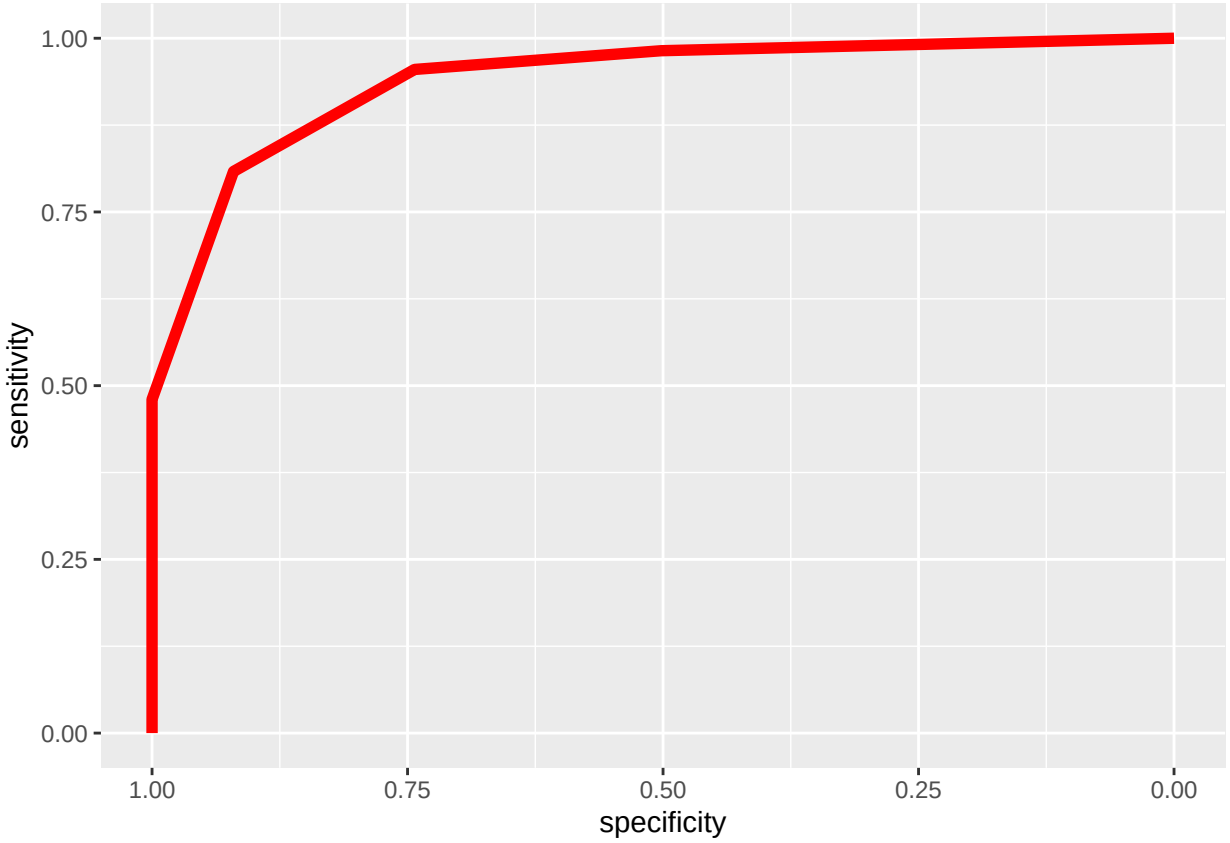
Bootstrap Confidence Interval for β_3 or for adults ages (5-11) {#6}

We perform a bootstrapped 95% confidence interval test for adults.

```
##       2.5%      97.5%
## -1.7124159 -0.6915422
```

Thus, we are 95% confident that the difference in logs odds for age group age5-11yrs and age0-4yrs is between (-1.7125006, -0.6940476)

ROC Curve {#7}




```
## Area under the curve: 0.9386
```

The area under our ROC is 0.9384 which is quite close to the total area of 1, thus we conclude that our model has a very good discrimination ability.

K-fold Cross Validation {#8}

```
## [1] 0.9363645 0.9342610 0.9437735 0.9368863
```

```
## [1] 0.9378213
```

we get 4 different summary measure of the model performance. By taking the average, we get that the out of sample predictive ability of our model is 0.9380531.

Summary of Research

1. Average death rates are significantly higher for persons that have a received a lower amount of vaccines compared to persons who have received more. This is shown from our hypothesis testing, hence vaccines are effective.
2. Year 2022 experienced a drastically higher amount of covid deaths in comparison to other years.
3. The COVID-19 death rate is notably higher for 60+ year olds compared to all other age groups. The general trend is that younger people suffer less deaths than older people.
4. A vast majority of the deaths come from people who were not fully vaccinated.

Appendix

Description of the Variables and Data

```
library(tidyverse)
library(knitr)
library(pROC)
d <- read_csv("covid_data.csv")

summary(d)
```

Tables

```
d <- d  %>% filter(!age_group=='ALL')
# Removed the ALL age group observations
#Because it actually did not represent all age groups.
#Thus, it was an unknown factor to us.

d<-d %>% mutate(date_var = as.Date(date),
                year=year(date_var),
                month=month(date_var),
                day=day(date_var)) # Creating columns for year,month, and day

t1 <- d  %>% group_by(year) %>%
  summarise(death_rate_boost_vac =
              sum(deaths_boost_vac_rate_7ma),
              death_rate_full_vac =
              sum(deaths_boost_vac_rate_7ma),
              death_rate_not_full_vac =
              sum(deaths_not_full_vac_rate_7ma),
              total_death_rate =
              death_rate_boost_vac +
              death_rate_full_vac +
              death_rate_full_vac)
kable(t1)
```

```
t2 <- d %>% group_by(age_group) %>%
  summarise(death_rate_boost_vac =
              sum(deaths_boost_vac_rate_7ma),
              death_rate_full_vac =
              sum(deaths_boost_vac_rate_7ma),
              death_rate_not_full_vac =
              sum(deaths_not_full_vac_rate_7ma),
              total_death_rate =
              death_rate_boost_vac +
              death_rate_full_vac +
              death_rate_not_full_vac)  %>%
  arrange(desc(total_death_rate))

kable(t2)
```

```
d <- d %>% mutate(`(not_full - boost) vac diff` =
                    deaths_not_full_vac_rate_7ma-deaths_boost_vac_rate_7ma,
                  `(not_full - full) vac diff` =
```

```
                    deaths_not_full_vac_rate_7ma-deaths_full_vac_rate_7ma,
                  `(full - boost) vac diff` =
                    deaths_full_vac_rate_7ma-deaths_boost_vac_rate_7ma)

t3 <- d  %>% group_by(year) %>%
  summarise(`total (not_full - boost) vac diff` = sum(`(not_full - boost) vac diff`),
            `total (not_full - full) vac diff` = sum(`(not_full - full) vac diff`),
            `total (full - boost) vac diff` = sum(`(full - boost) vac diff`))


kable(t3)
```

Graphs

```
d <- d %>% mutate(month_name = month(date,label=T), .after=month)

g1 <- d %>% group_by(year,month_name) %>%
  summarise(total_death_rate =
              sum(deaths_boost_vac_rate_7ma) +
              sum(deaths_full_vac_rate_7ma) +
              sum(deaths_not_full_vac_rate_7ma))

ggplot(g1,aes(x=month_name,y=total_death_rate,col=factor(year),group=factor(year))) +
  geom_point()+geom_line() +
  labs(x='Month', y= 'Death Rate per 100,000',
       title = 'Total Death Rate by Months from 2021-2024',col='Year')
```

```
g2 <- t2 %>% slice(1:2) %>%
  select(!total_death_rate) %>%
  pivot_longer(cols=c('death_rate_boost_vac',
                      'death_rate_full_vac',
                      'death_rate_not_full_vac'),
               names_to='variable', values_to="value")


ggplot(g2, aes(x = age_group, y = value, fill = variable)) +
  geom_bar(position = "dodge", stat = "identity") +
  facet_wrap(~age_group) +
  labs(x='Age Group', y='Total Death Rate per 100,000',
       title='Death Rate per 100,000 By Age Group and Vaccine Type',
       fill='Vaccine Type')
```

```
g3 <- t1 %>% select(!total_death_rate) %>%
  pivot_longer(cols=c('death_rate_boost_vac',
                      'death_rate_full_vac',
                      'death_rate_not_full_vac'),
               names_to='variable', values_to="value")

ggplot(g3, aes(x = year, y = value, fill = variable)) +
  geom_bar(position = "stack", stat = "identity") +
  labs(x='Year', y='Total Death Rate per 100,000',
       title='Death Rate per 100,000 By Year and Vaccine Type',
       fill='Vaccine Type')
```

Hypothesis Testing and Confidence Intervals

```
t.test(d$deaths_full_vac_rate_7ma,d$deaths_boost_vac_rate_7ma,alternative = "greater")
```

```
t.test(d$deaths_not_full_vac_rate_7ma,d$deaths_boost_vac_rate_7ma,alternative = "greater")
```

```
t.test(d$deaths_not_full_vac_rate_7ma,d$deaths_full_vac_rate_7ma,alternative = "greater")
```

Logistic Regression

```
d <- d %>% mutate(total_death_rate = case_when(deaths_boost_vac_rate_7ma +
                                                 deaths_boost_vac_rate_7ma +
                                                 deaths_not_full_vac_rate_7ma>0~1,TRUE~0))
```

```
m = glm(total_death_rate ~ age_group, family = binomial, data = d)
summary(m)
```

Bootstrap Confidence Interval for β_3 or for adults ages (18-39)

```
set.seed(123)
boot_function=function(){
  boot_data = d %>% sample_n(size = nrow(d), replace = T)
  m2 = glm(total_death_rate ~ as.factor(age_group),
           family = binomial, data = boot_data)
  s = coef(m2)[3]
  return(s)
}

out = replicate(1000, boot_function())
#Bootstrap Confidence Interval
quantile(out, c(0.025,0.975))
```

ROC Curve

```
logit.mod <- glm(total_death_rate ~ age_group , family = binomial, data = d)
p = predict(logit.mod, type = "response")
# ROC object
roc_logit = roc(d$total_death_rate ~ p)
# ROC plot
ggroc(roc_logit, color="red", size=2)
auc(roc_logit)
```

K-fold Cross Validation

```
k=4
d = d %>% mutate(group_ind = sample(c(1:k),
size=nrow(d),
replace = T))
c.index = vector()
for (i in 1:k){
  d.train = d %>% filter(group_ind != i)
  d.test = d %>% filter(group_ind == i)
  logit.mod = glm(total_death_rate ~ as.factor(age_group) ,
```

```
                  family = binomial, data = d.train)

  pi_hat = predict(logit.mod, newdata=d.test, type = "response")
  m.roc=roc(d.test$total_death_rate ~ pi_hat)
  c.index[i]=auc(m.roc)
}
c.index
mean(c.index)
```