Covid-19 spread speed and vaccination

Yao Lu

4/8/23

# 1. Summary/Abstract

*Write a summary of your project.* This paper (Chenchula, Karunakaran, Sharma, & Chavan, 2022) discuss the effectiveness of vaccine to Omicron variant.

Do regression for the relationship between the covid infection and vaccination. Maybe add some other variables like population, age, gender, race. Plan to use county as individual observation.

# 2. Introduction

## 2.1 General Background Information

*The main data we use is the vaccination data and the data of the COVID cases in Georgia. Then after we make a summary of the data. We try some models which could guide us to do further study.Our research question is, does the coverage of vaccine decrease the spread of COVID. Under the background of the pandemic, people generate the vaccine to prevent the spread of the COVID. How many effect do the vaccine have? Now, the method we use here is using crude model to briefly test the relationship between coverage of the vaccine and incerasing speed of the COVID.*

## 2.2 Description of data and data source

*Firstly, we load data and see if there is the need of clean and transformation*

All data wrangling process are recorded in code/processing\_code.

covid vaccination data comes from https://experience.arcgis.com/experience/3d8eea39f5c1443db1743a4cb8948a9c covid symptom and death data comes from https://ga-covid19.ondemand.sas.com/docs/ga\_covid\_data.zip

All the raw data are saved at raw\_data

## 2.3 Questions/Hypotheses to be addressed

Can booster decrease the speed of COVID spread?

# 3. Methods

Start from linear regression. Then may try CV/bootstraping.

## 3.1 Data import and cleaning

## 3.2 Statistical analysis

# 4. Results

## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Data exploration through tables

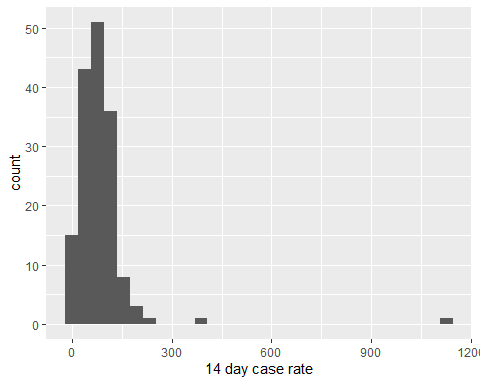
Vaccine-related tables

── Data Summary ────────────────────────  
 Values  
Name ...[]   
Number of rows 159   
Number of columns 5   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 numeric 5   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate mean sd p0 p25  
1 14 day case rate 0 1 83.4 97.3 0 41.8  
2 RTCUMVAXADMIN 0 1 127918. 27260. 60435 109456.   
3 PCTCUMPVAX 0 1 51.4 9.48 29.6 45.4  
4 PCTCUMPCVAX 0 1 46.9 8.88 24.9 41.5  
5 PCTBOOSTER 0 1 20.8 5.73 7 17   
 p50 p75 p100 hist   
1 74 104. 1126. ▇▁▁▁▁  
2 127981 143319 233717 ▂▇▇▁▁  
3 51.5 56.7 92.5 ▂▇▅▁▁  
4 47 52.1 83.1 ▂▇▆▁▁  
5 21 24 39 ▂▆▇▁▁

### 4.1.2 Data exploration through figures

Histogram plots for the continuous outcomes.

‘14 day case rate’ first.



Here we find there are two county which have high ‘14 day case rate’. We want to know which they are.

[1] "Quitman" "Stewart"

[1] 392.33

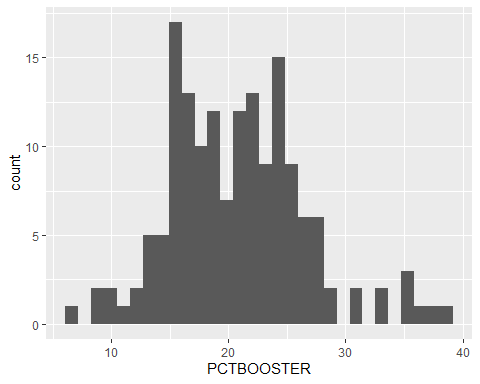
[1] 1125.8

They are ‘Quitman’ and ‘Stewart’ counties. In the previous 14 days/100,000 population, Quitman has 392.33 cases reported, and Stewart has 1125.8 cases reported. Beyond our current research, we should pay attention about what situations make these two counties have such a high level than other counties. Previous 14 days before 01/31/2023.

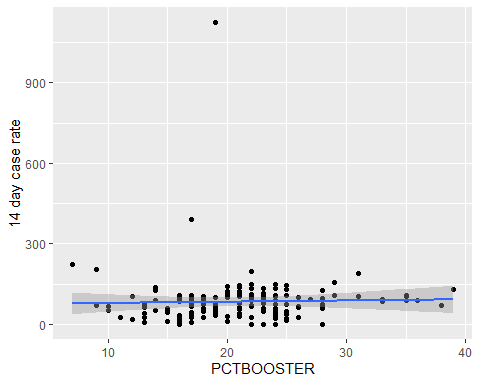
Now population, RTCUMVAXADMIN,PCTCUMPVAX,PCTCUMPCVAX and PCTBOOSTER. Since last four should be highly correlated. The analysis we will apply later should be same. Our final model should include two predictors which are population and one of these four.

Since the four choices share same statistics steps. We just pick population+PCTBOOSTER as example. I was thinking where could I find a categorical variables. Such as could I make population as ‘small’,‘medium’ and ‘large’. I think we can try both way how we define population. Since there is not other good choice of categorical variable I can find in this dataset. To make the diversity of our predictors. I decide to define population as categorical variables.(However,I am not saying this a better way.)

Now ‘PCTBOOSTER’.

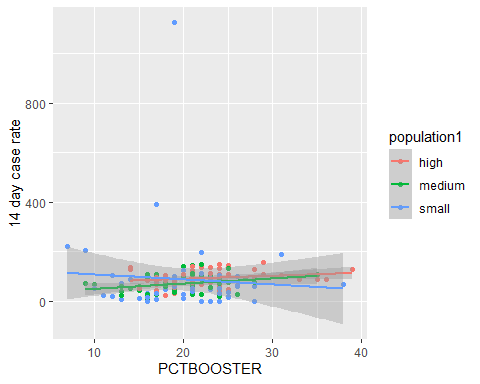


Now ‘14 day case rate’ as function of ‘PCTBOOSTER’.



Here we find that two outliers make our plot hard to see clearly. We can try another model removing these two outliers. Here I just leave it here to see what our model shows.

Once more cases as function of PCTBOOSTER, stratified by population.



### 4.1.3 Notes

Here we can see both the two outliers have small population.

## 4.2 Basic statistical analysis

Load the packages

Load the data.

Linear regression

Call:  
lm(formula = `14 day case rate` ~ PCTBOOSTER + population1, data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -88.77 -38.79 -7.89 19.12 1038.57   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 88.9698 34.4207 2.585 0.0107 \*  
PCTBOOSTER 0.1713 1.3962 0.123 0.9025   
population1medium -22.8006 19.4150 -1.174 0.2420   
population1small -4.9922 19.3080 -0.259 0.7963   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 97.68 on 155 degrees of freedom  
Multiple R-squared: 0.01056, Adjusted R-squared: -0.008594   
F-statistic: 0.5512 on 3 and 155 DF, p-value: 0.6481

Call:  
lm(formula = `14 day case rate` ~ PCTCUMPCVAX + population1,   
 data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -92.30 -38.83 -11.39 21.39 1032.29   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 126.5108 47.1870 2.681 0.00813 \*\*  
PCTCUMPCVAX -0.6718 0.9041 -0.743 0.45859   
population1medium -26.7886 19.5176 -1.373 0.17188   
population1small -8.4110 19.2612 -0.437 0.66295   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 97.51 on 155 degrees of freedom  
Multiple R-squared: 0.01397, Adjusted R-squared: -0.005112   
F-statistic: 0.7321 on 3 and 155 DF, p-value: 0.5343

These are the result from simple linear regression.

However, we would like to do linear regression with one predictor each time to explore the most significant ones. Since we haven’t find any significant even with the most possible significants ones.(fully vaccine and booster) So, we should expect we will have many diffculties. Here I will explore all variable in our data sets.

coln pvalue slope  
1 cases 0.446 0.000  
2 population 0.548 0.000  
3 confirmed\_case\_hospitalization 0.354 0.004  
4 deaths 0.476 0.016  
5 case rate 0.000 0.008  
6 death rate 0.902 0.006  
7 antigen\_cases 0.729 0.000  
8 probable\_deaths 0.692 -0.064  
9 antigen\_case\_hospitalization 0.431 -0.088  
10 VAXADMIN 0.572 0.000  
11 RTCUMVAXADMIN 0.996 0.000  
12 PERSONVAX 0.568 0.000  
13 PCTCUMPVAX 0.676 -0.343  
14 PERSONCVAX 0.569 0.000  
15 PCTCUMPCVAX 0.647 -0.401  
16 PERSONBVAX 0.580 0.000  
17 PCTBOOSTER 0.739 0.452  
18 population1 0.220 -23.303  
19 popusmall 0.771 -5.484

From one to one regression, only ‘case rate’ have a significant p-value. And all vaccine relevant predictors are irrelevant to “14 day case rate”. From here, I will adjust my predictors from vaccine-related to “case rate”.

And population1 is a categorical variable, so the single p-value is not enough to decide if should include it or not. So, we do anova F-test manually.

Analysis of Variance Table  
  
Response: mydata$`14 day case rate`  
 Df Sum Sq Mean Sq F value Pr(>F)  
mydata$population1 2 15635 7817.4 0.8245 0.4403  
Residuals 156 1479031 9481.0

Unfortunately, it is also insignificant. So, we should remove it in real data analysis. But here since I want to keep my predictors’ diversity. So, I assume it is significant.

Call:  
lm(formula = mydata$`14 day case rate` ~ mydata$`case rate` +   
 mydata$population1)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-283.73 -28.91 -7.44 13.14 796.83   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -93.727057 26.334946 -3.559 0.000494 \*\*\*  
mydata$`case rate` 0.008894 0.001135 7.836 6.92e-13 \*\*\*  
mydata$population1medium 6.980039 16.520571 0.423 0.673241   
mydata$population1small 30.332269 16.625454 1.824 0.070010 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 82.67 on 155 degrees of freedom  
Multiple R-squared: 0.2913, Adjusted R-squared: 0.2775   
F-statistic: 21.23 on 3 and 155 DF, p-value: 1.416e-11

Surprisingly, when we adjusted ‘case rate’ in our model, categorical population shows a little bit significance. Here is our initial conclusion. Vaccine cannot decrease the speed of COVID confirmed cases in 14 days. Since not all people who get infected will have test, there is bias. Here are the possible guess I made. Worst one, vaccine cannot prevent COVID. Neutral one, vaccine can smooth the symptom, but cannot reduce infection. Optimistic one, vaccine can smooth the symptom, but virus shows immune escape. The county who had experienced a high rate infection still keep a high ‘14 day case rate’, because the type of virus in that county escape the vaccine protection.

As the ‘case rate’ increase one unit, the ‘14 day case rate’ will increase 0.008894 unit. That is, if cumulative number of cases per 100,000 population increase one. Cases reported in the previous 14 days per 100,000 population will increase 0.008894. (t-value=7.836,p-value<0.001).

And the county with less population have higher ‘14 day case rate’. The initial guess is that, for the county with high population, people have already had herd immunity at the early stage of COVID. So, in 01/31/2023, the county with less population have higher ‘14 day case rate’.

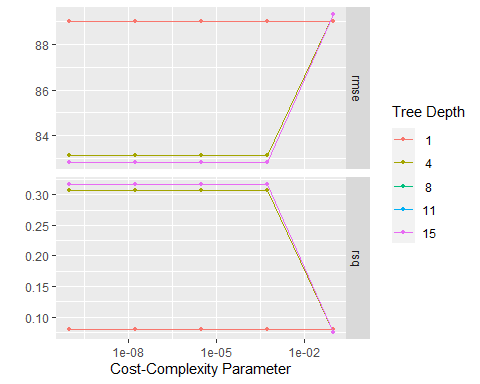
## 4.3 Full analysis

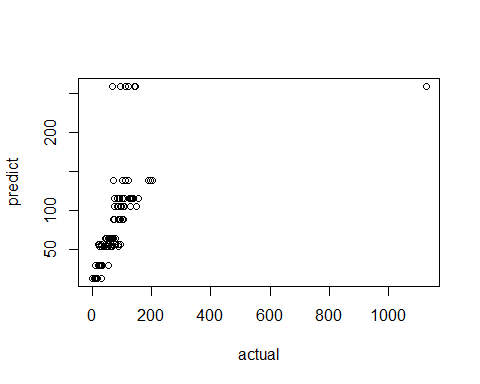
Here we will use machine learning method to analysis this data. We seperate our data for train and test strata by ‘14 day case rate’ which is our major interest outcome.

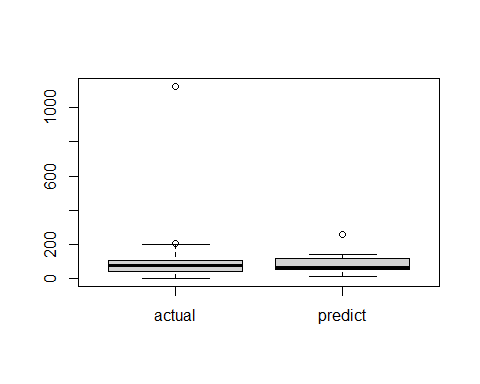
Then we check the performance of null model. If we just use mean, what is the rmse. Here we get if we use null model. Our rmse is 87.5.

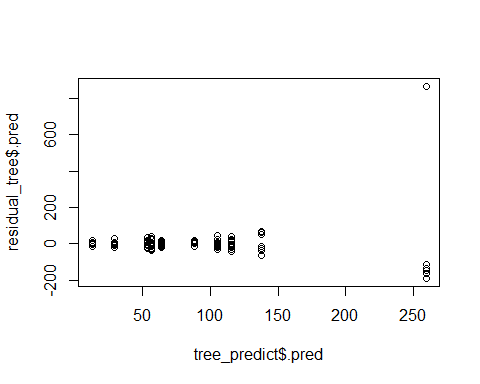
# A tibble: 1 × 6  
 .metric .estimator mean n std\_err .config   
 <chr> <chr> <dbl> <int> <dbl> <chr>   
1 rmse standard 80.5 25 14.6 Preprocessor1\_Model1

Then we setup three models to compare the performance of these models. The intermediate is hided. We will show important outcome which guide us to make further decision.

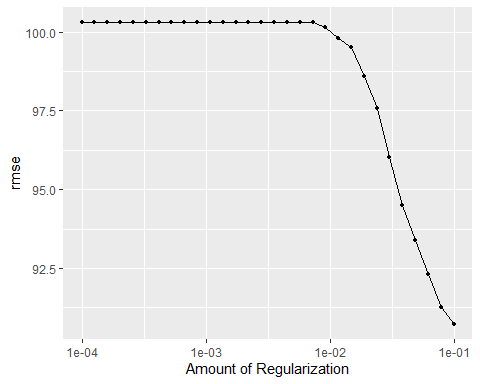


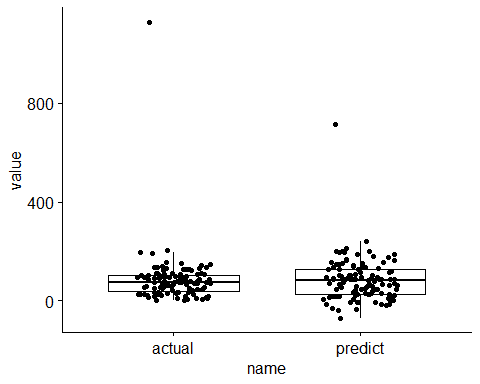


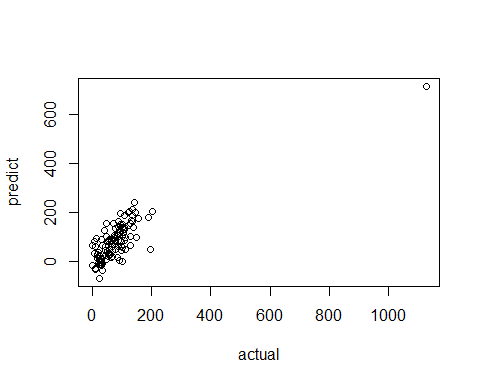


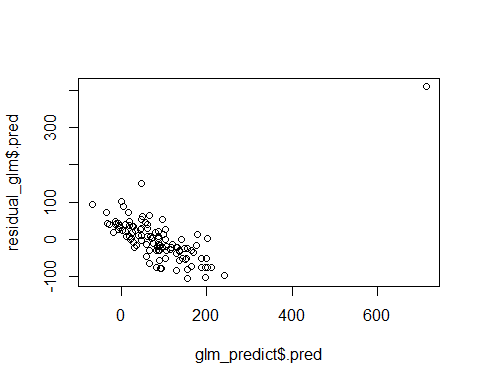


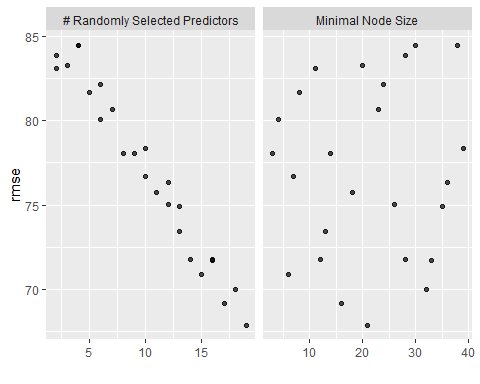
In our analysis using LASSO model, residual shows funnel shape. If we decide use LASSO as our final model, we should take the logarithm of our outcome. Or we can try getting the lambda in Box-Cox transformation. Since LASSO is not our best choice finally, so further step isn’t needed.

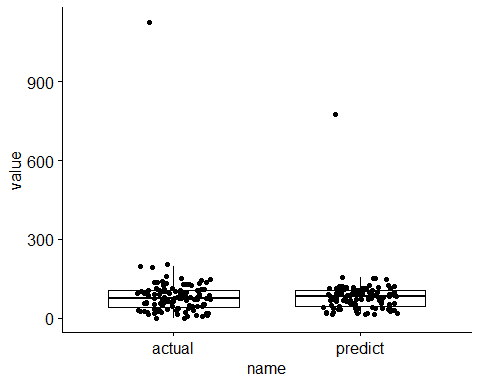


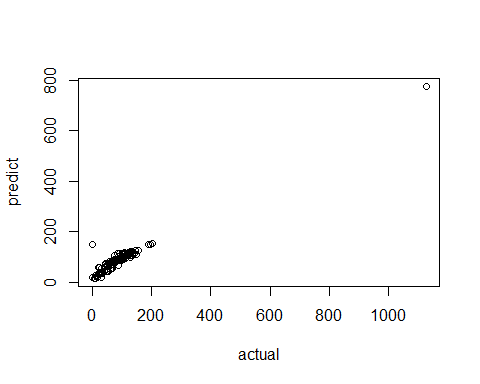


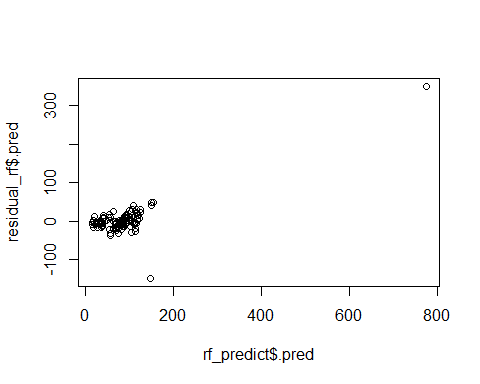








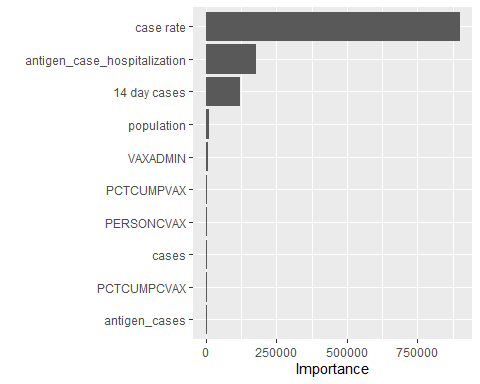


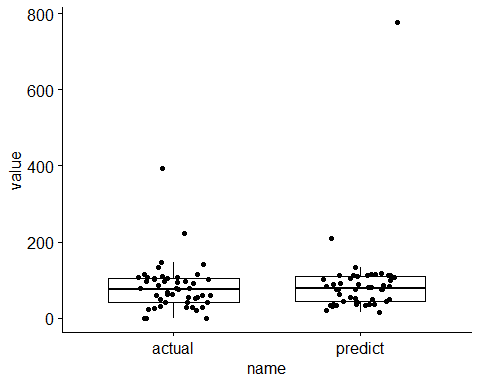


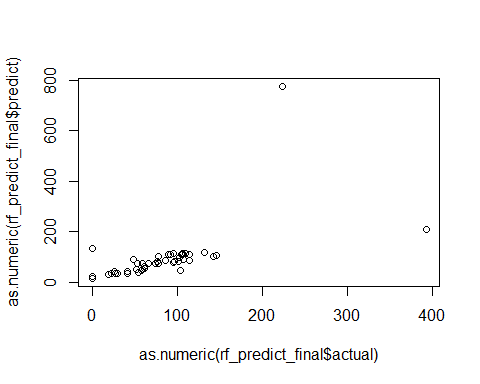
Here is the model comparision table.

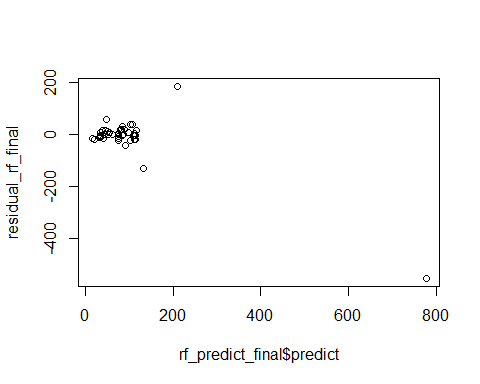
.metric .estimator mean n std\_err .config model  
1 rmse standard 80.51595 25 14.58213 Preprocessor1\_Model1 null  
2 rmse standard 82.83054 25 13.77485 Preprocessor1\_Model11 tree  
3 rmse standard 90.72367 25 10.17792 Preprocessor1\_Model30 lasso  
4 rmse standard 67.87188 25 14.63891 Preprocessor1\_Model05 random forest  
 lower upper 95%CI  
1 71.17457 92.70215 ( 71.1745696675898 , 92.7021512149816 )  
2 73.22063 95.36706 ( 73.2206280738665 , 95.3670639311271 )  
3 80.19800 104.45483 ( 80.1980010476191 , 104.45482501654 )  
4 59.99745 78.14439 ( 59.9974532329969 , 78.1443851096684 )

So, 95% CI of random forest for random forest looks the best. And we choose random forest model for this data.









The performance RMSE=81.3 RSQ=0.462. The predict vs actual plot looks fine. The residual plot looks fine except one outlier.

Base on the importance plot. We can see that ‘case rate’ plays an important role. Then the following important ones are ‘14 day cases’,‘antigen\_case\_hospitalization’,etc.

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References

Chenchula, S., Karunakaran, P., Sharma, S., & Chavan, M. (2022). Current evidence on efficacy of COVID-19 booster dose vaccination against the omicron variant: A systematic review. *J. Med. Virol.*, *94*(7), 2969–2976.