Covid-19 spread speed and vaccination

Yao Lu

4/21/23

# 1. Summary/Abstract

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

## 1.1 Importance

Many evidence shows that Covid-19 kill many people and influence people’s life quality.

## 1.2 Objective

Detect the importance of factors which influence the speed of COVID spread.

## 1.3 Result

‘Total Case rate’ is highly related with ‘14 day case rate’ at 01/31/2023. Second rank one is ‘antigen\_case\_hospitalization’. All vaccine related variables don’t have much importance.

## 1.4 Conclusion

In the county which has a high case rate. ‘14 day case rate’ will also be high at 01/31/2023.

# 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

Which factors influence the speed of COVID spread? We expect speed of COVID spread should be related with vaccine coverage.

# 3. Methods

## 3.1 Data import and clean dictionary

We combine these two sources by county. We have 159 counties information as our observation. Then we select part of the variables from the source. **County\_name**, cases(PCR confirmed cases + antigen positive cases), **population** (County population based on 2020 Georgia Population Estimates: https://opb.georgia.gov/census-data/population-projections), **confirmed\_case\_hospitalization** (Confirmed COVID-19 cases that were hospitalized at the time the case was reported to DPH or when the case was interviewed),**deaths**(Confirmed COVID-19 cases that were either reported to DPH as deceased by healthcare providers or medical examiners/coroners, identified by death certificates with COVID-19 indicated as the cause of death, or there is evidence that COVID-19 contributed to the individual’s death),**case rate**(Cases/100,000 population),**death rate**(Deaths/100,000 population),**14 day case rate**(Cases reported in the previous 14 days/100,000 population),**14 day cases**(Cases reported in the previous 14 days),**antigen\_cases**(Antigen cases include individuals with a positive antigen test. Only antigen test results are used in identifying these cases. These test results are reported through multiple sources including electronic lab reporting (ELR), State Electronic Notifiable Disease Surveillance System (SendSS), faxed case reports and calls from providers to DPH),**probable\_deaths**(Probable deaths include: individuals who were antigen positive or had compatible illness and known close contact to a case, and either they were reported to DPH as deceased by healthcare providers or medical examiners/coroners, they were identified by death certificates with COVID-19 indicated as the cause of death, or there is evidence that COVID contributed to their deaths; and individuals with a death certificate that has COVID-19 indicated as the cause of death and there is no laboratory evidence for SARS-CoV-2”),**antigen\_case\_hospitalization**(Antigen positive COVID-19 cases that were hospitalized at the time the case was reported to DPH or when the case was interviewed),**newid**(county ID),**VAXADMIN**(Number of vaccines administered (may include non-Georgia residents and persons with unknown residence)),**RTCUMVAXADMIN**(Vaccine administration rate: vaccines administered per 100,000 population (may include non-Georgia residents and persons with unknown residence)),**PERSONVAX**(Number of residents with at least one vaccine (does not include non-Georgia residents and persons with unknown residence)),**PCTCUMPVAX**(Percent of residents with at least one vaccination (does not include non-Georgia residents and persons with unknown residence)),**PERSONCVAX**(Number of residents that are fully vaccinated (does not include non-Georgia residents and persons with unknown residence)),**PCTCUMPCVAX**(Percent of residents that are fully vaccinated (does not include non-Georgia residents and persons with unknown residence)),**PERSONBVAX**(Number of residents that that have additional dose after fully vaccinated (does not include non-Georgia residents and persons with unknown residence)),**PCTBOOSTER**(Percent of residents that have additional dose after fully vaccinated (does not include non-Georgia residents and persons with unknown residence))

## 3.2 Statistical analysis

‘14 day case rate’ is the major outcome for my analysis. I fitted a basic linear model ‘14 day case rate’~percentage of booster+population of county. Here I categorize population to high,medium,low. The I did a pre-screen to all the variables base on p-value. I used machine learning engine to build four models which are null, tree,lasso,random forest. The compare all these model performance base on RMSE.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Data exploration through tables

[Table 1](#tbl-resulttable1) is the brief view of ‘14 day case rate’ and the vaccine related variables. We can see that it’s about half people have at least one vaccine. And about 20% people have a booster.

Table 1: Brief view of outcome and vaccine related variables

| skim\_variable | numeric.mean | numeric.sd | numeric.hist |
| --- | --- | --- | --- |
| 14 day case rate | 83.41528 | 97.262071 | ▇▁▁▁▁ |
| RTCUMVAXADMIN | 127918.37736 | 27259.909570 | ▂▇▇▁▁ |
| PCTCUMPVAX | 51.44340 | 9.484134 | ▂▇▅▁▁ |
| PCTCUMPCVAX | 46.93774 | 8.881083 | ▂▇▆▁▁ |
| PCTBOOSTER | 20.82390 | 5.734125 | ▂▆▇▁▁ |

### 4.1.2 Data exploration through figures

[Figure 1](#fig-result1) shows the detail of 14 day case rate distribution. It is almost normally distributed except two outliers. Then I check which county they are. 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.

|  |
| --- |
| Figure 1: 14 day case rate distribution |

Now population, RTCUMVAXADMIN,PCTCUMPVAX,PCTCUMPCVAX and PCTBOOSTER. Since last four should be highly correlated. The analysis we will apply later should be similar. Our exploratory model will 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 want to explore if population have linear influence or curve influence. So, I add a new population variable as ‘small’,‘medium’ and ‘large’. I think we can try both way how we define population.

|  |
| --- |
| Figure 2: Percentage of booster distribution |

[Figure 2](#fig-result2) is the percentage of booster distribution. Though the mean is about 20 percent, the peaks located at about 15 percent and 25 percent.

|  |
| --- |
| Figure 3: Percentage of booster distribution |

[Figure 3](#fig-result3) are two figures which show the trend between percentage of booster and 14 day case rate. With and without population stratification respectively. We can see that without stratification, the trend is not clear to say. With stratification, in high and medium population county, 14 day case rate increase as the percentage of booster increase. In small population county, 14 day case rate decrease as the percentage of booster increase. Here we can see both the two outliers have small population.

## 4.2 Basic statistical analysis

[Table 2](#tbl-resulttable2) is the basic linear regression with percentage of booster and categorical population. It shows that none of these variables are significant.

Table 2: Linear regression coefficient Booster and categorical population

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 88.970 | 34.421 | 2.585 | 0.011 |
| PCTBOOSTER | 0.171 | 1.396 | 0.123 | 0.903 |
| population1medium | -22.801 | 19.415 | -1.174 | 0.242 |
| population1small | -4.992 | 19.308 | -0.259 | 0.796 |

[Table 3](#tbl-resulttable3) is the basic linear regression with percentage of booster and numeric population.It shows that none of these variables are significant.

Table 3: Linear regression coefficient Booster and numeric population

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 78.710 | 30.724 | 2.562 | 0.011 |
| PCTBOOSTER | 0.131 | 1.497 | 0.088 | 0.930 |
| population | 0.000 | 0.000 | 0.507 | 0.613 |

[Table 4](#tbl-resulttable4) is the basic linear regression with percentage of fully vaccine and categorical population.It shows that none of these variables are significant.

Table 4: Linear regression coefficient fully vaccine and categorical population

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 78.710 | 30.724 | 2.562 | 0.011 |
| PCTBOOSTER | 0.131 | 1.497 | 0.088 | 0.930 |
| population | 0.000 | 0.000 | 0.507 | 0.613 |

Since all these three analysis shows large p values. So, I decide to seek another variables which can account for our outcome rather than vaccine related variables. 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.

[Table 5](#tbl-resulttable5) shows all the one to one regression results. 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.

Table 5: Prescreen table

| coln | pvalue | slope |
| --- | --- | --- |
| cases | 0.446 | 0.000 |
| population | 0.548 | 0.000 |
| confirmed\_case\_hospitalization | 0.354 | 0.004 |
| deaths | 0.476 | 0.016 |
| case rate | 0.000 | 0.008 |
| death rate | 0.902 | 0.006 |
| antigen\_cases | 0.729 | 0.000 |
| probable\_deaths | 0.692 | -0.064 |
| antigen\_case\_hospitalization | 0.431 | -0.088 |
| VAXADMIN | 0.572 | 0.000 |
| RTCUMVAXADMIN | 0.996 | 0.000 |
| PERSONVAX | 0.568 | 0.000 |
| PCTCUMPVAX | 0.676 | -0.343 |
| PERSONCVAX | 0.569 | 0.000 |
| PCTCUMPCVAX | 0.647 | -0.401 |
| PERSONBVAX | 0.580 | 0.000 |
| PCTBOOSTER | 0.739 | 0.452 |
| population1 | 0.220 | -23.303 |
| popusmall | 0.771 | -5.484 |

[Table 6](#tbl-resulttable6) shows in ANOVA table, categorical population is not significant. We should have removed it, but let’s keep it here to see if situation will change in future.

Table 6: Prescreen table

|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| mydata$population1 | 2 | 15634.83 | 7817.414 | 0.8245376 | 0.4403392 |
| Residuals | 156 | 1479031.03 | 9480.968 | NA | NA |

Surprisingly, [Table 7](#tbl-resulttable7) shows that when we adjusted ‘case rate’ in our model, categorical population shows a little bit significance.

Table 7: Prescreen table

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -93.727 | 26.335 | -3.559 | 0.000 |
| case\_rate | 0.009 | 0.001 | 7.836 | 0.000 |
| population\_medium | 6.980 | 16.521 | 0.423 | 0.673 |
| poplation\_high | 30.332 | 16.625 | 1.824 | 0.070 |

## 4.3 Full analysis(machine learning method)

Here we apply machine learning method to our entire data.We use 5-fold cross-validation, 5 times repeated. We use 4 different models which are null, tree, lasso and random forest. We separate our data for train(70%) and test(30%) strata by ‘14 day case rate’ which is our major interest outcome. Then we check the performance of each model. [Table 8](#tbl-resulttable8) is the final performance table. Here we think random forest model is the best model. Since comparing with null or tree model. They have close std\_err, but mean is lower than those. Comparing with lasso model, though lasso have less std\_err, their 95%CI is not overlapped. So, we only display the statistical result for random forest model only. Other model’s will be hided in **products/manuscript/machinelearning.R**.

Table 8: Prescreen table

| .metric | .estimator | mean | n | std\_err | model | 95%CI |
| --- | --- | --- | --- | --- | --- | --- |
| rmse | standard | 80.51595 | 25 | 14.58213 | null | ( 71.175 , 92.702 ) |
| rmse | standard | 82.83054 | 25 | 13.77485 | tree | ( 73.221 , 95.367 ) |
| rmse | standard | 90.72367 | 25 | 10.17792 | lasso | ( 80.198 , 104.455 ) |
| rmse | standard | 67.87188 | 25 | 14.63891 | random forest | ( 59.997 , 78.144 ) |

[Figure 4](#fig-result4) Top 2 are the real plot. And the bottom 2 are zoomed up ones. Looks like this model is acceptable. The prediction is good.

|  |
| --- |
| Figure 4: Actual vs predict value in random forest model |

[Figure 5](#fig-result5) Left is the real plot. And right one is zoomed up. Looks like fine except two outliers.

|  |
| --- |
| Figure 5: Residual plot in random forest model |

[Table 9](#tbl-resulttable9) is the performance table of random forest model.

Table 9: Prescreen table

| .metric | .estimator | .estimate |
| --- | --- | --- |
| rmse | standard | 86.9243257 |
| rsq | standard | 0.3405395 |

[Figure 6](#fig-result6) is the importance plot. Here we can find case\_rate is the most important factor. Second important one is antigen\_case\_hospitalization. Third is 14 day cases which we should be careful. Since our outcome is 14 day cases devided by population with appropriate units.

|  |
| --- |
| Figure 6: Importance plot |

[Figure 7](#fig-result7) is predict and actual value in cross-validation plot. Looks like the prediction is fine.

|  |
| --- |
| Figure 7: Predict and actual value in cross-validation |

In [Figure 8](#fig-result8), left is the real plot. Right is zoomed up. Looks like the residual is fine except few outliers.

|  |
| --- |
| Figure 8: Residual plot in cross-validation |

# 5. Discussion

## 5.1 Summary and Interpretation

### 5.1.1 Basic analysis

Here is our initial interpretation from basic linear analysis. Vaccine coverage is not related with 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’.

### 5.1.2 Machine learning

In the random forest model. We can find case\_rate is the most important factor. Second important one is antigen\_case\_hospitalization. Third is 14 day cases which we should be careful. Since our outcome is 14 day cases devided by population with appropriate units. All the statistical plot looks fine in random forest model. But there are still some outliers. So, there may exist some important factor which we miss. Such as if people get together in those county close before 01/31/2023.

## 5.2 Strengths and Limitations

The strength of this analysis, the model prediction is fine to predict the ‘14 day case rate’. And checking plots are OK as well. Limitation is that we still not discover the mechanism of COVID spread. Many references show evidence that vaccine can protect people. Vaccine can smooth the symptoms. Maybe RNA virus are so unstable. They get variation so quick to hide itself from vaccine.

## 5.3 Conclusions

As this model shows, total case rate is related with 14 day case rate. One hypothesis is the virus get vary to escape the vaccine protection. So, if there is a spread-powerful virus appear in those counties, the speed of spread will keep high in those counties. And this data also shows that, vaccine coverage related variables don’t have much importance to the speed of spread.

# 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.