Infections and Vaccine Rates in Different Regions

An Examination of Worldwide COVID-19 Data

Weiqiang Wang
The Department of Computer Science
University of Colorado, Boulder
Boulder Co United States
wewa2282@colorado.edu

Reid Pritchard
The Department of Computer Science
University of Colorado, Boulder
Boulder CO, United States
repr0811@colorado.edu

Haifeng Jiang
The Department of Computer Science
University of Colorado, Boulder
Boulder CO United States
haji9034@colorado.edu

Abstract:

It has been 1.5 years since the the World Health Organization (WHO) declared covid-19 is a pandemic, and one year since covid-19 vaccine has despite into markets. So far, there are more than 49.8 million confirmed cased and 796 thousand death cases. Our project is trying to compare the vaccine rate in the past one year in the United States, and study the effect of different vaccines. However, our result cannot represent the efficacy of vaccines; The confirmed cases and death cases seems to decrease overall, while the number of both confirmed and death cases suddenly increase since September. One reasons for this unexpected change is that we did not considered about the mutation of virus and the imagiration between different states. In order to have more accurate result, we need to considered the data for imagiration between different states, and compare our results with other pandemic in the past, such as SARS in 2003.

1 Introduction

Coronaviruses are a large family of viruses that are known to cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). A novel coronavirus (COVID-19) was identified in 2019 in Wuhan, China. Due to the wild-spread nature of the COVID-19 and the varietas of COVID-19, vaccine effectiveness is increasingly important when monitoring COVID-19 and it's variants. The difference between hospitalization rates and transmission rates can also give insight to how different variants work and help predict the need for booster/third vaccine administration. Finding trends in data can also help prepare for high occupancy of ICU beds and the need for preventative measures to be put in place. The dataset we are using contains the numbers of confirmed covid-19 cases, number of deaths from covid-19, number of recover cases from covid-19 cases, total vaccination, ICU admission rates, and the number of current occupancy rate for covid 19. After processing the data, we are trying to use machine learning to build a model to predict the new covid-19 cases of several regions.

On the other hand, the number of migration between states and international migration are hard to perceive and predict. As a result, the confirmed cases will have larger increments as the number of migrations increase, such as September, and end of November. Furthermore, the mutation of covid-19 is also unpredictable, and the vaccine efficacy for COVID-19 variant is unknown. While, we believe comparing the increase of confirmed cases and vaccine rate can be considered as one way to observe the possible start date of COVID-19 variant.

2 Related work

Due to the global impact of COVID-19, there are lots of examples of data analysis. The first example is one many people in the U.S. have probably used to keep track of COVID-19 over the past months. The Center for Disease Control and Prevention has collected and visualized data for the United States. Their website[1] has a huge variety of visualization from line graphs to heat maps allowing users to drill down into state data or track total country tallies. While this site is a great example of the potential effectiveness of data visualizations, the data is only focused around the United States, leaving out the global impact. Using larger global datasets we are creating similarly styled graphs on a broader scale. We are also trying to incorporate machine learning to predict COVID-19 cases on a local scale.

Another site that has done an excellent job with data analysis and visualization is Our World in Data. COVID-19 data is far from their first attempt at data visualization and analysis and thus the graphs found are quite informative. However, missing is the ability to "stack" graphs to look for possible causal relationships. While this is a small missing feature, when it comes to vaccine effectiveness, especially as new variants come into play, finding and predicting correlated data is increasingly important.

Finally, a study by Serkan Ballı is exemplary work in the COVID-19 data analysis field. In this study, they used, at the time, up-to-date data to visualize and predict COVID-19 cases in the US, Germany, as well as global numbers. The one downside to this study is simply how fast COVID has changed over the past year. The study focuses on data from 01/20/2020 to 09/18/2020, as vaccines roll out globally and the delta variant spreads faster and faster, it's clear this study is in need of an update.

3 Methodology

Dataset overview

We use the COVID-19 dataset from GitHub which records the time series data tracking the number of people affected by COVID-19 worldwide, including the number of people who died of coronavirus, the number of confirmed tested cases of coronavirus, and the number of people who have recovered from coronavirus. We only use the US data from this dataset, which contains one million rows and 6 columns from 2020/01/22 to 2021/10/15.

COVID-19 Vaccinations in the United States is a dataset directly from the Centers for Disease Control and Prevention. This dataset records the total number of distributed doses and the number of administered vaccines from each state from 2020/12/20 to 2020/12/12. There are four different kinds of vaccines: Janssen, Moderna, Pfizer, and unknown manufacturers. We will figure the administered vaccine rate of different brands in different states, and evaluate the effectiveness of each vaccine. There are 23.3K rows, and 80 columns in total.

COVID-19 Reported Patient Impact and Hospital Capacity by Facility is another dataset that comes from CDC. Due to the larger number of records and complex features in this dataset, we will only cite the average number of occupied inpatient beds reported during the 7-day period from 2020/12/20 to 2020/12/12.

Data preprocess

We have 3 datasets: COVID-19 dataset, COVID-19 Vaccinations in the United State, and COVID-19 Reported Patient Impact and Hospital Capacity from the CDC dataset. Since we are trying to explore the correlation between covid-19 cases and vaccine rate in different States, and there were no records of vaccinations

before 12/14/2020, we only select information from 01/01/2021 to the newest data in the dataset, which is 10/10/2021. For the convenience of future work, we uniform the formation of date into Year-month-day.

Then we look over each dataset and remove some unrelated columns or columns with a larger number of error values; In the COVID-19 Vaccinations dataset, the number of Delivered doses per 100,000 census population and its sub-columns, and the Total number of doses administered per 100,000 census population based on the jurisdiction where the recipient lives. The features left in the Vaccinations data are Date, Sate, MMWR week, number of different Vaccinations, and the number of people who are fully vaccinated, 16 features in total. For the COVID-19 Reported Patient Impact and Hospital Capacity from the CDC dataset, we only keep the date, locations, and the average number of a total number of staffed inpatient ICU beds that are occupied reported in the 7-day period, due to the inconsistency of other data, and a large number of error or empty data in this dataset.

For the covid19 cases dataset, its columns include Date, State, Confirm case, Death Case. In order to have a clear and easy way of understanding the increase of covid19 cases, we create a new column that simply presents that daily increasing covid19 case from 01/01/2021 to 10/15/2021.

Because the original dataset is too large to upload and manipulate with other members, each dataset is divided into a smaller size dataframe type and group by the State and date so that the operations will be much faster for each member. “states_data” is a dictionary type variable that contains the data of covid19 cases in one State. ”vac_states” is a dictionary type variable that contains the data of covid19 cases in one State. “hosp” is a dictionary type variable that contains the data of the number of ICU beds in one State.

Overall, our final data frame only contain 299 rows and 17 columns: 'Date', 'daily_increase', 'Confirmed','Deaths','MMWR_week','Location','Administered','Administered_Janssen','Administered_Moderna','Administered_Pfizer','Administered_Unk_Manuf','Administered_Dose1_Recip','Series_Complete_Yes','Series_Complete_Pop_Pct','Series_Complete_Janssen','Series_Complete_Moderna', 'Series_Complete_Pfizer','Series_Complete_Unk_Manuf', for each states.

Data process

After preprocessing, first we got the following table data. We processed data from six states, Arizona, Colorado, California, Florida, New York and Washington. Here we take California as an example. Then, We visualized part of the data.

Table 1, we can see the overall information.

	Date	MMWR_week	Location	Distributed	Distributed_Janssen	Distributed_Moderna	Distributed_Pfizer	Distributed_Unk_Manuf	Administered	Administe
45	2021-10-09	40	CA	59,727,285	2,830,400	22,768,300	34,128,585	0	51,722,146	
97	2021-10-08	40	CA	59,589,885	2,807,100	22,724,380	34,038,405	0	51,583,481	
168	2021-10-07	40	CA	59,306,315	2,801,600	22,690,320	33,814,395	0	51,458,972	
238	2021-10-06	40	CA	59,085,375	2,800,300	22,643,840	33,641,235	0	51,335,279	
256	2021-10-05	40	CA	58,979,155	2,783,500	22,609,860	33,585,795	0	51,250,507	
...	...	...	...	...	...	...	...	...	...	
18868	2020-12-18	51	CA	326,625	0	0	0	0	273	
18939	2020-12-17	51	CA	326,625	0	0	0	0	273	
18995	2020-12-16	51	CA	195,000	0	0	0	0	4	
19045	2020-12-15	51	CA	148,200	0	0	0	0	0	
19105	2020-12-14	51	CA	33,150	0	0	0	0	0	

300 rows × 22 columns

Table 2, we highlight the COVID-19 information that different hospitals have been exposed to on different dates.

	Date	state	hospital_name	city	hospital_subtype	total_beds_7_day_avg	all_adult_hospital_beds_7_day_avg	all_adult_hospital_inpatient_be
0	2021-05-28	CA	PORTERVILLE DEVELOPMENTAL CENTER	PORTERVILLE	Short Term	14.0	14.0	
1	2021-02-19	CA	SUTTER SURGICAL HOSPITAL-NORTH VALLEY	YUBA CITY	Short Term	14.0	14.0	
2	2021-01-22	CA	SUTTER SURGICAL HOSPITAL-NORTH VALLEY	YUBA CITY	Short Term	14.0	14.0	
3	2020-12-25	CA	STANISLAUS SURGICAL HOSPITAL	MODESTO	Short Term	23.0	23.0	
4	2020-11-13	CA	HEALTHBRIDGE CHILDREN'S HOSPITAL-ORANGE	ORANGE	Childrens Hospitals	27.0	0.0	
...	...	...	...	...	...	...	...	
21925	2020-07-31	CA	SAN ANTONIO REGIONAL HOSPITAL	UPLAND	Short Term	431.9	312.9	
21926	2020-07-31	CA	SUTTER COAST HOSPITAL	CRESCENT CITY	Short Term	43.7	51.0	
21927	2020-07-31	CA	NORTHBAY MEDICAL CENTER	FAIRFIELD	Short Term	209.1	208.0	
21928	2020-07-31	CA	PROVIDENCE SAINT JOHN'S HEALTH CENTER	SANTA MONICA	Short Term	153.5	153.5	

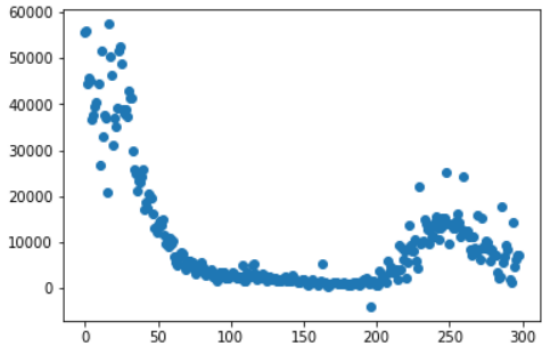
In Table 3, we highlight the daily increase, the number of confirmed cases, and the number of

deaths.

	Date	Confirmed	Deaths	MMWR_week	Location	Distributed	Distributed_Janssen	Distributed_Moderna	Distributed_Pfizer	Distributed_Unk_Manuf	...	Ac
0	2020-12-14	1669913	21153	51	CA	33,150	0	0	0	0	...	
1	2020-12-15	1725399	21488	51	CA	148,200	0	0	0	0	...	
2	2020-12-16	1781355	21832	51	CA	195,000	0	0	0	0	...	
3	2020-12-17	1825730	22152	51	CA	326,625	0	0	0	0	...	
4	2020-12-18	1871302	22441	51	CA	326,625	0	0	0	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
295	2021-10-05	4751206	69492	40	CA	58,979,155	2,783,500	22,609,860	33,585,795	0	...	
296	2021-10-06	4757349	69663	40	CA	59,085,375	2,800,300	22,643,840	33,641,235	0	...	
297	2021-10-07	4764554	69820	40	CA	59,306,315	2,801,600	22,690,320	33,814,395	0	...	
298	2021-10-08	4771626	69966	40	CA	59,569,885	2,807,100	22,724,380	34,038,405	0	...	
299	2021-10-09	4773284	69990	40	CA	59,727,285	2,830,400	22,768,300	34,128,585	0	...	

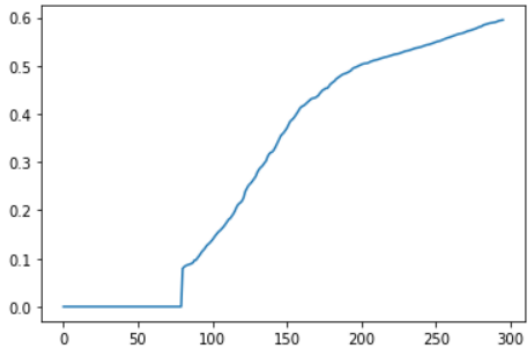
300 rows × 24 columns

At the same time, in order to better observe the trend of daily new cases, we use a scatter plot to visualize the daily new cases.



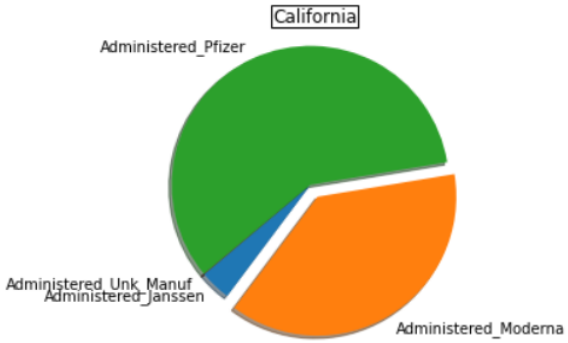
Judging from the trend of this scatter plot, it is not difficult for us to judge that the number of new cases per day has dropped from tens of thousands to thousands, which is directly related to the administration of vaccines.

As the vaccine is administered, we have also drawn a line chart to observe the vaccine injection rate in California.



This line chart shows the population that has been vaccinated in California as a percentage of the total population. As of the date of the chart data, we see that the number of people vaccinated in California is close to 60%.

Based on the injection rate obtained, we plotted the proportion of different vaccines.



It is not difficult to see from the picture that the Pfizer vaccine and Modena are the main vaccines.

4 Evaluation

In our project, we first have processed all the data: total vaccination, covid-19 confirmed cases, new covid-19 cases, and hospital ICU admissions rate in different regions. Then we have visualized all the data we have to make us have a better understanding of the trends of covid-19 cases, total vaccination and ICU admission rate of covid-19 patients over time. We are trying to conclude a summary of daily covid-19 new cases, the relationship of vaccination rate, covid-19 death cases, and recovered cases in the 2-3 different cities.

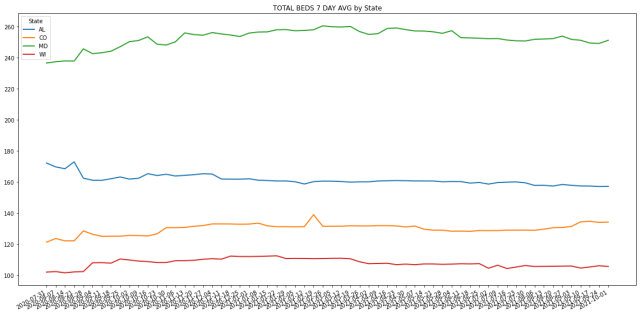
We have finished analyzing parts of data, we started to use machine learning to build up a model to predict the daily new cases, and ICU admission rates in different regions based on the vaccine rate and confirmed cases in the past 1 year. In the perdition program, we will import RandomForestClassifier, AdaBoostClassifier or Decision Tree from scikit-learn library in Python. The features of the first part of our program use total vaccination, covid-19 confirmed cases, and new covid-19 cases to predict the number

of new covid-19 cases in 3 days. So far, we are trying to use the ICU admissions rate and our predictions of new covid-19 cases to calculate the possibility of ICU bed shortage in different regions.

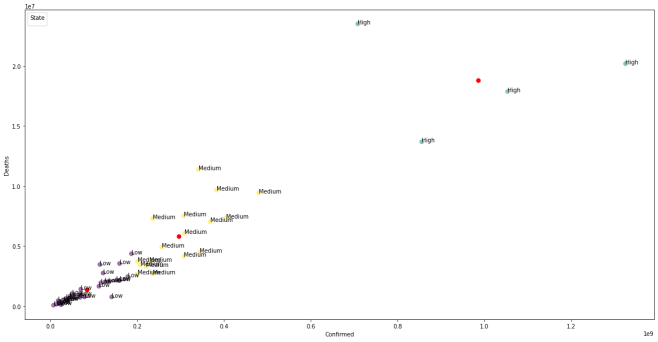
After we finish our predict program, we will take the data from the 3 cities in the United States as features and get the predictions. We will verify our predictions with really new cases, and calculate the accuracy of our program.

5 Discussion:

Graph: Comparing ICU bed occupancy over time

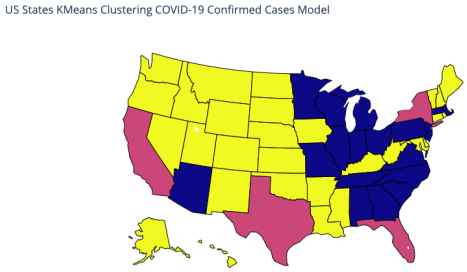


Graph: Categorizing and comparing states success in managing COVID



This initial visualization gave us an idea of which states had the best and worst COVID-19 responses, but the visualization wasn't the easiest to follow, so we recreated a similar KMeans grouping using a map view instead. Here we can find the states with the most COVID-19 Confirmed cases were New York, California, Texas, and Florida. From here we decided to look into the comparisons between these categories to look for possible places where things went wrong.

Map: Grouping COVID-19 Confirmed cases using KMeans clustering. This visualizes different state's success with COVID-19 restrictions and management



We began visualizing more detailed comparisons between states. Taking one state from each previous cluster, we ended up with California, Virginia, and Colorado in order from most cases to least.

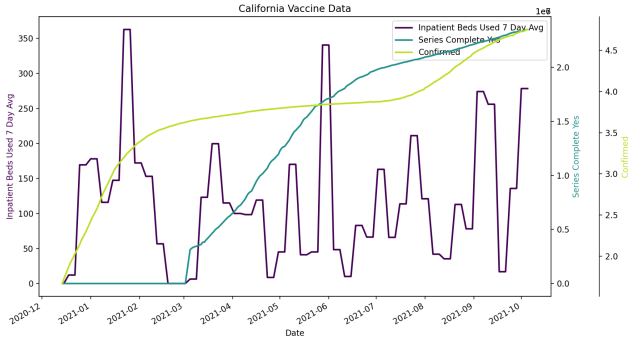
There was one main observation we found. We found the rate of cases before vaccines were released to be much greater in California than in Colorado. Virginia was in the middle of these two states. This was the first indication that COVID had a greater impact. This rate along with the vaccination rate can be compared when observing the intersection of the Confirmed case line and the Vaccine completion line in the graphs below. California's intersection was nearly a month after Colorado's intersection, while Virginia's took the middle ground. This intersection represents both the faster increasing case count as well as the slower vaccination completion rate, resulting in an overall later date.

Part of this can be explained by a much larger population in California and thus vaccine rollouts took more time to reach the critical mass. This also could account for the faster infection rate pre-vaccine.

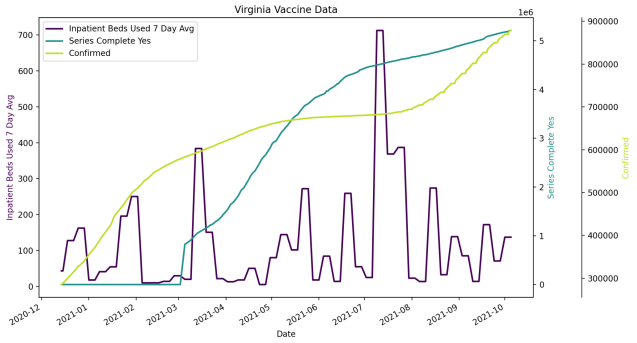
Another observation was the overall lower ICU usage in California than in both Colorado and Virginia. While you would expect a higher populated state with more cases to take up more hospital beds, these graphs tell a different story. This could be attributed to a larger geographical state, with more spread out hospitals and thus smaller amounts in each. However, the reason for the large difference in numbers is still unexpected. While this is one theory, it is completely possible other factors contribute to this difference,

including but not limited to possible errors in data collection or processing.

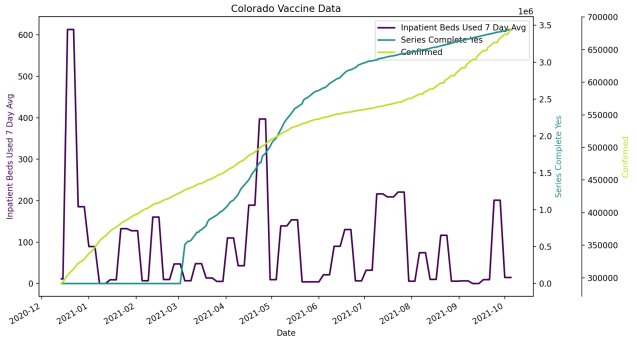
Graph: Plotting ICU usage, fully completed vaccines, and confirmed cases over time in California.



Graph: Plotting ICU usage, fully completed vaccines, and confirmed cases over time in Virginia.



Graph: Plotting ICU usage, fully completed vaccines, and confirmed cases over time in Colorado.

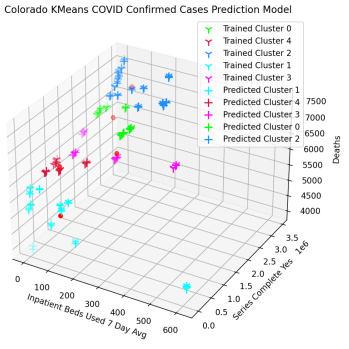


Attempting to find reasons why the discrepancy in ICU usage exists, we attempted to use KMeans to compare these states. First we trained the KMeans model on Colorado data, then used KMeans prediction on both California and Virginia's data. This was then visualized in a 3 dimensional graph, in hopes to highlight the differences between states.

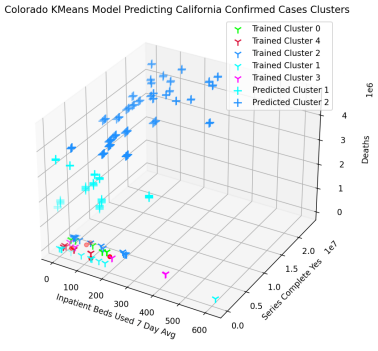
Using these graphs we found Colorado's high ICU usage could be attributed to low vaccination rates. Strangely Virginia's high ICU usage seemed to come even after the majority of vaccines had been completed. The one explanation for this is the reduction of COVID-19 restrictions as the state's total vaccine rate increased. Another reason could be schools returning to session. It's difficult to tell with high-level data and a more thorough analysis, with the inclusion of demographic or geographic data, would be needed to pinpoint what happened in Virginia around July and August this past year.

Overall these graphs show a huge difference between Colorado and California and a more minimal change between Colorado and Virginia. This is expected as the initial cluster map placed these three states in different categories, though it seems the top cluster has a much greater distance between it and the second cluster, while the first and second cluster are much more similar.

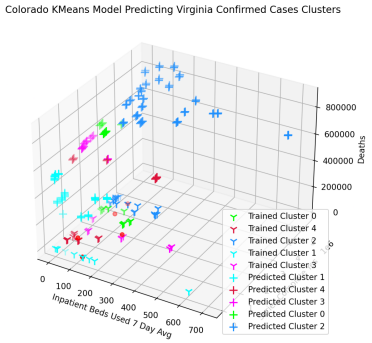
Graph: KMeans Confirmed Case Model trained on a subset of Colorado data and predicting Colorado Data.



Graph: Colorado KMeans Model used to predict California data for comparisons



Graph: Colorado KMeans Model used to predict Virginia data for comparisons



6 Conclusion:

The COVID-19 pandemic has lasted for almost 2 years, and it seems that it will continue to exist for a period of time due to its high infestation rate and multiple methods of infestation. Overall, the vaccines inhibit the rapid growth of confirmed cases and death cases for in 2021; The number of daily increase cases and daily death cases have dropped a lot since when vaccines have applied in the market. Meanwhile, these numbers seem to decrease as the rate of fully vaccinated people increases in different States. On the other hand, the COVID-19 vaccine doses administered by different manufacturers are very similar in different states, therefore, we are not able to judge the effectiveness of different vaccines.

Almost all the graphs show a rapid increase around September and July. In other words, the number of confirmed cases suddenly increases ,despite the increase of higher vaccine rate. Two factors for these points can be the migration of people and mutation of coronavirus. As we are writing this report, a new variant, omicron, has been reported in the US, and it has been found in 25 states. It seems like the old vaccines do not work well for this new variant, then the number of daily reported will continue to grow, but we still need more observation to evaluate if old vaccines can reduce the number of death with this omicron variant

Because we are lacking of dataset of migration and more detailed information of each patient,our results cannot represent the real effectiveness of vaccines. In order to have more accurate results, we will import the state-to-state migration dataset, and covid 19 variants. Further, we will also do the same research for SARS pandemic in 2003 to have a clear understanding of effectiveness of covid 19 vaccines.

REFERENCES

[1] CDC. “CDC Covid Data Tracker.” Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, October 6, 2021. https://covid.cdc.gov/covid-data-tracker/#datatracker-home

[2] Ritchie, Hannah, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. “Coronavirus Pandemic (COVID-19) - Statistics and Research.” Our World in Data, March 5, 2020. https://ourworldindata.org/coronavirus.

[3] Ballı S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos Solitons Fractals. 2021;142:110512. doi:10.1016/j.chaos.2020.110512