IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Identifying Factors Affecting the Spread of Variants of COVID-19 by Regression Models to Inform Policymaking

Peter Ma

Beijing No.4 High School International Campus No.2 XiTieJiang HuTong, XiCheng District Beijing, 100022, China, Peterma2020@outlook.com

ABSTRACT: Variants of SARS-CoV-2 continue to put stress on economies and healthcare worldwide. Governments are trying to take necessary steps to slow down the dynamic transmission of variants. In order to find out the potential key factors that drive the spread of variants, like demographics, health, lockdown, and vaccine-related factors, I proposed a study based on machine learning (ML) methodologies to construct predictive models. Those models revealed and ranked possible discriminatory features. What's more, as stated by The World Bank, since the shortage of vaccines is a major problem for the whole world, lockdown policies remain a significant part of the comprehensive strategy. Thus, I attempted to conduct regression analysis to find out the key pre-lockdown factors that could help maximize the effectiveness of lockdowns. This study seeks to figure out the discriminatory factors of variants and to guide strategies for combating variants based on those findings. Additionally, it provides guidance on the rollout of vaccines and how to maximize the efficacy of lockdowns to slow down the spread of variants of SARS-CoV-2.

KEYWORDS: Systems Software; Machine Learning; SARS-CoV-2; Variants, Spread; Policymaking.

Introduction

The coronavirus disease 2019 (COVID-19), which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread worldwide since the first known case was identified in December 2019.[1] On March 11, 2020, the World Health Organization (WHO) declared it as a pandemic.[1] At the time of writing, almost 200 million global cases have been confirmed and more than 4.23 million deaths have been reported according to the database of Johns Hopkins University.[2] The pandemic is not only a global health crisis, but has also caused unprecedented economic losses due to loss of life, business lockdowns, tourism disruptions, and reduced trade. It was estimated by the WHO that, due to the economic depression caused by the pandemic, nearly 820 million people are undernourished.[3]

Currently, there are two major strategies to handle the pandemic. The first strategy is based on the approved vaccines. For example, four COVID-19 vaccines (Moderna, Pfizer, Johnson & Johnson and AstraZeneca) have been authorized by Sweden.[4] Because of their high effectiveness, they are believed to be significant in halting the pandemic. For instance, the vaccine effectiveness of Pfizer is about 86% among the general adult population in Sweden.[5] The second strategy is national or regional lockdowns. At the beginning of the pandemic, lockdown efforts and social distancing were the major strategies for governments to handle the spreading of SARS-CoV-2 due to lack of efficient therapeutic treatments and vaccination strategies. Generally, the lockdown strategy works by cutting off the channels of transmission in the pandemic. Admittedly, these efforts mitigated the global health crisis to some extent. However, unprecedented economic losses were caused by those lockdown efforts, like business shutdowns, tourism disruptions, and reduced trade. Usually, most governments, like the U.S., preferred to lockdown nationally or regionally at the beginning of the pandemic to earn time for the development of vaccines. After the majority get vaccinated, governments are expected to handle the pandemic by building up herd immunity and reopening their countries.

Besides those two strategies, other demographic or social factors may curb the pandemic as well. Satyaki Roy applied several supervised machine learning algorithms and found that population density, testing numbers, and airport traffic emerge as the most discriminatory factors which affect COVID-19 infection rates.[6] Apart from that, Hannah *et al.* indicated that males are more likely to get COVID-19 than the female group according to meta-analysis,[7] which implies that gender may be a factor. Also, people with diabetes mellitus tend to have a higher chance to get infected.[8] All in all, because COVID-19 is relatively new to the whole world, it is still unknown what are the most efficient strategies to halt its spread. Fortunately, these strategies above flattened the pandemic curve successfully. Take the U.S. as an example. Thanks to the lockdown and vaccination, the three hundred thousand new cases per day in early Jan 2021 decreased to six thousand daily in 5 months.[9]

However, most countries in the world are experiencing another resurgence which is driven by variants of SARS-CoV-2. There are four kinds of variants of concern (Table 1). They are believed to be highly transmissible which increases hospitalizations and lead to higher mortality.[10] Thus, the emergence of the new variants posits several new problems to the current preventive strategies for global public health.

The first problem is that we need to create new models to determine how the mutations drive the transmissibility of SARS-CoV-2. In other words, we must examine if the

© 2022 Terra Science and Education

DOI: 10.36838/v4i3.14

previously established discriminatory factors, like gender and population density, are still critical in the transmissibility of the variants.

Table 1: Characteristics of four kinds of variants of concerns.

Variant	Date of First identification	Location of First identification	Characteristic mutations
B.1.1.7 (Alpha)	Dec 2020[10]	The United Kingdom[10]	1.There are 17 mutations in its genome among which 8 mutations form the basis of the 3 vaccines in the UK. [10] 2. 50% increased transmission.[13] 3. Increased hospitalizations and mortality.[13]
B.1.351 (Beta)[13]	Dec 2020[13]	South Africa[13]	1. 50% increased transmission.[13] 2. Greatly reduced susceptibility to monoclonal antibody treatment.[13]
B.1.617.2 (Delta)[13]	Dec 2020[13]	India[13]	1. increased transmission.[13] 2. potential reduced susceptibility to monoclonal antibody treatment.[13]
P.1 (Gamma)[13]	Jan 2021[13]	Brazil[13]	1. increased transmission.[13] 2. greatly reduced susceptibility to monoclonal antibody treatment.[13]

Secondly, it is significant for us to demonstrate the efficacy of current vaccines in the context of the new model. As shown in Table 1, there are mutations in the variants forming the basis of current vaccines, like the Alpha variant. So, those mutations will likely affect the efficacy of vaccines. For example, the AstraZeneca vaccine was deemed to be only 66% effective against the Alpha variant and 66% effective against the Delta variant after the second dose.[11] As a result, the reduced effectiveness will in return affect our current strategies which are mainly based on the high efficacy of vaccines. One of the questions most countries are facing is whether we need to lockdown again. In May 2021, several European countries started their new lockdown due to the variants at the cost of economic revival while some countries, like Norway, kept opening without regard to the quick spread of variants. Due to the shortage of scientific strategies, all governments are struggling in finding a balance between controlling the pandemic and reopening businesses. Therefore, a new model should be created to evaluate the effects of vaccines and lockdowns on the infection rates of variants to inform lockdown-vaccine-related policymaking.

Thirdly, identifying the pre-lockdown features which could greatly reduce the post-lockdown infected rates is extremely critical for some areas where they have limited number of vaccines and where lockdowns appear to be inevitable. WHO indicates that in parts of the third world, only less than 1% of the populations get vaccinated.[12] The low vaccination rate implies the importance of lockdowns to slow down the spreading of variants. So, national administrations should try to enhance the effectiveness of lockdowns by controlling the meaningful pre-lockdown features. A new model could help with figuring out those beneficial pre-lockdown features.

COVID-19 models are created using machine learning (ML) techniques and are used to predict the spread dynamics of COVID-19 successfully. Zoabi *et al*. analyzed 51,831 COVID-19 patients to study the effect of demographic factors, like gender, age, and social interactions on the transmissibility of COVID-19 and concluded that close social interactions are critical to the spreading based on ML approaches.[13] Satyaki Roy *et al*. applied several supervised ML approaches to examine the datasets of each U.S. state and they determined population density, testing rates, and airport traffic emerge as the discriminatory factors for their forecasting models.[6] Khan *et al*. used regression tree, cluster analysis, and principal component analysis (PCA) methodologies in their studies.[14] However, rare ML models for the purpose of predicting the variants of SARS-CoV-2 have been created or reported so far. Therefore, I hypothesize that discriminatory factors influencing the infection rates of the variants of COVID-19 could be identified by ML approaches.

While more and more knowledge about the spreading characteristics of COVID-19 have been investigated, there is no comprehensive study focusing on that of the variants. To solve this gap, I conducted this analysis and attempted to create prediction models utilizing ML technologies. In this study, I made use of datasets collected from the European Union (EU) and the European Economic Area (EEA). There are several reasons why I selected those 30 countries as my samples. The most important reason is that all those 30 countries are experiencing the spread of the four main types of variants. The data was collected from the 40th week of 2020 to the 16th week of 2021. During the 29 weeks, EU/EEA was experiencing a typical period: from the emergence of variants, followed by starting vaccination, then encountering another resurgence caused by variants. However, their anti-epidemic measures were not the same. They were distributing different types of vaccines. Also, some countries avoided re-lockdown while some countries kept lockdowns throughout the entire 29 weeks. Besides, an agency of the EU called European Centre for Disease Prevention and Control is tracking and monitoring the COVID-19 pandemic and building up concrete and accurate datasets.[15] So, my models and analysis would be more convincing if they are based on these official datasets.

In this study, I attempted to achieve the following objectives.

1.Developing ML models to predict the discriminatory factors which drive the transmissibility of variants. I selected several candidates to be examined in this study based on previous studies. All candidates have been demonstrated to be discriminatory factors influencing the previous infection rates of COVID-19. I tried to test to determine if they are still critical in the models of variants.

2.Determining the effects of lockdowns and vaccinations on the infection spread dynamics of variants in the EU/EEA. This will guide further policymaking. If either vaccinations or lockdown measures could be demonstrated to weigh more on controlling the pandemic of variants, the governments could adjust their anti-epidemic measures based on this study.

3.Analyzing the critical pre-lockdown factors which influence the post-lockdown infection rate to guide policymaking. This objective is essential for those countries that do not have enough vaccines and where lockdown is the first option to minimize the new variant cases.

Here I tried to address these issues by using five different ML structures and analyzing their performances. Compared with previous research, this is the first research to evaluate the feasibility of vaccination and lockdown policies for

controlling the variants of COVID-19 and to guide further policymaking.

Dataset

The dataset used in this study contains diverse attributes from 30 countries in the EU/EEA. Most of those attributes were demonstrated to be discriminatory factors which drive the transmissibility of COVID-19 in previous studies. Besides, this dataset is quite unique. Data for different features were integrated from diverse open sources which have been carefully curated. Last but not least, all data has been preprocessed using Pandas, which is a powerful open-source data manipulation tool in Python. This dataset is available on GitHub (https://github.com/mmm-y/covid-19virus.git). I will introduce all features in the following section in detail.

Features Description:

- *Population (2020), Male, Female: Population (2020)* is the population in 2020. And Male and Female stand for the male and female population of each country in 2020 respectively. (source:https://www.statista.com/statistics/611318/population-of-europe-by-country-and-gender/)
- *Male%, Female%, Gender*：*Male% and Female%* are the proportion of male and female among the whole population. *Male%=Male/Population (2020). Female%=Female/Population (2020)*. Also, Gender is the ratio between the population of males and females.
- *Urban%: Urban%* is the fraction of urban population and total population in 2020. (source: https://www.kaggle.com/tanuprabhu/population-by-country-2020)
- *GDP*：*GDP* stands for Gross Domestic Product in 2020 (million US dollars). The source is the World Bank. (source: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
- *GDP_per_capita: GDP_per_capita* stands for Gross Domestic Product per capita in 2020 (US dollars). The source is the World Bank. (source: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
- *Extreme_poverty:* This refers to the fraction of the total population with an income below the international poverty line of 1.90 dollars per day. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
- *Land Area*: This attribute is the land area in each country (square kilometer) (source: https://www.kaggle.com/tanuprabhu/population-by-country-2020)
- *Population Density*: Population Density stands for the density of the population and its unit is People/Km^2
- *Med_age:* This feature represents the median age lifespan of the country (source: https://www.kaggle.com/tanuprabhu/population-by-country-2020)
- *Health_care_index:* This is an estimation of the overall quality of the health care system, including equipment, professionals, doctors, staff and so on. A higher index indicates better quality of the health care system in a particular country. (source: https://www.numbeo.com/health-care/rankings_by_country.jsp?title=2021-mid&region=150)
- *Hospital_beds_per_thousand:* This index represents the number of hospital beds per thousand people in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
- *Diabetes_prevalence*: This is measured by the percentage of the population that have diabetes. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
- *Lockdown_date*: The date of imposition of lockdown at the national level between the 40th week of 2020 to the 16th week of 2021. The lockdown date was obtained from the websites of the official institutions of each country.
- *Lockdown_start_from*: This attribute shows the difference in days from the 40th week of 2020 to the week of imposition of lockdown at the national level.
- *Total_vaccinations*: The data is from Kaggle which represents the number of total immunizations in the country. The owner collected data from *Our World in Data* GitHub repository for covid-19 every day.[20] (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *Johnson&Johnson, Pre_Johnson, Post_Johnson: Johnson&Johnson* represents the total number of immunizations of the JNJ-78436735 by Janssen Pharmaceuticals Companies of Johnson & Johnson. This data is from the same dataset as *Total_vaccinations. Pre_Johnson* and *Post_Johnson* represents the total number of immunizations of the Johnson & Johnson's Janssen COVID-19 vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *Moderna, Pre_Moderna, Post_Moderna:* Moderna represents the total number of immunizations of the mRNA-1273 by ModernaTx, Inc. This data is from the same dataset as *Total_vaccinations*. *Pre_Moderna* and *Post_Moderna* represent the total number of immunizations of the Moderna vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *Oxford/AstraZeneca,Pre_AstraZeneca, Post_AstraZeneca*: *Oxford/AstraZeneca* represents the total number of immunizations of the AZD1222 vaccine by AstraZeneca-Oxford. This data is from the same dataset as *Total_vaccinations*. *Pre_AstraZeneca* and *Post_AstraZeneca* represent the total number of immunizations of the AstraZeneca vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *Pfizer/BioNTech,Pre_Pfizer, Post_Pfizer: Pfizer/BioNTech* represents the total number of immunizations of the BNT162b2 by Pfizer,inc. and BioNTech. This data is from the same dataset as *Total_vaccinations*. *Pre_Pfizer* and *Post_Pfizer* represent the total number of immunizations of

DOI: 10.36838/v4i3.14

the Pfizer vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from on Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)

- *Sputnik V, Pre_Sputnik, Post_Sputnik: Sputnik V* represents the total number of immunizations of the *Sputnik V* vaccines by the Gamaleya Research Institute of Epidemiology and Microbiology in Russia. This data is from the same dataset as *Total_vaccinations*. *Pre_Sputnik* and *Post_Sputnik* represent the total number of immunizations of the Sputnik V vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *Sinopharm/Beijing,Pre_Sinopharm,Post_Sinopharm: Sinopharm/Beijing* represents the total number of immunizations of the BIBP vaccine, which is inactivated virus COVID-19 vaccines by Sinopharm's Beijing Institute of Biological Products. This data is from the same dataset as *Total_vaccinations*. *Pre_Sinopharm* and *Post_Sinopharm* represent the total number of immunizations of the BIBP vaccine before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
- *People_fully_vaccinated_per_hundred:* This index represents the number of people who got fully vaccinated per hundred in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
- *People_vaccinated_per_hundred:* This index represents the number of people who get at least one dose of the COVID-19 vaccine per hundred in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
- *Total_variants, Pre_variants, Post_variants: Total_variants* shows the total number of variants confirmed in the EU/EEA country. This data is from European Centre for Disease Prevention and Control (ECDC). Pre_variants and Post_variants represent the number of variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *B.1.1.7_variants, Pre_B_1_1_7, Post_B_1_1_7: B.1.1.7_variants* indicates the total number of Alpha variants confirmed in the EU/EEA country. This data is from the ECDC. *Pre_B_1_1_7* and *Post_B_1_1_7* represent the number of Alpha variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *B.1.351_variants, Pre_B_1_351, Post_B_1_351: B.1.351_variants* indicates the total number of Beta variants confirmed in the EU/EEA country. This data is from the ECDC. *Pre_B_1_351* and *Post_B_1_351* represent the number of Beta variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *P.1_variants, Pre_P_1, Post_P_1: P.1_variants indicates* the total number of Gamma variants confirmed in the EU/EEA country. This data is from the ECDC. *Pre_P_1* and *Post_P_1* represent the number of Gamma variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *Other_variants, Pre_other_variant, Post_other_variant*: *Other_variants* indicates the total number of variants except from Alpha, Beta and Gamma variants confirmed in the EU/EEA country. This data is from the ECDC. *Pre_other_variant* and *Post_other_variant* represents the number of variants except Alpha, Beta and Gamma variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected manually from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *Peak_infected_week, Peak_infected_cases: Peak_infected_cases* measures the number of variants at the infected peak. *Peak_infected_week* shows the difference in weeks from the 40th week of 2020 to the week the peak was reached. Data is collected from the ECDC and rearranged manually. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)
- *Total_tests, Pre_tests, Post_tests*: *Total_tests* measures the test number during those 29 weeks. *Pre_tests* and P*ost_tests* represents the test number before lockdown and after lockdown respectively. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

Data Description:

A summary statistic description of the above data is shown below (Table 2):

Table 2: Summary of features and their statistics. The features shown in Table 2 will be used to build the ML models.

Feature	Mean	Standard Deviation	Min	Max
Gender	0.98	0.035	0.56	1.51
Urban%	72.73	16.95	15.00	98.00
GDP_per_capita	43220.77	30410.60	18563.31	181402.83
Extreme_poverty	1.39	2.82	0.20	15.00
Population_Density	167.47	253.87	3.00	1380.00
Med_age	42.73	2.55	37.00	47.00
Health_care_index	66.88	9.21	52.82	80.56
Hospital_bed_per_thousand	4.79	1.75	2.22	8.00
Diabetes_prevalence	6.17	1.81	3.28	9.85
Lockdown_start_from	75.50	63.11	0	203.0
Total_vaccine	3.14*10^7	4.10*10^7	5.63*10^4	1.48*10^8
Johnson&Johnson	5850.36	18934.27	0	99683.00

Moderna	1.57*10^6	2.02*10^6	2.68*10^4	7.39*10^6
Oxford/AstraZeneca	4.67*10^6	6.49*10^6	0	2.56*10^7
Pfizer/BioNTech	2.11*10^7	2.98*10^7	2.95*10^4	1.16*10^8
Sputnik V	1.29*10^5	6.57*10^5	0	3.60*10^6
Sinopharm/Beijing	2.21*10^5	1.13*10^6	0	6.20*10^6
People_fully_vaccinated_per_hundred	5.79	2.65	1.37	13.84
People_vaccinated_per_hundred	15.68	6.97	6.13	36.67
Total_variants	9684.86	15122.01	5	70444.00
B.1.1.7_variants	4908.20	8895.64	0	46379.00
B.1.351_variants	176.33	314.46	0	1236.00
P.1_variants	54.36	129.30	0	553.00
Other_variants	4545.97	8058.07	5.00	39730.00
Peak_infected_week	19.83	6.86	1.00	26.00
Peak_infected_cases	1120.17	1539.79	3.00	7358.00
Total_tests	1.14*10^7	1.48*10^7	3.13*10^4	5.60*10^7

Data Pre-processing:

There are null values for the following attributes: *Johnson&Johnson, Moderna, Sputnik V, Oxford/AstraZeneca, Pfizer/BioNTech* and *Sinopharm/Beijing* in Norway and Greece. I replaced the null values with the mean value of the respective columns.

Additionally, when building the ML models, the feature values will be preprocessed by StandardScaler in the Scikit-learn library. The StandardScaler will transform all values so that the distribution of the data will have a mean value 0 and standard deviation of 1. This is required prior to model fitting is because there are diverse variables at different scales, and those variables will cause a bias to the models because of their unequal contributions.

Methods

In this study, I created ML models in the context of the regression algorithm, including support vector regression (SVR), multiple regression (MR), decision tree regression (DTR), random forest regression (RFR), and Bayesian regression (BR). All models were constructed using the Scikit-learn library. The regression problem was to forecast the target label based on existing data. It trained a learner/model based on a training dataset. Then, the model automatically found a corresponding rule to map the input to its output.

Support Vector Regression (SVR)[21]:

Theoretically, the input dataset to train the SVR model is usually non-linear with respect to its labels (the output). Take our dataset **X** as an example, it has several input features (N), like the *Total vaccine*, *Urban*% and so on. It also has several instances (M), like Norway, Greece and so on. So, X is formed with a N x M matrix and an output column $\mathbf{y}_i$, where $\mathbf{y}_i$ represents the total case number of variants. Obviously, it is hard to build a linear relationship between the input matrix (N x M) with respect to $\mathbf{y}_i$.

So, how can SVR models help us to dig the relationship rather than the non-linear relationship? Generally, the successfully trained SVR model means to map the input features into a higher dimensional space by a mapping function. There are several mapping functions with the kernel function being the most used. These include polynomial kernel functions, radial basis kernel functions, and multi-layer perceptron kernel functions. Then in the higher dimensional space, we are expected to find a linear regression function f(x) between the projected input and output, and f(x) could predict the output of an unknown object. To be more specific, in our study, **X** will be projected onto a hyperplane by a kernel function. In this hyperplane, a linear regression function f(x) could be found and used to predict the $\mathbf{y}_i$ of an unknown object.

Multiple Regression (MR)[27]:

A limit of linear regression is that it can only be trained by a dataset with one dependent variable and one independent variable (One-One structure). However, there are several independent variables, like the Total vaccine and Urban%, in our dataset which forms a N-One structure. So, one of the solutions is to use multiple regression.

Understanding the formula of the MR equation is a good way to illustrate the purpose of MR, which is shown below:

$X_0=\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+...+\beta_n X_n+\alpha$, where X_0 denotes the dependent variable, β's denote the coefficients for the different independent variables (X's) and α is an error term suggesting the random sampling noise.

As an example, Total variants is our dependent variable as X_0 in our study. It probably depends on the first independent variable *Urban%* as X_1, and the second independent variable Diabetes prevalence as X_2, and so on. The formula chain will keep going as more and more independent variables are added. Each independent variable has their coefficients to indicate the weight they contribute to the MR equation. So, based on the equation, the objective of the MR equation is to fit a hyperplane to the training dataset based on several independent variables (like the *Urban%*, *Diabetes prevalence* and so on) with different weights. As a result, when a new data point is added, we can predict the output (like the number of infection) based on the independent variables of it.

Decision Tree Regression (DTR)[27]:

The decision tree algorithm is based on a tree-like structure, which consists of a root node, internal nodes/splits, and terminal nodes/leaves. A typical structure of the decision tree is shown in Figure 1. Generally, decision tree structures could represent a disjunction of conjunctions of constraints on the features of instances. For example, in Figure 1, Li represents 'constrain 1(f_h <t_h) ∪constrain 2 (f_i<t_i) ∪ constrain 3(f_k<t_k)', in which this decision tree structure splits a complicated decision into a disjunction of simpler decisions (like constrain 1∪constrain 2 ∪constrain 3) and each of the simpler decisions (like constraint 1) could be in a conjunction form, like an "∩" relationship. When the leaf nodes or our target variables are discrete values, like class attributes, it is called a decision tree classification. And when the leaf nodes are continuous values, like this COVID-19 variants dataset, it is called a decision tree regression.

Basically, how DTR works is to split the space of the original dataset into several sub-spaces by the DTR model. Then, when a new data point is predicted by the DTR, it will assign this new point into the split space to make forecasting. A good example is cited from Georgios Drakos *et al.*[29] Each of the

DOI: 10.36838/v4i3.14

instances or data points in the dataset in Figure 2 has two independent variables (X_1 and X_2) and a target variable which are continuous values, although there are only 2 values to simplify the situation. The blue point represents 1 and the red point represents 0. Based on the original dataset, the DTR model was created in Figure 2.a, and the split sub-space is shown in Figure 2.b. The value of each leaf is the average value of each subspace. So, when a new data point, like (0.38,0.5), is predicted by the model, it will be predicted as 0.3333. For my dataset, each instance has five independent features (more details could be found in 5 results) and the target variable is continuous value, which is the number of confirmed cases of variants in each country in the defined time frame. The trained DTR model will split the whole dataset into several sub-spaces (leaves) and the value of each leaf is the mean value of the corresponding subspace. As a result, when a new data point comes, the target variable could be predicted by this trained DTR.

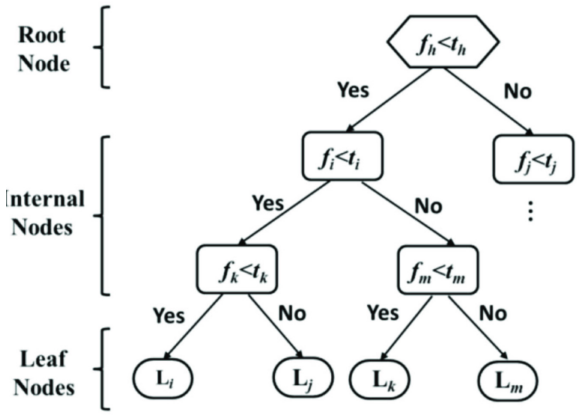

Figure 1: A typical decision tree structure[28].

It was shown that the root node contains the entire sample. Each node could make binary or multiple decision to split the dataset within the node into two or more classes. The whole splitting process will keep going until the algorithm reaches the leaf nodes.

Random Forest Regression (RFR)[30]:

A big problem in the decision tree regression model is overfitting. To address it, I will try random forest regression which uses ensemble learning methods in this study. The simulation of RFR is shown in Figure 3. The RFR will create several independent decision regression models based on subsampling of the entire dataset. The sampling process is conducted at a random and replacement manner. Finally, each decision tree regression will make a prediction and the final prediction is based on the average values of all predictions. In other words, I will select several groups of the training dataset randomly and create more DTR models in the way I discussed in 3.3. So, each model will have an output indicating the predicted number of confirmed cases of variants. The final decision will be made based on the average predictions of all DTR models.

Bayesian Regression (BR)[32]:

To illustrate BR, some background knowledge is crucial.

Multiple regression:

First, let us recall the multiple regression in 3.2. Take our dataset **X** as an example, **X** is in the form of $\{\mathbf{x_i},\mathbf{y_i}\}$, where $\mathbf{x_i}$

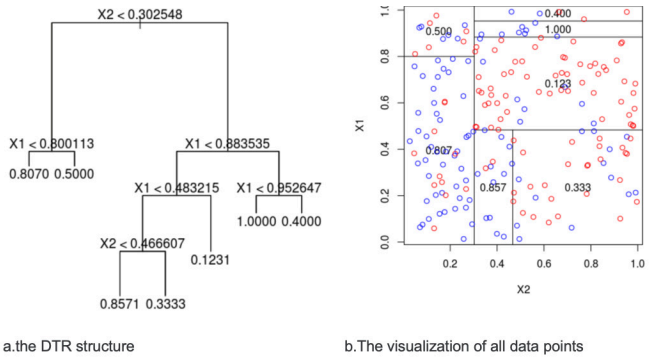

Figure 2: The relationship between DTR structure and its visualization.[29]

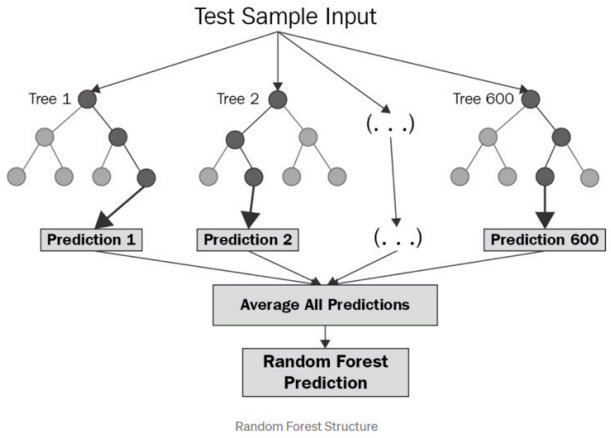

Figure 3: A simulation of the RFR.[31]

represents the input variables (like the *Urban%*, *Diabetes prevalence* and so on), $\mathbf{y_i}$ is the target labels (like the number of infection), and i is the index of the instance in the **X**. Then, a linear regression model could be created based on **X,** and the best model parameters could be set by minimizing the cost function. Finally, it could be used to forecast the predictive value $\mathbf{w_t}$ (like the number of infection), given any input variable $\mathbf{x_t}$.

A mathematical explanation for this linear regression model could be expressed as $Y=X_0=\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\ldots+\beta_n X_n+\alpha$, where X_0 denotes the dependent variable, β's denote the coefficients for the different independent variables (X's) and α is an error term suggesting the random sampling noise. The objective is to find β's to fit a multiple linear function to minimize the sum of the square error between the predictive value and the target value.

Bayes Theorem:

Given the dataset **X** and each created model represents a corresponding hypothesis, there are various hypotheses that could be created based on **X**. Our objective is to find the most probable hypothesis (h) fit to **X**. The Bayes Theorem could be shown as a mathematical formula as well:

$$P(h|X) = \frac{P(X|h)*p(h)}{p(X)}$$

In this equation, $P(h)$ is called the prior probability of h, which represents the initial probability that h holds before we have observed **X**. P(**X**) is the prior probability that **X** will be observed. $P(X|h)$ is the likelihood of the data **X** given h. We are looking for $P(X|h)$, which is the probability of h given **X** and is called the posterior probability.

For my understanding, this Bayes theorem could be organized into:

$$P(model|X) = \frac{P(X|Model)*p(Model)}{p(X)}.$$

Therefore, I am looking for the model with maximized posterior probability given our training dataset **X**.

Bayesian Ridge Regression:

Then, we are looking for the coefficient vector β's which minimizes the loss function.

Consider the probability distributions of y, $y|\mathbf{X},\beta \sim N(\beta^T\mathbf{X}, \sigma^2 I)$, where y is the response feature distribution, N represents that the conditional is Gaussian distributed, $\beta^T\mathbf{X}$ represents the conditional expectation in ML, and $\sigma^2 I$ represents the homoscedastic variance. In this equation, β is the fixed unknown vector to be estimated.

Furthermore, if we consider the posterior distribution of β which is also known as maximum a posteriori (MAP) is $\beta=\arg\max(y-\boldsymbol{X\beta})^T (y-\boldsymbol{X\beta})+\lambda|\ |\beta||^2_2$ where $=\sigma^2/\tau^2$. In this study, I will create a Bayesian linear regression model based on this equation.

Evaluation

The evaluation criteria are significant after creating the models. For regression problems, I used two evaluation indicators: the coefficient of determination (R^2) and Pearson Correlation Coefficient (PCC).

Coefficient of determination (R^2):

R^2 is to determine the percentage variation in the dependent variable that can be explained by other independent variables within a range from 0 to 1.[22] It suggests how many data points fall within the regression model. A formula of R^2 in mathematics is more straightforward.

The formula for R^2 is[22]:

$$R^2=1-\frac{Unexplained\ Variation}{Total\ Variation}$$

To be more specific, R^2 represents if the total number of variants (**yi**) owns a direct relationship to other feature variables (**X**), like vaccines. If R^2 is 0%, then it means 0% of the variation in yi can be explained by **X** or it means none of the data fall within the regression model. Also, if R^2 is 100%, then it means 100% of the variation in yi can be explained by **X** or it means all the data fall within the regression model. An ideal model prefers the R^2 closer to 100%. Additionally, R^2 can be negative. The negative R^2 in our model means this model is a poor fit for **X** or this model cannot set an intercept.[22]

Although it is hard to visualize how data is fitted into the regression model on a space more than 3 dimensions, there is a good example by John McDonal to show the relationship between R^2 and linear regression[23] which can be better understood (Figure 4).

Pearson Correlation Coefficient (PCC):

PCC measures how closely the two variables are related. PCC can be either positive or negative to indicate a positive or negative correlation respectively. The absolute value of PCC suggests the degree of correlation with a range from 0 to 1 inclusive (Figure 5). One is the strongest correlation and zero is the weakest correlation (Table 3).

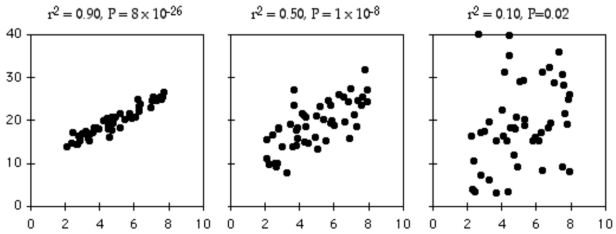

Figure 4: The relationship between R^2 and linear regression.[23]
It was shown that the data will fit the linear regression model better with a higher R^2.

Consider there are two sets of data (x column and y column), and each set has n items. Mathematically, we attempt to measure their correlation by the equation below, which means the "covariance of two variables divided by the product of their respective standard deviations"[25]:

$$PCC=\frac{n(\sum xy)-(\sum x)*(\sum y)}{\sqrt{[n*\sum x^2-(\sum x)^2]*[n*\sum y^2-(\sum y)^2]}}$$

Table 3: The relationship between the degree of correlation with the absolute value of Pearson correlation.[26]

Abs(Pearson correlation)	[0.8,1]	[0.6,0.8)	[0.4,0.6)	[0.2,0.4)	[0.0,0.2)
The degree of correlation	very strong	Strong	moderate	weak	very weak

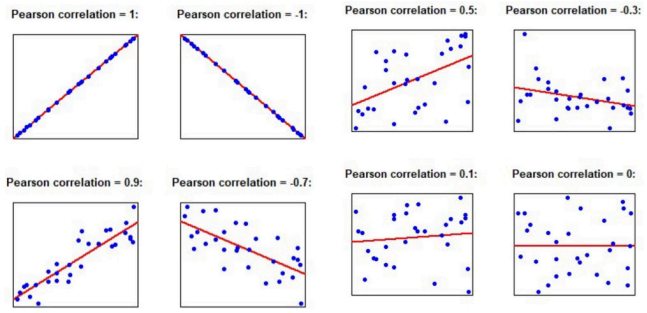

Figure 5: The relationship between Pearson correlation and linear regression.[26]
It was shown that the data will fit the regression model better with the absolute value of Pearson correlation closer to 1. And the positive and negative sign imply the direction of the relationship.

Results

The parameter values for all ML models are shown in Table 4. The set of features to create the ML models are *Gender*, *Urban%*, *GDP_per_capita*, *Extreme_poverty*, *Population_Density*, *Med_age*, *Health_care_index*, *Hospital_bed_per_thousand*, *Diabetes_prevalence*, *Lockdown_start_from*, *Total_vaccine*, *Johnson&Johnson*, *Moderna*, *Oxford/AstraZeneca*, *Pfizer/BioNTech*, *Sputnik V* ,*Sinopharm/Beijing*, *People_fully_vaccinated_per_hundred*, *People_vaccinated_per_hundred*, *Peak_infected_week*, *Peak_infected_cases* and *Total_tests*. The target value is the number of confirmed cases of variants in each country in the defined time frame.

Identification of the key factors:

The objective of this section is to identify the factors that influence the spread of variants. I found all possible combinations (C_{22}^5=26334) of any five features as the feature subset from the whole feature set. To avoid bias, every possible feature subset was used to build each ML model 5 times and the PCC and R^2 will be calculated each time. For example, subset *Gender*, *Urban%*, *GDP_per_capita*, *Extreme_poverty* and

DOI: 10.36838/v4i3.14

Table 4: Values of parameters of all ML models.

ML models	Parameters
SVR	kernel = 'rbf' ; C = 1.0 ;degree = 3;
MR	fit_intercept=True ; normalize=False ;
DTR	criterion = 'mse' ; splitter = 'best' ; max_depth=None;
RFR	n_estimators = 100; criterion = 'mse' ; max_depth=None
BR	n_iter = 300; alpha_=1e-06 ; lambda_=1e-06 ;

Population_Density will be used to create an SVR model 5 times with 5 different values of PCC and R^2. Then the mean values and the standard deviation of those 5 PCCs and R^2s will be calculated. For each model, I selected the feature subset which yielded the best performance. The results are shown in Table 5.

Ranking the key factors:

In this section, I applied the extra tree classifiers to rank the features based on impurity. Different features and their impurity are shown in Figure 7. It shows *Diabetes_prevalence*, *Hospital_bed_per_thousand* and *Lockdown_start_from* are the most important three features. *Med_age*, *Moderna* and *Total_vaccine* follows as their impurities are extremely close to each other. Then, the next echelon are *Peak_infected_cases*, *Pfizer/BioNTech* and *Total_test*. The last group includes *Johnson&Johnson*, *Sputnik V* and *Sinopharm/Beijing*.

Table 5: For each model, the feature subsets yielded the best result.

ML models	Feature Subset	mean(PCC), std(PCC)	mean(R^2), std(R^2)
SVR	*Med_age, Johnson&Johnson, Moderna, Pfizer/BioNTech, Peak_infected_cases*	0.9667, 0.0151	0.8963, 0.0363
MR	*Sputnik V, Total_tests, Hospital_bed_per_thousand, Sinopharm/Beijing, Peak_infected_cases*	0.9830, 0.0180	0.9271, 0.0582
DTR	*Med_age, Sinopharm/Beijing, People_vaccinated_per_hundred, Total_tests, Peak_infected_cases*	0.9632, 0.0223	0.8567, 0.0572
RFR	*GDP_per_capita, Total_vaccine, Johnson&Johnson, Sputnik V, Peak_infected_cases*	0.9454, 0.0717	0.8630, 0.1519
BR	*Diabetes_prevalence, Johnson&Johnson, Lockdown_start_from, People_fully_vaccinated_per_hundred, Peak_infected_cases*	0.9905, 0.0034	0.9687, 0.0189

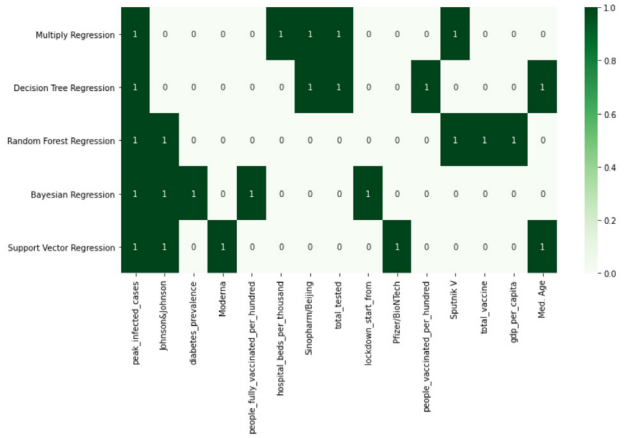

Figure 6: Different combinations of features were used to create different models. Some features were shared by various models.

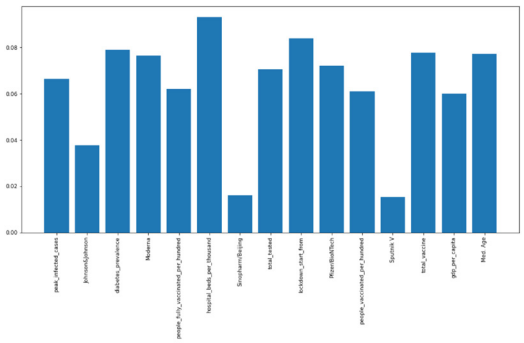

Figure 7: Ranking the key factors affecting the spreading of variants based on impurity.

Identify the key factors:

Firstly, every feature was transformed into percentiles with respect to their respective column. After that, I sorted all countries in decreasing order based on *Post_variants*. The top 5 and the bottom 5 countries are shown as a sample in Figure 8.

Urban population(%)	Health_care_index	...	pre_variants	Total_variant_cases
58.333333	70.000000	...	76.666667	70444.0
83.333333	93.333333	...	96.666667	44631.0
83.333333	60.000000	...	100.000000	24495.0
90.000000	73.333333	...	56.666667	21306.0
66.666667	93.333333	...	63.333333	18792.0
		.....		
11.666667	43.333333	...	30.000000	484.0
50.000000	20.000000	...	50.000000	218.0
33.333333	20.000000	...	43.333333	91.0
3.333333	30.000000	...	25.000000	42.0
93.333333	43.333333	...	20.000000	5.0

Figure 8: A sample of the top 5 and bottom 5 countries based on the *Total_variants* in decreasing order. For each country, their respective features were transformed into percentiles with respect to the column total.

Then, I selected the top 10 and the bottom 10 countries and used the formula from Satyaki Roy *et al*. to calculate the weighted average percentile of their features.[6] The formula is shown below:

$$\frac{1}{\sum_{i=1}^{r}\rho(f_i)}\sum_{i=1}^{k}p(f_i)\times(r-\rho(f_i))$$

where $p(f_i)$ and $p(f_i)$ are the rank and the percentile of the i^{th} feature value, and r is the total number of countries in EEA (r=30 here).

I assumed that for the top and bottom 10 countries, the key factors affecting the spread of variants post-lockdown will show the maximum difference in the weighted average percentiles. I plotted the difference in Figure 9. It suggests that *Diabetes_prevalence*, *Lockdown_start_from*, *Urban%* and *Pre_tests* are the top four key factors.

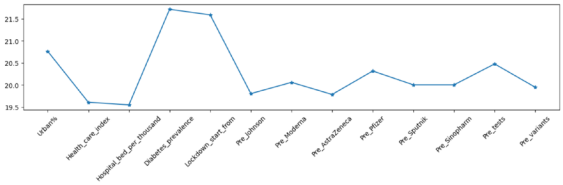

Figure 9: The difference in the weighted average percentile. The first three features are Diabetes–prevalence, Lockdown–start–from, Urban% and Pre–tests.

Creating the MR model based on the selected features:

In this section, I would like to measure the weight of those four features above based on the MR model and I ignored the least important features from the feature pool. Pairwise Pearson correlations were calculated for each pair of features to figure out if there are any correlated features (Figure 10). From the heatmap, no strong correlation has been found, which indicated I do not need to eliminate any features and I could create the MR model based on those four features. The result of the MR model is shown in Table 6. It indicates that *Total_variants* has a negative relationship with Lockdown_start_from and a positive relationship with *Diabetes_prevalence, Urban%* and *Pre_tests*.

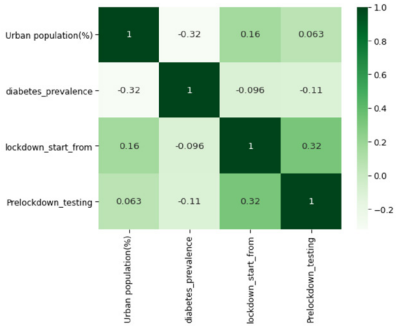

Figure 10: The heat-map for each pairwise Pearson correlation. From the

Table 6: The result of the MR model.

	Diabetes_prevalence	*Lockdown_start_from*	*Urban%*	*Pre_tests*
Coefficient	0.0372	-0.2511	0.0763	2.3248
PCC	0.8820			

Discussion

Nowadays, the variants of SARS-CoV-2 have become one of the biggest problems worldwide. They are threatening our current anti-epidemic strategies. Two major strategies to handle the pandemic are lockdowns and vaccinations. For vaccines of SARS-CoV-2, it was shown that only 15% were fully vaccinated and 29% of the world population had received at least one dose of a COVID-19 vaccine.[17] Unfortunately, the speed of vaccinations is not close to catching up with the spread dynamics of variants of SARS-CoV-2. Also, there are questions concerning the use of old target proteins to produce protective antibodies against those variants. As for the lockdown policy, its objective is to cut off the channels of transmission in the pandemic. However, there are issues with the policy, and the most severe one is the economic depression. Take South Africa as an example, Dorrit Posel *et al.* estimated that there are around 2.5 million adults in South Africa who lost their jobs due to the lockdown policy from February to April 2020.[18] Another issue besides economic loss is the psychological distress of people in lockdown areas. A systematic review and meta-analysis done by Nader Salari *et al.* indicates the prevalence of stress in a sample size of 9,074 is 29.6%, the prevalence of depression in a sample size of 44,531 is 33.7% and the prevalence of anxiety in a sample size of 63,439 is 31.9%.[19] So, how to maximize the positive effects of lockdown when lockdown cannot be avoided is our priority. In this study, I tried to address three questions by utilizing computer science.

The discriminatory factors influence the transmissibility of variants:

There are 15 features that have been selected as the key factors that drive the dynamics of spread of variants. They could be organized into 3 categories.

Vaccine-related features, like *People_fully_vaccinated_per_hundred, Johnson&Johnson, People_vaccinated_per_hundred, Moderna, Sinopharm/Beijing, Pfizer/BioNTech, SputnikV, Total_vaccine.*

It showed that vaccines have a close relationship with the spread of variants. Besides, the total amount of vaccines and the proportion of people who have gotten at least one dose of the vaccine could affect the pandemic. This implies the vaccination strategy could still suppress the pandemic, which could be demonstrated by several clinical trials that current vaccines could still offer adequate protection. However, there is no denying that the efficacy of several kinds of vaccines to the variants has been reduced. So, a modified vaccination strategy should be considered. For example, it was reported that some governments, like the U.S., were trying to strengthen vaccine immunogenicity by providing a third dose. Greatly accelerating the speed of vaccinations to ensure more people could be covered under the vaccines plan is another strategy. This is consistent with my research. Even though vaccine-related features are the key factors influencing the spread of variants, they are not the only determinants and other factors are showing more powerful effects which we will discuss in the following sections in detail. Besides, when I attempted to rank those key features, the top three features do not contain vaccine-related features. Thus, to handle the pandemic of variants of SARS-CoV-2, vaccines are one of our priorities. However, the governments should adjust their vaccine strategy to adapt to the new situation with other key factors as well.

Demographic features, like *Diabetes_prevalence, Hospital_bed_per_thousand, Med_age, GDP_per_capita , Peak_infected_cases* and *Total_tests*.

Among the features, Diabetes_prevalence is the most important. There is evidence that people with diabetes have an increased incidence of COVID-19 and blood glucose control is extremely important for them. In a study by Singh *et al.*, they studied 2,209 COVID patients in China and found 11% of them were suffering from diabetes.[33] And a study processed by Onder *et al.* found that nearly 36% of 355 COVID-19 patients in Italy were had diabetes.[34] Although the prevalence rates are different in different reports, those indicate patients with diabetes tend to be infected by SARS-CoV-2. Combined with previous studies, I concluded that the prevalence of diabetes is also an important feature for creating the ML model of variants. And I hypothesized that people with diabetes will have a higher chance to infect variants. The biological mechanisms behind this is not clear yet. However, it is reported that the increased glucose levels will elevate the replication level of SARS-CoV-2 through 'the production of mitochondrial reactive oxygen species and activation of hypoxia-inducible factor 1α', which could be demonstrated by a T2DM mice model.[35]

DOI: 10.36838/v4i3.14

Other potential mechanisms include immunomodulation, renin-angiotensin-aldosterone system, inflammation and insulin resistance.[35] Based on those discoveries, patients with diabetes may be the population most susceptible to variants. So, the government should pay more attention to them when making plans.

Hospital_bed_per_thousand is the second most important feature, which indicates that the spread of variants has a relationship with the hospital capacity. There is other evidence showing the same relationship between SARS-CoV-2 and hospital capacity by using a mathematical model. Dipo Aldila *et al.* created a "modified susceptible exposed infectious recovered compartmental model" based on the cases from Jakarta, Indonesia.[36] They found that the medical resources, like hospital capacity, are necessary to reduce the burden of COVID.[36] So, hospital capacity may be a crucial feature to slow down the spread of SARS-CoV-2. Admittedly, more investigations should be conducted to find out the possible mechanism. However, I hypothesize that if patients could be admitted into hospitals earlier and if they can stay at a hospital for a longer period, especially during highly contagious periods, the spread of variants of SARS-CoV-2 could be slowed down due to efficient quarantine in the hospital. That suggests one way to reduce the cases of variants is to increase hospital capacity.

Med_age implies that age is as significant a feature for the spread of variants as it is for the SARS-CoV-2. However, a limit in my study is that I only investigated if there is a relationship between the median age and the spread of variants. It is difficult to figure out which age group is more susceptible to the variants. Based on recent CDC reports, it was said that the younger people in the U.S. are taking the place of the older people as the biggest group of newly hospitalized. So, my study could provide a direction for further study of the variants that aims to figure out the age of the susceptible population.

As for the GDP per capita, my study could only demonstrate that the GDP per capita are important features to construct the ML models of variants and indicates that this feature is associated with the spread of the variant cases. This is consistent with a previous study carried out by Shahina Pardhan *et al*.[38] They demonstrated a negative relationship between the change in COVID-19 cases with the GDP per capita in 38 European countries during the first wave of the pandemic and concluded that the economic performance of a country should be an important consideration for policymakers.[36] However, it is hard to conclude if they have a causal relationship. And if there is a causal relationship, which one is the cause and which one is the result. Because the worsening economic situation may lower the quality of the healthcare system, or the uncontrolled pandemic can hurt the economy significantly.

The Peak_infected_cases indicates that it is possible to predict the total cases of variants by using the number of cases at peak point. In other words, *the Peak_infected_cases* could be an indicator for the government to evaluate the efficiency of their current strategy in the middle of a pandemic of variants. *The lower Peak_infected_cases* is associated with a positive output. Thus, the governments could adjust their strategies as soon as possible.

The last feature is the *Total_tests*. Our study shows the total tests have a relationship with the spread of variants. It is reasonable to deduce their relationship by reviewing the relationship between COVID-19 Cases and testing. Surprisingly, the total number of COVID-19 cases can be reduced with the help of testing. A study conducted by Umit Cirakli *et al*. demonstrated that 1% increase in the total tests contributed to reducing new cases by 1.45% as analyzed by four models.[39] Before the study, I thought that more new cases of variants will be detected as the outlay of tests, so the total cases of variants will increase. However, based on my study and previous studies, it seems like the spread of variants will be controlled by large-scale testing. The reason is because large-scale testing could detect and isolate risk groups to cut off the channels of transmission. So, governments need to put more effort into large-scale testing.

Lockdown-related features, like *Lockdown_start_from*. The number of days from our starting point to the day when national lockdowns were carried out is the third important feature in our study. More details will be discussed in section 6.2.

Unimportant features, like *Gender, Extreme_poverty, Population_Density, Health_care_index*. Those features are deemed to be important for the transmission of COVID-19. Interestingly, in my study, those features are not the crucial factors for the spread of variants. In the later section, I illustrate that the proportion of urban people is important to predicting the post-lockdown infection rate. Combined with the *Population_Density* here, it seems like it is the density of local areas rather than that of the national level affecting the spread of variants.

Effect of lockdown and vaccination on the infection spread dynamics of variantss:

This study shows that both lockdowns and vaccinations are significant strategies to handle the pandemic of variants. Besides, after ranking, I found that lockdown plays a more vital role than vaccinations. That implies the governments should not expect an ideal result by solely depending on vaccinations without any lockdown policies. In addition, my study found that the lockdown length is different in different countries. That's because countries set the start and end date according to their subjective perspective. Also, no objective criterion existed to guide when they should extend their lockdown date or when they could relieve the lockdown. Furthermore, each country implemented different measures during the lockdowns. For some countries, a national lockdown may simply imply a vacation for schools or working from home. However, for other countries, they may carry out a stricter policy, like quarantining regions.

An important part of the study is to analyze the critical factors which influence the post-lockdown. I constructed a MR model based on four features, *Diabetes_prevalence, Lockdown_start_from, Urban% and Pre_tests,* and discovered their coefficients. It shows that the coefficient of *Lockdown_start_from* is -0.251, which demonstrates that the lockdown length has a negative relationship with the spread of variants. Thus, to facilitate a better outcome, governments should lengthen the lockdown length, in other words they should lock down earlier and keep the lockdown period longer. Given that I could deduce the relationship between the lockdown and the

spread of variants based on previous studies of COVID-19 by assuming they are sharing the same pattern, my conclusion could be demonstrated by other studies as well. Based on a study by Atalan *et al.*, they found that the lockdown length during a particular time period without any interruption has a very strong negative correlation with the number of new confirmed cases, with the correlation value being -0.9126.[40] That implies that if the country starts to lock down earlier and keeps it longer, the efficacy of lockdown policy will be maximized.

However, my study only focused on lockdowns nationally. Further studies should concentrate on if lockdowns locally work and what is the criterion for the countries to implement lockdown policies. Because, obviously, lockdowns at a regional scale rather than a national scale would be less harmful. Furthermore, more details about the lockdown could be discussed in a further study, like what kind of measures should be included in the lockdown policy.

Thankfully, vaccinations contribute to the development of the variants' models as well. Even if they are not the top key factor, eight features out of the fifteen features are vaccine-related features, which suggests that the total effect of all vaccine-related features plays a vital role in the spreading of variants and the vaccines are still useful for the variants. That could guide the creation of policy.

Firstly, a single type of vaccine is far from being able to control the spread of variants and a combination of all types of vaccines is needed. Because each vaccine has different targets, and the mixed use of different vaccines offers better protection for a group of people at risk of being infected by different types of variants. The mixture strategy could be accomplished in two ways. The first one is combining different vaccines in one person. For example, the two doses of Sputnik V vaccine use Ad26 and Ad5 respectively, and its high efficacy has been demonstrated.[43] Besides, John *et al.* put forward that "using a mRNA vaccine or protein vaccine to boost the first dose of the Johnson & Johnson or AstraZeneca adenovirus vectors could possibly be more effective than giving a second dose of the same dose of the same adenovirus".[41] Another possible way is to individualize vaccination so that different groups could receive different kinds of vaccines to maximize the performance of vaccines and lower the resistant viruses. For instance, John *et al.* put forward that it could decrease the resistant viruses by only allowing the younger people to get Johnson & Johnson vaccines.[41] However, this scenario needs to be examined in further studies.

Secondly, in my study, it seems like two features, *People_fully_vaccinated_per_hundred* and *People_vaccinated_per_hundred* have equal significance in the spread of variants. That implies some groups of people may not need the second dose or even the third dose to protect them from infection of variants. An investigation by Stamatatos *et al.* could probably explain that.[42] Their study showed that a single mRNA vaccine dose could boost the antibody level rapidly and significantly in people who have recovered from COVID-19.[42] Of course, more designed trials should be carried out for this question. However, it awakens us to the possibility of giving different groups of people different doses of vaccines, which will significantly relieve the stress from the shortage of vaccines.

Analyzing the critical pre-lockdown factors which influence the post-lockdown infected rate :

To identify pre-lockdown factors, I constructed a MR model with a Pearson correlation of 0.8820. The model shows that the prevalence of diabetes, the proportion of population in urban areas, and the total testing before the lockdown are three important key factors that positively influence the post-lockdown infection rate. And the lockdown length has a negative relationship with our target label. Thus, for governments who decide to lockdown their countries, they should increase the testing, reduce the density of urban areas, pay more attention to people with diabetes, and, if possible, lockdown the countries as early as possible and avoid locking down too late.

The limitations of the study:

There are some limitations to this study. Firstly, we can enlarge the EEA or include more countries into our dataset, like countries in Africa. The reason why I did not include those in this study is partly because of the incomplete information. The capacity of testing in some countries lags behind the demands of our dataset. In addition, I only identified the key factors affecting the spreading of variants. In further studies, I could analyze each factor in depth. For example, further study could concentrate on weighing the merits and downsides of lockdowns locally or nationally, what is the criterion for the countries to start lockdown policies and what kind of measures could be included in the lockdown policy to maximize its efficacy. Additionally, how to modify the current vaccination strategy to variants is another problem. Currently, some countries, like the U.S., are expected to authorize the third vaccine dose to suppress the spread of variants. However, those policies are not supported by enough scientific evidence.

Conclusions

Machine learning is a significant way for governments to predict the spread of variants of COVID-19. By constructing ML models, we could identify and rank the key factors affecting the transmission of variants. Besides, we could figure out critical factors to improve the lockdown effectiveness. While existing studies focus on constructing the ML model of COVID-19, my work presents a possible way to examine that of variants and help guide policymaking. My study illustrates that the spread of variants are affected by multiple factors. Among them, both lockdowns and vaccinations are important to stopping the pandemic. That implies governments should pay attention to every possible factor and not ignore the importance of lockdowns when they are designing the anti-epidemic plan. In addition, special attention should be paid to the people with diabetes and people who live in the urban areas in the blueprint. Individualized vaccination plans are needed, including how many doses of vaccines and what kind of vaccines a person should get. Though this needs to be investigated further. To maximize the performance of lockdowns, the governments, before locking down, should design a strategy based on increasing testing, reducing the density of urban areas, paying more attention to people with diabetes, and, if

DOI: 10.36838/v4i3.14

some possible solutions or future areas of study. More research should be completed in future studies.

Acknowledgements

In writing this research, I have been fortunate to be assisted by my tutor, Dr. Wei Yang, who is the computer teacher in my high school. This research could not have been written without her help. I am deeply indebted to her work to review my drafts and guide my research. As always, I would like to thank my family. For me, I cannot imagine a more supportive family than the one I have.

References

1. Iboi, E. A.; Ngonghala, C. N.; Gumel, A. B. Will an Imperfect Vaccine Curtail the COVID-19 Pandemic in the U.S.? *Infect. Dis. Model*. 2020, 5, 510–524.
2. COVID-19 map-johns Hopkins Coronavirus resource Center. https://coronavirus.jhu.edu/map.html (accessed Aug 2, 2021).
3. Impact of COVID-19 on people's livelihoods, their health and our food systems.https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems (accessed Aug 2, 2021).
4. Sweden https://covid19.trackvaccines.org/country/sweden/ (accessed Aug 2, 2021).
5. Björk J. IM, Moghaddassi M., et al. Effectiveness of the BNT162b2 vaccine in preventing COVID-19 in the working age population – first results from a cohort study in Southern Sweden. medRxiv. 2021;https://www.medrxiv.org/content/10.1101/2021.04.20.21254636v1
6. Roy, S.; Ghosh, P. Factors Affecting COVID-19 Infected and Death Rates Inform Lockdown-Related Policymaking. *PLoS One* 2020, 15 (10), e0241165.
7. Peckham, H.; de Gruijter, N. M.; Raine, C.; Radziszewska, A.; Ciurtin, C.; Wedderburn, L. R.; Rosser, E. C.; Webb, K.; Deakin, C. T. Male Sex Identified by Global COVID-19 Meta-Analysis as a Risk Factor for Death and ITU Admission. *Nat. Commun*. 2020, 11 (1), 6317.
8. Lim, S.; Bae, J. H.; Kwon, H.-S.; Nauck, M. A. COVID-19 and Diabetes Mellitus: From Pathophysiology to Clinical Management. *Nat. Rev. Endocrinol*. 2021, 17 (1), 11–30.
9. Covid in the U.s.: Latest Map and Case Count. The New York times. The New York Times March 3, 2020.
10. CDC. About Variants of the Virus that Causes COVID-19 https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html (accessed Aug 3, 2021).
11. Iacobucci, G. Covid-19: Single Vaccine Dose Is 33% Effective against Variants from India, Data Show. *BMJ* 2021, 373, n1346.
12. Vaccine inequity undermining global economic recovery https://www.who.int/news/item/22-07-2021-vaccine-inequity-undermining-global-economic-recovery (accessed Aug 3, 2021).
13. CDC. SARS-CoV-2 Variant Classifications and Definitions https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (accessed Aug 3, 2021).
14. Zoabi, Y.; Shomron, N. COVID-19 Diagnosis Prediction by Symptoms of Tested Individuals: A Machine Learning Approach. *bioRxiv*, 2020. https://doi.org/10.1101/2020.05.07.20093948.
15. Rahaman Khan, M. H.; Hossain, A. Countries Are Clustered but Number of Tests Is Not Vital to Predict Global COVID-19 Confirmed Cases: A Machine Learning Approach. *bioRxiv*, 2020. https://doi.org/10.1101/2020.04.24.20078238.
16. Homepage https://www.ecdc.europa.eu/en (accessed Aug 4, 2021).
17. Ritchie, H.; Ortiz-Ospina, E.; Beltekian, D.; Mathieu, E.; Hasell, J.; Macdonald, B.; Giattino, C.; Appel, C.; Rodés-Guirao, L.; Roser, M. Coronavirus Pandemic (COVID-19). *Our World in Data* 2020.
18. Posel, D.; Oyenubi, A.; Kollamparambil, U. Job Loss and Mental Health during the COVID-19 Lockdown: Evidence from South Africa. PLoS One 2021, 16 (3), e0249352.
19. Salari, N.; Hosseinian-Far, A.; Jalali, R.; Vaisi-Raygani, A.; Rasoulpoor, S.; Mohammadi, M.; Rasoulpoor, S.; Khaledi-Paveh, B. Prevalence of Stress, Anxiety, Depression among the General Population during the COVID-19 Pandemic: A Systematic Review and Meta-Analysis. *Global. Health* 2020, 16 (1), 57.
20. *Covid-19-Data: Data on COVID-19 (Coronavirus) Cases, Deaths, Hospitalizations, Tests • All Countries • Updated Daily by Our World in Data*.
21. Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput*. 2004, 14 (3), 199–222.
22. Huang, J.-C.; Ko, K.-M.; Shu, M.-H.; Hsu, B.-M. Application and Comparison of Several Machine Learning Algorithms and Their Integration Models in Regression Problems. *Neural Comput. Appl*. 2020, 32 (10), 5461–5469.
23. McDonald, J. H. *Handbook of Biological Statistics*; Lulu.com: Morrisville, NC, 2012.
24. *Example of Statistical Techniques Applied to Analysis of Measurements of the Landing Airborne Manoeuvre. (Multiple Linear Regression with Two Independent Variables and One Dependent Variable.)* 92022; ESDU International PLC: London, England, 1992.
25. Emerson, R. W. Causation and Pearson's Correlation Coefficient. *J. Vis. Impair. Blind.* 2015, 109 (3), 242–244.
26. Szczepanek, A. Pearson correlation calculator https://www.omnicalculator.com/statistics/pearson-correlation (accessed Aug 6, 2021).
27. Darlington, R. B. Multiple Regression in Psychological Research and Practice. *Psychol. Bull*. 1968, 69 (3), 161–182.
28. Wang, D.; Rueda Torres, J. L.; Rakhshani, E.; van der Meijden, M. MVMO-Based Identification of Key Input Variables and Design of Decision Trees for Transient Stability Assessment in Power Systems with High Penetration Levels of Wind Power. Front. *Energy Res*. 2020, 8. https://doi.org/10.3389/fenrg.2020.00041.
29. Drakos, G. Decision Tree Regressor explained in depth https://gdcoder.com/decision-tree-regressor-explained-in-depth/ (accessed Aug 7, 2021).
30. Lindner, C.; Bromiley, P. A.; Ionita, M. C.; Cootes, T. F. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. *IEEE Trans. Pattern Anal. Mach. Intell*. 2015, 37 (9), 1862–1874.
31. Chakure, A. Random Forest regression - the Startup - medium https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f (accessed Aug 8, 2021).
32. Saqib, M. Forecasting COVID-19 Outbreak Progression Using Hybrid Polynomial-Bayesian Ridge Regression Model. *Appl. Intell*. 2021, 51 (5), 2703–2713.
33. Yang, J.; Zheng, Y.; Gou, X.; Pu, K.; Chen, Z.; Guo, Q.; Ji, R.; Wang, H.; Wang, Y.; Zhou, Y. Prevalence of Comorbidities and Its Effects in Patients Infected with SARS-CoV-2: A Systematic Review and Meta-Analysis. *Int. J. Infect. Dis*. 2020, 94, 91–95.
34. Onder, G.; Rezza, G.; Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA 2020*. https://doi.org/10.1001/jama.2020.4683.
35. Lim, S.; Bae, J. H.; Kwon, H.-S.; Nauck, M. A. COVID-19 and Diabetes Mellitus: From Pathophysiology to Clinical Management. Nat. *Rev. Endocrinol*. 2021, 17 (1), 11–30.
36. Aldila, D.; Khoshnaw, S. H. A.; Safitri, E. et al. A Mathematical Study on the Spread of COVID-19 Considering Social Distancing and Rapid Assessment: The Case of Jakarta, Indonesia. *Chaos Solitons Fractals* 2020, 139 (110042), 110042.

37.Brown, C. M.; Vostok, J.; Johnson, H.; Burns, M. et al. Outbreak of SARS-CoV-2 Infections, Including COVID-19 Vaccine Breakthrough Infections, Associated with Large Public Gatherings - Barnstable County, Massachusetts, July 2021. *MMWR Morb. Mortal. Wkly.* Rep. 2021, 70 (31), 1059–1062.
38.Pardhan, S.; Drydakis, N. Associating the Change in New COVID-19 Cases to GDP per Capita in 38 European Countries in the First Wave of the Pandemic. *Front. Public Health* 2020, 8, 582140.
39.Cirakli, U.; Dogan, I.; Gozlu, M. The Relationship between COVID-19 Cases and COVID-19 Testing: A Panel Data Analysis on OECD Countries. *J. Knowl. Econ*. 2021. https://doi.org/10.1007/s13132-021-00792-z.
40.Atalan, A. Erratum to "Is the Lockdown Important to Prevent the COVID-19 Pandemic? Effects on Psychology, Environment and Economy-Perspective" [Ann. Med. Surg. 56 (2020) 38-42]. *Ann. Med. Surg*. (Lond.) 2020, 56, 217.
41.Moore, J. P. Approaches for Optimal Use of Different COVID-19 Vaccines: Issues of Viral Variants and Vaccine Efficacy: Issues of Viral Variants and Vaccine Efficacy. *JAMA* 2021, 325 (13), 1251–1252.
42.Stamatatos, L.; Czartoski, J.; Wan, Y.-H.;et al.A Single MRNA Immunization Boosts Cross-Variant Neutralizing Antibodies Elicited by SARS-CoV-2 Infection. *medRxiv* 2021. https://doi.org/10.1101/2021.02.05.21251182.
43.Logunov, D. Y.; Dolzhikova, I. V.; Shcheblyakov, D. V. et al. Gam-COVID-Vac Vaccine Trial Group. Safety and Efficacy of an RAd26 and RAd5 Vector-Based Heterologous Prime-Boost COVID-19 Vaccine: An Interim Analysis of a Randomised Controlled Phase 3 Trial in Russia. Lancet 2021, 397 (10275), 671–681.

Author

Peter Ma, Grade 11 student studying at Beijing No. High School International Campus, which is one of the top public international schools in China. Before this, He studied at EagleBrook School in USA, where he got Academic Award issued by the Headmaster and was the leader of the school's Math team.

DOI: 10.36838/v4i3.14