World COVID-19 Death Project

Chaytanya Kumar, Dennis Poludnev, Youjia (Yuka) Chen

American University

STAT614- Statistical Method

Dr. Mary Gray

December 13th, 2021

Introduction

Covid-19 has shaken the entire world and its impact is unprecedented in epidemic calamity history of the last 50 years which will be felt for next 50 years. With over 240 million cases and 4+ million deaths, the covid-29 still grabs news headlines . With the recent development of Covid vaccines and its rapid distribution around the world to eliminate covid-19 has shown substantial progress in the vaccination rate. However , the goal to fully-vaccinate every 5 billion on earth and eradicate Covid-19 is far from achieved. It will be interesting to study if getting fully-vaccinated has really helped in reduction of the covid cases in a particular country. While the vaccination is pivotal ,many individuals sceptical about the efficacy and after-effects of the vaccination.Therefore many are skipping or procrastinating on the idea of getting vaccinated.

It will be interesting to study if the percentage increase in vaccination has really brought down the number of positive cases in Covid.Also what factors like population density, GDP, stringency index, poverty and people dealing with different complexities of related health issues (like Smokers,diabetes etc) have what effect on the percentage of total death rate.

Supposedly, the vaccination should help reduce the death rate of COVID-19. However, we want to first find out what is the strongest independent variable besides vaccine numbers that influences our response variable of the covid death rate. Later to examine if vaccination actually helps reduce the COVID-19 cases in both developed countries and developing countries. This study is to show people that despite the vaccination might help us to prevent COVID-19 spreading, there are still other factors we should focus on. We predict among all explanatory variables of Smokers, Poverty, Icu Patients,Gdp Per Capita, Diabetes Prevalence, Life Expectancy, Population Density, and Vaccination rate, vaccination number would be the strongest independent variable.

Literature Review

There have been recent studies and data analysis that support the hypothesis of vaccination greatly reducing the COVID related deaths. For example, a recent scholarly article published in October 22, 2021 by *ContentEngine LLC, a Florida limited liability* publishing company, reports the association between vaccination and COVID related deaths. The published article explains how the researchers in the UK analyzed data from 5.4 million people in Scotland from1 April to 27 September and discovered that there is a strong correlation between vaccination and death related to SARS-CoV-2. The data comes from analyzing 114,706 adults who tested positive for SARS-CoV-2 which shows a sharp decline in death incidents once the majority of the population received vaccination for COVID. In fact, the study shows that there was a steep decline in Delta variant deaths among the population that received a second dose of vaccination. The analysis shows that chances of surviving Delta variant after receiving vaccination are as high as 90-98%. With that in mind we have more reasons to believe that after analyzing the World Covid data from github we can find similar correlations between the variables covid-deaths and vaccinations, in fact we are certain that if we conduct bivariate or ANOVA tests between these variables the result will be significant enough to prove that there is a correlation between these two variables.

Another hypothesis is that covid related deaths are higher in developing countries, therefore by using the data we acquired from github we are trying to find a negative correlation between growth development index and covid deaths. In search of literature review on similar subject we stumbled upon an article from International Journal of Social Economics titled, *Determinants of the number of deaths from COVID-19: differences between low-income and high-income countries in the initial stages of the pandemic,* by Valero, M., & Valero-Gil, J. The article explains interesting findings regarding COVID spread in developing countries vs developed countries. One of the findings proves contrary to our beliefs and explains that some countries which have low economic development have less covid related death. However, the explanation brings in secondary variables which play a role in COVID spread, such as transportation and population density. For example, some underdeveloped countries that do not have good transportation infrastructure and less clustered communities have less COVID related deaths due to low spread.

However, other factors such as the healthcare system in developed countries has a higher chance of preventing deaths. In our earlier analysis based on GDP and covid deaths shows sufficient data that proves the hypothesis of lower deaths in developed countries vs developing based on the population proportion. On the other hand, the outlier countries could be explained by outside variables that affect the covid spread, as was mentioned in the article the transportation infrastructure and the clustering of population has an impact on COVID spread and therefore covid related deaths. We will take that under consideration as we run some tests with the Github data and implement bivariate tests between GDP and Vaccination, chances are these variables have an effect on the response variable: covid deaths.

Government policies also seem to have a strong effect on COVID-19 infection rate. Hansen et. al (2021) found the people in Denmark (developed country) mobility were significant factors that also interacted with government policies' restrictions. When restrictions were strong, an increased residential mobility resulted in decreased COVID-19 incidence, suggesting residential mobility as a proxy for compliance. With our data, we can believe that even when a country has low GDP and high property, as long as they have strict policies, there is a possibility of them getting a lower rate of covid death.

Experimental design

From the studies we've read, it seems that there are some contradictions. It seems possible when a country's GDP is higher, the more COVID19 cases they would have due to the population density, which eliminates our team's assumption that lower GDP countries have more COVID cases because of poverty. Moreover, one of the articles mentions that when the government has strict policies, the lower COVID19 cases they would get. It was believed that vaccination would reduce the spread of COVID and lower the COVID death rate. Therefore we want to see if GDP, poverty level, policy restriction level, as well as the vaccination as any association with GDP.

Data Source:

For our project, we are using the data from github named *covid-19-data*. It was a corporately work collected by COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University; European Centre for Disease Prevention and Control; and government sources

for the United Kingdom, the United States, Canada, Israel, Algeria, Switzerland, Serbia, Malaysia; as well as *Our World in Data* team.

Link to the data: https://github.com/owid/covid-19-data

For our project, we define our dependent variable (response variable) : total death number due to COVID-19 and 6 Independent variables (explanatory variable), including Vaccination, COVID test number, GDP Per Capita, Population density, Extreme poverty, Stringency.

Dependent variables	Independent variables	
Total Death Number of Covid-19 [total_deaths_per_million] Total deaths attributed to COVID-19 per 1,000,000 people ***new_deaths_per_million***	***Vaccination***	Total number of COVID-19 vaccination doses administered
	COVID testing number	Total tests for COVID-19 per 1,000 people
	GDP Per Capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
	Population density	Number of people divided by land area, measured in square kilometers, most recent year available
	Extreme poverty	Share of the population living in extreme poverty, most recent year available since 2010
	Stringency	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
	Female/Male Smoker	Smoker rate

Content Analysis

Overall Data Since Nov 7.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
GDP	661.2	4449.9	12951.8	19238.3	27216.4	116935.6
Deaths	1	79	725	49868	6406	5038372

Vaccinations	0.000e+00	3.493e+05	2.635e+06	1.044e+08	1.594e+07	7.227e+09
Pop., Density	0.137	36.253	83.479	399.810	209.588	20547.766
Extreme Poverty	0.1	0.6	2.2	13.5	21.2	77.6

Based on the both outputs of significance level, we can see that p-value is less than 0.05, therefore we reject the null hypothesis that there's no difference between the means and conclude that a significant difference does exist.

We then decided to further run the correlation tests between the variables gdp and vaccination and place totla_death variable as a factor.

H0: The variables are independent.
Ha: There is a correlation between the variables.

Based on the outcome for both total_deaths and total_vaccination the population correlation coefficient is not equal to 0, but rather equal to 1. This means we reject the null hypothesis, and have significant results to favor the alternative hypothesis, which means that variables are not independent.

Data Analysis using R Studio

We decided to use a random sample of 30 countries and run a linear regression model analysis in R studio by placing total_death_per_million as a response variable with multiple linear regression variables, then run forward and back analysis and use the stepwise regression method to narrow down results for the causation of covid related deaths.

Based on the random selection in R, we decided to focus on the thirty countries: Paraguay, Kazakhstan,Austria,Croatia,Ecuador,Bulgaria,Ireland,Norway,United States,Italy,Uruguay,Israel,Ukraine,Canada,Lithuania,Denmark,Portugal,Chile,Colombia,Argentina,Latvia,Panama,Mexico,Malta,Slovakia,Malaysia,Russia,Estonia,United Kingdom, and Turkey.

Once we filtered the thirty countries, we filtered the data and ran a multi linear regression model on all the variables we selected. We also combined female smokers and male smoker rate to create a new variable called smoker_total. From the initial result, we only had two variables that had a p-value lower than 0.05 and was statistically significant - which are smoker total and extreme poverty. Moreover, we only used the records from December 13th, 2020 to November 1st, 2021 to ensure our results of analysis consistency are consistent. To compare GDP per

capita, total smokers and extreme poverty rate, we decided to focus till April 24th, 2021, because of repeated same independent variables that can cause graphs hard to read, since we are looking for the total death per millions.

Multiple linear regression :

Dependent variable: Total Deaths per million

```
Call:
lm(formula = total_deaths_per_million ~ smoker_total + extreme_poverty +
    gdp_per_capita + total_vaccinations + stringency_index, data = mydata3)

Residuals:
    Min      1Q  Median      3Q     Max
-681.01 -317.92  -45.95  254.78  956.83

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.381e+02  7.504e+02   0.184  0.85553
smoker_total        1.528e+01  6.337e+00   2.410  0.02395 *
extreme_poverty     3.436e+02  1.171e+02   2.936  0.00723 **
gdp_per_capita      2.613e-04  7.847e-03   0.033  0.97371
total_vaccinations  2.931e-06  2.334e-06   1.256  0.22121
stringency_index   -1.977e+00  8.154e+00  -0.242  0.81051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 509.5 on 24 degrees of freedom
Multiple R-squared:  0.3764,    Adjusted R-squared:  0.2465
F-statistic: 2.898 on 5 and 24 DF,  p-value: 0.0348
```

Multicollinearity test :

After we ran the multiple linear regression model, we did a multicollinearity test to find if any of the independent variables are correlated.

```
> vif(peq1)
      smoker_total    extreme_poverty     gdp_per_capita
          1.540656           1.793067           1.533799
total_vaccinations   stringency_index
          1.217805           1.186933
>
```

Since all independent variables’s VIF values are less than 5 therefore, there is no multicollinearity in the model.

Stepwise Regression:

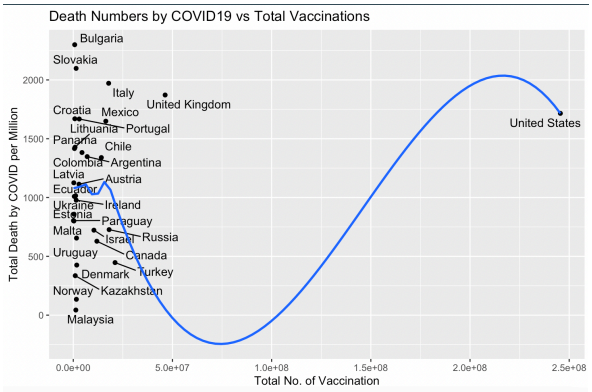
Stepwise regression to find the best fitted regression model.

```
Step:  AIC=375.39
total_deaths_per_million ~ extreme_poverty + smoker_total + total_vaccinations

                     Df Sum of Sq     RSS    AIC
<none>                             6246370 375.39
+ population_density  1   25024.6 6221346 377.27
+ stringency_index    1   15066.3 6231304 377.32
+ gdp_per_capita      1      95.5 6246275 377.39
> x3

Call:
lm(formula = total_deaths_per_million ~ extreme_poverty + smoker_total +
    total_vaccinations, data = mydata3)

Coefficients:
       (Intercept)     extreme_poverty        smoker_total  total_vaccinations
         -5.208e+00           3.432e+02           1.570e+01           3.038e-06
```

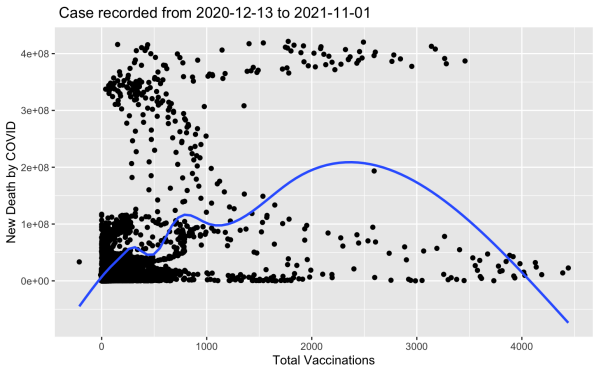
Based on the R result, the best fitted regression model is with the main dependent variable of total_deaths_per_million and three independent variables: extreme_poverty, smoker_total, and total_vaccinatinos.

Total Deaths per million = 343.2(Extreme Poverty) + 15.7(Total Smokers) + 0.000003038(Vaccination Rate)

Data visualization:



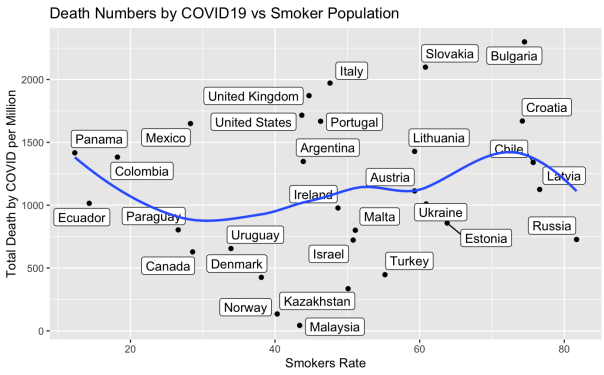
[Image1: Death based on vaccination till April 24th, 2021]



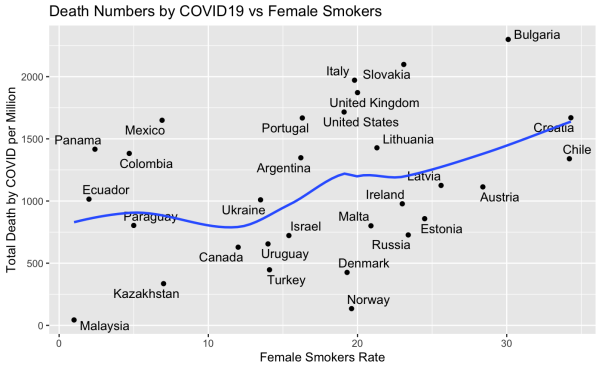
[Image 2: Death based on vaccination with date range of 1 year]

The possible reason why vaccination has such a small prediction value is because not all thirty selected countries have access to vaccination nor recorded and updated the vaccination records. Moreover, the dataset was recorded before the vaccination for COVID19 was available for commercial use. The timeline gap could also lead to increasing numbers of deaths aligned with a very small rate of fully vaccinated.

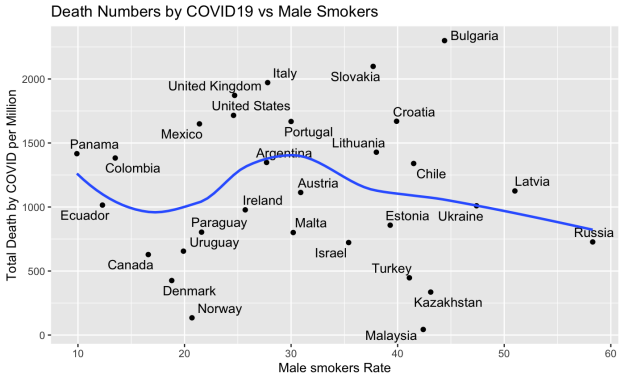
The regression statistical analysis shows that smoker_total is correlated with total_death_per_million, however, when we look at the smokers by genders, we have very different patterns.



[Image3 Total Smokers Rate versus Total Death per Million]



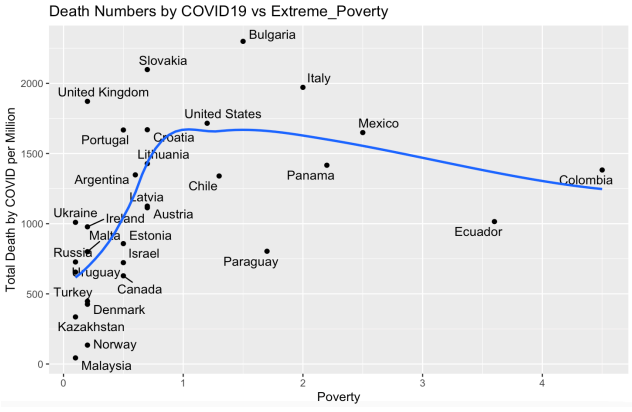
[Image 4 Female Smokers Rate versus Total Death per Million]



[Image 5 Make Smokers Rate versus Total Death per Million]

From the graph, the higher female smokers rate, the more deaths by COVID 19 per million. However, in the case for male smokers, it shows the opposite pattern: the higher the smoking rate, the less total death by COVID19. One of the possible explanations is that there are limited numbers of female smoker records. With a small sample size, there could be some major errors in the dataset.

Our initial prediction was that GDP per capita is strongly correlated with total death by COVID19. However, R does not recognize it as the best fitted model, instead we have extreme_poverty that is correlated with total death by COVID. Our explanation is that even though some countries have higher GDP per capita, they have extreme wealth or income gaps. For instance, the United States (Ewing, 2020). Therefore, it is possible that some countries are wealthy per individual, however, have high extreme poverty rates. The extreme poverty from this dataset is based on shares of the population that have the daily consumption less than $ 1.90. Therefore in a sense our initial hypothesis including GDP although was not significant, the extreme poverty variable was found statistically significant after the stepwise analysis.



[Image 4 poverty versus total death by COVID19 per million]

Limitations

Bias: 8 types of possible bias in the dataset.

1. Propagating the current data.
   The data for **confirmed death** variables is acquired from multiple sources such as COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. This could create few limitations on how JHU has the valid death data from multiple countries, the end source is not exactly clear and could be a subject to question.
   Similar to the variable data for **ICU admission**, the data is gathered from European Centre for Disease Prevention and Control (ECDC) for a select number of European countries; and government sources for the United Kingdom, the United States, Canada, Israel, Algeria, Switzerland, Serbia. Which means there are multiple other sources from countries that could be missing. Other data for variables such as: **Testing for COVID, Vaccinations** and other variables are collected from various sources like United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.
2. Data tailored on the assumption.
   Data structure is built with the assumption that all variables relate to one another. For example, data collected on confirmed deaths from John Hopkins could not be the same data of actual confirmed deaths collected by the government of the represented country of the variable, the current data set is tailored for the best fit and narrow assumption.
3. Underrepresenting populations.
   There is a chance that some crucial variables like gender, age and preexisting conditions are not represented in the current data set. Let alone the outside factor variables like GDP of certain countries and medical facility capacity.
4. Faulty Interpretation.
   This part of bias can be closely related to the underrepresented populations, for example the types of vaccines or deaths could be different. There are possible cases where people who died from other illnesses or conditions while being tested positive for COVID, does not necessarily mean they died directly from COVID, or because the country applying vaccination does not imply that the number of deaths is correlated to the death rate.
5. Analytics bias.
   There is much more data on the developed countries due to technology and vigorous data recording of the said developed countries. Which makes other underdeveloped countries in question of the data validity. For example, COVID data on Afghanistan could be not as accurate as the data acquired from the U.S., since there are many more data recordings in the U.S. on COVID than on Afghanistan.
6. Confirmation bias.
   There is a chance that the sources for the said data like United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc. do not give a full scope of the valid death cases and ICU admissions, these could be just generalized estimations based on the best data resources the said institutions have, there is a lot of missing data in

the table marked as N/A, which could mean that some of these institutions did not bother collecting data on some countries or simply did not have access to it.

7. Cognitive bias.
   This applies to the selected variables of this data set. Since the creators focused on variables such as: Vaccinations, Tests, Hospital & ICU, Confirmed Cases, Confirmed Deaths, Reproduction rate, Policy Response and other variables, does leave out other important variables which could act as a bivariate response variable. For the main example, confirmed deaths could be explained by multiple other variables not represented in this dataset; variables such as: age, pre existing conditions, country GDP, anti-COVID mandates etc. Using this dataset to prove hypotheses might leave out some crucial explanatory variables out of the picture.
8. Outlier Bias.
   This bias could act in the form of that some countries have more data on COVID than the others, or some countries have different data dependable variables, for example GDP or Vaccination which creates higher than usual death rate. Again, for example if we try to compare data from Afghanistan and the U.S. there will be some sound outliers due to difference in data scope and outside factors.

Subjective perception.

This dataset is well organized and has some very interesting data which is updated daily and could be used to test various hypotheses related to COVID epidemic. For example, one can compare various countries and see which had more deaths and try to find out what variables contributed to more or less COVID casualties. However, hypotheses like comparing COVID related deaths between U.S. and Canada will not answer the reason for the said results. The data might show enough correlation between vaccination and hospitalization, but it will not account for the outside factors like quarantine measures and available hospital equipment. Also, some crucial variables like age, gender, pre existing conditions play a huge role in COVID death rates. This data lacks the experiment implementation and therefore it does better serve at documenting the COVID pandemic and proving various hypotheses and comparison on COVID related deaths, this data will not answer questions like: What is the best action to stem COVID spread or lower the COVID related deaths.

Pseudoscience.

There is not much room for pseudoscience or IFO evaluation in this dataset, most of the variables are related to COVID pandemic. The main question could lie in the confirmed COVID deaths, as it might be possible that some of that data might have faulty entries, let's say if the person died from other causes but tested positive for covid and was recorded as a COVID related death. The data in this dataset is mined from big global organization sources which take pseudoscience out of the equation when they collect data. There are famous cases of outside sources where communities believed that they are cured by their faith instead of vaccination or safety measures, however this dataset does not include such variables. Perhaps the pandemic is relatively new and there is not much room in distinguishing between pseudoscience and real scientific proven variables which were used in this dataset.

Conclusion

The analysis showed that extreme poverty, smokers rate, and vaccinations are correlated with total death by COVID19. Even though our initial hypothesis is that rich countries would have more control on dealing COVID19 situation, the results show that GDP per capita is not statistically correlated with COVID19 death. Instead, extreme poverty is more correlated and statistically significant with dependent variables. Interesting, the smoker rate is strongly correlated with our dependent variable, however, it has different outcome when it splitted by sex female and male. Vaccination is correlation, however, due to limited dataset, we could not get a better input on how strongly it is actually correlated with our dependent variable. To predict the total death by COVID19, the best prediction independent variable is extreme poverty, smokers rate, and vaccinations.

Reference

Food and Drug Administration. (2021, August 23). *FDA approves first COVID-19 vaccine*. U.S. Food and Drug Administration. Retrieved November 1, 2021, from https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine.

Ewing, J. (2020). United States is the richest country in the world, and it has the biggest wealth gap. *New York Times.* Retrieved from https://www.nytimes.com/2020/09/23/business/united-states-is-the-richest-country-in-the-world-and-it-has-the-biggest-wealth-gap.html

Translated by Content Engine, L. L. C. (2021, Oct 22). COVID vaccines prevent death from delta variant in more than 90% of cases. CE Noticias Financieras Retrieved from http://proxyau.wrlc.org/login?url=https://www.proquest.com/wire-feeds/covid-vaccines-prevent-death-delta-variant-more/docview/2585038271/se-2?accountid=8285

World Health Organization. (2021, November). *Timeline: Who's covid-19 response*. World Health Organization. Retrieved November 1, 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=CjwKCAjwoP6LBhBlEiwAvCcthK1n740KnrK3GtyK1LTkYQHri1zsfRPNNefssBDV23F2juN2iz9fQBoC5uYQAvD_BwE#event-193.

World Health Organization. (2021, November). *Who has a coronavirus (COVID-19) dashboard*. World Health Organization. Retrieved November 1, 2021, from https://covid19.who.int/.

Valero, M., & Valero-Gil, J. (2021). Determinants of the number of deaths from COVID-19: Differences between low-income and high-income countries in the initial stages of the pandemic. International Journal of Social Economics, 48(9), 1229-1244. doi:http://dx.doi.org/10.1108/IJSE-11-2020-0752

Hansen, L. H., Rasmussen, T. L., & Villesen, P. (2021). *Social Compliance During High Stringency Periods Efficiently Reduces COVID-19 Incidence: Evidence from Google Mobility Reports* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-501561/v1