Predicting new COVID-19 cases in the US based on trends in google searches for COVID-19 symptoms and related topics using a multiple linear regression model

Medha Srivastava

2020-12-08

Abstract

COVID-19 has been spreading rapidly throughout the world since late 2019, becoming a major pandemic and a major point of research. A crucial goal that could aid the world's ability to minimize damage from the pandemic is to be able to predict increases in cases before they happen. Multiple linear regression models are useful for using data to predict an outcome. This report employs an MLR model using Google Trends search data on COVID-19 related terms, and data on the number of new COVID-19 tests per day in the US, to predict the number new COVID-19 cases per day in the US. The model finds Google searches for "Covid testing" and "Covid vaccine" to be the most significant predictors for new cases per day, along with the number of tests per day. Propensity score analysis found a non-causal relationship between high searches and new cases, showing a correlational relationship between the GT searches and new COVID-19 cases in the US. Code and data supporting this analysis is available at: https://github.com/medhasrivastavaa/STA304Final/

Key words: Multiple Linear Regression, Propensity Score, Causal Inference, Observational Study, COVID-19, Google Trends, USA

Introduction

The recent COVID-19 pandemic has been a focal point of research since it gained traction in early 2020. There is an abundant amount of data surrounding all aspects even distantly related to the topic, from data on patient demographics to data on social distancing measures, etc. This also comes with a thirst for trends and patterns that can lead to useful information in preventing further negative consequences. Understanding factors that correlate with high COVID-19 numbers, and especially factors that cause high COVID-19 numbers, can lead to understanding what to do to get lower rates (Corsi, de Souza & Pagani, 2020). The USA has been especially suffering from high COVID-19 rates and a consistent prevalence of this disease, and given its significant amount of up-to-date data it serves as a useful starting point for analyzing such factors.

Beyond the logged official data on COVID-19 cases, there are often useful predictors for high cases found in more casual factors- such as increasingly common internet searches on certain topics (Kurian et al., 2020). In the case of a highly contagious disease and pandemic, it is plausible that many outbreaks and cases begin with individual research and assessment by people socially isolating in their homes. In such cases, more google searches on common COVID-19 symptoms or about COVID-19 testing in the area could indicate a high prevalence of COVID-19. Alternatively, higher searches for COVID-19 testing and face masks might indicate greater caution towards the pandemic, and therefore reflect lower COVID-19 rates. Recently, google trends data showed a higher search interest in terms such as "testing" and "vaccine", etc., and an interesting point of research would be determining whether those higher search rates are correlated with higher COVID-19 rates (Google, 2020). Being able to predict COVID-19 rates based on internet searches data would be

intriguing not only due to its epidemiological study benefits, but also due to its insight on the populations mindsets and independent actions in response to serious global situations.

This report will use multiple linear regression model using the daily number of new tests in the US and google search trends in the US for 8 key words ("covid testing", "covid symptoms", "face mask", "lysol", "covid vaccine", "fever", "sore throat", and "shortness of breath") as predictors for the number of new COVID-19 cases in the US over time. To confirm the correlational relationship between the google trends (GT) and new COVID-19 cases, a propensity score matching analysis will be used to check for a causal link between the predictors and the outcome. The Methodology section explains the data collection process and the model used to describe the correlation relationship between the predictors and new COVID-19 cases, and the model used in the propensity score matching analysis. The Results section discusses the results from the models, and their interpretations. Finally, the Discussion section will elaborate on the significance of the analysis, relevant conclusions, and the limitations of the study.

Methodology

Data:

This analysis used 2 different datasets. The first dataset containing the daily number of new COVID-19 cases, new tests, and average country demographics for all countries in the world, was obtained through the Our World In Data database (OWID) (Hasell et al., 2020). The OWID data was obtained through official government and hospital reports of COVID-19 cases collected by JHU, which is updated daily (Hasell et al., 2020). This dataset was cleaned to include only information for the US, and to remove all variables except the date, number of new COVID-19 cases, new tests, and total COVID-19 cases per day. Demographic variables such as diabetes prevalence, GDP, ages over 70, or female smokers were removed due to the fact that the values for each demographic variable different only by country, not by date. Specifically, the values for each variable were exactly the same for all observations (dates) for the US data, and therefore could not be used as predictors of the number of new cases.

The second dataset was collected through the gtrendsR package in R, allowing access to information about trends in Google searches for specific keywords in specific regions. The keywords searched in this case were "covid testing", "covid symptoms", "face mask", "lysol", "coronavirus cases", and "covid vaccine", along with the most commonly searched COVID-19 symptoms "fever", "sore throat", and "shortness of breath" (Google, 2020). The keywords were chosen based on the Google data for most commonly searched queries related to "COVID-19" in the past months (Google, 2020), as well as results from previous correlation studies that found a high correlation between these keywords and COVID-19 cases in the world (Kurian et al., 2020) (Ayyoubzadeh et al., 2020). The geographic region for all of these searches was set to the US, for the time period of March 22, 2020 to December 16, 2020. The dates were chosen as they matched the dates in the OWID dataset, allowing comparison between the case data and the Google trend data. This data extraction process was done in R and can be seen in the Appendix.

It should be noted that the GT results are provided as scores from 1-100, and are relative search volumes rather than the number of actual search results. The score of 1-100 is given by the number of searches for the keyword relative to the total number of Google searches in the given date and location (Rogers, 2016). This is ideal for an analysis such as this which aims to see patterns in searches for a term, as spikes in the GT data indicate higher than usual searches for the keyword. Additionally, this ensures the trends for specific keywords are not shown as highest for locations with consistently more searches in general, or for dates where there were simply more Google searches than other days (such as holidays or weekends). This is a major strength of the GT data. The OWID dataset and the GT dataset were then combined into 1 complete dataset, which had the number of new cases and new tests per day in the US, as well as the Google trend search scores for each of the keywords per day in the US.

The target population in this study is the entire population of the US, as the aim of the study is to measure the relationship between Google searches (aiming to represent people's interests and curiosity regarding

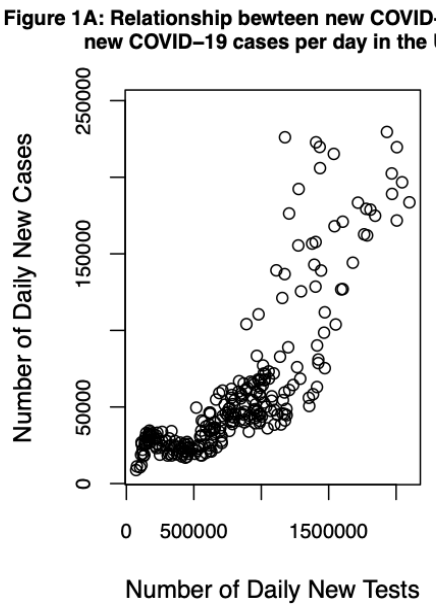
certain topics) by everyone in the US and the number of positive COVID-19 cases in the US. In this study, the sampling frame and the sample is significantly smaller than the target population, as it is limited to people in the US that have regular access to the internet and use Google. This analysis is largely a correlation analysis, since the data is observational data on COVID-19 cases and Google search trends, and because the population involved in the Google search trends is not necessarily the same population in the positive COVID-19 cases data. For example, someone searching "covid symptoms" on Google is not necessarily someone that also tests positive for COVID-19, on any date. This is a potential weakness of the study, especially for attempting to make causal links between a higher search volume for a keyword and new cases. However, the goal of obtaining predictions of overall new cases in the US based on keyword searches is based on the hypothesis that general search trends in a geographical area such as the US indicate changes in the overall case rates in the region. The data is for all dates from March 22 until December 16, which is slightly limited since the first case of COVID-19 in the US was recorded on January 20, 2020 (Holshue et al., 2020). This analysis could be improved with data from all possible dates.

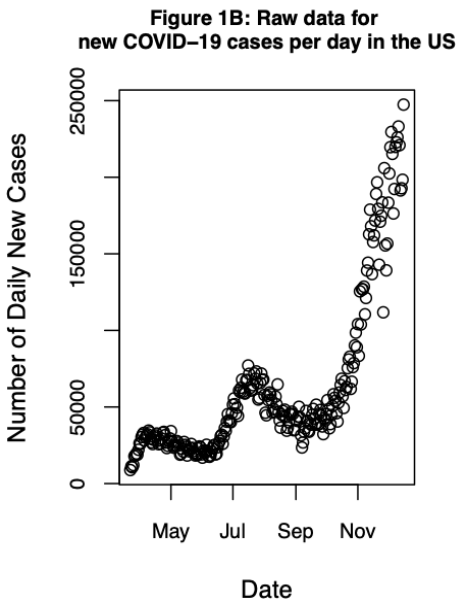
```
## Factors are dropped from the summary
```

Table 1: Summary Characteristics of All data

	N	Mean	SD	Min	Q1	Median	Q3	Max
date	270	18477.50	78.09	18343	18410.0	18477.5	18545	18612
total_cases	270	5337550.40	4245083.85	34855	1727357.0	4728870.5	7705153	16964180
total_tests	264	75605161.59	62683334.93	461024	16934942.5	65245732.5	122357174	213273421
new_cases	270	62733.91	53961.83	8830	27764.0	44480.5	66508	247403
new_tests	264	806380.97	464748.34	72178	444103.5	784156.0	1074440	2098517
covidsymptoms	270	10.63	3.99	4	8.0	10.0	14	20
lysol	270	3.67	2.74	1	2.0	3.0	4	21
facemask	270	14.42	10.48	3	7.0	12.0	19	100
covidtesting	270	15.47	10.57	3	7.0	13.0	21	49
covidvaccine	270	3.56	5.05	1	2.0	2.0	3	35
fever	270	39.49	12.85	24	33.0	36.0	41	100
shortnessbreath	270	4.04	1.95	2	3.0	4.0	4	13
sorethroat	270	14.07	5.12	7	11.0	13.0	15	40

Table 1 shows the baseline characteristics of the data, showing the highest mean values are for the number of total cases and new cases, and for the number of new and total tests. The cases and tests values are in primarily in the millions, whereas the values for the GT scores are in the range of 1-100. The maximum values for the different GT keywords are very different, where "covid symptoms" has a maximum of 20 while "face mask" has a maximum of 100, because of their relative scoring. The maximum of 100 for "face mask" does not indicate it was searched significantly more than "covid symptoms" overall, but instead it means "face mask" was searched more than other things on a specific day (however the actual amount it was searched could still be less than the amount "covid symptoms" was searched on the date it reached its maximum). Based on standard deviation, "face mask", "covid testing", and "fever" appear to be the least variable in their scores.





The number of new tests per day in the US was also used as a predictor because it was obtained from the same source as the number of new cases per day, serving as a control for bias due to the varying source for the data. Figure 1A shows a relatively linear, positive relationship between the number of new tests and the number of new cases per day. There is a clear spike in the linear relationship where small numbers of new tests correlate with high cases, although this may be due to the late introduction of widespread COVID testing in the US (AJMC, 2020). This linear relationship suggests new tests should be a good predictor for the number of new cases using a linear model. Figure 1B shows the spread of the number of new cases in the US over time, from March 22 to December 16. There is a clear increasing trend throughout, although the number of new cases tends to peak at different times. This is common for the spread of a pandemic over time, as separate outbreaks or different preventative measures or lack of can lead to different peaks in cases in time, creating the epidemiological curve generally seen in Figure 1B (Baldwin, 2020). The notable peaks are around March/April, July, and presumably at November and December.

Figure 2

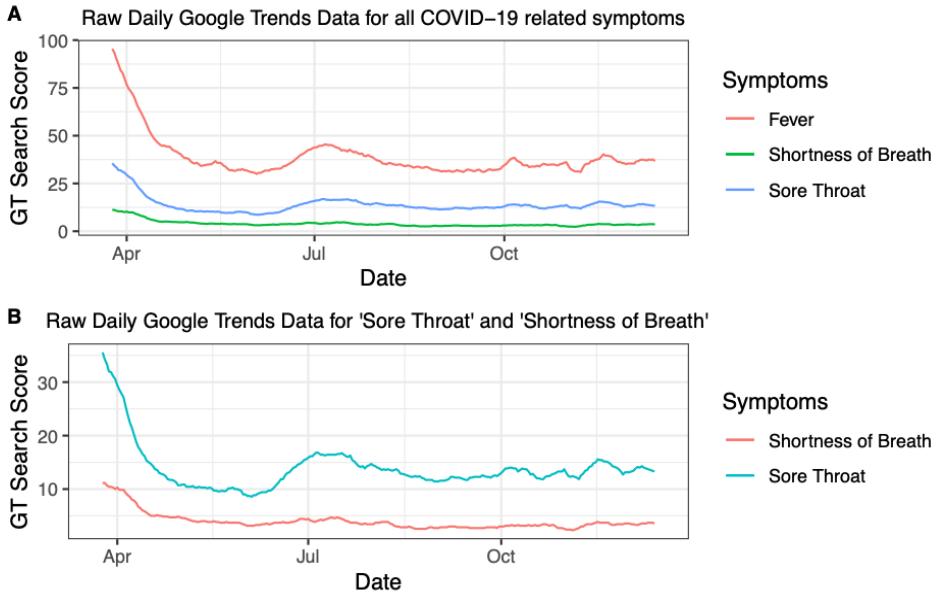


Figure 2A shows a plot of the GT scores per day for the symptom keywords. All the symptom keywords can be seen to have peaked right before April, which was soon after when the COVID-19 crisis was declared a pandemic (in March) and once the virus had been in the US for 4 months (AJMC, 2020). It is possible the symptoms for COVID-19 became more well-known around this time, as it was declared a national US emergency in the end of March (AJMC, 2020). "Fever" was consistently the highest searched symptom out of the 3 symptoms in this study. The notable peaks for the "Fever" data is around April, July, and November. The scores for the other 2 symptoms are shown more clearly in Figure 2B, where the peaks for "Sore Throat" appear at the same timeline.

Figure 3: Raw Daily Google Trends Data for COVID−19 related keywords

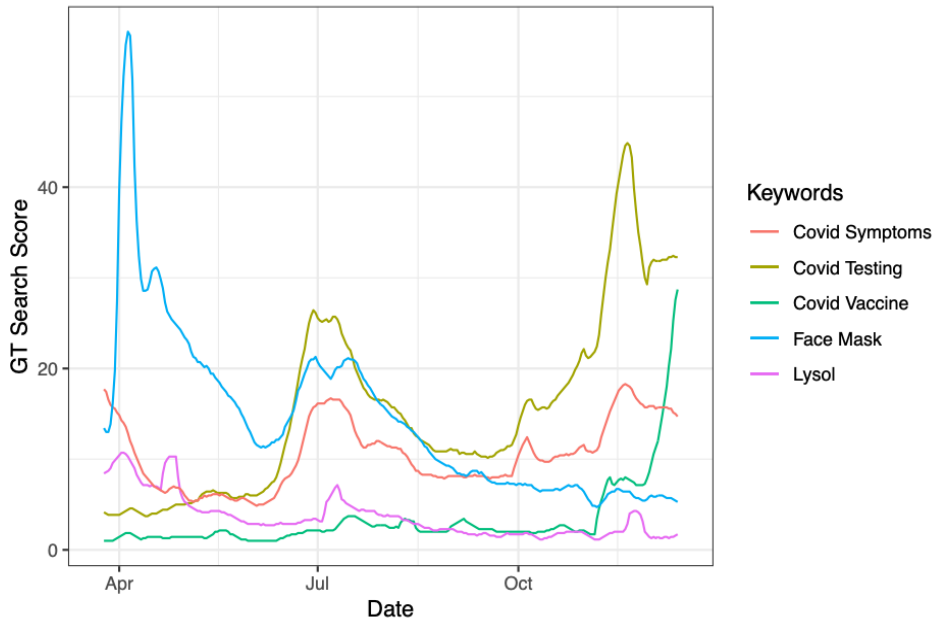


Figure 3 shows the GT scores for the 5 COVID-19 related queries, which follow a similar epidemiological curve as the number of new cases data in Figure 1B showing peaks in July and November, with an increasing rate in December. Searches for "Covid Vaccine" remained relatively low throughout the timeline and proved an exception to the epidemiological pattern, but significantly increased in November and December. This is very likely due to the recent approval and administration of COVID-19 vaccines in December (BBC, 2020). Conversely, searches for "face mask" followed the general peaks as the other keywords until around August, where searches tapered off and have remained low. This suggests "face mask" may have been a useful predictor in the earlier timeline of the pandemic, but may not be as significant now. "Covid testing" has the most searches from all 5 keywords in recent months.

Model

This analysis uses a multiple linear regression (MLR) model, which makes use of multiple predictors to predict the outcome of the dependent variable. In this case, the dependent variable is the number of new COVID-19 cases per day in the US, while the predictors are the GT search scores of each of the 8 keywords related to COVID-19 and its symptoms ("covid testing", "covid symptoms", "face mask", "lysol", "covid vaccine", "fever", "sore throat", and "shortness of breath"), as well as the number of new tests daily in the US. The general form of an MLR model is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

A MLR model was chosen because the outcome of interest, the number of new cases per day, is a continuous numeric variable, while the predictors are also all numeric continuous predictors. This makes predictions from an MLR model useful. Additionally, an MLR model with many different predictors can indicate which

predictors are useful, strong predictors. With Google trends, there has been shown correlation between popular COVID-related keywords and COVID-19 numbers, but using the trends to forecast future COVID-19 case numbers would be useful, and an MLR can accomplish this and indicate which keywords predictors are most useful predictors (Kurian et al., 2020). In addition to the predictions from the MLR model, an additional column was added to the dataset showing the 7-day moving average. The purpose of including a 7-day moving average for the predicted number of new cases and the actual number of new cases was because epidemiological data such as this dataset tends to have many minor fluctuations, so a moving average finds the average for a 7 day period, using the nearest 7 days to any specific date. This was used to create a smoother curve to visualize the COVID-19 pattern more clearly, and to partially minimize the effect of a lag between GT searches and COVID-19 new cases (HealthData, 2020).

To predict the number of new cases a day in the US, there are 9 total predictors, with the MLR model described below:

$$y_i = -24990 + 0.05245NT - 541CS + 1329L + 209.8FM + 934.9CT + 6857CV - 244.6F + 1012SB + 161.8ST$$

Where: y_i is the number of new COVID-19 cases per day in the US NT is the number of new COVID-19 tests per day in the US CS is the GT search score for "Covid Symptoms" L is the GT search score for "Lysol" FM is the GT search score for "Face Mask" CT is the GT search score for "Covid Testing" CV is the GT search score for "Covid Vaccine" F is the GT search score for "Fever" SB is the GT search score for "Shortness of Breath" ST is the GT search score for "Sore Throat"

In this case, the intercept of -24990 means that the mean predicted value for the number of new COVID-19 cases in the US in a day where the number of searches for all 8 key words is 0, and there are no new tests, would be -24990. The slope of 0.05245 for the number of new tests per day indicates that with all other variables constant, an increase of 1 new test in a day would lead to a mean increase of 0.05245 in the number of cases per day. Similarly, the slope of 1329 for the GT score for "Lysol" indicates with all other variables constant, a single unit increase in the GT score for "Lysol" would correlate with an increase of 1329 in the average number of new COVID-19 cases per day in the US. A similar interpretation applies to the other GT search scores as well, where the only 2 key words negatively correlated with the number of new cases per day are "Covid Symptoms" and "Fever", where an increase in a single unit in the scores for those words would cause a decrease in the average number of new cases per day.

After fitting a MLR model to the data, backward elimination by AIC was used to eliminate the least useful predictors and create the model with the best possible fit. Backward elimination by AIC begins with a MLR model with all predictors, and removes the predictors that lead to a model with the lowest AIC, which indicates the best model (Zhang, 2016). This process of variable selection was chosen because ideally this model will contain as many useful predictors as possible, since the prediction of COVID-19 cases is an important outcome that should use all relevant information. However, too many variables leads to higher variance for estimated regression coefficients, and can overfit the data, so variable selection is a necessary step in fitting the best MLR model. Through this process, the variables "covid symptoms", "fever", and "sore throat" were removed and the following final MLR model was fit:

$$y_i = -27280 + 0.05273NT + 1334L + 211FM + 761.1CT + 6822CV + 1110SB$$

Using the significant predictors identified in the MLR model, a propensity score analysis was performed. Propensity score matching is useful in observational data studies in interpreting causal links between a treatment and outcome, by matching and comparing pairs of treated observations with untreated observations that have the same propensity (Austin, 2011). In this analysis, since the data is observational and there is no specific treatment group, the treatment group was formed to be "high searches", i.e. the dates that had relatively high searches for all of the predictors. The propensity score analysis aimed to see if the outcome of high COVID-19 cases could be explained by the treatment of high searches, with a causal relationship. This propensity score analysis was performed in R (Alexander, 2020). The number of new tests was not used as a variable in this part of the analysis, as the treatment group being analyzed was high searches for GT keywords.

Results

Table 2: Multiple linear regression model using all 9 predictors

term	estimate	std.error	statistic	p.value
(Intercept)	-3.405466e+04	5831.8551063	-5.8394216	0.0000000
new_tests	5.203610e-02	0.0049513	10.5096574	0.0000000
covidsymptoms	-6.383479e+02	888.7621239	-0.7182439	0.4732671
lysol	1.676822e+03	634.9576544	2.6408402	0.0087826
facemask	1.280203e+02	148.3964558	0.8626908	0.3891213
covidtesting	9.862634e+02	337.1718478	2.9251060	0.0037547
covidvaccine	7.630380e+03	453.7387380	16.8166811	0.0000000
fever	5.982113e+02	326.7083715	1.8310253	0.0682682
shortnessbreath	-7.501001e+02	1439.6301579	-0.5210367	0.6027952
sorethroat	-5.736326e+02	672.3272156	-0.8532045	0.3943498

Table 3: Reduced multiple linear regression model with 6 predictors

term	estimate	std.error	statistic	p.value
(Intercept)	-2.775462e+04	4651.1352809	-5.9672790	0.0000000
new_tests	5.127310e-02	0.0048941	10.4764735	0.0000000
lysol	1.905744e+03	621.3405309	3.0671490	0.0023920
facemask	8.500633e+01	144.9063651	0.5866293	0.5579676
covidtesting	7.998644e+02	199.7852794	4.0036202	0.0000817
covidvaccine	7.570082e+03	450.0244082	16.8214914	0.0000000
shortnessbreath	7.106409e+02	750.3602987	0.9470663	0.3444946

Table 2 shows the initial MLR model with all 9 predictors, where the predictors with the lowest p-values are the number of new tests, Covid Vaccine, Covid Testing, Shortness of Breath, Lysol, and Sore Throat.

Table 3 shows the final MLR model fit to the data using backward AIC elimination. Based on Table 3, the most significant predictors for the number of daily new cases of COVID-19 in the US are the number of new tests per day and the GT searches for "Covid Testing" and for "Covid Vaccine". Slightly less significant predictors are GT searches for "Lysol" and for "Shortness of Breath" as a symptom. From the final MLR model, "Face Mask" was the only predictor shown to be not significant, with a high p-value of 0.52, supporting the prediction (made in the Data section of this report, based on Figure 3) that searches for "Face Mask" may have been a useful predictor early in the pandemic but no longer is correlated with higher COVID numbers. The regression coefficients for Covid Testing is 761.1, indicating that with all other variables constant an increase in the GT search score for "Covid Testing" by 1 unit would lead to a 761.1 increase in the mean number of new cases per day in the US, while a unit increase in GT search scores for "Covid Vaccine" would result in an increase of 6822 in the mean number of new cases per day. A notable finding from the initial MLR model and the final MLR model is that the only COVID-19 symptoms whose GT is a significant predictor for the daily number of new cases in the US is "Shortness of Breath", and even that is relatively less strong predictor than other keywords like "Covid Testing" and "Covid Vaccine". Searches for COVID-19 symptoms are not the strongest predictors for new COVID-19 cases based on this model.

The propensity score analysis found that the treatment variable, high searches (for all keywords), was not a good predictor for the number of new cases daily in the US. "Covid Testing" and "Covid Vaccine" were again found to be strong predictors for the number of new cases, supporting the MLR model chosen. This indicates there is no causal link between the GT search numbers and the number of new cases of COVID-19 in the US. Since there is no causal link found, the MLR model that best predicts the outcome of interest is

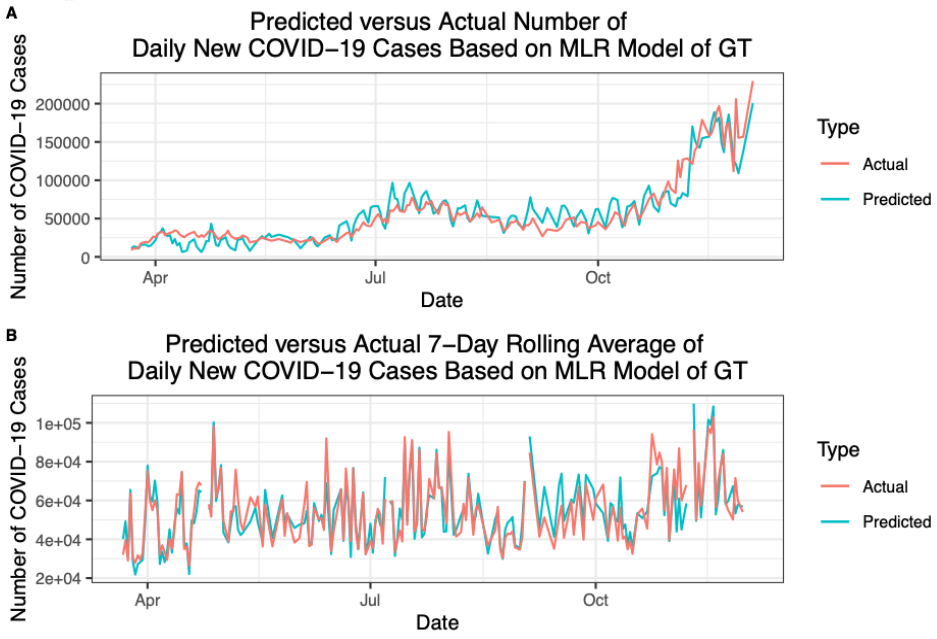
Table 4: Propensity Score Regression Analysis

	(1)
(Intercept)	5798.807
	(4471.095)
shortnessbreath	-1062.817
	(890.838)
lysol	110.131
	(653.557)
covidtesting	1975.411 ***
	(206.056)
covidvaccine	8919.449 ***
	(997.872)
highsearches1	2261.613
	(2777.859)
N	198
R2	0.796
logLik	-2232.833
AIC	4479.665

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

the model shown in Table 3, and that model will be used in the remainder of this report as the best fit for relating GT data to the number of new cases in the US.

Figure 4:



The predicted number of new cases per day in the US based on the MLR model fit in this analysis appears to correlate strongly with the actual number of new cases per day, based on Figure 4A. The predicted model shows greater fluctuation than the actual numbers, but overall predicts the curves and peaks relatively closely to the actual. This is seen more clearly in Figure 4B, which graphs the 7-day moving average for the predicted number of new cases and the 7-day moving average for the actual number of new cases. The predicted values in Figure 4B very closely predict the actual values, indicating the predictors found through the MLR model are good predictors for the number of new cases.

Discussion

Summary

To predict the number of new cases per day in the US, Google search trends data and data on the number of new tests per day were used as predictors in a multiple linear regression model. A propensity score matching analysis was done to find if the GT predictors were causally related to the number of new cases as an outcome, and no causational link was found indicating high searches for the GT keywords tested do not cause the observed changes in the number of new cases per day in the US. The MLR model used to fit the data predicted the actual number of new cases per day relatively well, seen in Figure 4, and the 7-day moving average for both predicted and actual number of new cases per day was calculated and plotted as well to show a smoother curve.

Conclusions

The results from this MLR analysis show that the number of new tests per day in the US and the Google search trends for the terms "Covid testing" and "Covid vaccine" are the most significant predictors for the number of new cases in the US per day. This is an interesting finding as the use of Google search data to predict COVID-19 rates and other endemic or pandemic rates is a very useful ability that could predict major outbreaks and allow for preparations and minimizing its impact. The MLR model showed that the only symptom that was a significant predictor for the number of new cases was Shortness of Breath, and overall searches for symptoms were not strong predictors for the outcome. This finding is contrary to hypotheses based on previous research, where many correlation analyses of Google Trends data related to COVID-19 symptoms found a relatively strong correlation between searches for symptoms and cases (Kurian et al., 2020). This could possibly be explained by the use of lag correlations in the research mentioned, as it is likely that GT searches for COVID-19 symptoms might increase earlier than the actual number of cases increase, rather than on the same day. The use of a 7-day moving average was aiming to minimize this effect. This model can be used to predict the number of new COVID-19 cases in the next month using GT search scores throughout the month, making this a very useful model for epidemiological research for COVID-19. This has a significant global impact, as it can be further extended to other countries to predict new cases using the same or additional different GT key words. This leads to the conclusion that GT trends data, specifically searches for "Covid testing" and "Covid vaccine" as well as the number of new tests per day, can be used to predict the number of new COVID-19 cases per day in the US. The correlation analysis results further support this conclusion, showing that GT searches for "Covid testing" and "Covid vaccine" are the most correlated with higher daily new cases in the US.

The results of the propensity score analysis showing a lack of causation between high searches for all the keywords and the outcome lead to the relationship between the number of new COVID-19 cases and the GT searches for the 8 keywords to be correlation-based. Considering the nature of the data, this finding is not surprising. It is unlikely that higher searches for COVID-19 related topics on Google would directly be a cause for changes in the number of new cases per day, but rather it is more likely the Google trends are reflective of upcoming changes in COVID-19 numbers and therefore serve as important predictors that correlate with the COVID-19 cases.

Overall, although Figure 4 shows the predicted number of new cases over time to fluctuate more than the actual number, the model is highly beneficial in its accurate prediction of the peaks in daily new COVID-19 cases in the US. Predicting the peaks in new cases is useful in minimizing economic and public health damage for countries, so this analysis shows GT trends can be useful in adding to existing official data to predict peaks through a more holistic analysis. In modern times, it is important to note the role of external data outside of the collected demographic and hospital data in every country in reflecting changes in the population. Although the results show specific GT keywords to be strong predictors of the COVID-19 daily cases, they would be ideal predictors alongside additional data such as age groups, average economic standings of countries, or availability of medical resources, etc. (Ossola, 2020).

Weaknesses & Next Steps

There are several limitations and weaknesses to this analysis. Firstly, the data used from March 22, 2020 until December 16, 2020. Although this data is recent and covers the majority of available data, the COVID-19 crisis began in the US in January and therefore there are 2 months of data missing from this analysis (Holshue et al., 2020). This was a limitation of the provided data, however since COVID-19 only entered the US in 2020, the data so far still is biased due to the small sample of only being 1 year of data. As more months of information occur, this model could be updated and improved to include new patterns and trends. Secondly, the data extracted from Google Trends only includes data for 8 keywords. This analysis could be significantly improved by beginning the model with many more keywords, and using variable selection techniques to identify the relevant useful predictors. Additionally, GT data may be biased in that it can be influenced by external factors, such as the media or social media trends. For example, previous uses

of a similar model attempting to predict flu rates was occasionally incorrectly predicting high rates due to higher GT data, however the higher GT searches was due to the flu being covered more heavily in the media leading to more people researching it (Ossola, 2020). The GT search keywords used in this study may also be reflective of policies in the country rather than actual circumstances—such as higher searches for face masks when they are required by the government rather than when the cases are higher. Usually policies are reflective of actual circumstances so the use of GT search keywords is still useful, however in the US the discrepancy between COVID-19 cases and policies has been noted as differing from other countries (Yong, 2020). This could be minimized in future studies by additional keywords to minimize external influences on the search rates of specific keywords, and by using internet data outside of just Google searches such as Youtube searches or Facebook searches.

A non-linear model could also be considered in a future study analysing GT trends as predictors for COVID-19. Additionally, adding a second layer of comparison to the model by investigating the number of new cases and GT trends by geographic regions (by states in the US or by countries, or regions within other countries) would minimize the effect of certain geographic regions being more concentrated in GT searches or in COVID-19 cases and impacting the overall distribution. Another major limitation is due to the relatively small sample size for this data existing so far, a useful training set and test set split was not made in this analysis. A promising future step would be performing a similar analysis including a training and test set split to give a more accurate model and prevent overfitting. Considering the high number of small fluctuations in this data and most epidemiological data, overfitting is a high possibility. Another potential next step would be to use internet data such as GT to predict individual risk for COVID-19 or other diseases, based on geographic search information in their area as well as demographic information.

Overall, the use of Google Trends data to predict COVID-19 cases over time is a useful model with some strong predictors, the identification of which is very beneficial in predicting and preventing the dangers of the pandemic. However, there are potential sources of bias that should be addressed in future studies and there are many additional factors that would aid the predictions of this model, such as geographic data, demographic data, and other internet data.

References:

1. Corsi, A., de Souza, F.F., Pagani, R.N. et al. Big data analytics as a tool for fighting pandemics: a systematic review of literature. J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-02617-4

2. Kurian, S. J., Bhatti, A., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., . . . Bydon, M. (2020, November 1). Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis. Retrieved December 23, 2020, from https://www.mayoclinicproceedings.org/article/S0025-6196(20)30934-4/fulltext

3. Google. (2020). Coronavirus Search Trends. Retrieved December 23, 2020, from https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en

4. Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). https://doi.org/10.1038/s41597-020-00688-8

5. Rogers, S. (2016, July 01). What is Google Trends data - and what does it mean? Retrieved December 23, 2020, from https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8

6. Holshue, M.L.; DeBolt, C.; Lindquist, S.; Lofy, K.H.; Wiesman, J.; Bruce, H.; Spitters, C.; Ericson, K.; Wilkerson, S.; Tural, A.; et al. First Case of 2019 Novel Coronavirus in the United States. N. Engl. J. Med. 2020, 382, 929–936.

7. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. JMIR Public Health Surveill 2020; 6(2). DOI: 10.2196/18828

8. AJMC. (2020, November 25). A Timeline of COVID-19 Developments in 2020. Retrieved December 23, 2020, from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

9. Baldwin, R. (2020, March 12). It's not exponential: An economist's view of the epidemiological curve. Retrieved December 23, 2020, from https://voxeu.org/article/it-s-not-exponential-economist-s-view-epidemiological-curve

10. BBC. (2020, December 08). Covid-19 vaccine: First person receives Pfizer jab in UK. Retrieved December 23, 2020, from https://www.bbc.com/news/uk-55227325

11. Austin P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research, 46(3), 399–424. https://doi.org/10.1080/00273171.2011.568786
Zhang Z. (2016). Variable selection with stepwise and best subset approaches. Annals of translational medicine, 4(7), 136. https://doi.org/10.21037/atm.2016.03.35

12. Alexander, R. (2020, November 05). Difference in differences. Retrieved December 23, 2020, from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html

13. HealthData. (2020, December 22). COVID-19 Daily Rolling Average Case, Death, and Hospitalization Rates. Retrieved December 23, 2020, from https://healthdata.gov/dataset/covid-19-daily-rolling-average-case-death-and-hospitalization-rates

14. Ossola, A. (2020, September 23). Here's what Google trends can tell us about the spread of coronavirus. Retrieved December 23, 2020, from https://www.weforum.org/agenda/2020/09/google-search-trend-data-coronavirus-health-global-epidemiology/

15. Yong, E. (2020, August 06). How the Pandemic Defeated America. Retrieved December 23, 2020, from https://www.theatlantic.com/magazine/archive/2020/09/coronavirus-american-failure/614191/

16. B. Hofner (2019). papeR: A Toolbox for Writing Pretty Papers and Reports, R package version 1.0-4, https://CRAN.R-project.org/package=papeR.

Appendix:

```
## Main Dataset Creation Code and Cleaning Code

#US COVID cases dataset
usdaily <- read.csv("owid-covid-data.csv")
usdaily <- usdaily %>%
  filter(location == "United States") %>%
  select(date, total_cases, total_tests, new_cases, new_tests)
usdaily$date <- as.Date(usdaily$date)

#Google Trends data extraction and variables setup
g_trend <-  gtrends(
  keyword = c("covid symptoms", "covid testing", "face mask", "lysol", "covid vaccine"),
  geo = "US",
  time = "2020-3-22 2020-12-16"
  )

g_trend2 <-  gtrends(
  keyword = c("fever", "shortness of breath", "sore throat"),
  geo = "US",
```

```
  time = "2020-3-22 2020-12-16"
  )

bytime <- g_trend$interest_over_time
bytime <- bytime %>%
  select(date, hits, keyword)
bytime$date <- as.Date(bytime$date)
bytimekw1 <- bytime %>%
  filter(keyword == "covid symptoms")
bytimekw2 <- bytime %>%
  filter(keyword == "lysol")
bytimekw3 <- bytime %>%
  filter(keyword == "face mask")
bytimekw4 <- bytime %>%
  filter(keyword == "covid testing")
bytimekw5 <- bytime %>%
  filter(keyword == "covid vaccine")

#Creating the combined dataset
alldata <- merge(usdaily, bytimekw1, by= "date")
alldata <- merge(alldata, bytimekw2, by= "date")
alldata <- merge(alldata, bytimekw3, by= "date")
alldata <- merge(alldata, bytimekw4, by= "date")
alldata <- merge(alldata, bytimekw5, by= "date")

colnames(alldata)[6] <- "covidsymptoms"
colnames(alldata)[8] <- "lysol"
colnames(alldata)[10] <- "facemask"
colnames(alldata)[12] <- "covidtesting"
colnames(alldata)[14] <- "covidvaccine"

alldata <- alldata %>%
  select(-keyword.x, -keyword.y, - keyword)

btt <- g_trend2$interest_over_time
btt <- btt %>%
  select(date, hits, keyword)
btt$date <- as.Date(btt$date)
btt1 <- btt %>%
  filter(keyword == "fever")
btt2 <- btt%>%
  filter(keyword == "shortness of breath")
btt3 <- btt%>%
  filter(keyword == "sore throat")
alldata <- merge(alldata, btt1, by= "date")
alldata <- merge(alldata, btt2, by= "date")
alldata <- merge(alldata, btt3, by= "date")
colnames(alldata)[11] = "fever"
colnames(alldata)[13] = "shortnessbreath"
colnames(alldata)[15] = "sorethroat"

alldata <- alldata %>%
  select(-keyword.x, -keyword.y)
```