The Causality Inference of Public Interest in Restaurants and Bars on COVID-19 Daily Cases in the US: A Google Trends Analysis

Milad Asgari Mehrabadi, Nikil Dutt, Amir M. Rahmani

Submitted to: JMIR Public Health and Surveillance
on: July 25, 2020

Table of Contents

The Causality Inference of Public Interest in Restaurants and Bars on COVID-19 Daily Cases in the US: A Google Trends Analysis

Milad Asgari Mehrabadi[1] BSc, MSc; Nikil Dutt[1, 2] PhD; Amir M. Rahmani[2, 3] PhD

[1]Department of Electrical Engineering and Computer Science, University of California Irvine Irvine US
[2]Department of Computer Science, University of California Irvine Irvine US
[3]School of Nursing, University of California Irvine Irvine US

Corresponding Author:
Milad Asgari Mehrabadi BSc, MSc
Department of Electrical Engineering and Computer Science, University of California Irvine
Berk Hall, 1st Fl.
Irvine
US

Abstract

Background: The COVID-19 coronavirus pandemic has affected virtually every region of the globe. At the time of conducting this study, the number of daily cases in the United States is more than any other country, and the trend is increasing in most of its states. Google trends provide public interest in various topics during different periods. Analyzing these trends using data mining methods might provide useful insights and observations regarding the COVID-19 outbreak.

Objective: The objective of this study was to consider the predictive ability of different search terms (i.e., bars and restaurants) with regards to the increase of daily cases in the US. In particular, we were concerned with searches for dine-in restaurants and bars. Data were obtained from Google trends API and COVID tracking project.

Methods: To test causation of one time series on another, we used Granger's Causality Test. We considered the causation of two different search query trends, namely restaurant and bars, on daily positive cases in top-10 states/territories of the United States with the highest and lowest daily new positive cases. In addition, to measure the linear relation of different trends, we used Pearson correlation.

Results: Our results showed for states/territories with higher numbers of daily cases, the historical trends in search queries related to bars and restaurants, which mainly happened after re-opening, significantly affect the daily new cases, on average. California, for example, had most searches for restaurants on June 7th, 2020, which affected the number of new cases within two weeks after the peak with the P-value of .004 for Granger's causality test.

Conclusions: Although a limited number of search queries were considered, Google search trends for restaurants and bars showed a significant effect on daily new cases for regions with higher numbers of daily new cases in the United States. We showed that such influential search trends could be used as additional information for prediction tasks in new cases of each region. This prediction can help healthcare leaders manage and control the impact of COVID-19 outbreaks on society and be prepared for the outcomes.

(JMIR Preprints 25/07/2020:22880)
DOI: https://doi.org/10.2196/preprints.22880

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

The Causality Inference of Public Interest in Restaurants and Bars on COVID-19 Daily Cases in the US: A Google Trends Analysis

Abstract

Background: The COVID-19 coronavirus pandemic has affected virtually every region of the globe. At the time of conducting this study, the number of daily cases in the United States is more than any other country, and the trend is increasing in most of its states. Google trends provide public interest in various topics during different periods. Analyzing these trends using data mining methods might provide useful insights and observations regarding the COVID-19 outbreak.

Objective: The objective of this study is to consider the predictive ability of different search terms not directly related to COVID-19 with regards to the increase of daily cases in the US. In particular, we are concerned with searches for dine-in restaurants and bars. Data were obtained from Google trends API and COVID tracking project.

Methods: To test causation of one time series on another, we used Granger's Causality Test. We considered the causation of two different search query trends related to dine-in restaurant and bars, on daily positive cases in top-10 states/territories of the United States with the highest and lowest daily new positive cases. In addition, to measure the linear relation of different trends, we used Pearson correlation.

Results: Our results showed for states/territories with higher numbers of daily cases, the historical trends in search queries related to bars and restaurants, which mainly happened after re-opening, significantly affect the daily new cases, on average. California, for example, had most searches for restaurants on June 7th, 2020, which affected the number of new cases within two weeks after the peak with the *P*-value of .004 for Granger's causality test.

Conclusions: Although a limited number of search queries were considered, Google search trends for restaurants and bars showed a significant effect on daily new cases for states/territories with higher numbers of daily new cases in the United States. We showed that such influential search trends could be used as additional information for prediction tasks in new cases of each region. This prediction can help healthcare leaders manage and control the impact of COVID-19 outbreaks on society and be prepared for the outcomes.

Keywords: Bars; Coronavirus; COVID-19; Deep Learning; Google Trends; LSTM; Machine Learning; Restaurants.

Introduction

The entire world has been affected significantly by a global virus pandemic. The first case of this virus was reported in China during December 2019, and the first case outside China was discovered in January 2020 [1]. During February, the World Health Organization called this virus coronavirus disease 2019 (COVID-19) [2].

Worldwide, there have been about 14,400,000 confirmed cases, with 604,000 deaths, as of 19th of July 2020 [3]. The United States of America, with 3,830,000 confirmed cases and 143,000 deaths, is the most affected country around the world. In some states, the numbers are still increasing (e.g.,

California), while in some other states such as New York, the peak has passed, and the average daily new cases are decreasing.

Due to the rapid spreading of this virus, finding effective reasons can play a significant role in prevention policies. Using data mining and time series analysis methods, it is possible to investigate the impact of different phenomena on time series data. In economics, as an example, there are different studies that model the temporal relationship of two or more time series (e.g., the relationship between oil and gold price) using the same methods [4]. Wang et al. [5] uses the same causality inference methods to determine whether the main air pollutions and the mortality rate of respiratory diseases have relationship.

Owing to infodemiology, which was first introduced by Eysenbach [6], it is now possible to extract knowledge from real time and inexpensive data from online sources. These sources reflect public health and answer the question of “what people are doing?” [7]. Conventionally, collection of such information was based on the data collected by public health agencies and personnel [8]. However, it is now possible to extract global health information using web-based data mining [9]. Google search trends, for instance, can be a useful tool for reflecting public interests/concerns during different periods [10–12]. Morsy et al. [13] considered the searches related to Zika virus to predict confirm cases in Brazil. During the COVID-19 outbreak, different studies have investigated the correlation of web-based data and cases of this virus. Kutlu et al. [14] investigated the correlation of dermatological diseases obtained by specific Google search trends with the COVID-19 outbreak. In addition, Google trends have been utilized to predict and monitor COVID-19 cases around the world [10,15–20]. Multiple studies analyzed the data related to the US to correlate the search trends and COVID-19 cases [21–26]. Although these studies consider the predictive ability of search trends on future confirmed cases, their search queries were limited to the symptoms and keywords related to the virus. For example, Ayyoubzadeh et al. [10] have investigated the concepts related to COVID-19, such as hand washing, hand sanitizer, and antiseptic, as the input features to predict the incidence of COVID-19 in Iran. Besides, such studies have only considered the correlation of search trends to the spread of the virus and not the causality analysis.

In this paper, we were interested to investigate the effect of the re-opening of in-store shopping on COVID-19 cases, rather than searches directly related to the virus. Therefore, we considered the causality effect and predictive ability of search terms related to bars and restaurants on the daily new cases of the US in different states/territories. We analyzed the states/territories with the highest and lowest daily new cases to investigate the effect of Google searches with higher confidence.

Along with the linear correlation analysis between search trends and COVID-19 cases, we have utilized the statistical causality methods to investigate the influential confidence of these methods on COVID-19 daily new cases.

Methods

Datasets

For our analysis, we obtained the daily cases of COVID-19 in the US using the COVID tracker project [27] which is publicly available. This project compiles the daily statistics, including the number of positive/negative tests, hospitalization, available ventilators, and the number of deaths from each US state and territory. For this study, we considered the data of approximately three months starting April 09, 2020, to July 07, 2020, which contains 5040 data points for 56 states/territories.

As infodemiology suggests, there are multiple sources of information for health informatics. Twitter and Google Trends are among the most popular sources that have been used to track outbreaks [18]. Although there are studies that leverage social media posts (e.g., Twitter) for time series forecasting (e.g., stock market [38]), in this research we selected Google Trends for the following reasons. First, our analysis needs access to location (i.e., state) information, however, location is not available by default in social media platforms. More precisely, social media users must opt-in to use location feature (e.g., tweeting with location) which limits the number of available data. Second, search engines (e.g., Google Trends) represent a wider scope of participants (e.g., age, ethnicity, socioeconomic status) and are more universal than social media platforms (e.g., Twitter) requiring memberships. In other words, Google Trends is a better proxy for the entire population in this case [37]. Lastly, social media is often used for idea and news sharing, whereas search engines are more informative with respect to searching for venues such as bars and restaurants.

For these reasons, we decided to use Google Trends to obtain the public interest in bars and restaurant categories with daily resolution. We followed the methodology presented in [28] to obtain the results. We used queries for each state/territory from April 09, 2020, to July 07, 2020, for 45 available states/territories in Google trends API. For restaurants and bars, we chose “dine-in restaurants that are open near me” and “bars near me” as our queries, respectively. Throughout the remainder of this paper, we refer to “bar and restaurant” searches as the Google trends data for the queries used to retrieve data related to dine-in restaurants and bars.

We did not narrow down the category, as the keywords were specific [28]. Google trend does not provide the number of queries per day. Instead, it provides a normalized number between 0 and 100, where 0 refers to “low volume of data for the query” while 100 refers to the “highest popularity for the term” [29]. To be consistent with Google trend values, we normalized the US daily new cases between 0 and 100 in our analysis.

Aggregating data from Google trends results and COVID-19 daily cases, and removing missing values, resulted in available data for 45 states/territories in the US. Although all the results for the entire states/territories are provided in supplemental material, we categorized our analysis to two different groups: First, top-10 states/territories with the highest number of daily new cases as of July 7th, 2020 which consist of Texas (TX), Florida (FL), California (CA), Arizona (AZ), Georgia (GA), Louisiana (LA), Tennessee (TN), North Carolina (NC), Washington (WA) and Pennsylvania (PA). Second, top-10 states/territories with the lowest number of daily new cases as of July 7th, 2020: Kansas (KS), Hawaii (HI), New Hampshire (NH), Maine (ME), West Virginia (WV), Rhode Island (RI), Connecticut (CT), Montana (MT), Nebraska (NE) and Delaware (DE).

All the data used in this study is publicly available and is therefore exempted from the requirements of the Federal Policy for the Protection of Human Subjects under Category 4.

Statistical Analysis

Correlation and Causation

To analyze the linear correlation of two time-series, the Pearson correlation has been utilized. The value of such a correlation ranges from -1 to 1, which shows a negative and positive correlation, respectively. Our analysis measured the Pearson correlation between the trends of search queries (i.e., restaurants and bars) and the daily new cases of COVID-19 in each state.

In addition, we used Granger’s causality [30] to model the influence of a time series’ past values on the new values of another time series. The cross-correlation (lag correlation) is not the appropriate method in this context since due to its symmetrical measurement, it does not explain the causation. However, Granger’s causality tests whether the past values of a time series X cause the current values of another time series Y. Hence, in this study, the null hypothesis is that X’s past values do not

affect Y's current values. If the *P*-value is less than the marginal value (.05), we can reject the null hypothesis. In our analysis, we reported *P*-values for each aforementioned search query's influence on the daily new cases. One of the main assumptions of modeling the influence of time series on each other is their stationarity. To test such a characteristic, we used the Augmented Dickey-Fuller (ADF) test [31] as our unit root test (Table 10 in the appendix). This test determines the effect of a trend in the creation of the time series. In other words, it determines how strongly a trend defines a time series. The alternative hypothesis in the ADF test is the stationarity of the time series.

In this study, since the time series were not stationary, we applied first differencing on search trends and second differencing on daily new cases to make all of the three series stationary. For statistical analysis, we used the Python Statsmodel package [32].

Vector Autoregression

In our study, we leveraged the fact that search trends might impact the daily new cases in the future; hence a Vector Autoregression (VAR) [33] model for each region was fitted to the data. A VAR model takes into account the influence of the past values of time series X and Y on current values of time series Y with a given lag order. Lag order with the lowest Akaike's Information Criterion (AIC) was picked in this study. Since symptoms may appear within 2-14 days after exposure to the COVID-19 virus [34], a maximum of 14 lags was used. The equation for the VAR model with two lags is summarized below:

$$Y_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_1^{'} Y_{t-1} + \beta_2^{'} Y_{t-2} + \epsilon_t \quad (1)$$

In equation 1, Y_t represents the value of time series Y at time t, which consists of a combination of previous lag values from Y and X with different weights $\beta, \beta^{'}$ and random white noise, ϵ_t. In other words, this equation models the importance of past values of the considering time series, as well as a secondary time series, for the estimation of current value. We fitted a VAR model with different lag orders to perform Granger's causality test. Although the VAR model was used to compute the Granger's causality, we did not use such model for the prediction task. Instead, we utilized a deep learning architecture for our prediction task.

Long Short-Term Memory

Long Short-Term Memory (LSTM) [35] models are a type of recurrent neural network useful for time series prediction. These models capture the long-term effect of time series as well as their most recent values. In this study, we utilized LSTMs to predict the daily new cases using two sets of features: 1) the historical values of the new cases time series and 2) using additional information from searching query time series. We used 70% of the data for training, and the rest were used for evaluation of the model. Root mean square error (RMSE) was selected as the performance metric. RMSE can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum \left(Y_{predict} - Y_{actual} \right)^2} \quad (2)$$

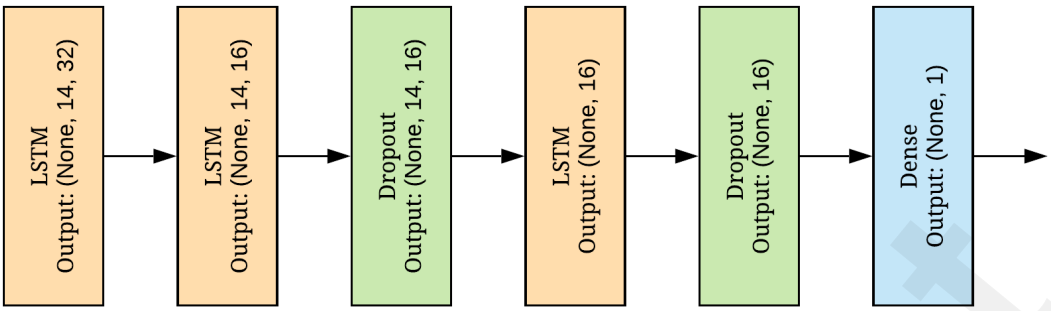
In equation 2, N is the number of samples, $Y_{predict}$ is the predicted value, and Y_{actual} is the actual value of the time series.

We calculated RMSE for three models: 1) the baseline model which uses only the past values of new cases time series for the prediction, 2) the model that uses the past values of restaurant searches along with the past values of new cases time series, 3) and, finally, the model that combines the information from the daily cases and bar searches time series.

The architecture of the used model is illustrated in Figure 1. It consists of three LSTM layers along with dropout layers, and a fully connected layer at the end. Dropout layers were utilized to avoid

overfitting, which is a typical problem in Machine Learning tasks. To train such a model, we used the TensorFlow package of Python.

Figure 1. The proposed model architecture.

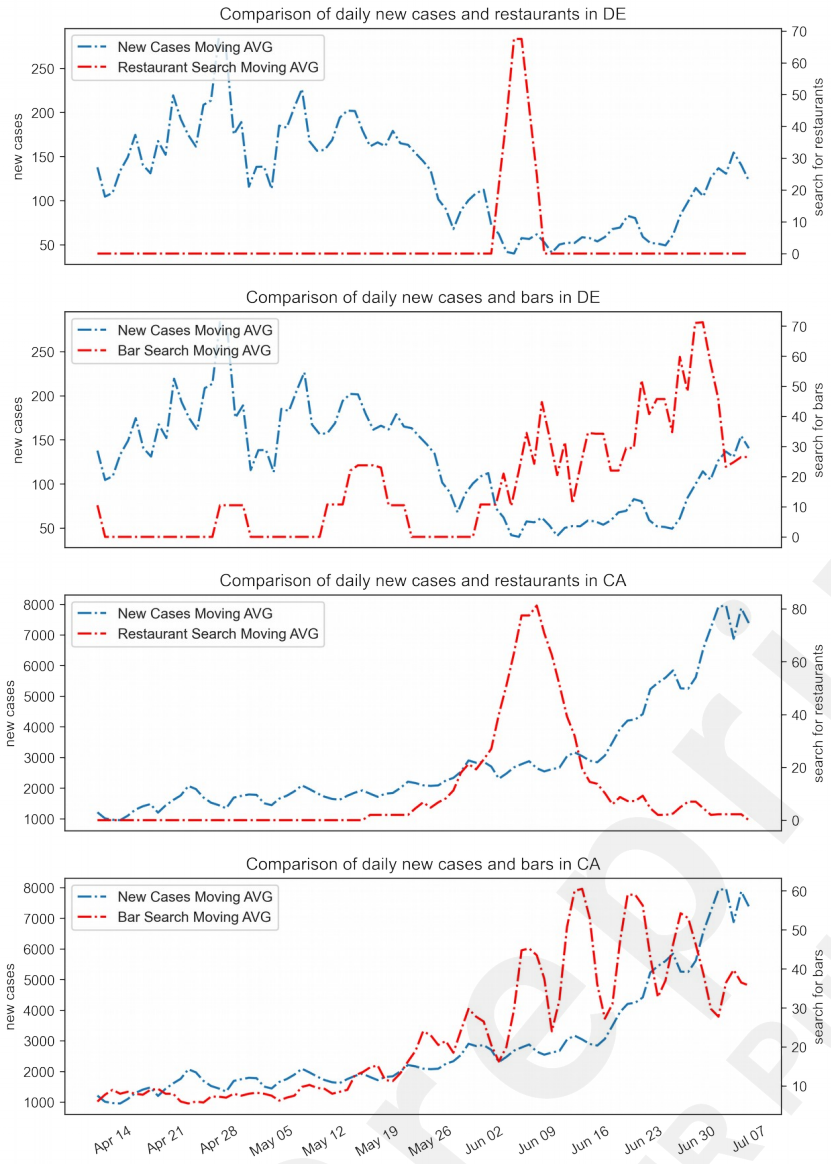


Results

Observation

Investigation of daily new cases and historical trends in search queries related to bars and restaurants showed correlation in some of the states/territories in the US. For states/territories such as CA, there was a steep rise of restaurant searches peaking on June 7th. The daily new cases have a drastic increase within two weeks of such a peak. Considering the bar searches in CA, the plot shows an increasing trend with peak value on June 13th. However, in DE, the daily new cases are not profoundly affected by such search trends (Figure 2).

Figure 2. Comparison between the effect of California (CA) and Delaware (DE) restaurant and bar search trends on daily cases of COVID-19 during April 09, 2020, to July 07, 2020.



Granger's Causality Test

In this section we provided the results of top-10 states/territories in the US with the highest and lowest daily new cases as of July 7th, 2020.

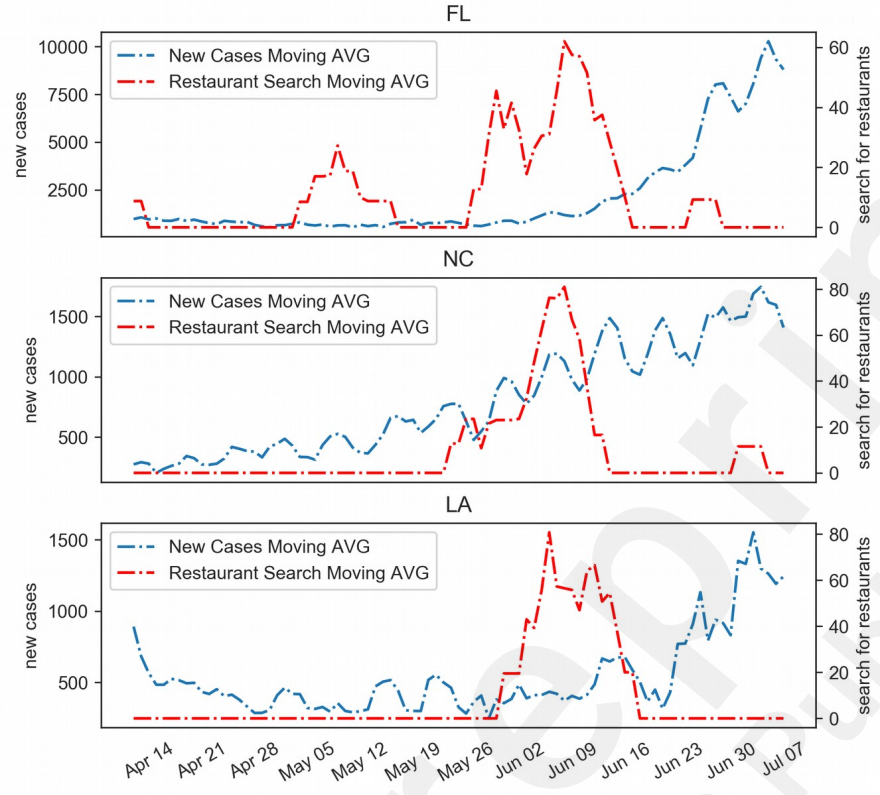
Table 1. Granger's causality test (*P*-values) on daily new cases of COVID-19 for top-10 states/territories with the most daily new cases in the US during April 09, 2020, to July 07, 2020.

Causing -> Caused	TX	FL	CA	AZ	GA	LA	TN	NC	WA	PA
Restaurant search -> New cases										
	.108	.35	.004	.003	.30	<.001	.091	.53	<.001	.108
Bar search -> New Cases										
	.019	.159	<.001	.042	.001	<.001	.075	.197	.016	.013

California has small *P*-values (significant effect of the search queries); hence, they can be used to predict daily new cases. Florida and North Carolina are two examples of states that the effect of restaurants is rejected with the Granger's Causality Test; however, Louisiana is affected by restaurant

searches (Table 1). Figure 3 illustrates the moving average of daily new cases and restaurant search trends for these three states. The high *P*-value for Florida is because of the first peak in the restaurant search, which did not change the daily new cases trends. North Carolina has an overall increasing trend, causing the effect of the search to be marginal. However, Louisiana is influenced by the sudden changes in restaurant search trends, which have affected the daily new cases (Figure 3).

Figure 3. Comparison of restaurant search effect on daily new cases of COVID-19 in Florida (FL), North Carolina (NC), and Louisiana (LA) during April 09, 2020, to July 07, 2020.



Similarly, Table 2 summarizes the *P*-values for Granger's causality test for the second group (i.e., top-10 states/territories with the lowest daily new cases). Most of the *P*-values for these states/territories is not significant.

Table 2. Granger's causality test (*P*-values) on daily new cases of COVID-19 for top-10 states/territories with the lowest daily new cases in the US during April 09, 2020, to July 07, 2020.

Causing -> Caused	KS	HI	NH	ME	WV	RI	CT	MT	NE	DE
Restaurant search -> New cases										
	.99	<.001	.88	.077	.081	.54	.99	<.001	.99	1.0
Bar search -> New Cases										
	.014	.001	.50	.11	.45	.28	.008	.073	.083	<.001

Pearson Correlation

In this section we have provided the Pearson correlation results. Tables 3 and 4 summarize these correlations with corresponding *P*-values for each group. Based on these two tables, the linear correlation between the search trends related to bars/restaurants and daily new cases in

states/territories with a higher number of daily cases is more substantial, on average, compared to states/territories with lower daily cases.

Table 3. Pearson correlation between search trends and daily new cases of COVID-19 for top-10 states/territories with the most daily new cases in the US during April 09, 2020, to July 07, 2020.

Correlation (r [*P*-value])	TX	FL	CA	AZ	GA	LA	TN	NC	WA	PA
Restaurant vs. New cases										
	-0.17 [.111]	-0.19 [.072]	0.0 [.96]	-0.11 [.30]	-0.2 [.065]	-0.13 [.23]	-0.18 [.081]	0.17 [.10]	-0.11 [.29]	-0.23 [.027]
Bar vs. New Cases										
	0.11 [.28]	0.41 [<.001]	0.47 [<.001]	0.31 [.003]	0.31 [.003]	0.12 [.26]	0.39 [<.001]	0.73 [<.001]	0.13 [.20]	-0.52 [<.001]

Table 4. Pearson correlation between search trends and daily new cases of COVID-19 for top-10 states/territories with the lowest daily new cases in the US during April 09, 2020, to July 07, 2020.

Correlation (r [*P*-value])	KS	HI	NH	ME	WV	RI	CT	MT	NE	DE
Restaurant vs. New cases										
	-0.05 [.62]	-0.08 [.43]	-0.08 [.45]	-0.08 [.42]	0.09 [.35]	-0.08 [.42]	-0.06 [.55]	-0.01 [.85]	-0.05 [.61]	-0.17 [.097]
Bar vs. New Cases										
	-0.20 [.057]	0.22 [.030]	-0.11 [.27]	0.13 [.21]	0.11 [.28]	-0.61 [<.001]	-0.22 [.035]	0.19 [.070]	0.007 [.94]	-0.18 [.087]

New Cases Prediction

The prediction results of daily new cases using our deep neural network architecture is provided in this section. The RMSE scores for test data for top-10 highest and lowest daily new cases are summarized in Tables 5 and 6 for each model.

Table 5. RMSE scores for new cases time series of COVID-19 (Baseline), Baseline + Restaurants time series, and Baseline + Bars time series for top-10 states/territories with the most daily new cases in the US during April 09, 2020, to July 07, 2020.

Model	TX	FL	CA	AZ	GA	LA	TN	NC	WA	PA
Baseline										
	18.00	48.21	24.19	31.35	29.90	39.84	35.88	19.74	26.44	18.70
Baseline + Restaurants										
	32.44	43.84	21.86	45.32	33.46	29.36	32.51	22.91	23.92	18.10

Baseline + bars										
	44.50	32.55	19.89	26.20	36.39	43.51	38.09	26.68	22.75	24.68

The states/territories with a significant causality effect, the RMSE improves on average. CA is an example of such an improvement (Table 5). Similarly, Figure 4 illustrates the prediction performance with and without considering restaurants search trends. The predicted values are closer to the actual values when taking into consideration the effect of restaurant searches in the prediction model.

Figure 4. Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for California (CA) during April 09, 2020, to July 07, 2020.

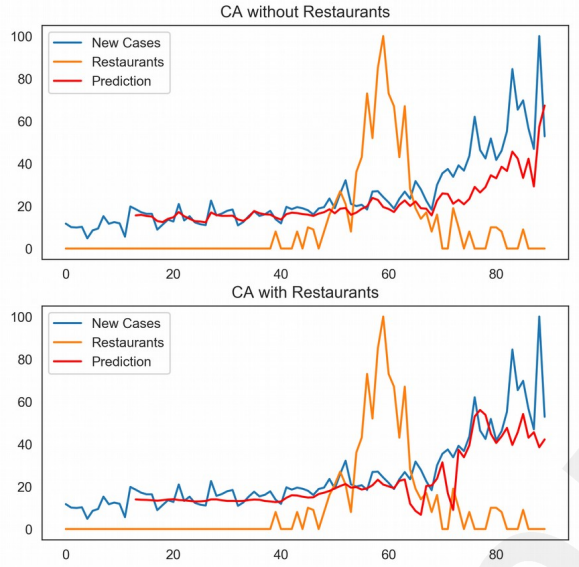


Table 6. RMSE scores for new cases time series of COVID-19 (Baseline), Baseline + Restaurants time series, and Baseline + Bars time series for top-10 states/territories with the lowest daily new cases in the US during April 09, 2020, to July 07, 2020.

Model	KS	HI	NH	ME	WV	RI	CT	MT	NE	DE
Baseline										
	28.41	51.49	12.09	20.92	26.18	5.37	3.47	29.58	5.49	20.73
Baseline + Restaurants										
	25.56	43.64	8.10	14.57	22.55	8.88	3.91	43.34	8.22	20.42
Baseline + bars										
	34.43	49.01	15.30	21.96	24.15	6.01	4.68	43.27	8.67	12.81

For some states, although there is no causality effect for the restaurants, the value of RMSE improves. On the other hand, for states like Montana (MT), which Granger's Causality Test shows a significant effect, the RMSE has been increased (Table 6). By investigating the time series for these two states (Figures 5 and 6), we can interpret such inconsistencies for two reasons. First, for states such as Kansas (KS), the improved value is because of the fluctuation in the new cases time series, making the prediction unreliable. Second, as Figures 5 and 6 show, the impulses in restaurant searches for KS and MT are point impulses. These unit jumps cannot improve the prediction of the

time series significantly, although they appear in causality tests.

Figure 5. Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for Kansas (KS) during April 09, 2020, to July 07, 2020.

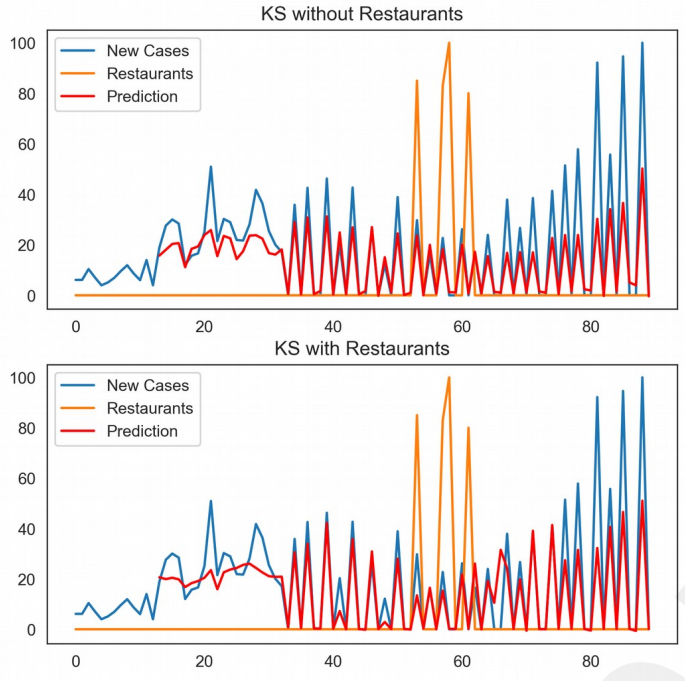
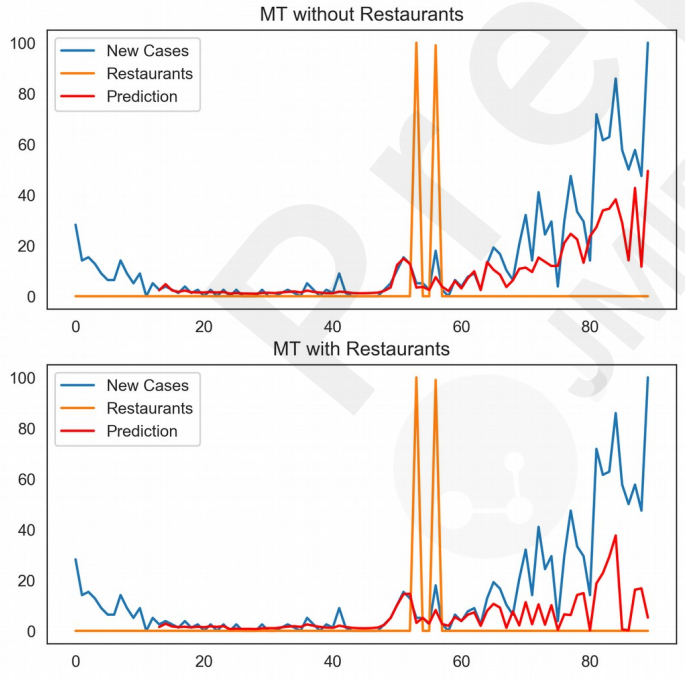


Figure 6. Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for Montana (MT) during April 09, 2020, to July 07, 2020.



Discussion

Principal Results

To the best of our knowledge, this study is the first analysis that considers the predictive ability of Google search trends related to the dine-in restaurants and bars, on daily new cases of COVID-19 in

the US. Our main findings show that in states/territories with higher numbers of daily cases, the historical trends in search queries related to bars and restaurants (queries related to dine-in places), which happened primarily after re-opening, significantly correlates with the daily new cases, on average. This study uses statistical methods to validate such an effect on daily new cases. One potential reason could be the lower population as it is reflected in the number of daily new cases. The other reason can be the high number of new daily cases, California for instance, at the time of re-opening restaurants and bars (+2000).

Granger's causality test shows that in some states/territories, the effect of restaurants on daily new cases is significant. California is an example of such states. On May 18th, the governor of California announced the easing of criteria for counties to re-open faster than the state, and on May 25th he announced plans for the re-opening of in-store shopping [36]. Consequently, there was an increase in restaurant searches, and the peak of the searches happened on June 7th. The daily new cases drastically increased within two weeks of the escalation in dine-in restaurant searches.

Similarly, such a trend in bar searches happened in California. Regardless of the seasonal effect of time series, which shows a higher number of searches for bars during weekends, the average trend in bar searches increased. However, North Carolina is not influenced by restaurant searches. The reason is that this state has an increasing average trend regardless of the other time series. Therefore, the *P*-value for Granger's causality is high (.53). In summary, Granger's causality shows significant results for states/territories with higher daily new cases, on average.

This study suggests that the effect of restaurant and bar searches is higher in the states/territories with higher daily new cases compared to the states/territories that have a lower number of positive cases reporting every day. On average, in the states/territories with a higher number of daily new cases, more significant Granger's casualties and higher values of Pearson correlation support this fact. Besides, by taking restaurants and bar searches into account, we can improve the underestimation of the prediction task. We used artificial intelligence models to improve the prediction results of new cases using additional information, namely Google trends. These Google trends for restaurants and bars can be useful depending on the time series structure.

Owing to infodemiology, capturing the real time information and public attitude could help decision makers to be prepared based on the feedback loop on public data and disease spread [7], and have a better estimation of the deadly disease like COVID-19 in each state to distribute healthcare-related utilities such as ventilators.

Limitations

There are several limitations to this study. We only used the specific search queries for each category. People use different search terms to find the information they are looking for. Moreover, we only considered the effect of restaurants and bars on daily cases. Further research could aim considering the effect of other public places such as gyms and adventure parks. The other limitation of our study is the limited number of data points for each region (88 samples on average). This limitation, which is the consequence of the daily report data structure affects the prediction results to a certain degree.

Conclusions

In conclusion, we investigated the causality effect and correlation of search queries related to dine-in restaurants and bars on daily new cases of COVID-19 in the US states/territories with high and low daily cases during April 09, 2020, to July 07, 2020. We showed that for most of the states/territories

with a high number of daily new cases, the effect of search queries on bars and restaurants is higher; hence, they can be used as additional information for prediction tasks.

Acknowledgements

No funding was received for this project.

Conflicts of Interest

None declared.

Abbreviations

ADF: Augmented Dickey-Fuller
AIC: Akaike's Information Criterion
COVID-19: Coronavirus disease
LSTM: Long Short-Term Memory
RMSE: Root Mean Square Error
VAR: Vector Autoregression

References

1. A comprehensive timeline of the coronavirus pandemic at 6 months, from China's first case to the present [Internet]. [cited 2020 Jul 21]. Available from: https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3
2. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, Tan K Sen, Wang DY, Yan Y. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- an update on the status. Mil Med Res. 2020. PMID:32169119
3. Worldometer. COVID coronavirus Outbreak [Internet]. 2020 [cited 2020 Dec 1]. Available from: https://www.worldometers.info/coronavirus/
4. Simakova J. Analysis of the relationship between oil and gold prices. Econ J 2012; PMID:3118295
5. Wang Q, Liu Y, Pan X. Atmosphere pollutants and mortality rate of respiratory diseases in Beijing. Sci Total Environ 2008; PMID:18061245
6. Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med. 2002. PMID:12517369
7. Eysenbach G. Infodemiology and infoveillance: Tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011. PMID:21521589
8. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A. Digital epidemiology. PLoS Comput Biol 2012; PMID:22844241
9. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 2008. PMID:18613747
10. 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 Heal Surveill 2020; [doi: 10.2196/18828]
11. MURDOCK J. COVID-19 Pandemic Can Now Be Tracked Through Google Searches [Internet]. 2020 [cited 2020 Dec 1]. Available from: https://www.newsweek.com/research-coronavirus-covid19-google-search-data-tracking-pandemic-1500444

12. Cuthbertson A. CORONAVIRUS TRACKED: COULD GOOGLE SEARCH TRENDS HELP PREDICT A RISE IN COVID-19 CASES? [Internet]. 2020 [cited 2020 Dec 1]. Available from: https://www.independent.co.uk/life-style/gadgets-and-tech/news/coronavirus-second-wave-us-google-trends-covid-19-symptoms-a9559371.html
13. Morsy S, Dang TN, Kamel MG, Zayan AH, Makram OM, Elhady M, Hirayama K, Huy NT. Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. Epidemiol Infect 2018; PMID:30056812
14. Kutlu Ö. Analysis of dermatologic conditions in Turkey and Italy by using Google Trends analysis in the era of the COVID-19 pandemic. Dermatol Ther 2020; PMID:32614116
15. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance. 2020. PMID:32183935
16. Effenberger M, Kronbichler A, Shin J Il, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis. Int J Infect Dis. 2020. PMID:32305520
17. Ciaffi J, Meliconi R, Landini MP, Ursini F. Google trends and COVID-19 in Italy: could we brace for impact? Intern Emerg Med. 2020. PMID:32451932
18. Mavragani A. Tracking COVID-19 in Europe: Infodemiology Approach. JMIR Public Heal Surveill 2020; [doi: 10.2196/18941]
19. Husnayain A, Fuad A, Su ECY. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. Int J Infect Dis 2020; PMID:32173572
20. Ortiz-Martínez Y, Garcia-Robledo JE, Vásquez-Castañeda DL, Bonilla-Aldana DK, Rodriguez-Morales AJ. Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia. Travel Med Infect Dis. 2020. PMID:32360323
21. Yuan X, Xu J, Hussain S, Wang H, Gao N, Zhang L. Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model. Explor Res Hypothesis Med 2020; [doi: 10.14218/erhm.2020.00023]
22. Hong Y-R, Lawrence J, Williams Jr D, Mainous III A. Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. JMIR Public Heal Surveill 2020; PMID:32250963
23. Walker A, Hopkins C, Surda P. The use of google trends to investigate the loss of smell related searches during COVID-19 outbreak. Int Forum Allergy Rhinol 2020; PMID:32279437
24. Husain I, Briggs B, Lefebvre C, Cline DM, Stopyra JP, O'Brien MC, Vaithi R, Gilmore S, Countryman C. How COVID-19 public interest in the United States fluctuated: A Google Trends Analysis (Preprint). JMIR Public Heal Surveill 2020; [doi: 10.2196/19969]
25. Jacobson NC, Lekkas D, Price G, Heinz M V, Song M, O'Malley AJ, Barr PJ. Flattening the Mental Health Curve: COVID-19 Stay-at-Home Orders Are Associated With Alterations in Mental Health Search Behavior in the United States. JMIR Ment Heal 2020; PMID:32459186
26. Rajan A, Sharaf R, Brown RS, Sharaiha RZ, Lebwohl B, Mahadev S. Association of Search Query Interest in Gastrointestinal Symptoms With COVID-19 Diagnosis in the United States: Infodemiology Study. JMIR Public Heal Surveill 2020;6(3). [doi: 10.2196/19354]
27. The COVID Tracking Project [Internet]. [cited 2020 Jul 21]. Available from: https://covidtracking.com/
28. Mavragani A, Ochoa G. Google trends in infodemiology and infoveillance: Methodology framework. J Med Internet Res 2019; [doi: 10.2196/13439]

29. FAQ about Google Trends data [Internet]. [cited 2020 Jul 21]. Available from: https://support.google.com/trends/answer/4365533?hl=en
30. Granger CWJ. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 1969; [doi: 10.2307/1912791]
31. Chatfield C, Fuller WA. Introduction to Statistical Time Series. J R Stat Soc Ser A 1977; [doi: 10.2307/2344931]
32. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Proc 9th Python Sci Conf 2010;
33. Johansen S. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Likelihood-Based Inference Cointegrated Vector Autoregressive Model. 2003. [doi: 10.1093/0198774508.001.0001]
34. Symptoms of Coronavirus [Internet]. [cited 2020 Jul 20]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
35. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput 1997; [doi: 10.1162/neco.1997.9.8.1735]
36. Timeline of COVID-19 policies, cases, and deaths in your state - Johns Hopkins Coronavirus Resource Center [Internet]. [cited 2020 Jul 20]. Available from: https://coronavirus.jhu.edu/data/state-timeline/new-confirmed-cases/california/53
37. Mavragani A. Infodemiology and infoveillance: scoping review. J Med Internet Res 2020;22(4). [doi:10.2196/16206]
38. Guo X, Li J. A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. IEEE SNAMS 2019;472-477. [doi: 10.1109/SNAMS.2019.8931720]

Appendix

Table 7. Granger's causality test (*P*-values) on daily new cases of COVID-19 for the rest of the states/territories in the US during April 09, 2020, to July 07, 2020.

State/Territory	Restaurant search -> New cases	Bar search -> New Cases
SC		
	.078	.002
MS		
	.085	<.001
OH		
	.99	<.001
AL		
	.003	.06
NV		
	.11	<.001
OK		
	<.001	<.001
MO		
	.12	.16
VA		
	.89	.83
MI		
	.77	.16
NY		
	1.0	.75
IL		
	.38	.91
UT		
	<.001	.17
MN		
	.002	.066
WI		
	<.001	<.001
MD		
	.82	.69
IA		
	.075	.001
KY		
	.18	.47
ID		
	.55	.013
IN		
	.086	.31
NJ		
	.96	.46
AR		
	<.001	.11
NM		
	<.001	.33
OR		

	<.001	.90
MA		
	<.001	.44
CO		
	.80	0.99

Table 8. Pearson correlation between search trends and daily new cases of COVID-19 for the rest of the states/territories in the US during April 09, 2020, to July 07, 2020.

State/Territory	Restaurant vs. New cases (r [P-value])	Bar vs. New cases (r [P-value])
SC		
	-0.06 [.52]	0.48 [<.001]
MS		
	-0.12 [.23]	0.16 [.12]
OH		
	-0.21 [.044]	.04 [.69]
AL		
	0.01 [.92]	.41[<.001]
NV		
	-0.07 [.45]	0.42 [<.001]
OK		
	-0.11 [.27]	0.20 [.069]
MO		
	-0.08 [.45]	0.11 [.29]
VA		
	0.16 [.109]	-0.13 [.20]
MI		
	-0.44 [<.001]	-0.54 [<.001]
NY		
	-0.26 [.012]	-0.47 [<.001]
IL		
	-0.19 [.067]	-0.57 [<.001]
UT		
	0.03 [.71]	0.37 [.002]
MN		
	-0.19 [.07]	-0.26 [.010]
WI		
	-0.01 [0.89]	0.20 [.047]
MD		
	-0.06 [.51]	-0.53 [<.001]
IA		
	-0.14 [.16]	0.07 [.48]
KY		
	-0.10 [.30]	0.03 [.77]
ID		
	0.0 [.99]	.22 [.03]

IN		
	-0.10 [.34]	-0.21 [.043]
NJ		
	-0.12 [.23]	-0.47 [<.001]
AR		
	0.12 [.23]	0.25 [.015]
NM		
	0.04 [.65]	0.17 [.10]
OR		
	-0.04 [.70]	0.28 [.006]
MA		
	-0.06 [.55]	-0.54 [<.001]
CO		
	-0.15 [.15]	-0.26 [.011]

Table 9. RMSE scores for new cases time series of COVID-19 (Baseline), Baseline + Restaurants time series, and Baseline + Bars time series for the rest of states/territories in the US during April 09, 2020, to July 07, 2020.

State/Territory	Baseline	Baseline + Restaurants	Baseline + Bars
SC			
	27.87	55.66	55.15
MS			
	27.02	24.90	27.43
OH			
	22.61	23.11	27.66
AL			
	28.26	46.12	41.79
NV			
	26.01	35.64	32.77
OK			
	19.14	36.80	38.02
MO			
	34.69	31.49	35.08
VA			
	8.55	10.75	54.49
MI			
	15.00	15.16	18.88
NY			
	3.99	2.07	19.90
IL			
	6.09	4.65	25.05
UT			
	25.29	29.61	49.20
MN			
	19.08	17.48	28.64
WI			
	38.42	28.32	32.60

MD			
	5.12	12.08	13.41
IA			
	27.08	21.46	19.29
KY			
	19.65	18.25	23.02
ID			
	50.33	44.76	52.78
IN			
	18.31	22.79	12.76
NJ			
	6.09	13.53	21.76
AR			
	26.70	45.51	38.47
NM			
	26.02	26.07	28.85
OR			
	38.59	24.31	24.92
MA			
	4.60	4.30	11.57
CO			
	11.01	7.03	7.39

Table 10. *P*-values of ADF statistics for stationarity test for the states/territories in the US.

State/Territory	Daily New Cases	Restaurants Trend	Bars Trend
TX			
	.039	<.001	.011
FL			
	<.001	<.001	<.001
CA			
	<.001	<.001	.002
AZ			
	<.001	<.001	<.001
GA			
	<.001	<.001	<.001
LA			
	<.001	.015	<.001
TN			
	<.001	<.001	.001
NC			
	<.001	<.001	.16
WA			
	<.001	<.001	<.001
PA			
	<.001	<.001	<.001

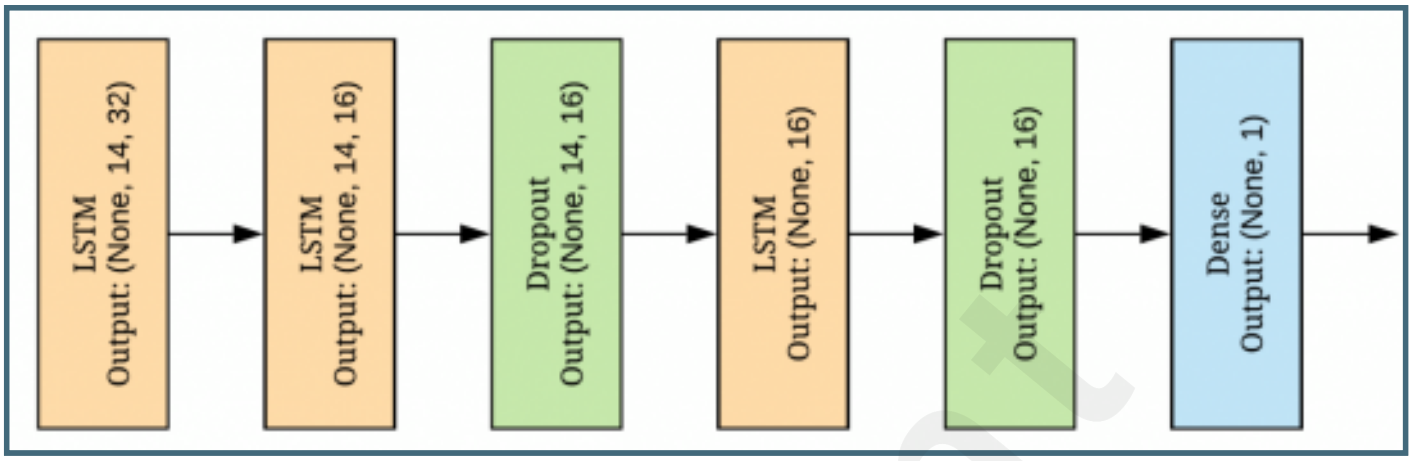
SC			
	.25	<.001	<.001
MS			
	<.001	<.001	<.001
OH			
	<.001	<.001	<.001
AL			
	<.001	<.001	<.001
NV			
	<.001	<.001	<.001
OK			
	<.001	<.001	<.001
MO			
	<.001	<.001	<.001
VA			
	<.001	<.001	<.001
MI			
	<.001	<.001	.002
NY			
	<.001	<.001	.080
IL			
	<.001	<.001	<.001
UT			
	<.001	<.001	<.001
MN			
	<.001	<.001	<.001
WI			
	<.001	<.001	<.001
MD			
	<.001	<.001	<.001
IA			
	<.001	<.001	<.001
KY			
	<.001	<.001	<.001
ID			
	<.001	<.001	<.001
IN			
	<.001	<.001	<.001
NJ			
	<.001	<.001	.071
AR			
	<.001	<.001	<.001
NM			
	<.001	<.001	<.001
OR			
	<.001	<.001	<.001
MA			
	<.001	<.001	<.001

CO			
	<.001	<.001	<.001
DE			
	<.001	<.001	<.001
NE			
	<.001	<.001	<.001
MT			
	<.001	<.001	<.001
CT			
	<.001	<.001	<.001
RI			
	<.001	<.001	<.001
WV			
	<.001	<.001	<.001
NH			
	<.001	<.001	<.001
ME			
	<.001	<.001	<.001
HI			
	<.001	<.001	<.001
KS			
	<.001	<.001	<.001

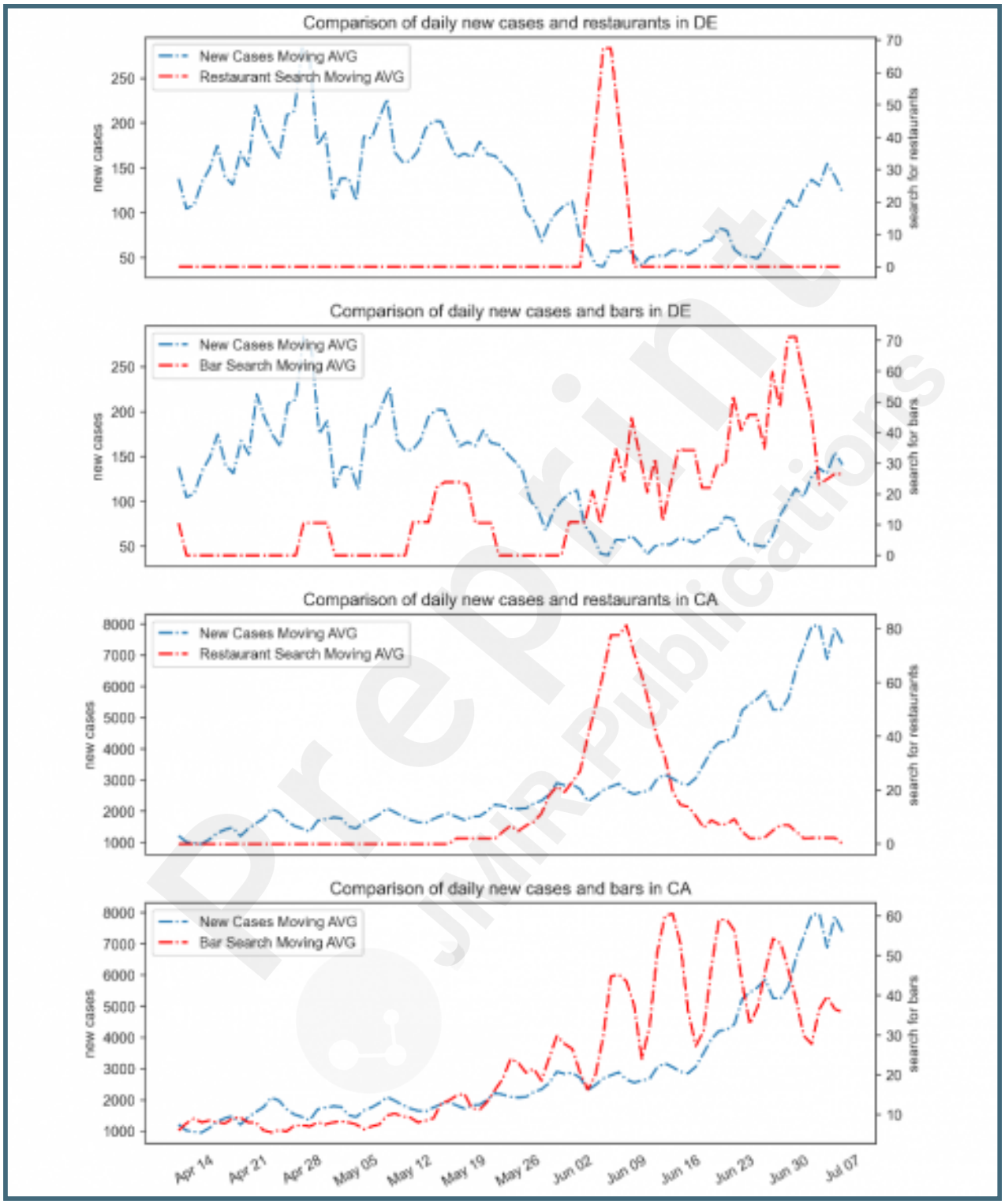
Supplementary Files

Figures

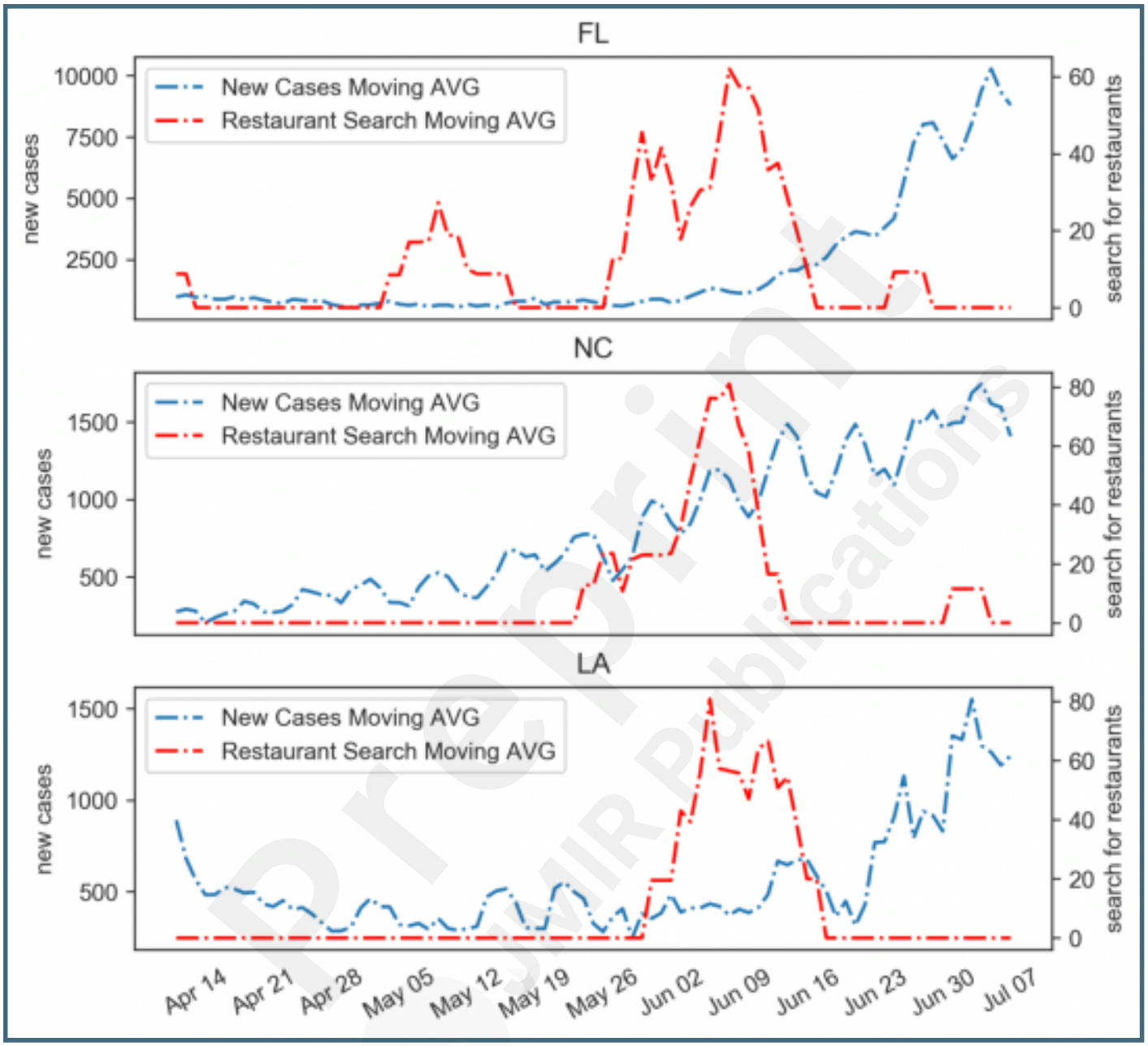
The proposed model architecture.



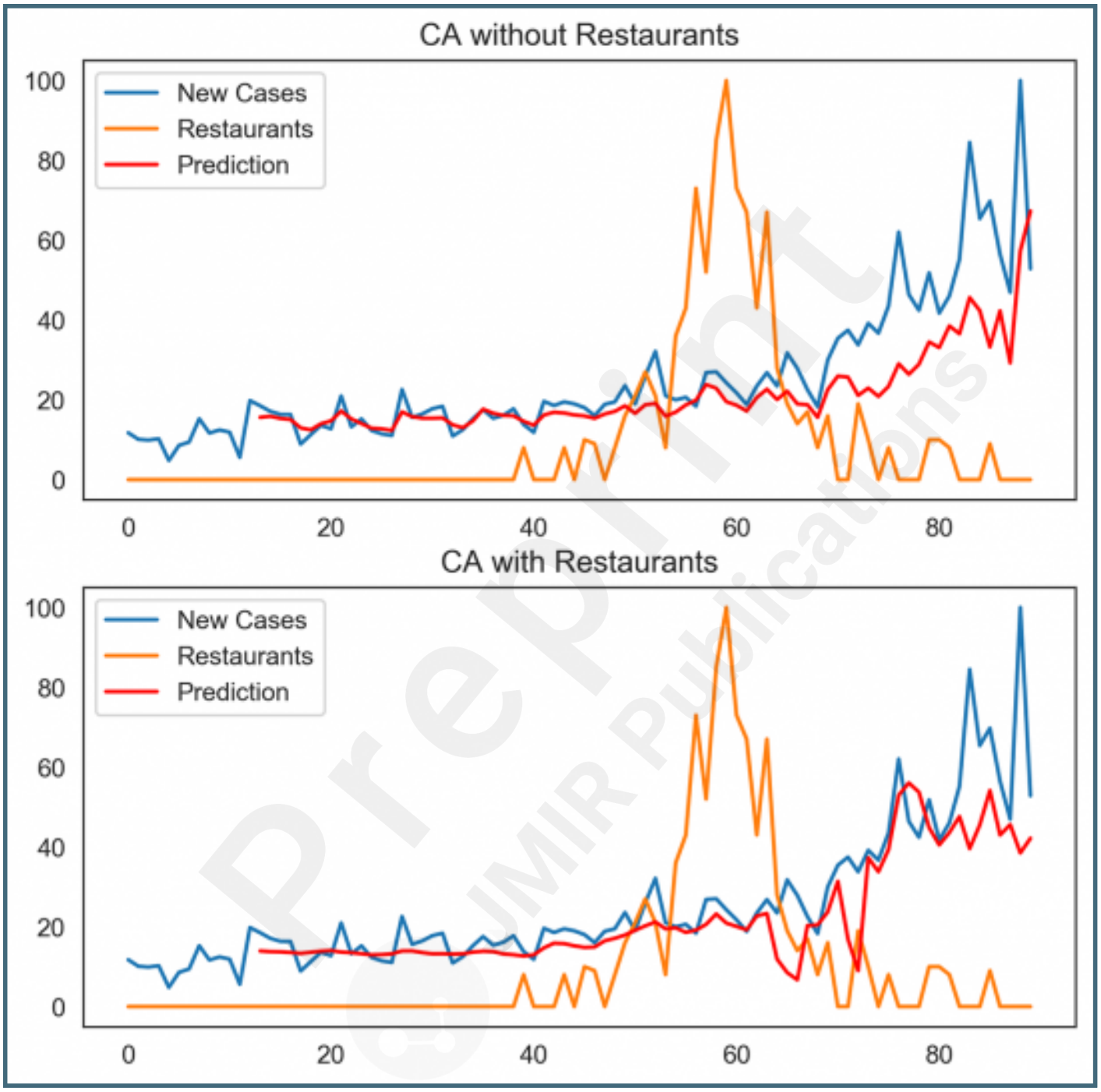
Comparison between the effect of California (CA) and Delaware (DE) restaurant and bar search trends on daily cases of COVID-19 during April 09, 2020, to July 07, 2020.



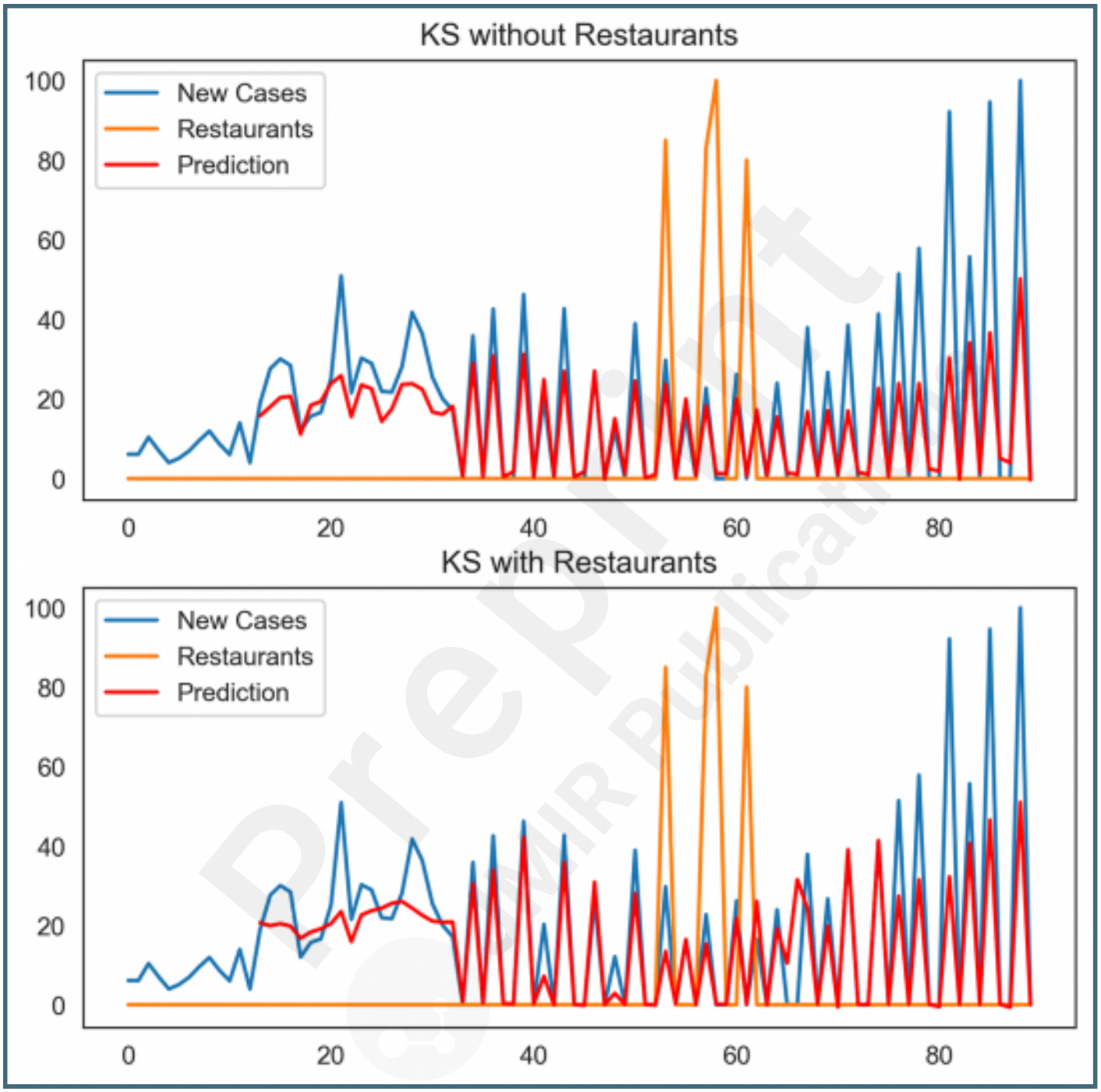
Comparison of restaurant search effect on daily new cases of COVID-19 in Florida (FL), North Carolina (NC), and Louisiana (LA) during April 09, 2020, to July 07, 2020.



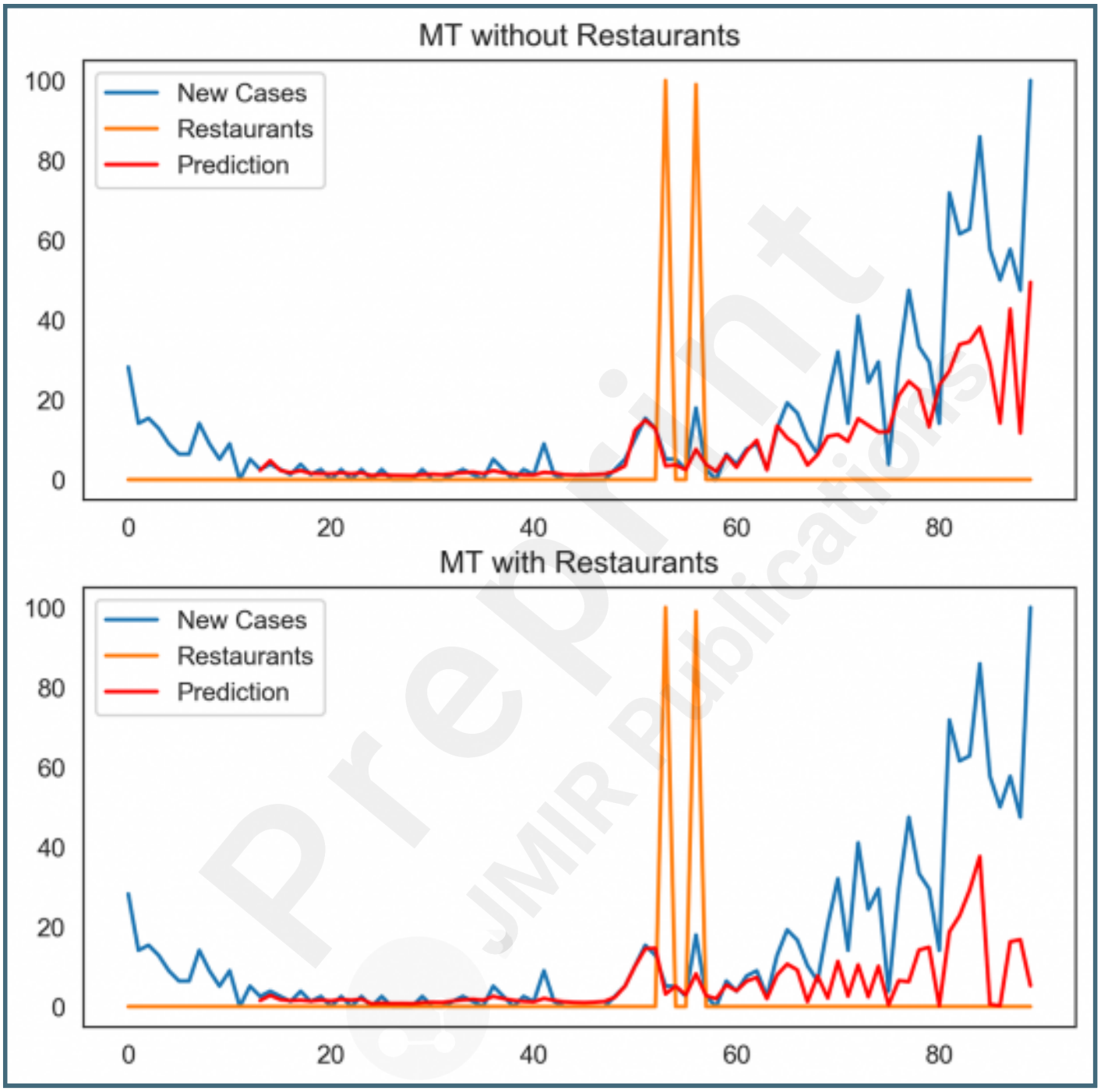
Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for California (CA) during April 09, 2020, to July 07, 2020.



Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for Kansas (KS) during April 09, 2020, to July 07, 2020.



Prediction values for daily new cases of COVID-19 with/without using restaurant search trends for Montana (MT) during April 09, 2020, to July 07, 2020.



Multimedia Appendixes

Tracked version of the manuscript for reviewers convenience.
URL: http://asset.jmir.pub/assets/9d0f85f8c3beebf26ffda21d89f043ef.docx

Granger's causality test (P-values) on daily new cases of COVID-19 for the rest of the states/territories in the US during April 09, 2020, to July 07, 2020.
URL: http://asset.jmir.pub/assets/c8de5a18bd937939c3be495b65bb70f1.docx

Pearson correlation between search trends and daily new cases of COVID-19 for the rest of the states/territories in the US during April 09, 2020, to July 07, 2020.
URL: http://asset.jmir.pub/assets/e841a0fba8a33721f2de6b1a0af55305.docx

RMSE scores for new cases time series of COVID-19 (Baseline), Baseline + Restaurants time series, and Baseline + Bars time series for the rest of states/territories in the US during April 09, 2020, to July 07, 2020.
URL: http://asset.jmir.pub/assets/7c0e0cbe610d3b7871045f08d88ef29e.docx

P-values of ADF statistics for stationarity test for the states/territories in the US.
URL: http://asset.jmir.pub/assets/3c11afbc140685d48bbf633c829eafd3.docx