Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Infectious Diseases 109 (2021) 269–278

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

High variability in model performance of Google relative search volumes in spatially clustered COVID-19 areas of the USA

Atina Husnayain[a,b], Ting-Wu Chuang[c], Anis Fuad[b], Emily Chia-Yu Su[a,d,*]

[a] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[b] Department of Biostatistics, Epidemiology and Population Health, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia
[c] Department of Molecular Parasitology and Tropical Diseases, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[d] Clinical Big Data Research Centre, Taipei Medical University Hospital, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 27 March 2021
Revised 22 June 2021
Accepted 11 July 2021

Keywords:
COVID-19
United States
Spatial analysis
Google Trends
Predictability performance
Infodemiology

ABSTRACT

Objective: Incorporating spatial analyses and online health information queries may be beneficial in understanding the role of Google relative search volume (RSV) data as a secondary public health surveillance tool during pandemics. This study identified coronavirus disease 2019 (COVID-19) clustering and defined the predictability performance of Google RSV models in clustered and non-clustered areas of the USA.

Methods: Getis-Ord General and local G statistics were used to identify monthly clustering patterns. Monthly country- and state-level correlations between new daily COVID-19 cases and Google RSVs were assessed using Spearman's rank correlation coefficients and Poisson regression models for January–December 2020.

Results: Huge clusters involving multiple states were found, which resulted from various control measures in each state. This demonstrates the importance of state-to-state coordination in implementing control measures to tackle the spread of outbreaks. Variability in Google RSV model performance was found among states and time periods, possibly suggesting the need to use different frameworks for Google RSV data in each state. Moreover, the sign of correlation can be utilized to understand public responses to control and preventive measures, as well as in communicating risk.

Conclusion: COVID-19 Google RSV model accuracy in the USA may be influenced by COVID-19 transmission dynamics, policy-driven community awareness and past outbreak experiences.

Introduction

Spatial spread is one of the most important aspects in understanding disease epidemics (Franch-Pardo et al., 2020), including the coronavirus disease 2019 (COVID-19) pandemic. During the outbreak, multiple studies have discussed COVID-19 spatial patterns in the USA using both state- (Cordes and Castro, 2020; Maroko et al., 2020; Ramírez and Lee, 2020) and county-level analyses (CDC COVID-19 Response Team, 2020; Dasgupta et al., 2020; Desjardins et al., 2020; Mollalo et al., 2020; Oster et al., 2020a,b; Snyder and Parks, 2020; Wang et al., 2020; Andersen et al., 2021). Most of these studies dealt with cluster detection analyses, a necessary approach in allocating resources, implementing strict control measures, and evaluating currently implemented policies (Desjardins et al., 2020). Disease mapping also enables targeted public health responses (Oster et al., 2020b) through assessment of the distribution of high-risk areas and their progression throughout the outbreak period (Desjardins et al., 2020).

Countrywide analyses have described COVID-19 clusters in the USA (CDC COVID-19 Response Team, 2020; Dasgupta et al., 2020; Desjardins et al., 2020; Oster et al., 2020a,b), vulnerability assessments (Snyder and Parks, 2020; Wang et al., 2020) and spatial modelling which employed various explanatory variables (Mollalo et al., 2020; Andersen et al., 2021) for the first 3–6 months of the outbreak. State-level studies also characterized emerging clusters (Cordes and Castro, 2020; Maroko et al., 2020; Ramírez and Lee, 2020). However, few studies have analysed a

* Corresponding author. Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 172-1 Keelung Rd., Sec. 2, Taipei 106, Taiwan. Tel.: +886-2-66382736 ext. 1515.
E-mail address: emilysu@tmu.edu.tw (E.C.-Y. Su).

https://doi.org/10.1016/j.ijid.2021.07.031

Table 1
Dataset description.

Dataset	Data description	Data unit	Source	Utilization
Case data	Cumulative daily cases New daily cases	County-level data State-level data	https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series https://covidtracking.com/data	Hot spot analysis State-level correlation and prediction analysis
Spatial data	Spatial features and population numbers (of 48 contiguous states and the District of Columbia)	County-level data	https://hub.arcgis.com/datasets/48f9af87daa241c4b267c5931ad3b226_0	Spatial analysis and visualization
Google RSV data	Google RSV data ranged from 0~100	Country- and state-level data	https://trends.google.com/trends	Country- and state-level correlations, state-level prediction analysis
Mobility data	Changes in time spent in six categorized places (retail and recreation, grocery and pharmacy, parks, transit stations, workplaces and residential) compared with baseline days (median value from 3 January to 6 February 2020)	State-level data	https://www.google.com/covid19/mobility/	State-level prediction analysis

RSC, relative search volume.

year of COVID-19 spatiotemporal patterns along with temporal predictability performances of Google relative search volume (RSV) models in clustered and non-clustered areas.

Google RSVs are emerging digital data that are being used as a secondary public health surveillance tool during the COVID-19 pandemic. These data are collected during information-seeking activities on Google search engines that are normalized during a specified period (Google, 2020). These online search data potentially depict patterns of information-seeking behaviours that represent the public's concerns, awareness or restlessness (Ayyoubzadeh et al., 2020; Husnayain et al., 2020a). This approach was part of an infodemiological study that examined the determinants and distributions of health information for public health purposes (Eysenbach, 2006). It may capture wider population events than conventional surveillance systems (Milinovich et al., 2014), as people who are ill may not contact local healthcare facilities, but they may still search for online health information.

In the case of COVID-19, various studies in the early phase of the outbreak suggested that Google searches peaked earlier than newly confirmed cases (Effenberger et al., 2020; Strzelecki, 2020) and correlated well with the rise of COVID-19-related data (Husnayain et al., 2020a,b; Li et al., 2020; Ortiz-Martínez et al., 2020). Similar results were also reported by several studies in the USA (Bento et al., 2020; Panuganti et al., 2020; Yuan et al., 2020). Certain studies also assessed the predictability performance of Google RSVs at national and regional levels, which resulted in high correlations (the highest correlation coefficients were 0.71 and 0.88) (Kurian et al., 2020; Mavragani and Gkillas, 2020). Moreover, a high accuracy of Google search models was also found in an earlier state-level analysis (Cousins et al., 2020). However, all of these studies were undertaken in the first 3 months of the outbreak, which potentially resulted in high performance of the models. Thus, an extensive study covering a longer-term assessment of the predictability of the Google RSV model, specifically in clustered areas, is needed urgently. Such a study is necessary to understand the role of Google RSV data as a secondary public health surveillance tool during a pandemic, and to be better prepared for future outbreaks. Therefore, this study aimed to identify COVID-19 hot and cold spots of disease clustering, and define the predictability performance of the Google RSV model in clustered and non-clustered areas of the USA.

Materials and methods

Study area and data acquisition

County-level data of cumulative daily COVID-19 cases from 48 states (all US contiguous states except Alaska and Hawaii) and the District of Columbia were collected from Johns Hopkins University's Center for Systems Science and Engineering GIS dashboard (Dong et al., 2020), along with new state-level daily COVID-19 cases from the COVID tracking project (The Atlantic, 2021). Data from 20 January to 31 December 2020 were used. Google RSV data were retrieved from the Google Trends website (Google Trends, 2020) for the USA at country and sub-regional level for health categories and web search type. Data were queried for COVID-19-related terms, topics and disease; the top related queries; and most-searched COVID-19 terms in 2020 with a lag of 7 days. This dataset gives the number of search activities made through Google search engines. Data were retrieved for the overall time period (on a weekly basis) and in monthly periods (on a daily basis) for the time frame of the entire study. The daily data were adjusted with weekly-based data to obtain adjusted daily data for the overall study period, as used in previous approaches (Bewerunge, 2018; Rengasamy et al., 2019). In addition, Google mobility data were used in constructing Google RSV models. These mobility data represent changes in time spent in categorized places. Data were queried with a lag of 7 days from COVID-19 Community Mobility Reports (Google, 2021). The datasets used for this analysis are listed in Table 1. All datasets were aggregated into monthly subsets to describe epidemic progression patterns over time.

Data analysis

Getis-Ord General G and local G statistics were utilized to identify monthly hot and cold spots for COVID-19 incidence rate clustering patterns. G statistics are a distance-based approach (Ord and Getis, 1995) that estimate a z-score from observed and expected spatial clustering patterns. The general G statistic was calculated as follows:

$$G = \frac{\sum_{1=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j},\ \forall j \neq i$$

where x_i and x_j are attribute values for features i and j, $w_{i,j}$ is the spatial weight between features i and j, n is the number of features in the dataset, and $\forall j \neq i$ indicates that features i and j cannot be the same feature (Esri, 2021).

A positive z-score indicates spatial clustering in the dataset, whereas negative values represent low clustering patterns. In addition, a z-score close to zero may represent a random spatial pattern in the observation (Getis and Ord, 1992). In this study, the monthly COVID-19 incidence rate was used as an input feature, and spatial relationships between spatial features were determined as contiguity edge corners. Furthermore, an optimized hot spot analysis of local G values was used to identify distributions of monthly COVID-19 hot and cold spots. $P<0.05$ was considered to indicate statistical significance. A clustered state was defined as the presence of hot spot counties, cold spot counties or both.

Monthly country-level correlations between new daily COVID-19 cases and Google RSVs were assessed using Spearman's rank correlation coefficients due to the small numbers of observations and non-normal distributions of the response variables. $P<0.05$ was considered to indicate statistical significance. A moderate correlation was determined as Spearman's rank correlation coefficient of ≥ 0.5, with ≥ 0.7 considered a strong correlation. The term 'COVID testing' (search term) was chosen to assess monthly state-level correlations. This term was used as it may reflect the important issue of COVID-19 testing during the research period.

Moreover, Google RSV models employing highly correlated search data with a lag time of 7 days were calculated using Poisson regressions in a generalized linear model to predict current state-level new daily COVID-19 cases. A Poisson regression was used as a response variable for count data that did not follow a normal distribution (Johnston, 1993). Models were constructed using Google RSVs and mobility data (with the highest correlation coefficient with case data). Model performance in the in-sample data was determined by root mean squared error (RMSE) values, Akaike information criterion (AIC) and Bayesian information criterion (BIC) to compare the performance between models. Multi-layer maps were also created to define monthly predictability performances of Google RSV models in clustered and non-clustered areas of the USA for a 1-year analysis. All spatial analyses and visualizations were conducted using ArcGIS Pro Version 2.6.1 (ESRI, Redlands, CA, USA), and statistical analyses were performed using SAS Version 9.4 (SAS Institute, Cary, NC, USA).

Results

COVID-19 spatial clusters in the USA

In the early stage of the disease outbreak, country-level incidence rates were 0.002~0.005 per 100,000 population, with higher incidence rates in county-level data, which ranged from 0.129 to 3.370 per 100,000 population, as shown in Table 2. However, starting in March 2020, huge increases in cases raised the country-wide incidence rate to 57.110 per 100,000 population and the county-level incidence rate to 1011.124 per 100,000 population. This increasing trend led to a massive national incidence rate, reaching more than 1000 cases in November 2020. Furthermore, counties with the highest incidence rates differed from month to month, indicating the rapid spread of the disease throughout the country.

The Getis-Ord General G test (Table 3) showed clustered patterns in all months during the study period, except in January 2020 due to the limited case count and distribution. Local G exhibited the first cluster identified in California in February 2020 (Figure 1). Afterwards, clusters appeared in neighbouring states, including Washington, Idaho and Colorado, as well as a cluster in the eastern part of the country that grew until May 2020. During this period, two clusters were also found in counties in the southern USA that expanded into large clusters from April to August 2020. However, beginning in September 2020, clusters were circulating in counties in the central USA, and then progressed into more-northerly parts of the country. In contrast, cold spots formed constantly in eastern counties from June to December 2020.

Predictability performance of Google RSV models

During the study period, low to high significant correlations between new daily COVID-19 cases and Google RSVs were found in country-level data (Table 4). Strengths of correlations were increased to the highest point in June 2020 and decreased as the outbreak progressed. For the state-level analysis (Table 5), significant correlations began to emerge in March 2020 (38.78%) and this was the highest point. Percentages of significant correlations fluctuated and increased in June 2020 (22.45%) and in November 2020 (26.53%). While the number of states with clustered areas increased, numbers of significant correlations were only found in low percentages, ranging from 4.08% to 24.49%.

Strong significant correlations were found in several states with clustered and non-clustered counties during the research period, including California, Florida, Illinois, New York and Texas in March 2020, and Texas and South Carolina in June 2020 (Figure 2). These findings suggest that strong correlations were rarely found in clustered areas in the USA during the COVID-19 outbreak. Moreover, the strength of the correlations tended to decrease as the outbreak progressed.

In terms of correlation signs (positive or negative), weak negative correlations were found in several clustered areas, as shown in Table 5. A negative correlation in this study illustrates a declining trend in information searches as the number of cases increased. Furthermore, to understand the pattern of correlations over time and time series of cases, data from three states are presented in Figure 3 as examples. This figure shows time series patterns of new daily COVID-19 cases per 100,000 population in Florida, Illinois and Maryland, along with their monthly correlations with Google search volumes during the study period. Their cluster characteristics as a hot spot, cold spot or non-significant area were determined based on Table 5. Figure 3 demonstrates that linearity between the strength of the correlation and the increase in cases and cluster characteristics differed between states. Significant correlations only tended to be found in the early stages of the outbreak. This finding suggests diverse performance of Google RSV data among states and outbreak periods.

Furthermore, the performance of the Google RSV models in strongly correlated areas (Table 6) resulted in RMSE values in unclustered areas ranging from 81.94 to 95.87, while in clustered areas (hot spots, cold spots and both), RSME values ranged from 61.92 to 1629.92. These findings suggest that Google RSV models may have performed slightly better in clustered areas, but model performances tended to be unstable, as illustrated by the large RMSE range. In addition, mobility variables, particularly transit stations, workplaces and parks, were identified as important variables in model development. However, huge RMSE values may suggest the absence of other important explanatory variables in the models.

Discussion

Spatial heterogeneity of COVID-19 cases at state level

As of 27 December 2020, new cases of COVID-19 in the USA accounted for 68% of all new cases in the Americas, placing the USA as the country with the highest number of new cases and deaths (World Health Organization, 2020). The rapid spread of this

Table 2
Monthly incidence rates of coronavirus disease 2019 in the USA.

Month, 2020	Country-level incidence rate[a]	Counties with the highest incidence rate (state)	County-level incidence rate[a]
January	0.002	Suffolk (MA)	0.129
		Santa Clara (CA)	0.051
		King (WA)	0.046
		Cook (IL)	0.038
		Orange (CA)	0.031
February	0.005	San Benito (CA)	3.370
		Humboldt (CA)	0.720
		King (WA)	0.231
		Washington (OR)	0.170
		Sacramento (CA)	0.132
March	57.110	Westchester (NY)	1011.124
		Blaine (ID)	886.508
		Rockland (NY)	872.079
		Nassau (NY)	624.511
		Richmond (NY)	596.974
April	271.981	Lincoln (AR)	5764.018
		Bledsoe (TN)	3951.936
		Nobles (MN)	3403.514
		Marion (OH)	3324.470
		Dakota (NE)	3081.114
May	218.057	Trousdale (TN)	15,385.550
		Colfax (NE)	5159.589
		Dakota (NE)	4729.698
		Lake (TN)	4616.770
		Buena Vista (IA)	3776.787
June	253.575	Lee (AR)	6528.712
		Buena Vista (IA)	4421.604
		East Carroll (LA)	3797.139
		Lake (TN)	3549.383
		Chattahoochee (GA)	3132.424
July	581.399	La Salle (TX)	4373.808
		Madison (TX)	4331.901
		Crockett (TX)	3847.181
		Chicot (AR)	3320.243
		Columbia (FL)	3153.315
August	440.344	Lafayette (FL)	13,001.420
		Wayne (TN)	6234.385
		Issaquena (MS)	5811.321
		Chattahoochee (GA)	4909.811
		Chicot (AR)	4069.975
September	362.574	Emmons (ND)	4612.707
		Woodward (OK)	4565.296
		Chattahoochee (GA)	4522.657
		Rosebud (MT)	4075.067
		Pawnee (KS)	3717.633
October	577.555	Bon Homme (SD)	12,994.680
		Norton (KS)	12,112.020
		Sheridan (KS)	6856.455
		Faulk (SD)	6250.000
		Buffalo (SD)	6200.787
November	1351.011	Crowley (CO)	20,244.420
		Lee (KY)	10,434.420
		Childress (TX)	10,114.060
		Foster (ND)	8712.459
		Jones (IA)	8642.276
December	1926.729	Bent (CO)	14,336.190
		Lincoln (CO)	11,687.610
		Pershing (NV)	11,075.760
		Alfalfa (OK)	9086.337
		Lassen (CA)	8743.610

[a] Incidence rate per 100,000 population.

disease was observed from geographic variations of the most affected counties in Table 2, which is in line with a previous report (Oster et al., 2020b). In addition, COVID-19 spatial clusters in the USA began to emerge in March 2020 (Figure 1) as a national emergency was declared and widespread testing was implemented (Taylor, 2020). However, some clusters continued to expand with the rise of protests, social distancing restrictions, and the re-opening of public facilities in April 2020 (Hauck et al., 2020; Taylor, 2020). Conditions worsened with the end of national social distancing guidelines on 30 April 2020, which led to the implementation of re-opening policies in various states in May 2020, but conditions varied between counties and cities (Hauck et al., 2020). As a consequence, multiple new clusters began to arise in June 2020, as the highest numbers of new daily cases occurred in the south, west and midwest regions of the country (Taylor, 2020).

The US Government also loosened travel restrictions at the end of June 2020 (US Department of Defense, 2021). During this period, clusters were found in southern and western counties, as reported

Table 3
Results of the global spatial autocorrelation test.

Month, 2020	Observed General G	z-score	*P*-value	Result
January	0.002	0.659	0.510	Random
February	0.011	3.247	0.001	Clustered
March	0.002	49.168	<0.001	
April	0.001	30.914	<0.001	
May	0.001	14.150	<0.001	
June	0.001	28.119	<0.001	
July	0.001	51.984	<0.001	
August	0.000	36.725	<0.001	
September	0.000	34.670	<0.001	
October	0.000	49.418	<0.001	
November	0.000	49.154	<0.001	
December	0.000	27.123	<0.001	

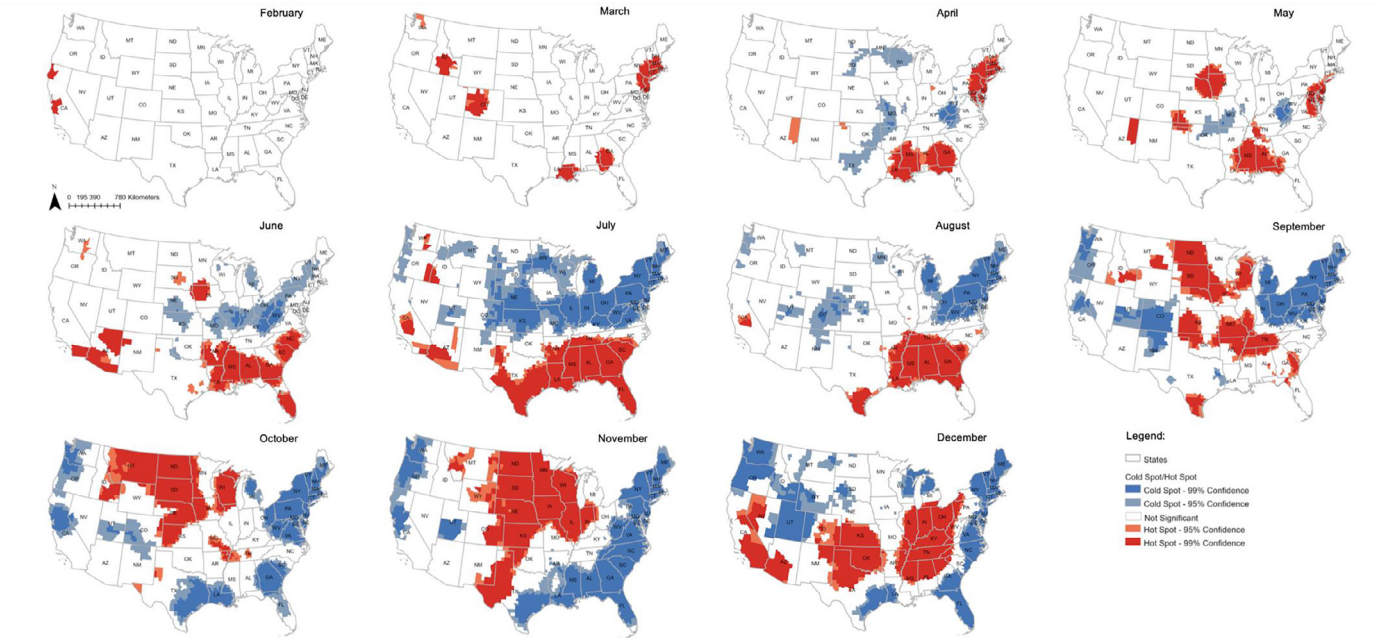

Figure 1. Distribution of coronavirus disease 2019 hot and cold spots in the USA.

previously (Oster et al., 2020b). Massive clusters continued to grow in those areas as positive tests increased in older age groups, leading to higher numbers of hospitalizations, severe outcomes and fatal cases (Oster et al., 2020a). The high COVID-19 incidence rate continued to cause huge clusters in southern counties, which then circulated into central US counties and progressed into northern parts of the country. In addition, better control measures implemented in eastern counties may have been responsible for cold spots arising in those areas.

Research findings showed that small clusters in one or several neighbouring states in the early stage of the outbreak began to develop into larger clusters, involving multiple states, as the outbreak progressed. These results demonstrate the importance of state-to-state coordination in implementing control measures to tackle the spread of new infectious disease outbreaks. Having various preventive policies in neighbouring areas may have promoted the massive growth of clusters. As control measures at state and local levels directly influence the disease incidence and cluster magnitude (CDC COVID-19 Response Team, 2020; Desjardins et al., 2020), coordinated responses are needed urgently. Moreover, this study illustrates that spatial analyses provided clear spatial patterns of disease spread, which could lead to the timely implementation of control measures before high-level community transmission has occurred. Therefore, this type of analysis should be considered as a crucial approach in public health surveillance during outbreak situations to implement focused public health actions. However, spatial clusters may not be induced by the time variable alone, and incorporating other explanatory variables would be beneficial in understanding differences in spatial patterns.

Factors that may affect the predictability performance of Google RSV models

Furthermore, as described in the Results section above, correlations between RSVs and COVID-19 varied in space and time, and the strength of the correlations also tended to decrease as the outbreak progressed. Similar results were found in a previous study, which reported that COVID-19 Google searches did not correspond with actual disease dynamics in 40 European countries (Szmuda et al., 2020). Diverse performances of Google RSV models found in this study suggest that the model performance in predicting new cases can be affected by several aspects, including COVID-19 transmission dynamics, policy-driven community awareness, and past outbreak experiences.

COVID-19 transmission dynamics may affect how the accuracy of the Google RSV model differed month to month as the outbreak progressed. In the early phase, high correlations may have appeared as a result of massive searches from affected communities and groups of people who were concerned about the emerging outbreak. However, with the extensive spread of the dis-

Table 4
Correlations between new daily cases of coronavirus disease 2019 (COVID-19) and Google relative search volumes of county-level data in the USA.

Term	Month, 2020 Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
A		0.568	0.497	0.476		0.909		0.547				
B		0.698	0.848	0.442	0.401	0.907	0.543	0.539				
C		0.568	0.512	0.484		0.897		0.568	-0.386	0.433		
D		0.831	0.945	0.429	0.370	0.910	0.478	0.524		0.376		
E		0.671	0.714	0.471		0.874	0.430	0.615				
F		0.568	0.512	0.458		0.902		0.522		0.422		
G			0.768			0.902		0.408				0.405
H			0.773	-0.460		0.929	0.531	0.509		0.747		

Terms for data query
A: Coronavirus (virus)
B: 'Coronavirus disease 2019' (disease)
C: coronavirus (search term)
D: covid (search term)
E: covid-19 (search term)
F: coronavirus + 'coronavirus update' + 'coronavirus
G: symptoms' (search terms)
H: 'covid symptoms' (search term) 'covid testing' (search term)

Strength of correlation
Weak correlation (r=0~≤0.49)
Moderate correlation (r=0.50~≤0.69)
Strong correlation (r=0.70~≤1)
All reported correlations were significant at P≤0.05

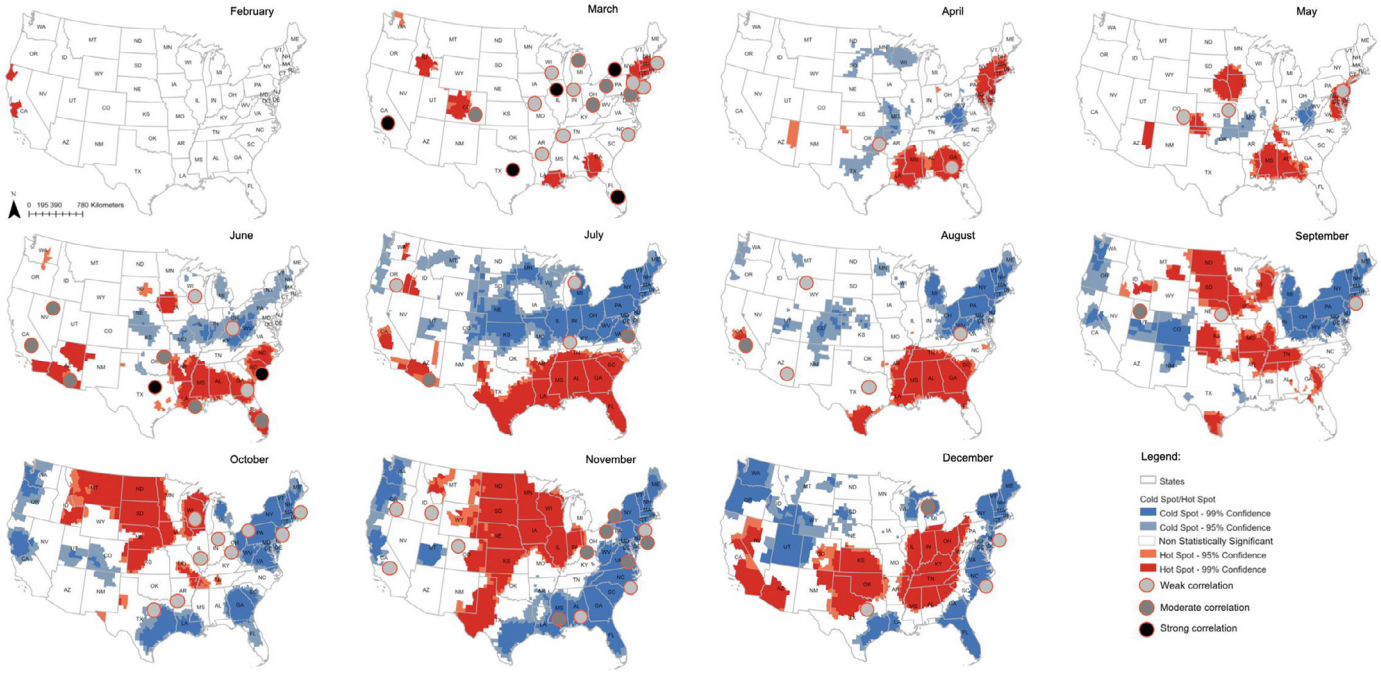

Figure 2. Correlations between new daily cases of coronavirus disease 2019 and Google relative search volumes in clustered and unclustered areas in the USA.

ease, people may have been overwhelmed by the enormous volumes of circulating information, and stopped searching COVID-19-related issues. This may have decreased the volume of information searches and the correlation strength, as observed in earlier studies (Husnayain et al., 2020a,b). At this point, the Google RSV model should have been built based on specific terms rather than using general keywords. This study showed that the use of general terms of COVID-19 may have been robust only in the first 5 months after the outbreak began (February–June 2020), as shown in Table 4. Beginning in July 2020, the more specific term of 'covid testing' (search term) had an increasing correlation coefficient. This possibly illustrates that more specific terms, such as vaccines, current control measures and preventive measures, should be used to better represent the public's current concerns, awareness or restlessness. Consequently, routine keyword identification is important to ensure precise analyses when utilizing Google RSV data.

The performance of the Google RSV model may also have been affected by policy-driven community awareness. This means that policies implemented in response to COVID-19 may have influenced public awareness towards the growing outbreak. As state-level policies are primarily affected by governors' decisions, governors' perceptions will contribute directly to the formation of community perceptions and reactions. However, these may also be influenced by the governor's political affiliation, which has been discussed in several previous articles (Green and Tyson, 2020; Jiang et al., 2020; Adolph et al., 2021). Hence, public perceptions and reactions may have altered COVID-19 online information searches to a certain degree. A previous study showed that COVID-19 queries in the USA increased more slowly than they did in other countries (Husain et al., 2020), which may also describe how the public responded to the degree of the emergency.

Finally, past experience with an outbreak may affect the robustness of the Google RSV model. As COVID-19 was a new outbreak that had global impacts, the public may have responded in diverse manners. Countries which were highly affected by the previous severe acute respiratory syndrome and Middle Eastern respiratory

Table 5
Correlations between new daily cases of coronavirus disease 2019 and Google relative search volumes of state-level data in the USA.

State	Month, 2020											
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Alabama											0.443	
Arizona						0.567	0.557	0.440				
Arkansas			0.461							0.374		
California			0.704			0.577		0.525			0.421	
Colorado			0.590		-0.470						0.462	
Connecticut									0.427			
Delaware											0.480	-0.369
Florida			0.746			0.592						
Georgia				0.461		0.442						
Idaho											0.383	
Illinois			0.716							0.448		
Indiana			0.430							0.449		
Iowa												
Kansas					0.445							
Kentucky							0.383					
Louisiana						0.547						
Maine												
Maryland			0.668									
Massachusetts			0.405							0.427		
Michigan			0.502				-0.361					0.541
Minnesota												
Mississippi											0.572	
Missouri			0.443									
Montana												
Nebraska									0.402			
Nevada						0.612						
New Hampshire												
New Jersey			0.435								0.509	
New Mexico												
New York			0.753								0.590	
North Carolina			0.480								0.366	-0.383
North Dakota												
Ohio			0.544			0.418				0.409	0.525	
Oklahoma				-0.377		0.649						
Oregon							-0.463				0.385	
Pennsylvania			0.643							0.373	0.532	
Rhode Island												
South Carolina						0.740						
South Dakota												
Tennessee			0.470									
Texas			0.756			0.769		0.400		0.377		0.401
Utah									0.570			
Vermont												
Virginia							0.500				0.699	
Washington			0.408									
West Virginia								0.416				
Wisconsin			0.392			0.380				0.416		
Wyoming								-0.365				
District of Columbia					-0.358					-0.436		
Number of states with a significant correlation [*n* (%)]	0 (0.000)	0 (0.000)	19 (38.776)	2 (4.082)	3 (6.122)	11 (22.449)	5 (10.204)	5 (10.204)	3 (6.122)	9 (18.367)	13 (26.531)	4 (8.163)
Number of states with clustered counties[a] [*n* (%)]		1 (2.041)	17 (34.694)	35 (71.429)	30 (61.224)	36 (73.469)	49 (100)	44 (89.796)	48 (97.959)	45 (91.837)	46 (93.878)	48 (97.959)
Number of states with a significant correlation and clustered counties[a] [*n* (%)]		0 (0.000)	6 (12.245)	2 (4.082)	3 (6.122)	10 (20.408)	5 (10.204)	4 (8.163)	3 (6.122)	8 (16.327)	12 (24.490)	4 (8.163)
Note:		Hot spot areas. Cold spot areas. Hot and cold spot areas. Non-significant areas at $P \leq 0.05$.										

[a] States with hot spot counties, cold spot counties and both.

syndrome outbreaks may have exhibited high numbers of searches and strong predictability performance of Google RSV models, particularly China (Li et al., 2020), Taiwan (Husnayain et al., 2020a) and South Korea (Husnayain et al., 2020b).

In brief, as the accuracy of the COVID-19 Google RSV model may be influenced by these three major aspects, the Google RSV model derived from general terms in the USA was only valid for use in the first 5 months of the outbreak. More specific keywords should be used in later stages of the outbreak. Moreover, because of the limited strong correlations found in clustered areas, the Google RSV model in the USA may be better utilized for designing risk communication rather than for predictive purposes. The sign (positive or negative) of correlations can be utilized to understand public responses to control and preventive measures, as well as for communicating risk. Negative correlations could be used as an alert, indicating the need for intensive risk communication and a campaign of preventive measures.

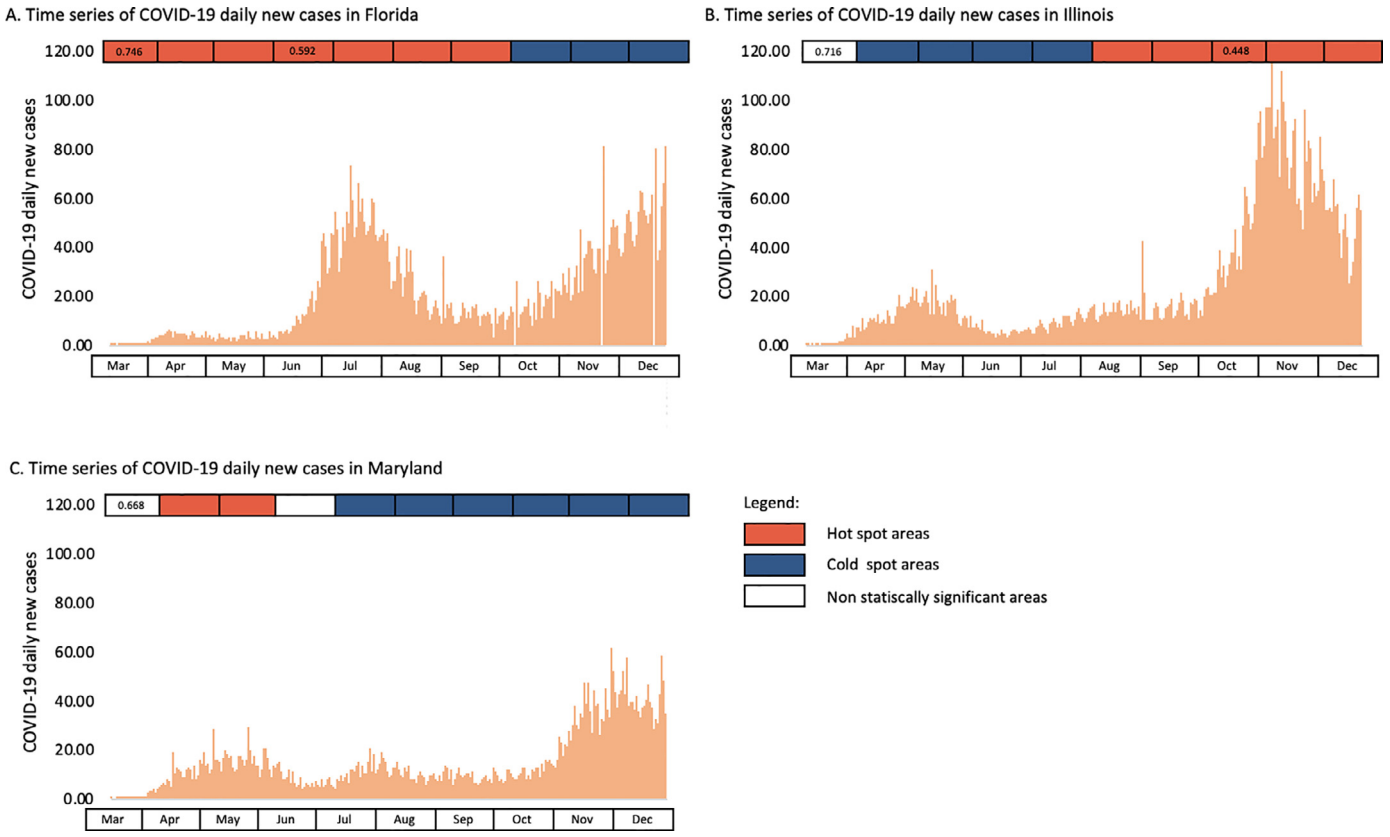

Figure 3. Time series of new daily cases of coronavirus disease 2019 (COVID-19) per 100,000 population in Florida, Illinois and Maryland.

In addition, this study may be subject to several limitations resulting from errors in reporting case data and limited terms used for the data query. This study only used English terms, and did not consider Spanish or other indigenous languages which are also used in the USA. Future studies could incorporate spatial modelling tasks for predicting active clusters that combine distributions of Google RSVs with other significant explanatory variables. Such variables might include income inequalities, median household incomes, the proportion of black females, the proportion of nurse practitioners (Mollalo et al., 2020), age, disability, language, race, occupation, urban status (Andersen et al., 2021) and crowded housing conditions (Dasgupta et al., 2020). However, more dynamic variables may be required to increase the performance of the model.

This study found that mobility variables are important variables in model development. Transit stations, workplaces and parks became the most common variables included in the models for a few months in the early stage of the outbreak, as working from home was widely implemented. However, the model should be constructed carefully to prevent the introduction of biases when designing the models. Furthermore, this study only used Google RSVs and mobility data with a lag of 7 days for analysis. This period was chosen to prevent a mass media reporting effect on Google searches over longer lag periods. Further analysis in defining the best lag period is needed to increase the accuracy of the study.

Several considerations when utilizing Google search data as a public health surveillance tool

Utilizing Google RSV data as a secondary public health surveillance tool is promising for the future. Google search data are publicly available at low cost, and potentially cover online information-seeking behaviours of the majority of people as most people use the Internet to search for specific terms in search engines (Mavragani, 2020; Schneider et al., 2020). Therefore, internet search data could potentially provide patterns unreported by traditional surveillance measures, such as the number of ill people who did not seek medical treatment but searched for health-related information (Barros et al., 2020). This method can potentially be used as an online surveillance tool in countries with limited resources (Schneider et al., 2020). Online queries also offer anonymous data that can potentially assess a large population (Mavragani, 2020). These opportunities make this infodemiological method a valuable approach in understanding the occurrence of illnesses circulating in the general population that can be inspected promptly. However, the findings of this study suggest the variability of Google RSV model performance between states and time periods (Figures 2 and 3; Tables 4–6). Different states may utilize Google RSVs in different frameworks. In highly correlated states, Google searches may be used for prediction tasks, while other states may use them to understand public responses and design risk communication.

Although promising, some issues need to be considered when employing information search data. Changes in online information and communication patterns that reflect user-generated data in infodemiology need to be validated to distinguish a true epidemic from an epidemic of fear (Eysenbach, 2011). People searching for flu information do not always reflect people suffering from flu, and can be affected by sudden incidents or events (Barros et al., 2020; Eysenbach, 2011; Mavragani, 2020). Recent studies have shown that Google Trends data cannot distinguish whether searches represent public concern or interest (Springer et al., 2020a,b), and the surge in online information searches related to coronaviruses for particular terms was irrespective of the time occurrence of the outbreak, which indicates that Google Trends data were closely af-

Table 6
Performance of the Google relative search volume (RSV) models in strongly correlated areas (with Spearman's rank correlation coefficients of ≥0.7).

Model	Coef. (95% CI)	*P*-value	RMSE	AIC	BIC
February					
California[a]					
Intercept	3.582 (3.518∼3.647)	<0.001	81.942	1244.475	1248.777
Google RSVs	0.056 (0.029∼0.083)	<0.001			
Mobility (transit stations)	-0.063 (-0.065∼-0.063)	<0.001			
Florida[b]					
Intercept	3.673 (3.611∼3.734)	<0.001	61.920	1295.194	1299.496
Google RSVs	-0.235 (-0.284∼-0.185)	<0.001			
Mobility (transit stations)	-0.072 (-0.074∼-0.070)	<0.001			
Illinois[a]					
Intercept	3.822 (3.764∼3.880)	<0.001	95.865	2010.341	2014.643
Google RSVs	0.193 (0.151∼0.235)	<0.001			
Mobility (transit stations)	-0.050 (-0.051∼-0.048)	<0.001			
New York[b]					
Intercept	6.259 (6.242∼6.276)	<0.001	1629.921	27386.572	27390.874
Google RSVs	-0.152 (-0.160∼-0.145)	<0.001			
Mobility (transit stations)	-0.055 (-0.056∼-0.055)	<0.001			
Texas[a]					
Intercept	2.665 (2.565∼2.766)	<0.001	84.144	1367.539	1371.841
Google RSVs	0.325 (0.285∼0.366)	<0.001			
Mobility (workplaces)	-0.086 (-0.089∼-0.082)	<0.001			
June					
South Carolina[b]					
Intercept	5.876 (5.839∼5.914)	<0.001	294.305	3182.487	3186.691
Google RSVs	0.030 (0.028∼0.032)	<0.001			
Mobility (parks)	0.012 (0.011∼0.013)	<0.001			
Texas[c]					
Intercept	10.458 (10.401∼10.514)	<0.001	961.395	8381.994	8386.198
Google RSVs	0.020 (0.019∼0.020)	<0.001			
Mobility (transit stations)	0.101 (0.099∼0.103)	<0.001			

Coef., coefficient; RMSE, root mean squared error; AIC, Akaike information criterion; BIC, Bayesian information criterion.
[a] Non-significant areas.
[b] Hot spot areas.
[c] Hot and cold spot areas.

fected by media coverage (Sousa-Pinto et al., 2020). Therefore, this proxy should be used with caution because it could be affected by false-positive events, such as in the case of an infodemic where Google searches may more closely represent the public's fear instead of disease dynamics.

Regular updates of keywords used in search query monitoring are necessary proxies to maintain the validity of emerging trends and changes in a population's health-seeking information behaviours. Other issues in the infodemiological approach are related to internet penetration and access problems, preferences used by certain age groups, and transparency in how internet search data are collected (Barros et al., 2020). In addition, information search data may leak from future to past observations in the case of retrospective analyses. Thus, future research should consider weekly data retrieval during the season to prevent information leaks from future to past observations (Schneider et al., 2020). Other emerging data sources, including Twitter, websites/platforms, blogs/forums, Facebook, reviews, mobile apps, Instagram, news/media, Wikipedia, health records and online surveys, are also important in conducting digital surveillance.

Conclusions

Small clusters in one or several neighbouring states in the early stage of the outbreak triggered larger clusters involving multiple states as the outbreak progressed. In the later phase of the outbreak, clusters circulated in counties located in the middle of the country, and progressed into northern parts. These results demonstrate the importance of state-to-state coordination in implementing control measures to tackle the spread of new infectious disease outbreaks. In addition, better control measures may have been performed in eastern counties based on the rise of cold spots in those areas.

Variabilities in Google RSV model performances were found among states and time periods. This suggests that different frameworks need to be implemented in each state when utilizing Google RSV data. In addition, mobility variables were identified as important variables in predicting new daily COVID-19 cases. Google searches may be used in prediction tasks in highly correlated states, while they can be used in other areas to understand public responses and design risk communication. Moreover, the sign (positive or negative) of the correlation can be utilized to understand public responses to control and preventive measures, as well as in communicating risk.

Declaration of Competing Interest

None declared.

Funding

This work was supported by the Ministry of Science and Technology in Taiwan (Grant No. MOST109-2221-E-038-018 and MOST110-2628-E-038-001) and the Higher Education Sprout Project by Ministry of Education in Taiwan (Grant No. DP2-110-21121-01-A-13) to Emily Chia-Yu Su.

Ethical approval

Not required.

Acknowledgments

The authors wish to acknowledge Johns Hopkins University's Center for Systems Science and Engineering GIS dashboard and the COVID tracking project for providing open access data on daily COVID-19 cases in the USA. In addition, the authors wish to acknowledge Google for allowing access to freely available data on Google RSVs and community mobility.

References

Adolph C, Amano K, Bang-Jensen B, Fullman N, Wilkerson J. Pandemic politics: timing state-level social distancing responses to COVID-19. J Health Polit Policy Law 2021;46:211–33.

Andersen LM, Harden SR, Sugg MMP, Runkle JDP, Lundquist TE. Analyzing the spatial determinants of local Covid-19 transmission in the United States. Sci Total Environ 2021;754.

Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SRN. 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:e18828.

Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020;22:e13680.

Bento AI, Nguyen T, Wing C, Lozano-Rojas F, Ahn YY, Simon K. Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. Proc Natl Acad Sci USA 2020;117:11220–2.

Bewerunge F. Google Trends: how to acquire daily data for broad time frames. 2018. Available at: https://medium.com/@bewerunge.franz/google-trends-how-to-acquire-daily-data-for-broad-time-frames-b6c6dfe200e6 (last accessed 2 May 2020).

CDC COVID-19 Response Team. Geographic differences in COVID-19 cases, deaths, and incidence – United States, February 12–April 7, 2020. MMWR Morb Mortal Wkly Rep 2020;69:465–71.

Cordes J, Castro MC. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spat Spatiotemporal Epidemiol 2020;34.

Cousins HC, Cousins CC, Harris A, Pasquale LR. Regional infoveillance of COVID-19 case rates: analysis of search-engine query patterns. J Med Internet Res 2020;22:e19483.

Dasgupta S, Bowen VB, Leidner A, Fletcher K, Musial T, Rose C, et al. Association between social vulnerability and a county's risk for becoming a COVID-19 hotspot – United States, June 1–July 25, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1535–41.

Desjardins MR, Hohl A, Delmelle EM. Rapid surveillance of COVID-19 in the United States using a prospective space–time scan statistic: detecting and evaluating emerging clusters. Appl Geogr 2020;118.

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4.

Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with internet search volumes: a Google Trends™ analysis. Int J Infect Dis 2020;95:192–7.

Esri. How high/low clustering (Getis-Ord General G) works. 2021. Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-high-low-clustering-getis-ord-general-g-spat.htm (Accessed 2 May 2021).

Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc 2006;2006:244–8.

Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011;40(2):S154–8 Suppl..

Franch-Pardo I, Napoletano BM, Rosete-Verges F, Billa L. Spatial analysis and GIS in the study of COVID-19. A review. Sci Total Environ 2020;739.

Getis A, Ord JK. The analysis of spatial association by use of distance statistics. Geogr Anal 1992;24:189–206.

Google. FAQ about Google Trends data. 2020. Available at: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (Accessed 15 June 2020).

Google. COVID-19 community mobility reports. 2021. Available at: https://www.google.com/covid19/mobility/(Accessed 22 February 2021).

Google Trends. Google Trends. 2020. Available at: https://trends.google.com/(Accessed 15 June 2020).

Green TV, Tyson A. 5 facts about partisan reactions to COVID-19 in the U.S. 2020. Available at: https://www.pewresearch.org/fact-tank/2020/04/02/5-facts-about-partisan-reactions-to-covid-19-in-the-u-s/(Accessed 3 February 2021).

Hauck G, Gelles K, Bravo V, Thorson M. Five months in: a timeline of how COVID-19 has unfolded in the US. 2020. Available at: https://www.usatoday.com/in-depth/news/nation/2020/04/21/coronavirus-updates-how-covid-19-unfolded-u-s-timeline/2990956001/(Accessed 3 February 2021).

Husain I, Briggs B, Lefebvre C, Cline DM, Stopyra JP, O'Brien MC, et al. Fluctuation of public interest in COVID-19 in the United States: retrospective analysis of Google Trends search data. JMIR Public Health Surveill 2020;6:e19969.

Husnayain A, Fuad A, Su EC. Applications of Google search trends for risk communication in infectious disease management: a case study of COVID-19 outbreak in Taiwan. Int J Infect Dis 2020a;95:221–3.

Husnayain A, Shim E, Fuad A, Su EC. Understanding the community risk perceptions of the COVID-19 outbreak in South Korea: infodemiology study. J Med Internet Res 2020b;22:e19788.

Jiang J, Chen E, Lerman K, Ferrara E. Political polarization drives online conversations about COVID-19 in the United States. Hum Behav Emerg Technol 2020;2. doi:10.1002/hbe2.202.

Johnston G. SAS software to fit the Generalized Linear Model. SUGI Proceedings. Cary, NC: SAS Institute Inc., 1993.

Kurian SJ, Bhatti AUR, Alvi MA, Ting HH, Storlie C, Wilson PM, et al. Correlations between COVID-19 cases and Google Trends data in the United States: a state-by-state analysis. Mayo Clin Proc 2020;95:2370–81.

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. Euro Surveill 2020;25. doi:10.2807/1560-7917.ES.2020.25.10.2000199.

Maroko AR, Nash D, Pavilonis BT. COVID-19 and inequity: a comparative spatial analysis of New York City and Chicago Hot Spots. J Urban Health 2020;97:461–70.

Mavragani A. Infodemiology and infoveillance: scoping review. J Med Internet Res 2020;22:e16206.

Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. Sci Rep 2020;10:20693.

Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014;14:160–8.

Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Sci Total Environ 2020;728.

Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 1995;27:286–306.

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;37.

Oster AM, Caruso E, DeVies J, Hartnett KP, Boehmer TK. Transmission dynamics by age group in COVID-19 hotspot counties – United States, April–September 2020. MMWR Morb Mortal Wkly Rep 2020a;69:1494–6.

Oster AM, Kang GJ, Cha AE, Beresovsky V, Rose CE, Rainisch G, et al. Trends in number and distribution of COVID-19 hotspot counties – United States, March 8–July 15, 2020. MMWR Morb Mortal Wkly Rep 2020b;69:1127–32.

Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 incidence using anosmia and other COVID-19 symptomatology: preliminary analysis using Google and Twitter. Otolaryngol Head Neck Surg 2020;163:491–7.

Ramírez IJ, Lee J. COVID-19 emergence and social and health determinants in Colorado: a rapid spatial analysis. Int J Environ Res Public Health 2020;17:3856.

Rengasamy V, Dougherty P, Elkins E, Sonnek P. Pytrends daily data. 2019. Available at: https://github.com/GeneralMills/pytrends/blob/master/pytrends/dailydata.py (Accessed 2 May 2021).

Schneider PP, van Gool CJ, Spreeuwenberg P, Hooiveld M, Donker GA, Barnett DJ, et al. Using web search queries to monitor influenza-like illness: an exploratory retrospective analysis, Netherlands, 2017/18 influenza season. Euro Surveill 2020;25.

Snyder BF, Parks V. Spatial variation in socio-ecological vulnerability to COVID-19 in the contiguous United States. Health Place 2020;66.

Sousa-Pinto B, Anto A, Czarlewski W, Anto JM, Fonseca JA, Bousquet J. Assessment of the impact of media coverage on COVID-19-related Google Trends data: infodemiology study. J Med Internet Res 2020;22:e19611.

Springer S, Menzel LM, Zieger M. Google Trends provides a tool to monitor population concerns and information needs during COVID-19 pandemic. Brain Behav Immun 2020a;87:109–10.

Springer S, Menzel LM, Zieger M. Google Trends reveals: Focus of interest in the population is on treatment options rather than theories about COVID-19 animal origin. Brain Behav Immun 2020b;87:134–5.

Strzelecki A. The second worldwide wave of interest in coronavirus since the COVID-19 outbreaks in South Korea, Italy and Iran: a Google Trends study. Brain Behav Immun 2020;88:950–1.

Szmuda T, Ali S, Hetzger TV, Rosvall P, Słoniewski P. Are online searches for the novel coronavirus (COVID-19) related to media or epidemiology? A cross-sectional study. Int J Infect Dis 2020;97:386–90.

Taylor DB. A timeline of the coronavirus pandemic. 2020. Available at: https://www.nytimes.com/article/coronavirus-timeline.html (Accessed 3 February 2021).

The Atlantic. The COVID Tracking Project. 2021. Available at: https://covidtracking.com/data. (Accessed 10 January 2021).

US Department of Defense. Coronavirus: DOD response timeline. 2021. Available at: https://www.defense.gov/Explore/Spotlight/Coronavirus/DOD-Response-Timeline/(last accessed 3 February 2021).

Wang C, Li Z, Clay Mathews M, Praharaj S, Karna B, Solís P. The spatial association of social vulnerability with COVID-19 prevalence in the contiguous United States. Int J Environ Health Res 2020. doi:10.1080/09603123.2020.1847258.

World Health Organization. COVID-19 weekly epidemiological update. Geneva: WHO; 2020.

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;5:1–6.