



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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis

Shyam J. Kurian, BA; Atiq ur Rehman Bhatti, MD; Mohammed Ali Alvi, MBBS; Henry H. Ting, MD, MBA; Curtis Storlie, PhD; Patrick M. Wilson, MPH; Nilay D. Shah, PhD; Hongfang Liu, PhD; and Mohamad Bydon, MD

Abstract

Objective: To evaluate whether a digital surveillance model using Google Trends is feasible for obtaining accurate data on coronavirus disease 2019 and whether accurate predictions can be made regarding new cases.

Methods: Data on total and daily new cases in each US state were collected from January 22, 2020, to April 6, 2020. Information regarding 10 keywords was collected from Google Trends, and correlation analyses were performed for individual states as well as for the United States overall.

Results: Among the 10 keywords analyzed from Google Trends, *face mask*, *Lysol*, and *COVID stimulus check* had the strongest correlations when looking at the United States as a whole, with *R* values of 0.88, 0.82, and 0.79, respectively. Lag and lead Pearson correlations were assessed for every state and all 10 keywords from 16 days before the first case in each state to 16 days after the first case. Strong correlations were seen up to 16 days prior to the first reported cases in some states.

Conclusion: This study documents the feasibility of syndromic surveillance of internet search terms to monitor new infectious diseases such as coronavirus disease 2019. This information could enable better preparation and planning of health care systems.

© 2020 Mayo Foundation for Medical Education and Research ■ Mayo Clin Proc. 2020;95(11):2370-2381



From the Neuro-Informatics Laboratory, Department of Neurologic Surgery (S.J.K., A.R.B., M.A.A., M.B.), Mayo Clinic Alix School of Medicine (S.J.K.), Department of Cardiovascular Medicine (H.H.T.), Department of Health Sciences Research (C.S., P.M.W., N.D.S., H.L.), and Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery (P.M.W., N.D.S.), Mayo Clinic, Rochester, MN.

Cases of pneumonia of unknown etiology appeared at the end of 2019 in Wuhan, China.¹ Further sequencing analysis revealed the involvement of a novel strain of virus named *severe acute respiratory syndrome coronavirus 2* obtained from the samples of the lower respiratory tract of infected patients.² The number of cases quickly accelerated, and eventually the disease spread to the United States, with the first confirmed case announced in January 2020; the World Health Organization labeled the situation a pandemic on March 11, 2020.

Web-based big data analytics has been gaining popularity in its potential to predict the distribution of infectious diseases.³ Internet usage has brought about a revolution when it comes to health care knowledge accessibility to the public. Monitoring and

analysis of Internet data has come under the research field known as *infodemiology*, defined as obtaining data from Web-based resources and repurposing it to inform public health and health policymaking.⁴ Web-based activity detection tools can play a vital role in early detection of infectious events and help in the timely preparedness of respective health care systems in order to avoid the adverse consequences of being caught by surprise. Among these Web-based surveillance tools, one of the most prominent is Google Trends.

Google Trends is one of the most efficient trend analyzers to determine Internet search behavior. Google search is based on pattern analysis focused on the most searched keywords that are centered around concerns of the general public. Google Trends provides valuable insights into

community dynamics and health-related problems, particularly in the area of infectious diseases. Big data produced by Google Trends has proved to be valuable for correlation assessments and forecasting models of a number of infectious diseases including influenza, Middle East respiratory syndrome (MERS), Zika virus, and more; it has also been found to be a useful tool for the assessment of dementia cases in the population.⁵⁻⁸

Since the first case of coronavirus disease 2019 (COVID-19) appeared in the United States, there has been an exponential increase in the daily number of cases. The United States now has the highest number of cases in the world, with the most deaths globally.⁹ The purpose of this study was to explore whether there is a correlation between certain keywords searched by the general public in Google and the number of COVID-19 cases in the United States on a state-by-state basis. Significant correlations could suggest the utilization of Google Trends to predict new COVID-19 case locations and hotspots.

METHODS

Google Trends Data

Google Trends processes the magnitude of Web searches performed for a specified keyword, among other searches, providing the relative search volume (RSV) for each keyword. This standardized value is calculated by dividing the total number of searches for a keyword by the total searches of the geography and time range it represents to compare relative popularity. The resultant number ranges from 0 to 100 and is based on the topic's daily popularity compared with its search popularity over a given time frame.¹⁰ Trend changes are displayed online for time series of interest. Keywords can be filtered by location (worldwide, country, state, city) and time span. Data are collected in a time series presented on a normalized scale of 0 to 100, where 0 represents no search and 100 represents the peak search activity for a particular keyword or string. Data can be downloaded as a ".csv" (comma-separated values) file.

Google Trends' daily base data were mined in our study from January 22, 2020, to April 6, 2020. In total, 10 keywords related to COVID-19 were chosen on the basis of popularity and increasing patterns on the Internet and Google News in the study period. The following keywords were searched: *COVID symptoms*, *coronavirus symptoms*, *sore throat+shortness of breath+fatigue+cough*, *coronavirus testing center*, *loss of smell*, *Lysol* (sanitizer), *antibody*, *face mask*, *coronavirus vaccine*, and *COVID stimulus check*. Keyword categories included disease symptoms, prevention, testing, and possible treatments. Our search method was to perform a query for each keyword for each US state individually. In total, we obtained data for 50 states for each selected keyword.

COVID-19 Case Data

Data for the daily new and total number of confirmed cases and deaths has been tracked and reported by Johns Hopkins University Center for Systems Science and Engineering. At the time of this study, the data provided included COVID-19 case data on a county-by-county basis for each of the 50 states. Total US cases reported from January 22, 2020, to April 6, 2020, were available; this is the time frame utilized in this study. County data for each state were combined to create a state-by-state data set.

Statistical Analyses

To assess the relationship between COVID-19 cases and keyword patterns in Google Trends, correlation analysis was performed using R version 3.6.2 (R Foundation for Statistical Computing). Ten keywords in Google Trends were searched and data were collected from January 22, 2020, to April 6, 2020. We plotted each keyword's RSV from January to April of 2020. Pearson correlation coefficients were calculated between each keyword's standardized RSV and the number of daily new COVID-19 cases, and 95% CIs were also calculated. We used the correlation coefficients of selected keywords and daily new COVID-19 cases to create a heat map for each of the 50 states at time zero (the

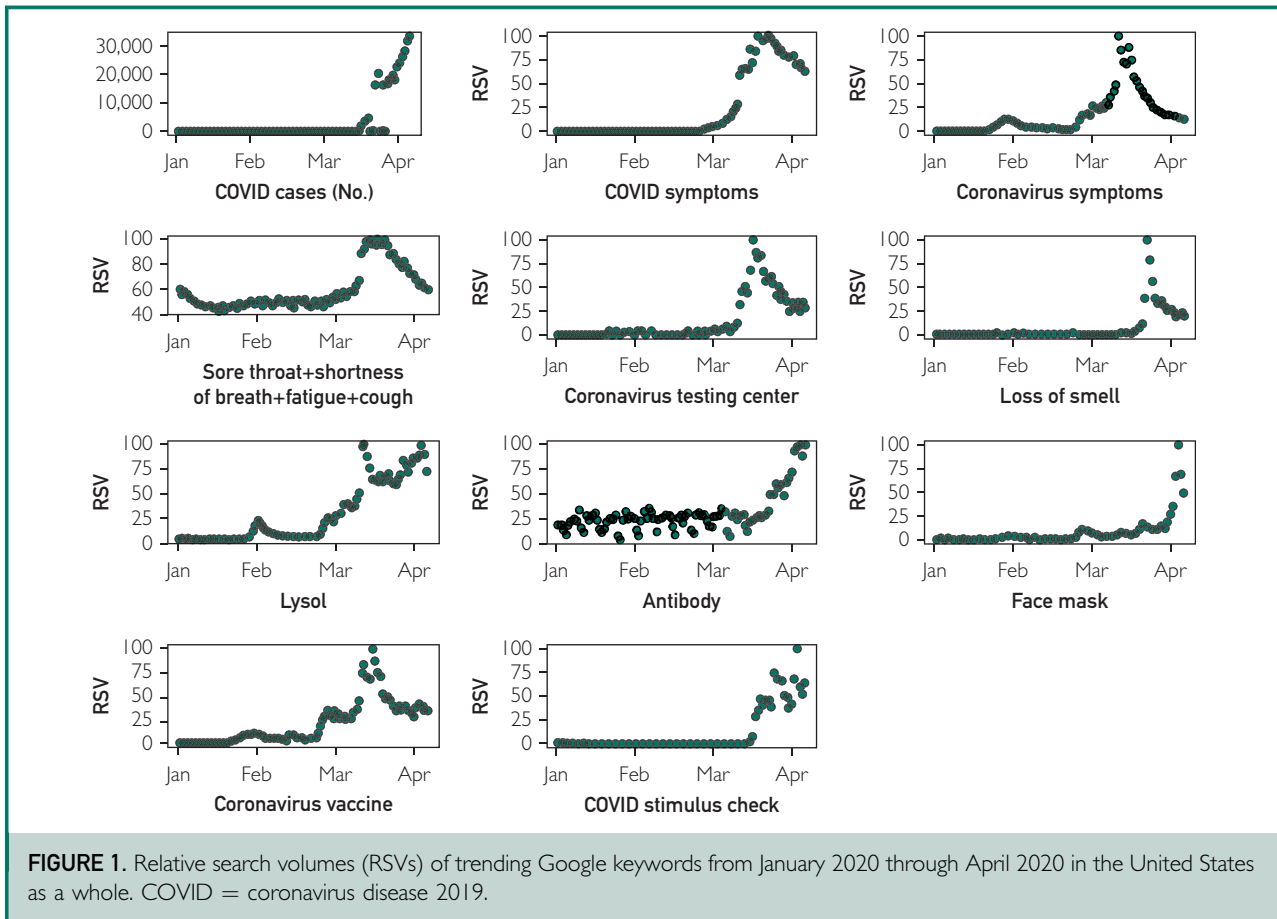


FIGURE 1. Relative search volumes (RSVs) of trending Google keywords from January 2020 through April 2020 in the United States as a whole. COVID = coronavirus disease 2019.

day of the first case in the state). To study the association between COVID-19 cases and Google search trends for each of the 10 keywords, we created scatterplots showing the number of COVID-19 cases against a standardized daily Google search RSV value.

Lag and lead Pearson correlation coefficients were calculated for all 50 states as well as the United States as a whole. The lag/lead times for each state started 16 days prior to time zero (day of the first case in that state) and 16 days after time zero. We compared the correlation coefficients for each keyword's RSV and daily new COVID-19 cases between day -16 and day $+16$ in all 50 states as well as the United States as a whole.

RESULTS

Ten keywords were searched in Google Trends and data were compounded from

January 22, 2020, to April 6, 2020. Keywords generally increased in search popularity over time compared with baseline; some keywords, such as *COVID symptoms*, peaked in popularity toward mid-March, while others, such as *face mask*, continued to increase in popularity into April (Figure 1). Correlation coefficients were calculated between each keyword and each of the 50 states' daily new COVID-19 cases as well as the daily new COVID-19 cases in the United States as a whole. When looking at the United States as a whole, keyword correlations ranged from $R=0.06$ (*coronavirus symptoms*) to $R=0.88$ (*antibody*); 6 of the 10 keywords had moderate correlations ($R=0.3$ to 0.7) with daily new COVID-19 cases in the United States, while 3 of the 10 keywords had strong correlations ($R=0.7$ to 1) (Table 1). When looking at correlations on a state-by-state basis, 4 keywords with considerable correlations nationwide

TABLE 1. Overall US Correlation Coefficients for 10 Google Keywords and Daily New COVID-19 Cases

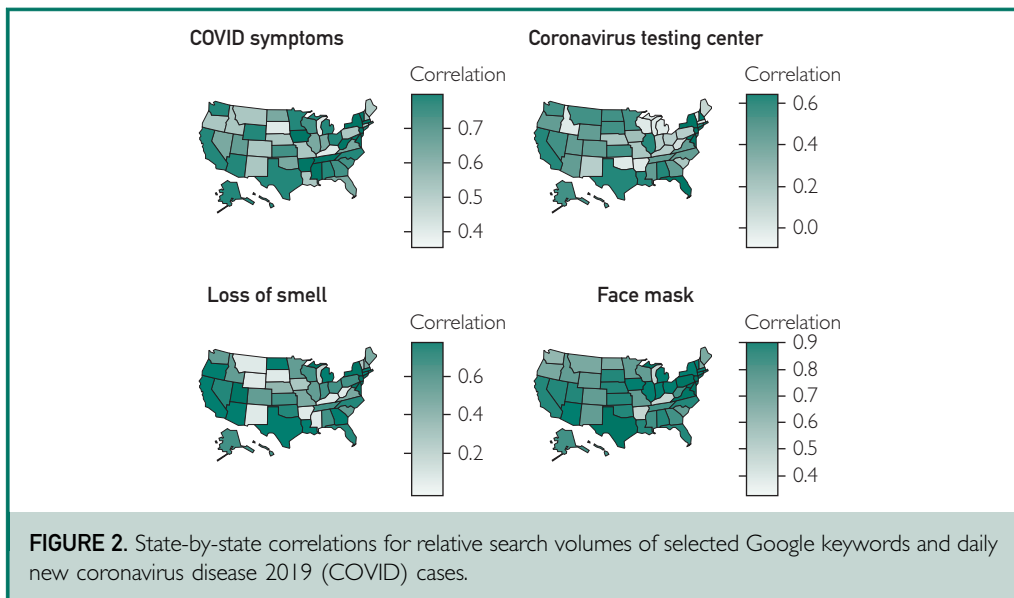
Keyword	R	95% CI (lower)	95% CI (upper)
COVID-19 symptoms	0.61	0.47	0.73
Coronavirus symptoms	0.06	-0.14	0.26
Sore throat + shortness of breath + fatigue + cough	0.31	0.12	0.48
Coronavirus testing center	0.39	0.2	0.54
Loss of smell	0.61	0.46	0.72
Lysol	0.66	0.52	0.76
Antibody	0.88	0.83	0.92
Face mask	0.82	0.75	0.88
Coronavirus vaccine	0.3	0.11	0.47
COVID-19 stimulus check	0.79	0.7	0.85

COVID-19 = coronavirus disease 2019.

included *COVID symptoms*, *coronavirus testing center*, *loss of smell*, and *face mask*. The 3 keywords with strong correlations when looking at the United States as a whole include "face mask," "Lysol," and "COVID stimulus check," which have R values of 0.88, 0.82, and 0.79 respectively. *COVID symptoms* had correlations ranging from 0.37 to 0.80, *coronavirus testing center* had correlations ranging from -0.06 to 0.63, *loss of smell* had correlations ranging from 0.02 to 0.76, and *face mask* had correlations

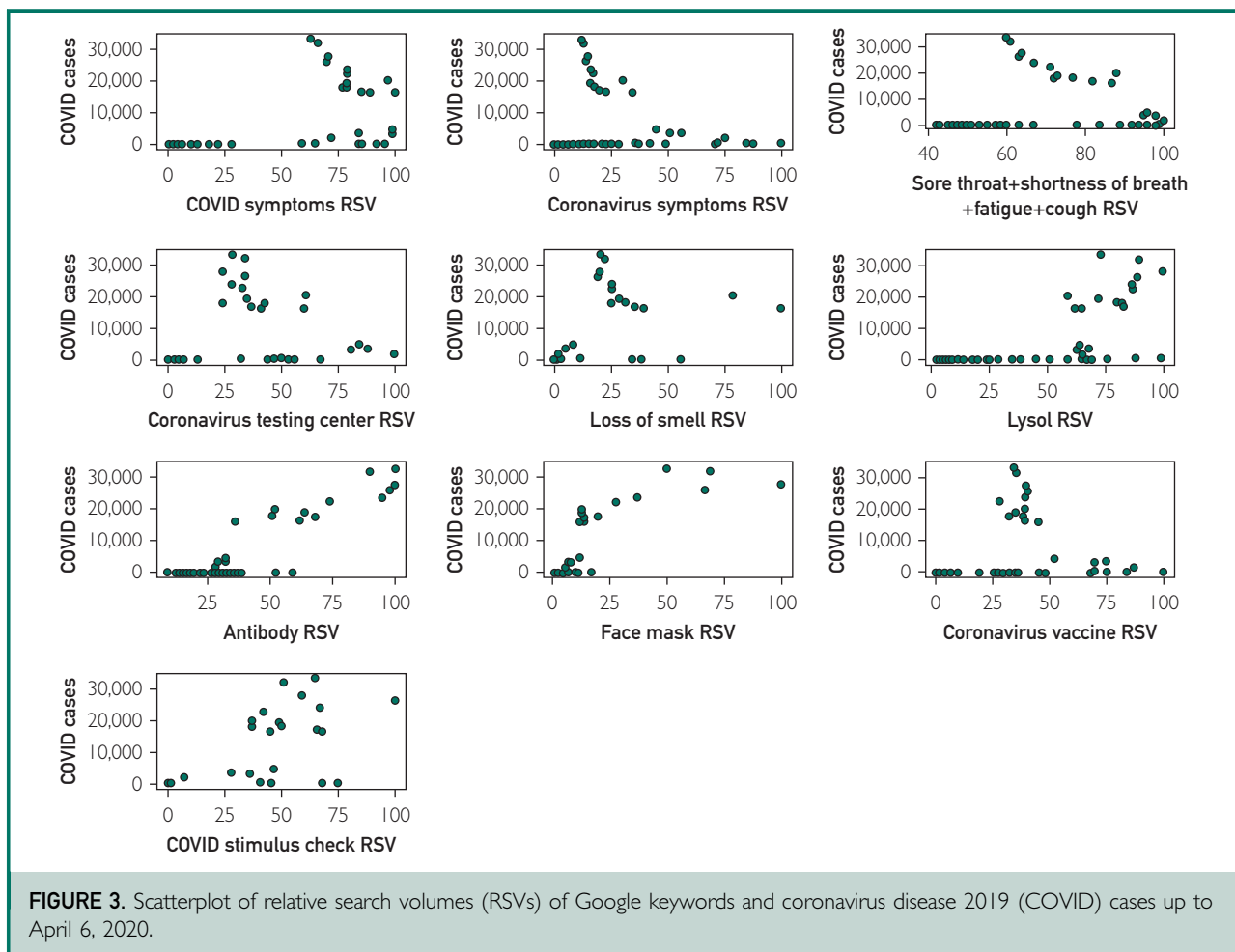
ranging from 0.35 to 0.90 (Supplemental Table 1, available online at <http://www.mayoclinicproceedings.org>). These correlations are further represented in Figure 2 as a United States heat map.

Search popularity for each keyword varied with COVID-19 case numbers. Some keywords such as *antibody* and *Lysol* had higher popularity as COVID-19 cases increased; other keywords such as *COVID symptoms* and *coronavirus vaccine* had higher popularity when COVID-19 case numbers were lower (Figure 3). To further assess this difference, lag and lead Pearson correlation coefficients were calculated for all 10 keywords and each of the 50 states, along with the United States as a whole. Lag correlations were calculated up to 16 days before the first case, and lead correlations were calculated up to 16 days after the first case. Most of the keywords had moderate to strong correlations days before the first COVID-19 cases appeared, with diminishing correlations following the first case (Figure 4). *Coronavirus symptoms*, for example, had its strongest correlations 16 days prior to the first case in the United States (R=0.77) and in most of the 50 states individually. All calculated lag and lead correlation coefficients for each of the 10 keywords and the 50 states, as well as the



United States overall, are displayed in [Supplemental Table 1](#) (available online at <http://www.mayoclinicproceedings.org>). When looking at Minnesota, Arizona, Florida, and New York, strong keyword correlations were seen up to 16 days prior to the first reported cases in each of these states. These 4 states are reported here individually because our institution (Mayo Clinic) has campuses in 3 (Minnesota, Arizona, and Florida) and New York was selected because it was the most strongly impacted area during the beginning of the pandemic in the United States. For Minnesota, the strongest correlations for *COVID symptoms*, *coronavirus symptoms*, *Lysol*, and *coronavirus vaccine* were seen on lag day 8 ($R=0.87$), lag day 14 ($R=0.85$), lag day 15 ($R=0.70$), and lag day 16 ($R=0.82$), respectively ([Table 2](#)). For

Arizona, the strongest correlations for *COVID symptoms*, *coronavirus symptoms*, *sore throat + shortness of breath + fatigue + cough*, *loss of smell*, *Lysol*, *coronavirus vaccine*, and *COVID stimulus check* were seen on lag day 9 ($R=0.80$), lag day 16 ($R=0.82$), lag day 11 ($R=0.73$), lag day 3 ($R=0.66$), lag day 1 ($R=0.73$), lag day 14 ($R=0.69$), and lag day 2 ($R=0.84$), respectively ([Table 3](#)). For Florida, nearly every keyword had strong correlations prior to the first case in the state; the strongest correlations for *COVID symptoms*, *coronavirus symptoms*, *loss of smell*, and *coronavirus vaccine* were seen on lag days 10 and 11 ($R=0.74$), lag day 16 ($R=0.78$), lag day 8 ($R=0.70$), and lag day 15 ($R=0.75$), respectively ([Table 4](#)). For New York, the strongest correlations for *COVID symptoms*,



coronavirus symptoms, coronavirus testing center, loss of smell, and coronavirus vaccine were seen on lag days 5, 6, and 7 ($R=0.87$), lag day 16 ($R=0.87$), lag day 9 ($R=0.76$), lag days 2, 4, and 5 ($R=0.78$), and lag day 15 ($R=0.80$), respectively (Table 5).

DISCUSSION

Our study found moderate to strong correlations between data obtained from searching COVID-19– related keywords in Google Trends and total COVID-19 cases in the United States as obtained from national data aggregators. Strong correlations were seen up to 16 days prior to the first reported cases in some states. This finding emphasizes the importance of digital surveillance and suggests that it can be a useful addition to our toolbelt when trying to monitor new infectious disease outbreaks.

Over the years, several studies have pointed to the role of Internet surveillance

in helping with early prediction of other infectious disease outbreaks, including diseases such as dengue fever, Zika virus, H1N1, influenza, measles, and MERS.^{5,6,11-14} There are several benefits to utilizing Internet surveillance methods vs traditional methods, and employing a combination of the two is likely the key to an effective surveillance system. One benefit to an Internet model is minimal costs because all of the data gathered from Google Trends were available free. Furthermore, the data are made available to the public in real time, with near-instant updates in regard to search results. This factor is extremely important when attempting to predict outbreaks and new hotspots for a pandemic because any delay in information could potentially miss the “golden window” that would allow for preparation prior to an outbreak in a certain location. Several other articles focusing on influenza have emphasized the pitfalls of traditional surveillance and how the US Centers for Disease Control

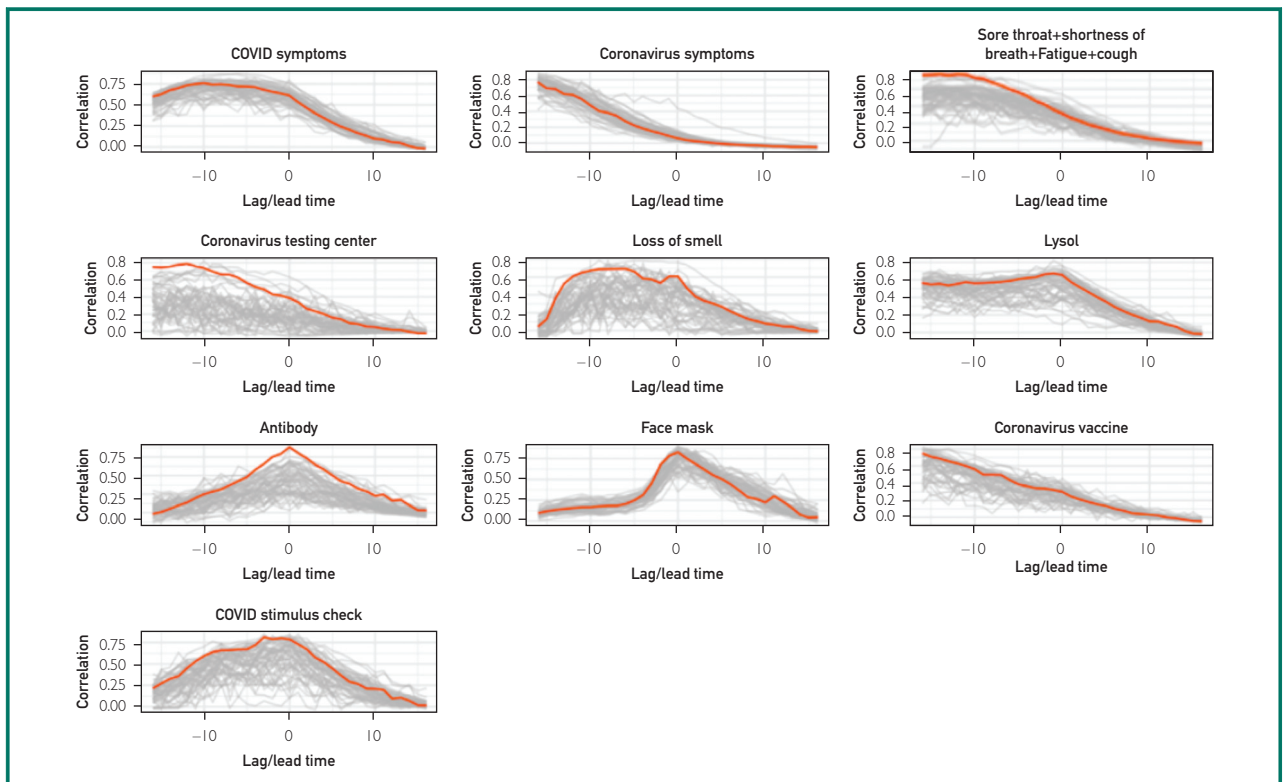


FIGURE 4. Lag/lead correlation coefficients of relative search volumes of Google keywords and new coronavirus disease 2019 (COVID) cases from January 22, 2020, to April 6, 2020. Red line = United States overall; gray lines = each of the 50 states.

TABLE 2. Minnesota Lag and Lead Correlation Coefficients for Each Google Keyword's Relative Search Volume and New COVID-19 Cases^{a,b}

Day	R (1)	R (2)	R (3)	R (4)	R (5)	R (6)	R (7)	R (8)	R (9)	R (10)
-16	0.60	0.83	0.58	0.24	-0.05	0.66	0.10	0.08	0.82	0.13
-15	0.65	0.83	0.62	0.26	-0.05	0.70	0.11	0.11	0.79	0.17
-14	0.71	0.85	0.64	0.31	0.00	0.66	0.14	0.12	0.78	0.19
-13	0.76	0.80	0.62	0.24	0.17	0.57	0.18	0.13	0.80	0.26
-12	0.79	0.75	0.62	0.19	0.23	0.60	0.18	0.14	0.75	0.34
-11	0.83	0.75	0.63	0.23	0.28	0.63	0.19	0.14	0.66	0.37
-10	0.85	0.70	0.62	0.26	0.26	0.57	0.22	0.15	0.69	0.42
-9	0.86	0.63	0.58	0.17	0.32	0.67	0.22	0.17	0.65	0.43
-8	0.87	0.57	0.57	0.28	0.37	0.67	0.22	0.18	0.59	0.55
-7	0.87	0.53	0.54	0.37	0.41	0.56	0.32	0.19	0.61	0.55
-6	0.86	0.47	0.53	0.26	0.51	0.56	0.39	0.20	0.59	0.58
-5	0.84	0.40	0.47	0.26	0.48	0.63	0.38	0.23	0.48	0.62
-4	0.83	0.36	0.43	0.18	0.43	0.59	0.42	0.26	0.46	0.62
-3	0.80	0.29	0.40	0.20	0.50	0.53	0.47	0.34	0.46	0.73
-2	0.77	0.24	0.39	0.27	0.42	0.59	0.48	0.50	0.39	0.80
-1	0.71	0.19	0.33	0.26	0.32	0.55	0.52	0.63	0.36	0.81
0	0.71	0.16	0.31	0.36	0.46	0.46	0.64	0.71	0.37	0.81
1	0.65	0.12	0.28	0.21	0.50	0.45	0.66	0.66	0.30	0.81
2	0.59	0.07	0.22	0.16	0.36	0.42	0.59	0.64	0.25	0.80
3	0.53	0.05	0.18	0.17	0.35	0.35	0.55	0.62	0.22	0.69
4	0.48	0.02	0.15	0.24	0.32	0.34	0.55	0.64	0.18	0.72
5	0.42	0.00	0.13	0.16	0.21	0.34	0.50	0.66	0.16	0.65
6	0.37	-0.01	0.08	0.12	0.27	0.27	0.46	0.62	0.15	0.62
7	0.34	-0.03	0.07	0.17	0.23	0.20	0.50	0.58	0.13	0.57
8	0.28	-0.04	0.03	0.11	0.23	0.19	0.42	0.52	0.07	0.50
9	0.22	-0.05	0.00	0.12	0.14	0.15	0.32	0.48	0.06	0.44
10	0.20	-0.05	-0.03	0.04	0.15	0.11	0.33	0.41	0.05	0.33
11	0.16	-0.06	-0.02	0.08	0.11	0.10	0.29	0.38	0.02	0.39
12	0.11	-0.06	-0.06	0.11	0.05	0.09	0.18	0.39	0.01	0.29
13	0.09	-0.06	-0.07	0.05	0.06	0.05	0.19	0.34	0.01	0.25
14	0.08	-0.07	-0.09	0.05	0.08	0.02	0.22	0.28	-0.01	0.23
15	0.04	-0.07	-0.10	0.02	0.02	0.01	0.10	0.19	-0.03	0.15
16	0.01	-0.07	-0.12	0.01	0.03	-0.02	0.05	0.14	-0.04	0.11

^aCOVID-19 = coronavirus disease 2019.

^bGoogle keywords: (1) = COVID symptoms; (2) = coronavirus symptoms; (3) = sore throat + shortness of breath + fatigue + cough; (4) = coronavirus testing center; (5) = loss of smell; (6) = Lysol; (7) = antibody; (8) = face mask; (9) = coronavirus vaccine; (10) = COVID stimulus check.

and Prevention surveillance reports were often weeks behind search engine results and estimates because traditional systems take 1 to 2 weeks to gather and process surveillance data.^{5,13}

This type of lag was further supported in our study of COVID-19, as Google data on

search trends predated the first reports of cases on a state-by-state basis. In a study on MERS reported in 2016, Shin et al⁶ found a similar lag pattern, with social media and search engine data reflecting disease outbreak earlier than conventional surveillance models. Scientists in China also looked

TABLE 3. Arizona Lag and Lead Correlation Coefficients for Each Google Keyword's Relative Search Volume and New COVID-19 Cases^{a,b}

Day	R (1)	R (2)	R (3)	R (4)	R (5)	R (6)	R (7)	R (8)	R (9)	R (10)
-16	0.57	0.82	0.60	0.49	-0.07	0.44	0.09	0.09	0.65	0.25
-15	0.60	0.80	0.68	0.38	-0.07	0.46	0.03	0.11	0.68	0.30
-14	0.66	0.77	0.71	0.30	0.06	0.45	0.03	0.13	0.69	0.30
-13	0.69	0.71	0.72	0.38	0.15	0.46	0.04	0.14	0.64	0.33
-12	0.71	0.67	0.71	0.28	0.19	0.46	0.00	0.15	0.63	0.43
-11	0.73	0.61	0.73	0.23	0.33	0.49	-0.01	0.16	0.64	0.44
-10	0.77	0.55	0.70	0.21	0.48	0.55	0.04	0.17	0.57	0.43
-9	0.80	0.49	0.68	0.19	0.50	0.54	0.12	0.18	0.52	0.53
-8	0.79	0.42	0.68	0.13	0.51	0.62	0.10	0.20	0.54	0.52
-7	0.78	0.37	0.63	0.09	0.59	0.63	0.06	0.21	0.49	0.58
-6	0.79	0.30	0.54	0.07	0.54	0.63	0.09	0.22	0.47	0.58
-5	0.76	0.26	0.50	0.14	0.50	0.61	0.12	0.26	0.44	0.61
-4	0.78	0.21	0.45	0.13	0.59	0.63	0.13	0.32	0.39	0.57
-3	0.76	0.17	0.39	0.07	0.66	0.65	0.19	0.46	0.35	0.70
-2	0.72	0.13	0.37	0.17	0.48	0.66	0.25	0.66	0.35	0.84
-1	0.72	0.10	0.33	0.16	0.49	0.73	0.38	0.78	0.32	0.74
0	0.68	0.07	0.31	0.15	0.49	0.71	0.50	0.85	0.29	0.75
1	0.62	0.05	0.25	0.09	0.39	0.60	0.54	0.79	0.29	0.69
2	0.53	0.03	0.24	0.12	0.37	0.56	0.46	0.72	0.24	0.63
3	0.47	0.02	0.18	0.12	0.35	0.52	0.40	0.67	0.18	0.52
4	0.41	0.00	0.15	0.05	0.26	0.43	0.42	0.61	0.17	0.52
5	0.35	0.00	0.14	0.13	0.21	0.40	0.39	0.60	0.14	0.45
6	0.30	-0.01	0.10	0.07	0.18	0.35	0.32	0.49	0.09	0.35
7	0.25	-0.02	0.08	0.04	0.16	0.28	0.34	0.46	0.10	0.35
8	0.19	-0.02	0.05	0.08	0.10	0.22	0.26	0.39	0.06	0.27
9	0.16	-0.03	0.04	0.05	0.08	0.19	0.23	0.33	0.03	0.21
10	0.13	-0.04	0.03	0.03	0.08	0.14	0.22	0.29	0.03	0.19
11	0.08	-0.04	0.01	0.03	0.04	0.11	0.16	0.24	0.00	0.15
12	0.06	-0.04	0.00	0.03	0.02	0.08	0.14	0.19	-0.02	0.11
13	0.04	-0.05	-0.01	0.01	0.02	0.05	0.13	0.15	-0.03	0.08
14	0.01	-0.05	-0.02	0.00	0.01	0.02	0.10	0.10	-0.04	0.05
15	-0.01	-0.05	-0.02	-0.01	0.00	0.00	0.07	0.05	-0.05	0.02
16	-0.02	-0.06	-0.03	-0.01	-0.01	-0.02	0.06	0.02	-0.06	0.00

^aCOVID-19 = coronavirus disease 2019.

^bGoogle keywords: (1) = COVID symptoms; (2) = coronavirus symptoms; (3) = sore throat + shortness of breath + fatigue + cough; (4) = coronavirus testing center; (5) = loss of smell; (6) = Lysol; (7) = antibody; (8) = face mask; (9) = coronavirus vaccine; (10) = COVID stimulus check.

for this data lag with COVID-19 in their country and had similar results.¹⁵ They looked back 14 days prior to the first reported cases and found that “the peak Internet searches and social media data about the COVID-19 outbreak occurred 10-14 days earlier than the peak of daily incidences in China.”¹⁵

We suspect that our US data reveal similar lags in traditional surveillance data for a number of reasons. First, hospital reporting can vary from state to state and even county to county. Although we try to standardize reporting guidelines, during a time of a pandemic when hospital systems and the country are becoming increasingly

TABLE 4. Florida Lag and Lead Correlation Coefficients for Each Google Keyword's Relative Search Volume and New COVID-19 Cases^{a,b}

Day	R (1)	R (2)	R (3)	R (4)	R (5)	R (6)	R (7)	R (8)	R (9)	R (10)
-16	0.55	0.78	0.56	0.53	-0.03	0.55	0.09	0.09	0.70	0.17
-15	0.62	0.74	0.59	0.56	0.02	0.56	0.09	0.11	0.75	0.28
-14	0.68	0.74	0.60	0.59	0.16	0.55	0.13	0.12	0.74	0.21
-13	0.70	0.66	0.58	0.56	0.26	0.51	0.12	0.14	0.63	0.27
-12	0.71	0.58	0.59	0.62	0.42	0.52	0.19	0.15	0.63	0.37
-11	0.74	0.53	0.58	0.57	0.47	0.49	0.21	0.16	0.59	0.38
-10	0.74	0.48	0.53	0.70	0.68	0.48	0.22	0.15	0.55	0.45
-9	0.73	0.41	0.54	0.56	0.52	0.50	0.29	0.17	0.52	0.49
-8	0.73	0.35	0.50	0.68	0.70	0.50	0.32	0.18	0.49	0.62
-7	0.73	0.30	0.48	0.55	0.57	0.52	0.33	0.19	0.45	0.53
-6	0.70	0.26	0.46	0.57	0.67	0.52	0.43	0.21	0.42	0.58
-5	0.67	0.23	0.44	0.57	0.60	0.53	0.45	0.24	0.39	0.61
-4	0.65	0.18	0.42	0.54	0.55	0.52	0.52	0.27	0.32	0.61
-3	0.62	0.15	0.38	0.54	0.58	0.54	0.58	0.36	0.34	0.69
-2	0.63	0.12	0.37	0.52	0.48	0.58	0.65	0.52	0.28	0.68
-1	0.60	0.10	0.35	0.55	0.48	0.60	0.60	0.62	0.28	0.73
0	0.60	0.07	0.33	0.52	0.47	0.57	0.69	0.74	0.28	0.79
1	0.54	0.05	0.29	0.46	0.38	0.60	0.69	0.80	0.23	0.78
2	0.50	0.03	0.26	0.53	0.40	0.53	0.61	0.79	0.22	0.61
3	0.43	0.02	0.23	0.37	0.34	0.44	0.58	0.64	0.18	0.67
4	0.37	0.01	0.17	0.41	0.31	0.41	0.54	0.65	0.14	0.61
5	0.29	0.00	0.13	0.27	0.21	0.32	0.37	0.52	0.11	0.39
6	0.27	-0.01	0.11	0.27	0.22	0.29	0.38	0.48	0.10	0.45
7	0.22	-0.02	0.07	0.24	0.19	0.23	0.30	0.42	0.07	0.36
8	0.16	-0.02	0.04	0.15	0.12	0.18	0.21	0.33	0.05	0.28
9	0.14	-0.03	0.02	0.16	0.10	0.15	0.20	0.30	0.04	0.23
10	0.09	-0.03	0.00	0.10	0.07	0.10	0.14	0.20	0.02	0.17
11	0.07	-0.03	-0.02	0.08	0.06	0.07	0.10	0.15	0.00	0.14
12	0.04	-0.04	-0.04	0.04	0.03	0.04	0.05	0.10	-0.01	0.09
13	0.03	-0.04	-0.06	0.03	0.01	0.02	0.03	0.09	-0.02	0.07
14	0.02	-0.04	-0.07	0.02	0.01	0.01	0.02	0.08	-0.03	0.06
15	0.00	-0.05	-0.08	0.01	0.00	0.00	-0.01	0.06	-0.04	0.04
16	-0.01	-0.05	-0.09	0.00	-0.01	-0.02	-0.02	0.04	-0.04	0.03

^aCOVID-19 = coronavirus disease 2019.

^bGoogle keywords: (1) = COVID symptoms; (2) = coronavirus symptoms; (3) = sore throat + shortness of breath + fatigue + cough; (4) = coronavirus testing center; (5) = loss of smell; (6) = Lysol; (7) = antibody; (8) = face mask; (9) = coronavirus vaccine; (10) = COVID stimulus check.

stressed, appropriate reporting can break down. In fact, inappropriate reporting can lead to significant inaccuracies when data is released using traditional surveillance models. For example, on April 17, 2020, China raised its coronavirus death toll in Wuhan by 50% in comparison with their

previously reported numbers.¹⁶ A second important source of data lag using traditional surveillance in the United States is the lack of testing required for the current pandemic. Testing is evolving on a day-by-day basis, and, thankfully, we are moving in the right direction; however, the United States and

TABLE 5. New York Lag and Lead Correlation Coefficients for Each Google Keyword's Relative Search Volume and New COVID-19 Cases^{a,b}

Day	R (1)	R (2)	R (3)	R (4)	R (5)	R (6)	R (7)	R (8)	R (9)	R (10)
-16	0.56	0.87	0.64	0.49	0.05	0.61	0.11	0.20	0.78	0.16
-15	0.63	0.86	0.67	0.59	0.10	0.63	0.13	0.22	0.80	0.23
-14	0.69	0.85	0.69	0.59	0.21	0.64	0.15	0.25	0.79	0.28
-13	0.73	0.81	0.70	0.62	0.36	0.65	0.18	0.26	0.78	0.33
-12	0.78	0.78	0.71	0.69	0.47	0.66	0.20	0.26	0.76	0.41
-11	0.80	0.74	0.71	0.72	0.55	0.66	0.22	0.27	0.75	0.50
-10	0.81	0.68	0.71	0.71	0.60	0.65	0.30	0.28	0.72	0.52
-9	0.84	0.64	0.72	0.76	0.64	0.68	0.33	0.29	0.68	0.55
-8	0.86	0.57	0.72	0.74	0.74	0.68	0.37	0.31	0.64	0.61
-7	0.87	0.53	0.72	0.75	0.76	0.70	0.42	0.32	0.61	0.62
-6	0.87	0.47	0.71	0.74	0.77	0.70	0.44	0.34	0.58	0.65
-5	0.87	0.40	0.68	0.69	0.78	0.71	0.50	0.37	0.56	0.64
-4	0.86	0.35	0.65	0.69	0.78	0.70	0.54	0.41	0.52	0.66
-3	0.85	0.29	0.62	0.66	0.77	0.72	0.56	0.52	0.49	0.77
-2	0.81	0.24	0.59	0.70	0.78	0.73	0.65	0.65	0.46	0.80
-1	0.79	0.19	0.56	0.67	0.76	0.75	0.70	0.77	0.43	0.81
0	0.76	0.15	0.53	0.63	0.77	0.76	0.76	0.83	0.41	0.84
1	0.70	0.11	0.48	0.59	0.74	0.72	0.74	0.80	0.36	0.78
2	0.63	0.08	0.43	0.52	0.68	0.66	0.70	0.74	0.32	0.75
3	0.56	0.06	0.37	0.46	0.60	0.60	0.64	0.69	0.28	0.69
4	0.49	0.04	0.32	0.41	0.53	0.54	0.59	0.63	0.24	0.59
5	0.44	0.02	0.27	0.34	0.48	0.49	0.53	0.59	0.21	0.54
6	0.38	0.00	0.23	0.32	0.41	0.44	0.50	0.55	0.18	0.51
7	0.33	-0.01	0.18	0.28	0.34	0.39	0.43	0.51	0.15	0.44
8	0.28	-0.02	0.15	0.23	0.29	0.34	0.39	0.49	0.13	0.39
9	0.23	-0.03	0.11	0.20	0.24	0.30	0.35	0.43	0.10	0.33
10	0.19	-0.04	0.07	0.16	0.20	0.25	0.29	0.38	0.08	0.30
11	0.15	-0.04	0.04	0.13	0.15	0.20	0.24	0.32	0.05	0.25
12	0.11	-0.05	0.01	0.11	0.11	0.15	0.18	0.27	0.03	0.20
13	0.08	-0.06	-0.02	0.07	0.09	0.12	0.14	0.23	0.01	0.16
14	0.05	-0.06	-0.05	0.06	0.06	0.08	0.10	0.18	-0.01	0.13
15	0.02	-0.06	-0.07	0.03	0.04	0.04	0.05	0.13	-0.03	0.10
16	0.00	-0.07	-0.09	0.02	0.02	0.01	0.02	0.08	-0.04	0.06

^aCOVID-19 = coronavirus disease 2019.

^bGoogle keywords: (1) = COVID symptoms; (2) = coronavirus symptoms; (3) = sore throat + shortness of breath + fatigue + cough; (4) = coronavirus testing center; (5) = loss of smell; (6) = Lysol; (7) = antibody; (8) = face mask; (9) = coronavirus vaccine; (10) = COVID stimulus check.

the world still have a ways to go. Testing capabilities were sparse at the beginning of the US outbreaks, and many areas were backlogged in their abilities to test for COVID-19. Even if patient samples were available, the time to test that sample and report the diagnosis back to the physician and patient

were delayed because testing capabilities were not robust. This issue, of course, results in a delay in reported cases and is where Internet surveillance could add value. As the pandemic continues to evolve, the need for quicker testing and an increase in the quantity of testing for COVID-19 is

paramount. In an article by Gottlieb et al¹⁷ regarding the reopening of the United States, the authors stated, “We estimate that a national capacity of at least 750,000 tests per week would be sufficient. In conjunction with more widespread testing, we need to invest in new tools to make it efficient for providers to communicate test results and make data easily accessible to public-health officials working to contain future outbreaks.” Data accessibility and speed of communication are key; search engine surveillance meets both of these criteria and thus provides important up-to-date information while traditional models catch up.

It is important to note that our study looked at 10 keywords, and each had varying strengths of correlation with case numbers. If we had looked at 100 keywords, even stronger correlations may have been found. Search terms will also evolve as a pandemic progresses. Furthermore, Google itself is widely used in the United States, which makes it a good candidate for digital surveillance, but this is not the case for every country. For example, Google is not a major search engine in China.¹⁵ It would be important to utilize sites relevant to each country when developing predictive models, and using multiple sites could further improve predictions. Shin et al⁶ utilized Google and Twitter when conducting their study on MERS and found strong correlations using both sites. One other limitation of Google Trends is the granularity it provides. Although it does provide information on some cities, it does not currently provide a comprehensive town-by-town breakdown of its data. This issue would make it difficult to create appropriate forecast models on a town-by-town basis, and individuals would have to rely on broader state-wide predictions.

CONCLUSION

This study reveals the benefits of internet surveillance models and the use of Google Trends to monitor new infectious diseases such as COVID-19. For the United States, Google Trends data were highly correlated with cases of COVID-19 on a state-by-state basis and could potentially be used to predict

new areas of outbreak and possible high-impact zones as the disease progresses. Furthermore, this study documents that there is information present in Google Trends that precedes outbreaks, and these data should be utilized to allow for better resource allocation in regard to tests, personal protective equipment, medication, and more.

SUPPLEMENTAL ONLINE MATERIAL




Supplemental material can be found online at <http://www.mayoclinicproceedings.org>. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: COVID-19 = coronavirus disease 2019; MERS = Middle East respiratory syndrome; RSV = relative search volume

Potential Competing Interests: The authors report no competing interests.

Correspondence: Address to Mohamad Bydon, MD, Department of Neurologic Surgery, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (Bydon.mohamad@mayo.edu; Twitter: @MayoNeuroInfo).

ORCID

Atiq ur Rehman Bhatti:  <https://orcid.org/0000-0002-2500-0192>; Mohammed Ali Alvi:  <https://orcid.org/0000-0002-7131-079X>; Mohamad Bydon:  <https://orcid.org/0000-0002-0543-396X>

REFERENCES

1. Lu H, Stratton CW, Tang Y-W. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. *J Med Virol*. 2020;92(4):401-402.
2. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [published correction appears in *Lancet*. 2020;395(10223):496]. *Lancet*. 2020;395(10223):497-506.
3. Otter JA. What's trending in the infection prevention and control literature? from HIS 2012 to HIS 2014, and beyond. *J Hosp Infect*. 2015;89(4):229-236.
4. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res*. 2009;11(1):e11.
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012-1014.
6. Shin S-Y, Seo D-W, An J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Sci Rep*. 2016;6:32920.
7. Wang H-W, Chen D-R, Yu H-W, Chen Y-M. Forecasting the incidence of dementia and dementia-related outpatient visits

- with Google Trends: evidence from Taiwan. *J Med Internet Res*. 2015;17(11):e264.
8. Bragazzi NL, Alicino C, Trucchi C, et al. Global reaction to the recent outbreaks of Zika virus: Insights from a Big Data analysis. *PLoS One*. 2017;12(9):e0185263.
 9. COVID-19 United States Cases by Country. <https://coronavirus.jhu.edu/us-map>. Accessed April 20, 2020.
 10. Google. FAQ about Google Trends data. https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052. Accessed July 4, 2020.
 11. Majumder MS, Santillana M, Mekanu SR, McGinnis DP, Khan K, Brownstein JS. Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 Colombian Zika virus disease outbreak. *JMIR Public Health Surveill*. 2016;2(1):e30.
 12. Marques-Toledo CA, Degener CM, Vinhal L, et al. Dengue prediction by the web: tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLoS Negl Trop Dis*. 2017;11(7):e0005729.
 13. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLoS One*. 2011;6(4):e18687.
 14. Santangelo OE, Provenzano S, Piazza D, Giordano D, Calamusa G, Firenze A. Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy. *Ann Ig*. 2019;31(4):385-391.
 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. *Euro Surveill*. 2020;25(10):2000199.
 16. Qin A. China raises coronavirus death toll by 50% in Wuhan. *New York Times* website. <https://www.nytimes.com/2020/04/17/world/asia/china-wuhan-coronavirus-death-toll.html>. Published April 17, 2020. Accessed April 21, 2020.
 17. Gottlieb S, Rivers C, McClellan M, Silvis L, Watson C. National coronavirus response: a road map to reopening. American Enterprise Institute website. Published March 29, 2020. <https://www.aei.org/research-products/report/national-coronavirus-response-a-road-map-to-reopening/>. Accessed April 20, 2020.