JMIR Publications
The Leading eHealth Publisher

Regional Infoveillance of COVID-19 Case Rates: Analysis of Search-Engine Query Patterns

Henry Cousins, Clara Cousins, Alon Harris, Louis Pasquale

Submitted to: Journal of Medical Internet Research
on: April 20, 2020

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Regional Infoveillance of COVID-19 Case Rates: Analysis of Search-Engine Query Patterns

Henry CousinsMPhil, ; Clara Cousins; Alon HarrisMS, PhD, ; Louis PasqualeMD,

Corresponding Author:
Louis PasqualeMD,
Phone: +1212-979-4500
Email: louis.pasquale@mssm.edu

Abstract

Background: Timely allocation of medical resources for COVID-19 requires early detection of regional outbreaks. Internet browsing data, such as search activity levels, may provide predictive ability for estimating cases in a local population that are yet to be confirmed.

Objective: The objective of our study was to determine whether search-engine query patterns can forecast COVID-19 case rates at the state and local levels in the United States.

Methods: We used regional confirmed case data from the New York Times and Google Trends results from 50 states and 203 county-based designated market areas (DMA). We identified search terms whose activity precedes and correlates with confirmed case rates at the national level, using univariate regression to construct a composite explanatory variable based on top-scoring search queries offset by temporal lags. We measured the correlation of the explanatory variable with out-of-sample case rate data at the state and DMA level.

Results: Forecasts were highly correlated with confirmed case rates at the state and local level, using search data available up to 10 days in advance of confirmed case rates. They predicted case activity in 49 of 50 states and in 128 of 203 DMA at a significance level of .05 and were robust to differences in regional location, population, and date of outbreak.

Conclusions: Identifiable patterns in search query activity may be used to forecast emerging regional outbreaks of COVID-19.

(JMIR Preprints 20/04/2020:19483)
DOI: https://doi.org/10.2196/preprints.19483

Preprint Settings

Original Manuscript

Regional Infoveillance of COVID-19 Case Rates: Analysis of Search-Engine Query Patterns

Henry C. Cousins, MPhil[1]; Clara C. Cousins, BA[2,3,4]; Alon Harris, MS, PhD, FARVO[5]; Louis R. Pasquale, MD, FARVO[5]

[1] Department of Genetics, Stanford School of Medicine, Stanford, CA
[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA
[3] Department of Data Sciences, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA
[4] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA
[5] Department of Ophthalmology, Icahn School of Medicine at Mount Sinai, New York, NY

Corresponding author
Louis R. Pasquale, MD, FARVO
Professor of Ophthalmology
Icahn School of Medicine at Mount Sinai
One Gustave L. Levy Place
Box 1183
New York, NY 10029
Louis.Pasquale@mssm.edu
212-241-6752

Abstract

Background: Timely allocation of medical resources for COVID-19 requires early detection of regional outbreaks. Internet browsing data may predict case outbreaks in local populations that are yet to be confirmed.

Objective: We investigated whether search-engine query patterns can help to predict COVID-19 case rates at the state and metropolitan-area level in the United States.

Methods: We used regional confirmed case data from the New York Times and Google Trends results from 50 states and 166 county-based designated market areas (DMA). We identified search terms whose activity precedes and correlates with confirmed case rates at the national level. We used univariate regression to construct a composite explanatory variable based on best-fitting search queries offset by temporal lags. We measured the raw and z-transformed Pearson correlation and root-mean-square error (RMSE) of the explanatory variable with out-of-sample case rate data at the state and DMA level.

Results: Predictions were highly correlated with confirmed case rates at the state (mean r = .69; 95% confidence interval (CI): .51-.81; median RMSE 1.27; interquartile range (IQR) 1.48) and DMA level (mean r = .51, 95% CI .39-.61; median RMSE 4.38; IQR: 1.80), using search data available up to 10 days prior to confirmed case rates. They fit case-rate activity in 49 of 50 states and in 103 of 166 DMA at a significance level of .05.

Conclusions: Identifiable patterns in search query activity may help to predict emerging regional outbreaks of COVID-19, although they remain vulnerable to stochastic changes in search intensity.

Keywords: epidemiology, COVID-19, internet activity, Google Trends, disease surveillance, public health

Introduction

Early detection of regional COVID-19 outbreaks is essential for efficient medical resource allocation, public health messaging, and implementation of infection prevention and control strategies [1]. It is particularly important given the probability of future waves of COVID-19 cases and the difficulty of applying traditional epidemiological forecasting models in areas with low case levels [2,3]. However, laboratory testing capacity is limited, and confirmed case reports lag behind underlying infections, decreasing their predictive capacity in the early days of an outbreak or resurgence.

Internet browsing data, such as search-engine query results, can provide a real-time indication of symptoms in a population and have been used extensively to predict and model outbreaks like influenza and dengue [4-7]. Such methods generally assume that specific and detectable patterns in internet behavior, such as search trends or social media postings, reflect health-seeking behavior in real time at the population level. Forecasting models based on search queries, such as Google Flu Trends, have shown predictive value without direct reliance on formal case reports, although historical inaccuracies mean that they can only supplement, not replace, traditional forecasting methodologies based on confirmed cases [8-10].

COVID-19 case rates display significant regional heterogeneity that require locally tailored containment strategies. Google search trends, encompassing a majority of internet queries in the United States and publicly available through Google Trends (GT), provide a powerful resource for systematic comparison of browsing behavior between US regions. We hypothesized that keyword libraries could be screened for specific terms whose aggregate activity would reflect regional differences in COVID-19 case rates, as has been demonstrated for influenza [4]. While several studies have previously attempted to model the COVID-19 pandemic using search query data, such attempts have largely focused on specific regions, like Taiwan and Iran, and a limited number of individually selected search terms [11-14]. We explored the potential of large-scale, publicly accessible search query data to signal new COVID-19 cases at the state and metropolitan-area levels in the US.

Methods

Data collection and processing

We obtained confirmed case data for US states and counties from the New York Times (NYT) dataset from January 21, the date of the first confirmed US case, to April 2, 2020, comprising county-specific, lab-confirmed COVID-19 case reports compiled daily from local and state health authorities [15]. We used the NYT dataset because of both its inclusion of county-level case geotags and its strong correlation with other case tracking sources [16]. Next, we used GT to compile a library of 463 unique search queries and their associated daily activity levels over the same time period. Library terms were automatically retrieved based on likelihood of user association with a set of prespecified coronavirus-related seed terms (Table S1), using the GT “Related Queries” function.

We compared the z-transformed correlation of each query’s search activity with an in-sample dataset comprising daily confirmed national cases rates for days through March 10 with >100 new cases per day. Each query’s search activity was offset by temporal lags of between 0 and 14 days, generating a list of best-fitting queries and their associated optimal lag times. To focus on terms with early predictive power, we excluded queries whose optimal lag was less than 9 days. We selected the five best-fitting queries and constructed a single explanatory variable by summing the lag-adjusted, relative activity levels of each query. Finally, we linearly fit the explanatory variable to national data through March 10 to generate a single scalar coefficient.

Data analysis

We measured the correlation of state-specific activity levels for our explanatory variable with daily reported case levels in individual states using out-of-sample data from March 11 through April 2. We also measured how well the explanatory variable explained out-of-sample case rates in 166 designated market areas (DMA), which are collections of approximately 15 counties each constituting the highest-resolution regional data available on GT. Means and confidence intervals for correlation coefficients were calculated using the inverse z-transformation of the averaged z-transformed coefficients. The strength of model predictions over time was measured using a partial correlation of first confirmed case dates with z-transformed correlation coefficients in all regions with >100 cases, controlling for regional population. We used root-mean-square error (RMSE) as an additional measure of model performance. Model predictions were adjusted for regional population and internet access [17]. All data were anonymous, and the study protocol was approved by the institutional review board of the Icahn School of Medicine at Mount Sinai.

Results

Search query characteristics

Queries incorporated into the final explanatory variable were highly correlated with national case data, with correlation coefficients ranging from .996 to .999 on the in-sample data. The optimal temporal lags for incorporated queries were from 11 to 12 days, indicating a prediction horizon of up to 10 days (assuming that a day’s full GT query results become available on the subsequent day). The final variable, the linear sum of weighted, lag-adjusted activity levels for the five best-fitting terms from the 463-term library, fit the in-sample data with a correlation of 0.998.

Characteristics of additional screened queries validated our methodology. For instance, acute topics like medical care and testing had smaller associated lag times with confirmed case rates, as would be expected for urgent inquiries (Table 1). Queries unrelated to COVID-19 had correspondingly weaker correlations with the observed data. The best-fitting category of queries was “COVID-19 guidance,” which included terms related to coronavirus-specific medical advice from health authorities. Relative levels of search activity had no significant effect on fit with case data.

Table 1. Characteristics of query topics screened for fit with COVID-19 case data.

Search query category	Number (%) of unique queries[a]	Mean correlation with national case rate[b]	Mean associated lag time (days)[c]	Activity weighting[d]
COVID-19 guidance	32 (6.9)	.96	9.1	.38
COVID-19 news	57 (12.3)	.96	8.3	1.00

COVID-19 symptoms	91 (19.7)	.94	8.9	.41
Medical treatments	34 (7.3)	.93	10.1	.31
COVID-19 testing	58 (12.5)	.89	5.4	.11
Medical care	33 (7.1)	.89	7.2	.60
Nonspecific symptoms	62 (13.4)	.89	6.8	.57
Economic effects	28 (6.0)	.86	5.9	.12
Unrelated to illness	51 (11.0)	.86	6.6	.76
Symptoms of other illnesses	17 (3.7)	.84	8.3	.77

[a] Number of queries of each type in the query library (for example, the category “COVID-19 testing” would include the specific query “coronavirus test near me”, and the category “nonspecific symptoms” would include the query “cough”)

[b] Expressed as the inverse z-transformation of the averaged z-transformed correlations with in-sample national data

[c] Mean lag time between best-fitting query activity and confirmed case rate, in days

[d] Relative mean search activity levels, normalized

Regional case-rate predictions

The query-based predictions fit well with out-of-sample case rate data at the national level, with a correlation of 0.84 (*P* < .001) for out-of-sample data and 0.83 (*P* < .001) for all available data. The predictions were also well correlated at the state level in nearly all cases, fitting case data in 49 of 50 states at a significance level of α = .05 and 41 of 50 states at α = .005 (mean r = .69; 95% confidence interval (CI): .51-.81; Figure 1A; Table S2). RMSE was less than 4 cases per 100,000 residents for model predictions in 44 of 50 states (median 1.27; interquartile range (IQR) 1.48; Figure 1B).

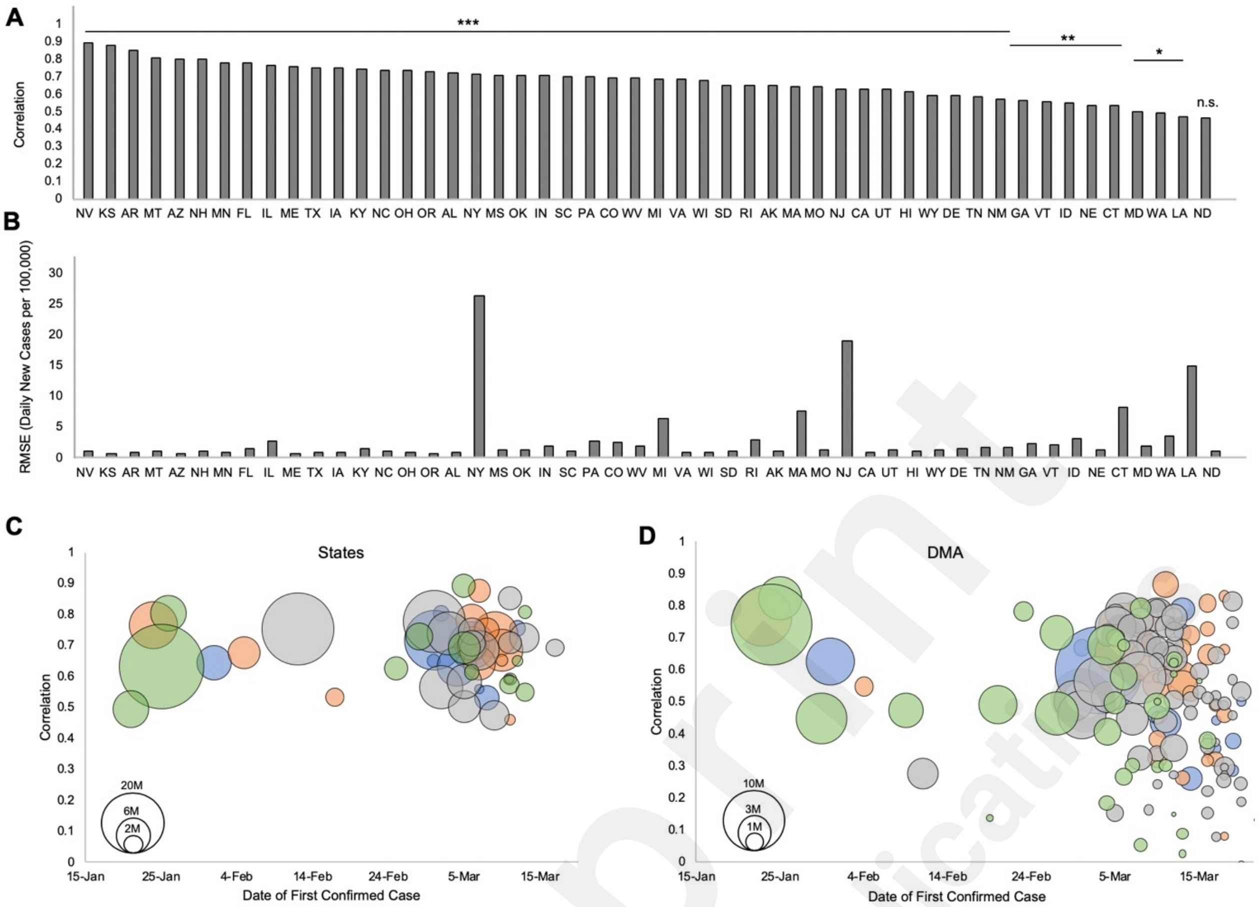


Figure 1. Correlation of query predictions with regional COVID-19 confirmed case rates

(A) Correlation of predicted case rates with actual case rates for the 50 states. Values are Pearson correlation coefficients. * indicates significance at α = .05; ** at α = .01; *** at α = .005.

(B) RMSE between predicted case rates and actual case rates for the 50 states, in units of daily new cases per 100,000 population.

(C) Prediction correlations at the state level do not depend on outbreak timing, as measured by date of first confirmed case. Circle size indicates relative population of state. Color indicates census-designated region of the US (blue: Northeast, orange: Midwest, gray: South, green: West).

(D) Prediction correlations at the DMA level do not depend on outbreak timing, as measured by date of first confirmed case. Circle size indicates relative population of DMA. Color indicates census-designated region of the US, as described.

At the DMA level, the query-based predictions fit with daily case data for 62% (103/166) of regions at α = .05, or 79% (84/107) excluding DMA with fewer than 100 cases (mean r for all DMA = .51; 95% CI: .39-.61; Table S3). RMSE was slightly higher for DMA-level compared to state-level predictions but was less than 7 for 92% (152/166) of DMA (median 4.38; IQR: 1.80). Furthermore, at both the state and DMA level, strength of correlation was not significantly associated with the date of a region's first confirmed case (*P* = .51 for states and .71 for DMA for partial correlations in regions with >100 cases, controlling for population), suggesting that predictive search behaviors may precede new cases regardless of the timing of a regional outbreak (Figure 1C,D). The explanatory variable consistently produced well-fitting predictions with data available 10 days in advance of predicted new case activity (Figure 2), even in regions where fewer than 100 new cases were confirmed per day.

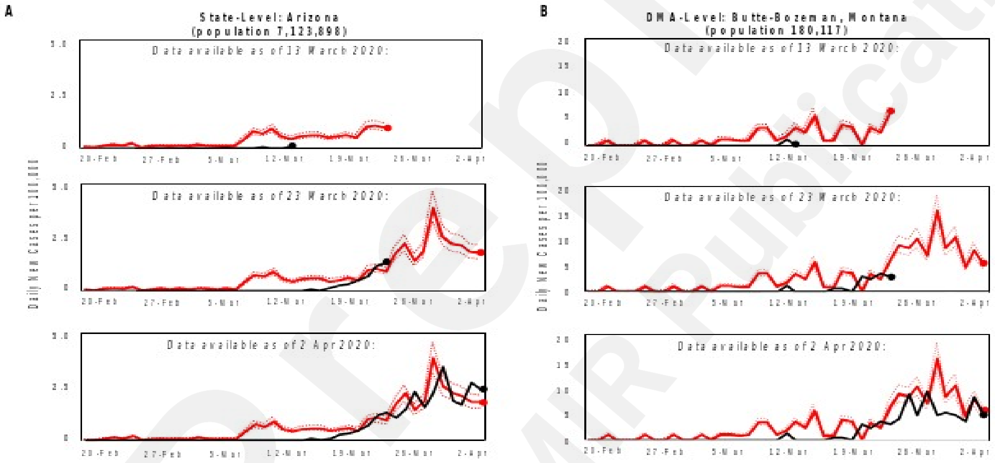


Figure 2. Correlation of query predictions (red) with regional COVID-19 case rates (black) at the state and DMA level, February 20–April 2, 2020
(A) Comparison of predicted case rates (red) with actual case rates (black) at the state level, with Arizona shown as an example. Dashed lines indicate 95% confidence intervals.
(B) Comparison at the DMA level, with the Butte-Bozeman area shown as an example of predictions in a low-population region.

Discussion

These data suggest that specific patterns of internet search behavior, which can be curated automatically from libraries of search terms, precede and correlate with regional case rates of COVID-19. Such patterns, which we capture using a single explanatory variable, remain correlated with case rates in regions with a broad range of populations, locations, and outbreak times, making aggregate search trends a useful tool for estimating regional COVID-19 outbreaks in the days preceding confirmed case reports. Correlation strengths were not significantly associated with the date of onset of regional outbreaks, making it unlikely that a single national event, such as a press release, could explain the strength of model predictions in all regions. Furthermore, search queries explicitly related to COVID-19 have more predictive power than unrelated keywords, and acute queries, such as those concerning testing or medical care, have smaller associated lag times.

Taken together, these results suggest that systematic screening of key term libraries can identify search queries reflecting real-time health-seeking behaviors at the regional level, expanding the suite of “infoveillance” methods that may assist in monitoring COVID-19 cases. This type of approach does not directly depend on either regional testing capacity or local media reports, making it particularly relevant in areas with small populations, limited medical infrastructure, or low case numbers. Such information can supplement traditional epidemiological approaches, such as estimates based on a compartmental framework, to guide community health interventions in the early days of an outbreak.

Several aspects of query-based approaches to case estimation, such as this work, must be further characterized for COVID-19. First, while correlations were statistically strong across most US regions, elevated RMSE indicated lower accuracy for predictions in the New York City and New Orleans areas, both regions with major outbreaks. However, comparable losses in accuracy were not observed for other major outbreak sites, such as Philadelphia, Los Angeles, or Chicago. This may reflect region-specific differences in both internet browsing behavior and patterns of community infection and may be a limitation of query-based models using fixed terms. As evidenced by previous attempts to predict influenza outbreaks based on search data, browsing behavior will also likely change as public understanding evolves over the course of disease spread [18]. Therefore, search-term relevance is likely to vary with time, which may require continuous supplementation or reselection of query terms to ensure representativeness of current population behaviors. Furthermore, although we observed strong historical correlations generally, query-based models must also be monitored for sudden changes in COVID-related query activity due to external events, such as unrelated news reports. Such distortions would be particularly important in regions with limited internet access. Future models incorporating learning and real-time updating of region-specific search terms may improve query-based prediction efforts for future COVID-19 outbreaks.

Acknowledgments

HCC, CCC, and LRP conceived of the study and designed the methodology. HCC and CCC conducted the data analysis. HCC, CCC, LRP, and AH wrote the manuscript. All authors met the ICMJE criteria for authorship. The corresponding author had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of interest

Unrelated to this work, AH would like to disclose that he receives remuneration from AdOM for serving as a consultant and a board member, and from Thea for a speaking engagement. AH also holds an ownership interest in AdOM, Luseed, Oxymap, and QuLent. Unrelated to this work, LRP is a consultant to Eyenovia, Bausch+Lomb, Nicox, Emerald Bioscience and Verily. No relevant financial relationship exists for any of the other authors.

Abbreviations

CI: confidence interval
DMA: designated market area
GT: Google Trends
IQR: interquartile range
NYT: New York Times
RMSE: root-mean-square error

References:

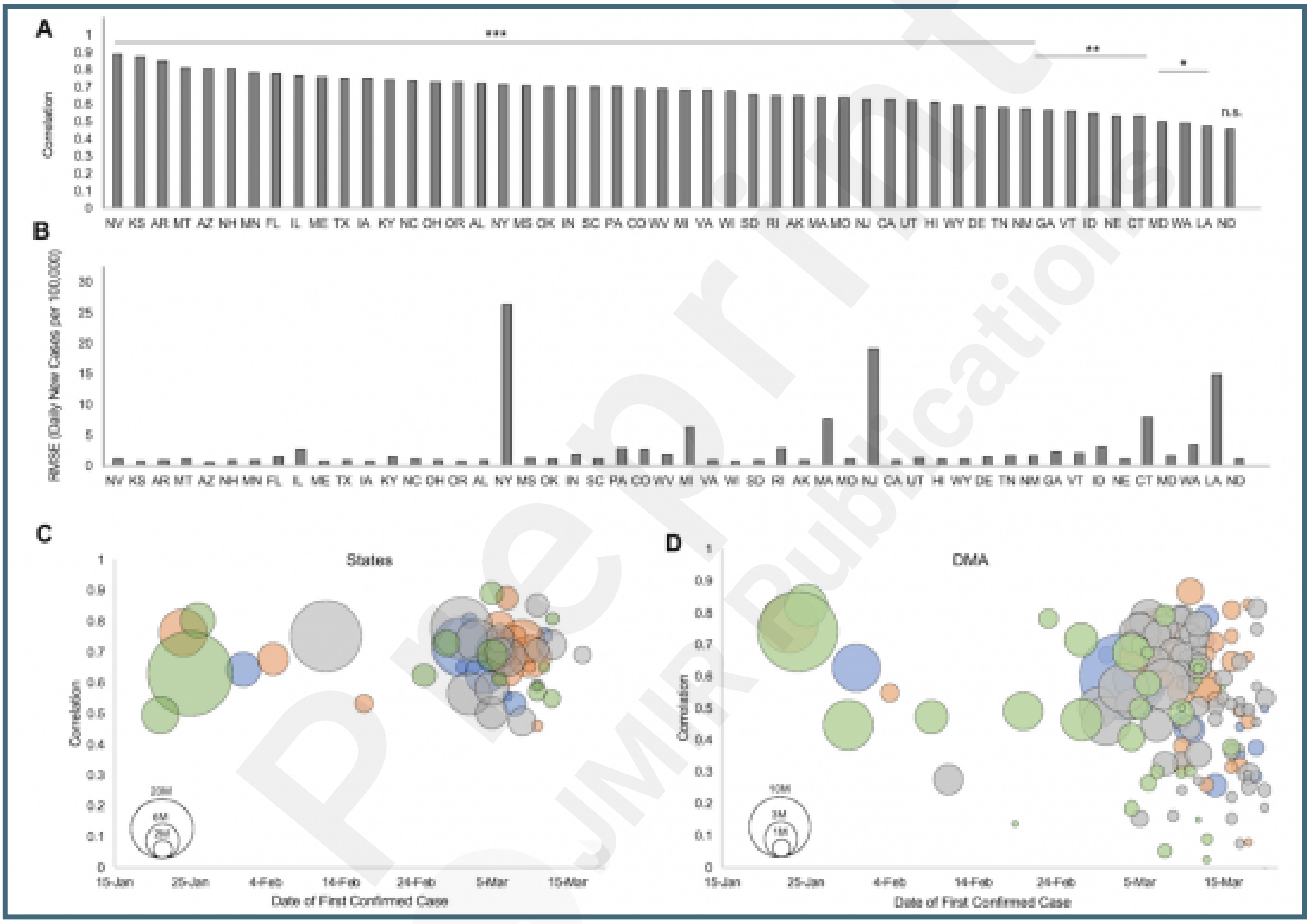
1. Heymann D, Shindo N. COVID-19: what is next for public health? Lancet. 2020;395(10224):542-545. doi: 10.1016/S0140-6736(20)30374-3. PMID: 32061313
2. Gander K. CDC director says there may be another coronavirus wave in late fall and a 'substantial portion of Americans' will be susceptible. *Newsweek*. https://www.newsweek.com/cdc-director-coronavirus-wave-late-fall-substantial-portion-americans-will-susceptible-1495401. Published April 1, 2020. Accessed April 6, 2020.
3. Bertozzi A, Franco E, Mohler G, Short M, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proc Natl Acad Sci. 2020:202006520. doi: 10.1073/pnas.2006520117. PMID: 32616574
4. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012-1014. doi: 10.1038/nature07634. PMID: 19020500
5. Woo H, Cho Y, Shim E, Lee J-K, Lee C-G, Kim S. Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea. J Med Internet Res. 2016;18(7):e177. doi: 10.2196/jmir.4955. PMID: 27377323
6. Yang S, Kou S, Lu F, Brownstein J, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLoS Comput Biol. 2017;13(7):e1005607. doi: 10.1371/journal.pcbi.1005607. PMID: 28727821
7. Cooper C, Mallon K, Leadbetter S, Pollack L, Peipins L. Cancer internet search activity on a major search engine, United States 2001-2003. J Med Internet Res. 2005;7(3):e36. doi: 10.2196/jmir.7.3.e36. PMID: 15998627
8. Pervaiz F, Pervaiz M, Rehman N, Saif U. FluBreaks: Early epidemic detection from google flu trends. J Med Internet Res. 2012;14(5):e125. doi: 10.2196/jmir.2102. PMID: 23037553
9. Dugas A, Jalalpour M, Gel Y, et al. Influenza Forecasting with Google Flu Trends. PLoS One. 2013;8(2):e56176. doi: 10.1371/journal.pone.0056176. PMID: 23457520
10. Olson D, Konty K, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. PLoS Comput Biol. 2013;9(10):e1003256. doi: 10.1371/journal.pcbi.1003256. PMID: 24146603
11. Husnayain A, Fuad A, Chia-Yu Su E. 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. 2020;S1201-9712(20):30140-5. doi: 10.1016/j.ijid.2020.03.021. PMID: 32173572
12. Ayyoubzadeh S, Zahedi H, Ahmadi M, Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google Trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill. 2020;6(2):e18828. doi: 10.2196/18828. PMID: 32234709
13. Li C, Chen L, Chen X, Zhang M, Pang C, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance. 2020;25(10):2000199. doi: 10.2807/1560-7917.ES.2020.25.10.2000199. PMID: 32183935
14. Walker A, Hopkins C, Surda P. Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak. Int Forum Allergy Rhinol. 2020;10(7):839-847. doi: 10.1002/alr.22580. PMID: 32279437

15. The New York Times. We’re Sharing Coronavirus Case Data for Every U.S. County. *The New York Times*. https://www.nytimes.com/article/coronavirus-county-data-us.html. Published March 28, 2020. Accessed April 6, 2020.
16. Smith J. Comparison of COVID-19 case and death counts in the United States reported by four online trackers: January 22-May 31, 2020. medRxiv. 2020. doi: 10.1101/2020.06.20.20135764.
17. Internet World Stats - Usage and Population Statistics. Accessed April 6, 2020. https://www.internetworldstats.com/.
18. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science. 2014;343(6176):1203-1205. doi: 10.1126/science.1248506. PMID: 24626916

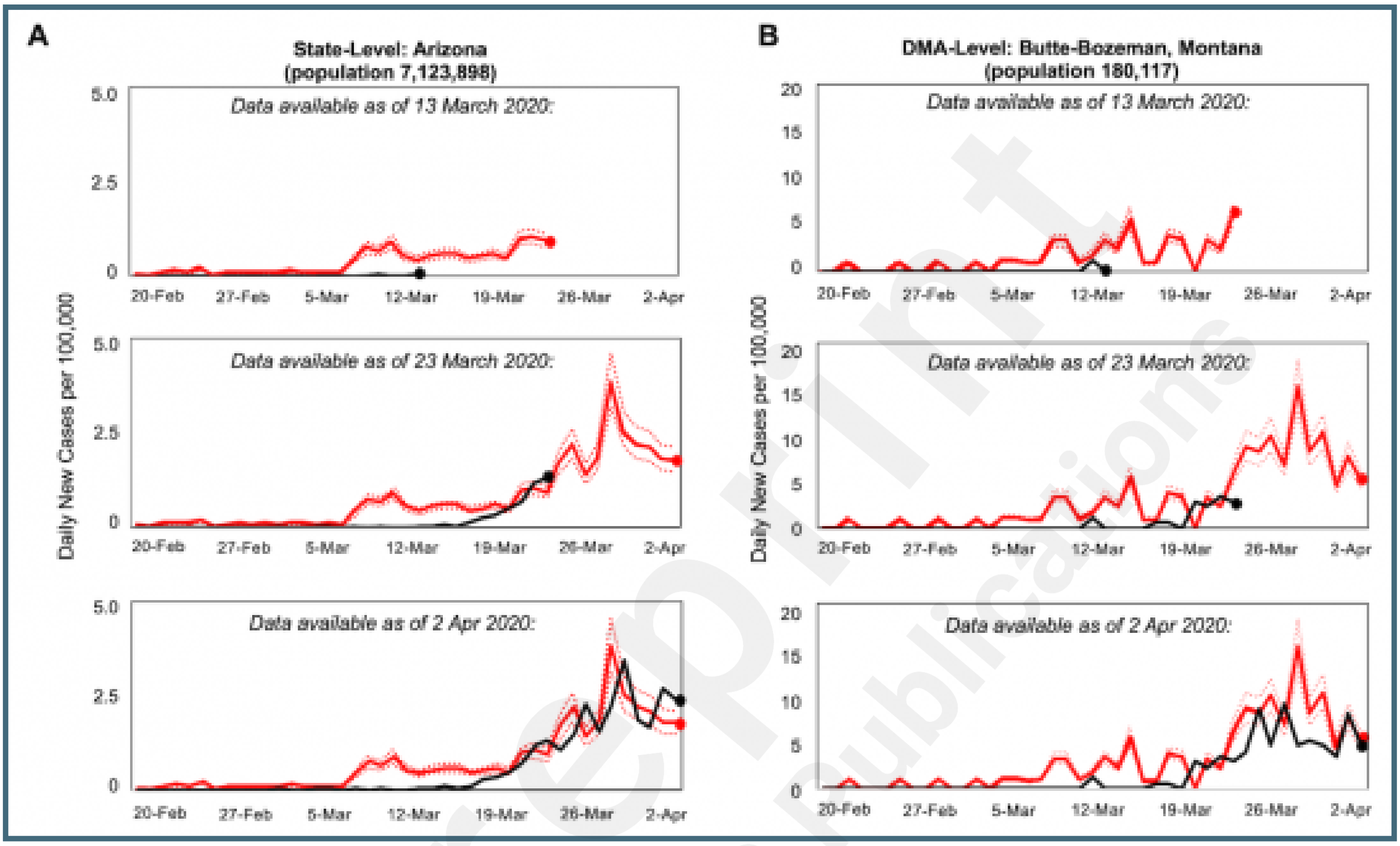
Supplementary Files

Figures

Correlation of query predictions with regional COVID-19 confirmed case rates (A) Correlation of predicted case rates with actual case rates for the 50 states. Values are Pearson correlation coefficients. * indicates significance at alpha = .05; ** at alpha = .01; *** at alpha = .005. (B) RMSE between predicted case rates and actual case rates for the 50 states, in units of daily new cases per 100,000 population. (C) Prediction correlations at the state level do not depend on outbreak timing, as measured by date of first confirmed case. Circle size indicates relative population of state. Color indicates census-designated region of the US (blue: Northeast, orange: Midwest, gray: South, green: West). (D) Prediction correlations at the DMA level do not depend on outbreak timing, as measured by date of first confirmed case. Circle size indicates relative population of DMA. Color indicates census-designated region of the US, as described.



Correlation of query predictions (red) with regional COVID-19 case rates (black) at the state and DMA level, February 20–April 2, 2020 (A) Comparison of predicted case rates (red) with actual case rates (black) at the state level, with Arizona shown as an example. Dashed lines indicate 95% confidence intervals. (B) Comparison at the DMA level, with the Butte-Bozeman area shown as an example of predictions in a low-population region.



Multimedia Appendixes

Supplementary Materials.
URL: https://asset.jmir.pub/assets/ef49bd1716349645b1165d43256835fd.docx