Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model

Xiaoling Yuan[1,2,#], Jie Xu[1,#], Sabiha Hussain[3], He Wang[4], Nan Gao[2,5] and Lanjing Zhang[2,5,6,7*]

[1]*Department of Infectious Disease, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China;* [2]*Department of Biological Sciences, Rutgers University Newark, NJ, USA;* [3]*Department of Medicine, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, USA;* [4]*Department of Pathology, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, USA;* [5]*Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA;* [6]*Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA;* [7]*Department of Chemical Biology, Rutgers Ernest Mario School of Pharmacy, Piscataway, NJ, USA;* [#]*These authors made equal contributions to the works and should be considered as co-first authors.*

Abstract

Background and objectives: The daily incidence and deaths of coronavirus disease 2019 (COVID-19) in the USA are poorly understood. Internet search interest was found to be correlated with COVID-19 daily incidence in China, but has not yet been applied to the USA. Therefore, we examined the association of internet search-interest with COVID-19 daily incidence and deaths in the USA.

Methods: We extracted COVID-19 daily new cases and deaths in the USA from two population-based datasets, namely 1-point-3-acres.com and the Johns Hopkins COVID-19 data repository. The internet search-interest of COVID-19-related terms was obtained using Google Trends. The Pearson correlation test and general linear model were used to examine correlations and predict trends, respectively.

Results: There were 636,282 new cases and 28,325 deaths of COVID-19 in the USA from March 1 to April 15, 2020, with a crude mortality of 4.45%. The daily new cases peaked at 35,098 cases on April 10, 2020 and the daily deaths peaked at 2,494 on April 15, 2020. The search interest of COVID, "COVID pneumonia" and "COVID heart" were correlated with COVID-19 daily incidence, with 12 or 14 days of delay (Pearson's r = 0.978, 0.978 and 0.979, respectively) and deaths with 19 days of delay (Pearson's r = 0.963, 0.958 and 0.970, respectively). The 7-day follow-up with prospectively collected data showed no significant correlations of the observed data with the predicted daily new cases or daily deaths, using search interest of COVID, COVID heart, and COVID pneumonia.

Conclusions: Search terms related to COVID-19 are highly correlated with the COVID-19 daily new cases and deaths in the USA.

Keywords: Trend; Incidence; COVID-19; USA; Pandemic; Model; Search interest.

Abbreviations: CDC, Centers for Disease Control and Prevention; COVID-19, coronavirus disease 2019; WHO, World Health Organization.

Received: April 14, 2020; Revised: April 16, 2020; Accepted: April 17, 2020

[*]**Correspondence to:** Lanjing Zhang, Department of Pathology, Princeton Medical Center, 1 Plainsboro Rd., Plainsboro, NJ 08563, USA. Tel: +1-609-853-6833; Fax: +1-609-853-6841; E-mail: lanjing.zhang@rutgers.edu

How to cite this article: 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. *Exploratory Research and Hypothesis in Medicine* 2020;000(000):000–000. doi: 10.14218/ERHM.2020.00023.

Introduction

Coronavirus disease 2019 (COVID-19) has been pandemic in the world.[1–4] It has now affected more than 560,000 Americans.[3,5] Several attempts were successfully made to model COVID-19 daily incidence in China.[1,6] However, the trends of daily incidence and deaths of COVID-19 in the USA are still poorly understood. Recently, internet search-interest was found to be correlated with daily incidence of COVID-19 in China, with the lag time of 8 to 10 days.[7] Google search-interest was also used to track or model COVID-19 trends in Europe, Iran, and Taiwan.[8–10] Indeed, internet search-interest has been used for modelling and detecting influenza epidemics in the USA and Australia.[11,12] We, therefore, aimed to examine the association of search-interest with daily incidence/new cases and deaths of COVID-19 in the USA, using population-based data and a semiparametric model.

Methods

The data of daily new cases and new deaths of COVID-19 in the USA were extracted from the 1-point-3-acres.com[5] and the Johns Hopkins COVID-19 data repository[3] on April 9, 2020, respectively, for modelling. We later obtained additional data from these sites to evaluate our models' accuracies using Pearson's correlation coefficients. We used a semiparametric model, including prediction of the daily new-case or new-death value based on a given Google Trends search-interest using Pearson's correlation (the parametric component), as well as assigning such a predicted value to the corresponding date of the given Google Trends search interest. Owing to no finite dimensionality of Google Trends search-interest versus time, the second component thus is non-parametric.

Data from the World Health Organization (WHO) Situation Reports appeared significantly inconsistent, and thus were not used.[13] According to the 1-point-3-acres.com website, their data were extracted from various media and government websites, have been manually verified,[5] and have been used by various parties, including Johns Hopkins COVID-19 data repository, WHO, and many others. Due to the use of publicly available, de-identified data and lack of protected health information, the study is exempted from requiring an Institutional Review Board approval (Category 4).

We used the Google Trends function to extract the data of search-interest with the search period of March 1 to April 7, 2020 and COVID-19-related search terms. Based on the COVID-19 symptoms, common terms for COVID-19 and common diseases in the USA, we chose the search terms of "COVID-19," "COVID," "coronavirus," "SARS-CoV2," "pneumonia," "high temperature," "cough," "COVID heart," "COVID pneumonia," and "COVID diabetes." Google Trends search-interest represented search interest relative to the highest search-interest for a given time and region.[7,12] A value of 100 is the peak popularity for the term, while a score of 0 means there were not enough data for this term.

We then examined the lag correlations of the terms' search interests with COVID-19 daily new cases and deaths as described before,[7] whereas the lag time was defined as the difference between a data point's original corresponding time and the shifted one in the lag correlation study. The lag times of our interest were up to 20 days for daily new cases and 23 days for daily death, respectively. The terms with the top-3 correlation coefficients were used to build respective generalized linear models. Based on these models, we used the existing search interests to predict future COVID-19 daily new cases and new deaths in the USA, which would be compared with the prospectively collected data for assessing prediction accuracies.

All statistical analyses were carried out using Stata (version 15). The models' accuracies were assessed using Pearson's r. All p values were two-sided. Only a $p<0.05$ was considered statistically significant.

Results

The Johns Hopkins data repository and 1-point-3-acres.com provided slightly different estimates of COVID-19 daily new cases and deaths in the USA, although they claimed to share data. The data of a given date from 1-point-3-acres.com dataset varied by the release dates. Considering the data inconsistency, we chose the John Hopkins' data for modelling, and the 1-point-3-acres.com data for a sensitivity study. There were 636,282 new cases and 28,325 deaths of COVID-19 reported in the USA from March 1 to April 15, 2020, with a crude mortality of 4.45%. The daily new cases peaked at 35,098 cases on April 10, 2020 and the daily deaths peaked at 2,494 on April 15, 2020.

Google Trends search-interests had a 2-day delay in reporting (*i.e.* a search on April 9 yielded data up to April 7). COVID-19 has a much lower search interest score than COVID (Fig. 1), and was excluded from additional analysis also owing to its close relationship with COVID. As reported before, the correlation coefficients of search terms changed with lag time (Fig. 2). Among the nine terms we searched, COVID, "COVID pneumonia" and "COVID heart" had the top-3 correlation coefficients for the correlation with daily incidence and new deaths (Table 1). Our predicted COVID-19 daily new cases and new deaths would plateau for about 12 days (Fig. 3), suggesting a possible 12-day plateau of these epidemiologic parameters in the future.

The sensitivity study using 1-point-3-acres' data revealed the correlation coefficients that were similar to those produced using Johns Hopkins' data (Table 1). The 7-day follow-up with prospectively collected data showed no significant correlations of the observed data with the predicted daily new cases using search-interest of COVID, COVID heart and COVID pneumonia ($p = 0.178$, 0.480 and 0.094, respectively) nor with the predicted daily new deaths using search interest of COVID, COVID heart and COVID pneumonia ($p = 0.267$, 0.222 and 0.841, respectively).

Discussion

This population-based study shows that there were 636,282 new cases and 28,325 deaths of COVID-19 reported in the USA from March 1 to April 15, 2020. It also shows that the search-interest of COVID, COVID pneumonia, and COVID heart were highly correlated with COVID-19 daily new cases and new deaths, with a delay of 12 days and 19 days, respectively. However, the prediction accuracies of these models appeared low during a 7-day follow-up.

To our knowledge, this study provided, for the first time, evidence that search-interest pertinent to COVID-19 is highly correlated with the trends in COVID-19 daily new cases and new deaths in the USA. The approximately 7 days of difference in lag time between daily new cases and deaths suggest the possibility of a 7-day interval between COVID-19 diagnosis and death in some patients. Additional studies are warranted to investigate this hypothesis. The findings of our study enable us to model daily new cases and deaths in the USA during the early phase (March 1 to April 8) of the COVID-19 outbreak and may greatly help prevent and prepare for any upcoming pandemic and burdens of COVID-19 in the future.

The 12 days of lag time in the USA, as shown by us, was longer than the previously reported 9 days in China.[7] Several factors may contribute to this difference but should be subject to additional studies. First, there was a significant delay in testing for COVID-19 in the USA,[14] which might subsequently lead to longer lag time between the trends of search-interest and daily incidence. Second, the U.S. Centers for Disease Control and Prevention (CDC) recommended a priority-based testing strategy and allowed for not testing some subjects considered low-priority when the COVID-19 tests are short in supply.[15] The criteria for testing COVID-19 in the USA, therefore, were different from those in China and Europe,

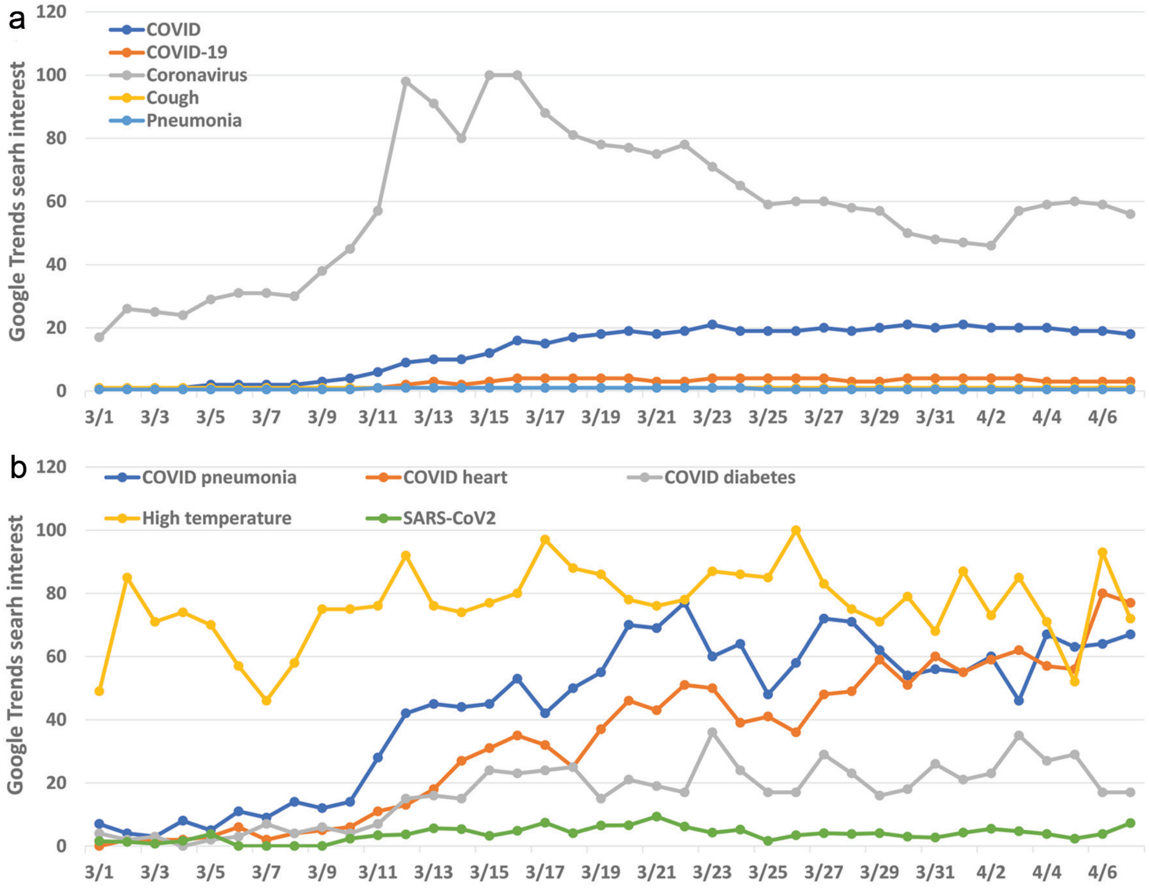

Fig. 1. Trends in search-interest of COVID-19-related terms. The numbers represented the search-interest relative to the term of the highest search-interest in the USA from March 1 to April 7, 2020.

where the WHO criteria were adopted.[16–18] Thus, the patients, who met the WHO criteria, may not be tested and subsequently not included in the daily incidence in the USA; this could lead to underreporting of daily incidence. Third, the biological and socioeconomic differences between the USA and Chinese patients may also contribute to the difference. Finally, the prevalent COVID-19 subtypes in the USA may also be different from those in China and result in different lag times.[19]

This study provides several lines of valuable evidence. First, COVID-19 daily new deaths in the USA are poorly understood, and are here described and studied using a semiparametric model. Second, we extensively examined nine COVID-19-related search terms, which are more than the two used in a previous study.[7] Our data also suggest that pneumonia and heart problems were highly relevant to the daily new cases and deaths in the USA. This finding may be explained by the frequent pneumonia and cardiac injuries seen in COVID-19 patients.[20,21] Third, the lag time in our study was longer than that previously reported in China (12 days vs. 9 days). However, the 12 and 19 days of lag time also afforded us the opportunity to assess a model's prediction accuracy for a longer period of future trends. Fourth, the comparison of predicted values and prospectively collected data will significantly reduce the recall and selection biases.

We will continue updating the models' accuracies as more data become available (see https://github.com/thezhanglab/COVID-US-google). Indeed, we found very high correlation in retrospective modelling but low accuracy in prediction, suggesting that the search-interest based model may be more helpful in predicting daily-incidence peak or early outbreak than post-peak or post-intervention trends. The unexpected low accuracy of model prediction was due to significant attenuation of trend plateau. It may be linked to the April 3 recommendation of wearing masks by the U.S. CDC,[22] which was 5 days before our model's peak time and matched the COVID19's median incubation time of 5 days.[20] Finally, to our knowledge, we are first to examine the correlations of search interest with the COVID-19 daily new cases and deaths in the USA and show greater correlations (Pearson's $r > 0.97$) than reported in the Chinese data.[7]

This study is limited by the retrospective nature of the modeling part and may have some related biases. Moreover, due to the different testing strategies and criteria used in the USA and other countries,[15–18] the comparison of our findings to those of other countries should be interpreted with caution. Finally, the data from Johns Hopkins' data repository was not independently validated or authenticated. However, our sensitivity study using the 1-point-3-acres' data confirms a similar correlation of search-interest with COVID-19 daily new cases and deaths in the USA.

Future directions

Despite the high correlation coefficients in retrospective study/

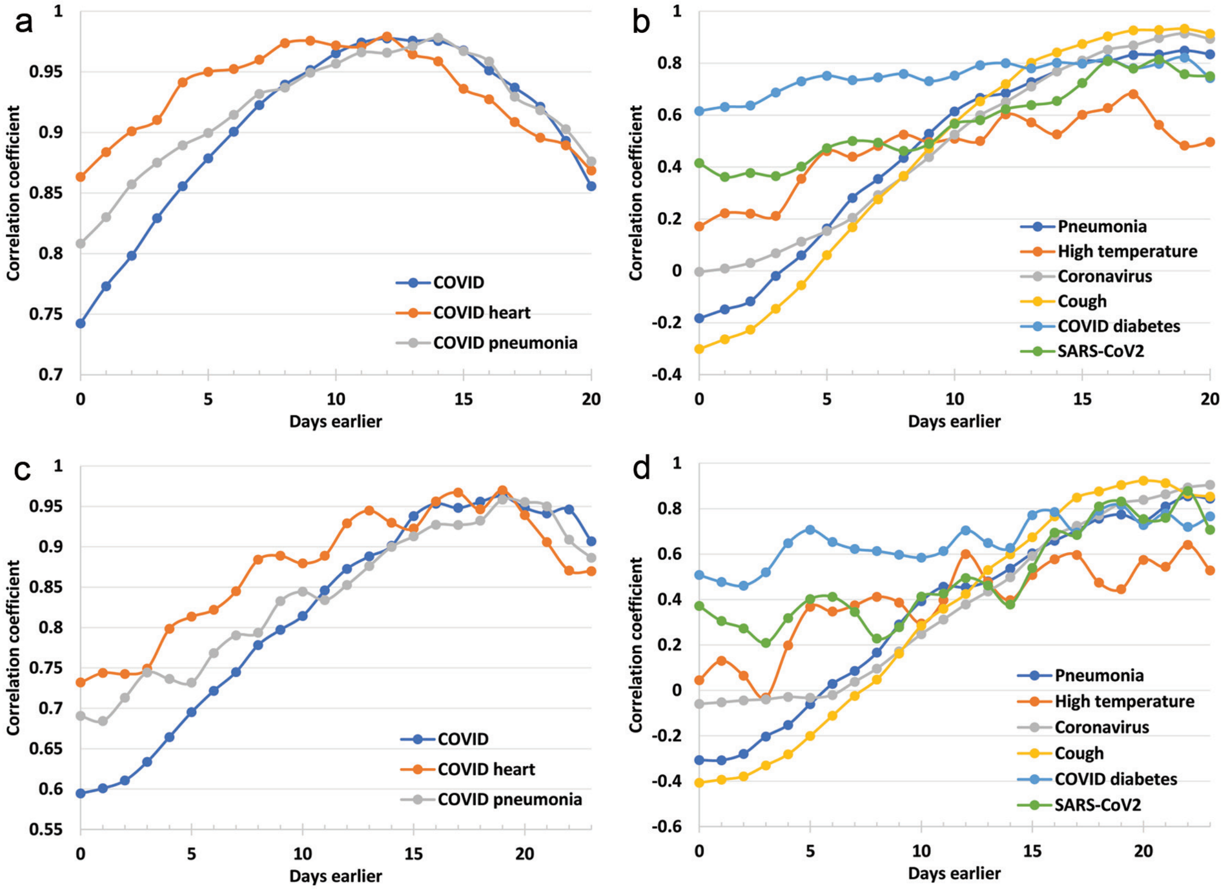

Fig. 2. Lag correlations between Google Trends search-interest of the terms "COVID," "COVID heart," "COVID pneumonia," and others, and the daily new cases and deaths of COVID-19 in the USA, March 1 to April 8, 2020. (a, c) The search terms with the highest Pearson's correlation coefficients for daily new cases and new deaths, respectively; (b, d) The rest of the search terms.

modeling, the prediction-models based on the search-interest trend reached low accuracies during a 7-day follow-up. Additional studies are warranted to understand and improve these models. Why the prediction model failed should also be examined. The April 3 CDC recommendation of more indications for mask-use might be one of the reasons. Finally, the factors linked to and the epidemiological sig-

Table 1. The search term of the top-3 correlation coefficients for correlations with COVID-19 daily new cases and deaths, March 1 to April 8, 2020

Search term	Johns Hopkins Data Repository						1-point-3-acres.com					
	Daily new cases			Daily new deaths			Daily new cases			Daily new deaths		
	Days earlier	r^a	*p*	Days earlier	r^a	*p*	Days earlier	r^a	*p*	Days earlier	r^a	*p*
COVID heart	12	0.979	<0.001	19	0.970	<0.001	12	0.982	<0.001	19	0.977	<0.001
COVID pneumonia	14	0.978	<0.001	19	0.958	<0.001	12	0.977	<0.001	19	0.967	<0.001
COVID	12	0.978	<0.001	19	0.963	<0.001	13	0.973	<0.001	20	0.972	<0.001
Cough	19	0.932	<0.001	20	0.923	<0.001	19	0.935	<0.001	20	0.945	<0.001
Coronavirus	19	0.914	<0.001	23	0.905	<0.001	19	0.909	<0.001	22	0.925	<0.001
Pneumonia	19	0.848	<0.001	22	0.854	<0.001	19	0.832	<0.001	22	0.897	<0.001
COVID diabetes	18	0.821	<0.001	19	0.816	<0.001	18	0.812	<0.001	19	0.801	<0.001
SARS-CoV2	18	0.814	<0.001	22	0.877	<0.001	18	0.805	<0.001	22	0.856	<0.001
High temperature	17	0.681	<0.001	22	0.641	0.006	16	0.667	<0.001	22	0.650	0.005

[a] The highest correlation coefficients among the correlation coefficients of a given search term by various lag times.

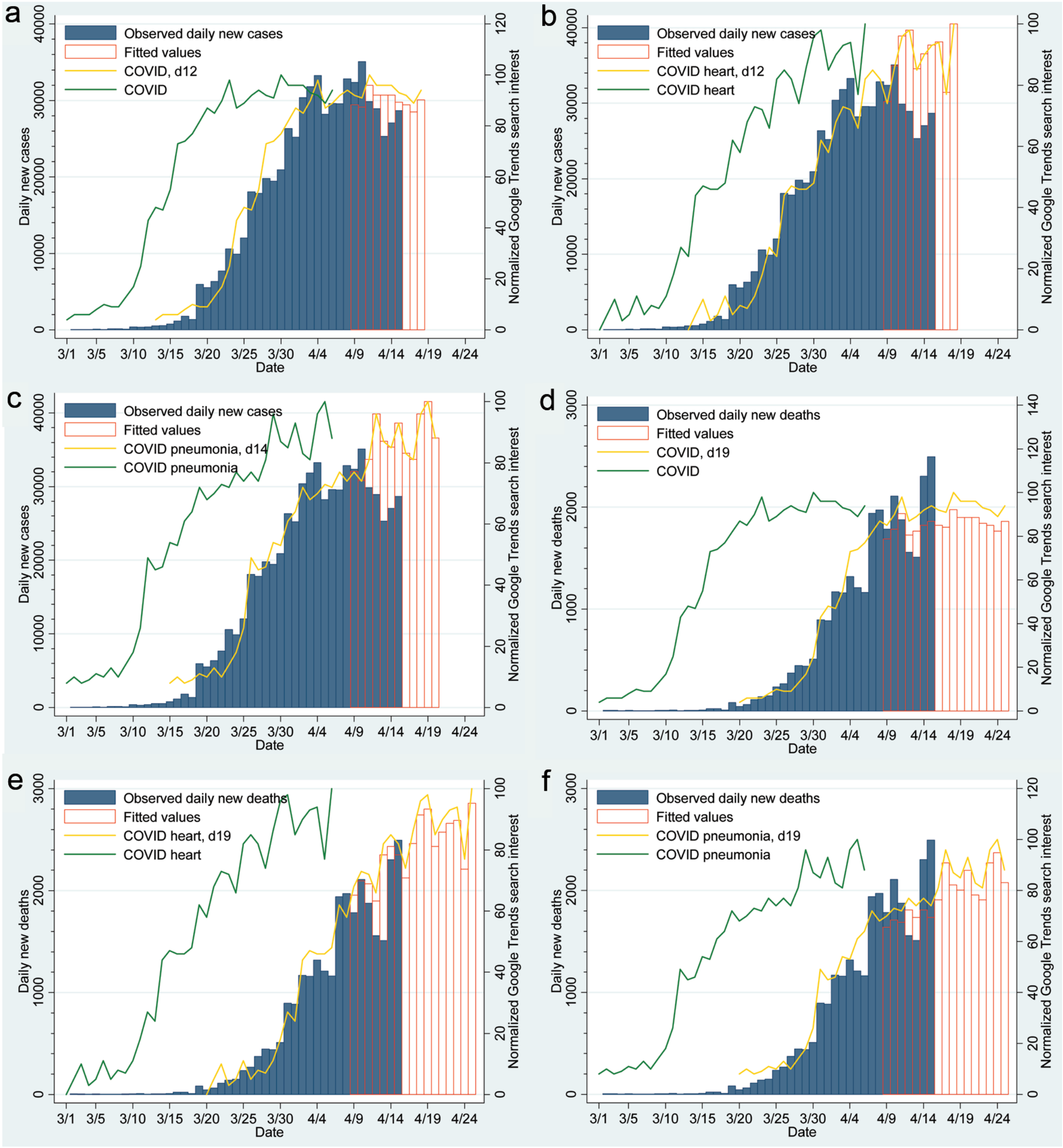

Fig. 3. Google Trends search-interest and the trends in COVID-19 daily new cases and new deaths in the USA, March 1 to April 15, 2020. (a–c) The search-interests of “COVID,” “COVID heart,” and “COVID pneumonia” in Google Trends were 12 to 13 days lagged from COVID-19 daily new cases/incidence (Pearson’s r = 0.977, 0.982 and 0.973, respectively, $p < 0.001$ for all). (d–f) The search interests of “COVID,” “COVID heart,” and “COVID pneumonia” in Google Trends were 19 to 20 days lagged from COVID-19 daily new deaths (Pearson’s r = 0.967, 0.977 and 0.972, respectively, $p < 0.001$ for all). Note, d12, d14 and d19 indicate the trend curves were shifted for 12, 14 and 19 days, respectively, to compensate for lag time. The 7-day follow-up with prospectively collected data showed no significant correlations of observed data with the predicted daily new cases using search interest of “COVID,” “COVID heart,” and “COVID pneumonia” search (p = 0.178, 0.480 and 0.094, respectively), or with predicted daily new deaths (p = 0.267, 0.222 and 0.841, respectively).

nificance of lag time revealed by this study should also be further explored.

Conclusions

This population-based observational study shows that search terms related to COVID-19 are highly correlated with the trends in daily new cases and new deaths of COVID-19 in the USA. Therefore, an internet search-interest based model may be used to predict development and peak-time of COVID-19 outbreak.

Acknowledgments

The data will be regularly updated at https://github.com/thezhanglab/

COVID-US-google, as new prospectively collected incidence data become available.

Funding

The work was supported in part by a grant from the National Institutes of Health (R01DK119198 to NG and LZ).

Conflict of interest

The authors have no conflicts of interest related to this publication.

Author contributions

LZ had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. XY and JX contributed equally and should be considered co-first authors. Concept and design (LZ, NG); drafting of the manuscript (XY, JX); statistical analysis (XY, LZ); supervision (LZ); acquisition, analysis, or interpretation of data (all authors); critical revision of the manuscript for important intellectual content (all authors).

References

[1] Xu J, Cheng Y, Yuan X, Li WV, Zhang L. Trends and prediction in daily incidence of novel coronavirus infection in China, Hubei Province and Wuhan City: an application of Farr law. medRxiv 2020:20025148. doi:10.1101/2020.02.19.20025148.

[2] WHO. Coronavirus disease (COVID-19) outbreak. Available from: https://www.who.int/westernpacific/emergencies/covid-19. Accessed February 23, 2020.

[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020. doi:10.1016/S1473-3099(20)30120-1.

[4] Zhang L. Blind Spots in Fighting the Outbreak of Coronavirus Disease 2019. Explor Res Hypothesis Med 2020;5(1):6–7. doi:10.14218/ERHM.2020.00012.

[5] 1Point3Acres. COVID-19 in US and canada: Real time updates with credible sources. Available from: https://coronavirus.1point3acres.com/en. Accessed April 13, 2020.

[6] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020;395(10225):689–697. doi:10.1016/S0140-6736(20)30260-9.

[7] 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. doi:10.2807/1560-7917.Es.2020.25.10.2000199.

[8] Mavragani A. Tracking COVID-19 in Europe: an infodemiology study. JMIR Public Health Surveill 2020. doi:10.2196/18941.

[9] 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 2020. doi:10.1016/j.ijid.2020.03.021.

[10] Ayyoubzadeh SMo, Ayyoubzadeh SMe, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. JMIR Public Health Surveill 2020;6(2):e18828. doi:10.2196/18828.

[11] Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: the New Zealand experience. Euro Surveill 2009;14(44):19386. doi:10.2807/ese.14.44.19386-en.

[12] 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. doi:10.1038/nature07634.

[13] Ritchie H. Coronavirus Source Data. Available from: https://ourworldindata.org/coronavirus-source-data. Accessed April 12, 2020.

[14] Buchanan L, Lai KKR, Allison McCann A. U.S. lags in coronavirus testing after slow response to outbreak. New York Times 2020. Available from: https://www.nytimes.com/interactive/2020/03/17/us/coronavirus-testing-data.html. Accessed April 16, 2020.

[15] CDC. Evaluating and Testing Persons for Coronavirus Disease 2019 (COVID-19). Available from: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-criteria.html. Accessed April 16, 2020.

[16] WHO. Global Surveillance for human infection with coronavirus disease (COVID-19). Available from: https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov). Accessed April 16, 2020.

[17] McIntosh K. Coronavirus disease 2019 (COVID-19): Epidemiology, virology, clinical features, diagnosis, and prevention. Uptodate 2020. Available from: https://www.uptodate.com/contents/coronavirus-disease-2019-covid-19-epidemiology-virology-clinical-features-diagnosis-and-prevention/print. Accessed April 16, 2020.

[18] European CDC. Case definition and European surveillance for COVID-19, as of 2 March 2020. Available from: https://www.ecdc.europa.eu/en/case-definition-and-european-surveillance-human-infection-novel-coronavirus-2019-ncov. Accessed April 16, 2020.

[19] Nextrain.org. Genomic epidemiology of novel coronavirus - Global subsampling. Available from: https://nextstrain.org/ncov/global. Accessed April 13, 2020.

[20] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, *et al*. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020;382(13):1199–1207. doi:10.1056/NEJMoa2001316.

[21] Shi S, Qin M, Shen B, Cai Y, Liu T, Yang F, *et al*. Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China. JAMA Cardiol 2020. doi:10.1001/jamacardio.2020.0950.

[22] CDC. Recommendation Regarding the Use of Cloth Face Coverings, Especially in Areas of Significant Community-Based Transmission. Available from: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cloth-face-cover.html. Accessed April 16, 2020.