Tu et al. BMC Infectious Diseases (2021) 21:98
https://doi.org/10.1186/s12879-020-05740-x

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Using Baidu search values to monitor and predict the confirmed cases of COVID-19 in China: – evidence from Baidu index

Bizhi Tu[1†], Laifu Wei[1†], Yaya Jia[2†] and Jun Qian[1*]

Abstract

Background: New coronavirus disease 2019 (COVID-19) has posed a severe threat to human life and caused a global pandemic. The current research aimed to explore whether the search-engine query patterns could serve as a potential tool for monitoring the outbreak of COVID-19.

Methods: We collected the number of COVID-19 confirmed cases between January 11, 2020, and April 22, 2020, from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The search index values of the most common symptoms of COVID-19 (e.g., fever, cough, fatigue) were retrieved from the Baidu Index. Spearman's correlation analysis was used to analyze the association between the Baidu index values for each COVID-19-related symptom and the number of confirmed cases. Regional distributions among 34 provinces/regions in China were also analyzed.

Results: Daily growth of confirmed cases and Baidu index values for each COVID-19-related symptom presented robust positive correlations during the outbreak (fever: r_s=0.705, p=9.623× 10^{-6}; cough: r_s=0.592, p=4.485× 10^{-4}; fatigue: r_s=0.629, p=1.494× 10^{-4}; sputum production: r_s=0.648, p=8.206× 10^{-5}; shortness of breath: r_s=0.656, p=6.182× 10^{-5}). The average search-to-confirmed interval (STCI) was 19.8 days in China. The daily Baidu Index value's optimal time lags were the 4 days for cough, 2 days for fatigue, 3 days for sputum production, 1 day for shortness of breath, and 0 days for fever.

Conclusion: The searches of COVID-19-related symptoms on the Baidu search engine were significantly correlated to the number of confirmed cases. Since the Baidu search engine could reflect the public's attention to the pandemic and the regional epidemics of viruses, relevant departments need to pay more attention to areas with high searches of COVID-19-related symptoms and take precautionary measures to prevent these potentially infected persons from further spreading.

Keywords: COVID-19, Web-based data, Internet searching, Baidu index

* Correspondence: qjpaper@sina.cn
[†]Bizhi Tu, Laifu Wei, and Yaya Jia contributed equally to this work and should be considered as co-first authors.
[1]Department of Orthopedics, The First Affiliated Hospital of Anhui Medical University, 218 Jixi Road, Hefei 230022, Anhui, China
Full list of author information is available at the end of the article

Background

The outbreak of new coronavirus disease 2019 (COVID-19) was characterized by fever, cough, fatigue, sputum production, and shortness of breath, receiving people's attention globally [1, 2]. As of April 22, 2020, COVID-19 had spread to more than 188 countries and regions, resulting in over 9.6 million confirmed cases and 490 thousand deaths worldwide [3]. The astonishing spread speed of the epidemic, to some extent, was failing to monitor and manage the potentially infected persons, which may pose a substantial infection control challenge [4]. Therefore, recognizing the potential quantity of infected persons timely and taking corresponding management measures to control the further spread of COVID-19 is in urgent need.

Because of the unpredictability of international public health emergency, novel methods for monitoring the epidemic's development are substantial. Network real-time data can be easily obtained from the web due to the quick availability of the Internet. According to the 45th China statistical report on internet development, there were over 904 million Internet users in China, with the penetration rate of search engine use reached 83% [5]. Among Internet users, 80% of them tended to use electronic devices to acquire the information they are interested in [6].

Recently, people can easily get health-related information via Internet search engines, which could greatly reflect the searches' physical condition or the relatives and friends the searchers concerned [7]. Moreover, to interrupt the transmission of the epidemic, the Chinese government had put in place strong quarantine measures, which also influences the routinely outpatient service process. Reported studies showed that public search behaviors have already been used to predict some epidemic diseases, such as influenza [8], epidemic erythromelalgia [9], dengue [10], and HIV/AIDS [11].

The surveillance of network searches about clinical symptoms of COVID-19 is more predictable and timely compared to previous detection surveillance (e.g., official announcements, news reports, and mass media) [12–14]. Baidu serves as the most popular search engine, occupies more than 90% of Internet users in China [15]. In this study, we obtained the Baidu index values of COVID-19-related symptoms and the data of confirmed cases of COVID-19 across China to analyze the association between these variables and explore whether the Baidu index could act as a novel tool for monitoring and predicting the epidemic of COVID-19 in China.

Methods

Data from Baidu index

More than 90% of Chinese search engine users tend to use Baidu to retrieve their interesting information [16, 17]. The weighted sum of the Baidu search values can describe the characteristics of people's search behaviors [18]. Baidu Index value is obtained by calculating the number of searches of specific keywords input by the searchers [18]. Using the keywords analysis function, Baidu Index automatically matches its related words according to the keywords typed by users. Previous studies had reported that the top five most common symptoms of the COVID-19 were fever (which accounted for 88.7% of the confirmed cases during hospitalization), cough (67.8%), fatigue (38.1%), sputum production (33.7%), and shortness of breath (18.7%) [1]. Accordingly, we selected those symptoms as the keywords in the current study. Based on the keyword analysis function, 26 search terms that could represent the most common symptoms of COVID-19 were selected (Table S1). We added the search values of each symptom and its related keywords together to get the composite Baidu Index values to perform our research. Besides, we compared the search values of 5 keywords between 2011 and 2020 vertically to investigate whether the Baidu Index changes were an accidental event during the outbreak (Figure S1). To explore whether people's search behaviors appear earlier than the epidemic of COVID-19, we defined a definition to examine our hypothesis: search-to-confirmed interval (STCI). The values of STCI can be obtained by calculating the time interval between the peak growth rate (daily Baidu Index values (DBIV) minus its previous day's values as the growth rate) of the Baidu Index and the peak daily growth of confirmed cases (DGCC). The top ten provinces/regions ranked by the cumulative confirmed cases were selected for STCI analysis.

Confirmed cases of COVID-19

We obtained the data of confirmed cases of COVID-19 from accessible official channels, including the official website of Hopkins University [2], the world health organization (WHO) [19], and the National Health Commission of the People's Republic of China [20]. Since China's epidemic had been gradually controlled after April 22, 2020, we divided the COVID-19 pandemic (January 11, 2020, to April 22, 2020) into a growth period and a decline period. February 10, 2020, was set as the cut-off date, when the government announced the road closures re-opened and fully production resumed [21].

Statistical analysis

Using SPSS (version 23.0), we applied a Spearman correlation analysis to explore the relationships between DGCC and DBIV of COVID-19-related symptoms from January 11, 2020, to April 22, 2020. Using the same statistical methods, we also explored the time lag pattern between DGCC and DBIV of COVID-19-related

symptoms. $P< 0.05$ was set as the level of statistical significance (two-sided test). Besides, GraphPad Prism 8.2 was used to draw figures.

Results

Correlation analysis among search values of Baidu index, cumulative confirmed cases and DGCC in China

As shown in Fig. 1, nationwide cumulative confirmed cases were strongly negative correlated to DBIV (fever: r_s=− 0.455, p=1.206× 10^{-6}; cough: r_s=− 0.923, p=4.958× 10^{-44}; fatigue: r_s=− 0.425 p=7.041× 10^{-6}; sputum production: r_s=− 0.749, p=8.585× 10^{-24}; shortness of breath: r_s=-0.428, p=5.786× 10^{-6}). Taking the cut-off date (February 10, 2020) as the demarcation point, the cumulative confirmed cases and DBIV of fever (r_s=0.705, p=9.623× 10^{-6}), cough (r_s=0.592, p=4.485× 10^{-4}), fatigue (r_s=0.629 p=1.494× 10^{-4}), sputum production (r_s=0.648, p=8.206× 10^{-5}), shortness of breath (r_s=0.656, p=6.182× 10^{-5}) had a strong positive correlation during the growth period and a significantly negative correlation during the decline period (fever: r_s=− 0.971, p =5.850× 10^{-46}; cough: r_s=− 0.967, p=8.601× 10^{-44}; fatigue: r_s=− 0.937 p = 3.948× 10^{-34}; sputum production: r_s=− 0.770, p=1.604 × 10^{-15}; shortness of breath: r_s=-0.930, p=1.333× 10^{-32}) (Figures S2 and S3).

Table 1 and Fig. 2 shows that there were strong statistically positive correlations among the DGCC and Baidu search values of fever (r_s=0.768, p=8.013× 10^{-23}), cough (r_s=0.556, p=1.087× 10^{-9}), fatigue (r_s=0.763, p=7.930× 10^{-21}), sputum production (r_s=0.665, p=1.793× 10^{-14}), and shortness of breath (r_s=0.780, p=2.673× 10^{-22}), nationwide. Among the 34 provinces/regions in China, we found significant correlations between DGCC and DBIV; we observed that the number of daily confirmed cases tended to increase when Baidu searches for terms related to fever, cough, fatigue, and shortness of breath were increasing (Table 1). For Hong Kong, Macao, Taiwan, and Tibet, no consistent correlation was detected between DGCC and Baidu search values of COVID-19-related symptoms. However, DBIV of cough in Shanghai was not correlated to DGCC (r_s=0.133, p= 0.184). Besides, the correlations between sputum production and DGCC in several provinces/regions were inconspicuous (e.g., Beijing: r_s=0.249, p=0.012; Guangdong: r_s=0.262, p=0.008, Hunan: r_s=-0.244, p=0.014) (Table 1).

STCI analysis for people's search behaviors of COVID-19-related symptoms and the epidemic of COVID-19

Figure 3 shows that the peak of the growth rate of the Baidu Index occurred 19–22 days earlier than the peak of DGCC across china (STCI for fever: 22 days; cough: 19 days; fatigue: 20 days; sputum production: 19 days; shortness of breath: 19 days). Moreover, the top 10 provinces/regions ranked by confirmed cases presented similar results except for sputum production (Fig. 3). Oddly, the peak of the Baidu Index's growth rate occurred 17 days later than the peak of DGCC in Heilongjiang.

Lag correlation between the DGCC and search index values of COVID-19 related symptoms

Figure 4 and Table S2 manifests the lag correlation between DBIV of the different keywords and DGCC. We found the highest lag correlation with DBIV were 4 days earlier (r_s =0.574, p=1.826 × 10^{-10}) for cough, 2 days earlier (r_s =0.778, p=2.434 × 10^{-22}) for fatigue, 3 days earlier (r_s =0.664, p=1.630 × 10^{-14}) for sputum production, 1 day earlier (r_s =0.804, p=9.707 × 10^{-25}) for shortness of breath, and 0 days earlier for fever compared with the number of DGCC (r_s =0.791, p=1.623 × 10^{-23}).

Discussion

People with the travel and exposure history of high-risk areas with COVID-19 patients will be required quarantined to control the spread of the pandemic. Since the understanding of the new coronavirus's characteristics and the effective treatments remains uncertain, people usually compared COVID-19 with the SARS, which outbroke in 2003 in China with a mortality rate of 11% [19, 22]. Due to the separate isolation precautions policy and the fear of an unknown virus, people with exposure history are likely to conceal their own and their family's high-risk behaviors, which undermines the government's early attempts to control the suspected cases of COVID-19 [23]. Using Internet search engines, we could predict the potential quantity of affected persons; and the real-time data of the Baidu Index helps monitor the epidemic development and formulates the corresponding government policies.

China had achieved preliminary success in controlling the COVID-19 pandemic by April 22, 2020. The correlation analysis between Chinese public searches of COVID-19-related symptoms and the actual number of confirmed cases will be helpful for exploring the relationships between Internet search values and COVID-19 pandemic and provide novel insights for controlling the epidemic of COVID-19.

The current research shows that the related DBIV reached a peak earlier than the DGCC, and the dynamic changes of DBIV were also earlier than DGCC. We noticed that the higher the search values, the higher the cumulative confirmed cases will be during the growth period, which indicated that the searchers could be the potential infectors of the virus. Besides, DGCC and DBIV presented with a positive correlation during the whole observation period (even in the decline period), which implied the DBIV declined with a decreased number of DCGG. However, when

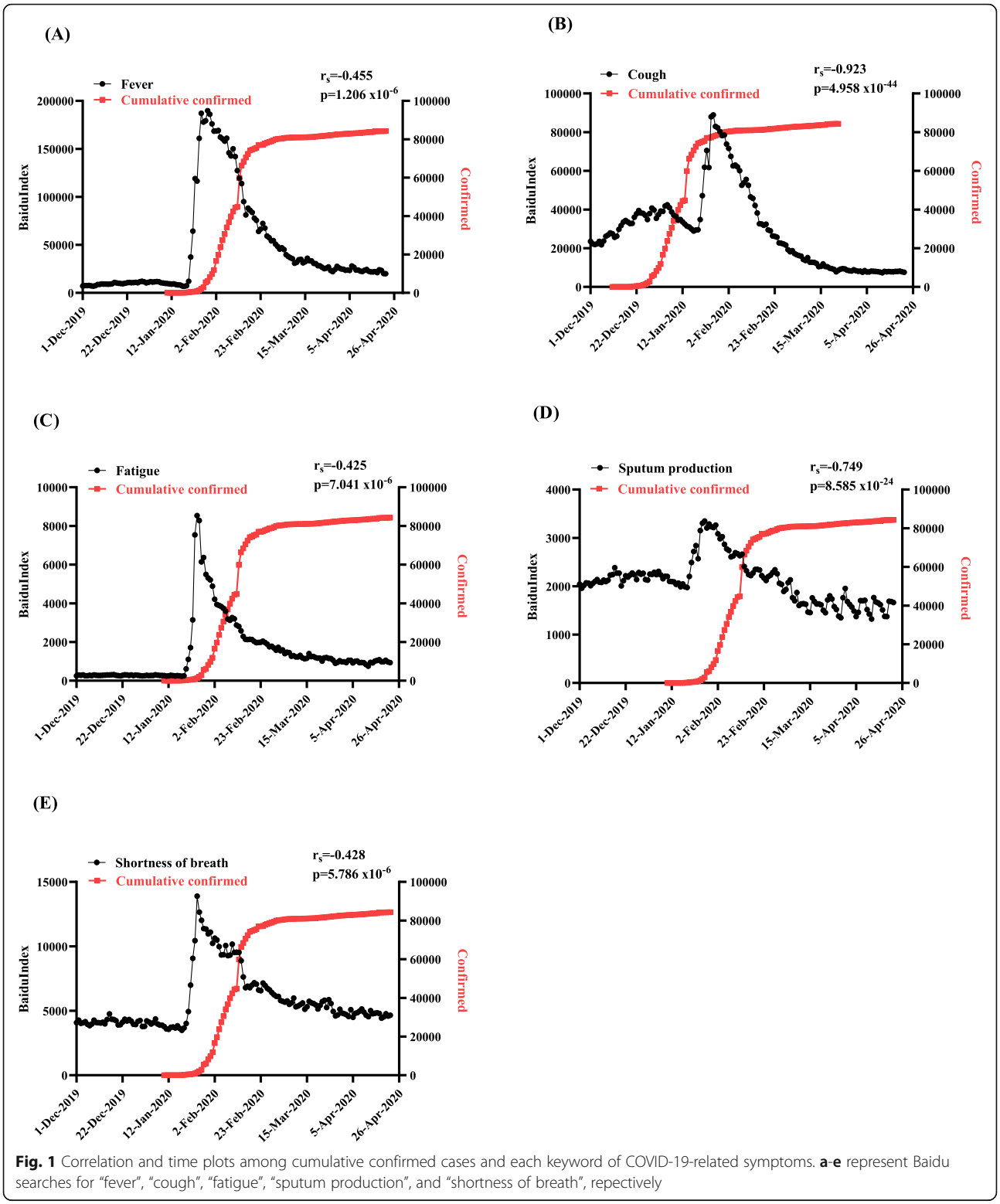

Fig. 1 Correlation and time plots among cumulative confirmed cases and each keyword of COVID-19-related symptoms. **a**-**e** represent Baidu searches for "fever", "cough", "fatigue", "sputum production", and "shortness of breath", repectively

DGCC was declining, the number of cumulative cases continued to increase instead, which could be an explanation for the negative correlation between cumulative cases and DBIV during the decline period. The public's search behaviors for health-related information can reflect their potential physical and psychological problem [7, 24]. The declined searches of COVID-19-related symptoms indicated that the

Table 1 Correlation between daily growth of confirmed cases (DGCC) across China and values of Baidu index (BI)

Region	Values of BI				
	Fever	Cough	Fatigue	Sputum production	Shortness of breath
China					
r_s	0.768	0.556	0.763	0.665	0.780
p	8.013×10^{-23}	1.087×10^{-9}	7.930×10^{-21}	1.793×10^{-14}	2.673×10^{-22}
Anhui					
r_s	0.801	0.770	0.760	− 0.028	0.775
p	5.39×10^{24}	3.131×10^{-21}	2.172×10^{-20}	0.782	1.205×10^{-21}
Beijing					
r_s	0.657	0.431	0.582	0.249	0.610
p	6.336×10^{-14}	6.502×10^{-6}	1.358×10^{-10}	0.012	1.040×10^{-11}
Chongqing					
r_s	0.796	0.769	0.740	0.572	0.738
p	1.542×10^{-23}	1.647×10^{-23}	6.057×10^{-19}	3.389×10^{-10}	8.809×10^{-19}
Fujian					
r_s	0.588	0.471	0.705	0.367	0.537
p	8.473×10^{-11}	5.677×10^{-7}	1.388×10^{-16}	1.485×10^{-4}	5.809×10^{-9}
Gansu					
r_s	0.527	0.444	0.373	− 0.150	0.484
p	1.277×10^{-8}	3.008×10^{-6}	1.112×10^{-4}	0.133	2.586×10^{-7}
Guangdong					
r_s	0.535	0.336	0.527	0.262	0.506
p	7.113×10^{-9}	1.564×10^{-4}	1.287×10^{-8}	0.008	5.598×10^{-8}
Guangxi					
r_s	0.766	0.754	0.760	0.287	0.731
p	7.075×10^{-21}	5.780×10^{-20}	1.904×10^{-20}	0.004	6.872×10^{-8}
Guizhou					
r_s	0.673	0.657	0.622	0.355	0.629
p	9.182×10^{-15}	6.433×10^{-14}	2.921×10^{-12}	2.555×10^{-4}	1.388×10^{-12}
Hainan					
r_s	0.717	0.735	0.694	−0.354	0.693
p	2.474×10^{-17}	1.468×10^{-18}	6.080×10^{-16}	2.673×10^{-4}	6.597×10^{-16}
Hebei					
r_s	0.731	0.635	0.662	0.040	0.705
p	2.622×10^{-18}	7.392×10^{-13}	3.396×10^{-14}	0.691	1.297×10^{-16}
Heilongjiang					
r_s	0.413	0.201	0.453	0.089	0.345
p	1.590×10^{-5}	0.042	1.710×10^{-6}	0.375	2.669×10^{-4}
Henan					
r_s	0.771	0.766	0.728	0.655	0.759
p	2.652×10^{-21}	6.291×10^{-21}	4.647×10^{-18}	7.887×10^{-14}	2.288×10^{-20}
Hong Kong					
r_s	− 0.094	− 0.514	− 0.282	0.517	− 0.085
p	0.349	3.394×10^{-8}	0.004	2.676×10^{-8}	0.398

Table 1 Correlation between daily growth of confirmed cases (DGCC) across China and values of Baidu index (BI) *(Continued)*

Region	Values of BI				
	Fever	Cough	Fatigue	Sputum production	Shortness of breath
Hubei					
r_s	0.709	0.745	0.631	0.614	0.704
p	7.410×10^{-17}	2693×10^{-19}	1.131×10^{-12}	6.640×10^{-12}	1.640×10^{-16}
Hunan					
r_s	0.813	0.797	0.738	−0.244	0.759
p	2.942×10^{-25}	1.256×10^{-23}	9.111×10^{-19}	0.014	2.300×10^{-20}
Inner Mongolia					
r_s	0.322	0.129	0.369	0.385	0.316
p	0.001	0.197	1.384×10^{-4}	6.326×10^{-5}	0.001
Jiangsu					
r_s	0.695	0.565	0.629	0.502	0.630
p	5.378×10^{-16}	5.918×10^{-10}	1.441×10^{-12}	7.609×10^{-8}	1.306×10^{-12}
Jiangxi					
r_s	0.692	0.672	0.686	− 0.317	0.640
p	7.678×10^{-16}	1.052×10^{-14}	1.861×10^{-15}	0.001	4.433×10^{-13}
Jilin					
r_s	0.538	0.446	0.626	0.323	0.355
p	5.415×10^{-9}	2.646×10^{-6}	1.925×10^{-12}	0.001	2.472×10^{-4}
Liaoning					
r_s	0.575	0.425	0.486	− 0.221	0.513
p	2.685×10^{-10}	8.698×10^{-6}	2.179×10^{-7}	0.026	3.436×10^{-8}
Macau					
r_s	0.105	0.016	0.093	0.204	0.015
p	0.293	0.872	0.354	0.040	0.882
Ningxia					
r_s	0.696	0.649	0.541	−0.389	0.503
p	4.495×10^{-16}	1.656×10^{-13}	4.279×10^{-9}	5.317×10^{-5}	7.051×10^{-8}
Qinghai					
r_s	0.461	0.465	0.428	0.297	0.396
p	1.115×10^{-6}	8.234×10^{-7}	7.029×10^{-6}	0.002	3.833×10^{-5}
Shaanxi					
r_s	0.637	0.607	0.606	− 0.157	0.670
p	5.969×10^{-13}	1.319×10^{-11}	1.494×10^{-11}	0.115	1.406×10^{-14}
Shandong					
r_s	0.706	0.584	0.702	0.528	0.708
p	1.230×10^{-16}	1.217×10^{-10}	2.135×10^{-16}	5.238×10^{-7}	9.317×10^{-17}
ShanghaiShanghai					
r_s	0.331	0.133	0.379	− 0.020	0.391
p	0.001	0.184	8.633×10^{-5}	0.841	4.810×10^{-5}
Shanxi					
r_s	0.380	0.275	0.313	0.001	0.365
p	8.102×10^{-5}	0.005	0.001	0.991	2.382×10^{-4}

Table 1 Correlation between daily growth of confirmed cases (DGCC) across China and values of Baidu index (BI) *(Continued)*

Region	Values of BI				
	Fever	Cough	Fatigue	Sputum production	Shortness of breath
Sichuan					
r_s	0.775	0.687	0.720	0.681	0.771
p	1.247×10^{-21}	1.565×10^{-15}	1.588×10^{-17}	3.530×10^{-15}	2.517×10^{-21}
Tianjin					
r_s	0.483	0.424	0.517	0.295	0.453
p	2.675×10^{-7}	9.050×10^{-6}	2.624×10^{-8}	0.003	1.755×10^{-6}
Tibet					
r_s	0.167	0.139	0.173	−0.003	0.043
p	0.093	0.165	0.082	0.973	0.670
Xinjiang					
r_s	0.737	0.704	0.593	−0.284	0.504
p	9.948×10^{-19}	1.642×10^{-16}	4.944×10^{-11}	0.004	6.872×10^{-8}
Yunnan					
r_s	0.689	0.616	0.635	−0.340	0.638
p	1.274×10^{-15}	5.308×10^{-12}	7.636×10^{-13}	4.776×10^{-14}	5.252×10^{-13}
Zhejiang					
r_s	0.592	0.530	0.628	0.349	0.618
p	5.553×10^{-11}	1.026×10^{-8}	1.569×10^{-12}	3.250×10^{-4}	4.461×10^{-12}
Taiwan					
r_s	−0.111	− 0.428	−0.242	0.523	−0.019
p	0.269	7.105×10^{-6}	0.014	1.699×10^{-8}	0.854

public's mentality might be more relaxing in the decline period compared with the growth period.

We can tell from Baidu's time plots for COVID-19-related symptoms and the number of confirmed cases that the former dynamic changes appeared earlier than the latter. Among 34 provinces/regions in China, although most areas in this research showed statistical correlations among the DBIV and DGCC (except sputum production), Hong Kong, Macao, Taiwan, and Tibet did not present with such correlations. One possible reason could be that the Baidu search engine is not the primary search tool in these places [4]. Additionally, there was only one confirmed case in Tibet, which was insufficient to conduct the statistical analysis. Besides, there was no correlation between DGCC and DBIV of cough in Shanghai, which might owe to the incompleteness of search keywords related to cough. Of interest, no correlation between DBIV of sputum production and DGCC was observed. A reasonable explanation could be that sputum production is more common in the elderly with chronic respiratory diseases, and such searches might be correlated to seasonal influenza every year in the late autumn to early spring [25]. Based on our research, the increase in the DBIV of COVID-19-related symptoms could be treated as an abnormal signal worthy of government departments' corresponding action in advance.

The increased number of relevant searches indicates there are more potentially infected candidates. Around 97.5% of people with identifiable exposure history would develop symptoms within 11.5 days, and 1% of them had a more extended incubation period of more than 14 days [26]. We found that the average maximum of DBIV's growth rate was 20 days earlier than DGCC in most areas except Heilongjiang. On May 10, 2020, the Heilongjiang government reported that the pandemic had relapsed; thus, the apex of DBIV appeared later compared with other provinces [27]. Compared with the traditional diagnosis and treatment process, most potential patients are inclined to search the Internet for help, indicating the difference to publicly reported overrepresent severe cases of COVID-19 [7, 28, 29]. Those potential infectors were likely to use search engines (usually Baidu in China) to search for the related information, so the Baidu index could reflect the approximate number of these potential infectors. The mild potential infectors may possess a more extended incubation period theoretically on account of several days lags before being confirmed [30]. The soaring DBIV of COVID-19-related symptoms in a certain area might be a precursor for the

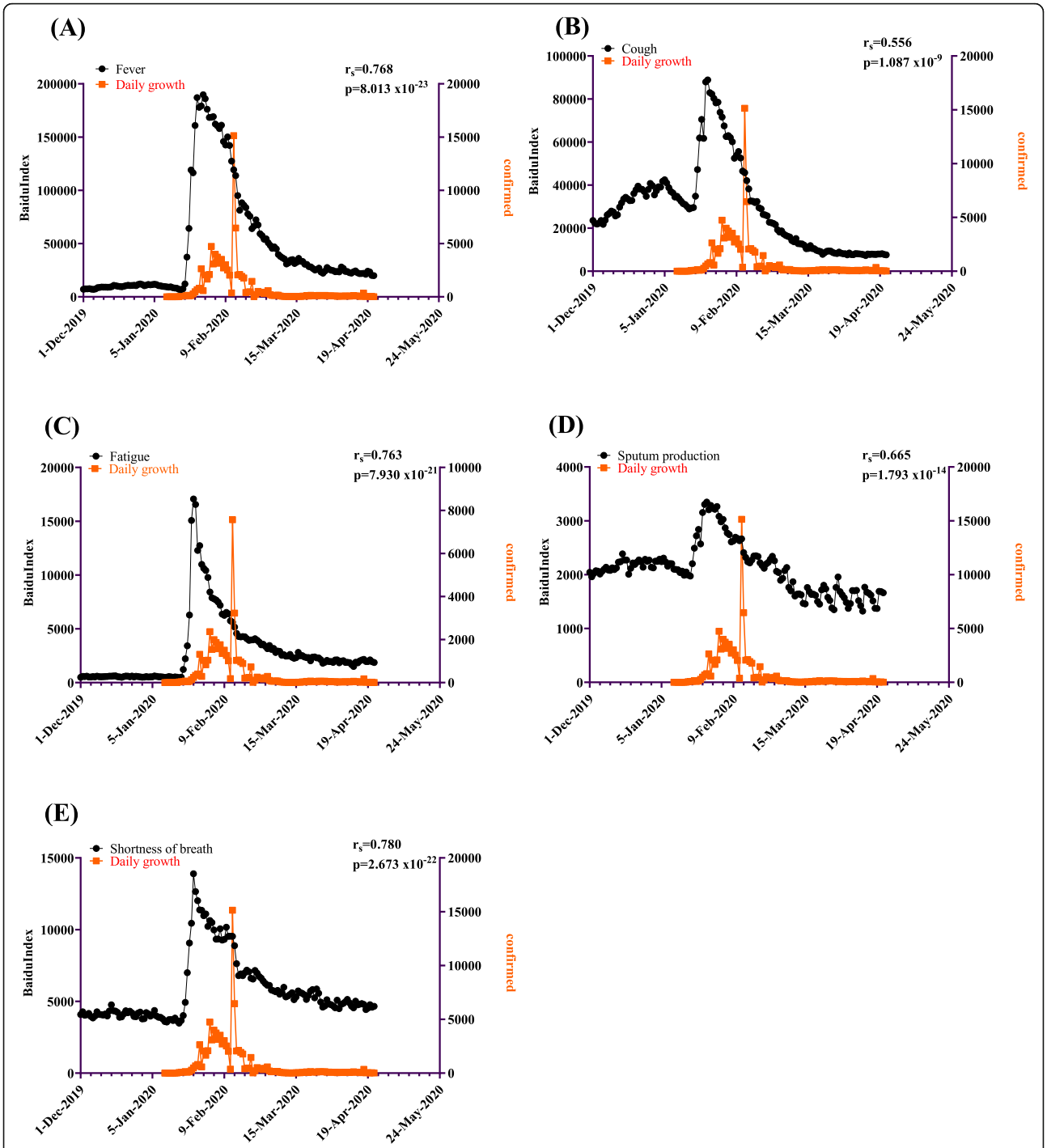

Fig. 2 Correlation and time plots among daily confirmed cases and each keyword of COVID-19-related symptoms. **a**-**e** represent Baidu searches for "fever", "cough", "fatigue","sputum production", and "shortness of breath", repectively

future outbreak of the epidemic. The STCI analysis shows that the peak DBIV of COVID-19-related symptoms appeared 19–22 days earlier than the peak DGCC. However, the results of the time-lag correlation analysis delivered a shorter lag than STCI. Since the STCI study only compared the interval between the peak DBIV of COVID-19-related symptoms and DGCC, it did not take other data into account. Therefore, time-lag correlation analysis could be better to explore the lag patterns of DBIV and DGCC. We found that the optimal time lag of DBIV of fever, cough, fatigue, sputum production, and shortness of breath was 0, 4, 2, 3, 1 day/days,

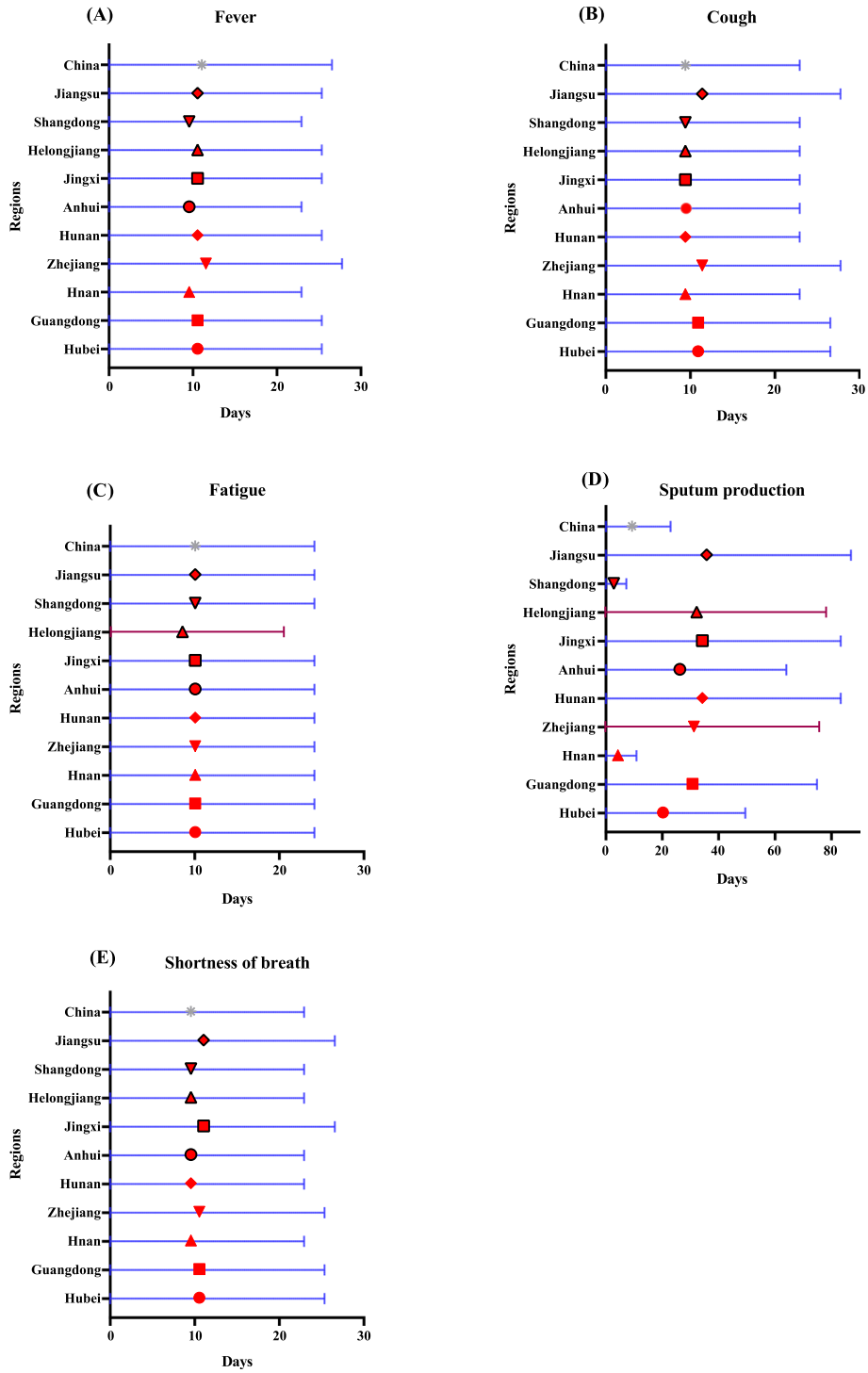

Fig. 3 Days earlier of serarch-to-comfrimed interval (STCI) among the top ten provinces in the cumulative number of confirmed cases. e.g., The red line represents the absolute value of the negative value

respectively. According to Cuilian et al., the peak of Internet searches about COVID-19 appeared 10–14 days earlier than the peak of reported daily growth cases in China [31], and 10 days earlier in America [32]. People who searched the terms of "新冠" or "冠状病毒" (keywords in Cuilian's study) were more likely to experience the incubation period, while the searchers querying the COVID-19-related symptoms were likely those who

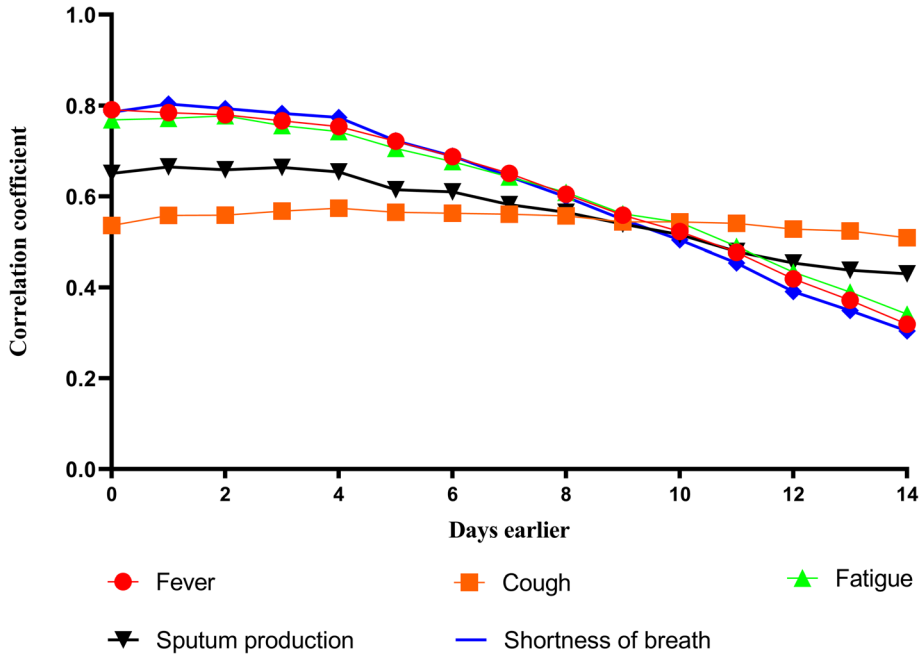

Fig. 4 Lag correlations between the DGCC and each search topics of COVID-19-related symptom

were infected and had already experienced the incubation period. Moreover, there is no time-lag for "fever"; this may attribute to the body temperature reporting mechanism adopted by both the Chinese government and local institutions. This reporting system required that people with fever be actively isolated and quarantined immediately to prevent the potential further spread of COVID-19 [33, 34]. Therefore, people with fever would be isolated and confirmed subsequently. As a result, no time lag was observed.

Limitations

There are some limitations needed to be recognized. Firstly, we only utilized Baidu's data to perform our research; other search engines, such as Weibo and Twitter, were not included. Secondly, some keywords related to the symptoms of COVID-19 were not included in the current study, and the keywords utilized in the current work could not guanine the consistency and efficiency of the long-term prediction in the future. Therefore, future studies are suggested to add or delete the corresponding keywords of COVID-19-related symptoms to confirm that the time lag patterns exist between DBIV and DGCC. Thirdly, the detailed information about the individual searchers remains unclear, so it is impossible to identify the specific potential infectors. Besides, there were several documented issues with the predictability of disease incidence trends using search engines. To avoid the failure of predicting an epidemic with the utilization of the Internet search engine, a random forest regression model is suggested in the future study to facilitate our observing results [35].

Conclusion

Our research suggested that there was a significant correlation between DBIV of COVID-19-related symptoms and DGCC. The dynamic changes of DGCC showed several days lags compared with the DBIV. Besides, DBIV of COVID-19-related symptoms could serve as a potential indicator for predicting the epidemic of emerging infectious diseases and guide targetable intervention and prevention of COVID-19 to further assist in the overall control of the pandemic.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-020-05740-x.

Additional file 1: Table S1. Search terms for topics regarding top five symptoms of COVID-19 in Chinese.

Additional file 2: Figure S1. Search trend of keywords related to COVID-19 symptoms among 2011~2020.

Additional file 3: Figure S2. Correlation and time plots among cumulative confirmed cases and each keyword of COVID-19-related symptoms during the growth period. (A), (B), (C), (D), and (E) represent Baidu searches for "fever", "cough", "fatigue", "sputum production", and "shortness of breath", respectively.

Additional file 4: Figure S2. Correlation and time plots among cumulative confirmed cases and each keyword of COVID-19-related symptoms during the decline period. (A), (B), (C), (D), and (E) represent Baidu searches for "fever", "cough", "fatigue", "sputum production", and "shortness of breath", respectively.

Additional file 5: Table S2. Lag correlation coefficients and *p* values between search index values of each keyword and daily confirmed cases.

Abbreviations

CSSE: Center for Systems Science and Engineering; JHU: Johns Hopkins University; COVID-19: New coronavirus disease 2019; STCI: Search-to-confirmed interval; DBIV: Daily Baidu Index values; DGCC: Daily growth of confirmed cases; WHO: World Health Organization

Acknowledgments

We thank all the people who offer help for this study. And thank the Department of Orthopedics, The First Affiliated Hospital of Anhui Medical University, for its grateful supports.

Authors' contributions

JQ conceived the study idea. BZ collected the data. BZ, YY, and LF contributed to the analysis of the data as well as wrote the initial draft with all authors providing critical feedback and edits to subsequent revisions. All authors approved the final draft of the manuscript. All authors are accountable for all aspects of the work in ensuring related questions accuracy or integrity. Any parts of the work are appropriately investigated and resolved. JQ is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

This project was supported by the National Natural Science Foundation of China (grant numbers 81471273 and 81671204), the Foundation of Supporting Program for the Excellent Young Faculties in the University of Anhui Province in China. Grants for Scientific Research of BSKY from the First Affiliated Hospital of Anhui Medical University; and Grants for Outstanding Youth from the First Affiliated Hospital of Anhui Medical University.
The funding body has neither role in the design of the study, collection, analysis, interpretation of data, nor in writing the manuscript.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Author(s) declare(s) that there is no conflict of interest.

Author details

[1]Department of Orthopedics, The First Affiliated Hospital of Anhui Medical University, 218 Jixi Road, Hefei 230022, Anhui, China. [2]Department of Pediatrics, The Shanxi Medical University, Taiyuan, Shanxi, China.

Received: 17 July 2020 Accepted: 26 December 2020
Published online: 21 January 2021

References

1. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020;382(18):1708–20.
2. Wuhan Municipal Health Commission. http://wjw.wuhan.gov.cn/. Accessed 26 June 2020.
3. COVID-19 Dashboard by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://www.arcgis.com. Accessed 26 June 2020.
4. Al-Tawfiq JA. Asymptomatic coronavirus infection: MERS-CoV and SARS-CoV-2 (COVID-19). Travel Med Infect Dis. 2020;35:101608.
5. Search engines in China - statistics & facts. Available at: https://www.statista.com/topics/1337/search-engines-in-china/. Accessed 26 June 2020.
6. Fox S. Health information online. Washington, DC: Pew Internet & American Life Project; 2005.
7. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health. 2017;7(3):185–9.
8. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PLoS One. 2013;8(5):e64323.
9. Gu YZ, Chen FL, Liu T, Lv XJ, Shao ZM, Lin HL, Liang CB, Zeng WL, Xiao JP, Zhang YH, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. Sci Rep-Uk. 2015;5:12649.
10. Guo P, Liu T, Zhang Q, Wang L, Xiao JP, Zhang QY, Luo GF, Li ZH, He JF, Zhang YH, et al. Developing a dengue forecast model using machine learning: a case study in China. Plos Neglect Trop D. 2017;11(10):e0005973.
11. He GY, Chen YS, Chen BW, Wang H, Shen L, Liu L, Suolang DJ, Zhang BY, Ju GD, Zhang LL, et al. Using the Baidu search index to predict the incidence of HIV/AIDS in China. Sci Rep-Uk. 2018;8(1):1–10.
12. Freifeld CC, Mandl KD, Ras BY, Bronwnstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc. 2008;15(2):150–7.
13. van de Belt TH, van Stockum PT, Engelen LJLPG, Lancee J, Schrijver R, Rodriguez-Bano J, Tacconelli E, Saris K, van Gelder MMHJ, Voss A. Social media posts and online search behaviour as early-warning system for MRSA outbreaks. Antimicrob Resist Infect Control. 2018;7:69.
14. Li CL, Chen LJ, Chen XY, Zhang MZ, Pang CP, Chen HY. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance. 2020;25(10):7–11.
15. China Internet Network Information Center. Available at: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/. Accessed 26 June 2020.
16. China Search Engine Market Overview (2015). Available at: https://www.chinainternetwatch.com/17415/search-engine-2012-2018e/. Accessed 26 June 2020.
17. China Internet Network Information Center. Chinese internet users search behavior study. Beijing; 2014. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/ssbg/201910/P020191025506904765613.pdf.Htm. Accessed 26 June 2020. [WebCite Cache ID 75IWlqvRn].
18. Baidu. Baidu Index. https://index.baidu.com/. Accessed 26 June 2020. [WebCite Cache ID 6yOtOa7p9].
19. World Health Organization. Available at: https://www.who.int/. Accessed 26 June 2020.
20. National Health Commission of the People's Republic of China. Available at: http://www.nhc.gov.cn/. Accessed 26 June 2020.
21. State Council of the PRC. Available at: http://www.gov.cn/guowuyuan/. Accessed 26 June 2020.
22. Ashraf H. Investigations continue as SARS claims more lives. Lancet. 2003;361(9365):1276.
23. China Central Television. Available at: https://www.cctv.com/. Accessed 26 June 2020.
24. Hu D, Lou X, Xu Z, Meng N, Xie Q, Zhang M, Zou Y, Liu J, Sun G, Wang F. More effective strategies are required to strengthen public awareness of COVID-19: evidence from Google Trends. J Glob Health. 2020;10(1):011003.
25. Uyeki TM, Bernstein HH, Bradley JS, Englund JA, File TM, Fry AM, Gravenstein S, Hayden FG, Harper SA, Hirshon JM, et al. Clinical practice guidelines by the Infectious Diseases Society of America: 2018 update on diagnosis, treatment, chemoprophylaxis, and institutional outbreak management of seasonal influenza. Clin Infect Dis. 2019;68(6):895–902.
26. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med. 2020;172(9):577–82.
27. Health Commission of Heilongjiang Province. Available at: http://wsjkw.hlj.gov.cn/. Accessed 3 Oct 2020.
28. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009;49(10):1557–64.
29. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron AJ. More diseases tracked by using Google Trends. Emerg Infect Dis. 2009;15(8):1327–8.
30. Press Conference of the Joint Prevention and Control Mechanism of the State Council. Available at: http://www.gov.cn/xinwen/gwylflkjz18/index.htm. Accessed 26 June 2020.
31. 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. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000199.

32. 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(7):e19483.
33. National Health Commission of the People's Republic of China. http://www.nhc.gov.cn/xcs/zcwj2/new_zcwj.shtml. Accessed 3 Oct 2020.
34. Lin RT, Cheng Y, Jiang YC. Exploring public awareness of overwork prevention with big data from Google Trends: retrospective analysis. J Med Internet Res. 2020;22(6):e18181.
35. Kandula S, Shaman J. Reappraising the utility of Google Flu Trends. PLoS Comput Biol. 2019;15(8):e1007258. https://doi.org/10.1371/journal.pcbi.1007258 eCollection 2019 Aug.

Publisher's Note