Investigating the correlation of COVID-19 spread with Baidu search data : Infoveillance Study

Xiao Qi, Su-Zhen WANG, Jia-Ning Feng, Gao-Pei ZHu, Yu-Jie Liu, Qian Mao, Zhe Wang, Pei-Xia Guan

Submitted to: Journal of Medical Internet Research
on: April 15, 2021

Table of Contents

Investigating the correlation of COVID-19 spread with Baidu search data : Infoveillance Study

Xiao Qi[1] MD, MSc; Su-Zhen WANG[1] MD, MSc; Jia-Ning Feng[1] MPH; Gao-Pei ZHu[1] MD, MSc; Yu-Jie Liu[1] MSc; Qian Mao[1] MPH; Zhe Wang[1] MPH; Pei-Xia Guan[1] MPH

[1]School of Public Health Weifang Medical University Wei Fang CN

Corresponding Author:
Su-Zhen WANG MD, MSc
School of Public Health
Weifang Medical University
7166 Baotong West Street
Wei Fang
CN

Abstract

Background: The sudden outbreak of COVID-19 has placed an unprecedented pressure on China's public health system. It is imperative to strengthen the capacity of early surveillance and early warning to build a sound public health system. Therefore, it is necessary to improve the multi-channel monitoring and early warning mechanism to improve the ability of real-time analysis and judgment.

Objective: To explore the correlation of COVID-19 spread with Baidu search data in Beijing, so as to evaluate the possibility of monitoring the epidemic situation of COVID-19 with Baidu search data.

Methods: This study compared the daily case counts of COVID-19 outbreak from January 20 to March 1, 2020 with Baidu search data for the same period in Beijing. After keyword selection, filtering and composition, the most correlated lag of the COVID-19 Baidu Search Index (CBSI) was used for comparison and linear regression model development.

Results: Our findings showed a positive relationship of CBSI and the confirmed cases of COVID-19 (?=0.711, P < .001). The strongest correlation between COVID-19 confirmed cases and indices, CBSI, was at a lag of -11 days. The regression coefficient ?1 of the established regression model was equal to 1.042 (P<.001), R2 was equal to 0.7, which indicated that Baidu search data could reflect 70% of the variation in COVID-19 cases.

Conclusions: COVID-19 Baidu Search index may be a good monitoring indicator for early detection of COVID-19 outbreaks.

(JMIR Preprints 15/04/2021:29578)
DOI: https://doi.org/10.2196/preprints.29578

Preprint Settings

Original Manuscript

Investigating the correlation of COVID-19 spread with Baidu search data : Infoveillance Study

Xiao Qi, MSc,MD;Jianing Feng,MPH; Gaopei ZHu, MSc,MD;Yujie Liu, MSc ;Qian Mao,MPH; Zhe Wang,MPH; Peixia Guan,MPH;Suzhen WANG, MSc, MD
(School of Public Health, Weifang Medical University, Shandong, China)
Corresponding author：
Suzhen WANG, MSc, MD
School of Public Health
Weifang Medical University
7166 Baotong West Street
Shandong
China
Phone: 86 05368462437
E-mail：wangsz@wfmc.edu.cn

Abstract：Background: The sudden outbreak of COVID-19 has placed an unprecedented pressure on China's public health system. It is imperative to strengthen the capacity of early surveillance and early warning to build a sound public health system. Therefore, it is necessary to improve the multi-channel monitoring and early warning mechanism to improve the ability of real-time analysis and judgment.

Objective To explore the correlation of COVID-19 spread with Baidu search data in Beijing, so as to evaluate the possibility of monitoring the epidemic situation of COVID-19 with Baidu search data.

Methods This study compared the daily case counts of COVID-19 outbreak from January 20 to March 1, 2020 with Baidu search data for the same period in Beijing. After keyword selection, filtering and composition, the most correlated lag of the COVID-19 Baidu Search Index (CBSI) was used for comparison and linear regression model development.

Results Our findings showed a positive relationship of CBSI and the confirmed cases of COVID-19 (ρ=0.711, P < .001). The strongest correlation between COVID-19 confirmed cases and indices, CBSI, was at a lag of -11 days. The regression coefficient $\beta 1$ of the established regression model was equal to 1.042 (P<.001), R^2 was equal to 0.7, which indicated that Baidu search data could reflect 70% of the variation in COVID-19 cases.

Conclusion COVID-19 Baidu Search index may be a good monitoring indicator for early detection of COVID-19 outbreaks.

Key words ：COVID-19; SARS-CoV-2; public health; Baidu search index (BSI); monitor; digital surveillance;correlation; Internet

Introduction

Since the beginning of the new millennium, communicable diseases have emerged continuously. The severe acute respiratory syndrome (SARS) in 2002 and the Middle East respiratory syndrome (MERS) in 2012 were all spread around the world [1,2]. In 2019, the novel coronavirus (SARS-CoV-2) which caused the Novel Coronavirus Disease 2019 (COVID-19), is considered to be the third highly pathogenic coronavirus after MERS-CoV and SARS-CoV in the 21st century[3] . The continuous and regular emergence of coronavirus has posed a major threat to human health and economy.

However, we still have many unknowns about SARS-CoV-2, such as its intermediate host, its animal-to-human transmission mode and so on[4]. Therefore, it is very important to develop a monitoring system to monitor or track the spread of COVID-19 so as to make a correct report and an instant reaction. Traditional surveillance systems, which are based on passive or sentinel surveillance in outpatient clinics or hospitals, are limited by insufficient reporting, delayed diagnosis and inadequate laboratory services, which may resulting in getting inaccurate number of the cases[5,6]. The development of real-time and accurate surveillance of communicable diseases remains a real challenge for the world.

Digital surveillance systems based on Internet search data can provide important information about the emergence and spread of diseases[7] and can be used to complement traditional health care-based surveillance systems[8]. For example, some studies use Google flu trends to accurately track influenza outbreaks in real time[9,10], The occurrence of a number of infectious diseases, including influenza, gonorrhea and erythema limb pain, has been linked to the Baidu searching Index (BSI) [11,12]. BSI is the data that Baidu's searching volume provided to the public in the form of a weighted index. Baidu is the world's largest Chinese search engine with the highest domestic market share[13], and its search data is highly representative in China. However, to our knowledge, previous studies had focused on the possibility of monitoring infectious diseases by using searching data from previous several years.There were few studies on the feasibility of using short-term and timely data to monitor the spread of communicable diseases. Therefore, this study used BSI to obtain timely searching data about COVID-19, then explored the relationship between COVID-19 searching behavior and COVID-19, so as to evaluate the feasibility of using search data to monitor the epidemic situation of COVID-19.

At the same time, this study also explored the relationship between temperature changes and COVID-19 to determine whether temperature can be used in conjunction with searching data to monitor COVID-19 epidemic situation.

Methods

Data Sources

Outbreak data. COVID-19 is diagnosed based on epidemiological history, clinical manifestations and etiological evidence[14].In this study, the daily autochthonous case counts of COVID-19 from January 20 to March 1, 2020 in Beijing were collected from Beijing Municipal Health Commission.

Baidu search data Baidu is the most popular Internet search engine in China. Baidu search index (BSI)[15] contains the search volume of a large number of keywords entered by Baidu users since June 2006. User’s privacy is also maintained because only term frequency data is available. Daily search data can be provided at the municipal, provincial, and national levels. This study collected Baidu search data of COVID-19-related keywords in Beijing for a total of 42 days from January 20, 2020 to March 1, 2020 from BSI.

Temperature data According to previous studies, low temperature is conducive to the transmission of some Virus [16]. Therefore, this study collected the low temperature data from January 20, 2020 to March 1, 2020 in Beijing to directly explore the relationship between low temperature and COVID-19 cases. The temperature data were obtained from a free weather inquiry website in

Chinese, the Weather Post report [17].

Keywords selection and filtering

Keywords selection is a crucial issue in monitoring disease development based on Internet search data, which directly affects the ability and accuracy of monitoring. However, there are no clear guidelines or criteria for the selection of keywords[18]. Previous studies generally chose the name, clinical symptoms or diagnosis of the disease as its primary keywords[19,20]. This study used the name of COVID-19 in Chinese and Baidu keyword search website[21] to obtain COVID-19-related keywords in order to minimize the omission of major terms. The relevant keywords in the site are recommended by different sites: Baidu, portals, blogs, and online reports using semantic correlation analysis, etc. After inputting the core terms, we obtained 70 related keywords with search volume(Supplementary Table 1). However, more keywords do not necessarily result in better results[22], because some of the recommended keywords are not closely related to the epidemic situation of COVID-19 in Beijing, which may reduce the monitoring capacity of the surveillance system. Therefore, we collected Beijing's search data from Baidu and filtered the keywords according to the following two steps:

1) We deleted keywords that were not related to the COVID-19 epidemic in Beijing, and 41 keywords left (Supplementary Table 2).

2) The Spearman rank correlation coefficient (ρi) between the BSI of each keyword and the daily new cases of COVID-19 during the study period was then calculated. Keywords with correlation coefficients less than 0.4 and those whose correlations were not statistically significant ($P > .05$) were excluded. Finally, there remained 25 keywords (Supplementary Table 2)

COVID-19 composite Baidu search index (CBSI) calculation

The last 25 keywords left were used to calculate the COVID-19 composite Baidu search index (CBSI). The weight of each keyword was determined by the correlation coefficient (ρi). The calculation formula of CBSI is as follows:

$$weight_i = \frac{\rho i}{\sum_{i=1}^{n} \rho i} \quad (1)$$

$$CBSI = \sum_{i=1}^{n} weight_i \times keyword_i \quad (2)$$

Where n is the number of keywords, $keyword_i$ and $weight_i$ represent the Baidu search index of the i^{th} keyword and the weight of the i^{th} keyword, respectively.

Statistical analysis

Spearman rank correlation analysis was used to evaluate the correlation between CBSI, low temperature and the daily confirmed COVID-19 cases in Beijing during the study period. Time-series cross-correlation analysis was applied to evaluate and quantify the time-lag linear correlation between two time series data. In the present study, the time-series cross-correlation analysis was carried out between BSI and daily confirmed cases in Beijing. Then the time lag with the largest correlation coefficient was selected to establish a linear regression model as follows:

$$Daily\ new = \beta_0 + \beta_1 * \mathrm{CBSI}_l + \varepsilon \quad (3)$$

Where Daily new represents COVID-19 case counts, $CBSI_l$ denotes the lag CBSI with the largest correlation, β_1 as the regression coefficient. The model estimates the case count l days later, based on the Baidu search data for the current day

All analyses were conducted by R software version 4.0.0, and *P*< .05 indicated a statistically significant difference.

Results

General description

Based on the filtering analysis, 29 out of the 70 keywords were not closely related to the epidemic situation of COVID-19 in Beijing, 16 keywords were excluded because the correlation with the case data was not statistically significant or the correlation coefficient was less than 0.4, and at last 25 keywords were left (Supplementary Table 2). Among all the remaining keywords, 52% were about symptoms of COVID-19.

The overall trend of daily confirmed COVID-19 cases, CBSI and low temperature in Beijing from January 20, 2020 to March 1, 2020 were shown in figure 1 and figure 2. The peak of CBSI appeared on January 21, and the daily confirmed cases reached the peak after 11 days.The overall trend of the two figures was similar (Figure 1). Figure 2 showed us the opposite trend between low temperature and newly confirmed cases.

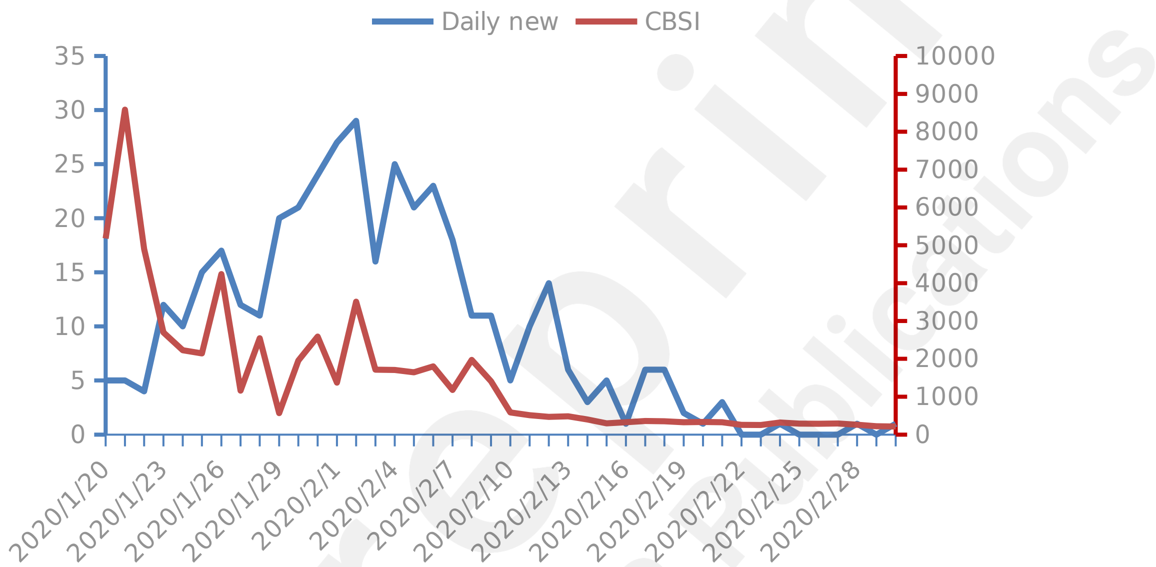


Figure 1. CBSI and Daily COVID-19 case counts during the study period.

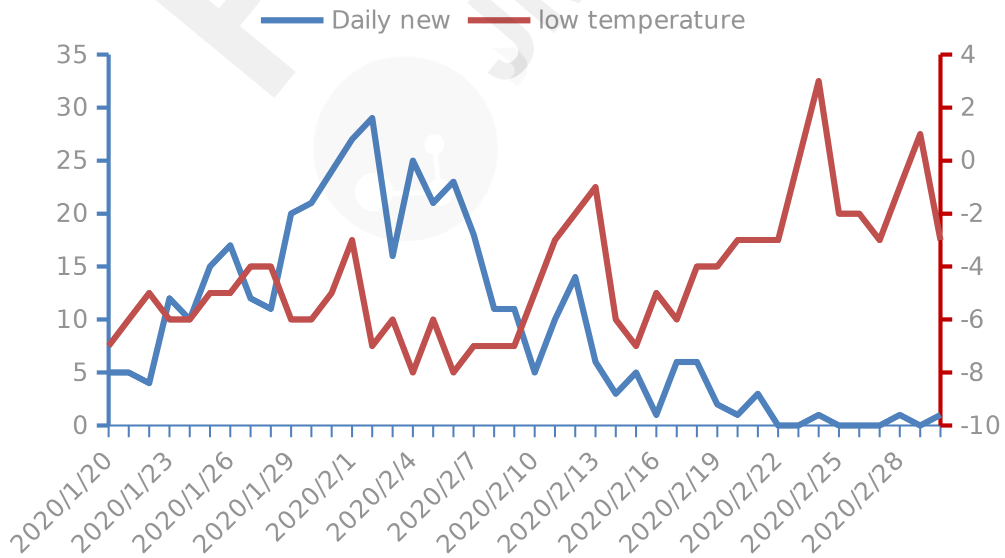
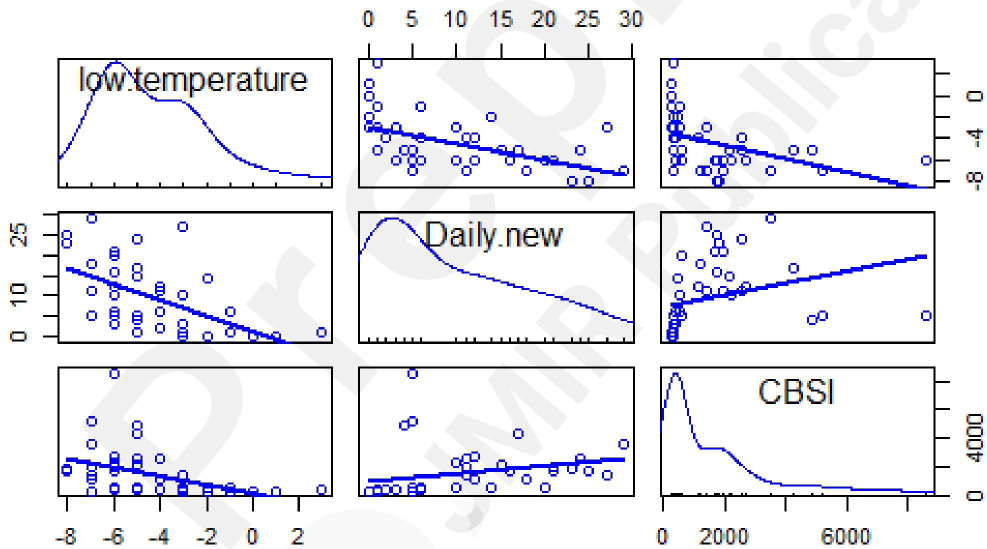


Figure 2. Low temperature and Daily COVID-19 case counts during the study period.

Spearman rank correlation results

The low temperature data, CBSI and the daily COVID-19 case counts in Beijing from January 20 to March 1, 2020 were shown in scatter plot matrix and fitted to the regression line (Figure 3). The relationship between low temperature, CBSI and COVID-19 was analyzed through Spearman correlation analysis. The picture and the results of analysis showed that the low temperature was statistically negatively correlated with the daily confirmed COVID-19 cases (ρ =-0.61 P < .001, see Table 1), and the CBSI value was positively correlated with the daily confirmed COVID-19 cases (ρ = 0. 711, P < .001, see Table 1).



Figue 3. Scatterplot matrix among CBSI, low temperature and Daily COVID-19 case counts

Table 1. The results of Spearman correlation analysis between low temperature, CBSI and Daily COVID-19 case counts are presented

subject	ρ	P-value
low temperature	-0.61	2.98×10^{-5}
CBSI	0.711	4.701×10^{-7}

Time-series cross correlation analysis

Time-series cross correlation analysis demonstrated that daily COVID-19 occurrence to be positively correlated with daily CBSI at the negative time lags of 2–13 days (Figure 4). That is, search data for CBSI 2-13 days in advance was positively correlated with the current number of COVID-19 cases. The strongest correlation between CBSI and daily COVID-19 case counts was found at negative lag of 11 days. We then graphed the curves of daily COVID-19 case counts and CBSI at negative lag 11 over the outbreak period (Figure 5). It was clear that the search data accurately capture the changes in the daily case count.

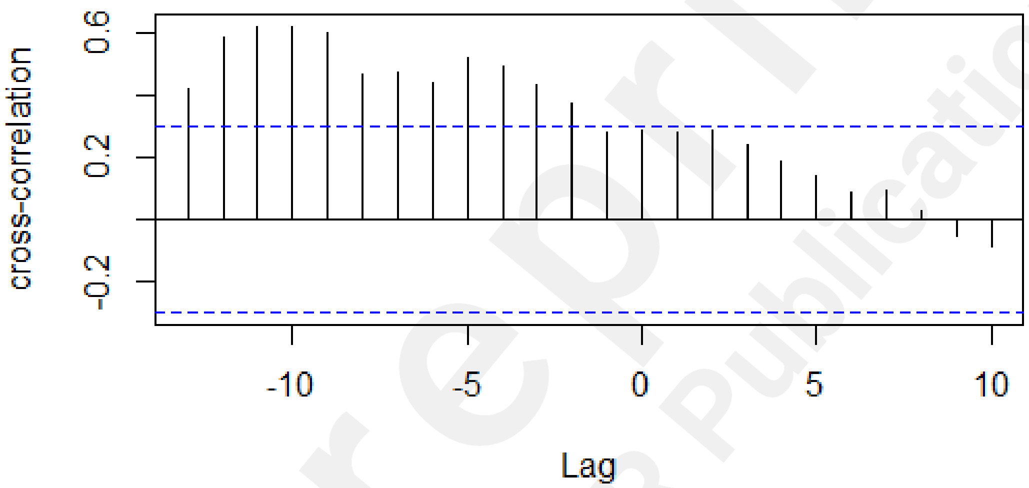


Figure 4. Time series cross-correlation between CBSI and Daily COVID-19 case counts. Confidence intervals (95%) are indicated by the dashed lines (X axis: lag value, Y axis:CCF value).

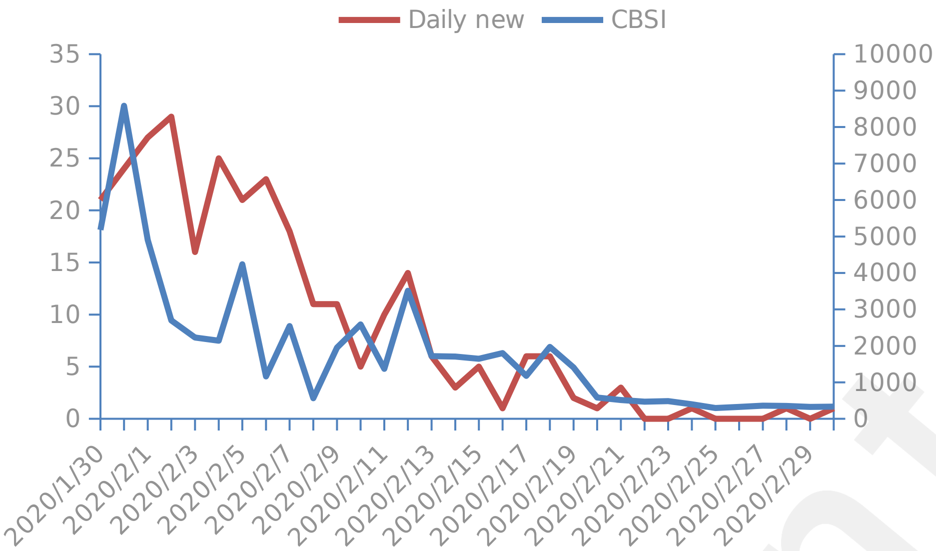


Figure 5. CBSI and Daily COVID-19 case counts at lag -11days

The linear regression model

A linear regression model was established to predict the daily new cases with CBSI. The values of CBSI at negative 11-day lag and COVID-19 case counts was taken and then transformed into logarithms. The model was fitted with independent variable of logarithm of CBSI and dependent variable of logarithm of COVID-19 case counts according to equation (3). The coefficient (β 1) for the linear regression model was 1.042 (P < .001). The R^2 was 0.7, suggesting that the search Index could explain 70% of the variation in daily case counts. The scatter fitting regression diagram was shown in figure 6.

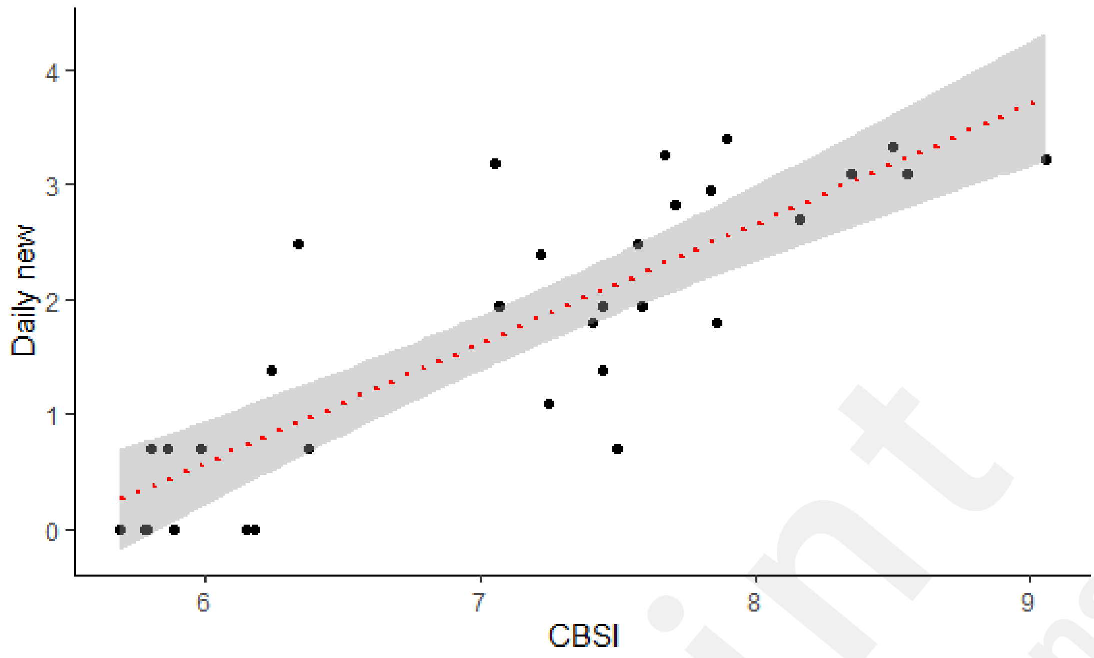


Figure 6. Linear regression fitting diagram of logarithm of CBSI at negative 11-day lag and logarithm of COVID-19 case counts

Discussion

Main Results

This paper mainly studied the correlation between the number of daily confirmed COVID-19 cases and the composite Baidu search index of COVID-19(CBSI) related keywords from January 20, 2020 to March 1, 2020 in Beijing, to evaluate the feasibility of using search data to monitor the development of the epidemic. A preliminary study had also been made on the correlation between low temperature and the number of COVID-19 cases to assist in epidemic surveillance. The results showed that there was a high negative correlation between low temperature and the daily confirmed COVID-19 cases, while a high positive correlation between CBSI and the daily COVID-19 case counts. The positive correlation between CBSI search data 2-13 days in advance and the current number of daily COVID-19 cases was statistically significant, among which the CBSI 11 days earlier had the strongest correlation with the current number of daily COVID-19 cases. After taking the logarithm of the former two, a linear regression model was established, and the regression coefficient β 1 was equal to 1.042 and R^2 was equal to 0.7.

Temperature is an essential factor in people's living environment and plays a significant role in the development and control of epidemic in public health[16,23,24]. A specific temperature may be most suitable for the reproduction of the virus, and the lower temperature contributes to its spread. Chin et al. [25] showed that SARS-CoV-2 survived for a long time at 4 ℃, but its resistance was 5 minutes at 70 ℃, and confirmed that SARS-CoV-2 was more resistant than other viruses on smooth surfaces, such as steel and plastic. DavidN.Prata et al[26]. studied the relationship between temperature in the capitals of Brazilian states and COVID-19 infection, and found a negative linear relationship.

The results of this study indicated that temperature played an important role in the outbreak of COVID-19 in Beijing. The lower the temperature is, the higher the number of daily confirmed cases. The reason for this may be that the virus survives longer at low temperatures and is more likely to spread through large droplets or contact. This result was consistent with previous studies on SARS. In particular, the analysis of SARS data and climate in four Chinese cities revealed that temperature was a powerful indicator of the spread of SARS-CoV, and low temperature increased the risk of daily incidence [27]. Similar results have also been obtained from the study by AurelioTobías et al. on the relationship between COVID-19 and temperature in Barcelona, Spain[28]. This could provide a clue for understanding the temperature-transmission relationship of COVID-19. So the predicted weather conditions, together with other monitoring results, can be used to estimate the propagation of COVID-19.

In recent years, the Internet-based monitoring system, as an effective and innovative method to improve the prevention and control of infectious diseases, has been increasingly explored. For example, online digital disease surveillance tools based on Google trends and Google observations have been mined and reported by some studies[9,29,30]. However, investigations on Baidu search data are not numerous. The current research attempted to explore the possibility and application of Baidu search data for timely and sensitive monitoring of the COVID-19 epidemic in Beijing. Our results clearly showed that there was a positive correlation between the occurrence of COVID-19 and Baidu search data, which provided the possibility to use search query monitoring data to further supervise the occurrence of COVID-19.

The results of cross-correlation can indicate the extent to which Internet-based data can give early warning of disease outbreaks. Our findings indicated that the CBSI of COVID-19 related search term can be present and be increasing 2-13 days before the epidemic occurs. The peak emerged at the 11 day earlier. To some extent, this result is consistent with that of Cuilian Li's research which found that the peak of the online search data about COVID-19 was 10-14 days ahead of the daily incidences peak in China[31]. Generally, because a confirmatory laboratory test takes two or three days at first, the real lag time could be two or three days. If the negative results was false due to improper sampling site, low viral load and virus mutation were taken into account[32], then the diagnosis lag might be longer. Thus, search engine data may reflect the actual disease outbreaks earlier than conventional monitoring, as many people use internet searches to obtain health information before consulting a doctor[33-35]. If people experienced symptoms similar to COVID-19, such as fever, cough, or fatigue, or suspected that they were close contacts, people usually wanted to determine whether they may be COVID-19, so they as well as their relatives often used online search engines, such as Baidu, for information retrieval[36]. In this study, 52% of the remaining keywords were about symptoms, which to some extent confirmed the above view. Given the uncertain conditions associated with emerging diseases, such earlier information for infectious diseases surveillance will help make decisions related to disease prevention and treatment. This kind of digital monitoring system has a more obvious effect on the communicable diseases little known by population. Through the submitted operation of the public, a good disease surveillance effect can be achieved. In addition, there are many advantages of using search engines for digital surveillance: the data can be acquired earlier, more easily, and at a lower cost than by traditional surveillance techniques[9,37,38]. Therefore, in order to improve the performance of disease surveillance, it is essential to combine digital monitoring systems on the basis of traditional monitoring systems.

The linear regression model based on the data with logarithm of CBSI and COVID-19 showed that R^2 was 0.7, indicating that Baidu search data could reflect 70% of the variation in the number of COVID-19 cases 11 days later. That is to say, the change in the number of COVID-19 cases was relevant to the increased behavior of searching for keywords about the disease on the Internet 11 days before. This suggests that COVID-19 composite Baidu search index may be a good monitoring index for early detection of COVID-19 epidemic.

Limitations

This study has certain limitations. First, Baidu will not release search data for keywords when there is not enough search volume, which may lead to underestimation of relevance. Second, although the selected keywords capture the trend of epidemic data well, due to the diversity of online search habits, there may be some omissions. Third, many factors affect individual search behavior, for example, different Internet access levels may affect the accuracy of BSI[8]. Previous research reports have also shown that media biases can adversely affect Internet-based surveillance systems[39,40].

Conclusions and Recommendations

In summary, the search data of COVID-19 obtained by Baidu search index may be a good monitoring indicator for early detection of an outbreak of COVID-19, especially when combined with temperature change monitoring.

So far, most research on Internet-based surveillance systems is a retrospective analysis of performance, and few studies have explored how to transform Internet-based surveillance systems into public health responses. Therefore, the key to future research is to integrate the digital surveillance system into the traditional surveillance system. So, how to choose keywords more scientifically, how to incorporate influencing factors into models to improve the accuracy and sensitivity of surveillance, and how to conduct public health responses based on surveillance data and so on are all problems to be solved. Moreover, global surveillance is also a future research direction.

Supplementary Table 1. Initial keywords and Baidu recommended keywords are presented

Initial keywords	Baidu recommended keywords	
1 □□□□□□□□ (Novel coronavirus pneumonia)	1 □□□□□□□□□□ (Symptoms of novel coronavirus pneumonia)	37 □□□□□□□□□□□□ (Early symptoms of novel coronavirus pneumonia)
2 □□□□□□ (COVID-19)	2 □□□□□□□□ (Novel coronavirus pneumonia)	38 □□□□□□□□□□□□ (How to prevent and treat COVID-19)
3 □□□□ (NCP)	3 □□□□□□□□ (News of COVID-19)	39 □□□□□□□□□□ (Symptoms of COVID-19 virus)
	4 □□□□□□□□□□ (Latest news of COVID-19)	40 □□□□□□□□□□□ (Handwritten report on prevention of COVID-19)
	5 □□□□□□□□□□ (The symptoms of COVID-19)	41 □□□□□□□□□□□ (How long is the incubation period for COVID-19)
	6 □□□□□□□□□□□□□□ (What are the symptoms of COVID-19)	42 □□□□□□□□□□ (Early stage of novel coronavirus pneumonia)
	7 □□□□□□□□□□□□ (Latest news on the novel coronavirus pneumonia)	43 □□□□□□□□□□□ (Incubation symptoms of COVID-19)
	8 □□□□□□□□ (COVID-19 self-test)	44 □□□□□□□□ (Fujian COVID-19)
	9 □□□□□□□□□□ (Novel Coronavirus Pneumonia outbreak)	45 □□□□□□□□□□□ (Handwritten report on fighting COVID-19)
	10 □□□□□□□□□□ (Initial symptoms of COVID-19)	46 □□□□□□□□□□□ (The incubation period for novel coronavirus pneumonia)
	11 □□□□□□ (COVID-19)	47 □□□□□□□□□□□ (The characteristics of COVID-19)
	12 □□□□□□□□□□□□ (Can a dry cough be COVID-19)	48 □□□□□□□□□□□□□□□□ (What symptoms are novel Coronavirus infected pneumonia)
	13 □□□□□□□□ (COVID-19 viruses)	49 □□□□□□□□□□□□ (The symptoms of novel coronavirus pneumonia are infectious)
	14 □□□□□□□□□□□ (What is COVID-19)	50 □□□□□□□□□ (The incubation period for COVID-19)
	15 □□□□□□□□□ (COVID-19 Manuscript)	51 □□□□□□□□□□ (Symptoms of novel coronavirus pneumonia)
	16 □□□□□□□□□□□ (Temperature of fever in COVID-19)	52 □□□□□□□□ (The performance of COVID-19)
	17 □□□□□□□□□□□□ (COVID-19 fever intensity)	53 □□□□□□□□□□□□□ (Pneumonia symptoms of novel coronavirus infection)
	18 □□□□□□□□□□□ (COVID-19 Manuscript Picture)	54 □□□□□□□□□□□□ (Symptoms of pneumonia infected by novel coronavirus)
	19 □□□□□□□□□□□□ (COVID-19 epidemic progress)	55 □□□□□□□□□□□□□□□□ (What are the symptoms of COVID-19)
	20 □□□□□□□□□□□□ (Early symptoms of COVID-19)	56 □□□□□□□□□□□□ (What is the transmission of new coronary pneumonia?)
	21 □□□□□□□□ (Wuhu COVID-19)	57#□□□□□□□□□□□# (# Novel Coronavirus infected pneumonia #)
	22 □□□□□□□□□□ (Real-time dynamics of CIVID-19)	58 □□□□□□□□□□ (Diagnosis of novel coronavirus pneumonia)
	23 □□□□□□□□□□□ (Novel Coronavirus Pneumonia Manuscript)	59 □□□□□□□□□□□ (COVID-19's incubation period)
	24 □□□□□□□□□□□□□□ (COVID-19 epidemic situation combing)	60 □□□□□□□□□□□□□□□□ (Medical staff were infected with COVID-19)
	256 □□□□□□□□□□□□ (Latest situation in noval coronary pneumonia)	61 □□□□□□□□□□□□□□□□□□□ (COVID-19 is included in the management of legal infectious diseases)
	26 □□□□□□□□□□ (Yancheng COVID-19)	62 □□□□□□□□□□□□□□□□□
	27 □□□□□□□□□□□□ (Latest news of COVID-19)	
	28 □□□□□□□□□□□	

	(Pneumonia infected by Novel Coronavirus) 29#□□□□□□□□□□□□□# (#How to prevent and treat COVID-19#) 30 □□□□□□□□ (COVID-19 symptoms) 31 □□□□□□□□□□ (Novel coronavirus-infected pneumonia) 32 □□□□□□□□□□□□ (Real-time status of COVID-19 epidemic) 33 □□□□□□□□□□ (Latest situation of COVID-19) 34 □□□□□□□□□□ (How is COVID-19 diagnosed) 35 □□□□□□□□□□□ (Content of COVID-19 handwritten report) 36 □□□□□□□□□□□□ (Covid-19 mamuscript of Elementary school students)	(Ministry of Foreign Affairs responds to COVID-19 epidemic) 63 □□□□□□□□□□ (Real-time rumor refutation of COVID-19) 64 □□□□□□□□□□□□□□□ (U.S. confirmed second case of COVID-19) 65 □□□□□□□□□□□□□□□□ (France confirms two cases of COVID-19) 66 □□□□ (NCP) 67 □□□□□□□□□□□□ (What are the symptoms of novel coronavirus pneumonia) 68 □□□□□□ (The symptoms of COVID - 19) 69 □□□□□□□□□□□ (Is cough COVID-19?) 70 □□□□□□□□□□□ (Could the cough be COVID-19)

Supplementary Table 2. Preliminary screening keywords and remaining keywords are presented

Preliminary screening keyword	Remaining keywords
1 □□□□□□□□□□ (Symptoms of novel coronavirus pneumonia)	1 □□□□□□□□ (Novel coronavirus pneumonia)
2 □□□□□□□□ (Novel coronavirus pneumonia)	2 □□□□□□□□□□□□□□ (What are the symptoms of COVID-19)
3 □□□□□□□□□□ (The symptoms of COVID-19)	3 □□□□□□ (COVID-19)
4 □□□□□□□□□□□□□□ (What are the symptoms of COVID-19)	4 □□□□□□□□□□□ (Pneumonia infected by Novel Coronavirus)
5 □□□□□□□□ (COVID-19 self-test)	5#□□□□□□□□□□□□# (# How to prevent and treat COVID-19#)
6 □□□□□□□□□□ (Initial symptoms of COVID-19)	6 □□□□□□□□ (COVID - 19 symptoms)
7 □□□□□□ (COVID-19)	7 □□□□□□□□□□ (Novel coronavirus-infected pneumonia)
8 □□□□□□□□□□□□ (Can a dry cough be COVID-19)	8 □□□□□□□□□□ (How is COVID-19 diagnosed)
9 □□□□□□□□□□□ (What is COVID-19)	9 □□□□□□□□□□□□ (Early symptoms of novel coronavirus pneumonia)
10 □□□□□□□□□□□ (How many degrees does COVID-19 fever have)	10 □□□□□□□□□□□□ (How to prevent and treat COVID-19)
11 □□□□□□□□□□□□ (Covid-19 fever intensity)	11 □□□□□□□□□□ (Symptoms of COVID-19 virus)
12 □□□□□□□□□□□□ (Early symptoms of COVID-19)	12 □□□□□□□□□□□ (How long is the incubation period for COVID-19)
13 □□□□□□□□□□□ (Pneumonia infected by Novel Coronavirus)	13 □□□□□□□□□□□ (Incubation symptoms of COVID-19)
14#□□□□□□□□□□□□# (#How to prevent and treat COVID-19#)	14 □□□□□□□□□□□ (The characteristics of COVID - 19)
15 □□□□□□□□ (COVID-19 symptoms)	15 □□□□□□□□□□□□□□□□ (What symptoms are novel Coronavirus infected pneumonia)
16 □□□□□□□□□□ (Novel coronavirus-infected pneumonia)	16 □□□□□□□□□□□□ (The symptoms of novel coronavirus pneumonia are infectious)
17 □□□□□□□□□□ (How is COVID-19 diagnosed)	17 □□□□□□□□□ (The incubation period for COVID-19)
18 □□□□□□□□□□□□ (Early symptoms of novel coronavirus pneumonia)	
19 □□□□□□□□□□□□	18 □□□□□□□□□□

(How to prevent and treat COVID-19)
20 □□□□□□□□□□
(Symptoms of COVID-19 virus)
21 □□□□□□□□□□□
(How long is the incubation period for COVID-19)
22 □□□□□□□□□□
(Early stage of novel coronavirus pneumonia)
23 □□□□□□□□□□□
(Incubation symptoms of COVID-19)
24 □□□□□□□□□□□
(The incubation period for novel coronavirus pneumonia)
25 □□□□□□□□□□□
(The characteristics of COVID-19)
26 □□□□□□□□□□□□□□□□
(What symptoms are novel Coronavirus infected pneumonia)
27 □□□□□□□□□□□□
(The symptoms of novel coronavirus pneumonia are infectious)
28 □□□□□□□□□
(The incubation period for COVID-19)
29 □□□□□□□□□□
(Symptoms of novel coronavirus pneumonia)
30 □□□□□□□□
(The performance of COVID-19)
31 □□□□□□□□□□□□□
(Pneumonia symptoms of novel coronavirus infection)
32 □□□□□□□□□□□□
(Symptoms of pneumonia infected by novel coronavirus)
33 □□□□□□□□□□□□□□□□
(What are the symptoms of COVID-19)
34#□□□□□□□□□□□#
(# Novel Coronavirus infected pneumonia #)
35 □□□□□□□□□□
(Diagnosis of novel coronavirus pneumonia)
36 □□□□□□□□□□□
(COVID-19's incubation period)
37 □□□□
(NCP)
38 □□□□□□□□□□□□
f(What are the symptoms of novel coronavirus pneumonia))
39 □□□□□□
(The symptoms of COVID - 19)
40 □□□□□□□□□□□
(Is cough COVID-19?)
41 □□□□□□□□□□□
(Is coughing a COVID-19)

(Symptoms of novel coronavirus pneumonia)
19 □□□□□□□□
(The performance of COVID - 19)
20 □□□□□□□□□□□□□
(Pneumonia symptoms of novel coronavirus infection)
21 □□□□□□□□□□□□
(Symptoms of pneumonia infected by novel coronavirus)
22 □□□□□□□□□□□□□□□□
(What are the symptoms of COVID-19)
23#□□□□□□□□□□□#
(#Novel Coronavirus infected pneumonia#)
24 □□□□□□□□□□
(Diagnosis of novel coronavirus pneumonia)
25 □□□□□□□□□□□
(COVID-19's incubation period)

Acknowledgements
WSZ was funded by the National Nature Science Foundation of China (Grant Number 81872719) and the Shandong Province Natural Science Foundation (ZR201807090257).

Authors' Contributions
QX completed the conception and design of the manuscript, data acquisition and sorting, statistical analysis, and the manuscript writing. FJN and ZGP conducted the feasibility of the project. LYJ, MQ, WZ, and GPX extracted and sorted the materials and reviewed the manuscript. WSZ carried out the revision, quality control and proofreading of the manuscript. All authors dedicated to the editing and development of the paper.

Conflicts of Interest
None declared.

Abbreviations
COVID-19: coronavirus disease
CBSI: COVID-19 composite Baidu search index
MERS: Middle East respiratory syndrome
SARS: severe acute respiratory syndrome

References

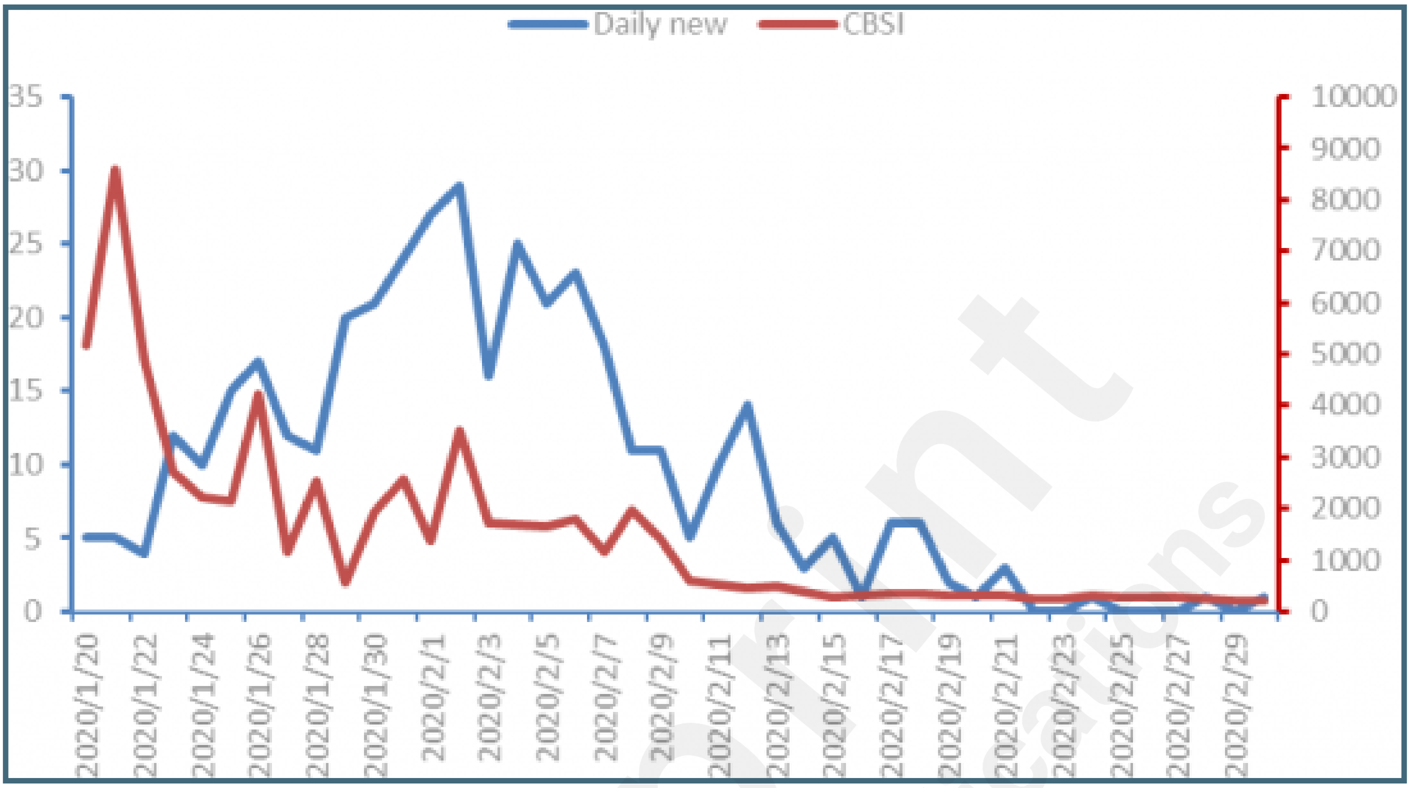
1. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. *Nat Med.* Dec 2004;10(12 Suppl):S88-97.PMID:15577937
2. Shin SY, Seo DW, An J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Sci Rep.* Sep 6 2016;6:32920.PMID:27595921
3. Shanmugaraj B, Siriwattananon K, Wangkanont K, Phoolcharoen W. Perspectives on monoclonal antibody therapy as potential therapeutic intervention for Coronavirus disease-19 (COVID-19). *Asian Pacific journal of allergy and immunology.* Mar 2020;38(1):10-18.PMID:32134278
4. Peng X, Xu X, Li Y, Cheng L, Zhou X, Ren B. Transmission routes of 2019-nCoV and controls in dental practice. *Int J Oral Sci.* Mar 3 2020;12(1):9.PMID:32127517
5. Garcell HG, Hernandez TMF, Abdo EAB, Arias AV. Evaluation of the timeliness and completeness of communicable disease reporting: Surveillance in The Cuban Hospital, Qatar. *Qatar medical journal.* 2014;2014(1):50-56.PMID:25320693
6. L WJLNCM. Epidemic prevention and control of COVID-19 brings thinking for the general medicine. *Chinese General Practice.* 2020;23(09):1090-1094
7. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital Health Strategies to Fight COVID-19 Worldwide: Challenges, Recommendations, and a Call for Papers. *Journal of Medical Internet Research.* 2020;22(6):e19284
8. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet Infectious Diseases.* 2014/02/01/ 2014;14(2):160-168
9. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature.* Feb 19 2009;457(7232):1012-1014.PMID:19020500
10. Zhang Y, Bambrick H, Mengersen K, Tong S, Hu W. Using Google Trends and ambient temperature to predict seasonal influenza outbreaks. *Environ Int.* Aug 2018;117:284-291.PMID:29778013
11. 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-e64323.PMID:23750192
12. Gu Y, Chen F, Liu T, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. *Sci Rep.* Jul 28 2015;5:12649.PMID:26218589
13. Wen X, Jiang Y. Research on the Distribution of Media Attention to the "Second Child" : Based on Baidu Index Big Data. *News Tribune.* 2018(6):24-27(in Chinese)
14. Committee GOoNH. Notice on the issuance of a program for the diagnosis and treatment of novel coronavirus (2019-nCoV) infected pneumonia (trial sixth edition). 2020-02-19; URL:http://yzs.satcm.gov.cn/zhengcewenjian/2020-02-19/13221.html.
15. Baidu index. 2020; URL:http://index.baidu.com, [Accessed 2020-03-05].
16. Hemmes JH, Winkler KC, Kool SM. Virus Survival as a Seasonal Factor in Influenza and Poliomyelitis. *Nature.* 1960/10/01 1960;188(4748):430-431
17. tianqihoubao. URL:http://www.tianqihoubao.com/, [Accessed 2020-03-05].
18. Liu Y, Lv B, Peng G, Yuan Q. A preprocessing method of internet search data for prediction improvement: application to Chinese stock market. *Proceedings of the Data Mining and Intelligent Knowledge Management Workshop*. Beijing, China: Association for Computing Machinery; 2012:Article 3.
19. Luo Y, Zeng D, Cao Z, et al. Using multi-source web data for epidemic surveillance: A case study of the 2009 Influenza A (H1N1) pandemic in Beijing. Paper presented at: international conference on service operations and logistics, and informatics2010; Beijing

,China.
20. Zhou X, Shen H. Notifiable infectious disease surveillance with data collected by search engine. *Journal of Zhejiang University Science C.* 2010;11(4):241-248
21. Baidu keyword mining. URL:http://stool.chinaz.com/baidu/words.aspx, [Accessed 2020-03-05].
22. Li J-xBB-fLGPN. Gonorrhea incidence forecasting research based on Baidu search data. Paper presented at: 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings; 17-19 July 2013, 2013.
23. McMichael A, Wilkinson P, Kovats S, et al. International study of temperature, heat and urban mortality: The 'ISOTHURM' project. *International journal of epidemiology.* 2008;37:1121-1131
24. Banu S, Hu W, Guo Y, Hurst C, Tong S. Projecting the impact of climate change on dengue transmission in Dhaka, Bangladesh. *Environment International.* 2014/02/01/ 2014;63:137-142
25. Chin A, Chu J, Perera M, et al. Stability of SARS-CoV-2 in different environmental conditions. *medRxiv.* 2020
26. Prata DN, Rodrigues W, Bermejo PH. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. *Sci Total Environ.* Aug 10 2020;729:138862.PMID:32361443
27. Tan J, Mu L, Huang J, Yu S, Chen B, Yin J. An initial investigation of the association between the SARS outbreak and weather: with the view of the environmental temperature and its variation. *Journal of Epidemiology and Community Health.* 2005;59(3):186
28. Tobias A, Molina T. Is temperature reducing the transmission of COVID-19 ? *Environ Res.* Apr 18 2020;186:109553.PMID:32330766
29. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. *PLoS Negl Trop Dis.* Feb 2014;8(2):e2713.PMID:24587465
30. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. *Science.* 2014;343(6176):1203-1205
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. *Eurosurveillance.* 2020;25(10)
32. Udugama B, Kadhiresan P, Kozlowski HN, et al. Diagnosing COVID-19: The Disease and Tools for Detection. *ACS Nano.* Apr 28 2020;14(4):3822-3835.PMID:32223179
33. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am J Prev Med.* May 2011;40(5 Suppl 2):S154-158.PMID:21521589
34. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J Med Internet Res.* Jul 18 2013;15(7):e147.PMID:23896182
35. Luo L, Zeng X, Liao X, Yang Y. Disease cognition, coping style and exercise behavior among the public during novel coronavirus epidemic: an online survey. *Chin J Public Health.* 2020;36(2):156-159(in Chinese)
36. Qiu H-j, .Yuan L-x, . Huang X-k, et al. Using the big data of internet to understand the characteristics of coronavirus disease 2019: a big data study. *Chin J Otorhinolaryngol Head Neck Surg.* 2020;55(6):569-575(in Chinese)
37. Seo D-W, Jo M-W, Sohn CH, et al. Cumulative Query Method for Influenza Surveillance Using Search Engine Data. *Journal of Medical Internet Research.* 2014;16(12):e289
38. Ortiz JR, HZ, David K. Shay, KMN, ALF, CH, Goss. Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. *PLoS ONE.* 2011;6(4):e18687-e18687
39. Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. *PLOS Neglected Tropical Diseases.* 2011;5(8):e1258
40. Zhang Y, Milinovich G, Xu Z, et al. Monitoring Pertussis Infections Using Internet Search Queries. *Scientific reports.* 2017;7(1):10437-10437.PMID:28874880

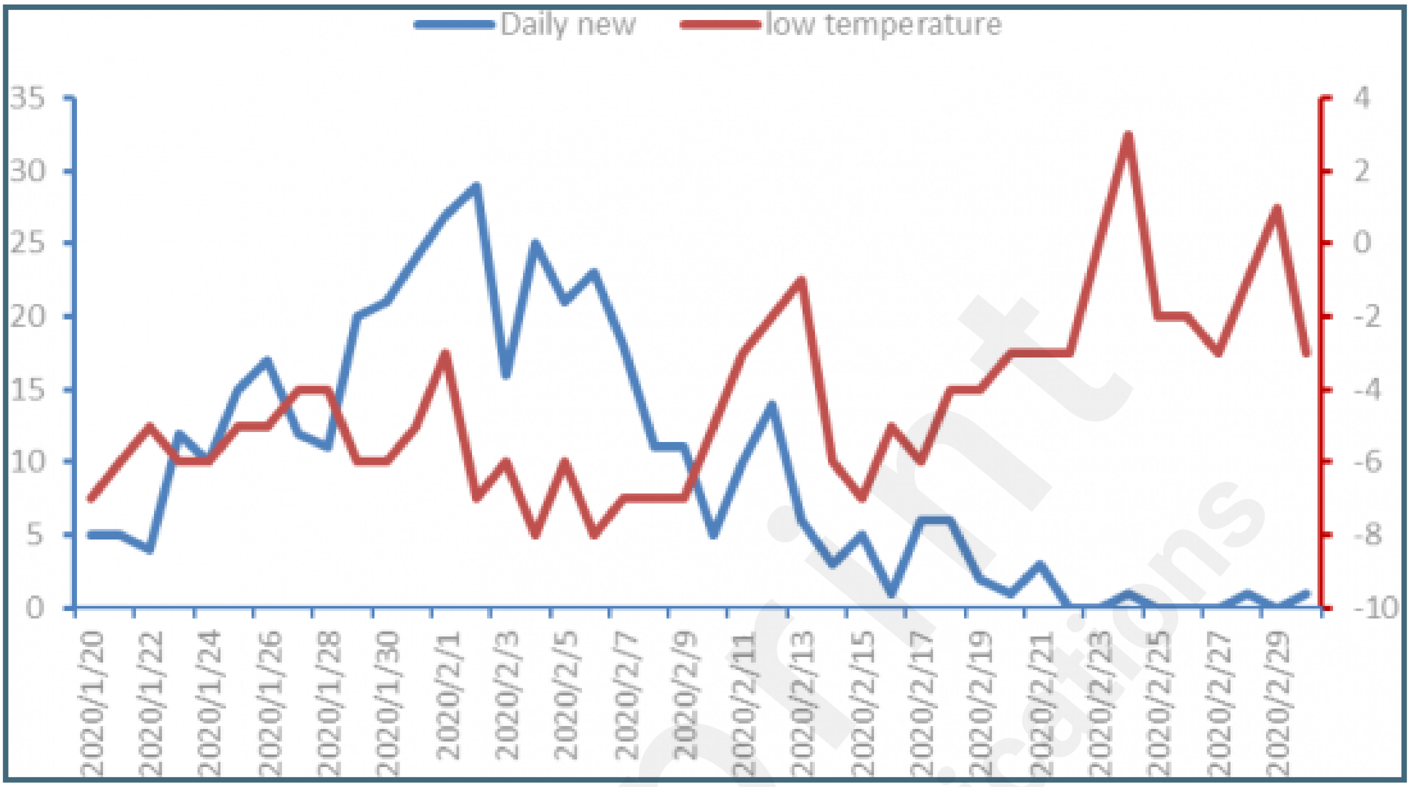
Supplementary Files

Figures

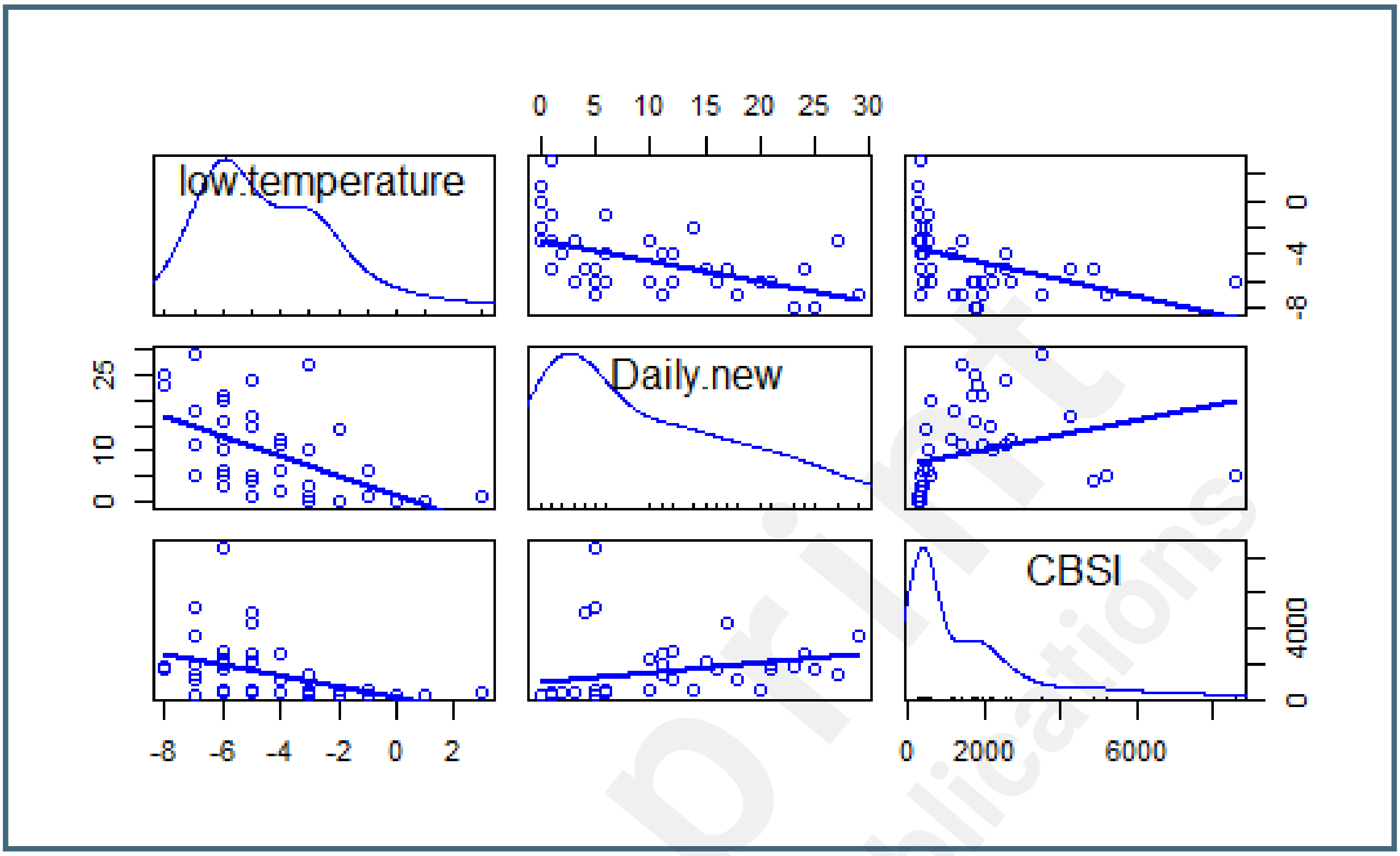
CBSI and Daily COVID-19 case counts during the study period.



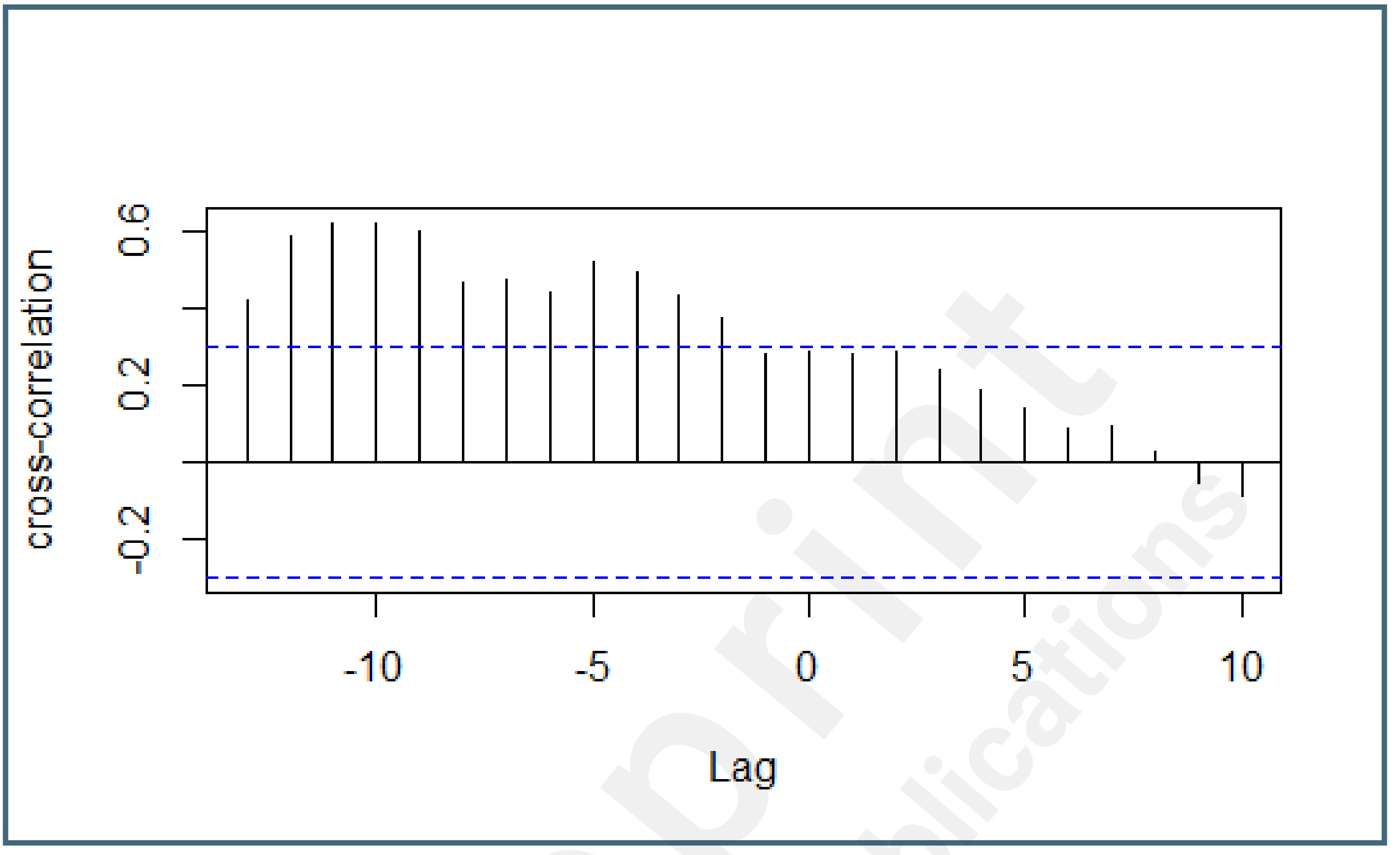
Low temperature and Daily COVID-19 case counts during the study period.



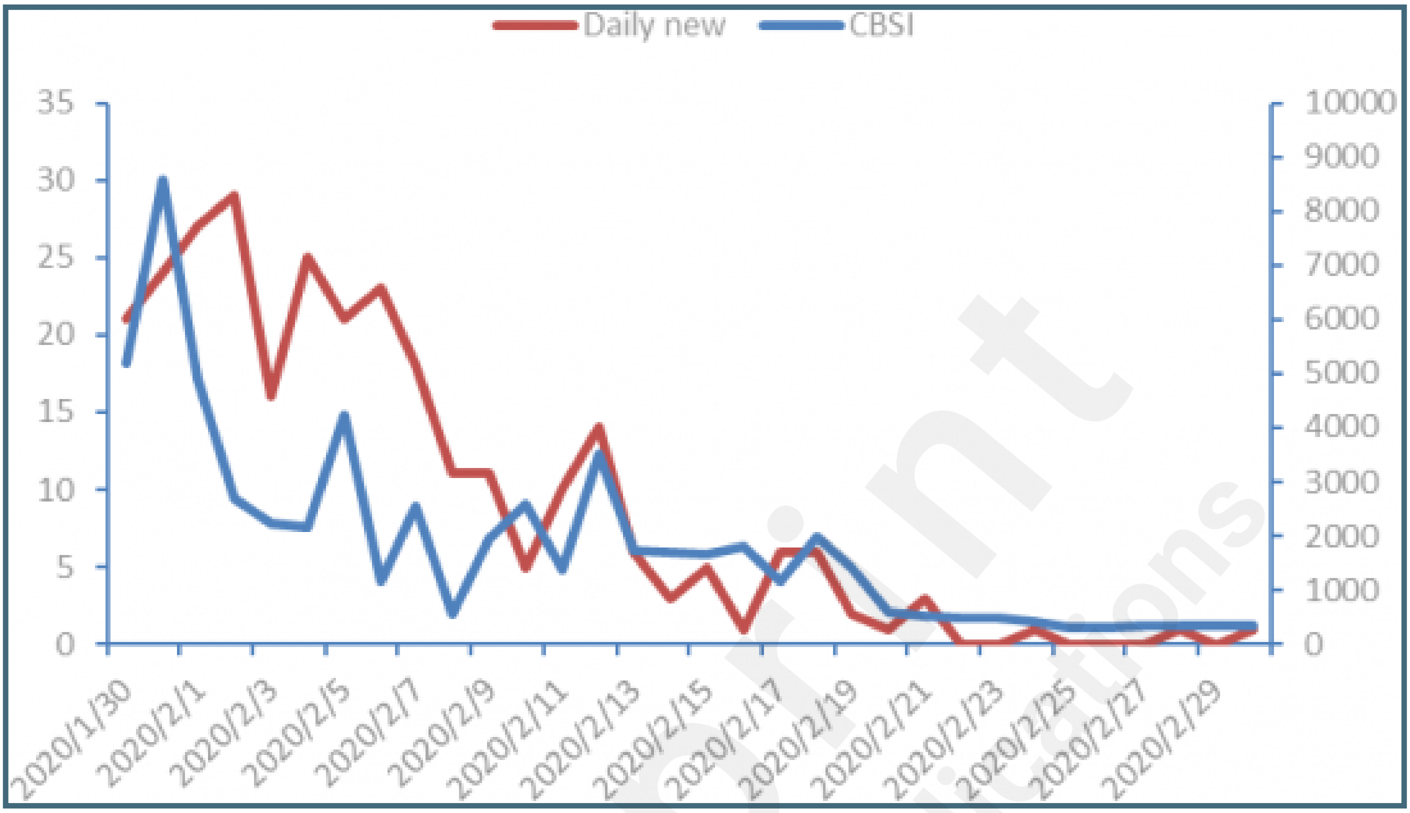
Scatterplot matrix among CBSI, low temperature and Daily COVID-19 case counts.



Time series cross-correlation between CBSI and Daily COVID-19 case counts.



CBSI at lag -11days and Daily COVID-19 case counts.



Linear regression fitting diagram of logarithm of CBSI at negative 11-day lag and logarithm of COVID-19 case counts.