Infectious Disease Modelling 5 (2020) 848–854

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Internet search data could Be used as novel indicator for assessing COVID-19 epidemic

Kang Li [a, b, 1], Yanling Liang [b, d, 1], Jianjun Li [c], Meiliang Liu [d], Yi Feng [b], Yiming Shao [a, b, c, *]

[a] Key Laboratory of Molecular Microbiology and Technology, Ministry of Education, College of Life Sciences, Nankai University, Tianjin, China
[b] State Key Laboratory for Infectious Disease Prevention and Control, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China
[c] Guangxi Center for Disease Prevention and Control, Nanning, Guangxi, China
[d] School of Public Health, Guangxi Medical University, Nanning, Guangxi, China

ARTICLE INFO

Article history:
Received 19 July 2020
Received in revised form 6 September 2020
Accepted 1 October 2020
Available online 3 October 2020
Handling editor: Dr. J Wu

Keywords:
COVID-19
Internet data
ARIMAX
Predict

ABSTRACT

The pandemic of the coronavirus disease (COVID-19) poses a huge challenge all countries, since no one is well prepared for it. To be better prepared for future pandemics, we evaluated association between the internet search data with reported COVID-19 cases to verify whether it could become an early indicator for emerging epidemic. After the keyword filtering and Index composition, we found that there were close correlations between Composite Index and suspected cases for COVID-19 ($r = 0.921$, $P < 0.05$). The Search Index was applied for the Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) model to quantify the relationship. Compared with the model based on surveillance data only, the ARIMAX model had smaller Akaike Information Criterion (AIC = 403.51) and the most accurate predictive values. Overall, the Internet search data could serve as a convenient indicator for predicting the epidemic and to monitor its trends.

Introduction

In the last 100 years, there have been several times large-scale outbreaks of a major epidemic causing by coronavirus, namely, the Spanish Influenza Pandemic of 1918–1919, the Asian Influenza Pandemic of 1957–1958, SARS in mainland China of 2003 and MERS in Saudi Arabia of 2012 (Bi et al., 2020; Yin and Wunderink, 2018). Since the first case of a new type coronavirus infections was found in Wuhan, Hubei, in December 2019 (Wang et al., 2020). The viral infection is spreading fast with an increasing number of infected patients nationwide within a short period of time, and it can also cause serious illness and death. On February 11, 2020, the new novel coronavirus pneumonia was officially named “Corona Virus Disease 2019”

* Corresponding author. Division of Research on Virology and Immunology National Center for AIDS/STD Control and Prevention China CDC, Beijing, 102206, China.
E-mail address: yshao@bjmu.edu.cn (Y. Shao).
Peer review under responsibility of KeAi Communications Co., Ltd.
[1] These authors contributed equally to this article.

https://doi.org/10.1016/j.idm.2020.10.001

(Wertheim et al., 2014) (COVID-19) by the World Health Organization. Various investigations and research indicate that the COVID-19 has higher infection rate and long incubation period characteristics. Low fever, cough and fatigue are the most common initial symptoms of infection (Chinazzi, Davis, & Ajelli, 2020; Huang et al., 2020), which were similar to the seasonal influenza symptoms, more serious is some infected patients are asymptomatic, so it could spread widely without protection consciousness. At the same time, the counts of infected patients is increasing in other countries, especially in South Korea, Japan, the United States, Iran and Italy (Brown et al., 2011; OrganizationCoronavi, 2020). The prevention and control of COVID-19 has become a global arduous task.

By reason of the relatively longer incubation period of COVID-19 and the poor sanitary inspection conditions in some regions, which result in a time lag between real epidemic and the numbers of suspected case (Backer et al., 2020; Linton et al., 2020). Moreover, the suspected cases also need a time to be final confirmed and reported in the official published numbers. Nowadays, comparing to other conventional infectious diseases, the outbreak is more severe in COVID-19 (Surveillances, 2020). On account of time and local resource restrains, insufficiently operational knowledge of reporting systems, and the lag in gather data and reporting is an obstacle to the real-time reliability evaluation and control of the COVID-19 epidemic (Lipsitch, Swerdlow, & Finelli, 2020; Milinovich et al., 2014). Therefore, in this situation when the illness does not have any specific treatment, we need to established an early assessment systems and preparation in healthcare services to avoid such disasters.

With the remarkable development of the Internet, search engine data have been increasingly used to track and predict the outbreak of diseases in recent years. The research by Polgreen (Polgreen et al., 2008) and Hulth (Hulth et al., 2009) used the Yahoo search engine and medical website to predict the incidence of influenza. In 2009, the Google Flu Trends, which successfully predicted flu outbreaks by assessed the search volume of several keywords (Ginsberg, 2009). Moreover, the real-time monitoring of disease trends explored on dengue fever in Singapore and Bangkok by using the Internet-based Google Insights from 2004 to 2011 (Althouse et al., 2011). Consequently, we can utilize the search data provided by the Baidu search index (BDI) to evaluate the trend of COVID-19 epidemic in China.

We compare the Internet keywords search data with the number of regular epidemic reports in the same period, to analyze and evaluate the correlation between of them. Moreover, the ultimate goal of this study is to examine whether an exogenous variable (BDI) could enhance the forecast accuracy and stability of the models based on suspected cases data from surveillance system of COVID-19. Our research will provide a new reference approaches and indicators for real-time monitoring and prevention of COVID-19, which will also help to more effectively promote the accurate assessment and effective intervention of the COVID-19 epidemic.

Materials and methods

Data sources. Two types of data are needed: one is numbers of suspected case of COVID-19 in China, and the other is search query data from Baidu for the nationwide. We gathered the two kinds of data for the outbreak periods from January 20, 2020 to February 20, 2020. The count of suspected cases is publicly available on China Centre for Disease Control and Prevention's (China CDC) daily status report of 2019-nCoV (COVID-19). Daily search engine query data were obtained through the Baidu index (http://index.baidu.com), a sharing platform of big data for public.

Keyword selection and Filtering As different key words have different search frequencies at different times and regions and diverse modeling outcomes can be obtained by selecting different keywords, thus, keywords selection is the crucial issue in internet search data-based surveillance. Although previous studies generally chose the relevant names or clinical characteristic of target illness as their critical keywords (Kang et al., 2013; Yuan et al., ; Zhou, Shen), however, there are still no principles or criteria for guidance. Therefore, we selected the core keywords carefully to reflect terms most likely associated with COVID-19. Primary keywords used in this study were obtaining from a Chinese website: https://ci.aizhan.com/and did some further analysis of gathered keywords (Yuan et al., ; Liu et al., 2016a). Keywords suggested by using semantic correlation analysis not only from Baidu, but also from Microblog, Wikipedia, WeChat and other portal websites (Yuan et al.,). Finally, we obtain 14 related keywords about COVID-19 search behavior. But more keywords do not necessarily produce a greater result in previous researches (Ginsberg, 2009; Hulth et al., 2009), since some suggested keywords are not closely related to COVID-19, which could reduce the ability of model fit. Hence, we gathered a diversity of COVID-19 related keywords and filtered it following two steps:

First, The Spearman correlation analysis was applied between the suspected case and the BDI of the keywords. We removed the keywords with a maximum correlation coefficient of less than 0.4 and those correlations with no statistical significance.

Second, we used time-series cross-correlation analysis to inspect the keywords whether having lag effects between the keywords and the suspected cases in different lag periods (Du et al., 2017; Gu et al., 1264). Then, the lag-value with the maximum correlation coefficient of each search keyword was considered to be included in the subsequent analysis of the COVID-19 Search Index composition.

COVID-19 Search Index composition. After screening and deleting, the remained keywords was applied for composition of COVID-19 Search Index for each time lag. Furthermore, we defined weights of keywords by the strength of the maximum correlation coefficient (ρ) (Du et al., 2017; Gu et al., 1264). This approach was generally combined with Analytic Hierarchy Process for a better result, however, it seems adequate to use only the correlation coefficient without adjustment for this

study (Althouse et al., 2011). The weights calculation and COVID-19 Search Index composition equation are as follows (I) and (II):

$$Weight_{ki} = \rho_{ki} \Big/ \sum_{i=1}^{n} \rho_{ki} \tag{I}$$

$$COVID-19\ Search\ Index_k = \sum_{i=0}^{n} weight_{ki}\ keyword_{ki} \tag{II}$$

where k is the potential time lag, n denotes the number of keywords, $keyword_{ki}$ and $weight_{ki}$ represent the ith keyword daily Baidu Index and the weight of it with specific time lag(k) (Gu et al., 1264).

Model Construction. The Autoregressive Integrated Moving Average Model (ARIMA) model was extensively used to predict infectious diseases incidence through the use of historically surveillance cases data (Du et al., 2017; Gu et al., 1264; Ebhuoma et al., 2018; Li et al., 2019). The ARIMAX model is an important technique, which is an extension of the ARIMA model. Compared with the ARIMA model, the ARIMAX model was the combination of multiple regression analysis and time series analysis, it can utilize more information and test the relationship between COVID-19 suspected case and multiple keywords Baidu search index (Granger and Swanson, 1996; Zhao et al., 2020). In addition, the multivariate models need to examine the stationarity, estimation of coefficients and post-model evaluation, moreover, detected the distribution of model residuals and goodness of fit in the process of the model development. The fitness of the ARIMA and ARIMAX models was estimated by Akaike Information Criterion (AIC) to reveal the fitness of models (Chadsuthi et al., ; Cornelsen, Normand). The lower AIC value, the better the goodness of fit of the representative model. Based on this, the following formula was applied to acquire the forecasting disease cases:

$$y_t = \mu + \sum_{i=1}^{k} \frac{\Theta_i(B)}{\Phi_i(B)} B^{l^i} x_{it} + r_i \tag{III}$$

$$r_i = \frac{\Theta_i(B)}{\Phi_i(B)} a_t \tag{IV}$$

In (III), index $\Phi_i(B)$ and l^i, represents the autoregressive coefficients polynomial, moving average coefficient polynomial and a lag operator of the ith input variables separately; x_{it} is external variables *Log* (*COVID-19 Search Index*$_k$); r_i was the regression residual sequence. In (IV), $\Theta_i(B)$ and $\Phi_i(B)$ are denotes the rest of autoregressive coefficient polynomial and the residual moving average coefficient polynomial, respectively; r_i is the residual sequence, and a_t is the white noise sequence with zero mean. y_t is the dependent variable. We tested the model residuals by the Augmented Dickey–Fuller Unit Root (ADF) test and the white noise test of residuals was also implemented according to the auto correlation plot residuals graph (Granger and Swanson, 1996). In this research, in order to construct the models and evaluate the prediction accuracy, we divided the overall time data into two sub-data sets: the training dataset (twenty-seven days) and the test dataset (five days).

Mean absolute percentage errors (MAPE) was used to examine the accuracy between the predicted and observed of suspected case of COVID-19 (Lee et al., 2005, Zhang et al., 2008), which was calculated as:

$$\mathrm{MAPE} = \sum_{t=1}^{n} \left| \frac{observed_t - predicted_t}{observed_t} \right| \times \frac{100}{n} \tag{V}$$

For the models, the smaller MAPE values revealed a more accurate prediction. All data descriptive analysis was calculated in SPSS (version 22.0). The modeling process of ARIMA and ARIMAX was analyzed by R statistical software (version 3.6.2) using packages of 'forecast', 'psych', 'zoo' and 'TSA'.

Result

Descriptive Analysis of Various Keywords. The COVID-19 relevant search statistics are shown in Table 1, which is performed using Baidu search index data from January 20, 2020 to February 20, 2020 in China. Internet users search messages in Baidu used Chinese and corresponding translation of each Chinese keywords was listed in English. We could see that the keyword "Novel coronavirus disease", "Real-time dynamics of novel coronavirus" and "Novel coronavirus characteristics" had the highest average index of daily search and the keyword "The latest situation of Novel coronavirus", "Fever degree of novel coronavirus pneumonia" and "Novel coronavirus" search index was at a slightly low level among the 14 keywords.

Correlation Analysis and Composite Index. Through the Spearman's rank analysis of correlation between keyword search index and COVID-19 suspected cases, and finally five search keywords with high correlation and statistical difference were selected (Table 2). The time-series comparison curve between the Baidu index of a few keywords and the daily counts of suspected cases in China were demonstrated in Fig. 1. Although the search frequency of the three keywords was different, the

overall trend was the same. Moreover, the search index showed an upward or downward trend when the number of suspected cases of COVID-19 increased or decreased.

Then, the cross-correlation function (CCF) was applied to analysis the correlation between daily suspected case numbers and keywords search index over a range of 10 days-lag values, moreover, we selected the maximum cross-correlation coefficients which reaching a peak at lag period (Table 2). We found that the keywords “Novel coronavirus characteristics”, “What day is the most serious novel coronavirus symptom” and “Early symptoms of Novel coronavirus” were calculated with distinct time-lags.

Ultimately, we estimated the weight of each keyword in the formula, and accumulate the search index according to the weight to form Baidu composite index (COVID-19 Search Index). The coefficient of the Spearman correlation analysis between COVID-19 suspected case and the composite index was 0.921($P < 0.001$), indicating that they were significantly correlated.

Differencing and Autoregressive Integrated Moving Average Model Construction. The temporal dependence change was observed by the plots of autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF). The time series displayed a clear upward trend and volatile trend, which indicated that the series was unstable. (Fig. 2a). Moreover, the unit root test also showed that the sequence was non-stationary ($P > 0.05$).

After the sequence differencing, a stable time series was formed. According to the ACF, PACF and the Akaike Information Criterion order criterion, the following univariate ARIMA (0, 2, 2) regression model was constructed to forecast the suspected case of COVID-19 (Fig. 2b). Furthermore, the model parameter tests were all statistically significant non-zero (Table 3, $P < 0.05$). The Ljung-Box test indicated that the residual series were the white noise sequence (Table 3, $P > 0.05$).

Multivariate time series regression model analysis. In the ARIMAX model, both the input sequence BDI and the output sequence COVID-19 suspected cases were significantly stationary and correlative. The parameters of the ARIMAX models was significantly non-zero ($P < 0.05$) and the residual autocorrelation test for the model couldn't reject the null hypothesis ($P > 0.05$), it exhibited in Table 3. Additionally, the ARIMAX model (AIC = 403.51) with BDI as an external variable revealed better goodness fit than the ARIMA model (AIC = 404.26). Moreover, Fig. 3 indicated that the ARIMAX model had a smaller MAPE than ARIMA model for the predict accuracy (MAPE = 21.5 vs 51.5, a lower value indicating a more accurate prediction).

Discussion

As of February 20, 2020, 75465 reported cases, 5206 suspected cases and 2236 deaths have been reported nationwide in China (Brown et al., 2011). With conducted series of measures such as closed all cities in Hubei province and strengthened travel constraints throughout the country, and now the disease spread is slowing down. Nonetheless, there new cases and new deaths still be reported every day and the epidemic may rebound as people return to work after the Spring Festival holiday. In this case, real-time assessment of the epidemic situation was particularly significant for the rehabilitation treatment of diseases and preparation of epidemic prevention materials in medical services (Chinazzi, Davis, & Ajelli, 2020). Previous researches have shown that collecting and analyzing data from social media, Internet search queries and portal websites maybe effective to early detect disease epidemic (Li et al., 2019; Zhang et al., 2008), which is a complementary to conventional surveillance systems and improve the evaluation methods of the disease.

In our study, by exploring daily internet engine search data from Baidu index, we developed a comprehensive approach and index for monitoring and modeling COVID-19 epidemics in China. Here we found that the search index of different keywords correlates positively with the number of suspected cases. That is, there is a significant correlation between the Baidu search index and the potential epidemic, which indicated that we can assess and forecast the potential epidemic by gathering and analyzing the internet search data. Meanwhile, our results were consistent with some earlier research indicating that using the ARIMAX model with BDI has smaller AIC value and MAPE value, which suggested that using a multivariate ARIMAX model provides greater prediction than a model without variation, and good predictability in terms of stability. Since search queries can be disposed rapidly, an early indicator for monitoring and detection of COVID-19 may

Table 1
Descriptive statistics of Baidu search index with COVID-19 related keywords.

Search Keywords (in Chinese)	Search Keywords (in English)	Minimum	Median	Maximum	Mean	SD
新型冠状病毒的特征	Novel coronavirus characteristics	115001	458759	626452	438054.0	124623.3
新冠状病毒症状	Novel coronavirus symptoms	12753	35416	50153	34220.5	10248.1
冠状病毒的症状	Coronavirus symptoms	54184	187419	322705	174285.1	55701.0
新型冠状病毒	Novel coronavirus disease	266892	1277132	2330851	1338312.7	452666.3
新冠状病毒	Novel coronavirus	16811	82357	173167	84299.0	26848.2
新型冠状病毒传播途径	Novel coronavirus transmission route	18719	35038	101933	38591.2	17011.7
新型冠状病毒肺炎症状	Novel coronavirus pneumonia symptoms	25489	75537	139003	83169.5	35459.7
冠状病毒第几天最严重	What day is the most serious novel coronavirus symptom	4860	32729	139781	38388.4	26144.0
新型肺炎实时动态	Real-time dynamics of novel coronavirus	1969	1157933	1679388	969591.9	503735.6
冠状病毒症状早期表现	Early symptoms of Novel coronavirus	3526	84902	150669	69961.2	44235.7
新型冠状病毒的临床表现	Clinical manifestations of novel coronavirus	4344	11933	23035	12612.2	5254.0
新型冠状感染最新情况	The latest situation of Novel coronavirus	257	1077	14926	3726.0	4819.2
新型冠状感染先兆	New signs of Novel coronavirus infection	11716	30736	36972	27766.2	7778.7
新型冠状性肺炎的发烧度数	Fever degree of novel coronavirus pneumonia	2275	4936	7668	4835.5	1792.3

Table 2
Correlation analysis and Cross-correlation of Baidu Search index and COVID-19.

Search Keywords	Correlation coefficient	*P* value	Maximum CCF	Lag (day)
Novel coronavirus characteristics	0.63	<0.001	0.706	lag1
Novel coronavirus symptoms	0.42	<0.05	0.522	lag0
What day is the most serious novel coronavirus symptom	0.65	<0.001	0.548	lag3
Real-time dynamics of novel coronavirus	0.89	<0.001	0.853	lag0
Early symptoms of Novel coronavirus	0.49	<0.001	0.644	lag3

CCF: Cross-Correlation Function coefficient

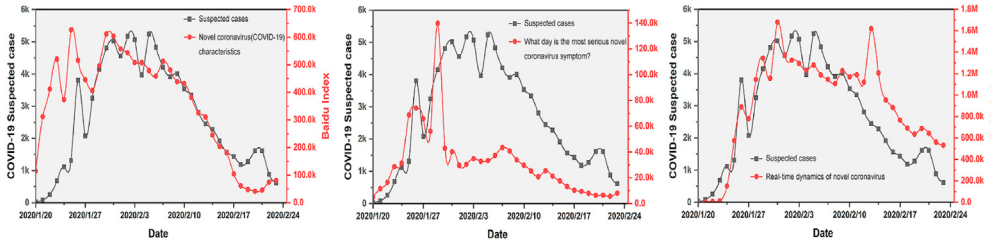

Fig. 1. Time series of Some Keywords Search Index and daily suspected cases for COVID-19 in China, from January 20, 2020 to February 20, 2020. This figure describes the time-series comparison curve between the Baidu search index and the daily suspected cases for the three keywords. The X-axis date interval is a week; The Y-axis use two coordinates, which the black Y-axis shows the number of daily suspected cases, the red Y axis is the Baidu search index of the keywords; BDI: Baidu Search index.

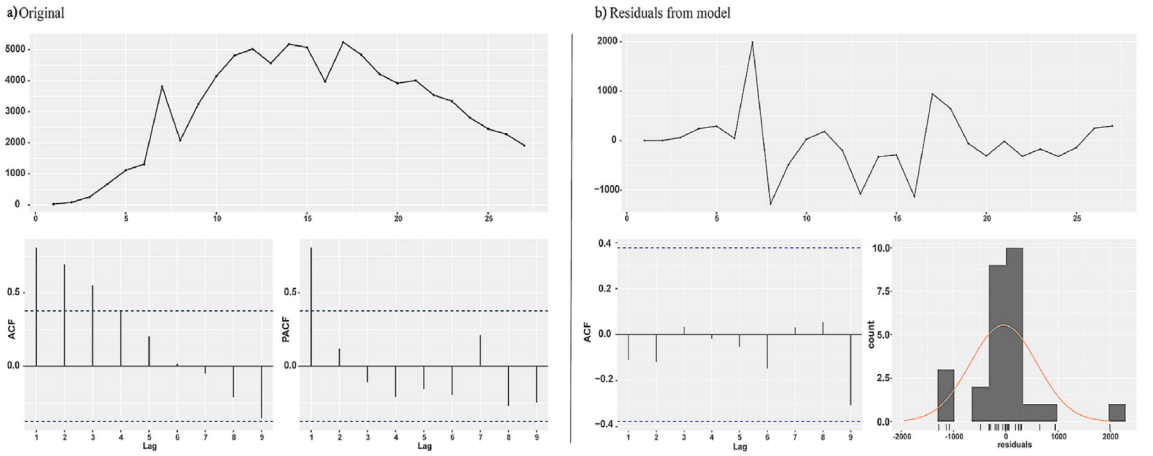

Fig. 2. Autocorrelation function (ACF) and partial ACF (PACF) plots of original COVID-19 suspected cases and Autocorrelation check of residuals for the model.

obtained by combining the ARIMA model with real time Baidu search engine query data, and the efficiency of the infectious disease surveillance system to greater assess the potential epidemic situation can be improved, which is vital for the prevention and control of COVID-19.

As we all known, the COVID-19 infectivity are strongly and has relatively long incubation (Backer et al., 2020; Guan et al., 2020; Liu et al., 2020), so it is a very urgent and important issue to discover and isolate the suspected cases. Only when quarantine measures are widely implemented can we better protect susceptible groups. This is essential for controlling the spread of the epidemic nationwide. Furthermore, this model was applied to the published data of Baidu search index to predict the trends of COVID-19 epidemic, and shown a precise forecasting of the trends with the keywords. Currently, COVID-19 has spread to nearly 200 countries and regions (OrganizationCoronavi, 2020). For countries or regions in the early stages of an outbreak or with low detection capacity, our model can estimate the epidemic through collecting and analysis the internet search data and provide direction on the control of the epidemic and reduce further transmission. In conclusion, this model, if it can be a supplement to support traditional surveillance systems, can provide sensitive monitoring of disease and epidemic situation information before the diagnosis of the disease is reported and discover more suspected cases, which can also contribute to allocate medical staff and material resources efficiently.

Table 3
Parameter characteristics of ARIMAX models for the relationship between COVID-19 suspected cases and Baidu Index variables.

Model	Variable	Parameter	Coefficients	*p* value	Ljung-Box test		AIC
					Q value	*P* value	
Model1	ARIMA	MA1	−1.458	<0.05	1.602	0.659	404.26
		MA2	0.853	<0.05			
Model2	ARIMAX + BDI	MA1	−1.4407	<0.05	1.532	0.675	403.51
		MA2	0.8493	<0.05			

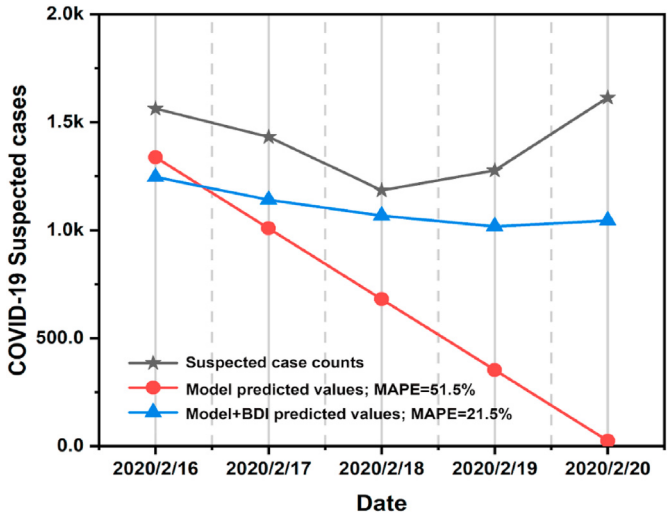

Fig. 3. Forecasting the counts of COVID-19 suspected cases from the two models in China. BDI: Baidu Index; MAPE: mean absolute percentage error.

However, there are several limitations of our study. As individual online search behavior constantly changing and there are some relevant keywords in the Baidu index are still not included, we haven't obtained enough data for the prediction model, which may lead to an undervaluation of the correlation (Kang et al., 2013; Eysenbach, 2006; Yoo et al., et al.). Another limitation base on its retrospective nature, we used the keywords in this study only represent the search behavior of individuals from January 20, 2020 to February 20, 2020, therefore, it cannot ensure consistent and effective prediction in the future. In time series analysis, long-term data support is required. Furthermore, according to the currently report of China Internet Network Information Center (CNNIC), Internet penetration rate reached 60% in China in 2019. This means that the Internet still unavailable access in some rural areas.

In summary, using data of the search engine query, may be a batter novel indicator for early monitoring and warning of the epidemic of COVID-19 outbreak, as well as help public health officials to assess and predict the progress of disease outbreaks in real time, which take effective early warning and intervention measures. Moreover, in this study the process of search keywords analysis detains can provide some reference value. Future reaches can prospectively collect and analyze data from Internet search index to early assess outbreaks of other diseases in China.

Author contributions

Kang Li and Yiming Shao were responsible for study design and results interpretation. Kang Li and Yanling Liang analyzed, interpreted the data and drafted the manuscript. Jianjun Li, Meiliang Liu and Yi Feng collected and supervised the data. All authors have read and reviewed the manuscript.

Funding

This work was supported by 13th Five-year National Major Project for HIV/AIDS and Hepatitis B Control and Prevention, Chinese Ministry of Science and Technology (2017ZX10202102005004 to P. M), Guangxi Bagui Honor Scholars, Ministry of Science and Technology of China(2017ZX10201101), and the National Natural Science Foundation International/Inter-Organization Cooperation and Exchange Study—NSFC—VR Project (China and Switzerland) (81861138011).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Department of Guangxi Universities Key Laboratory of Prevention and Control of Highly Prevalent Disease for assistance during the study and the Chinese Center for Disease Control and Prevention for providing the data materials.

References

Althouse, B. M., Ng, Y. Y., & Cummings, D. A. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis. Aug, 5*(8), e1258. https://doi.org/10.1371/journal.pntd.0001258

Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Euro Surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 25*(5). https://doi.org/10.2807/1560-7917.es.2020.25.5.2000062. Feb.

Bi, I., Alexander, W., Andrea, T.-B., Carmen, H., Kmu, G., & Kamran, K. (2020). Pneumonia of unknown aetiology in wuhan, China: Potential for international spread via commercial air travel. *Journal of Travel Medicine, 27*(2), taaa008.

Brown, A. J. L., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., & Dunn, D. T. (2011). Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Diseases, 204*(9), 1463.

Chadsuthi S, Modchang C, Lenbury Y, Iamsirithaworn S, Triampo W. Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. Asian Pacific Journal of Tropical Medicine. 000(007):539-546.

Cornelsen L, Normand C. Impact of the smoking ban on the volume of bar sales in Ireland − evidence from time series analysis. Health Economics. 21(5): 551-561.

Chinazzi, M., Davis, J. T., Ajelli, M., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*(6489), 395–400.

Du, Z., Xu, L., Zhang, W., Zhang, D., & Hao, Y. (2017). Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: An ecological study in Guangdong, China. *Bmj Open, 7*(10), Article e016263.

Ebhuoma, O., Gebreslasie, M., & Magubane, L. (Jun 26 2018). A Seasonal Autoregressive Integrated Moving Average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde. 108*(7), 573–578. https://doi.org/10.7196/SAMJ.2018.v108i7.12885

Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. *AMIA Annual Symposium Proceedings*, 244–248.

Ginsberg, J. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*.

Granger, C. W. J., & Swanson, N. R. (1996). Future developments in the study of cointegrated variables. *Oxford Bulletin of Economics & Statistics, 58*(3), 537–553.

Guan, W-j, Ni, Z-y, Hu, Y., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2002032

Gu Y, Chen F, Liu T, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. Scientific Reports. 5(1):12649.

Huang, C., Wang, Y., Li, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet, 395*(10223), 497–506. https://doi.org/10.1016/s0140-6736(20)30183-5. Feb 15.

Hulth, A., Rydevik, G., & Linde, A. (2009). Web queries as a source for syndromic surveillance. *PloS One, 4*(2), e4378.

Kang, M., Zhong, H., He, J., Rutherford, S., & Yang, F. (2013). Using Google trends for influenza surveillance in South China. *PloS One, 8*(1), Article e55205. https://doi.org/10.1371/journal.pone.0055205

Lee HS, Her M, Levine M, Moore GE. Time series analysis of human and bovine brucellosis in South Korea from 2005 to 2010. Preventive Veterinary Medicine. 110(2):190-197.

Li, K., Liu, M., Feng, Y., et al. (2019). Using Baidu search engine to monitor AIDS epidemics inform for targeted intervention of HIV/AIDS in China. *Science Reports, 9*(1), 320. https://doi.org/10.1038/s41598-018-35685-w. Jan 23.

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., & Akhmetzhanov, A. R. (Feb 17 2020). incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *9*(2). https://doi.org/10.3390/jcm9020538

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*.

Lipsitch, M., Swerdlow, D. L., & Finelli, L. (2020). Defining the epidemiology of Covid-19—studies needed. *New England Journal of Medicine, 382*(13), 1194–1196.

Liu, K., Wang, T., Yang, Z., et al. (2016a). Using Baidu search index to predict Dengue outbreak in China. *Scientific Reports, 6*, 38040.

Milinovich, G. J., Williams, G. M., Clements, A. C., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet Infectious Diseases, 14*(2), 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5. Feb.

Organization WH. Coronavirus disease 2019 (COVID-19) (Vol. 72) (2020). situation report.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, (11), 11.

Surveillances, V. (2020). The epidemiological characteristics of an outbreak of 2019 novel Coronavirus diseases (COVID-19)—China, 2020. *China CDC Weekly, 2*(8), 113–122.

Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *Lancet, 395*(10223), 470–473. https://doi.org/10.1016/s0140-6736(20)30185-9. Feb 15.

Wertheim, J. O., Brown, A. J., Leigh, N., Lance, H., et al. (2014). The global transmission network of HIV-1. *Journal of Infectious Diseases, 209*(2), 304.

Yin, Y., & Wunderink, R. G. (2018). MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology, 23*(2), 130–137. https://doi.org/10.1111/resp.13196. Feb.

Yoo HS, Park O, Park HK, et al. Timeliness of national notifiable diseases surveillance system in Korea: A cross-sectional study. BMC Public Health. 9(1):93-0.

Yuan Q, Nsoesie EO, Lv B, Peng G, Brownstein JS. Monitoring influenza epidemics in China with search query from Baidu. PloS One. 8(5):e64323-.

Zhang, Y., Bi, P., & Hiller, J. (2008). Climate variations and salmonellosis transmission in adelaide, South Australia: A comparison between regression models. *International Journal of Biometeorology, 52*(3), 179–187. https://doi.org/10.1007/s00484-007-0109-4. Jan.

Zhao, C., Yang, Y., Wu, S., et al. (2020). Search trends and prediction of human brucellosis using Baidu index data from 2011 to 2018 in China. *Scientific Reports, 10*(1), 5896. https://doi.org/10.1038/s41598-020-62517-7. Apr 3.

Zhou XC, Shen HB. Notifiable infectious disease surveillance with data collected by search engine. Frontiers of Information Technology & Electronic Engineering. (4):241-248.