RESEARCH Open Access

Early warning and predicting of COVID-19 using zero-inflated negative binomial regression model and negative binomial regression model

Wanwan Zhou[1], Daizheng Huang[2], Qiuyu Liang[3], Tengda Huang[1], Xiaomin Wang[1], Hengyan Pei[1], Shiwen Chen[1], Lu Liu[1], Yuxia Wei[1], Litai Qin[1] and Yihong Xie[1*]

Abstract

Background It is difficult to detect the outbreak of emergency infectious disease based on the exiting surveillance system. Here we investigate the utility of the Baidu Search Index, an indicator of how large of a keyword is in Baidu's search volume, in the early warning and predicting the epidemic trend of COVID-19.

Methods The daily number of cases and the Baidu Search Index of 8 keywords (weighted by population) from December 1, 2019 to March 15, 2020 were collected and analyzed with times series and Spearman correlation with different time lag. To predict the daily number of COVID-19 cases using the Baidu Search Index, Zero-inflated negative binomial regression was used in phase 1 and negative binomial regression model was used in phase 2 and phase 3 based on the characteristic of independent variable.

Results The Baidu Search Index of all keywords in Wuhan was significantly higher than Hubei (excluded Wuhan) and China (excluded Hubei). Before the causative pathogen was identified, the search volume of "Influenza" and "Pneumonia" in Wuhan increased with the number of new onset cases, their correlation coefficient was 0.69 and 0.59, respectively. After the pathogen was public but before COVID-19 was classified as a notifiable disease, the search volume of "SARS", "Pneumonia", "Coronavirus" in all study areas increased with the number of new onset cases with the correlation coefficient was 0.69 ~ 0.89, while "Influenza" changed to negative correlated (r_s: -0.56 ~ -0.64). After COVID-19 was closely monitored, the Baidu Search Index of "COVID-19", "Pneumonia", "Coronavirus", "SARS" and "Mask" could predict the epidemic trend with 15 days, 5 days and 6 days lead time, respectively in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei). The predicted number of cases would increase 1.84 and 4.81 folds, respectively than the actual number of cases in Wuhan and Hubei (excluded Wuhan) from 21 January to 9 February.

Conclusion The Baidu Search Index could be used in the early warning and predicting the epidemic trend of COVID-19, but the search keywords changed in different period. Considering the time lag from onset to diagnosis, especially in the areas with medical resources shortage, internet search data can be a highly effective supplement of the existing surveillance system.

Keywords COVID-19, Baidu search index, Early warning, Predicting, Zero inflation negative binomial regression, Negative binomial regression

*Correspondence:
Yihong Xie
gxxieyihong@163.com
Full list of author information is available at the end of the article

Introduction

Coronavirus disease 2019 (COVID-19) was a novel acute infectious disease reported in Wuhan, China in early December 2019. It spread quickly to all parts of the country in the next one month as the large population movement with the Chinese New Year coming. As of March 15, 2020, a total of 80,860 confirmed cases and 3,213 deaths have been reported in mainland China [1, 2]. COVID-19, as an emerging infectious disease, it was hard to detect based on the traditional surveillance systems [3]. The time lag for detecting the outbreak in the early stage was far from optimal for policymakers making decisions [2], especially for this fast and widely spread infectious disease.

Internet search engine data can play important role in predicting the diseases outbreak. It has been widely suggested as a supplement method to improve infectious disease surveillance [4, 5] and had successfully used in outbreak detection [6–8]. In China, Baidu is a leading internet search engine that offers similar features and services as Google, more than 90% of internet users prioritize using Baidu as a search tool to retrieve information of interest [9]. Baidu Index is a data analysis platform based on Baidu's massive Internet users' behavior data, which can provide the Baidu Search Index to tell users how large a keyword is in Baidu's search volume over a period of time (including overall trends, PC trends, and mobile trends) by calculating the weighted sum of the search frequency of each keyword in Baidu web search [9]. After the epidemic of COVID-19, several studies have focused on the prediction of COVID-19 in China using Baidu search data [10–12]. However, previous studies mainly focus on the period after COVID-19 was classified as a notifiable disease and the surveillance system was built in China. The search terms were limited to "COVID-19 (新冠肺炎) or (新型冠状病毒肺炎)" and "coronavirus (冠状病毒)". Whether the internet search data can be used in early warning of COVID-19 before the pathogen was confirmed and the disease was named hasn't been evaluated. Moreover, as the epidemic situation in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei) were different, the search behavior and the search terms may different, especially before and after the causative pathogen was confirmed. The objectives of this study were to investigate the predictive utility of Baidu Search Index in the early warning of COVID-19 and predicting the epidemic trends in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei). The effectiveness of comprehensive control measures was also be evaluated.

Methods

Data source

The study period was from December 1, 2019 to March 15, 2020. The daily number of new onset cases in Wuhan, Hubei and China before COVID-19 was notifiable in China (before January 20, 2020) was obtained from the publications [2, 13, 14] by using GetData Graph Digitizer software to capture and simulate the epidemic curve of COVID-19. The daily number of confirmed cases after January 20 (including the clinical diagnosis cases between February 12 to 14, 2020 in Wuhan and Hubei as the Fifth Edition of COVID-19 Diagnosis and Treatment Scheme in China changed the reporting criteria) was obtained from China National Health Commission (http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml) and Hubei Health Commission (http://wjw.hubei.gov.cn/bmdt/ztzl/fkxxgzbdgrfyyq/xxfb/). The daily search query data of the Baidu Search Index for individual keywords was achieved from the Baidu Index Platform (https://index.baidu.com/v2/index.html#/). The keywords mainly focus on some diseases with similar symptoms or caused public concern, including 8 Chinese terms: "Pneumonia (肺炎)", "SARS (非典) or (非典型肺炎)", "Coronavirus (冠状病毒) or (新型冠状病毒)", "COVID-19 (新冠肺炎) or (新型冠状病毒肺炎) or (新冠)", "MERS (中东呼吸综合征)", "Influenza (流感) or (流行性感冒)", "Avian Influenza (禽流感) or (人感染高致病性禽流感)" and "Mask (口罩) (including N95)". As the COVID-19 epidemic waves in Wuhan, Hubei and China were different, for comparison, all data analysis were separated into Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei), respectively. The Baidu Search Index per million population was used to evaluate the people's search behaviors. The population data of Wuhan and Hubei in year 2019 were achieved from Hubei Statistical Yearbook (http://tjj.hubei.gov.cn/tjsj/sjkscx/tjnj/qstjnj/) and for China was achieved from China Statistical Yearbook (http://www.stats.gov.cn/tjsj/ndsj/).

Statistical analysis

As the search keywords might change with more knowledge about COVID-19, the data analysis was classified into 3 phases according to the key events that may affect people's search behaviors. Phase 1 was from December 1 (7 days before the first case onset) to December 31, 2019 (Wuhan Municipal Health Commission issued an internal alert and the National Health Commission sent the rapid response team to Wuhan) [2], which caused the public's great concerned. Phase 2 was from January 1 to 20, 2020. During this period, the Huanan seafood market in Wuhan was closed on January 1 [2] and the causative pathogen was detected on January 7 [15]. The first announcement of human-to-human transmission was made on January 20, 2020 and COVID-19 was classified as a notifiable disease in China on the same day. Phase 3 was from January 21 to March 15, 2020, COVID-19 was closely monitored and strong intervention measures were taken in this period, and it being well controlled in China after mid-March.

The daily number of cases (by symptom onset date in the first two phases and by laboratory confirmed date in phase 3) and Baidu Search Index (per million population) was plot to explore the search keywords that can be used in the early warning of COVID-19. Their correlation coefficient was also explored with the results were shown without time lag in the first two phases while in different time lag in phase3 by considering the time lag from onset to laboratory confirmed. Kruskal–Wallis Rank Sum Test was used to compare the Baidu Search Index volume (per million population) in different areas and phases. In phase 1, as it was common for the daily number of cases to be zero, zero-inflated negative binomial regression model was used to predict the epidemic trend of COVID-19 [16]. The dependent variable was the number of cases. And the daily number of cases in China (excluded Hubei) used the data in Hubei to instead as it was all zero. In phase 2 and phase 3, negative binomial regression model was used to predict the epidemic trend of COVID-19 due to the data were over-dispersion [16, 17]. As the search keywords were moderately or highly correlated (Supplement Fig. 1), 8 search keywords were put in the factor analysis to obtain the a fewer number of uncorrelated factor scores to be used as independent variables in the negative binomial regression model. The accuracy of factor analysis was evaluated with Kaiser–Meyer–Olkin (KMO) index more than 0.8 and the number of factors were judged by eigenvalues more than 1 [17]. The best model fit was based on smallest Akaike Information Criterion (AIC) value [18]. The predicted number and the actual number of COVID-19 were plotted for visualization and comparison. To evaluate the effectiveness of the comprehensive intervention measures, we compared the accumulative predicted number and actual number of COVID-19 from 21 January to 9 February, a longest incubation period (14 days) after the Chinese government launched Public Health Events Level-I Emergency Response on January 23–25, 2020 [2]. Data analyses were performed using R 4.0.2 (R Foundation for Statistical Computing: Vienna, Austria) with "psych", "pscl", "MASS" and "ggplot2" packages. The statistical significance level was set as 0.05. Data collection and analysis of this study were from public data and were thus considered exempt from institutional review board approval.

Results

The time series of the number of cases and the Baidu Search Index (per million population)

In phase 1, the first case was onset on December 8, 2019 in Wuhan and there were only few sporadic cases before 16 December, 2019. The number of new onset cases in Wuhan increased gradually with around 2–13 cases per day on the second half of December. The Baidu Search Index of "Influenza" in Wuhan was also increased since December 8, 2019 and it was obviously higher than other keywords (Fig. 1A). The median (interquartile) search volume of "Influenza" in phase 1 was 72.15 (42.81, 87.98) per million population in Wuhan. While it was 2.89 (2.28, 4.49) per million population in Hubei (excluded Wuhan) and 6.74 (3.61, 8.04) per million population in China (excluded Hubei), where with only 2 cases and 0 case reported, respectively. An anomalous peak of the search volume of "SARS" and "Pneumonia" occurred on December 31, 2019 when Wuhan Municipal Health Commission issue an alert of "Urgent Notice on treating pneumonia of unknown cause" and the rapid response team of China CDC was sent for investigation. The search peak was as high as 8485.02 and 1008.21 per million population, respectively in Wuhan, while were only 313.39 and 28.51 per million population, respectively in Hubei (excluded Wuhan), 242.44 and 20.11 per million population, respectively in China (excluded Hubei) (Table 1, Fig. 1).

In phase 2, the number of cases in Wuhan increased rapidly with an average number of new onset cases was 192 per day (range from 47 to 802 cases). The search volume of "SARS" in Wuhan kept in relative high level. The search volume of "pneumonia" and "coronavirus" in Wuhan increasing gradually after the casual pathogen of COVID-19 was public on 8 January, 2020. In Hubei (excluded Wuhan) and China (excluded Hubei), the number of cases increased after January 1, 2020, the average daily number of new onset cases was 87 in Hubei (excluded Wuhan) and 47 in China (excluded Hubei). The search volume of "pneumonia" and "coronavirus" were also increased compare to phase 1 but were much lower than in Wuhan (P-value < 0.001).

In phase 3, the number of confirmed cases in Wuhan and Hubei (excluded Wuhan) increased substantially with the epidemic peak occurred on 1–19 February in Wuhan and on 28 January-15 February in Hubei (excluded Wuhan). The number of confirmed cases in China (excluded Hubei) was also increased but with a smaller peak lasted from 28 January to 8 February, 2020. The Baidu Search Index of "COVID-19", "Pneumonia", "Coronavirus", "SARS" and "Mask" were also substantially

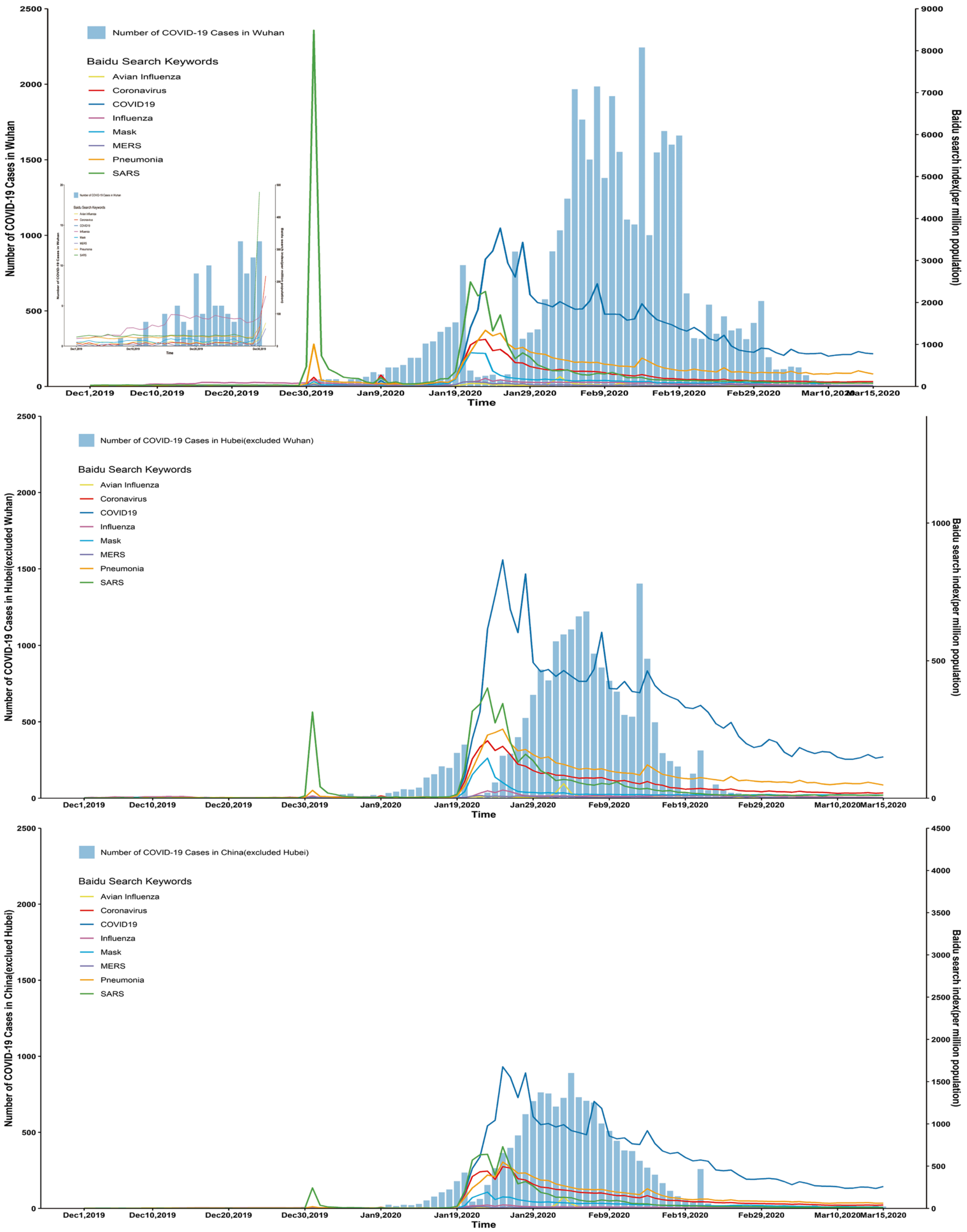

Fig. 1 The time series of the number of COVID-19 cases and the Baidu Search Index in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei); the COVID-19 cases by onset date before 20 January, 2020 and by laboratory confirmed date after 20 January, 2020; the Baidu Search Index (per million populations) from December 1,2019 to March 15, 2020

Table 1 The median (P_{25}, P_{75}) of Baidu Search Index volume (per million populations) of different keywords in phase 1&2 and phase 3

Number of cases/ Keywords	Phase 1 & 2					Phase 3				
	Wuhan	**Hubei (excluded Wuhan)**	**China (excluded Hubei)**	X^2	*P*值	**Wuhan**	**Hubei (excluded Wuhan)**	**China (excluded Hubei)**	X^2	*P*值
Number of cases	9 (1, 71)	0 (0, 23)	0 (0, 4)	17.6	< 0.001	370 (79, 1088)	89 (3, 610)	62 (10, 421)	16.7	< 0.001
Keywords										
Pneumonia	30.95 (27.65, 76.17)	1.75 (1.53, 3.10)	4.12 (3.78, 5.40)	43.5	< 0.001	440.10 (334.80, 632.50)	78.90 (59.16, 111.49)	120.13 (81.36, 237.76)	104.1	< 0.001
SARS	34.07 (31.30, 147.61)	1.33 (0.85, 4.07)	2.40 (1.99, 4.41)	36.0	< 0.001	149.70 (104.50, 353.60)	21.60 (11.21, 61.66)	33.09 (15.39, 112.83)	56.2	< 0.001
Coronavirus	10.88 (5.89, 37.77)	0.29 (0.06, 1.25)	0.24 (0.21, 0.88)	34.4	< 0.001	187.60 (126.50, 375.00)	37.00 (23.61, 76.16)	92.88 (51.92, 199.19)	72.9	< 0.001
COVID-19	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	-	-	1405.90 (831.50, 1911.70)	330.0 (184.2, 443.7)	603.50 (330.20, 919.90)	94.8	< 0.001
MERS	5.62 (5.08, 15.07)	0.08 (0.00, 0.45)	0.15 (0.13, 0.26)	40.0	< 0.001	27.65 (19.93, 35.54)	3.08 (1.80, 4.45)	3.18 (1.95, 5.25)	107.7	< 0.001
Influenza	56.99 (34.65, 85.08)	2.91 (2.43, 4.45)	6.14 (4.20, 7.89)	46.4	< 0.001	65.47 (44.69, 97.08)	7.99 (6.86, 9.23)	10.44 (5.87, 17.77)	106.5	< 0.001
Avian Influenza	11.68 (8.43, 15.12)	0.37 (0.29, 0.56)	0.63 (0.59, 0.65)	43.6	< 0.001	24.53 (16.99, 36.92)	2.58 (1.47, 4.97)	3.50 (1.28, 7.63)	91.8	< 0.001
Mask	19.98 (15.35, 28.36)	0.54 (0.45, 1.37)	0.81 (0.76, 0.87)	37.9	< 0.001	93.11 (82.77, 143.55)	12.63 (11.71, 14.61)	35.71 (28.95, 61.24)	110.0	< 0.001

increased in all the study areas, with the search peaks occurred on 21–29, January in Wuhan and on 23–28, January both in Hubei (excluded Wuhan) and in China (excluded Hubei). The search peaks occurred 5–21 days ahead of the epidemic peak of cases. The Baidu Search Index was in declining trend with the decreased of cases after 1 March, 2020.

The correlation between the number of cases and the Baidu Search Index (per million population)

In phase 1, there were significant associations between the search volume of "Influenza", "Pneumonia" and the daily number of new onset cases in Wuhan, with the correlation coefficients were 0.69 and 0.59, respectively, while no association was found in Hubei (excluded Wuhan) ($P > 0.05$). In phase 2, the association between the search volume of "Coronavirus", "COVID-19" and the daily number of new onset cases in all study areas were in moderate to high correlated (r_s: 0.69 ~ 0.89, $P < 0.05$), while "Influenza" changed to negative correlated (r_s: -0.56 ~ -0.64, $P < 0.05$). In phase 3, a high correlation ($r_s > 0.7$, $P < 0.05$) between the search volume of "Influenza", "Pneumonia", "Coronavirus", "SARS", "COVID-19" and the number of reported cases can be observed in 4–16 days time lag in Wuhan, 0–10 days lag time in Hubei (excluded Wuhan, except "Influenza") and 0–7 days time lag in China (excluded Hubei) (Table 2).

The prediction of COVID-19 using zero-inflated negative binomial regression model and negative binomial regression model

The factor analysis results showed that only one factor was identified in each study area in phase 2 and phase 3 (Supplement Table 1). The zero-inflated negative binomial regression model with 8 keywords as independent variable in phase 1, and the negative binomial regression model with the factor score as independent variable in phase 2 and phase 3 were conducted. As showed in Fig. 2, the cumulative number of predicted cases and actual cases in Wuhan (phrase 1: 103 vs. 115; phrase 2: 3932 vs. 3883), Hubei (excluded Wuhan) (phrase 1: 2 vs. 2; phrase 2: 1,762 vs. 1,742), and China (excluded Hubei) (phrase 1: 111 vs. 117; phrase 2: 909 vs. 905) in phase 1 and phrase 2 were highly consistent. In phase 3, the best model was showed in 15 days time lag in Wuhan, 5 days time lag in Hubei (excluded Wuhan) and 6 days time lag in China (excluded Hubei) (Table 3). The cumulative number of predicted cases and actual cases in phase 3 were 72,235 and 50,601, respectively in Wuhan, and 78,246 and 18,252, respectively in Hubei (excluded Wuhan), 13,055 and 13,062, respectively in China (excluded Hubei). The predicted peak occurred 5 days ahead the actual peak in Wuhan, and an abnormally high peak was observed in Hubei (excluded Wuhan) (Fig. 2, Supplement Fig. 2). For better comparison, we conducted the negative binomial regression model again only using the data from 21 January to 9 February, 2020. The predicted number of cases would be 2.84 (47,204 vs. 16,645) and 5.81 (74,372 vs. 12,803) folds of the actual number, respectively in Wuhan and Hubei (excluded Wuhan), while the predicted number of cases was slightly decrease in China (excluded Hubei) (8,524 vs. 10,538).

Discussion

This study demonstrated that the Baidu Search Index could be used in the early warning and predicting the COVID-19 epidemic with different keywords in different period. The Baidu Search Index of "Influenza" and "Pneumonia" could be used in the early warning of COVID-19 in Wuhan on December 2019. After the National Health Commission sent the rapid response team to Wuhan but before COVID-19 was notifiable in China, the continuously high search volume of "SARS", "Pneumonia" and "Coronavirus" in Wuhan and the increased concern of these search keywords in Hubei (excluded Wuhan) and China (excluded Hubei) indicated that these three keywords could be used in predicting the severity of COVID-19 in Wuhan and the further spread in China. After COVID-19 was closely monitor in China in phase 3, the Baidu Search Index of "COVID-19", "SARS", "Pneumonia", "Coronavirus" and "Mask" could be used in predicting the epidemic trends with 15 days, 5 days and 6 days lead time, respectively in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei).

With the popularity of internet, more and more people tend to search for health-related knowledge and information when getting sick. In this study, we found that the search volume of "Influenza", "Pneumonia" in phase 1 in Wuhan and the search volume of "SARS", "Pneumonia", "Coronavirus" in phase 2 in all study areas increased with the number of new onset cases. There were high correlated and consistent distribution between the search volume and the number of cases by symptom onset date. This implicated that public paid great concerned to this emergency infectious disease and searched for help from internet immediately once they develop the disease symptoms. This could be further approved as the search volume in Wuhan, the most severe epidemic area, was significantly higher than Hubei (excluded Wuhan) and China (excluded Hubei). As there is always time lag between onset to diagnosis, the Baidu search data could play important role in the early warning of COVID-19.

Previous studies about the early warning of COVID-19 only focused on the whole China and the period after COVID-19 was notifiable and with inconsistent findings. A study reported that the internet search data had

Table 2 The Spearman correlation coefficient of the Baidu Search Index of different keywords and the daily number of cases

	Influenza	**Pneumonia**	**SARS**	**Coronavirus**	**COVID-19**	**MERS**	**Avian Influenza**	**Mask**
Phase 1 (no lag)[a]								
Wuhan	0.68*	0.59*	0.01	0.45*	-	0.19	0.39△	0.41△
Hubei (excluded Wuhan)	0.31	0.31	0.31	0.31	-	0.20	0.31	0.29
Phase 2 (no lag)								
Wuhan	-0.56△	-0.06	-0.22	0.69*	0.85*	0.37	-0.28	0.20
Hubei (excluded Wuhan)	-0.61*	0.06	-0.06	0.69*	0.76*	0.47△	-0.33	0.45
China (excluded Hubei)	-0.64*	-0.25	-0.17	0.82*	0.89*	0.60*	-0.07	0.27
Phase 3								
Wuhan								
Lag 0 day	0.44*	0.39*	0.43*	0.43*	0.50*	0.28△	0.58*	0.34△
Lag 1 day	0.45*	0.40*	0.43*	0.43*	0.55*	0.31△	0.59*	0.35△
Lag 2 days	0.51*	0.48*	0.45*	0.49*	0.62*	0.29△	0.64*	0.36△
Lag 3 days	0.58*	0.56*	0.49*	0.54*	0.69*	0.36△	0.67*	0.43*
Lag 4 days	0.64*	0.64*	0.56*	0.61*	0.75*	0.44*	0.69*	0.49*
Lag 5 days	0.69*	0.69*	0.62*	0.68*	0.79*	0.48*	0.70*	0.54*
Lag 6 days	0.75*	0.72*	0.69*	0.73*	0.80*	0.53*	0.77*	0.60*
Lag 7 days	0.78*	0.75*	0.74*	0.79*	0.80*	0.56*	0.77*	0.66*
Lag 8 days	0.80*	0.76*	0.78*	0.80*	0.81*	0.58*	0.77*	0.70*
Lag 9 days	0.79*	0.78*	0.82*	0.80*	0.81*	0.57*	0.73*	0.67*
Lag 10 days	0.77*	0.79*	0.86*	0.82*	0.81*	0.60*	0.74*	0.71*
Lag 11 days	0.76*	0.79*	0.85*	0.82*	0.78*	0.57*	0.74*	0.71*
Lag 12 days	0.73*	0.77*	0.85*	0.80*	0.74*	0.54*	0.70*	0.72*
Lag 13 days	0.68*	0.74*	0.83*	0.76*	0.67*	0.53*	0.65*	0.71*
Lag 14 days	0.62*	0.70*	0.80*	0.73*	0.60*	0.51*	0.59*	0.70*
Lag 15 days	0.57*	0.65*	0.78*	0.69*	0.56*	0.48*	0.52*	0.67*
Lag 16 days	0.48*	0.57*	0.71*	0.63*	0.51*	0.45*	0.41*	0.63*
Lag 17 days	0.40*	0.47*	0.63*	0.57*	0.44*	0.41*	0.31△	0.56*
Lag 18 days	0.35△	0.41*	0.59*	0.56*	0.39*	0.45*	0.23	0.49*
Lag 19 days	0.27	0.33△	0.47*	0.45*	0.33△	0.36△	0.13	0.42*
Lag 20 days	0.22	0.24	0.37△	0.40*	0.28△	0.36△	0.07	0.34△
Hubei (excluded Wuhan)								
Lag 0 day	0.33△	0.76*	0.67*	0.73*	0.84*	0.56*	0.82*	0.50*
Lag 1 day	0.36△	0.81*	0.71*	0.77*	0.87*	0.65*	0.85*	0.52*
Lag 2 days	0.36△	0.84*	0.78*	0.84*	0.88*	0.68*	0.85*	0.63*
Lag 3 days	0.35△	0.85*	0.84*	0.87*	0.87*	0.69*	0.83*	0.68*
Lag 4 days	0.36△	0.84*	0.86*	0.87*	0.85*	0.70*	0.79*	0.71*
Lag 5 days	0.34△	0.83*	0.87*	0.87*	0.83*	0.67*	0.73*	0.73*
Lag 6 days	0.31△	0.81*	0.85*	0.85*	0.81*	0.64*	0.69*	0.70*
Lag 7 days	0.25	0.78*	0.83*	0.82*	0.75*	0.60*	0.66*	0.72*
Lag 8 days	0.19	0.73*	0.82*	0.78*	0.68*	0.58*	0.61*	0.70*
Lag 9 days	0.13	0.68*	0.77*	0.74*	0.62*	0.54*	0.56*	0.68*
Lag 10 days	0.06	0.60*	0.72*	0.68*	0.55*	0.48*	0.48*	0.63*
Lag 11 days	0	0.54*	0.66*	0.61*	0.45*	0.41*	0.43*	0.59*
Lag 12 days	-0.08	0.45*	0.58*	0.53*	0.35△	0.32△	0.31△	0.52*
China (excluded Hubei)								
Lag 0 day	0.79*	0.84*	0.80*	0.82*	0.87*	0.58*	0.82*	0.77*
Lag 1 day	0.80*	0.87*	0.83*	0.85*	0.87*	0.61*	0.83*	0.80*
Lag 2 days	0.83*	0.87*	0.85*	0.85*	0.86*	0.63*	0.81*	0.80*
Lag 3 days	0.84*	0.86*	0.85*	0.84*	0.81*	0.62*	0.76*	0.79*

Table 2 (continued)

	Influenza	Pneumonia	SARS	Coronavirus	COVID-19	MERS	Avian Influenza	Mask
Lag 4 days	0.83*	0.83*	0.84*	0.83*	0.77*	0.62*	0.71*	0.77*
Lag 5 days	0.81*	0.79*	0.82*	0.79*	0.72*	0.63*	0.66*	0.75*
Lag 6 days	0.78*	0.75*	0.79*	0.75*	0.66*	0.63*	0.60*	0.71*
Lag 7 days	0.72*	0.69*	0.73*	0.69*	0.57*	0.59*	0.53*	0.66*
Lag 8 days	0.65*	0. 61*	0.68*	0.63*	0.47*	0.55*	0.47*	0.60*
Lag 9 days	0.58*	0.53*	0.61*	0.56*	0.39*	0.48*	0.38*	0.52*
Lag 10 days	0.50*	0.44*	0.53*	0.48*	0.30$^{\triangle}$	0.41*	0.29$^{\triangle}$	0.42*
Lag 11 days	0.40*	0.34**	0.43*	0.38*	0.21	0.29$^{\triangle}$	0.18	0.31$^{\triangle}$
Lag 12 days	0.28$^{\triangle}$	0.24	0.33*	0.28$^{\triangle}$	0.10	0.19	0.08	0.22

Phase 1 by symptom onset date without time lag, phase 2 by symptom onset date without time lag, phase 3 by diagnosis date with different time lag; $^{\triangle}$*P*-value < 0.05; **P*-value < 0.001; [a] No case in China (excluded Hubei) as the number of COVID-19 on December 2019 was 0

10–14 days lead time in the prediction of COVID-19 [11], while other studies showed 18 days [19] and 0–4 days lead time [10]. However, the epidemic situation in different areas were different and the public concerned were closed related to severity of epidemic, population density, economic level, etc. [10]. In this study, we analyzed the data in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei) separately according to the epidemic situation and weight the search behavior by population. We found that the Baidu Search Index of "COVID-19", "SARS", "Pneumonia", "Coronavirus" and "Mask" was increase ahead of the number of reported cases, which showed 15 days, 5 days and 6 days lead time, respectively in Wuhan, Hubei (excluded Wuhan) and China (excluded Hubei) in phase 3. The long lead time in Wuhan may due to the severe shortage of medical resources. Lots of cases could not receive timely diagnosis and treatment and have to self-quarantine at home [13]. The lead time in Wuhan was consistent with the average interval of 12 days from symptom onset to laboratory confirmation [20]. But in Hubei (excluded Wuhan) and in China (excluded Hubei), the medical resource was relative sufficient, the lead time in these areas were shorter and consistent with the 3–7 days interval from onset to diagnosis [21].

To better control the wide spread of COVID-19, the Chinese government launched Public Health Events Level-I Emergency Response on January 23–25, 2020 [2]. A series of comprehensive intervention measures including lockdown of Wuhan [21, 22], travel restrictions [21], cases isolation and contact tracing [21, 23], etc. were implements in China. The epidemic being well control at the end of February. A study in Jilin Province used the Susceptible-Exposed-Infectious-Asymptomatic-Recovered/ Removed model to evaluate the effectiveness of local interventions implemented on February 1, 2020 and found the incidence of cases reduce 99.99% [24]. Shengjie Lai et al. [25] built a travel network-based stochastic susceptible-exposed-infectious-removed model to simulate the COVID-19 spread in mainland China. And found that if without non-pharmaceutical interventions, as of February 29, the number of COVID-19 cases would increase 51 folds in Wuhan, 92 folds in other cities of Hubei, and 125 folds in other provinces. Our study used Baidu Search Index to predicted the epidemic trend from 21 January to 9 February and showed that the number of cases would be 2.84 and 5.81 folds of the actual number, respectively in Wuhan and Hubei (excluded Wuhan), while it was slightly decrease in China (excluded Hubei). This may due to we only focused on the first 14 days after COVID-19 was notifiable, and the number of cases used in prediction in Wuhan might underestimated as a lot of cases being delayed diagnosis before February 18 with the serious shortage of medical resources [10]. The different between Wuhan, Hubei (excluded Wuhan), and China (excluded Hubei) may due to COVID-19 was notifiable more than 40 days after the first case onset in Wuhan, the disease already widely spread among the city. But it still timely prevented the large population movement out from Wuhan before Chinese New Year [25] and prevented the further spread to Hubei (excluded Wuhan) and China (excluded Hubei).

The limitations of this study were that we simulated the epidemic curve of COVID-19 before January 20, 2020 by using software to capture the daily onset data from the publications. The exact daily number of cases by onset date used in the analysis may deviate from the real situation. However, it would not change the epidemic trend for visualization in time series analysis. Another limitation is that the search volume of some keywords had anomalous peak such as "SARS" and "Pneumonia", which might affected the stability of the model. Moreover, we only focus on the largest search engine Baidu in China, other search engine such as

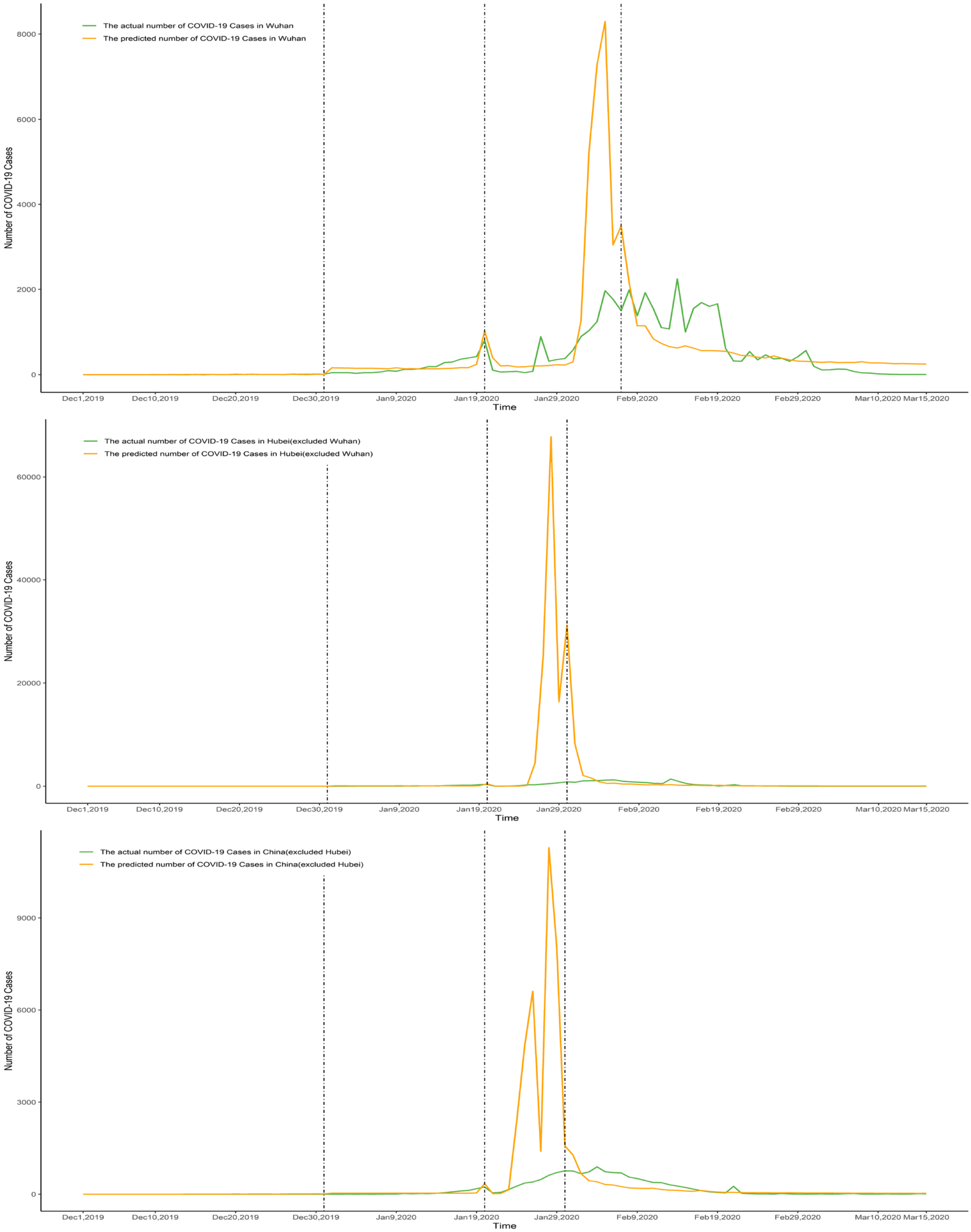

Fig. 2 The predicted and actual number of COVID-19 cases from December 1,2019 to March 15, 2020. Note: A) Wuhan, B) Hubei (excluded Wuhan), C) China (excluded Hubei)

Table 3 The AIC of the predicting model in phase 1, phase 2 and phase 3

	Wuhan	Hubei (excluded Wuhan)	China (excluded Hubei)
Phase 1 (no lag)[b]	137.79	24.61	120.84
Phase 2(no lag)[c]	151.14	73.70	10.90
Phase 3[b]			
Lag 0 day	582.02	307.28	50.58
Lag 1 day	582.54	304.79	49.65
Lag 2 days	583.25	299.90	48.43
Lag 3 days	583.17	294.81	47.28
Lag 4 days	581.83	289.61	46.51
Lag 5 days	579.32	285.64	46.04
Lag 6 days	576.80	286.01	45.89
Lag 7 days	573.91	287.58	46.04
Lag 8 days	572.79	291.15	46.43
Lag 9 days	569.15	294.56	47.45
Lag 10 days	564.49	298.43	48.85
Lag 11 days	564.32	302.21	49.92
Lag 12 days	563.60	305.32	51.00
Lag 13 days	562.27	/	/
Lag 14 days	561.60	/	/
Lag 15 days	559.65	/	/
Lag 16 days	563.15	/	/
Lag 17 days	562.41	/	/
Lag 18 days	561.00	/	/
Lag 19 days	565.43	/	/
Lag 20 days	564.47	/	/

Phase 1 by symptom onset date without times lag, phase 2 by symptom onset date without times lag and phase 3 by diagnosis date with different times lag; [b] The AIC of zero inflation negative binomial regression model; [c] The AIC of negative binomial regression model

Weibo, 360, Yahoo and Sogou should also be considered. And the Baidu Search Index is not reflective of the magnitude of cases. Lastly, we only considered the Baidu Search Index in the regression model, the population movement and impact of policy should also be considered.

Conclusion

The Baidu Search Index could be used in the early warning and predicting the epidemic trend of COVID-19, but the search keywords may change over time with more knowledge about COVID-19. The keywords selection should consider the public's concern in different period. Considering the time lag from onset to diagnosis, especially in the areas with medical resources shortage, internet search data can be a highly effective supplement of the existing surveillance system.

Abbreviations

COVID-19	Coronavirus disease 2019
KMO	Kaiser–Meyer–Olkin
AIC	Akaike Information Criterion

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-09940-7.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

Y.X. and W.Z. had full access to all of the data in the study, took responsibility for the integrity of the data and the accuracy of the data analysis, made a critical revision of the manuscript for important intellectual content, and supervised the manuscript; Y.X., D.H., Q.L., X.W. and W.Z.conceptualized and designed the study; T.H., H.P., S.C., L.L., Y.W.and L.Q.acquired analyzed or interpreted data, drafted the manuscript, conducted the statistical analysis, and provided administrative, technical, or material support. All authors read and approved the final manuscript.

Funding

This research was funded by Health and Emergency Skills Training Center of Guangxi, grant number HESTCG202104.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Epidemiology and Biostatistics, Guangxi Medical University, 22 Shuangyong Road, Qingxiu District, Nanning, Guangxi 530021, China. [2]Institute of Life Science, Guangxi Medical University, Nanning, China. [3]Department of Health Management, The People's Hospital of Guangxi Zhuang Autonomous Region & Research Center of Health Management, Guangxi Academy of Medical Sciences, Nanning, China.

Received: 24 April 2024 Accepted: 16 September 2024
Published online: 19 September 2024

References

1. The latest situation of COVID-19 as of 24:00, March 15,2020. China National Health Commission. 2020. http://www.nhc.gov.cn/xcs/yqtb/202003/114113d25-c1d47aabe68381e836f06a8.shtml. Accessed 16 Mar 2020.
2. The Novel Coronavirus Pneumonia Emergency Response Epidemiology T. The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) - China, 2020. China CDC Wkly. 2020;2(8):113–22.
3. Semenza JC, Rocklöv J, Penttinen P, Lindgren E. Observed and projected drivers of emerging infectious diseases in Europe. Ann N Y Acad Sci. 2016;1382(1):73–83. https://doi.org/10.1111/nyas.13132.

4. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. JMIR. 2020;22(5):e19421.
5. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q. 2014;92(1):7–33. https://doi.org/10.1111/1468-0009.12038.
6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012–4. https://doi.org/10.1038/nature07634.
7. Huang DC, Wang JF, Huang JX, Sui DZ, Zhang HY, Hu MG, et al. Towards Identifying and Reducing the Bias of Disease Information Extracted from Search Engine Data. PLoS Comput Biol. 2016;12(6):e1004876. https://doi.org/10.1371/journal.pcbi.1004876.
8. Wang J, Zou Y, Peng Y, Li K, Jiang T. On prediction of dengue epidemics based on Baidu index. Comp Appl Softw. 2016;33(07):e46.
9. Liu K, Li L, Jiang T, Chen B, Jiang Z, Wang Z, et al. Chinese Public Attention to the Outbreak of Ebola in West Africa: Evidence from the Online Big Data Platform. Int J Environ Res Public Health. 2016;13(8):780. https://doi.org/10.3390/ijerph13080780.
10. Gong X, Han Y, Hou M, Guo R. Online Public Attention During the Early Days of the COVID-19 Pandemic: Infoveillance Study Based on Baidu Index. JMIR Public Health Surveill. 2020;6(4): e23098. https://doi.org/10.2196/23098.
11. 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.
12. Tu B, Wei L, Jia Y, Qian J. Using Baidu search values to monitor and predict the confirmed cases of COVID-19 in China: - evidence from Baidu index. BMC Infect Dis. 2021;21(1):98. https://doi.org/10.1186/s12879-020-05740-x.
13. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, et al. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan. China JAMA. 2020;323(19):1915–23. https://doi.org/10.1001/jama.2020.6130.
14. Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan. China J Med Virol. 2020;92(4):441–7. https://doi.org/10.1002/jmv.25689.
15. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33. https://doi.org/10.1056/NEJMoa2001017.
16. Schober P, Vetter TR. Count Data in Medical Research: Poisson Regression and Negative Binomial Regression. Anesth Analg. 2021;132(5):1378–9. https://doi.org/10.1213/ane.0000000000005398.
17. Mahmoudi MR, Baleanu D, Band SS, Mosavi A. Factor analysis approach to classify COVID-19 datasets in several regions. Results Phys. 2021;25:104071. https://doi.org/10.1016/j.rinp.2021.104071.
18. Saleh F, Kitau J, Konradsen F, Kampango A, Abassi R, Schiøler KL. Epidemic risk of arboviral diseases: Determining the habitats, spatial-temporal distribution, and abundance of immature Aedes aegypti in the Urban and Rural areas of Zanzibar, Tanzania. PLoS Negl Trop Dis. 2020;14(12):e0008949. https://doi.org/10.1371/journal.pntd.0008949.
19. Li Z, Hu D. Forecast of the COVID-19 Epidemic Based on RF-BOA-LightGBM. Healthcare (Basel). 2021;9(9):1172. https://doi.org/10.3390/healthcare9091172.
20. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). World Health Organization. 2020. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf. Accessed 28 Feb 2020.
21. Press Conference of WHO-China Joint Mission on COVID-19. World Health Organization. 2020. https://www.who.int/docs/default-source/coronaviruse/transc-ripts/joint-mission-press-conference-script-englishfinal.pdf?sfvrsn=51c90b9e_2. Accessed 28 Feb 2020.
22. Kraemer MUG, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science (New York, NY). 2020;368(6490):493–7. https://doi.org/10.1126/science.abb4218.
23. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Health. 2020;8(4):e488–96. https://doi.org/10.1016/s2214-109x(20)30074-7.
24. Zhao Q, Wang Y, Yang M, Li M, Zhao Z, Lu X, et al. Evaluating the effectiveness of measures to control the novel coronavirus disease 2019 in Jilin Province, China. BMC Infect Dis. 2021;21(1):245. https://doi.org/10.1186/s12879-021-05936-9.
25. Lai S, Ruktanonchai NW, Zhou L, Prosper O, Luo W, Floyd JR, et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature. 2020;585(7825):410–3. https://doi.org/10.1038/s41586-020-2293-x.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.