Methods and Applications

Exploring the Lagged Correlation Between Baidu Index and Influenza-Like Illness — China, 2014–2019

Xuan Han[1]; Jiao Yang[1]; Yan Luo[1]; Dazhu Huo[2]; Xuya Yu[1]; Xuancheng Hu[3]; Ling Xin[1]; Liuyang Yang[1]; Hualei Xin[1]; Ting Zhang[1,#]; Zhongjie Li[1,#]; Weizhong Yang[1,#]

ABSTRACT

Introduction: This study investigated the lagged correlation between Baidu Index for influenza-related keywords and influenza-like illness percentage (ILI%) across regions in China. The aim is to establish a scientific foundation for utilizing Baidu Index as an early warning tool for influenza-like illness epidemics.

Methods: In this study, data on ILI% and Baidu Index were collected from 30 provincial-level administrative divisions (PLADs) spanning April 2014 to March 2019. The Baidu Index was categorized into Overall Index, Ordinary Index, Prevention Index, Symptom Index, and Treatment Index based on search query themes. The lagged correlation between the Baidu Index and ILI% was examined through the cross-correlation function (CCF) method.

Results: Correlating the Baidu Overall Index of 30 PLADs with ILI% revealed CCF values ranging from 0.46 to 0.86, with a median lag of 0.5 days. Subcategory analysis indicated that the Prevention Index and Symptom Index exhibited quicker responses to ILI%, with median lags of −9 and −0.5 days, respectively, compared to 0 and 3 days for the Ordinary and Treatment Indexes. The median lag days between the Baidu Index and the ILI% were earlier in the northern PLADs compared to the southern PLADs.

Discussion: The Prevention and Symptom Indexes show promising predictive capabilities for influenza-like illness epidemics.

Influenza, an acute respiratory infection caused by influenza viruses (1), affected approximately one billion people worldwide annually between 1999 and 2015 (2). Timely identification of the influenza season' s onset is crucial for preparing national and local healthcare resources (3), fostering vaccination uptake, strengthening public health measures (4), curtailing disease spread, and reducing the impact of seasonal influenza.

Surveillance plays a crucial role in the prevention and management influenza (5). Currently, the primary global surveillance methods for influenza include monitoring influenza-like illness and influenza virus activity. These methods aim to capture fluctuations in patient visits and the intensity of influenza viral circulation, providing insights into the onset, peak, and conclusion of seasonal influenza outbreaks (6). Traditional surveillance approaches, which involve weekly data reporting and case-based analysis, are prone to delays in detecting early signs of influenza epidemics (7). Previous research has demonstrated the utility of internet search data in identifying infectious disease outbreaks (8) and highlighted the potential of monitoring epidemic trends using information from social media and online activities of Internet users (9). However, accurate monitoring hinges on selecting and utilizing disease-specific web search keywords effectively.

This study examines the lagged correlation between influenza-like illness percentage (the ratio of the total number of influenza-like cases to the total number of outpatient emergency department visits; ILI%) and the Baidu Index of influenza-related keywords in 30 provincial-level administrative divisions (PLADs) in China. It further categorizes the Baidu Index to identify a specific subset that shows a strong correlation and timely response. These findings aim to establish a scientific foundation for enabling early detection of influenza-like disease outbreaks.

METHODS

This study utilized weekly influenza surveillance data, including the number of ILI cases and total outpatient emergency department visits, obtained from the National Influenza Center of China (10). The data were sourced from outpatient emergency departments in sentinel hospitals located across 30 PLADs (excluding Xizang Autonomous Region) spanning from April 2014 to March 2019. To calculate daily ILI%, a cubic spline function was employed to interpolate weekly ILI%. The cubic spline function is a widely accepted method for curve fitting and interpolation, commonly used to convert weekly influenza data into daily estimates based on weekly reports (*11*).

The Baidu Index is a calculated total of search frequencies for specific keywords on Baidu web pages, sourced from the public Baidu Index website (*12*). The selection and refinement of keywords for the Baidu Index were based on methodologies outlined in prior research, encompassing "influenza bidding terms, Baidu Index demand mapping, expert consultation, and literature summarization." Keywords unrelated to influenza, not included in the Baidu Index database, not searched for in the past year, or pertaining to specific strains were excluded (*13*). A total of 39 influenza-related keywords were compiled and categorized under the "Overall Index." Based on search patterns, these keywords were further segmented into four groups: basic terminology, prevention, symptoms, and treatment, designated as the Ordinary Index, Prevention Index, Symptom Index, and Treatment Index, respectively (Table 1).

The lagged correlation between the Baidu Index and ILI% was examined using the Cross-Correlation Function (CCF) method. This technique was utilized to assess the cross-correlation between the two time series, specifically to investigate if a particular pattern in one series tends to follow a pattern in the other series. The method generates a judgment indicator in the form of CCF values to determine the correlation between the two time series, with the formulas detailed as per reference (Table 2) (*14*).

Initially, using sample estimates of cross-covariance in Equation 1 to measure co-variation at different time points, offering preliminary insights.

$$\hat{\gamma}_{xy}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h}-\bar{x})(y_t-\bar{y}) \qquad (1)$$

Then the quantitative correlation strength was standardized by cross-correlation coefficient in Equation 2. This step normalized covariance, providing a clearer interpretation.

$$\rho_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\,\hat{\gamma}_y(0)}} \qquad (2)$$

Equation 3 finalized the expression of the cross-correlation coefficient, offering a systematic approach for a profound understanding of the dynamic relationship between the targeted time series.

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\,\gamma_y(t,t)}} \qquad (3)$$

The CCF values fall within the range of [-1,1], denoting the correlation of the series at various lag orders. Correlation was categorized into three groups: CCF values >0.4 were considered correlated (*14*); values >0.6 indicated strong correlation; while values <0.4 were deemed not correlated and were therefore excluded from the analysis.

Based on findings from pertinent studies, a maximum lag period of ±14 days was determined. The day displaying the highest values of CCF was identified as the optimal lag day. Optimal lag days were categorized into three groups: <0 days, =0 days, >0 days. These categories signify that the Baidu Index may have the capacity to anticipate fluctuations in influenza epidemics before, during, or after ILI%.

Data processing and graphical representation were performed using R (version 4.2.1) software, developed by R Core Team, headquartered in Vienna, Austria.

RESULTS

The findings from the lagged correlation analysis

TABLE 1. The classification of Baidu Index keywords related to influenza.

Classification	Keywords
Overall Index	Contains all the keywords below
Ordinary Index	Influenza; Cold; Viral Influenza; Seasonal Influenza; Influenza Virus; Influenza transmission route; How influenza is transmitted
Prevention Index	Flu Vaccine; Prevention of influenza; Precautions against influenza; Prevent influenza; How to prevent influenza; Flu Vaccine Side Effects; Flu Vaccine Prices; Is the Flu Vaccine Necessary; Influenza Prevention Knowledge
Symptom Index	Febrile; Fever; Cough; Pharyngalgia; Sore throat; Runny nose; Pneumonia; Chest tightness; Symptoms of influenza; Sneezing; Lacking in strength; Muscle soreness
Treatment Index	Flu Treatment; Cold Medicine; Antipyretic; Lianhuaqingwen*; What is the Most Effective Flu Medicine; Liuganwan; Ganmaoqingre*; Banlangen*; Baijiahei*; Oseltamivir; Tamiflu

* The names of traditional Chinese medicines.

TABLE 2. Description of formula components.

Equation component	Explanation
Equation 1	
$\hat{\gamma}_{\mathrm{xy}}$	The covariance between time series x and y at a lag of h time points.
$\mathrm{x}_{\mathrm{t+h}}$	The values of time series x and y at time points t+h and t.
$\bar{x}$ and $\bar{y}$	The means of time series x and y.
n	The length of the time series.
Equation 2	
$\rho_{xy}(h)$	The cross-correlation coefficient between time series x and y at a lag of h time points.
$\hat{\gamma}_{xy}(h)$	The covariance between time series x and y at a lag of h time points.
$\hat{\gamma}_x(0)$ and $\hat{\gamma}_y(0)$	The variances of time series x and y at time 0.
Equation 3	
$\rho_{xy}(s,t)$	The cross-correlation coefficient between time series x and y at time points s and t.
$\gamma_{xy}(s,t)$	The covariance between time series x and y at time points s and t.
$\gamma_x(s,s)$ and $\gamma_y(t,t)$	The variances of time series x at time s and y at time t.

between the Baidu Overall Index and ILI% reveal significant results. The CCF values fluctuate within the range of 0.46 to 0.86, indicating a lagged correlation pattern across 30 PLADs. Notably, 25 PLADs demonstrate a robust correlation trend (Figure 1). The median lag day between the Baidu Overall Index and ILI% is 0.5 days, with a range from −5 to 11 days. Specifically, 7 PLADs show a lag of less than 0 days, 7 PLADs display a lag of 0 days, and 16 PLADs exhibit a lag exceeding 0 days.

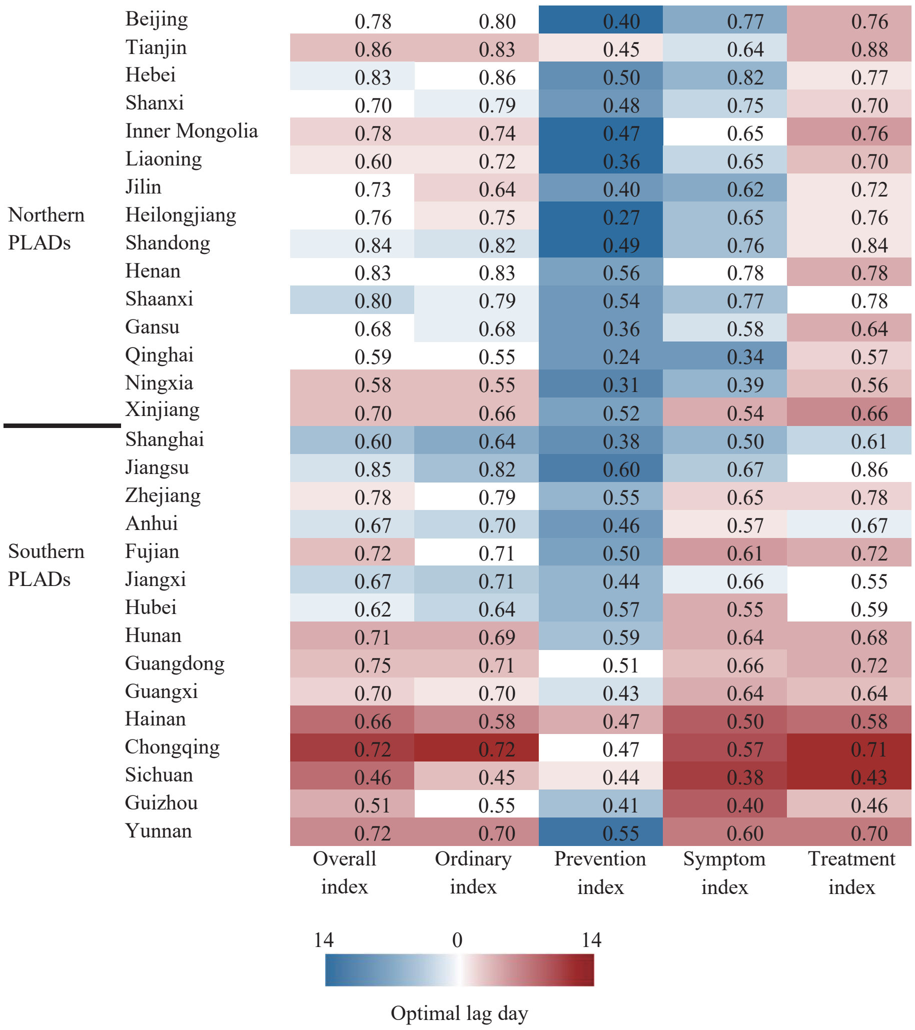
The Baidu Overall Index consists of four distinct subindices: Ordinary Index, Prevention Index, Symptom Index, and Treatment Index. The shortest median time discrepancy was observed in the Prevention Index at −9 days compared to the incidence of ILI%, with 18 PLADs registering a negative lag and 2 PLADs exhibiting no lag. The Symptom Index followed closely with a median lag of −0.5 days related to ILI%, where 13 PLADs displayed a negative lag and 2 PLADs recorded no lag. The Ordinary Index aligned most closely with ILI% showing a median lag of 0 days, with 9 PLADs experiencing a negative lag and 7 PLADs reporting no lag. In contrast, the Treatment Index showed the longest median lag of 3 days when compared to ILI%, with all 25 PLADs showing a lag exceeding 0 days.

When analyzed according to northern and southern regions, the median lag days for the Prevention Index, Symptom Index, and ILI% in the northern PLADs were −9 days and −5 days, while the Ordinary Index, Overall Index, and ILI% had median lag days of 0 days. The Treatment Index showed a median lag of 3 days behind ILI%. In the southern PLADs, the Prevention Index had a median lag of −6 days compared to ILI%, indicating its potential for early influenza outbreak detection. The Ordinary Index aligned closely with ILI%, with a median lag of 0 days. The Symptom, Treatment, and Overall Indexes showed median lags of 4, 3, and 3 days after ILI%, respectively (Figure 2).

DISCUSSION

This study reveals a significant correlation between the Baidu Overall Index and ILI%. Refining the classification of the Overall Index resulted in more advanced trends for the Prevention Index and Symptom Index, indicating potential value for warning of influenza disease epidemics. Notably, the Baidu Index trend advanced for a longer duration in the northern PLADs compared to the southern PLADs. This novel provincial-specific analysis of the Baidu Index deepens our insight into its intricate relationship with influenza transmission. Exploring geographic variations in the correlation between the Baidu Index and ILI% suggests regional differences in the utility of Internet search data for predicting future influenza-like illness epidemics.

The Prevention and Symptom Indexes demonstrate significant lead times, with median delays of −9 days and −0.5 days, respectively, compared to ILI%. On the other hand, the Treatment Index lags behind ILI% by a median lag of 3 days. Findings indicate that changes in the Prevention Index precede changes in the ILI% trend, possibly because individuals draw on past experiences and local outbreaks to proactively seek



Northern PLADs	Beijing	0.78	0.80	0.40	0.77	0.76
	Tianjin	0.86	0.83	0.45	0.64	0.88
	Hebei	0.83	0.86	0.50	0.82	0.77
	Shanxi	0.70	0.79	0.48	0.75	0.70
	Inner Mongolia	0.78	0.74	0.47	0.65	0.76
	Liaoning	0.60	0.72	0.36	0.65	0.70
	Jilin	0.73	0.64	0.40	0.62	0.72
	Heilongjiang	0.76	0.75	0.27	0.65	0.76
	Shandong	0.84	0.82	0.49	0.76	0.84
	Henan	0.83	0.83	0.56	0.78	0.78
	Shaanxi	0.80	0.79	0.54	0.77	0.78
	Gansu	0.68	0.68	0.36	0.58	0.64
	Qinghai	0.59	0.55	0.24	0.34	0.57
	Ningxia	0.58	0.55	0.31	0.39	0.56
	Xinjiang	0.70	0.66	0.52	0.54	0.66
Southern PLADs	Shanghai	0.60	0.64	0.38	0.50	0.61
	Jiangsu	0.85	0.82	0.60	0.67	0.86
	Zhejiang	0.78	0.79	0.55	0.65	0.78
	Anhui	0.67	0.70	0.46	0.57	0.67
	Fujian	0.72	0.71	0.50	0.61	0.72
	Jiangxi	0.67	0.71	0.44	0.66	0.55
	Hubei	0.62	0.64	0.57	0.55	0.59
	Hunan	0.71	0.69	0.59	0.64	0.68
	Guangdong	0.75	0.71	0.51	0.66	0.72
	Guangxi	0.70	0.70	0.43	0.64	0.64
	Hainan	0.66	0.58	0.47	0.50	0.58
	Chongqing	0.72	0.72	0.47	0.57	0.71
	Sichuan	0.46	0.45	0.44	0.38	0.43
	Guizhou	0.51	0.55	0.41	0.40	0.46
	Yunnan	0.72	0.70	0.55	0.60	0.70
		Overall index	Ordinary index	Prevention index	Symptom index	Treatment index

FIGURE 1. The optimal lag days and CCF values between Baidu Index and ILI% in 30 PLADs*.
Abbreviation: PLAD=provincial-level administrative division; CCF=cross-correlation function; ILI%=influenza-like illness percentage.
* The numbers on each plate represent the CCF values.

influenza prevention information before the onset of infection. The Symptom Index leads ILI% trends, potentially as individuals begin feeling unwell, search for symptom-related terms to identify potential causes, and subsequently seek care at outpatient clinics or emergency departments. In contrast, the Treatment Index trend lags behind ILI%, as individuals may search for influenza treatment-related keywords (such as medications) only when symptoms become severe or self-recovery is unsuccessful. It is important to acknowledge that fluctuations in epidemic surveillance data may introduce discrepancies in the correlation analysis results between the Baidu index and ILI%.

The predictive capability of the Baidu Index in anticipating changes in ILI% trends might be more effective in the northern region compared to the southern region. In the northern region, the median lag days between the Baidu Overall Index and ILI% were 0 days, while in the southern region they were 3 days. Upon further analysis of the subcategories of the four types of Baidu indexes, it was observed that two types of keywords (Prevention Index and Symptom Index) preceded the ILI% trend in the northern region, whereas only the Prevention Index did so in the southern region. This discrepancy could be attributed to the distinct seasonal influenza patterns in the northern region, characterized by more prominent single-peak epidemics than those in the southern region. These findings suggest that utilizing the Baidu Index for developing future influenza epidemic forecasting models may be more successful in the northern region than in the southern region.

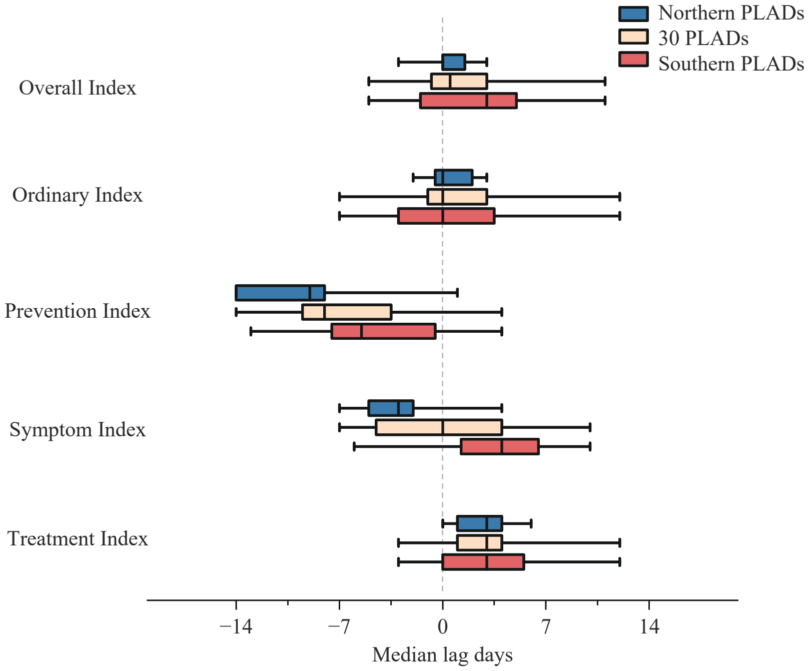


FIGURE 2. The median lag days between Baidu Index and ILI% in 30 PLADs, northern and southern PLADs.
Abbreviation: PLADs=provincial-level administrative divisions; ILI%=influenza-like illness percentage.

The Baidu Index demonstrated more prompt changes in trends compared to the ILI reporting process. According to the National Influenza Surveillance Technical Guidelines (2017 edition), surveillance sentinel sites report weekly cases every Monday (*15*), introducing a delay of 1–7 days from the case date. Conversely, the Baidu Index webpage displays data from the day before up to the following day, trailing just one day behind real-time conditions. Essentially, when the Baidu Index trend corresponds with the ILI% (with a lag of 0 days), it can provide insights into influenza epidemics 0–6 days in advance of the current reporting period, thereby circumventing delays in manual statistical reporting.

The study is subject to some limitations. Initially, it focused on categorizing and analyzing Baidu Index data relating to influenza-related keywords, overlooking an analysis of the lagged correlation between the Baidu Index of individual keywords and ILI%. Furthermore, the study did not account for discrepancies resulting from regional variations in the processes and methodologies of influenza-like illness reporting by provincial influenza surveillance sentinel sites. Future research could investigate the correlation between individual keywords and ILI% on a provincial or municipal level, leading to potential adjustments in the surveillance keywords used in each region based on findings. Additionally, a more in-depth exploration of keyword weighting may enhance surveillance accuracy and early warning capabilities.

In conclusion, a substantial correlation was observed between the Baidu Overall Index and ILI%. Additionally, the changes in trend for the Prevention Index and Symptom Index preceded those of the ILI%. Future research could explore the development of a predictive model using the Prevention Index and Symptom Index to anticipate and provide early warnings for influenza epidemics. The predictive nature of these indices could aid health authorities in identifying patient care requirements earlier in the influenza season, facilitating the prompt implementation of preventive measures. Furthermore, initiating early public health campaigns focused on influenza prevention could enhance public awareness.

Conflicts of interest: No conflicts of interest.

Funding: Supported by CAMS Innovation Fund for Medical Sciences (2021-I2M-1-044); National Key Research and Development Program of China (2023YFC2308701).

doi: 10.46234/ccdcw2024.084

# Corresponding authors: Ting Zhang, zt0416@126.com; Zhongjie Li, lizhongjiecdc@163.com; Weizhong Yang, yangweizhong@cams.cn.

[1] School of Population Medicine and Public Health, Chinese Academy of Medical Sciences (CAMS) & Peking Union Medical College, Beijing, China; [2] School of Health Policy and Management, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China; [3] Department of management science and information

system, Faculty of Management and Economics, Kunming University of Science and Technology, Kunming City, Yunnan Province, China.

Submitted: December 26, 2023; Accepted: March 26, 2024

REFERENCES

1. National Immunization Advisory Committee (NIAC) Technical Working Group (TWG), Influenza Vaccination TWG. Technical guidelines for seasonal influenza vaccination in China (2022-2023). Chin J Epidemiol 2022;43(10):1515 – 44. https://doi.org/10.3760/cma.j.cn112338-20220825-00734.
2. Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet 2018;391(10127):1285 – 300. https://doi.org/10.1016/S0140-6736(17)33293-2.
3. Lee VJ, Ho ZJM, Goh EH, Campbell H, Cohen C, Cozza V, et al. Advances in measuring influenza burden of disease. Influenza Other Respir Viruses 2018;12(1):3 – 9. https://doi.org/10.1111/irv.12533.
4. The Lancet. Preparing for seasonal influenza. Lancet 2018;391(10117):180. https://doi.org/10.1016/S0140-6736(18)30087-4.
5. WHO. Influenza (seasonal). 2018. http://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal). [2023-12-2].
6. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate statistics on COVID-19 are essential for policy guidance and decisions. Am J Public Health 2020;110(7):949 – 51. https://doi.org/10.2105/AJPH.2020.305708.
7. Paul R, Han D, DeDoncker E, Prieto D. Dynamic downscaling and daily nowcasting from influenza surveillance data. Stat Med 2022;41(21):4159 – 75. https://doi.org/10.1002/sim.9502.
8. Liang F, Guan P, Wu W, Huang DS. Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015. PeerJ 2018;6:e5134. https://doi.org/10.7717/peerj.5134.
9. Lai SJ, Feng LZ, Leng ZW, Lyu X, Li RY, Yin L, et al. Summary and prospect of early warning models and systems for infectious disease outbreaks. Chin J Epidemiol 2021;42(8):1330 – 5. https://doi.org/10.3760/cma.j.cn112338-20210512-00391.
10. Chinese National Influenza Center. Influenza weekly report of the national influenza center of China. https://ivdc.chinacdc.cn/cnic/zyzx/lgzb/. [2023-10-2]. (In Chinese).
11. Ali ST, Cowling BJ, Wong JY, Chen DX, Shan SW, Lau EHY, et al. Influenza seasonality and its environmental driving factors in mainland China and Hong Kong. Sci Total Environ 2022;818:151724. https://doi.org/10.1016/j.scitotenv.2021.151724.
12. Baidu Index Public Website. https://index.baidu.com. (In Chinese).
13. Yang LY, Zhang T, Han X, Yang J, Sun YX, Ma LB, et al. Influenza epidemic trend surveillance and prediction based on search engine data: deep learning model study. J Med Internet Res 2023;25:e45085. https://doi.org/10.2196/45085.
14. Shumway RH, Stoffer DS. Time series analysis and its applications. New York: Springer. 2000. http://dx.doi.org/10.1007/978-1-4757-3261-0.
15. Chinese National Influenza Center. National influenza surveillance technical guidelines (2017 edition). https://ivdc.chinacdc.cn/cnic/zyzx/jcfa/201709/t20170930_153976.htm. (In Chinese).