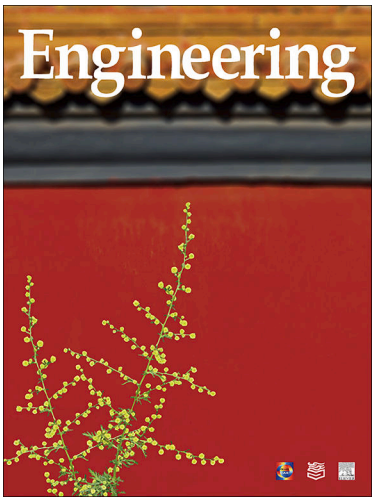
Journal Pre-proofs



Article

Methods on COVID-19 Epidemic Curve Estimation during Emergency Based on Baidu Search Engine and ILI Traditional Surveillance in Beijing, China

Ting Zhang, Liuyang Yang, Xuan Han, Guohui Fan, Jie Qian, Xuancheng Hu, Shengjie Lai, Zhongjie Li, Zhimin Liu, Luzhao Feng, Weizhong Yang

PII: S2095-8099(23)00375-2
DOI: https://doi.org/10.1016/j.eng.2023.08.006
Reference: ENG 1348

To appear in: *Engineering*

Received Date: 24 February 2023
Revised Date: 25 July 2023
Accepted Date: 28 August 2023

Please cite this article as: T. Zhang, L. Yang, X. Han, G. Fan, J. Qian, X. Hu, S. Lai, Z. Li, Z. Liu, L. Feng, W. Yang, Methods on COVID-19 Epidemic Curve Estimation during Emergency Based on Baidu Search Engine and ILI Traditional Surveillance in Beijing, China, *Engineering* (2023), doi: https://doi.org/10.1016/j.eng.2023.08.006

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company

Res Journal Pre-proofs

Public Health—Article

Methods on COVID-19 Epidemic Curve Estimation during Emergency Based on Baidu Search Engine and ILI Traditional Surveillance in Beijing, China

Ting Zhang [a,#], Liuyang Yang [a,b,#], Xuan Han [a,#], Guohui Fan [a], Jie Qian [a], Xuancheng Hu [b], Shengjie Lai [c], Zhongjie Li [a], Zhimin Liu [d,*], Luzhao Feng [a,*], Weizhong Yang [a,*]

[a] *School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China*

[b] *Department of management science and information system, Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650504, China*

[c] *WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK*

[d] *The Third Affiliated Hospital of Kunming Medical University, Kunming 650118, China.*

[#] These authors contributed equally to this work.

[*] Corresponding authors.

E-mail addresses: ych_lzm@163.com (Z. Liu), fengluzhao@cams.cn (L. Feng), yangweizhong@cams.cn (W. Yang)

ARTICLE INFO

Article history:

Received 24 February 2023

Revised 25 July 2023

Accepted 28 August 2023

Available online

Keywords

COVID-19

Epidemic curve

Baidu search engine

Influenza-like illness

Deep learning

Transmission dynamics model

Abstract

Surveillance is an essential work on infectious diseases prevention and control. When the pandemic occurred, the inadequacy of traditional surveillance was exposed, but it also provided a valuable opportunity to explore new surveillance methods. This study aimed to estimate the

2095-8099/© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Journal Pre-proofs

transmission dynamics and epidemic curve of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Omicron BF.7 in Beijing under the e[Journal Pre-proofs]gated recurrent unit (MABG)–susceptible–exposed–infected–removed (SEIR)) was developed, which leveraged a deep learning algorithm (MABG) to scrutinize the past records of ILI occurrences and the Baidu index of diverse symptoms such as fever, pyrexia, cough, sore throat, anti-fever medicine, and runny nose. By considering the current Baidu index and the correlation between ILI cases and coronavirus disease 2019 (COVID-19) cases, a transmission dynamics model (SEIR) was formulated to estimate the transmission dynamics and epidemic curve of SARS-CoV-2. During the COVID-19 pandemic, when conventional surveillance measures have been suspended temporarily, cases of ILI can serve as a useful indicator for estimating the epidemiological trends of COVID-19. In the specific case of Beijing, it has been ascertained that cumulative infection attack rate surpass 80.25% (95% confidence interval (95% CI): 77.51%–82.99%) since December 17, 2022, with the apex of the outbreak projected to transpire on December 12. The culmination of existing patients is expected to occur three days subsequent to this peak. Effective reproduction number (R_t) represents the average number of secondary infections generated from a single infected individual at a specific point in time during an epidemic, remained below 1 since December 17, 2022. The traditional disease surveillance systems should be complemented with information from modern surveillance data such as online data sources with advanced technical support. Modern surveillance channels should be used primarily in emerging infectious and disease outbreaks. Syndrome surveillance on COVID-19 should be established to following on the epidemic, clinical severity, and medical source demand.

1. Introduction

In recent years, emerging infectious diseases have been a persistent threat, causing harm to human life, health, economic development, and social order [1], and posing a potential risk to humankind. Disease surveillance is a fundamental element for preventing and controlling diseases and is also a requirement for eliminating infectious diseases. Therefore, establishing a surveillance and early-warning system is advantageous for detecting diseases earlier, thereby allowing for prompt response measures [2], which can diminish the peak of the epidemic and reduce the impact on health.

The current global coronavirus disease 2019 (COVID-19) outbreak has highlighted the inadequacies of traditional surveillance systems. With the policy of no longer considering infected individuals as the primary surveillance subjects, the reported cases cannot accurately reflect the actual infection rate, thus posing a challenge to traditional epidemic prevention and control. Nevertheless, the severity of the disease, the effects of symptoms on health, and the need for medical resources are still essential information that must be tracked. In this regard, it is necessary to reform traditional surveillance systems and pay attention to new types of surveillance, which may serve as a supplement to existing systems. The application of big data and the advancement of modern technology can help significantly in this regard.

The World Health Organization (WHO) proposed in May 2021 to develop a new model for surveillance of emerging threats, the Global Hub for Pandemic and Epidemic Intelligence [3]. This model aims to integrate traditional and modern big data surveillance methods, such as artificial intelligence, to combine different data sources and conduct interdisciplinary collaboration, thus increasing the availability of various data and connections. This project will make a significant leap forward in data analysis to aid decision-making [4]. Furthermore, media surveillance based on network search engines can make up for the shortcomings of traditional surveillance, especially in backward areas with underdeveloped surveillance networks or in periods of unstable surveillance due to major events and major infectious diseases. Studies have shown that Baidu, Daum, Twitter, Wikipedia, and other social media (including search engines) can be used to detect the prevalence of influenza, Zika virus [5], dengue fever [6], avian influenza [7], and hand, foot, and mouth disease [8].

Baidu is the most-used search engine in China. As of December 2021, the number of users is approximately 829 million, and 80.3% of them use search engines [9]. By July 2021, its monthly active users had exceeded 600 million, making it the largest search engine in the country with comprehensive coverage and usage. Thus, Baidu is an ideal choice to surveil the development of the epidemic due to its large population and widespread use, especially in Beijing. Given the prevalence of the Baidu search engine and the relatively stable usage habits of the population, this study verifies its effectiveness in surveilling the epidemic situation.

In the current global context, COVID-19 has been declared an end to the public health emergency of international concern [10]. The pathogenicity has weakened, vaccination rates have increased, and experience in prevention and control has accumulated. In China, the goal is to reduce influences on healthcare while considering economic and social impacts, given limited medical treatment and social prevention and control resources. To this end, greater attention should be paid to risk surveillance of key populations and treatment of severe and critical illnesses. Symptom surveillance can provide insight into the epidemic of infectious diseases and is an essential indicator of disease focus, which can also increase the demand for medical resources.

“In dealing with a complex crisis, we should establish upfront which dimension to prioritise, and adapt more quickly to changing situations to not allow the perfect to become the enemy of the good.” as the *White paper on Singapore’s response to COVID-19: lessons for the next pandemic summarized* [11]. When faced with an emergency outbreak, it becomes necessary to adopt innovative approaches to overcome the limitations of traditional surveillance methods. This study examined the use of modern surveillance channels alongside conventional methods in emergency situations to evaluate the scale of COVID-19 infection. The results provide a valuable methodological reference for future infectious disease surveillance, utilizing real-world observations of the pandemic to inform surveillance strategies.

2. Methods

Journal Pre-proofs

2.1. Data sources

This study used the daily number of influenza-like illness (ILI) cases in Beijing and the daily proportion of ILI among the outpatients (ILI%) as the dependent variables and the daily Baidu index as an independent variable. The research period was from July 1, 2013, to December 9, 2022. The ILI data were collected from 419 sentinel hospitals in 21 districts of Beijing, with a total of 1 275 742 samples. The Baidu index was formulated using six keywords, including fever, pyrexia, cough, sore throat, anti-fever medicine, and runny nose, which were sourced from both mobile and personal computer platforms.

WHO and the Centers for Disease Control and Prevention (CDC) define an ILI as an acute respiratory illness with a temperature of at least 38 °C (100 °F) and associated cough, with onset within the past ten days [12]. For the 2021–2022 influenza season, case definitions no longer require “no other known etiology other than influenza” [13]. The ILI definition issued by the Department of Disease Control and Prevention of the National Health Commission of China is: fever (body temperature ≥ 38 °C) accompanied by either cough or sore throat [14]. These definitions of ILI only differ slightly in body temperature, and the composition of symptoms is the same. Additionally, no etiological tests are conducted to confirm the diagnosis of ILI, which includes the current pandemic of COVID-19.

Data sharing statement：the Baidu search data in this study are publicly available, the influenza virological surveillance data in Beijing were retrieved from a previously published study [15].

2.2. Data preprocessing

Data standardization involves the process of adjusting the values in a dataset to a specific scale, thereby enabling different variables to be compared with one another while also eliminating the impact of varying magnitudes. This technique can enhance data quality, streamline data processing, improve model precision, expedite model convergence, reduce model training duration, and enhance the stability and reliability of the model.

In the current study, the data underwent pre-processing utilizing Min–Max scaling of the following aggregation. The normalization method adopted was off-difference, where the data underwent linear scaling based on the maximum and minimum values to ensure that the scaled data values fall within the range of [0, 1]. This range was deemed suitable for observation and training purposes. The normalized thermal distribution of each feature is presented in Fig. 1.

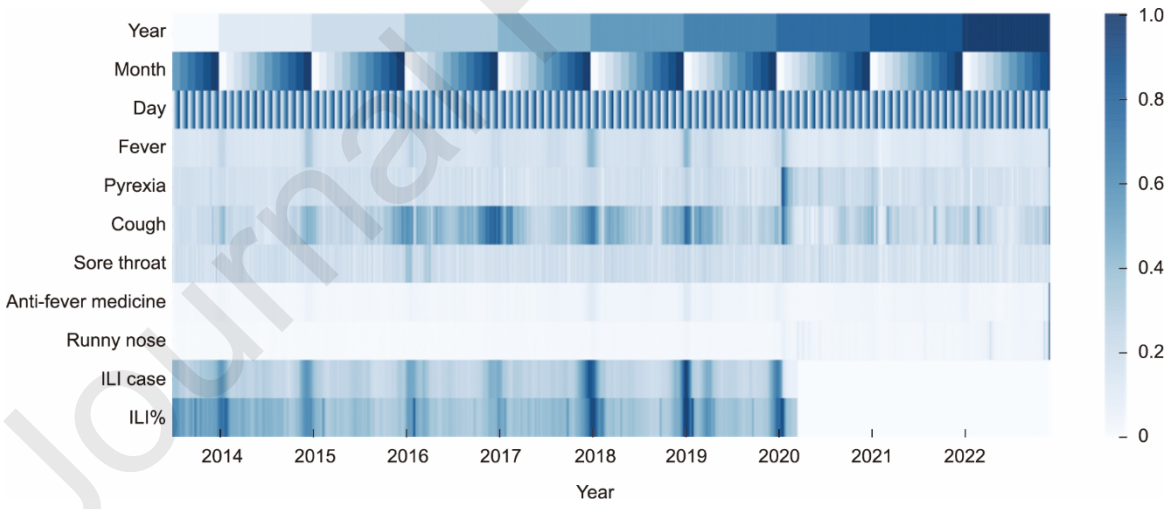


Fig. 1. Thermal distribution of each feature after standardization. To ensure equitable inclusion in model training, we normalize multi-source data using a Min–Max scale within the range of [0,1]. In the corresponding visual representation, lighter colors are indicative of values closer to 0, while darker colors signify values approaching 1, as illustrated in the legend.

2.3. Establishment of the dataset

(1) Training: July 1, 2013 to May 28, 2018 (1793 days).

(2) Validation: May 28, 2018 to March 24, 2019 (300 days).

(3) Testing: March 24, 2019 to March 23, 2020 (365 days).

(4) Prediction: October 10, 2022 to December 9, 2022 (60 days).

(5) Estimation: November 22, 2022 to January 20, 2023 (60 days)

Journal Pre-proofs

2.4. Modeling

This study employed a composite model that combined deep learning and a transmission dynamics model to predict the COVID-19 epidemic. First, we used the MABG model to predict the current ILI% and ILI case. Given the multidimensional nature of our data, we developed a prediction model based on the multiattention mechanism and bidirectional gated recurrent unit to handle multi-featured time series. By thoroughly exploring the inherent characteristics of multi-source heterogeneous data and establishing the connection between characteristics and results, the MABG model was able to complete the task of time series prediction effectively and reliably.

When a multi-featured time series was fed to the model, we first connected it to a bidirectional gated recurrent unit (GRU) layer, which was good at processing time series and capturing features between step intervals in the time series. The bidirectional GRU (BGRU) is an improved version of the GRU that offers several advantages, including a higher level of global information utilization, prediction capability, and modeling ability. Unlike traditional recurrent neural networks that can only consider the input of the current moment and the implied state of the previous moment, the BGRU can utilize the information of the before and after states of the current moment. This approach facilitates better global information capture and more accurate output prediction. The structure of the BGRU model is illustrated in Fig. 2.

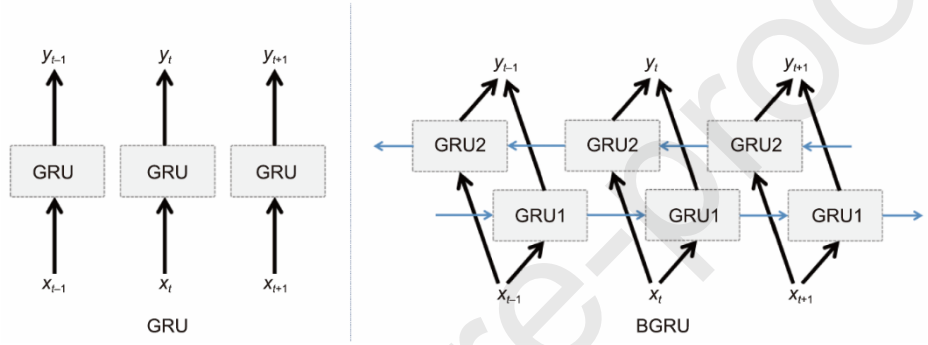


Fig. 2. GRU and BGRU.

Then, we employed three different attention mechanism modules simultaneously: squeeze and excitation attention [16], channel attention, and spatial attention [17]. Combining these three attention mechanisms, we extracted important information between different features and key information within the same feature. In addition, to prevent the gradient from disappearing, after concatenating the results of different attention modules, we connected the results with two pooling layers for residual connection and output the prediction results through the dense layer (Fig. 3).

Journal Pre-proofs

Journal Pre-proofs

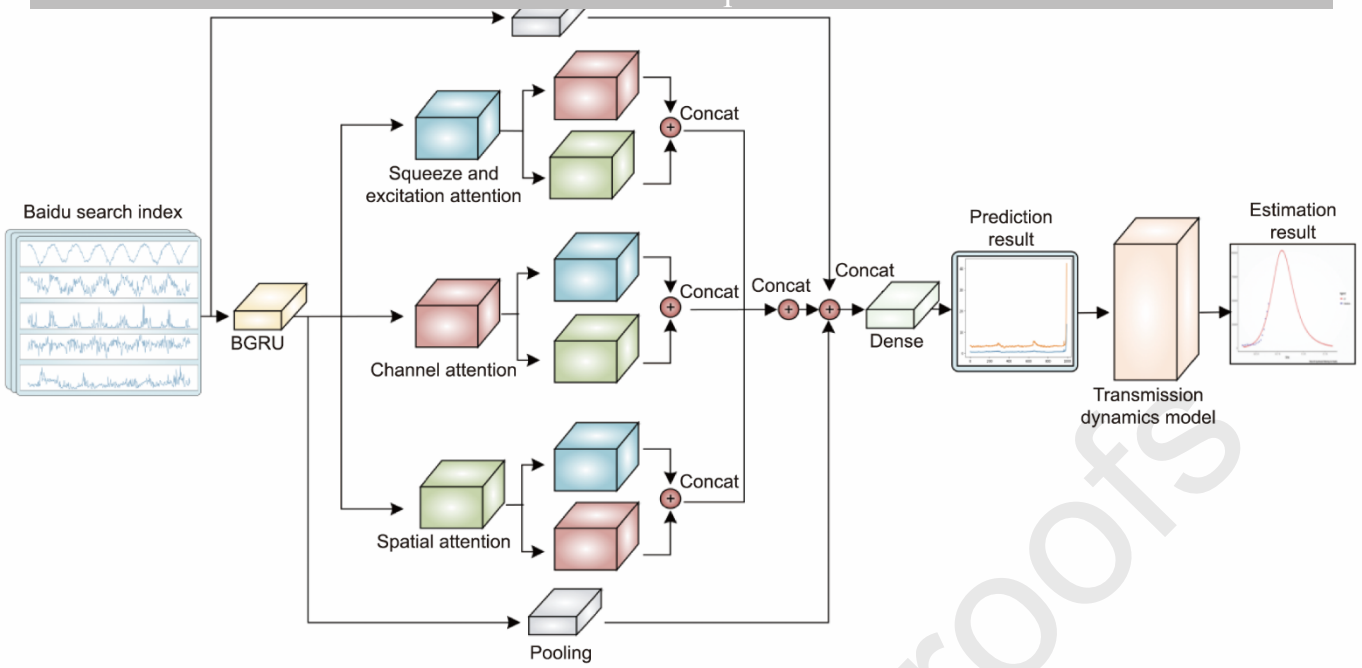


Fig. 3. MABG–susceptible–exposed–infected–removed (SEIR) model structure. Concat: contatenate.

Finally, the study utilized a classical transmission dynamics model to estimate the epidemic curve of COVID-19 infection in Beijing, incorporating predicted results. The transmission dynamics model has various versions, depending on the study's objectives, and requires defining related parameters to evaluate the effectiveness of pharmaceutical/non-pharmaceutical interventions. To predict the epidemic trend, essential factors must be considered. This study aimed to estimate the epidemic trend based on actual information, utilizing an optimal solution set based on real-time data. The equation used in this study marked the influence of different factors, but the focus was not to distinguish the impact of each factor. Therefore, the index of comprehensive effect was used as a substitute when seeking the optimal solution. The total population, N, was categorized into four classes: susceptible (S), exposed (E), infected (I), and recovered/removed (R). The governing differential equation (1) was as follows. A continuous time variable model was established to account for the continuous infection process, as expressed by the Eq. (1).

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \Lambda N - \mu S - (1-c)(1-v)\delta S\frac{I}{N} \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= (1-c)(1-v)\delta S\frac{I}{N} - (\mu+\alpha)E \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha E - (\mu+\gamma)I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R \\
N &= S + E + I + R + vN \\
\beta &= (1-c)(1-v)\delta
\end{aligned} \tag{1}$$

Journal Pre-proofs

where Eq. (1) are subject to the initial conditions S(0), E(0), I(0), and R(0). The parameters are defined as: t: time; A: per-capi the effectiveness of all kinds of pharmaceutical interventions; δ: the probability of disease transmission per contact (dimensionless) times the number of contacts per unit time; α: rate of progression from exposure to infectious (the reciprocal is the latent period); γ: recovery or death rate of infectious individuals (the reciprocal is the infectious period). In this study, we did not distinguish the effects of c, v, and δ, but considered their effects together, denoted by the rate per unit of time at which the susceptible become infected β, which could be calculated by R_0 depend on Eq. (2).

$$R_0 = \frac{\beta\alpha}{(\mu + \alpha)(\mu + \gamma)} \quad (2)$$

2.5. Assessing the scale of COVID-19 infections in comparison to ILI

In the past, surveillance of ILI in China did not include patients with COVID-19 infections. However, this study took into account those with ILI among the existing COVID-19-infected patients (Fig. 4). In addition to those with ILI symptoms, COVID-19 infection also includes asymptomatic cases. Therefore, the ILI estimated by the model was first adjusted according to the historical level and the prevalence level of people without COVID-19 infections. This allowed for the subtraction of the non-ILI population to derive the number of ILI populations infected with COVID-19. Then, based on the proportion of asymptomatic infections of Omicron, the adjustment was made to obtain a rough estimate of the scale of COVID-19. The proportion of asymptomatic infections concerning overall infections was subject to variables such as age distribution, general health status, underlying health conditions, and vaccination coverage. As per previous systematic reviews, meta-analyses [18,19], and official reports [20,21], the asymptomatic proportion ranged from 25.3% to 40.0%. This study was established based on an assumed a symptomatic proportions of 30%.

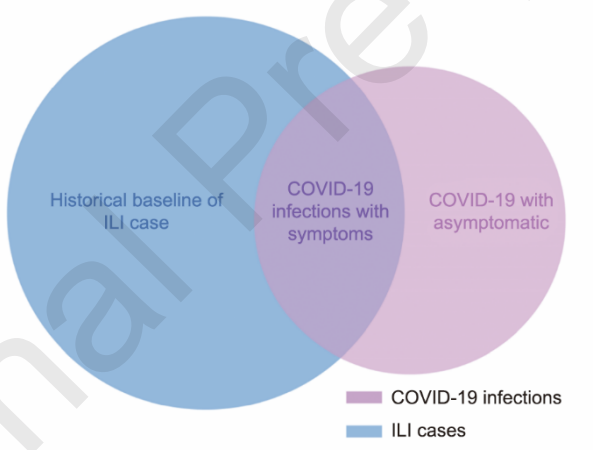


Fig. 4. Assessing the scale of COVID-19 infections based on ILI. The relationship between ILI and COVID-19 patients.

2.6. Study assumptions

(1) Assuming that the motivation of search behavior remains relatively constant once symptoms of ILI are present.

(2) The definition of ILI encompasses the primary symptoms of COVID-19.

(3) The assuming is that the current policy is maintained without considering the potential policy alterations as the epidemic peak approaches.

(4) The prevalence of other ILI diseases did not differ from historical levels.

3. Results

3.1. Model validation

This study was validated by comparing the predicted and actual values from May 28, 2018 to March 24, 2019 (Fig. 5). The R^2 values (a value between 0 and 1, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables in the model) of ILI cases and ILI% were 0.6540 and 0.6057, the explained variance scores (EVSs)

were 0.6596 and 0.6069, the mean absolute errors (MAEs) were 0.1145 and 0.5629, and the mean squared errors (MSEs) were

Journal Pre-proofs

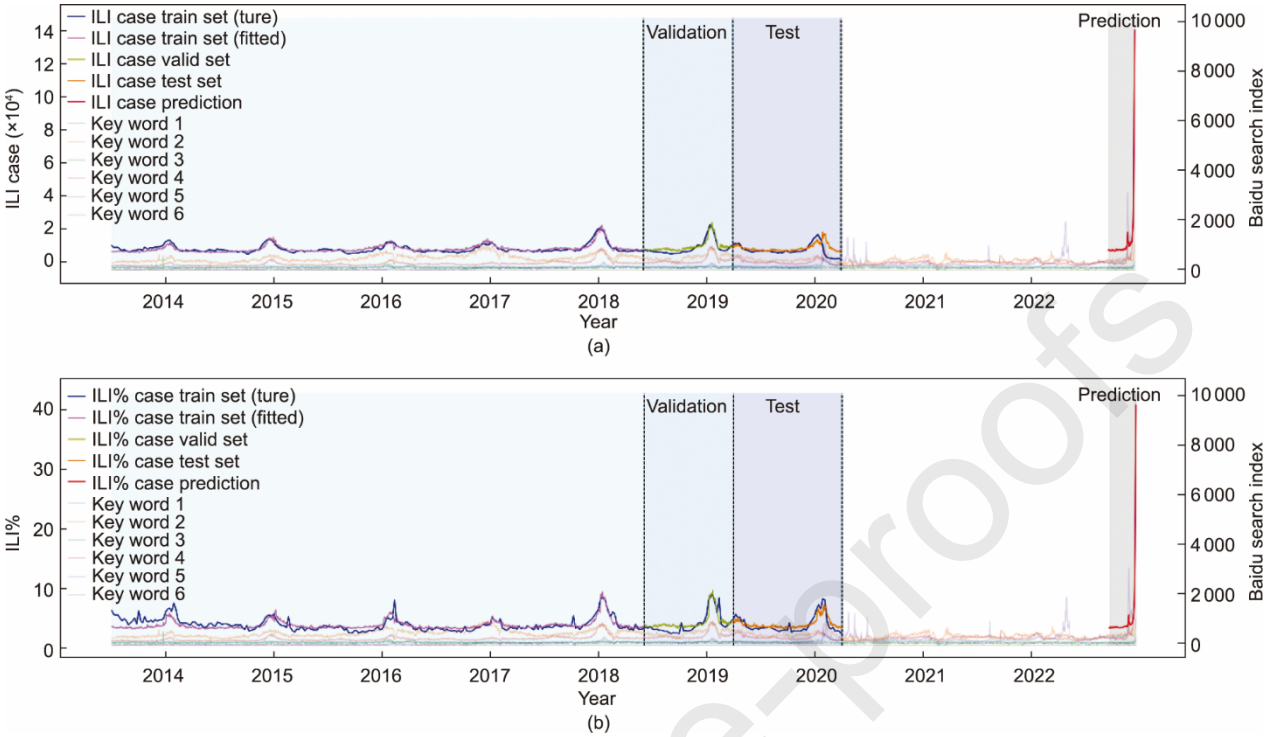


Fig. 5. Prediction of ILI using the Baidu index. (a) The number of ILI cases at Beijing Sentinel Hospital. (b) The percentage of ILI cases at Beijing Sentinel Hospital.

Table 1

Comparison of the ILI% and ILI case between different models.

Model category	R^2		EVS		MAE		MSE	
	ILI%	ILI case	ILI%	ILI case	ILI%	ILI case	ILI%	ILI case
ES	0.1801	0.5532	0.5747	0.5533	0.1002	0.0806	0.0194	0.0124
RF	0.5360	0.2337	0.5530	0.2468	0.6366	2.0300	0.9252	9.1966
XGB	0.4329	0.2499	0.5467	0.4247	0.0896	0.0941	0.0127	0.0171
LSTM	0.5128	0.5788	0.5223	0.5794	0.633	0.1314	0.7029	0.0362
BGRU	0.5752	0.6075	0.5801	0.6076	0.5896	0.1239	0.6128	0.0338
Informer [22]	0.3066	0.1341	0.3601	0.3307	0.5874	0.9668	0.7975	2.6335
MABG	**0.6057**	**0.6540**	**0.6069**	**0.6596**	**0.5629**	**0.1145**	**0.5688**	**0.0298**

ES: exponential smoothing; RF: random forest; XGB: extreme gradient boosting; LSTM: long short-term memory.

Journal Pre-proofs

3.2. ILI estimation results based on the Baidu index

Analysis of the Baidu index and ILI data concerning the emergence of COVID-19 since January 2020 revealed that ILI cases and ILI% had surpassed the historical baseline levels from December 1, 2022 ($p < 0.05$). Furthermore, the number of ILI cases surged in November and December, prior to the government's historic policy adjustments on December 7, 2022. These findings suggest that the epidemic had already reached a large scale before the official policy changes were enacted (Fig. 6(a)).

3.3. Comparison of ILI% and ILI cases among different models.

We also compared the MABG model with other standard traditional statistical models, machine learning, and deep learning models using four metrics R^2 (Eq. (3)), EVS (Eq. (4)), MAE (Eq. (5)), and MSE (Eq. (6)). The calculation methods of the four metrics are shown below. The results are shown in Table 1, from which we can see that the MABG model we used outperforms other models in most evaluation metrics.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{yi})^2}{\sum_{i=1}^{n}(yi - \overline{yi})^2} \quad (3)$$

$$\mathrm{EVS}(y,\hat{y}) = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)} \quad (4)$$

$$\mathrm{MAE}(y,\hat{y}) = \mathrm{median}\left(|y1 - \hat{y1}|,\ldots,|yn - \hat{yn}|\right) \quad (5)$$

$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{n_{\mathrm{samples}}}\sum_{i=0}^{n_{\mathrm{samples}}-1}(yi - \hat{yi})^2 \quad (6)$$

where y is the actual observed values of the dependent variable; $\hat{y}$ is the predicted or estimated values of the dependent variable based on the model; n is the total number of data points or observations in the dataset; i is an index that represents each individual data point in the dataset, ranges from 1 to n.

3.4. Model application on the epidemic curve estimation of COVID-19 infection in Beijing

The present study utilized a variation susceptible–exposed–infected–removed (SEIR) model to analyze the epidemiological characteristics of COVID-19 in Beijing. The parameters were calculated based on the infections estimated through the ILI model. The resident population of Beijing is 21 893 095 [23], with over 80% having received the COVID-19 vaccination booster [24]. The birth rate of Beijing in 2021 is 0.635%, and the death rate is 0.539% [25]. Approximately 30% of the population is assumed to be asymptomatic during infections. The transmission dynamics of COVID-19 were modeled to simulate the epidemic curve in Beijing. The relevant parameter settings are shown in Table 2. The results of the variation SEIR model suggest that the epidemic's peak is expected to occur on December 12, with about 1.66 (95% confidence interval (95% CI): 1.61–1.72) million new infections at peak time. The outbreak is expected to conclude in early January. The peak of existing patients' curve, which refers to the increase in new infections and decrease in recoveries/deaths, is expected to occur on December 15 with more than 5.47 (95% CI: 5.22–5.73) million existing patients at peak time (Fig. 6(a)). The duration between the peak of new infections and the peak of existing patients is estimated to be three days. We estimated that the cumulative infection attack rate was 80.25% (95% CI:77.51%–82.99%) on December 17, and 97.50% (95% CI: 97.00%–98.00%) on January 15, 2023 (Fig. 6(b)). The overall trend of corresponding estimated effective reproduction number (R_t) kept fluctuating dropping, and it remained below 1, 0.92 (95% CI: 0.90–0.95), since December 17, 2022 (Fig. 6(c)).

Table 2

Parameters for SEIR model to estimate epidemic curve of COVID-19 infection in Beijing.

Parameter	Value

Journal Pre-proofs

N	
Λ	0.635%
μ	0.539%
β	1.00[a]
α	0.50[a]
γ	0.25[a]

[a] These parameters were inferred optimal solutions based on the results of the MABG model.

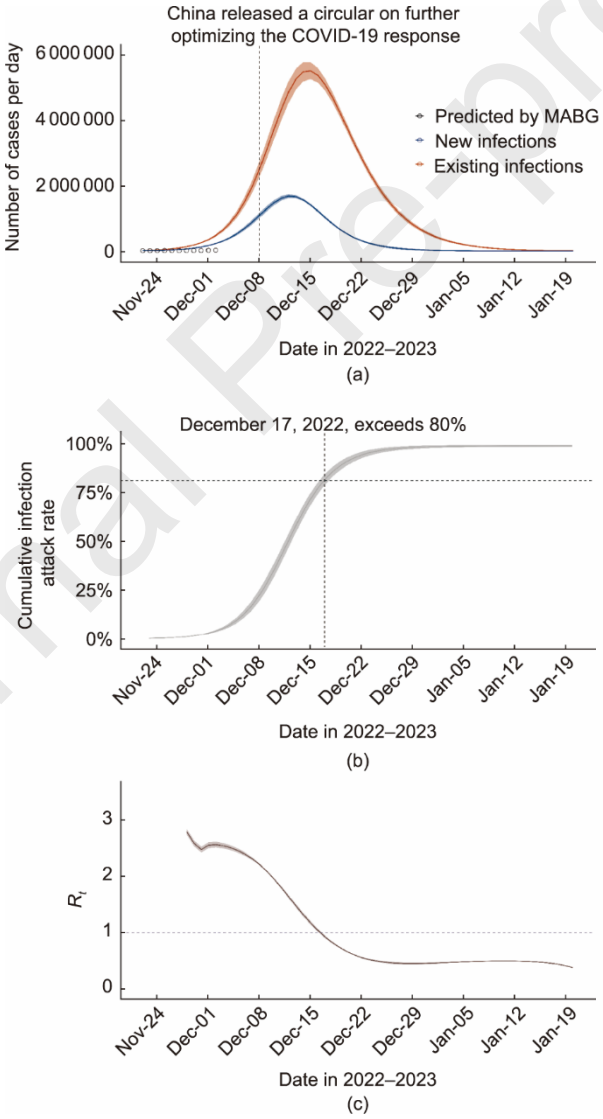


Fig. 6. Based on the Baidu search engine and ILI surveillance to simulate the COVID-19 epidemic curve in Beijing. (a) Existing and new infections per day. The dark black points are the estimated case by the MABG model, and the blue lines represent new infections per day while the orange line represents existing patients per day. (b) Cumulative infection attack rate per day. (c) R_t from November 28, 2022–January 20, 2023.

4. Discussion

This research investigated the implementation of the Baidu index to predict the magnitude of ILIs at sentinel hospitals in BeijJournal Pre-proofsobal surveillance. Additionally, the estimation of the size of the population infected by COVID-19 in cities with policy changes was also examined. The findings showed that the number of ILIs in Beijing has surpassed the historical average since December, a trend which could be attributed to the rise in COVID-19 cases. However, an increase in other respiratory infection cases could not be ruled out. Furthermore, the study also revealed changes in Beijing residents' medical-seeking behaviors and habits during the pandemic. At 419 sentinel hospitals included in the study, the number of people with ILI cases and related symptoms increased rapidly. Finally, Baidu provided new ideas for the surveillance of this round of the COVID-19 pandemic.

The positive nucleic acid testing rate [26] and Baidu search data were both peaked on December 14, providing a valuable cross-validation of the COVID-19 epidemic trend estimation based on two distinct data sources. The purpose of COVID-19 nucleic acid testing is to detect new cases of infection, and once a positive result is obtained, frequent testing is unlikely. Therefore, nucleic acid testing does not reflect the current infected individuals, but rather identifies newly infected individuals in the early stages of the disease. In this study, the peak of the positive rate of nucleic acid testing is compared with the peak of new infections daily. Since December 8, 2022, the nucleic acid testing strategy has shifted from population-wide testing to voluntary testing. Therefore, the absolute values presented in the nucleic acid testing data cannot represent the number of infections, and they are not directly comparable to the absolute values of infections in this study. To a certain degree, the concurrence of peak times provides empirical validation for the reliability of the study method. It is important to note that the model should be tailored to the specific application scenario of the transmission dynamics model, rather than striving for excessive complexity and detail.

This study aligns with Kathy Leung's research [27], which estimated the transmission dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Omicron BF.7 in Beijing from November to December 2022. Both studies indicate that the infection peaked before mid-December, 2022, with around 92% of the population infected as of December 22, 2022. But our study found a 97.50% infection rate (95% CI: 97.00%–98.00%) as of January 15, 2023, notably higher than Kathy Leung's estimates. This discrepancy may stem from our model's uniform assumptions about social interaction, which overlook subgroups like the self-isolating or those with limited mobility, potentially inflating the infection rate. Furthermore, our study observed the infection incidence peaking one day later and witnessed a similar rapid increase in the proportion of the population infected. However, the maximum values of Rt in this study (2.79) are lower than them (3.44). This discrepancy may be attributed to different assumptions, data sources, and model parameter errors between the two studies. Therefore, the significance and applicability of the study's results should be carefully considered in light of the based data source, research hypothesis, and model structure. It is important to acknowledge that this model encounter challenges when attempting to accurately reflect real-world circumstances.

WHO proposes that traditional surveillance of infectious diseases, such as ILI, includes patients receiving medical services, hospitalized patients, laboratory confirmation, gene sequencing, death estimation, active surveillance, tracking, etc. Modern surveillance techniques such as network information, animal health, occupational health, policy reports, community-reported cases, mobile data, public databases, and wearable devices are being employed to supplement these traditional methods. In particular, the use of the Baidu index as a supplementary means of ILI surveillance is an example of this modern surveillance. Studies have demonstrated that modern surveillance methods, such as Google Flu Trends (GFT), can detect signs of disease occurrence earlier than traditional methods, being able to detect the occurrence of ILI one week in advance [28]. These Internet-based systems improve the sensitivity of surveillance for developed countries and may be more effective for countries with underdeveloped traditional surveillance systems [8].

The significance of syndrome surveillance lies in its ability to quantify the magnitude of an outbreak and ascertain the demand for medical resources and strategize accordingly. The findings of this study demonstrate that following a surge in new infections at the 9-day mark, there was a subsequent surge in the number of existing patients, posing a significant challenge for the healthcare sector [29]. The severity of a disease's symptoms often leads to an increased likelihood of seeking medical treatment. In situations where laboratory testing is unavailable or unnecessary, it is still important to consider the health and recovery of those infected. Therefore, estimating the number of ILI cases in a particular area can help assess the demand for medical resources. However, it is essential to note that the predicted number of cases refers to the number of people seeking treatment at sentinel surveillance sites, not the total number of ILIs in the area. To obtain an accurate representation of the area's ILI rate, the hospital's coverage of services must be taken into account.

Syndrome surveillance is essential for the control and prevention of influenza at a global level [30]. The aim of these strategies should be to maximize the health benefits of the population while avoiding economic disruption. For this purpose, surveillance efforts should be concentrated on symptomatic infected individuals. A study [31] conducted in Chaoyang District, Beijing, demonstrated that intensifying influenza surveillance and conducting a comprehensive analysis of the surveillance results can assist in the timely detection of influenza and enable more precise measures to be taken. Additionally, public data from the Baidu search engine can be used to infer the prevalence of respiratory infectious diseases more comprehensively,

which can be utilized to anticipate any potential shortage of medical resources, thus allowing for timely adjustments to prev

Journal Pre-proofs

It is recommended to surveillance the symptoms of COVID-19 based on or in reference to the ILI system of influenza surveillance. The COVID-19 pandemic is expected to persist [32]. Surveillance of the symptoms of COVID-19 is essential to comprehend the magnitude of the disease, evaluate the epidemic trend, and assess the demand for medical resources and the burden of the disease. In the past, ILI surveillance sentinel sites in China [33], the United States [34], Japan [35], and the United Kingdom [36] have been instrumental in the surveillance of influenza. The population's susceptibility and the burden of the disease associated with COVID-19 are higher than those of influenza. Adjustment of preventive measures, preparation for a response, and virus mutation all depend on effective surveillance.

There are some limitations. This study has only estimated the number of people visiting a doctor or obtaining medication, which did not reflect the actual number of infections or symptoms. The SEIR model calculates certain parameters based on assumptions, which can limit their credibility in accurately representing the real world. As a result, not all parameters, such as the recovery rate, may be reliable indicators of real-world dynamics. Also, the SEIR model also could not incorporate all real-world factors into the estimation model. Various factors, such as weather conditions, traffic conditions, holidays, and the risk of cross-infection, influence this behavior. Additionally, this study did not include all Baidu indexes related to influenza-like cases because the Baidu index is subject to interference and guidance from numerous sources, thus introducing certain levels of uncertainty. Furthermore, this study did not differentiate between influenza virus infection, COVID-19, rhinovirus infection, and other specific diseases.

5. Conclusion

The Baidu index effectively gauges the quantity and proportion of individuals who manifest influenza-like symptoms and subsequently visit sentinel hospitals or procure medication within a reliable range. Additionally, Baidu index can be utilized to calculate the dissemination of a virus and the rate of contagion during a pandemic.

Acknowledgments

This study was supported by grants from the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (2021-I2M-1-044). All authors would extend thanks to Baidu for the data publication and Sinosoft Company Limited for technical support.

Authors' contribution

Weizong Yang, Luzhao Feng, and Ting Zhang contributed to the study design; Liuyang Yang, Xuan Han, and Xuancheng Hu were responsible for data collection and curation; Liuyang Yang, Ting Zhang, Zhongjie Li, and Zhimin Liu verified and analyzed the data; Jie Qian and Xuan Han conducted literature review; Ting Zhang, Xuan Han, and Liuyang Yang wrote the first draft of the manuscript; Weizhong Yang, Luzhao Feng, Zhimin Liu, Zhongjie Li, Shengjie Lai, and Guohui Fan reviewed and contributed to the writing of the manuscript. All authors had full access to all the data in the study, approved the revisions, and had final responsibility for the decision to submit for publication.

Compliance with ethics guidelines

Ting Zhang, Liuyang Yang, Xuan Han, Guohui Fan, Jie Qian, Xuancheng Hu, Shengjie Lai, Zhongjie Li, Zhimin Liu, Luzhao Feng, and Weizhong Yang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Ellwanger JH, Kaminski VL, Chies JAB. Emerging infectious disease prevention: where should we invest our resources and efforts? J Infect Public Health 2019;12(3):313–6.

[2] Son WS, Park JE, Kwon O. Early detection of influenza outbreak using time derivative of incidence. EPJ Data Sci 2020;9(1):28.

[3] Morgan OW, Abdelmalik P, Perez-Gutierrez E, Fall IS, Kato M, Hamblion E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. Nat Med 2022;28(8):1526–8.

[4] 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.

Journal Pre-proofs

[5] McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Negl Trop Dis 2017;11(1):e0005295.

[6] Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLOS Comput Biol 2017;13(7):e1005607.

[7] Chen Y, Zhang Y, Xu Z, Wang X, Lu J, Hu W. Avian influenza A (H7N9) and related Internet search query data in China. Sci Rep 2019;9(1):10434.

[8] Zhao Y, Xu Q, Chen Y, Tsui KL. Using Baidu index to nowcast hand-foot-mouth disease in China: a meta learning approach. BMC Infect Dis 2018;18(1):398.

[9] statcounter. Tablet search engine market share China [Internet]. Dublin: statcounter; [cited 2023 May 5]. Available from: https://gs.statcounter.com/search-engine-market-share/tablet/china#monthly-202111-202211.

[10] WHO. Statement on the fifteenth meeting of the IHR (2005) emergency committee on the COVID-19 pandemic [Internet]. Geneva: WHO; 2023 [cited 2023 May 5]. Available from: https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-%282005%29-emergency-committee-regarding-the-coronavirus-disease-%28covid-19%29-pandemic.

[11] Government of Singapore. White Paper on Singapore's response to COVID-19: lessons for the next pandemic. Singagpore: Government of Singapore; 2023.

[12] WHO. WHO surveillance case definitions for ILI and SARI [Internet]. Geneva: WHO; [cited 2023 May 5]. Available from: https://www.who.int/teams/global-influenza-programme/surveillance-and-surveillance.

[13] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases. U.S. Influenza surveillance: purpose and methods [Internet]. Atlanta: CDC; [cited 2023 May 5]. Available from: https://www.cdc.gov/flu/weekly/overview.htm.

[14] Feng L, Zhang T, Wang Q, Xie Y, Peng Z, Zheng J, et al. Impact of COVID-19 outbreaks and interventions on influenza in China and the United States. Nat Commun 2021;12(1):3249.

[15] Yang L, Li G, Yang J, Zhang T, Du J, Liu T, et al. Deep-learning model for influenza prediction from multisource heterogeneous data in a megacity: model development and avaluation. J Med Internet Res 2023;25:e44238.

[16] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 2020;42(8):2011–23.

[17] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. 2018. p. 3–19.

[18] Shang W, Kang L, Cao G, Wang Y, Gao P, Liu J, et al. Percentage of asymptomatic infections among SARS-CoV-2 Omicron variant-positive individuals: a systematic review and meta-analysis. Vaccines 2022;10(7):1049.

[19] Oran DP, Topol EJ. The proportion of SARS-CoV-2 infections that are asymptomatic: a systematic review. Ann Intern Med 2021;174(5):655–62.

[20] ECDC-EU. Clinical characteristics of COVID-19 [Internet]. Solna: ECDC-EU; 2022 [cited 2022 Aug 15]. Available from: https://www.ecdc.europa.eu/en/covid-19/latest-evidence/clinical.

[21] PHAC. COVID-19 signs, symptoms and severity of disease: a clinician guide [Internet]. Ottawa: PHAC; 2022 [cited 2022 Jul 1]. Available from: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/guidance-documents/signs-symptoms-severity.html#a5.

Journal Pre-proofs

[22] Clemons PA, Bittker JA, Wagner FF, Hands A, Danpšík V, Schreiber SL, et al. The use of informer sets in screening: perspectives on an efficient strategy to identify new probes. SLAS Discov 2021;26(7):855–61.

[23] National Bureau of Statistic. Major figures on 2020 population census of China. Beijing: China Statistics Press; 2021 [cited 2022 Jul 1]. Available from: https://www.gov.cn/guoqing/2021-05/13/content_5606149.htm?eqid=cf2ff410000631d70000000664560881. Chinese.

[24] Beijing Municipal Health Commission. COVID-19 vaccination in Beijing [Internet]. Beijing: Beijing Municipal Health Commission; 2022 [cited 2022 Apr 18]. Available from: http://wjw.beijing.gov.cn/xwzx_20031/wnxw/202204/t20220418_2680279.html. Chinese.

[25] NBS. Per-capita birth rate and per-capita natural death rate of Beijing [Internet]. Beijing: NBS; 2021 [cited 2023 Jul 23]. Available from: https://data.stats.gov.cn/search.htm?s=%E5%8C%97%E4%BA%AC%20%E4%BA%BA%E5%8F%A3%E5%87%BA%E7%94%9F%E7%8E%87. Chinese.

[26] CDC. COVID-19 clinical and surveillance data— December 9, 2022 to January 23, 2023, China [Internet]. Beijing: CDC; 2023 [cited 2022 Jul 1]. Available from: https://weekly.chinacdc.cn/news/covid-surveillance/bfa0d054-d5bf-42bb-b8b4-f7ce34539b74_en.htm.

[27] Leung K, Lau EHY, Wong CKH, Leung GM, Wu JT. Estimating the transmission dynamics of SARS-CoV-2 Omicron BF.7 in Beijing after adjustment of the zero-COVID policy in November–December 2022. Nat Med 2023;29(3):579–82.

[28] Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. PLoS One 2013;8(2):e56176.

[29] Zhang T, Wang Q, Leng Z, Yang Y, Yang J, Chen F, et al. A scenario-based evaluation of COVID-19-related essential clinical resource demands in China. Engineering 2021;7(7):948–57.

[30] Pariani E, Amendola A, Piatti A, Anselmi G, Ranghiero A, Bubba L, et al. Ten years (2004-2014) of influenza surveillance in northern Italy. Hum Vaccin Immunother 2015;11(1):198–205.

[31] Wen W, Ma J, Huang L, Wang H, Liu M. Epidemiological analysis of surveillance for influenza in Chaoyang District, Beijing, 2015–2016. Chin J Dis Control Prev. 2017;21(1):8. Chinese.

[32] Zheng C, Kar I, Chen CK, Sau C, Woodson S, Serra A, et al. Multiple sclerosis disease-modifying therapy and the COVID-19 pandemic: implications on the risk of infection and future vaccination. CNS Drugs 2020;34(9):879–96.

[33] Chan TC, Tang JH, Hsieh CY, Chen KJ, Yu TH, Tsai YT. Approaching precision public health by automated syndromic surveillance in communities. PLoS One 2021;16(8):e0254479.

[34] Budd AP, Abd Elal AI, Alabi N, Barnes J, Blanton L, Brammer L, et al. Influenza activity—United States, September 30–December 1, 2018. MMWR Morb Mortal Wkly Rep 2018;67(49):1369–71.

[35] Tsuzuki S, Yoshihara K. The characteristics of influenza-like illness management in Japan. BMC Public Health 2020;20(1):568.

[36] De Lusignan S, Sherlock J, Akinyemi O, Pebody R, Elliot A, Byford R, et al. Household presentation of influenza and acute respiratory illnesses to a primary care sentinel network: retrospective database studies (2013-2018). BMC Public Health 2020;20(1):1748.