Original Paper

Using Baidu Index Data to Improve Chickenpox Surveillance in Yunnan, China: Infodemiology Study

Zhaohan Wang[1]; Jun He[2]; Bolin Jin[1]; Lizhi Zhang[1]; Chenyu Han[1]; Meiqi Wang[1]; Hao Wang[1]; Shuqi An[1]; Meifang Zhao[1]; Qing Zhen[1*]; Shui Tiejun[2*]; Xinyao Zhang[3*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Jilin University, Changchun, China

[2]Yunnan Center for Disease Control and Prevention, Yunnan, China

[3]Department of Social Medicine and Health Management, School of Public Health, Jilin University, Changchun, China

[*]these authors contributed equally

Corresponding Author:
Xinyao Zhang
Department of Social Medicine and Health Management
School of Public Health
Jilin University
1163 Xinmin Street
Changchun, 130021
China
Phone: 86 0431 85619442
Email: 15892560@qq.com

Abstract

Background: Chickenpox is an old but easily neglected infectious disease. Although chickenpox is preventable by vaccines, vaccine breakthroughs often occur, and the chickenpox epidemic is on the rise. Chickenpox is not included in the list of regulated communicable diseases that must be reported and controlled by public and health departments; therefore, it is crucial to rapidly identify and report varicella outbreaks during the early stages. The Baidu index (BDI) can supplement the traditional surveillance system for infectious diseases, such as brucellosis and dengue, in China. The number of reported chickenpox cases and internet search data also showed a similar trend. BDI can be a useful tool to display the outbreak of infectious diseases.

Objective: This study aimed to develop an efficient disease surveillance method that uses BDI to assist in traditional surveillance.

Methods: Chickenpox incidence data (weekly from January 2017 to June 2021) reported by the Yunnan Province Center for Disease Control and Prevention were obtained to evaluate the relationship between the incidence of chickenpox and BDI. We applied a support vector machine regression (SVR) model and a multiple regression prediction model with BDI to predict the incidence of chickenpox. In addition, we used the SVR model to predict the number of chickenpox cases from June 2021 to the first week of April 2022.

Results: The analysis showed that there was a close correlation between the weekly number of newly diagnosed cases and the BDI. In the search terms we collected, the highest Spearman correlation coefficient was 0.747. Most BDI search terms, such as "chickenpox," "chickenpox treatment," "treatment of chickenpox," "chickenpox symptoms," and "chickenpox virus," trend consistently. Some BDI search terms, such as "chickenpox pictures," "symptoms of chickenpox," "chickenpox vaccine," and "is chickenpox vaccine necessary," appeared earlier than the trend of "chickenpox virus." The 2 models were compared, the SVR model performed better in all the applied measurements: fitting effect, R^2=0.9108, root mean square error (RMSE)=96.2995, and mean absolute error (MAE)=73.3988; and prediction effect, R^2=0.548, RMSE=189.1807, and MAE=147.5412. In addition, we applied the SVR model to predict the number of reported cases weekly in Yunnan from June 2021 to April 2022 using the same period of the BDI. The results showed that the fluctuation of the time series from July 2021 to April 2022 was similar to that of the last year and a half with no change in the level of prevention and control.

Conclusions: These findings indicated that the BDI in Yunnan Province can predict the incidence of chickenpox in the same period. Thus, the BDI is a useful tool for monitoring the chickenpox epidemic and for complementing traditional monitoring systems.

(J Med Internet Res 2023;25:e44186) doi: 10.2196/44186

KEYWORDS

Baidu index; chickenpox; support vector machine regression model; disease surveillance; disease; infectious; vaccine; surveillance system; model; prevention; control; monitoring; epidemic

Introduction

Chickenpox is an acute infectious disease caused by the varicella-zoster virus. Children and adolescents are highly susceptible to chickenpox infection, and it is characterized by maculopapular rashes on the skin and mucous membranes with mild systemic symptoms. Chickenpox has a high incidence but low mortality rate, and it is one of the most common childhood diseases [1]. Although most cases have mild clinical symptoms, if not treated promptly, the infection may lead to postherpetic neuralgia, which affects the quality of life and can lead to death in severe cases [2].

Chickenpox is prevalent worldwide, and it has a seasonal pattern. The European region has a single peak pattern from March to May, whereas Asian countries have a double peak pattern from March to May and from December to January [3]. According to the World Health Organization, it is estimated that there are approximately 4.2 million hospitalizations and 4200 deaths due to serious complications of varicella annually worldwide, and the danger of varicella has been seriously underestimated [4]. Despite widespread immunization, varicella continues to spread and develop in many countries, such as the United States, Italy, and various European countries, due to its high transmissibility [3-6]. In China, according to an epidemiological survey of varicella in China from 2005-2019, a total of 6,442,147 cases of varicella were reported nationwide. The reported incidence of varicella increased from 41,211 cases (3.17/100,000) in 2005 to 7,979,482 cases (70.14/100,000) in 2019. The average annual incidence of chickenpox showed an increasing trend each year from 2005 to 2019, and the number of reported cases of varicella increased nearly 22 times from 2005 to 2019. Chickenpox now has one of the highest incidences in terms of preventable diseases in China [7].

Yunnan Province has a high incidence of chickenpox in China, and it ranked among the top 5 provinces in the country in 2019 in terms of the incidence in the population aged under 14 years, accounting for 76.36% of all chickenpox cases. Chickenpox public health emergencies mainly occur in rural elementary schools, and the situation is serious as the chickenpox epidemic continues to grow year by year. However, attention to the management and elimination of chickenpox public health emergencies in rural elementary schools and kindergartens is still insufficient, and the problem of untimely reporting of chickenpox incidence remains. Yunnan Province needs to take effective measures to suppress the spread and prevalence of chickenpox [8].

Currently, problems such as low chickenpox vaccine coverage, the breakthrough of the chickenpox vaccine, the lack of attention to chickenpox disease, imperfect chickenpox surveillance reports, and the lack of public information on chickenpox epidemics are still present in China [9]. In China, chickenpox is reported through the Chinese Disease Prevention and Control Information System. Because chickenpox is a viral disease with mild and self-limiting symptoms, it has not been included in the Chinese National Disease Reporting System for statutory infectious diseases. There is no uniform standard for reporting chickenpox cases across China. However, the public health emergency management information system has focused on some aggregated outbreaks [9,10], but some disseminated cases or subclinical infections may be overlooked.

At present, the means of chickenpox prevention and control in China are relatively singular and limited to the traditional reported incidence monitoring and isolation treatment of patients with chickenpox [2]. The actual incidence of chickenpox in most provinces is not publicly available, and the public cannot obtain timely information on the chickenpox epidemic. Apart from public health departments, the public does not have effective channels to obtain real-time information on local chickenpox incidence, which is not conducive to taking active self-protective measures against chickenpox.

The internet has become an increasingly popular means of accessing health information. The increase in web-based information provides a potentially useful source of data for disease surveillance. Due to its real-time nature and ease of access, internet data can be used to fill the gaps in traditional public health surveillance [11]. Therefore, the possibility of using internet data for the surveillance of various diseases is being increasingly explored as an aid to improve disease prevention and control. The use of internet search data as a complementary means to traditional infectious disease surveillance methods is particularly applicable to neglected diseases that are less affected by consultations [12].

Previous studies [12,13] have indicated that Google Trends can well demonstrate the epidemic characteristics of chickenpox abroad and is suitable for the simulation of features of infectious diseases in epidemic cycles. Moreover, Bakker et al [12] studied chickenpox incidence patterns using Google Trends and found that Google search data show a high correlation (R^2=0.65-0.71) with varicella outbreaks. Thus, Google Trends can be used for chickenpox incidence prediction and early warning. In China, more studies have focused on vaccine efficacy analysis and descriptive studies of chickenpox prevalence [7-10], and most of the exploration of chickenpox surveillance issues has been performed by using actual incidence report data to build models to predict the onset of chickenpox [2], whereas there are few studies exploring whether the Baidu search engine can be applied to chickenpox epidemic surveillance and early warning.

Google Trends is not highly used in China. Google search data do not reflect the true search tendencies of the Chinese public because Baidu is the most widely used search engine product in China. The Baidu index (BDI) [14], which was established based on Baidu search information, should reflect the search needs and awareness of internet users as well as Google Trends abroad in China [15,16]. A considerable number of studies have been conducted using the BDI for China-wide disease prediction. For example, one study has successfully predicted the epidemic

trend of influenza using the BDI [17]. Similarly, some surveillance studies using BDIs for diseases, such as brucellosis; dengue fever; and hand, foot, and mouth disease, have shown that BDIs can be used to reflect disease prevalence [18-20].

The current chickenpox surveillance system in China is flawed and incomplete. The flawed surveillance system and the lack of public information on chickenpox outbreaks have led to a weak response to chickenpox outbreaks in China. Public health events caused by chickenpox epidemics have persisted for a long time, generating a large number of patients with chickenpox, which disrupts the study, work, and life of patients while also affecting the normal production and everyday life of families, schools, and workplaces. Moreover, chickenpox epidemics impose a heavy burden on the healthy socioeconomic development of China.

This study focused on the following specific research questions:

1. What are the correlations between BDI scores and actual varicella incidence data in Yunnan Province?
2. Can internet data be used to predict future varicella disease epidemics?
3. Can big data be used as a supplement to traditional surveillance systems for early warning surveillance of infectious diseases as well as epidemics?

Methods

Real-World Databases

The data were divided into 2 parts as follows: the number of reported chickenpox cases in Yunnan Province from 2017 to 2021 and the chickenpox-related BDI search data in Yunnan Province from 2017 to April 2022. The chickenpox monitoring data in Yunnan Province were obtained from the chickenpox epidemic information of the Yunnan Province Center for Disease Control and Prevention.

BDI Databases

The chickenpox-related BDI search data were obtained from the official website of the BDI [14]. The BDI of "PC + mobile" from 2017 to April 2022 was collated as search data.

The principles of selecting keywords were as follows: (1) keywords must be closely related and specific to chickenpox; (2) selected longtail keywords were considered for inclusion; (3) keywords must have sufficient search volume in the mining module; and (4) the time series of each keyword must be complete and valid.

The following 12 keywords were selected: "chickenpox," "chickenpox pictures," "chickenpox symptoms," "chickenpox diet," "chickenpox infection period," "symptoms of chickenpox," "symptoms and treatment of chickenpox," "chickenpox vaccine," "is chickenpox vaccine necessary," "shingles," "chickenpox treatment," and "treatment of chickenpox."

Analysis

Keyword Screening and Relevance Test

First, Pearson and Spearman correlation tests were performed on the first 12 selected BDIs and the number of chickenpox cases, with Pearson correlation coefficient>0.3 as the criteria for determining the correlation. A time-lagged cross-correlation test was then performed on the screened keyword indices and the number of chickenpox cases to determine the time type of keywords (prior, simultaneous, and lagged).

Support Vector Machine Regression Model Construction

For variable selection, the number of chickenpox cases in Yunnan Province from January 2017 to May 2021 was used as the dependent variable, and the weekly BDIs of 8 keywords, such as "chickenpox," were used as the independent variables. The BDIs and incidence data of 209 weeks from 2017 to 2020 were used as the training set, and the data of 22 weeks from January to May 2021 were used as the test set. The data were fitted with the actual incidence data to analyze the accuracy.

Multiple Linear Regression Model Construction

The general expression was as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

where β_0, β_1, β_2, ..., β_k are the parameters of the model, and ε is the error term.

The variable selection was consistent with the support vector machine regression (SVR) model. For model construction, the stepwise regression method was used to eliminate the independent variables that had no significant effect (P>.05) on the dependent variable, and the model with the highest R^2 value of goodness-of-fit was selected for the next step of prediction according to the stepwise regression model construction results.

Determination Index of the Optimal Model

The optimal model for the present study was selected by comparing the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean square error (RMSE) of the predicted and actual values of the above 2 models in fitting and predicting data.

Data Prediction of the Chickenpox Epidemic in Yunnan Province

The optimal prediction model was selected by the above method. The BDI of the "chickenpox" keyword and the number of reported chickenpox cases in Yunnan Province for 231 weeks from January 2017 to May 2021 were imported as independent and dependent variables to establish the prediction model. The BDI data from June 2021 to the first week of April 2022 were obtained to predict the trend of chickenpox incidence in Yunnan Province.

Ethical Considerations

The data sources for this study included publicly available BDI search data and varicella incidence data provided by the Yunnan Province Center for Disease Control and Prevention. As this study did not involve human or animal experimental subjects, nor did it involve any ethical issues related to data collection

or use, ethical approval or a license was not required for this study.

Results

Cross-Correlation Analysis of Chickenpox and the BDI in Yunnan Province

After the cross-correlation test, the results showed that most BDI search terms, such as "chickenpox," "chickenpox treatment," "treatment of chickenpox," "chickenpox symptoms," and "chickenpox virus," trend consistently. Some BDI search terms, such as "chickenpox pictures," "symptoms of chickenpox," "chickenpox vaccine," and "is chickenpox vaccine necessary" appeared earlier than the trend of "chickenpox virus." In all, 8 keywords were screened from the 12 chickenpox-related keywords as variables required for modeling: "chickenpox picture," "chickenpox," "chickenpox symptoms," "symptoms of chickenpox," "symptoms and treatment of chickenpox," "chickenpox vaccine," "treatment of chickenpox," and "chickenpox treatment" (Table 1).

Table 1. Cross-correlation analysis of actual chickenpox cases and internet search terms from Yunnan, China[a].

Search terms	Lag (weeks)						
	–3	–2	–1	0	1	2	3
水痘 图 片 (chickenpox picture)							
r[b]	0.702	*0.730*[c]	0.715	0.705	0.636	0.555	0.462
P value[d]	<.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘 (chickenpox)							
r	0.656	0.718	0.727	*0.747*[c]	0.741	0.678	0.609
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘症状 (chickenpox symptoms)							
r	0.364	0.37	0.371	*0.372*[c]	0.331	0.287	0.237
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘的症状 (symptoms of chickenpox)							
r	0.550	*0.578*[c]	0.548	0.406	0.472	0.390	0.298
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘的症状和治 疗 (symptoms and treatment of chickenpox)							
r	0.302	0.351	0.379	*0.556*[c]	0.385	0.345	0.320
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘疫苗 (chickenpox vaccine)							
r	*0.353*[c]	0.345	0.317	0.317	0.282	0.215	0.150
P value	<.001	.001	<.001	<.001	<.001	.001	.02
水痘的治 疗 (treatment of chickenpox)							
r	0.259	0.261	0.308	*0.323*[c]	0.264	0.267	0.199
P value	<.001	<.001	<.001	<.001	<.001	<.001	.002
水痘治 疗 (chickenpox treatment)							
r	0.216	0.272	0.294	*0.325*[c]	0.293	0.283	0.278
P value	.001	<.001	<.001	<.001	<.001	<.001	<.001
水痘 饮 食 (chickenpox diet)							
r	0.194	0.196	0.187	*0.222*[c]	0.157	0.180	0.170
P value	.003	.003	.004	.001	.02	.006	.01
水痘 传 染期 (chickenpox transmission period)							
r	0.165	0.200	0.219	*0.280*[c]	0.275	0.292	0.295
P value	.01	.002	.001	<.001	<.001	<.001	<.001
水痘疫苗有必要打 吗 (is chickenpox vaccine necessary)							
r	0.264	*0.200*[c]	0.144	0.090	–0.001	–0.078	–0.161
P value	<.001	.002	.03	.17	.99	.24	.02
带 状疱疹 (herpes zoster)							
r	0.024	*0.022*[c]	0.008	0.011	0.000	–0.010	–0.017
P value	.72	.74	.91	.87	>.99	.88	.79

[a]Spearman correlation coefficient>0.3 was used as the criteria for inclusion.

[b]*r* values represented cross correlation coefficient.

[c]Italicized values showed the maximum of cross correlation coefficient.

[d]*P* values represented statistical significance between 2 variables.

Time Series Characteristics and Correlation Analysis of Varicella and BDI in Yunnan Province

The number of reported chickenpox cases in Yunnan Province showed an increasing trend year by year, especially during the winter chickenpox susceptibility period in 2020, with a peak higher than the general level in previous years. The trend of chickenpox incidence in 2017-2021 showed seasonality with double peaks in the number of cases from May to July and from November to January in each year. The BDI data for the same period also showed bimodal peaks from May to July and from November to January in each year. Comparison charts of the trend of 8 BDI keywords and actual occurrence are shown in Multimedia Appendix 1.

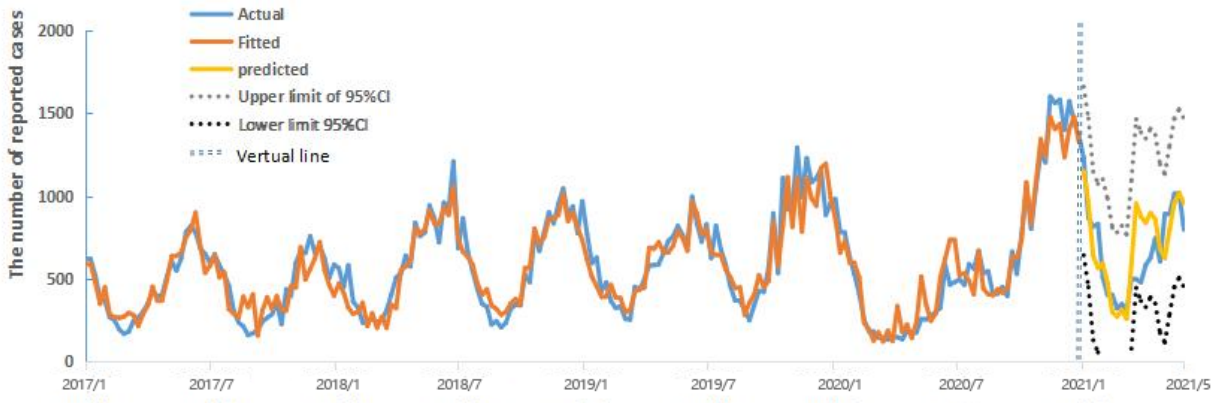
Construction and Prediction of Chickenpox Case Prediction Model

Establishment of the SVR Prediction Model

In the SVR model, the radial basis kernel function was used. We chose all 8 key terms for modeling: "chickenpox picture," "chickenpox," "chickenpox symptoms," "symptoms of chickenpox," "symptoms and treatment of chickenpox," "chickenpox vaccine," "treatment of chickenpox," and "chickenpox treatment." Using the grid search method for the hyperparameter search, the final obtained parameters were C=1 and γ=0.1.

The complete fitting and prediction effect graph is shown in Figure 1.

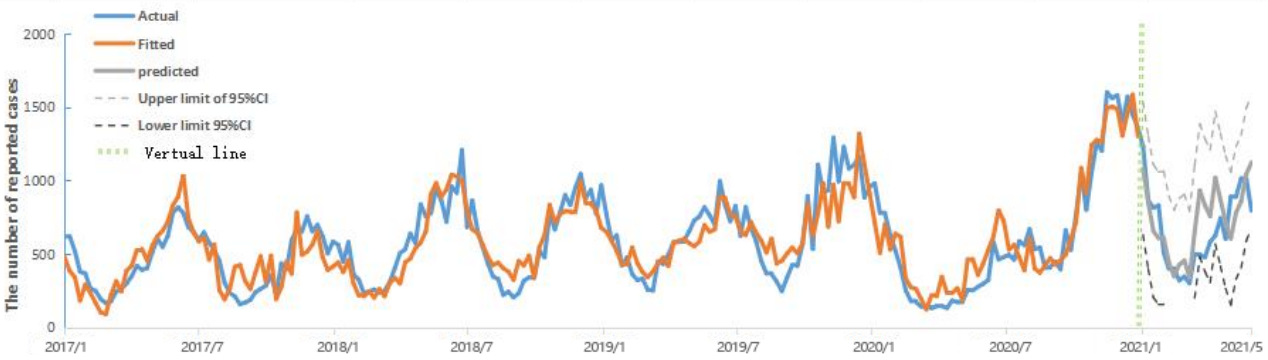
Figure 1. Predicted effect of support vector machine regression fitting for chickenpox in Yunnan Province (January 2017 to May 2021).



Multiple Linear Regression Prediction Modeling

In the multiple linear regression model, we also chose all 8 key terms for modeling. We used the stepwise regression method to fit and test the data. The model performed best when those 4 key terms were included, namely "chickenpox," "chickenpox symptoms," "symptoms of chickenpox," and "treatment of chickenpox." The detailed regression coefficients are as follows: b_1=0.243, b_2=0.495, b_3=0.381, and b_4=0.278. The complete fitting and prediction effect graph is shown in Figure 2.

Figure 2. Predicted effect of multiple linear regression fitting for chickenpox in Yunnan Province (January 2017 to May 2018).



Comparison of Multiple Regression Model and SVR Model Prediction Performance

To evaluate the fitting performance of the model, we used R^2, MAE, and RMSE to compare the advantages and disadvantages of the 2 models. The results showed that the SVR model had better performance in both fitting and prediction effects compared to the multiple linear regression model (Tables 2 and 3).

Therefore, SVR was selected as the best model for case number prediction.

Table 2. Comparison of model fitting effect indicators.

	MLR[a] model	SVR[b] model
R^2	0.833	0.911
RMSE[c]	130.3389	96.2995
MAE[d]	106.6526	73.3988

[a]MLR: multiple linear regression.
[b]SVR: support vector machine regression.
[c]RMSE: root mean square error.
[d]MAE: mean absolute error.

Table 3. Comparison of model prediction effect indicators.

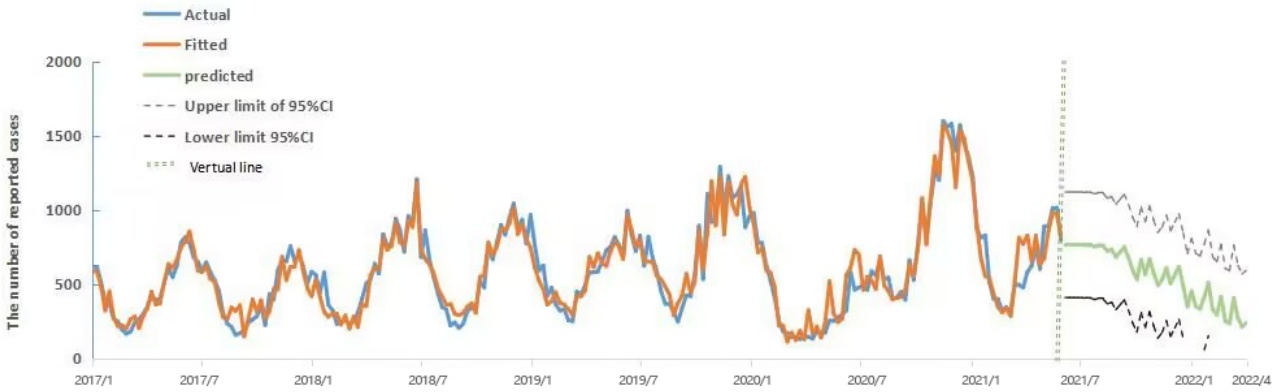
	MLR[a] model	SVR[b] model
R^2	0.459	0.548
RMSE[c]	204.2203	189.1807
MAE[d]	166.2412	147.5412
MAPE[e]	15%	9.1%

[a]MLR: multiple linear regression.
[b]SVR: support vector machine regression.
[c]RMSE: root mean square error.
[d]MAE: mean absolute error.
[e]MAPE: mean absolute percentage error.

SVR–Based Prediction of the Number of Chickenpox Cases

SVR prediction was performed using the BDI with the number of cases from January 2018 to May 2021 to predict the likely number of chickenpox cases from June 2021 to the first week of April 2022. The predicted trends are shown in Figure 3 and Multimedia Appendix 2.

Figure 3. Fitting effect and prediction results of chickenpox in Yunnan Province based on the support vector machine regression model.



Discussion

Principal Findings

In this study, we used the chickenpox BDI to build a model to predict the incidence of chickenpox during the same period. We concluded that it is feasible to use BDI data for varicella epidemic surveillance, which does not rely on actual reported data of varicella incidence.

The Principle of Using the BDI to Predict Incidence

The BDI reflects changes in public interest in various events over a period of time. As a relatively neglected disease, news reports about chickenpox are substantially fewer than those about other infectious diseases, such as influenza and AIDS. Possible reasons for public interest in chickenpox are news reports or illnesses in their own family, suggesting that the main source of web-based search data may come from patients with chickenpox and their family members. This hypothesis is the

reason why we believe that the chickenpox BDI can reflect the dynamics of chickenpox incidence through search information.

The use of internet data for disease surveillance has several advantages. For instance, it is more cost-effective and less time-consuming compared to traditional surveillance methods that rely on patient consultations and medical records [21,22]. Moreover, internet data can capture disease trends earlier than traditional surveillance methods because they are updated in real time. In addition, they can detect disease outbreaks that may go unnoticed using traditional surveillance methods. As neglected diseases often go unnoticed due to the limited health care resources and infrastructure in low-income countries, internet data can serve as a useful tool for monitoring these diseases. By leveraging internet data, health officials can detect disease outbreaks earlier, allocate resources more effectively, and ultimately prevent the spread of diseases more efficiently. Therefore, the use of internet data for disease surveillance has the potential to revolutionize disease prevention and control efforts.

Based on this principle, we believe that the BDI can be used as a new infectious disease surveillance tool to complement traditional public health surveillance systems. By combining big data epidemiological surveillance methods with traditional epidemiological surveillance tools, governments can establish a more sensitive surveillance tool. Real-time and low-cost internet data can help government health departments quickly identify diseases, populations, and areas with potential transmission risks and take effective measures.

Comparison With Prior Work

This study explores the potential of using internet data for disease surveillance, with a focus on chickenpox. Although previous research has shown success in using Google for disease prediction [21-23], we wanted to investigate whether Baidu, a search engine with higher usage rates in China, could also be used for this purpose. Our findings confirm that Baidu can be used for disease monitoring, providing an opportunity to better leverage internet data to predict disease occurrence and its spread. This discovery is particularly important for public health institutions, which can now make more accurate predictions and take more targeted preventive measures. Furthermore, this study highlights the growing importance of internet information as a source of data for disease surveillance in the future.

In this study, the chickenpox BDI was highly consistent with the actual occurrence trend of chickenpox (r=0.747), and compared to the disease studies conducted using the BDI [15,17-20,24-27], the correlation coefficients of these studies were generally between r=0.3 and r=0.93, indicating that the chickenpox BDI well reflects the actual trend of chickenpox occurrence. Thus, it is feasible to use the BDI for chickenpox epidemic surveillance. At the same time, studies exploring the correlation between Google Trends and chickenpox [12] have shown that the average correlation between chickenpox and Google Trends is approximately r=0.762 globally, indicating that the use of the BDI for chickenpox in China has similar surveillance effects as those obtained using Google Trends abroad. Therefore, in China, the methods and ideas of using Google for varicella incidence trend prediction in foreign countries can be applied to the BDI to conduct supplementary surveillance of chickenpox incidence.

In this study, the SVR model outperformed the multiple linear regression model both in terms of fitting and prediction (Tables 2 and 3), which showed a mean absolute percentage error of 9.1%, RMSE of 189, and R^2 of 0.548. Comparing it with previous studies [19,28], the SVR model outperformed the autoregressive model for sexually transmitted diseases and the autoregressive integrated moving average model for brucellosis in terms of mean absolute percentage error and RMSE. Overall, the SVR model showed promising results for predicting chickenpox incidence and can be further improved and compared with other models for infectious diseases in future studies.

The SVR model developed using the BDI accurately predicted the actual number of current varicella cases. Unlike previous chickenpox prediction models, the method developed in this study provided a rapid assessment of the current chickenpox epidemic without relying on actual incidence reporting data. The method allows for immediate calculation of chickenpox case prediction, which is more rapid than traditional prediction systems, allowing it to be used as a rapid method to help the public know the current chickenpox dynamics. Moreover, due to the simplicity of the method and easy access to data sources, it is likely to be applicable to most places in China. When model predictions increase, it may indicate a rise in the number of chickenpox cases, allowing for disease control and prevention–related authorities to prepare for potential chickenpox outbreaks.

Limitations

This study had several limitations that deserve further discussion. First, China is a vast country with regional differences in customs and culture as well as population distribution and geographic ethnicity [29]. Thus, it is not possible to predict the incidence trend in all provinces using one model. If data can be collected for each region, the prediction of chickenpox incidence trends can be tailored to different regions. Second, similar to other studies exploring the relationship between disease and the internet [23], when the media reports the chickenpox epidemic, people without chickenpox may search for chickenpox out of curiosity and fear, which would lead to a surge in the number of searches. Thus, further research is needed to eliminate the influence of the media on the results. Finally, multiple regression modeling based on BDI data alone does not allow for accurate prediction of chickenpox dynamics, and the volume of disease search data does not correspond one to one with the number of reported cases. Infectious disease prediction models based solely on internet search data may also be confounded by searchers' knowledge of the disease and local language restrictions.

Therefore, in future studies, we should also consider additional influencing factors related to chickenpox outbreaks, such as climate, economy, and vaccination status, for a comprehensive analysis to achieve accurate prediction of chickenpox incidence.

Conclusions

Based on the results, it is feasible to apply the BDI method to reflect the incidence of varicella. The fitted and predicted values of the SVR model were consistent with the actual incidence trend of chickenpox, indicating that the model based on the BDI can be used to reflect the actual local incidence trend of chickenpox in real time. Thus, the BDI can be used for disease surveillance. Internet search data can be used as a supplement to traditional surveillance systems to help with the early detection of potential disease outbreaks or disease epidemics.

Acknowledgments

We thank the Yunnan Province Center for Disease Control and Prevention for providing available data on chickenpox incidence.

Data Availability

The data sets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

ZW, QZ, and ST designed the main part of work. ZW, BJ, and LZ performed data collection and wrote the paper. ST and QZ edited and promoted the manuscript. We would like to acknowledge the efforts of CH, MW, SA, and MZ who helped complete the work of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparison charts of trends of 8 additional Baidu index keywords and actual occurrences.
[DOCX File , 937 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Support vector machine regression model prediction results.
[DOCX File , 15 KB-Multimedia Appendix 2]

References

1. Somekh E, Dalal I, Shohat T, Ginsberg GM, Romano O. The burden of uncomplicated cases of chickenpox in Israel. J Infect 2002 Nov;45(4):233-236. [doi: 10.1053/jinf.2002.1039] [Medline: 12423610]
2. Pang FR, Luo QH, Hong XQ, Wu B, Zhou JH, Zha WT, et al. The study on the early warning period of varicella outbreaks based on logistic differential equation model. Epidemiol Infect 2019 Jan;147:e70 [FREE Full text] [doi: 10.1017/S0950268818002868] [Medline: 30868977]
3. Freer G, Pistello M. Varicella-zoster virus infection: natural history, clinical manifestations, immunity and current and future vaccination strategies. New Microbiol 2018 Apr;41(2):95-105 [FREE Full text] [Medline: 29498740]
4. World Health Organization. Varicella and herpes zoster vaccines: WHO position paper, June 2014. Wkly Epidemiol Rec 2014 Jun 20;89(25):265-287 [FREE Full text] [Medline: 24983077]
5. Wutzler P, Bonanni P, Burgess M, Gershon A, Sáfadi MA, Casabona G. Varicella vaccination - the global experience. Expert Rev Vaccines 2017 Aug;16(8):833-843 [FREE Full text] [doi: 10.1080/14760584.2017.1343669] [Medline: 28644696]
6. Desai R, Lopman BA, Shimshoni Y, Harris JP, Patel MM, Parashar UD. Use of internet search data to monitor impact of rotavirus vaccination in the United States. Clin Infect Dis 2012 May;54(9):e115-e118. [doi: 10.1093/cid/cis121] [Medline: 22423140]
7. Sun X, Zhu Y, Sun H, Xu Y, Zhang L, Wang Z. Comparison of varicella outbreaks in schools in China during different vaccination periods. Hum Vaccin Immunother 2022 Nov 30;18(6):2114255 [FREE Full text] [doi: 10.1080/21645515.2022.2114255] [Medline: 35993917]
8. Liu H, Zhang H, Zhang M, Changzeng F, Cong S, Xu D, et al. Epidemiological and etiological characteristics of viral meningitis for hospitalized pediatric patients in Yunnan, China. Medicine (Baltimore) 2022 Jul 01;101(26):e29772 [FREE Full text] [doi: 10.1097/MD.0000000000029772] [Medline: 35777023]
9. Pan X, Shu M, Ma R, Fang T, Dong H, Sun Y, et al. Varicella breakthrough infection and effectiveness of 2-dose varicella vaccine in China. Vaccine 2018 Sep 05;36(37):5665-5670. [doi: 10.1016/j.vaccine.2018.05.025] [Medline: 30104113]
10. Zhang Z, Suo L, Pan J, Zhao D, Lu L. Two-dose varicella vaccine effectiveness in China: a meta-analysis and evidence quality assessment. BMC Infect Dis 2021 Jun 09;21(1):543 [FREE Full text] [doi: 10.1186/s12879-021-06217-1] [Medline: 34107891]

11. Ayers JW, Westmaas JL, Leas EC, Benton A, Chen Y, Dredze M, et al. Leveraging big data to improve health awareness campaigns: a novel evaluation of the Great American Smokeout. JMIR Public Health Surveill 2016 Mar 31;2(1):e16 [FREE Full text] [doi: 10.2196/publichealth.5304] [Medline: 27227151]
12. Bakker KM, Martinez-Bakker ME, Helm B, Stevenson TJ. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. Proc Natl Acad Sci U S A 2016 Jun 14;113(24):6689-6694 [FREE Full text] [doi: 10.1073/pnas.1523941113] [Medline: 27247405]
13. Berlinberg EJ, Deiner MS, Porco TC, Acharya NR. Monitoring interest in herpes zoster vaccination: analysis of Google search data. JMIR Public Health Surveill 2018 May 02;4(2):e10180 [FREE Full text] [doi: 10.2196/10180] [Medline: 29720364]
14. Baidu index. URL: https://index.baidu.com/v2/index.html#/ [accessed 2023-05-10]
15. Liu K, Wang T, Yang Z, Huang X, Milinovich GJ, Lu Y, et al. Using Baidu search index to predict dengue outbreak in China. Sci Rep 2016 Dec 01;6:38040 [FREE Full text] [doi: 10.1038/srep38040] [Medline: 27905501]
16. Newton MA, Wang Z. Multiset statistics for gene set analysis. Annu Rev Stat Appl 2015 Apr;2:95-111 [FREE Full text] [doi: 10.1146/annurev-statistics-010814-020335] [Medline: 25914887]
17. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from Baidu. PLoS One 2013 May 30;8(5):e64323 [FREE Full text] [doi: 10.1371/journal.pone.0064323] [Medline: 23750192]
18. Liu B, Wang Z, Qi X, Zhang X, Chen H. Assessing cyber-user awareness of an emerging infectious disease: evidence from human infections with avian influenza A H7N9 in Zhejiang, China. Int J Infect Dis 2015 Nov;40:34-36 [FREE Full text] [doi: 10.1016/j.ijid.2015.09.017] [Medline: 26432409]
19. Zhao C, Yang Y, Wu S, Wu W, Xue H, An K, et al. Search trends and prediction of human brucellosis using Baidu index data from 2011 to 2018 in China. Sci Rep 2020 Apr 03;10(1):5896 [FREE Full text] [doi: 10.1038/s41598-020-62517-7] [Medline: 32246053]
20. 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 Aug 13;18(1):398 [FREE Full text] [doi: 10.1186/s12879-018-3285-4] [Medline: 30103690]
21. Yuan K, Huang G, Wang L, Wang T, Liu W, Jiang H, et al. Predicting norovirus in the United States using Google trends: infodemiology study. J Med Internet Res 2021 Sep 29;23(9):e24554 [FREE Full text] [doi: 10.2196/24554] [Medline: 34586079]
22. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
23. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
24. Fung ICH, Fu KW, Ying Y, Schaible B, Hao Y, Chan CH, et al. Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks. Infect Dis Poverty 2013 Dec 20;2(1):31 [FREE Full text] [doi: 10.1186/2049-9957-2-31] [Medline: 24359669]
25. Jena AB, Karaca-Mandic P, Weaver L, Seabury SA. Predicting new diagnoses of HIV infection using internet search engine data. Clin Infect Dis 2013 May;56(9):1352-1353. [doi: 10.1093/cid/cit022] [Medline: 23334812]
26. 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 Oct 22;6(4):e23098 [FREE Full text] [doi: 10.2196/23098] [Medline: 32960177]
27. Fang J, Zhang X, Tong Y, Xia Y, Liu H, Wu K. Baidu index and COVID-19 epidemic forecast: evidence From China. Front Public Health 2021 May 5;9:685141 [FREE Full text] [doi: 10.3389/fpubh.2021.685141] [Medline: 34026721]
28. Huang R, Luo G, Duan Q, Zhang L, Zhang Q, Tang W, et al. Using Baidu search index to monitor and predict newly diagnosed cases of HIV/AIDS, syphilis and gonorrhea in China: estimates from a vector autoregressive (VAR) model. BMJ Open 2020 Mar 24;10(3):e036098 [FREE Full text] [doi: 10.1136/bmjopen-2019-036098] [Medline: 32209633]
29. Nan Y, Gao Y. A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS One 2018 Jul 11;13(7):e0199697 [FREE Full text] [doi: 10.1371/journal.pone.0199697] [Medline: 29995920]

Abbreviations

BDI: Baidu index
MAE: mean absolute error
RMSE: root mean square error
SVR: support vector machine regression model

Edited by A Mavragani; submitted 10.11.22; peer-reviewed by R Guo, Y Zheng, A Allam; comments to author 06.03.23; revised version received 21.03.23; accepted 04.05.23; published 16.05.23

Please cite as:
Wang Z, He J, Jin B, Zhang L, Han C, Wang M, Wang H, An S, Zhao M, Zhen Q, Tiejun S, Zhang X
Using Baidu Index Data to Improve Chickenpox Surveillance in Yunnan, China: Infodemiology Study
J Med Internet Res 2023;25:e44186
URL: https://www.jmir.org/2023/1/e44186
doi: 10.2196/44186
PMID: