






BMJ Open 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

Ruonan Huang ¹, Ganfeng Luo,² Qibin Duan,^{3,4} Lei Zhang ^{5,6,7,8}, Qingpeng Zhang,⁹ Weiming Tang ^{10,11}, M. Kumi Smith ¹², Jinghua Li,^{1,13} Huachun Zou ^{2,3}

To cite: Huang R, Luo G, Duan Q, *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;**10**:e036098. doi:10.1136/bmjopen-2019-036098

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-036098>).

JL and HZ contributed equally.

RH, GL and QD are joint first authors.

Received 29 November 2019
Revised 04 March 2020
Accepted 04 March 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Professor Huachun Zou;
zouhuachun@mail.sysu.edu.cn

ABSTRACT

Objectives Internet search engine data have been widely used to monitor and predict infectious diseases. Existing studies have found correlations between search data and HIV/AIDS epidemics. We aimed to extend the literature through exploring the feasibility of using search data to monitor and predict the number of newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea in China.

Methods This paper used vector autoregressive model to combine the number of newly diagnosed cases with Baidu search index to predict monthly newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea in China. The procedures included: (1) keywords selection and filtering; (2) construction of composite search index; (3) modelling with training data from January 2011 to October 2016 and calculating the prediction performance with validation data from November 2016 to October 2017.

Results The analysis showed that there was a close correlation between the monthly number of newly diagnosed cases and the composite search index (the Spearman's rank correlation coefficients were 0.777 for HIV/AIDS, 0.590 for syphilis and 0.633 for gonorrhoea, $p < 0.05$ for all). The R^2 were all more than 85% and the mean absolute percentage errors were less than 11%, showing the good fitting effect and prediction performance of vector autoregressive model in this field.

Conclusions Our study indicated the potential feasibility of using Baidu search data to monitor and predict the number of newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea in China.

INTRODUCTION

HIV/AIDS, syphilis and gonorrhoea are prevalent sexually transmitted diseases (STDs) in China, claiming to have heavy disease burdens. According to the national reports on notifiable infectious diseases in China (available at <http://www.chinacdc.cn/>), there were 134 512 newly diagnosed cases of HIV/AIDS, 475 860 cases of syphilis and 138 855 cases of gonorrhoea totally in 2017. These three STDs

Strengths and limitations of this study

- The vector autoregressive model can potentially predict the number of newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea in relatively longer future, by integrating internet search data with routine surveillance data.
- The applied exclusion criteria for each disease/infection's keywords set may need to be adjusted according to different characteristics in the future.
- The algorithm of Baidu search index on the official website is changing irregularly, thus the corresponding coefficients of the model need to be recalculated accordingly.

have significantly affected people's quality of life and caused immense economic burdens on the healthcare system.^{1 2} These STDs are also listed as category B notifiable infectious diseases in the *Law of the People's Republic of China on the Prevention and Treatment of Infectious Diseases*³, which are subject to strict monitoring and management by the health authorities.

Since the outbreak of severe acute respiratory syndrome in 2003,⁴ China's national infection diseases surveillance system has made substantial improvement. The data are currently being collected based on a timelier and geographically specific basis (formerly at city level, now at county/district level). Category B notifiable infectious diseases are required to be reported within 24 hours⁵; however there is a delay in the release of official statistics, that is, 1–3 weeks for syphilis and gonorrhoea, 2 months for HIV/AIDS. The use of search data can potentially overcome publication delays and help health facilities such as medical institutions and community

health organisations to make preemptive and preventative treatment plans. In addition, due to the secular stigma and discrimination against STDs, people at risk of infection tend to search for testing services, while infected people tend to seek treatment and care services on the internet as an initial step,⁶ instead of attending health-care facilities. This makes it difficult for health authorities to monitor true HIV/STDs epidemics. There is a possibility that more people's information can be covered like undiagnosed people as a supplement by following the tendency of relevant statistics based on the internet.

In the context of big data, attempts have been made to use internet-based information to provide early predictions of diseases, such as Google,⁷ Twitter⁸ and Baidu.⁹ Compared with traditional surveillance systems, this novel approach is more economical, simpler and timely.¹⁰ It has been proven that various infectious diseases, including influenza,^{11–14} dengue^{15–18} and H7N9,¹⁹ can be effectively monitored and predicted by models using internet data. The sensitivity and specificity of search engine data-based monitoring or predicted results is on par with that of the traditional surveillance system.^{11–20} It is invaluable to extend this innovative method to other diseases, including HIV/AIDS, syphilis and gonorrhoea that lead to significant morbidity, mortality, clinical outcomes, and with no existing vaccines.²¹

Google search data have already been applied in the monitoring and predicting of new diagnoses of HIV infection in USA at state level.^{22–23} In China there are also articles using Baidu search data to predict the incidence of HIV/AIDS.^{24–26} These studies made an inspiration for adopting more attempts, such as multivariate time series analysis, to explore more effective predictions. Baidu holds the highest search engine market penetration rate in China (93.1% in December 2015) with 94.6% of the total 566 million search engine users searching for information while using it.²⁷ It is currently the most representative tool for measuring users' behaviours in the country.

The vector autoregressive (VAR) model is one of the most flexible and comprehensible models for analysing multivariate time series data. It could involve more than one variable and explain past and causal relationships among multiple variables over time, as well as predict future observations.²⁸ The structure is that each variable is a linear function of its own past lags and the past lags of the other variables. The incidence of diseases/infections and search data were usually autocorrelated itself and cross-correlated with the other. By evaluating the time lags, the final VAR model would include real incidence data and internet search data which have the potential to make the predictions more stable. However, there has been no studies using VAR model to predict the incidence of HIV/AIDS, syphilis and gonorrhoea. We aimed to use Baidu search data in combination with previously diagnosed cases through a VAR model to monitor and predict newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea, and compare the different incidence-search patterns.

METHODS

Data sources

The number of newly diagnosed cases

Aggregated data on the number of newly diagnosed cases of HIV/AIDS, syphilis and gonorrhoea from January 2011 to October 2017 (online supplementary appendix 1, table S1) were retrieved from the monthly academic journal *Chinese Journal of AIDS&STD* and the official website of Chinese Center for Diseases Control and Prevention (<http://www.chinacdc.cn/>). The journal updates HIV/AIDS epidemics on a monthly basis with a time-lag of 2 months. Monthly syphilis and gonorrhoea epidemic statistics, from the Chinese Center for Diseases Control and Prevention, are posted on its official website every month for a delay of 1–3 weeks.

Search volume data from Baidu

The search volume data were obtained from the website of 'Baidu Index' (<http://index.baidu.com/>), which shows the search volume of Baidu's search engine using specific keywords at different time periods (the lowest level of each day) and regions (lowest at city level). In order to be consistent with the number of newly diagnosed cases, we selected the national average Baidu index data of each month (average of the daily total search index) for specific keywords from January 2011 to October 2017.

Keywords selection and filtering

The search volume of different keywords could vary to a large extent, thus we used the keywords selection and filtering methods based on a study on influenza epidemic monitoring.¹² This study applied inclusion criteria to exclude irrelevant keywords and performed a correlation analysis between the number of newly diagnosed cases and search volume, instead of including all disease names and symptom-relevant keywords.

The original keywords set derived from the internet (<http://tool.chinaz.com/>), where researchers could explore relevant keywords by typing disease names. The related keywords recommendations on the website include suggestions from Baidu and other resources, such as blogs and online reports using semantic correlation analysis. Using the Chinese equivalents of 'HIV/AIDS', 'Syphilis' and 'Gonorrhoea', we obtained 185 keywords in total from this website. In the next step, we used the following criteria to exclude keywords that were not closely related to the number of newly diagnosed cases of these diseases.

Exclusion criteria:

1. Words irrelevant to the epidemic information of AIDS, syphilis and gonorrhoea, such as 'The World AIDS Day' or 'The Population of AIDS patients in China' and so on.
2. Words with an interrupted time series representing Baidu search index, as a result of fake news or news irrelevant to diseases' epidemic information, that is, anecdotes about celebrities or media hype.

3. Words whose Baidu search index's Spearman's rank correlation coefficients with the monthly reported incident case counts was less than 0.4 (lower bound for moderate correlation, the guide for describing the strength of the correlation was shown in online supplementary appendix 1, table S2).

SEARCH INDEX COMPOSITION

After filtering, the remaining keywords (see online supplementary appendix 1, table S3) were used to build a search index composition for each HIV/STD. Spearman's rank correlation coefficient between monthly composite search index and the number of newly diagnosed cases was used to test the correlation. In the index composition formula, weights of each keyword were defined by the strength of the correlation coefficient. The detailed calculations are as follows:

$$Weight_i = \frac{\rho_i}{\sum_{i=1}^n \rho_i}$$

$$CompositeSearchIndex = \sum_{i=1}^n Weight_i * Keyword_i$$

where ρ_i represents the Spearman's rank correlation coefficient of the i^{th} word, $Keyword_i$ and $Weight_i$ represent the Baidu search index of i^{th} word and the weight of it, n is the number of the final selected keywords.

Model fitting and validation

The modelling procedure included: (1) testing the stationarity of a single regression variable; (2) performing a cointegration test if the data were unstable; (3) selecting a lag length; (4) establishing VAR models; (5) testing the residual autocorrelation; (6) assessing the stability of VAR models and (7) making the predictions.

Data set and data preprocessing

Once the data were ready for modelling, the entire data set was split into training data from January 2011 to October 2016 and validation data from November 2016 to October 2017. The ratio of training data versus validation data was approximately 6:1.

Testing stationarity and cointegration

Before estimating the model, we tested the stationarity characteristics of each variable, otherwise, the model's statistics, such as mean and correlations, would not be able to accurately describe the time series signal.²⁸ The augmented Dickey-Fuller (ADF) t-statistic value was used to check the stationarity and if the data were not stationary, the series would be differenced and the ADF test applied again on the differenced value.

After decomposing the number of newly diagnosed cases (NDC) and composite search index data, there was an uptrend and periodic character of both variables and heteroscedasticity of composite search index data. So the stationarity test was applied to the first-order differential NDC and \ln (composite search index) time series data. The result was shown in table 1. Then, the Engle-Granger method was used to test for co-integration because the NDC and composite search index data were all nonstationary. Testing for cointegration identified stable, long-run relationships between sets of variables. The Engle-Granger test was a two-step residual-based testing procedure on regression techniques which first estimating the long-run equation and then the error correction model. The result showed that these two variables were cointegrated.

Selecting lag length and building the VAR model

Selection of appropriate lag length was critical to inferring in VARs and it could be determined using many criteria. Here we used Akaike Information Criterion to choose the lag length and the final structure was described as follows:

$$NDC_t = \sum \alpha_k NDC_{t-k} + \sum \beta_k \ln BSI_{t-k} + \varphi t + cons + \sum_{d=1}^{11} season_d + \varepsilon_t$$

where NDC_t represents the number of newly diagnosed cases in t^{th} month, NDC_{t-k} represents the number of newly diagnosed cases in $(t-k)^{th}$ month and $\ln BSI_{t-k}$ represents the \ln (composite search index) in $(t-k)^{th}$ month. t represents the trend term, $cons$ represents the intercept and $season_d$ represents the mean value in d^{th} month. α_k , β_k and φ denote the coefficients. ε_t represents the error

Table 1 Stationary test results

	Regression variable	ADF test statistic	Critical values			Result
			1%	5%	10%	
HIV/AIDS	dfNDC	-11.1023	-4.04	-3.45	-3.15	Stationary
	dflogBSI	-9.6116	-4.04	-3.45	-3.15	Stationary
Syphilis	dfNDC	-7.1526	-4.04	-3.45	-3.15	Stationary
	dflogBSI	-7.4332	-4.04	-3.45	-3.15	Stationary
Gonorrhoea	dfNDC	-5.7524	-4.04	-3.45	-3.15	Stationary
	dflogBSI	-7.1976	-4.04	-3.45	-3.15	Stationary

*dfNDC and dflogBSI represent the first-order differential variables (monthly number of newly diagnosed cases and \ln (composite search index) time series).

ADF, augmented Dickey-Fuller.

Table 2 Results of the VAR models

HIV/AIDS			Syphilis			Gonorrhoea		
Regression variable	Coefficient	R ²	Regression variable	Coefficient	R ²	Regression variable	Coefficient	R ²
α_1	0.160	0.863	α_1	-0.290	0.853	α_1	0.368	0.887
α_2	-0.009		β_1	-408.723		α_2	0.504	
β_1	-0.004		ψ	81.853		β_1	-802.385	
β_2	5663.000		Cons	44 679.478		β_2	128.690	
ψ	46.040		Season1	634.254		ψ	13.225	
Cons	-0.002		Season2	-3074.942		Cons	5528.811	
Season1	-0.002		Season3	5283.978		Season1	-675.487	
Season2	-0.009		Season4	5982.820		Season2	-1900.432	
Season3	-0.001		Season5	8178.966		Season3	378.737	
Season4	-0.002		Season6	7281.723		Season4	723.264	
Season5	-0.002		Season7	8097.763		Season5	921.385	
Season6	-0.002		Season8	8038.559		Season6	698.750	
Season7	-0.002		Season9	4961.415		Season7	754.297	
Season8	-0.002		Season10	3895.552		Season8	750.918	
Season9	-0.002		Season11	2013.618		Season9	-202.920	
Season10	-0.004					Season10	302.411	
Season11	-0.002					Season11	366.978	

* α_1 represents the coefficient of NDC_{t-1} and α_2 represents the coefficient of NDC_{t-2} . β_1 represents the coefficient of $LnBSI_{t-1}$ and β_2 represents the coefficient of $LnBSI_{t-2}$. φ represents the coefficient of trend term and *cons* represents the result of intercept. *season*₁₋₁₁ represents the mean value in $(1 - 11)^{th}$ month.

term. The outcome of the VAR models for HIV/AIDS, syphilis and gonorrhoea was listed in [table 2](#). The equations were solved using ordinary squares estimation.

Testing the residual autocorrelation and evaluating the stability of the VAR model

We tested residuals correlation using a Portmanteau test and the residuals for the three models passed the serial correlation test. Then, we evaluated stability of the VAR systems using the roots of the characteristic polynomial of the coefficient matrix and all models were stable.

Making the prediction

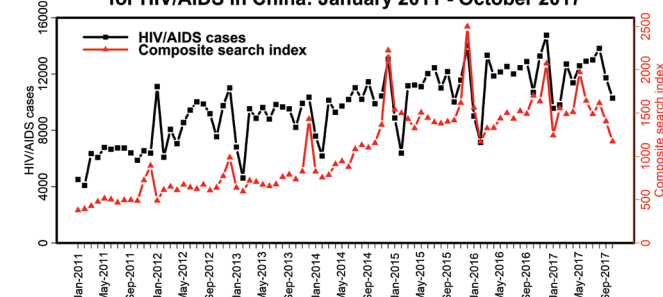
The fitted models were used to predict the number of newly diagnosed cases from November 2016 to October 2017. In the end, the prediction performance of the final model was assessed by two evaluation metrics: root mean square error (RMSE) and mean absolute percentage error (MAPE), and effect of model fitting by R^2 . These two parameters for prediction performance of model indicated that there is an error between true numbers of newly diagnosed cases and predicted ones in different forms. RMSE represented the absolute forecasting error. MAPE indicated the percentage prediction error.²⁹ The definition of RMSE, MAPE were outlined in online supplementary appendix 2.

Python V.3.5.2 was used to crawl the monthly search index data for a time frame between January 2011 and October 2017 nationwide. R V.3.4.1 was used to statistically analyse and graphically illustrate data.

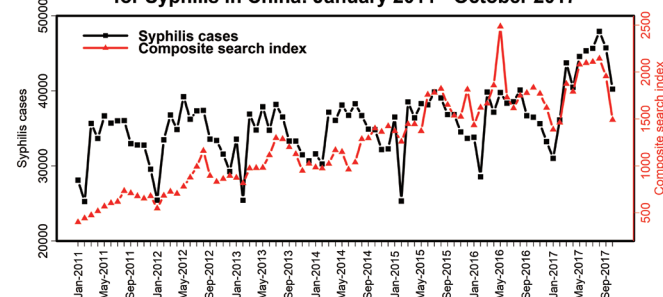
RESULTS

According to the keywords exclusion criteria, 44 keywords for HIV/AIDS, 10 for syphilis and 2 for gonorrhoea still need to be constructed to form a composite search index. The final remaining keywords set of the three HIV/STDs both in Chinese and English was shown in online supplementary appendix 1, table S3. Names and their symptoms all were shown in the final keywords set of each infection/disease, which were usually searched by the people at an earlier stage of these diseases' development.¹⁰ Additionally, the graphs, routes of transmission, testing and treatment were also included in the vocabulary of HIV/AIDS and syphilis respectively. Spearman's rank correlation coefficient between monthly composite search index and the number of newly diagnosed cases showed strong correlations with HIV/AIDS (0.777), syphilis (0.590) and gonorrhoea (0.633, $p < 0.05$ for all). As shown in [figure 1](#), the diagram of the number of newly diagnosed cases and search trend of these three HIV/STDs were annual periodicity and the search trend showed overall upward

Monthly number of newly diagnosed cases and composite search index for HIV/AIDS in China: January 2011 - October 2017



Monthly number of newly diagnosed cases and composite search index for Syphilis in China: January 2011 - October 2017



Monthly number of newly diagnosed cases and composite search index for Gonorrhea in China: January 2011 - October 2017

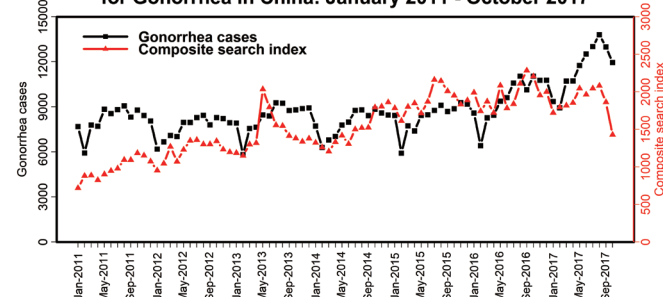
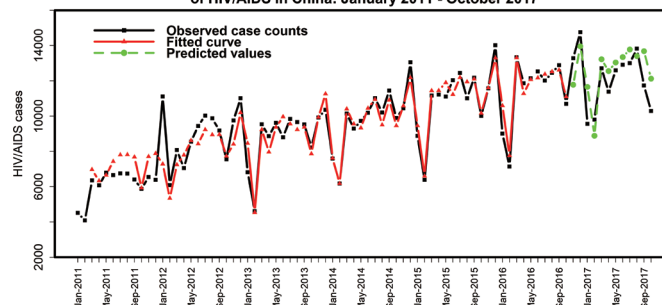


Figure 1 Trends of monthly number of newly diagnosed cases and composite search index.

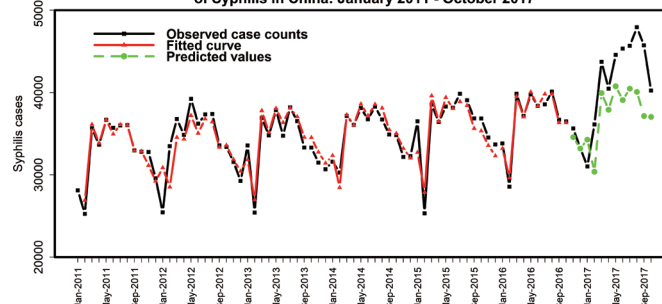
trend for HIV/AIDS and syphilis. Studies that used search data to predict newly diagnosed cases in China owed this phenomenon as health institutions increase their HIV prevention efforts and resources to make people aware of the risks.²⁵ The uptrend of syphilis showed the efforts and resources from government or non-governmental organisations (NGOs) may also have influence on people's awareness of other important STDs like syphilis. The trend of monthly composite search index was basically consistent with the number of newly diagnosed cases for the three HIV/STDs

The selected lag length of VAR models showed that the number of newly diagnosed cases was affected by the number of newly diagnosed cases and the composite search index in previous 2 months of HIV/AIDS and gonorrhoea, in last month for syphilis. For VAR model, it was difficult to interpret every coefficient of all regression variables. The coefficients of trend terms were all positive, which illustrated an overall uptrend of the number of newly diagnosed cases for three infections/diseases during 2011–2017. In a cycle of full year, the trend of all newly diagnosed cases showed cyclical change. All

Monthly observed and predicted number of newly diagnosed cases of HIV/AIDS in China: January 2011 - October 2017



Monthly observed and predicted number of newly diagnosed cases of Syphilis in China: January 2011 - October 2017



Monthly observed and predicted number of newly diagnosed cases of Gonorrhea in China: January 2011 - October 2017

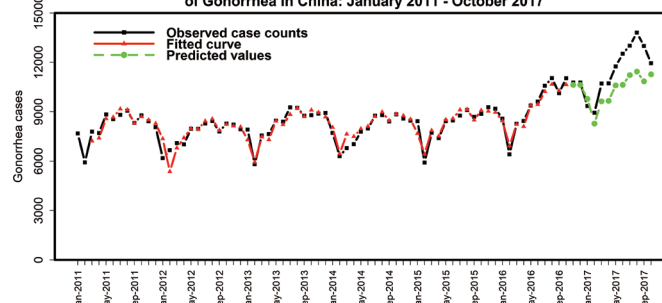


Figure 2 Monthly number of newly diagnosed cases and prediction result based on the models.

infections/diseases started to rise at the beginning of the year and decline at different rates by January next year. In addition, the peak value of HIV/AIDS occurred each December may be attributed to the free testing services on World AIDS Day. Figure 2 showed the VAR models fitted well with the real data and the prediction could catch the periodicity and uptrend character especially for HIV/AIDS and gonorrhoea. The RMSE was 1227.22 for HIV/AIDS, 4929.44 for syphilis and 1342.30 for gonorrhoea. The MAPE was 9.39% for HIV/AIDS, 10.11% for syphilis and 9.35% for gonorrhoea. The R^2 were all more than 85% and MAPE were less than 11%, which indicated the good applicability of VAR in this field.

DISCUSSION

In this paper, we demonstrated that the composite search index resembles the trends of the number of newly diagnosed cases for HIV/AIDS, syphilis and gonorrhoea to a large extent, and point out the potential feasibility of using search engine data to monitor and predict these three HIV/STDs in China. Official data on the number

of newly diagnosed cases are released with delay. Whereas with time series analysis, we could make long-term prediction to provide the health authorities and community health organisations with time to respond in advance. This advantage is of particular significance to the surveillance of infectious diseases.

For infectious diseases, changes in the number of newly diagnosed cases are influenced by changes in time trends, seasonal fluctuations and random disturbances.²⁶ Time series analysis could potentially deal with these features by adding periodic terms and error terms. Besides, the number of newly diagnosed cases of infectious diseases were autocorrelated and using multivariate model could explain the relationship by selecting an appropriate lag length. Figure 2 depicted there were seasonal periodicity for all three HIV/STD, dropping abruptly from January to February, which may attribute to the testing intention reduction of people during the Spring Festival, then going up from February and tending to be stable after several waves in December except for the cases of HIV/AIDS, it possibly due to test promotion activities during World AIDS Day. Compared with syphilis, HIV/AIDS and gonorrhoea models have a longer correlation between the number of newly diagnosed cases and search data. It shall be noted that because the uptrend was more stable, the prediction curve fitted well for HIV/AIDS. As shown in figure 2, there was a rise of incidence in 2017 for syphilis and gonorrhoea, which was different from past years, making the prediction curve lower than true incidence curve. This may be caused by government's effort on publicity of health education and people's attention to syphilis and gonorrhoea testing which were more likely to be deficient than HIV/AIDS in previous years.

It is recommended to merge traditional official data into the internet search data. This can not only heal the peak of search volume trends, but partially excluded from the search volume caused by headlines such as celebrity anecdotes. Reliance on internet data can potentially lead to miscalculation. With VAR model, we could avoid over-estimation by inputting historical data resources such as the number of newly diagnosed cases in the previous months.

An existing study predicted new HIV diagnoses only using Baidu search data,²⁵ whereas we monitored and predicted the overall number of newly diagnosed HIV/AIDS cases. For example, the keywords, 'How long can people survive with AIDS' and 'Is there a cure for AIDS', were more likely searched by people who have been infected for a period of time or AIDS patients. After filtering the keywords, the remaining keywords set showed variation in the aspects of the three diseases. The keywords set of HIV/AIDS and syphilis covered almost all domains relevant to disease name, symptom, transmission, testing and treatment which presented the procedure of the diseases' progress. The keywords set for gonorrhoea only included the name and symptom expressed in similar forms. The applied exclusion criteria varied for different infectious diseases/infection, which

indicated the potential necessity of adjusting the filtering methods according to the characteristics of individual disease/infection in the future.

People may search for relevant information on the internet when they have a specific illness or just curious about the disease. Even if the irrelevant keywords were excluded, there was no guarantee that the rest of the keywords involved in search behaviours were incurred by 'true illness' or real infection. One of the confounders caused by irrelevant search behaviours is social media hype. We used composite search index constructed from different keywords and models including real data resources to minimise the impact of huge peaks or valleys in the search trend caused by specific keywords as far as possible.

There were also many explorations we can do in the future. For example, characteristics of search engine users, such as age, gender and sexual orientation, which can potentially be obtained from search portals, are useful in the understanding of target population. It is important to take this opportunity to target populations at high risk of HIV/STD infection, such as men who have sex with men.³⁰ Similarly, there were limitations of this research. One of the limitations was the change of algorithm of Baidu search index on the official website, thus the according coefficients of the model needed recalculation. Restricted to the limited time span, whether the VAR model was applicable for the prediction in a much longer period of time needed further verification. Second, we used and made the prediction at national level even search index for specific keyword at province/city level could already be retrieved in Baidu Search Index website. The report of diagnosed cases could not updated timely and consistently for these STIs/diseases in less-developed districts such as Tibet, urging local official institutions to complete surveillance and reporting system.

Internet search data are readily available and may potentially be incorporated into the routine surveillance system in monitoring and predicting HIV/AIDS, syphilis and gonorrhoea epidemics in China. Additional socio-demographic characteristics of search engine users may provide more information.

Author affiliations

¹School of Public Health, Sun Yat-Sen University, Guangzhou, China

²School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, China

³The Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia

⁴School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia

⁵China-Australia Joint Research Center for Infectious Diseases, School of Public Health, Xi'an Jiaotong University, Xi'an, China

⁶Melbourne Sexual Health Centre, Alfred Health, Melbourne, Victoria, Australia

⁷Central Clinical School, Faculty of Medicine Nursing and Health Sciences, Monash University, Melbourne, Victoria, Australia

⁸Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, Zhengzhou, China

⁹School of Data Science, City University of Hong Kong, Kowloon, Hong Kong

¹⁰University of North Carolina Project China, Guangzhou, China

¹¹Southern Medical University Dermatology Hospital, Guangzhou, China

¹²Division of Epidemiology and Community Health, School of Public Health, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA

¹³Sun Yat-sen Global Health Institute, Sun Yat-Sen University, Guangzhou, China

Acknowledgements The authors would like to thank Jianwu Huang, an undergraduate student of School of Data and Computer Science, Sun Yat-sen University, China, for his great help with web crawler technology.

Contributors HZ designed and coordinated this study. Data were collected and analysed by RH and GL. RH and HZ drafted the manuscript. The manuscript was reviewed by MKS, JL, LZ, WT, QD, QZ and HZ. All authors read and approved the final manuscript.

Funding This publication was funded by National Natural Science Foundation of China (81703278, 81803334, 71672163), National Science and Technology Major Project of China (2018ZX10721102), Australian National Health and Medical Research Council Early Career Fellowship (APP1092621), Sanming Project of Medicine in Shenzhen, China (SZSM201811071), China Medical Board (18-301) and A Major Infectious Disease Prevention and Control of the National Science and Technique Major Project (2018ZX10715004).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval No work with human subjects was directly involved in our research. The search data were recorded from Baidu Search Index Website (<http://index.baidu.com/v2/index.html#/>) and the official diagnosed cases were extracted from the monthly report of Chinese Center for Disease Control and Prevention (China CDC) or academic journal *Chinese Journal of AIDS&STD*. No data involved any personal information. Permission to conduct the research was granted by School of Public Health, Sun Yat-sen University (2019008).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Original data are available in a public, open access repository. Data after processing are available upon reasonable request. All data relevant to the study are included in the article or uploaded as supplementary information. Aggregated data on the number of newly diagnosed cases of HIV/AIDS, syphilis and gonorrhea from January 2011 to October 2017 (Table S1, Appendix 1) were available from the monthly academic journal *Chinese Journal of AIDS&STD* and the official website of Chinese Center for Diseases Control and Prevention (<http://www.chinacdc.cn/>). The search index data for specific keyword were not shown in detail and could be accessed by sending email to corresponding author.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Ruonan Huang <http://orcid.org/0000-0003-4586-8172>

Lei Zhang <http://orcid.org/0000-0003-2343-084X>

Weiming Tang <http://orcid.org/0000-0002-9026-707X>

M. Kumi Smith <http://orcid.org/0000-0001-5861-8100>

Huachun Zou <http://orcid.org/0000-0002-8161-7576>

REFERENCES

- 1 Zou Y, Liao Y, Liu F, *et al.* The annual economic burden of syphilis: an estimation of direct, productivity, and intangible costs for syphilis in Guangdong initiative for comprehensive control of syphilis sites. *Sex Transm Dis* 2017;44:671–7.
- 2 Hu T, Du J. Progress on social economic burden and cost-effectiveness of comprehensive prevention of AIDS. *Int J Infect Dis* 2014;41:350–4.
- 3 Fung IC-H, Hao Y, Cai J, *et al.* Chinese social media reaction to information about 42 notifiable infectious diseases. *PLoS One* 2015;10:e0126092.
- 4 Wang L, Wang Y, Jin S, *et al.* Emergence and control of infectious diseases in China. *Lancet* 2008;372:1598–605.
- 5 Law of the People's Republic of China on prevention and treatment of infectious diseases (2004 revision) (issued on 08282004 effective on 12012004).
- 6 Johnson AK, Mikati T, Mehta SD. Examining the themes of STD-related Internet searches to increase specificity of disease forecasting using Internet search terms. *Sci Rep* 2016;6:36503.
- 7 Ginsberg J, Mohebbi MH, Patel RS, *et al.* Detecting influenza epidemics using search engine query data. *Nature* 2009;457:1012–4.
- 8 Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *Plos Curr* 2011;6.
- 9 Xiao QY, Liu HJ, Feldman MW. Tracking and predicting hand, foot, and mouth disease (HFMD) epidemics in China by Baidu queries. *Epidemiol Infect* 2017;145:1699–707.
- 10 Milinovich GJ, Williams GM, Clements ACA, *et al.* Internet-based surveillance systems for monitoring emerging infectious diseases. *Lancet Infect Dis* 2014;14:160–8.
- 11 Santillana M, Nguyen AT, Dredze M, *et al.* Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 2015;11:e1004513.
- 12 Yuan Q, Nsoesie EO, Lv B, *et al.* Monitoring influenza epidemics in China with search query from baidu. *PLoS One* 2013;8:e64323.
- 13 Guo P, Zhang J, Wang L, *et al.* Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. *Sci Rep* 2017;7:46469.
- 14 Davidson MW, Haim DA, Radin JM. Using networks to combine "big data" and traditional surveillance to improve influenza predictions. *Sci Rep* 2015;5:8154.
- 15 Crockett RJK, Althouse BM, YY N, *et al.* Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 2011;5:e1258.
- 16 Guo P, Wang L, Zhang Y, *et al.* Can internet search queries be used for dengue fever surveillance in China? *Int J Infect Dis* 2017;63:74–6.
- 17 Liu K, Wang T, Yang Z, *et al.* Using Baidu search index to predict dengue outbreak in China. *Sci Rep* 2016;6:38040.
- 18 Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 2011;5:e1258.
- 19 Xie T, Yang Z, Yang S, *et al.* Correlation between reported human infection with avian influenza A H7N9 virus and cyber user awareness: what can we learn from digital epidemiology? *Int J Infect Dis* 2014;22:1–3.
- 20 Liu K, Huang S, Miao Z-P, *et al.* Identifying potential norovirus epidemics in China via Internet surveillance. *J Med Internet Res* 2017;19:e282.
- 21 Chan EH, Sahai V, Conrad C, *et al.* Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis* 2011;5:e1206.
- 22 Young SD, Zhang Q. Using search engine big data for predicting new HIV diagnoses. *PLoS One* 2018;13:e0199527.
- 23 Jena AB, Karaca-Mandic P, Weaver L, *et al.* Predicting new diagnoses of HIV infection using Internet search engine data. *Clin Infect Dis* 2013;56:1352–3.
- 24 Li K, Liu M, Feng Y, *et al.* Using Baidu search engine to monitor AIDS epidemics inform for targeted intervention of HIV/AIDS in China 2019.
- 25 Zhang Q, Chai Y, Li X, *et al.* Using internet search data to predict new HIV diagnoses in China: a modelling study. *BMJ Open* 2018;8:e018335.
- 26 He G, Chen Y, Chen B, *et al.* Using the Baidu search index to predict the incidence of HIV/AIDS in China. *Sci Rep* 2018;8:9038.
- 27 Dong X, Boulton ML, Carlson B, *et al.* Syndromic surveillance for influenza in Tianjin, China: 2013–14. *J Public Health-Uk* 2016;39:fdw022.
- 28 Bose E, Hravnak M, Sereika SM. Vector autoregressive models and Granger causality in time series analysis in nursing research: dynamic changes among vital signs prior to cardiorespiratory instability events as an example. *Nurs Res* 2017;66:12–19.
- 29 Shcherbakov MV, Brebels A, Shcherbakova NL, *et al.* A survey of forecast error measures. *World Appl Sci J* 2013;24:171–6.
- 30 Zhang J, Xu J-J, Song W, *et al.* HIV incidence and care linkage among MSM First-Time-Testers in Shenyang, China 2012–2014. *AIDS Behav* 2018;22:711–21.