BMJ Open

Using internet search data to predict new HIV diagnoses in China: a modelling study

Qingpeng Zhang,[1,2] Yi Chai,[1,3] Xiaoming Li,[4] Sean D Young,[5] Jiaqi Zhou[1]

To cite: 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. doi:10.1136/bmjopen-2017-018335

► 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-2017-018335).

Received 21 June 2017
Revised 18 June 2018
Accepted 20 August 2018

[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[2]City University of Hong Kong Shenzhen Research Institute, Shenzhen, China
[3]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, Hong Kong SAR, China
[4]Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA
[5]University of California Institute for Prediction Technology, Department of Family Medicine, University of California Los Angeles, Los Angeles, California, USA

Correspondence to
Dr Qingpeng Zhang;
qpzhang@email.arizona.edu

ABSTRACT

Objectives Internet data are important sources of abundant information regarding HIV epidemics and risk factors. A number of case studies found an association between internet searches and outbreaks of infectious diseases, including HIV. In this research, we examined the feasibility of using search query data to predict the number of new HIV diagnoses in China.

Design We identified a set of search queries that are associated with new HIV diagnoses in China. We developed statistical models (negative binomial generalised linear model and its Bayesian variants) to estimate the number of new HIV diagnoses by using data of search queries (Baidu) and official statistics (for the entire country and for Guangdong province) for 7 years (2010 to 2016).

Results Search query data were positively associated with the number of new HIV diagnoses in China and in Guangdong province. Experiments demonstrated that incorporating search query data could improve the prediction performance in nowcasting and forecasting tasks.

Conclusions Baidu data can be used to predict the number of new HIV diagnoses in China up to the province level. This study demonstrates the feasibility of using search query data to predict new HIV diagnoses. Results could potentially facilitate timely evidence-based decision making and complement conventional programmes for HIV prevention.

Strengths and limitations of this study

- This work is one of the first data-driven studies on using search engine (Baidu) data to predict new HIV diagnoses in China.
- The proposed models can predict new HIV diagnoses in China with high accuracy.
- The proposed models are trained on historical data. Thus, they are ineffective if historical data are unavailable.
- This study focuses on predicting the number of new HIV diagnoses, instead of HIV incidence (the estimated number of people newly infected with HIV).

INTRODUCTION

HIV is a critical public health issue worldwide.[1] There are approximately 36.7 million people living with HIV worldwide (including 0.66 million in China) worldwide at the end of 2016.[2 3] The epidemics of HIV has caused huge burden on economy, society, politics and culture. The Chinese surveillance system consists of (1) a sentinel surveillance system, (2) a case-reporting system and (3) a behavioural surveillance and epidemiological survey.[4–7] By 2010, the number of sentinel surveillance system reached 1888 across the country for eight key populations, and the system involved various sampling methods: the populations are drug users (snowball sampling in communities or at detention/detoxification centres); men who have sex with men (MSM, stratified snowball sampling at venues or through internet/social networks); female sex workers (stratified sampling at venues or detention centres); sexually transmitted infection (STI) clinic attendees (consecutive sampling); antenatal care clinic attendees (consecutive sampling); long-distance truck drivers (consecutive sampling); mobile population (consecutive sampling); and young students (multistage clustered sampling at colleges).[4] The annual funding for HIV prevention and control increased dramatically in the previous decade (average increase of 8% per year).[8] Existing HIV prevention and surveillance programmes are based on official statistics from national monitoring authorities, which usually have a lag of 1 month. Recent works conducted a review on HIV prevention and control.[9–11]

Retrieving up-to-date information of new HIV diagnoses can help raise public awareness of its outbreaks and provide effective prevention efforts.[12] Although HIV interventions are not as time sensitive as influenza interventions due to the long incubation period of the virus, precise estimation of new HIV diagnoses can help authorities and public health officials to effectively allocate resources and schedule prevention programmes (eg, national and local campaigns).

In practice, if a model accurately predicts a burst of HIV diagnoses in the current month (nowcasting) or in a few months (forecasting), then decision makers can allocate resources (eg, test kit, fund for campaigns and antiretroviral therapy kit) to help national and local public health services satisfy the upcoming demand. Timely HIV test and early treatment can effectively prevent onward transmission of HIV.[7 13] In addition, targeted campaigns could help raise public care for HIV patients and reduce HIV stigma.[14 15] Thus, timely and quality treatment in early stages of the infection is important for people who are newly diagnosed with HIV and those who are at risk. Prediction-based campaigns may inform at-risk people of the risk of HIV transmission and allow them to prevent having risky behaviour during the upcoming outbreaks of new HIV diagnoses. These efforts may not directly prevent some at-risk people from being infected but can help them receive better and timely treatment services.

Simulation models can be used to estimate trends in the HIV epidemic and explore the effectiveness of interventions in preventing its spread.[16–18] Despite numerous studies on simulation models of HIV epidemics and transmissions, current decision makers do not always use them in practice partly because of the lack of confidence in them. In general, simulation models can be fitted well by using historical data. However, using these models to analyse future scenarios remains challenging given the limited information of future epidemics. If we can develop accurate prediction models for timely estimation of epidemics (such as new HIV diagnoses), then simulation models can be calibrated and applied to actual decision making. Hence, it is important to tune the parameters in simulation models using real-time prediction, so that decision makers could develop effective intervention plans.[17 18] As a result, new data sources are highly needed to complement conventional report of incidence statistics from Centers for Disease Control and Prevention (CDC), particularly in low/middle-income countries, such as China, where HIV prevention programmes targeting key populations have not been well-developed.

Internet data are important sources of abundant information regarding HIV epidemics.[19] In particular, internet searches have been found to be associated with the outbreaks of infectious diseases, such as influenza, hand, foot and mouth diseases and dengue fever.[12 20–35] Internet search data are relevant to determine HIV incidence because people, especially adolescents and young adults, rely on the internet as their primary source to acquire HIV-related information.[36–40] Such internet data are particularly relevant to young (under 18 years of age) people given their high prevalence of using the internet. Young people often encounter barriers (parent's consent for testing for HIV) in accessing corresponding services and may actively seek information from the internet.[41 42]

Previous studies used Wikipedia searches and social media data to predict HIV diagnoses.[19 43–45] People prefer the internet search engine as platform to obtain information about diseases that are highly stigmatised in the society (eg, sexually transmitted disease (STD) and mental health) because anonymous internet searches can better protect their privacy compared with other social media platforms (eg, Weibo, WeChat and blog).[46–50] Therefore, internet search data have strong predictive power for HIV diagnoses. Existing empirical studies reported a strong correlation between Google trends and incidence of other STDs, including HIV in the USA.[12 51–53] However, modelling research on using query data of search engines to predict HIV incidence has been rarely reported. Prediction models must be developed to help decision makers monitor HIV epidemics and estimate outbreaks. The requirement is vital in China, where the traditional HIV monitoring system is very costly because of the large population and the stigma towards HIV.[41 42 51 52] Therefore, it is sensible for Chinese authorities and decision makers to use internet-based data to supplement traditional surveillance to help with HIV intervention. To date, research on using internet search query data for HIV surveillance in China is scarce and should be further explored.

In this work, we developed six statistical models to estimate the number of new HIV diagnoses in China by incorporating search query data and historical records at the national and provincial levels. At the end of 2016, China had a total of 731 million Web users, among which the majority (82.4%) use search engines to acquire information.[54] Existing research suggests that the use of the internet and social media is correlated with HIV testing behaviour among key populations.[55] Our hypothesis is that people who suspect that they are infected with HIV (because of their at-risk behaviour or related symptoms) would search for HIV-related information in the internet search engine and subsequently undergo HIV testing. People infected with HIV will be counted as new diagnoses in the official statistics for the corresponding month. The internet searches represent a representative sample of the at-risk population, who are likely to undergo HIV testing.

Data

We retrieved two sets of official statistics as dependent variable. At the national level, we collected the monthly counts of new HIV diagnoses in China from the China CDC (http://www.chinacdc.cn) between 2010 and 2016 (84 months). The robustness of the proposed models should also be evaluated at the local level because public health interventions are usually targeted at specific subgroups in high-risk local geographic areas. Therefore, we retrieved the monthly counts of new HIV diagnoses in Guangdong province, which is the most populous province in China (with more than 100 million people), from the Health and Family Planning Commission of Guangdong Province (http://www.gdwst.gov.cn/) during the same period. These data are published with a lag of approximately 1 month. We did not investigate the difference between gender and age groups because of lack of demographic information.

Internet search query data were obtained from Baidu Index (https://index.baidu.com/), a Google Trend

equivalent to the most popular search engine (Baidu) in China. Similar to Google Trends, Baidu Index is a normalised value that measures the search volume of certain keywords at a specific time. Baidu is the largest search engine in China and has a market share of approximately 70% to 83%.[56] The majority (82.4%) of Chinese web users use search engines to acquire information.[54] Although not all people diagnosed with HIV used Baidu to search for HIV-related information, several lines of evidence indicate that the number of new diagnoses is correlated with the internet searches.[12] In particular, Baidu users are a meaningful sample of people diagnosed with HIV in China because we are modelling the number of new diagnoses at the aggregated level (national and provincial).

To ensure the quality of search query data, we adopted a data-driven approach for identifying HIV-related terms in Baidu searches. In addition to '艾滋病' (HIV/AIDS), we used the Baidu Index toolkit to retrieve terms searched by users immediately before and after searching for '艾滋病' (HIV/AIDS). The retrieved terms were manually cleaned to exclude ambiguous terms and those associated with other diseases, such as '途径' (ways) and '症状' (symptom). We finally collected data on the monthly frequencies of Baidu searches for the following eight HIV-related terms from 2011 (its inception) to 2016: '艾滋病' (HIV/AIDS), '艾滋病检测' (HIV testing), '艾滋病初期症状' (initial symptoms of HIV), '艾滋病窗口期' (HIV test window period), '艾滋病试纸' (HIV test kits/strips), '艾滋病潜伏期' (incubation period of HIV/AIDS), '艾滋病能活多久' (how long can HIV patients live) and '艾滋病传播途径' (ways of HIV transmission). We collected search query data generated by users located in the entire country and those in Guangdong province.

We performed a cross-correlation test with search query data and HIV statistics to filter terms that are not significantly correlated with new HIV diagnoses at the national level (correlation coefficient lower than 0.3).[28 29 35] For Guangdong province, all terms are correlated. We included seven terms in the composite search index at the national level and eight terms in the composite search index for Guangdong province (table 1). The composite search index $index(t)$ is the weighted summation of the query data of the selected terms, where the weight of a term is the correlation coefficient between the new HIV diagnoses curve $y(t)$ and the frequency of the search curve. This composition of search queries/terms produces the most correlative predictor of internet searches and has been widely used for disease surveillance.[29 35 57] We normalised the composite search indices according to the actual counts of new diagnoses by using equation (1) and present their time series in figure 1. The normalised composite search index is calculated as follows:

$$index(t)' = \frac{index(t) - \min index(i)}{\max index(i) - \min index(i)} \times [\max y(i) - \min y(i)] + \min y(i) \quad (1)$$

Table 1 Terms included in the composite search indices at the national and provincial levels

Chinese	English translation	Correlation coefficient (China)	Correlation coefficient (Guangdong)
艾滋病	HIV/AIDS	0.54 (p<0.001)	0.43 (p<0.001)
艾滋病初期症状	Initial symptoms of HIV	0.44 (p<0.001)	0.49 (p<0.001)
艾滋病试纸	HIV test kits/strips	0.36 (p<0.01)	0.42 (p<0.001)
艾滋病检测	HIV testing	0.35 (p<0.01)	0.41 (p<0.001)
艾滋病窗口期	HIV test window period	0.34 (p<0.01)	0.36 (p<0.01)
艾滋病能活多久	How long can HIV patients live	0.33 (p<0.01)	0.43 (p<0.001)
艾滋病传播途径	Ways of HIV transmission	0.31 (p<0.01)	0.37 (p<0.01)
艾滋病潜伏期	Incubation period of HIV	Excluded	0.31 (p<0.01)

Patient and public involvement

The patients and the public were not involved in this study.

Model development

The counts of the new HIV diagnoses generally follow a negative binomial distribution. We examined the relationship between new HIV diagnoses and the search query data (composite search index, figure 1). Similar patterns were obtained over time and had yearly seasonality. The search query data could effectively capture the outbreaks of new HIV diagnoses. Hence, search query data could be used to estimate new HIV diagnoses before official data are published (commonly named as 'nowcasting' task).

Although both curves show a continuous increase over time, the search query data curve increases faster, indicating the growing awareness of HIV in China and Guangdong province. Chinese authorities and non-governmental organisations (NGOs) provide significant resources towards national and local HIV prevention programmes. In particular, huge campaigns are conducted during World AIDS Day (December 1) every year. Such campaigns greatly help people learn about HIV. Many newly diagnosed people were educated by these campaigns before they realised their risk of contracting HIV. These campaigns are also the reason for the burst of new HIV diagnoses in December each year. In addition, efforts are exerted to provide test kits to young adults. For example, in a recent practice, the introduction of vending machines of HIV test kits in universities was found to increase the availability of self-testing

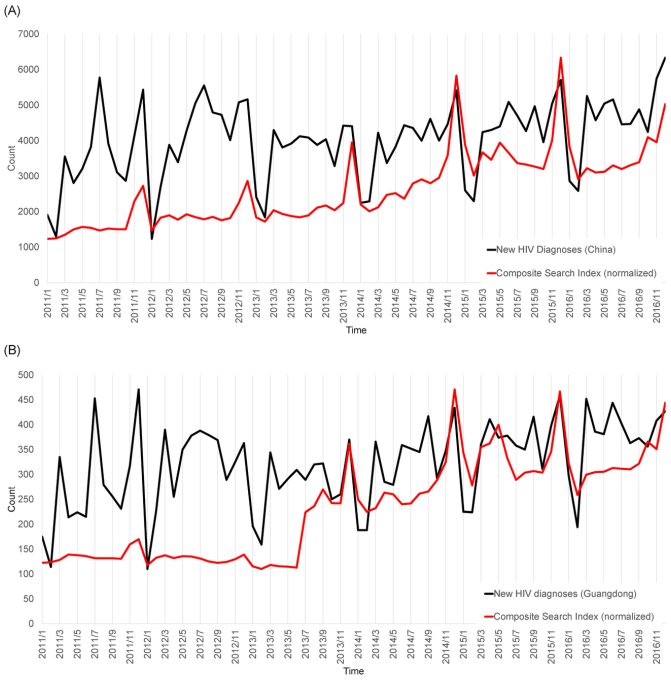

Figure 1 Number of new HIV diagnoses (black) and composite search index (red) from 2011 to 2016 in China (A) and Guangdong province (B).

among college students (goo.gl/k6E65). In general, the rapid increase in HIV-related internet searches is a result of the increasing effort and resources of authorities and NGOs for HIV prevention. The different increasing rates indicate that the prediction models should be calibrated frequently over time to capture the changing associations between the search query data and the count of new HIV diagnoses.

Basing on the empirical observations, we propose six competing models to estimate the number of new HIV diagnoses by using search query data. The first two models are a negative binomial generalised linear model (nbGLM) with a log link and autoregressive terms (nbGLM-AR) and a Bayesian negative binomial generalised linear model (BnbGLM) with the same variables (BnbGLM-AR). The third and fourth models are nbGLM and BnbGLM with a variable representing the composite Baidu Search Index (nbGLM-Baidu and BnbGLM-Baidu). The fifth and sixth models are nbGLM and BnbGLM with autoregressive terms and composite search index (nbGLM-AR-Baidu and BnbGLM-AR-Baidu). Here, we selected nbGLM as the base model because the data of HIV diagnoses followed a negative binomial distribution. The difference between nbGLM-based and BnbGLM-based models is the coefficient estimation approach. nbGLM uses least squares fitting approach, which might lead to overfitting problem.[58] BnbGLM uses Bayesian inference approach, which could fully use prior information for training and resolving the overfitting problem.[59] These two sets of models can be represented as follows:
nbGLM-AR and BnbGLM-AR:

$$\ln y(t) = \mu_0 + \sum_{n=1}^{12} \beta_n \ln y(t-n) + \varepsilon_t \quad (2)$$

nbGLM-Baidu and BnbGLM-Baidu:

$$\ln y(t) = \mu_0 + \mu_1 index(t)' + \varepsilon_t \quad (3)$$

nbGLM-AR-Baidu and BnbGLM-AR-Baidu:

$$\ln y(t) = \mu_0 + \mu_1 index(t)' + \sum_{n=1}^{12} \beta_n \ln y(t-n) + \varepsilon_t \quad (4)$$

where $y(t)$ represents the number of new HIV diagnoses in month t. As we found a monthly seasonality with a circle of a full year, a 12-month autoregressive component is adopted to determine the trend of the overall curve. $index(t)'$ represents the normalised composite search index of month t. ε_t represents Gaussian white noise.

We split the data into two sets, namely the training set (January 2010 to December 2014 for new HIV diagnoses data for search query data) and the test set (January 2015 to December 2016). We estimated the parameters of the statistical models using the training set and evaluated the prediction performance using the test set. MASS and BRMS packages in R were used for model fitting. We focused on two main tasks: (1) using search query of a certain month to predict the number of new HIV diagnoses in the same month ('nowcast' task) and (2) using search query data of a certain month to predict new HIV diagnoses in the upcoming 1 and 2 months ('forecast' task). For the nowcast task, we used equations (2)–(4). For the forecast task, we modified the models as follows:
nbGLM-AR and BnbGLM-AR (forecast):

$$\ln y(t+k) = \mu_0 + \sum_{n=1}^{12} \beta_n \ln y(t-n) + \varepsilon_t \quad (5)$$

nbGLM-Baidu and BnbGLM-Baidu (forecast):

$$\ln y(t+k) = \mu_0 + \mu_1 index(t)' + \varepsilon_t \quad (6)$$

nbGLM-AR-Baidu and BnbGLM-AR-Baidu (forecast):

$$\ln y(t+k) = \mu_0 + \mu_1 index(t)' + \sum_{n=1}^{12} \beta_n \ln y(t-n) + \varepsilon_t \quad (7)$$

where k is equal to 1 or 2, indicating 1-month or 2-month ahead forecast. For both tasks, we used an adaptive time window for model training; that is, each estimate was based on the model trained using the data of all previous months. This adaptive method is more appropriate than fixed time window method[60] or shifting time window method[23] because it can take advantage of all available training data and the seasonality of the data is consistent over the years. To confirm the robustness of the proposed models, we evaluated their performances by using fixed and shifting time windows. We obtained consistent but slightly less accurate results (see online supplementary materials).

We used commonly adopted tools, namely root mean square error (RMSE) and normalised root mean square error (NRMSE), to evaluate the accuracy of the nowcast and forecast results. RMSE calculates the SD of prediction errors (the smaller, the better), and NRMSE is a normalised version of RMSE for comparing performance at different scales:

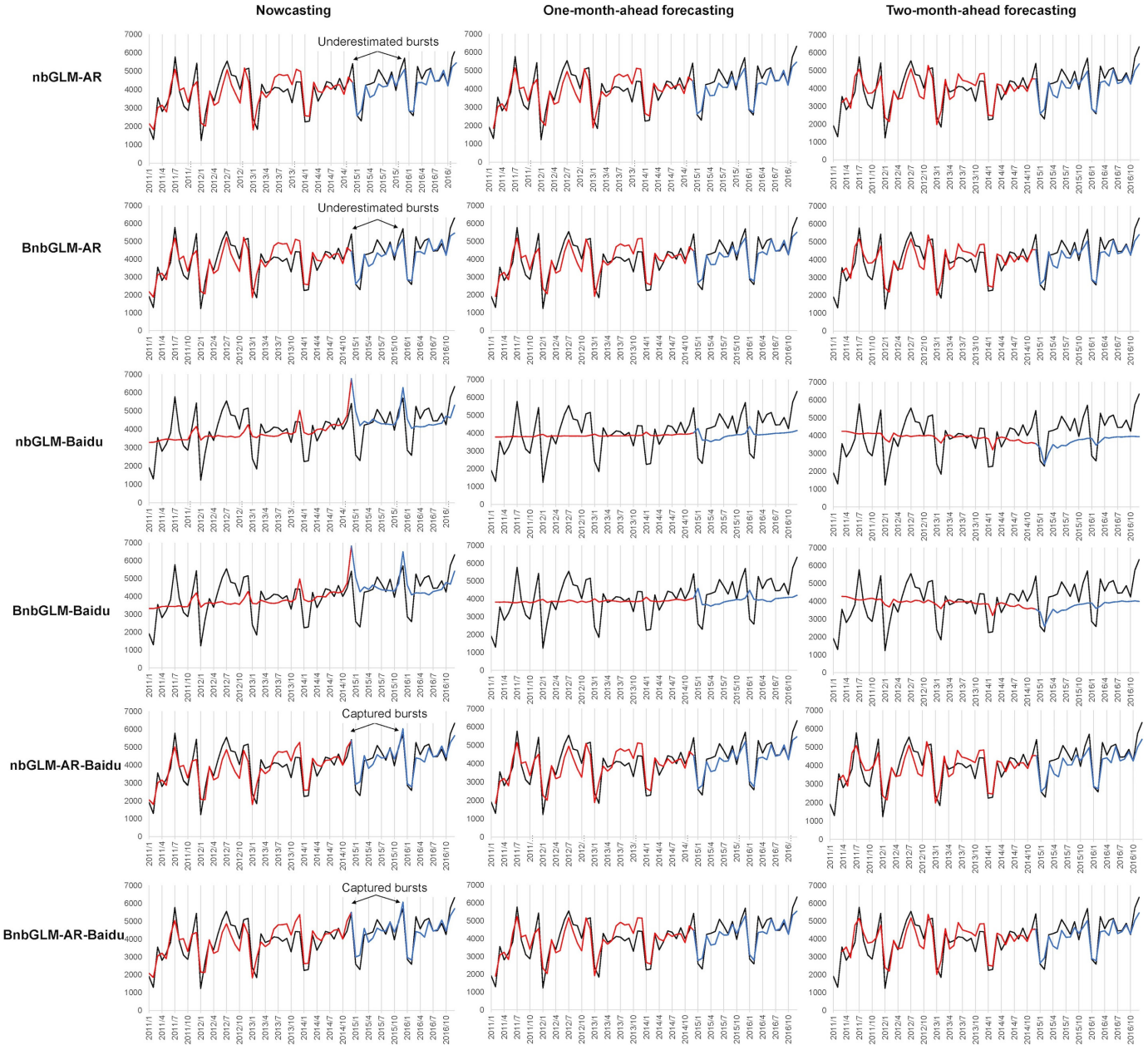

Figure 2 Actual number of new HIV diagnoses and prediction results of the six proposed models (China). The black curve represents the actual data of new HIV diagnoses. The red curve represents the fitted values. The blue curve represents the prediction result. BnbGLM-AR, Bayesian negative binomial generalised linear model (BnbGLM) with autoregressive terms; BnbGLM-AR-Baidu, BnbGLM with autoregressive terms and the composite Baidu Search Index; BnbGLM-Baidu, BnbGLM with a variable representing the composite Baidu Search Index; nbGLM-AR, negative binomial generalised linear model (nbGLM) with autoregressive terms; nbGLM-AR-Baidu, nbGLM with autoregressive terms and the composite Baidu Search Index; nbGLM-Baidu, nbGLM with a variable representing the composite Baidu Search Index.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{p}\left(\hat{y}_i - y_i\right)^2}{p}} \quad (8)$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (9)$$

where $\hat{y}_i$ represents the forecasted values of the corresponding observed value y_i. y_{max} and y_{min} represent the maximum and the minimum observed values, respectively.

RESULTS

We present the prediction results in figures 2 (China) and 3 (Guangdong province) and compared the performances of different models in tables 2 (China) and 3 (Guangdong province). All models performed reasonably well in predicting the number of new HIV diagnoses of the current month, 1 month in advance and 2 months in advance. The models with only the composite search index (nbGLM-Baidu and BnbGLM-Baidu) were

Table 2 Accuracy in predicting the number of new HIV diagnoses in China

	Nowcasting		One-month ahead forecasting		Two-month ahead forecasting	
Model	RMSE	NRMSE	RMSE	NRSME	RMSE	NRMSE
nbGLM-AR	473.35	11.71%	484.21	11.98%	528.38	13.08%
nbGLM-Baidu	957.58	23.7%	1166.38	28.86%	1176.06	29.1%
nbGLM-AR-Baidu	420.68	10.41%	482.79	11.95%	539.37	13.35%
BnbGLM-AR	455.65	11.27%	456.95	11.31%	497.11	12.3%
BnbGLM-Baidu	976.99	24.18%	1176.16	29.11%	1145.23	28.34%
BnbGLM-AR-Baidu	423.17	10.47%	451.75	11.18%	508.31	12.58%

BnbGLM-AR, Bayesian negative binomial generalised linear model (BnbGLM) with autoregressive terms; BnbGLM-AR-Baidu, BnbGLM with autoregressive terms and the composite Baidu Search Index; BnbGLM-Baidu, BnbGLM with a variable representing the composite Baidu Search Index; nbGLM-AR, negative binomial generalised linear model (nbGLM) with autoregressive terms; nbGLM-AR-Baidu, nbGLM with autoregressive terms and the composite Baidu Search Index; nbGLM-Baidu, nbGLM with a variable representing the composite Baidu Search Index; NRMSE, normalised root mean square error; RMSE, root mean square error.

less accurate than those with autoregressive terms. This finding is expected because the number of new HIV diagnoses exhibits a clear seasonality within a 12-month cycle. Our results indicated that using search query data only could generate a reasonable prediction for nowcasting tasks, with normalised RMSE ranging from 23% to 24% at the national and provincial levels. The prediction results in figures 2 and 3 indicated that the composite search index-based models can accurately capture the outbreaks of HIV diagnoses. Furthermore, we found that the variable of the composite search index is only statistically significant in the nowcasting tasks. Hence, internet search query data are useful in estimating the number of HIV diagnoses for the current month as well as in modelling the outbreaks of HIV diagnoses. A full list of the estimates of regression coefficients for all models is presented in online supplementary materials.

The integration of the composite search index and autoregressive terms led to excellent performance in all tasks at the national level. In particular, nbGLM-AR-Baidu performed the best in the nowcasting task. The performance of nbGLM-AR-Baidu was 11.8% more accurate than that of the nbGLM-AR model. BnbGLM-AR-Baidu performed almost identically (<1% difference) to nbGLM-AR-Baidu in the nowcasting task and slightly better in the forecasting task. BnbGLM-AR-Baidu was 7.2% and 3.9% more accurate than the nbGLM-AR in the 1-month and 2-month ahead forecasting, respectively.

For the study on Guangdong province, BnbGLM-AR-Baidu performed the best in the nowcasting task. The performance was 4.1% more accurate than that of nbGLM-AR. In the forecasting tasks, although BnbGLM-AR performed the best, the difference between BnbGLM-AR and BnbGLM-AR-Baidu was minimal (less than 1%). Thus, search query data are useful for the nowcasting task in Guangdong.

DISCUSSION

An interesting finding is that the number of new HIV diagnoses always surged in December, and dipped in January and February. The surge in HIV diagnoses in December may be due to the national campaign around the World AIDS Day. There are multiple effects[61 62]: (1) many people become aware of their risk of being infected, and may search for HIV-related information more often; (2) it is easier for at-risk people to assess HIV testing services, thus they may do the test and be diagnosed. The decrease in

Table 3 Accuracy in predicting the number of new HIV diagnoses in Guangdong province

	Nowcasting		One-month ahead forecasting		Two-month ahead forecasting	
Model	RMSE	NRMSE	RMSE	NRSME	RMSE	NRMSE
nbGLM-AR	56.39	21.28%	54.73	20.65%	54.66	20.6%
nbGLM-Baidu	64.08	24.18%	80.65	30.43%	82.84%	31.26%
nbGLM-AR-Baidu	54.64	20.62%	55.09	20.79%	55.25	20.85%
BnbGLM-AR	55.2	20.83%	53.75	20.28%	52.59	19.84%
BnbGLM-Baidu	63.03	23.79%	79.35	29.94%	81.38	30.71%
BnbGLM-AR-Baidu	54.18	20.45%	54.21	20.46%	53.01	20.04%

BnbGLM-AR, Bayesian negative binomial generalised linear model (BnbGLM) with autoregressive terms; BnbGLM-AR-Baidu, BnbGLM with autoregressive terms and the composite Baidu Search Index; BnbGLM-Baidu, BnbGLM with a variable representing the composite Baidu Search Index; nbGLM-AR, negative binomial generalised linear model (nbGLM) with autoregressive terms; nbGLM-AR-Baidu, nbGLM with autoregressive terms and the composite Baidu Search Index; nbGLM-Baidu, nbGLM with a variable representing the composite Baidu Search Index; NRMSE, normalised root mean square error; RMSE, root mean square error.

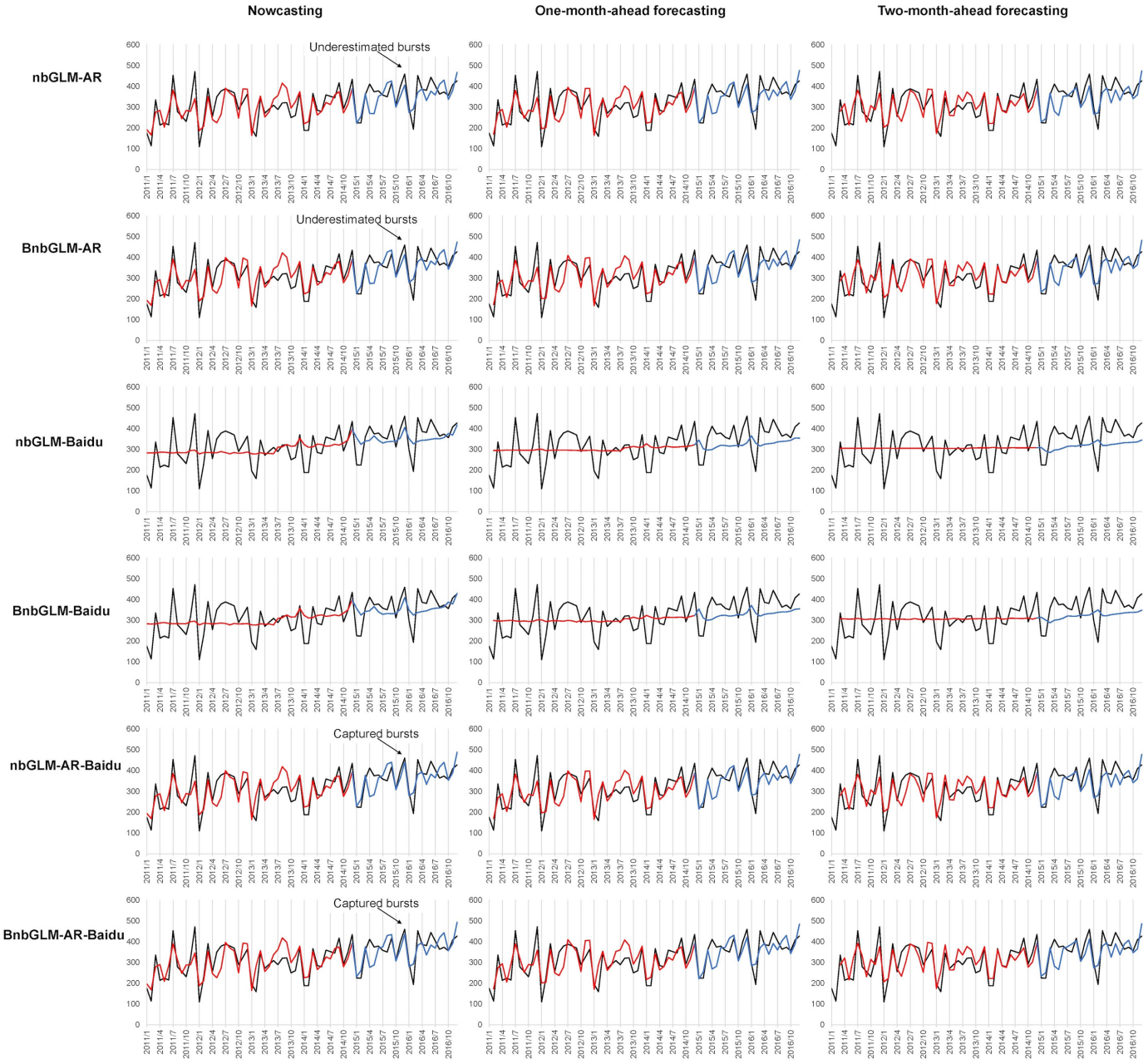

Figure 3 Actual number of new HIV diagnoses and prediction results of the six proposed models (Guangdong province). The black curve represents the actual data of new HIV diagnoses. The red curve represents the fitted values. The blue curve represents the prediction result. BnbGLM-AR, Bayesian negative binomial generalised linear model (BnbGLM) with autoregressive terms; BnbGLM-AR-Baidu, BnbGLM with autoregressive terms and the composite Baidu Search Index; BnbGLM-Baidu, BnbGLM with a variable representing the composite Baidu Search Index; nbGLM-AR, negative binomial generalised linear model (nbGLM) with autoregressive terms; nbGLM-AR-Baidu, nbGLM with autoregressive terms and the composite Baidu Search Index; nbGLM-Baidu, nbGLM with a variable representing the composite Baidu Search Index.

HIV diagnoses in January and February may be due to the fact that most Chinese have a long break (1–2 weeks for adults, and 3–4 weeks for students) to celebrate the Chinese Lunar New Year (usually between late January and late February) with family members in hometown. In addition, many people who suspected themselves to be with HIV may have already done the test in December because of the national campaign. However, this explanation is only our hypothesis. There are many other potential reasons such as the delay of reports and the accessibility of related services. New survey studies and randomised controlled trials are needed to identify the root cause of this phenomenon.

By incorporating search query data, we can accurately predict the number of new HIV diagnoses for the current month before the official statistics are available at the national and provincial levels. At the national level, incorporating search query data could also improve the performance in predicting the number of new HIV diagnoses in the near future. In addition to the benefits of improved

overall accuracy, nbGLM-AR-Baidu and BnbGLM-AR-Baidu could accurately capture the bursts of HIV diagnoses, whereas nbGLM-AR and BnbGLM-AR without the composite search index underestimated the bursts.

CONCLUSION

This research demonstrated the feasibility of using search query data to predict the number of new HIV diagnoses in China at the national and provincial levels. We provide a basis for developing low-cost methods for prediction. The proposed method could be applied to actual HIV surveillance and prevention programmes. The prediction models can serve as a forward-looking feedback loop to help decision makers respond timely through allocating resources. If the prediction results indicate that an outbreak is going on, then decision makers can allocate resources to initiate national campaigns and increase the inventory of test and treatment kits without having to wait for the official statistics. National campaigns should focus on approaching the at-risk web users by providing HIV-related information to social and mass media. This prediction-enabled online approach could destroy the barriers for people accessing HIV-related services, such as testing and treatment, and reduce the risk of new infections in the long run. To fulfil the potential of using internet search data for HIV prevention and surveillance, future research must complement conventional HIV prevention programmes with internet data for timely evidence-based decision making. In addition to internet searches, we will further research on the value of social media data (eg, Microblog, online discussion forums and online health communities) in HIV surveillance.

This research presents several limitations. First, considering the limited access to HIV diagnosis data at the province and city levels, we only validated the model for the entire country and Guangdong province. Second, the number of new HIV diagnoses is only a portion of the HIV incidence. Our future research will focus on using internet search query data to accurately estimate HIV incidence. Third, the surveillance data are subject to sampling bias.[4] For example, the system tends to oversample MSM who are younger and more interested in HIV testing because of continence sampling through the internet or social networks. Moreover, the selection of STI clinic attendees is based on the physicians' judgement of the risk of patients and the workload of physicians. Patients are less likely to be selected when they are seen during peak hours. Fourth, age disaggregation and other detailed demographic information of the users who searched for HIV-related terms are not available. Fifth, this research is based on data in China. The associations between search query data and HIV diagnoses could be different in other cultural context.

Contributors YC and QZ had equal contributions to the work. QZ, XL, and YC designed the study. YC, JZ and QZ collected data and performed the analysis. SY provided intellectual content and feedback on the manuscript. All authors contributed to the writing of the manuscript.

Funding This research is supported in part by The National Natural Science Foundation of China (NSFC) Grant Nos. 71402157 and 71672163, in part by The Theme-Based Research Scheme of the Research Grants Council of Hong Kong Grant No. T32-102/14N, and in part by the National Institutes of Health NIH/NIAID Grant Nos. R01AI127203, 1R01AI132020, and R56AI125105.

Competing interests None declared.

Patient consent Not requried.

Ethics approval Ethics approval has been obtained at City University of Hong Kong.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The Baidu Index data and the official statistics of new HIV diagnoses are freely available online. We also include the data we've used in the supplementary materials.

REFERENCES

1. WHO. AIDS related questions and answers. http://www.who.int/features/qa/71/en/ (accessed 20 Feb 2018).
2. UNAIDS, 2017. UNAIDS Data http://www.unaids.org/en/resources/documents/2017/2017_data_book
3. NCAIDS, NCSTD CC.Update on the AIDS/STD epidemic in China in December, 2016:93.
4. Lin W, Chen S, Seguy N, *et al*. Is the HIV sentinel surveillance system adequate in China? Findings from an evaluation of the national HIV sentinel surveillance system. *Western Pacific Surveillance and Response Journal* 2012;3:61–8.
5. Liu H, Yang H, Li X, *et al*. Men who have sex with men and human immunodeficiency virus/sexually transmitted disease control in China. *Sex Transm Dis* 2006;33:68–76.
6. Ge L, Li D, Li P, *et al*. Population specific sentinel surveillance for HIV infection, syphilis and HCV infection in China, during 2010-2015. *Dis Surveill* 2017;32:111–7.
7. Hall HI, Holtgrave DR, Maulsby C. HIV transmission rates from persons living with HIV who are aware and unaware of their infection. *AIDS* 2012;26:893–6.
8. Ma Y, Yu D, Gu R, *et al*. Analysis of fund inputs for HIV/AIDS prevention and control from 2010 to 2015 in China. *Chinese J AIDS STD* 2016;22:991–3.
9. Wang B, Li X, Stanton B, *et al*. Sexual attitudes, pattern of communication, and sexual behavior among unmarried out-of-school youth in China. *BMC Public Health* 2007;7:189.
10. Song Y, Li X, Zhang L, *et al*. HIV-testing behavior among young migrant men who have sex with men (MSM) in Beijing, China. *AIDS Care* 2011;23:179–86.
11. Zhang C, Li X, Liu Y, *et al*. Emotional, physical and financial burdens of stigma against people living with HIV/AIDS in China. *AIDS Care* 2016;28 :124–31.
12. 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.
13. Ulett KB, Willig JH, Lin HY, *et al*. The therapeutic implications of timely linkage and early retention in HIV care. *AIDS Patient Care STDS* 2009;23:41–9.
14. Mackellar DA, Hou SI, Whalen CC, *et al*. Reasons for not HIV testing, testing intentions, and potential use of an over-the-counter rapid HIV test in an internet sample of men who have sex with men who have never tested for HIV. *Sex Transm Dis* 2011;38:419–28.
15. Liu Y, Sun X, Qian HZ, *et al*. Qualitative assessment of barriers and facilitators of access to HIV testing among men who have sex with men in China. *AIDS Patient Care STDS* 2015;29:481–9.
16. Li J, Gilmour S, Zhang H, *et al*. The epidemiological impact and cost-effectiveness of HIV testing, antiretroviral treatment and harm reduction programs. *AIDS* 2012;26:2069–78.
17. Zhong L, Zhang Q, Li X. Modeling the intervention of HIV transmission across intertwined key populations. *Sci Rep* 2018;8:2432.
18. Zhang Q, Zhong L, Gao S, *et al*. Optimizing hiv interventions for multiplex social networks via partition-based random search. *IEEE Trans Cybern* 2018:1–9.

19. Young SD. A "big data" approach to HIV epidemiology and prevention. *Prev Med* 2015;70:17–18.
20. Polgreen PM, Chen Y, Pennock DM, *et al*. Using internet searches for influenza surveillance. *Clin Infect Dis* 2008;47:1443–8.
21. Liu Y, Lv B, Peng G, *et al*. A preprocessing method of internet search data for prediction improvement. *Proceedings of the data mining and intelligent knowledge management workshop on - DM-IKM' 12*, 2012:1–7.
22. Achrekar H, Gandhe A, Lazarus R, *et al*. Predicting flu trends using twitter data. *2011 IEEE Conference on computer communications workshops, INFOCOM WKSHPS 2011*, 2011:702–7.
23. Xu Q, Gel YR, Ramirez Ramirez LL, *et al*. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PLoS One* 2017;12:e0176690.
24. Lampos V, Miller AC, Crossan S, *et al*. Advances in nowcasting influenza-like illness rates using search query logs. *Sci Rep* 2015;5:12760.
25. 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.
26. Wang S, Paul MJ, Dredze M. Exploring Health Topics in Chinese Social Media : An Analysis of Sina Weibo. *workshops at the twenty-eighth aaai conference on artificial intelligence*, 2014:20–3.
27. Santillana M, Nsoesie EO, Mekaru SR, *et al*. Using clinicians' search query data to monitor influenza epidemics. *Clin Infect Dis* 2014;59:1446–50.
28. Ginsberg J, Mohebbi MH, Patel RS, *et al*. Detecting influenza epidemics using search engine query data. *Nature* 2009;457:1012–4.
29. Yuan Q, Nsoesie EO, Lv B, *et al*. Monitoring influenza epidemics in china with search query from baidu. *PLoS One* 2013;8:e64323.
30. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput Biol* 2014;10:e1003581.
31. Milinovich GJ, Avril SM, Clements AC, *et al*. Using internet search queries for infectious disease surveillance: screening diseases for suitability. *BMC Infect Dis* 2014;14:690.
32. 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.
33. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. *PLoS One* 2009;4:e4378.
34. 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.
35. Gu Y, Chen F, Liu T, *et al*. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. *Sci Rep* 2015;5:12649.
36. Reeves PM. How individuals coping with HIV/AIDS use the Internet. *Health Educ Res* 2001;16:709–19.
37. Swendeman D, Rotheram-Borus MJ. Innovation in sexually transmitted disease and HIV prevention: internet and mobile phone delivery vehicles for global diffusion. *Curr Opin Psychiatry* 2010;23:139–44.
38. Wilkerson JM, Smolenski DJ, Horvath KJ, *et al*. Online and offline sexual health-seeking patterns of HIV-negative men who have sex with men. *AIDS Behav* 2010;14:1362–70.
39. Kalichman SC, Weinhardt L, Benotsch E, *et al*. Internet access and Internet use for health information among people living with HIV-AIDS. *Patient Educ Couns* 2002;46:109–16.
40. Taggart T, Grewe ME, Conserve DF, *et al*. Social Media and HIV: a systematic review of uses of social media in HIV communication. *J Med Internet Res* 2015;17:e248.
41. Kubicek K, Carpineto J, McDavitt B, *et al*. Use and perceptions of the internet for sexual information and partners: a study of young men who have sex with men. *Arch Sex Behav* 2011;40:803–16.
42. Magee JC, Bigelow L, Dehaan S, *et al*. Sexual health information seeking online: A mixed-methods study among lesbian, gay, bisexual, and transgender young people. *Health Educ Behav* 2012;39:276–89.
43. Generous N, Fairchild G, Deshpande A, *et al*. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol* 2014;10:e1003892.
44. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med* 2014;63:112–5.
45. Young SD, Rice E. Online social networking technologies, HIV knowledge, and sexual risk and testing behaviors among homeless youth. *AIDS Behav* 2011;15:253–60.
46. Ayers JW, Althouse BM, Allem JP, *et al*. Seasonality in seeking mental health information on Google. *Am J Prev Med* 2013;44:520–5.
47. Wang HW, Chen DR, Yu HW, *et al*. Forecasting the incidence of dementia and dementia-related outpatient visits with Google trends: evidence from Taiwan. *J Med Internet Res* 2015;17:e264.
48. Yang AC, Tsai SJ, Huang NE, *et al*. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004-2009. *J Affect Disord* 2011;132:179–84.
49. Gunn JF, Lester D. Using google searches on the internet to monitor suicidal behavior. *J Affect Disord* 2013;148:411–2.
50. Ling R, Lee J. Disease Monitoring and Health Campaign Evaluation Using Google Search Activities for HIV and AIDS, Stroke, Colorectal Cancer, and Marijuana Use in Canada: A Retrospective Observational Study. *JMIR Public Health Surveill* 2016;2:e156.
51. 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.
52. Johnson AK, Mehta SD. A comparison of Internet search trends and sexually transmitted infection rates using Google trends. *Sex Transm Dis* 2014;41:61–3.
53. Young SD, Mercer N, Weiss RE, *et al*. Using social media as a tool to predict syphilis. *Prev Med* 2018;109:58–61.
54. CNNIC. *The 39th China statistical report on internet development*. CNNIC: Beijing, 2017.
55. Cao B, Liu C, Durvasula M, *et al*. Social media engagement and hiv testing among men who have sex with men in china: A nationwide cross-sectional survey. *J Med Internet Res* 2017;19:e251.
56. Statcounter. Search Engine Market Share in China. http://gs.statcounter.com/search-engine-market-share/all/china/ (accessed 22 Jan 2018).
57. Li Z, Liu T, Zhu G, *et al*. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China. *PLoS Negl Trop Dis* 2017;11:e0005354.
58. Cameron AC, Trivedi PK. Regression analysis of count data. 1998.
59. Bishop CM. Pattern recognition and machine learning. 2006.
60. Kristoufek L, Moat HS, Preis T. Estimating suicide occurrence statistics using Google trends. *EPJ Data Sci* 2016;5.
61. Wu Z, Sullivan SG, Wang Y, *et al*. Evolution of China's response to HIV/AIDS. *Lancet* 2007;369:679–90.
62. He N, Detels R. The HIV epidemic in China: history, response, and challenge. *Cell Res* 2005;15:825–32.