Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Notifiable infectious disease surveillance with data collected by search engine

Xi-chuan ZHOU[1], Hai-bin SHEN[†‡2]

([1]*Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*)

[†]E-mail: shenhb@yahoo.cn

Received June 25, 2009; Revision accepted Sept. 29, 2009; Crosschecked Jan. 29, 2010

Abstract: Notifiable infectious diseases are a major public health concern in China, causing about five million illnesses and twelve thousand deaths every year. Early detection of disease activity, when followed by a rapid response, can reduce both social and medical impact of the disease. We aim to improve early detection by monitoring health-seeking behavior and disease-related news over the Internet. Specifically, we counted unique search queries submitted to the Baidu search engine in 2008 that contained disease-related search terms. Meanwhile we counted the news articles aggregated by Baidu's robot programs that contained disease-related keywords. We found that the search frequency data and the news count data both have distinct temporal association with disease activity. We adopted a linear model and used searches and news with 1–200-day lead time as explanatory variables to predict the number of infections and deaths attributable to four notifiable infectious diseases, i.e., scarlet fever, dysentery, AIDS, and tuberculosis. With the search frequency data and news count data, our approach can quantitatively estimate up-to-date epidemic trends 10–40 days ahead of the release of Chinese Centers for Disease Control and Prevention (Chinese CDC) reports. This approach may provide an additional tool for notifiable infectious disease surveillance.

Key words: Notifiable infectious diseases, Disease surveillance, Search engine
doi:10.1631/jzus.C0910371 **Document code:** A **CLC number:** TP393.4; R51

1 Introduction

Traditional surveillance systems, including Chinese Centers for Disease Control and Prevention (Chinese CDC), rely on clinical data. A network of sentinel laboratories performs disease test, by counting and classifying pathogens collected from patients, while a network of sentinel physicians reports the number of people diagnosed with notifiable infectious diseases. Chinese CDC publishes the data collected on a monthly basis, typically with a 10-day reporting lag. In order to reduce the lag and detect possible disease outbreaks in an early stage, we aggregate the data collected by the Baidu search engine from the Internet to monitor the current activity of scarlet fever, dysentery, AIDS, and tuberculosis.

Now millions of Chinese people are believed to search online for information about medical problems each year. Searchers include patients and their families and health care professionals (Diaz *et al.*, 2002; Ybarra and Suman, 2006; Bundorf *et al.*, 2006; Fox, 2006). However, the large number of health-related sites has made it difficult to find specific information that is credible and reliable. Thus, Internet search engines (e.g., Baidu, Google, and Yahoo) are now essential for Internet users to find information. In fact, most people searching for medical information use a specific search engine (Fox, 2006). Since a large population of people search online for medical information, the pattern of how and when people search may provide clues or early indications about future concerns and expectations. An analysis of Internet search terms related to jobs and job opportunities has

‡ Corresponding author

produced accurate and useful statistics about the unemployment rate (Ettredge *et al.*, 2005). Similarly, searches for health-related information might also yield useful health statistics.

We examined the Baidu search database for infectious diseases surveillance. With millions of search queries collected in 2008, we counted unique queries associated with infectious-disease-related terms every day, and used the search frequency data as the first data source to predict disease occurrence. Prior to our research, Internet search data have been analyzed for surveillance purpose (Johnson *et al.*, 2004). Polgreen *et al.* (2008) used Yahoo searches associated with 'flu' or 'influenza' for influenza surveillance. Ginsberg *et al.* (2009) also started a research project to estimate influenza epidemic trend through Google search queries. Internet searches for specific cancers were also found to be correlated with estimated incidence and mortality (Cooper *et al.*, 2005). Lately, Wilson and Brownstein (2009) analyzed different Internet-based methods, including the search-engine based method, for disease outbreaks detection.

Despite the strong relation between search frequency and disease activity, people's Web search behavior can be affected by other factors, such as panic or disease-related events. For example, the World AIDS Day may attract healthy people's attention and generate a lot of searches irrelevant with disease occurrence. If we estimate the disease occurrence based on search frequency data alone, the irrelevant searches will decrease the estimation accuracy. Thus, we constructed a second data source with Internet-accessible news to estimate infectious disease trends.

Usually outbreaks of notifiable infectious diseases are heavily reported online. One case in point is the hand, foot and mouth disease (HFMD), which stroke China in early 2008 and reached its peak during May and June. In these two months, many websites published news reports about HFMD. The average number of news collected by Baidu's robot programs exceeded two hundred per day, which is ten times more than that in regular days (more information about Baidu News service can be found in http://www.baidu.com/search/news help.html). Thus, the number of disease-related news available in the Internet reveals important information about infectious disease activity. To approximate the number of disease-related news, we counted the Baidu News containing diseases-related keywords through 2008 and used the news count data as a second data source. Prior to this project, Internet-accessible news data have been aggregated and analyzed for surveillance purpose in the HealthMap project (Brownstein *et al.*, 2008). Brownstein and his colleagues extracted and integrated disease-related reports from sources like Google News with the intention to monitor and visualize emerging infectious diseases.

Though there are more and more papers published on this topic in recent two years, to our knowledge, no paper addressing this issue has been published in China. With the intention to design a syndromic surveillance system for early disease detection in China, we analyzed the patterns in Internet-based time series, i.e., Baidu search frequency and the amount of Baidu News related to infectious diseases.

2 Data sources

In our study, we aim to estimate the up-to-date number of people infected with scarlet fever, dysentery, AIDS, and tuberculosis, as well as the number of deaths attributable to AIDS and tuberculosis. In this section, we introduce the data sources constructed to build the surveillance system.

2.1 Chinese CDC reports

To measure disease occurrence, we used two types of data published by Chinese CDC (http://www.chinacdc.net.cn/n272562/n276018/index.html, accessed on Mar. 20, 2009). The first type of data is based on the number of people infected with scarlet fever, dysentery (including both bacillary dysentery and amebic dysentery), AIDS, and tuberculosis. Each month, the National Disease Reporting System reports the total number of specimens tested and the number that are positive for notifiable infectious diseases. Chinese CDC publishes the infection data on a monthly basis, usually with a 10-day reporting lag. Fig. 1b is an example of scarlet fever infection data for the year of 2008 extracted from Chinese CDCs reports.

The second type of data summarizes monthly mortality attributable to AIDS and tuberculosis. These data are collected and published by Chinese CDC from the Mortality Reporting System.

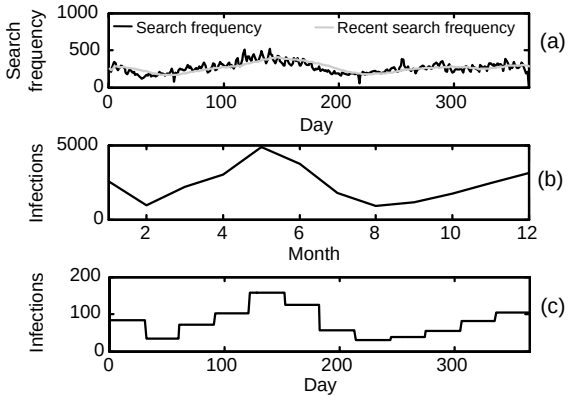

Fig. 1 Comparison of three temporal series associated with scarlet fever infection

(a) Search-frequency data (s_t) and recent-search-frequency data ($s_t^{(l)}$, l=30) associated with keyword 'xinghongre' (meaning scarlet fever); (b) The number of scarlet fever infections published by Chinese CDC on a monthly basis in 2008; (c) Monthly numbers of scarlet fever infections are averaged across days of the reporting month, which forms the disease occurrence measurement (y_t) of scarlet fever infection

Each month, Nationwide Disease Surveillance Points (DSPs) report the total number of death certificates received and the number of deaths that lists a notifiable infectious disease as the underlying and/or contributing cause. Based on these data, we obtained national mortality figures for AIDS and tuberculosis.

To match the date range of our data collected by a search engine, both infection data and mortality data were collected in 2008. Since we aim to achieve the up-to-date estimates, both types of data were averaged across the days of the reporting month. And we used the average number of infections and deaths caused by diseases as two measurements of the disease occurrence. Throughout this paper, we denote by y_t the disease occurrence measurements calculated for day t. Fig. 1c shows an example of y_t, the average daily number of people infected by scarlet fever in 2008.

2.2 Baidu search database

The Baidu search query database consists of over a million most common Web search queries submitted by Chinese Web users. By aggregating the log files in the database, the number of unique searches for selected keywords is computed every day during 2008 (http://index.baidu.com). For disease surveillance, we focus on the searches for Chinese terms of the diseases involved. The disease terms involved are listed in Table 1, along with their associated Chinese pinyins. All Chinese keywords are represented with pinyins throughout this paper.

Table 1 Chinese terms of each infectious disease we studied are used as keywords to construct the recent-search-frequency data and the recent-news-count data

Disease	Chinese name	Chinese pinyin
Scarlet fever	猩红热	xinghongre
Dysentery	痢疾	liji
AIDS	艾滋	aizi
Tuberculosis	肺结核	feijiehe

Polgreen *et al.* (2008) normalized the search frequency by the total number of searches submitted each year, so that the yearly growth of total search will not affect the estimation results. In contrast, the search frequency in our research is not normalized because Baidu supplies only one-year search data.

Here we use scarlet fever as an example to illustrate this data source (Fig. 1). Every day we count the number of unique searches for keyword 'xinghongre' (meaning scarlet fever). Suppose we denote by s_t the resulting search frequency data for day t (Fig. 1a, black). Our system is designed to estimate the epidemic trends through recent-search-frequency data. Thus we calculate the recent-one-day-averaged search-frequency of keyword 'xinghongre' as

$$s_t^{(l)} = \frac{1}{l}\sum_{i=t-l+1}^{t} s_i. \quad (1)$$

In our research, we use the recent-search-frequency data $s_t^{(l)}$ as the first data source to estimate the disease occurrence measurement y_t based on the linear correlation between them. Suppose we denote the $s_t^{(l)}$ series and the disease occurrence measurements calculated every day in 2008 by

$$\boldsymbol{s}^{(l)} = [s_1^{(l)}, s_2^{(l)}, ..., s_N^{(l)}],\ \boldsymbol{y} = [y_1, y_2, ..., y_N]^{\mathrm{T}},\ N = 365.$$

For the four infectious diseases we studied, the correlation coefficients between $\boldsymbol{s}^{(l)}$ and $\boldsymbol{y}$ are positive (p<0.001) for a range of l. More detailed results are listed in Table 2, where the correlation between $\boldsymbol{s}^{(l)}$

and $\boldsymbol{y}$ is denoted by $\rho_s^{(l)}$. As one can see from Table 2, the correlation between scarlet fever infection data and associated search frequency is high ($\max_l \rho_s^{(l)} = 0.879$, $p<0.001$); thus, with a linear method we can estimate the infection trend of scarlet fever using search frequency data. On the other hand, the correlations calculated for AIDS ($\max_l \rho_s^{(l)} = 0.510$, $p<0.001$) and tuberculosis ($\max_l \rho_s^{(l)} = 0.502$, $p<0.001$) are relatively low, suggesting that the problem of irrelevant searches is more serious for these two diseases. To further improve the estimation accuracy, we analyze the disease-related news articles in the Baidu News database and incorporate the recent news count data as a second data source.

Table 2 Correlation between input data series and different disease occurrence measurements

Measurement	$\max_{l=1:200}\rho_s^{(l)}$	$\max_{l=1:200}\rho_n^{(l)}$
Scarlet fever infection	0.879	0.319
Dysentery infection	0.735	0.931
AIDS infection	0.510	0.459
AIDS mortality	0.673	0.550
Tuberculosis infection	0.502	0.789
Tuberculosis mortality	0.586	0.161

$\rho_s^{(l)}$: correlation between $\boldsymbol{s}^{(l)}$ and $\boldsymbol{y}$; $\rho_n^{(l)}$: correlation between $\boldsymbol{n}^{(l)}$ and $\boldsymbol{y}$; l: length of lead time

2.3 Baidu News database

The Baidu News database is generally built using documents collected by Baidu's robot programs. Baidu selects over one thousand important and trustworthy Chinese websites as news sources, including national and local news websites, and networks for government health department, health care organizations, and traditional media companies. Robot programs are used to collect the latest news, about 120 000 to 130 000 articles per day, from selected websites (http://www.baidu.com/search/news_help.html). For disease surveillance, we count daily news containing the terms of diseases considered (Table 1) every day through 2008. Both Baidu search-frequency data and news-count data associated with disease-related keywords are now publicly available from http://index.baidu.com.

Here we use dysentery as an example to illustrate this data source (Fig. 2). Suppose we aggregate the daily news of day t and count the number of articles containing the name of dysentery ('liji'). We denote by n_t the news-count data (Fig. 2b, black). Then we can calculate the average number of news articles for the recent l days that contain 'liji' as

$$n_t^{(l)} = \frac{1}{l}\sum_{i=t-l+1}^{t} n_i. \tag{2}$$

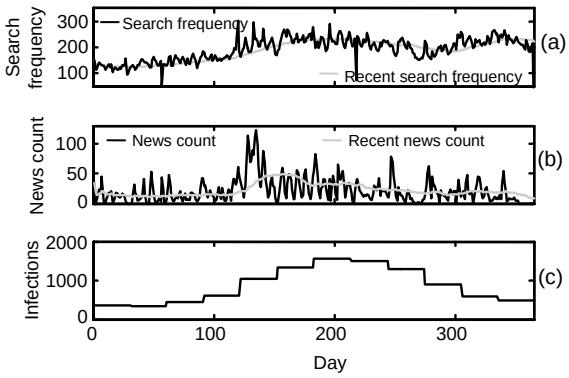

Fig. 2 Comparison of three temporal series associated with dysentery infection

(a) Search-frequency data (s_t) and recent-search-frequency data ($s_t^{(l)}$, l=30) of keyword 'liji' (meaning dysentery); (b) News-count data (n_t) and recent-news-count data ($n_t^{(l)}$, l=30), constructed by counting the Baidu News containing keyword 'liji'; (c) Monthly numbers of dysentery infections are averaged across days of the reporting month, which forms the disease occurrence measurement (y_t) of dysentery infection

Fig. 2c shows that the number of people infected by dysentery increased significantly after May. About 64.72% infections occurred during the dysentery season from May to September in China. Meanwhile, the news-count data n_t obviously increased at the beginning of the dysentery season. Most reports collected as Baidu News were published during the dysentery season as well (63.35%). In our research, the recent-news-count data $n_t^{(l)}$ (Fig. 2b, grey) defined in Eq. (2) are used to estimate disease trends based on its correlation with disease occurrence measurements proposed in Section 2.1. Suppose we denote the temporal recent-news-count series calculated every day in 2008 by

$$\boldsymbol{n}^{(l)} = [s_1^{(l)}, s_2^{(l)}, ..., s_N^{(l)}],\ N = 365.$$

The vector $\boldsymbol{n}^{(l)}$ associated with dysentery is positively correlated to dysentery infection trend $\boldsymbol{y}$ ($\min_l \rho_n^{(l)} = 0.314$, $\max_l \rho_n^{(l)} = 0.931$). Results about recent-news-count data for other diseases are listed in Table 2, where the correlation between $\boldsymbol{n}^{(l)}$ and $\boldsymbol{y}$ is denoted by $\rho_n^{(l)}$. Thus, by carefully choosing the length of lead time, l, we can estimate y_t using the recent-news-count data with a linear model.

Note that with the recent-news-count data we can reduce the influence of searches caused by irrelevant events. Fig. 3 shows AIDS as an example. The number of searches for AIDS ('aizi') was counted every day in 2008 to construct s_t (Fig. 3a, black) and $s_t^{(l)}$ (Fig. 3a, grey).

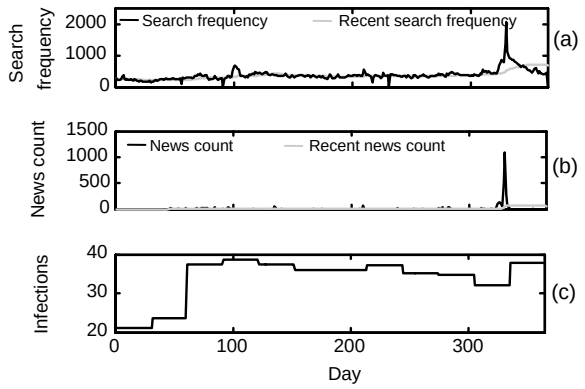

Fig. 3 Comparison of three temporal series associated with AIDS infection

(a) Search-frequency data (s_t) and recent-search-frequency data ($s_t^{(l)}$, l=30) of keyword 'aizi' (meaning AIDS); (b) News-count data (n_t) and recent-news-count data ($n_t^{(l)}$, l=30), constructed by counting the Baidu News containing keyword 'aizi'; (c) Monthly numbers of AIDS infections are averaged across days of the reporting month, which forms the disease occurrence measurement (y_t) of AIDS infection

The number of daily Baidu News containing 'aizi' was counted every day to construct n_t (Fig. 3b, black) and $n_t^{(l)}$ (Fig. 3b, grey). One can see that the number of AIDS-related news increased remarkably around the World AIDS Day on December 1. Meanwhile, there was an increase in searches around December 1. Fig. 3c shows that these searches have hardly any relation to AIDS occurrence. Actually most keywords related to AIDS were heavily searched around the World AIDS Day. If the estimation is based on the search frequency data alone, it will result in a false alarm and decrease the estimation accuracy. But by the combination of search frequency data and news count data, the effect of disease-related events can be reduced.

3 Estimation model

In summary, by analyzing the data collected from the Baidu search database and the Baidu News database in 2008, we constructed the recent-search-frequency data ($s_t^{(l)}$) and the recent-news-count data ($n_t^{(l)}$) associated with scarlet fever, dysentery, AIDS, and tuberculosis. Then at day t, we estimated the disease occurrence measurements proposed in Section 2.1 using the following linear model:

$$e_t = u_1 s_t^{(l)} + u_2 n_t^{(l)} + \varepsilon, \tag{3}$$

where u_1 and u_2 represent the weights across different sources. Obviously, Eq. (3) is a standard linear model; thus, the parameters can be calculated using the traditional linear regression analysis method. No constraint functions for the weights are included in Eq. (3) because firstly, we intended to compare our results with those of (Ginsberg *et al.*, 2009; Polgreen *et al.*, 2008), in both of which linear models were applied and no constraint functions were used, and secondly, a negative weight, such as u_2<0, probably indicates that some irrelevant events have been heavily reported over the Internet (such as the case of AIDS). And the negative weight could neutralize the increase of searches caused by news reports. However, as future work, we are considering designing more complex models, e.g., nonlinear or multiple-source models, to better fit the scenario.

By comparing the correlation between resulting estimates and the disease occurrence measurements, we can select the best value of l. Table 3 lists the optimal parameters and results calculated, where $\rho_t^{(l)}$ and $\rho_v^{(l)}$ represent the correlations between model estimates and disease occurrence measurements $\boldsymbol{y}$ over training and validation points respectively, $\rho_s^{(l)}$ represents the correlation coefficient between recent-search-frequency data series $\boldsymbol{s}^{(l)}$ and disease occurrence measurements, and $\rho_n^{(l)}$ represents the correlation between recent-news-count data series $\boldsymbol{n}^{(l)}$

Table 3 Correlation between input data series and different disease occurrence measurements

Measurement	l (d)	u_1	u_2	$\rho_s^{(l)}$	$\rho_n^{(l)}$	$\rho_t^{(l)}$	$\rho_v^{(l)}$
Scarlet fever infection	29	0.604	4.425	0.875	−0.124	0.897	0.885
Dysentery infection	107	0.134	64.312	0.656	0.930	0.921	0.913
AIDS infection	49	0.031	−0.256	0.468	0.234	0.601	0.594
AIDS mortality	41	0.065	−0.210	0.621	0.337	0.752	0.741
Tuberculosis infection	145	5.742	102.643	0.101	0.779	0.763	0.756
Tuberculosis mortality	49	0.044	0.017	0.581	0.156	0.634	0.609

$\rho_s^{(l)}$: correlation between $\boldsymbol{s}^{(l)}$ and $\boldsymbol{y}$; $\rho_n^{(l)}$: correlation between $\boldsymbol{n}^{(l)}$ and $\boldsymbol{y}$; l: length of lead time

and disease occurrence measurements. For all diseases we studied, $\rho_t^{(l)}$ and $\rho_v^{(l)}$ are larger than $\rho_s^{(l)}$ and $\rho_n^{(l)}$, which indicates that the estimation results are improved after combining the recent-search-frequency data and recent-news-count data. The source weight values u_1 and u_2 in Eq. (3) vary with $\rho_s^{(l)}$ and $\rho_n^{(l)}$. Specifically, our method puts a larger weight on the source of data with a higher correlation to disease occurrence measurements.

Note that due to searches generated on the World AIDS Day, the correlation between input data sources and the AIDS occurrence measurements are relatively low ($\max_l \rho_s^{(l)} = 0.510$ and $\max_l \rho_n^{(l)} = 0.459$ for infection measurement; $\max_l \rho_s^{(l)} = 0.673$ and $\max_l \rho_n^{(l)} = 0.550$ for mortality measurement). The influence of irrelevant searches is reduced by model 3 with the optimal u_1 and u_2 calculated, and the correlation between final estimates and disease occurrence measurements over the validation points rises up to 0.637 for infection estimation and 0.788 for mortality estimation.

4 Infection measurement estimation

In this section, we summarize the infection estimation results of scarlet fever, dysentery, AIDS, and tuberculosis. In the first experiment, all data were used and no temporal lag was considered. Firstly, we randomly split the 365-point set into three sets, a training set (219 points), an evaluation set (73 points), and a validation set (73 points). Secondly, we fit the models with different values of l using the training set. This resulted in the estimation of u_1 and u_2. Thirdly, we applied the models to the evaluation set to choose l. After u_1, u_2, and l were calculated, the model was tested over the validation set. The correlation coefficient between the estimates and the disease infection measurement data were computed repeatedly over both training ($\rho_t^{(l)}$) and validation points ($\rho_v^{(l)}$). The values of $\rho_t^{(l)}$ and $\rho_v^{(l)}$ depend on the value of lead time parameter l (Fig. 5). To determine an appropriate l, we examined 200 possible values and compared the correlation for each model. The parameters resulting in the highest correlation over validation points were selected and listed in Table 3. The infection estimates for each disease with the optimal l calculated are presented in Fig. 4.

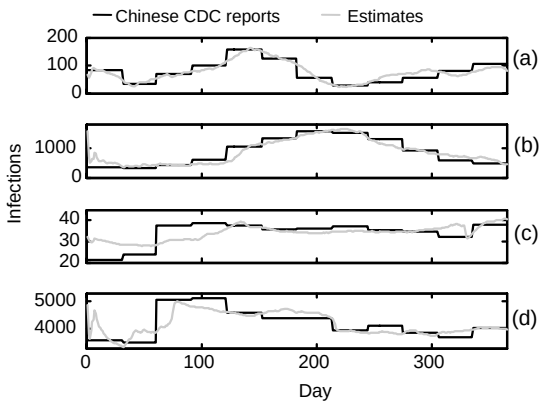

Fig. 4 A comparison of infection trend estimates for (a) scarlet fever (l=29), (b) dysentery (l=107), (c) AIDS (l=49) and (d) tuberculosis (l=145) against the disease occurrence measurement of infections extracted from Chinese CDC reports, including points over which the model was trained and validated

Note that our model seems to work better for dysentery and scarlet fever than for tuberculosis and AIDS. This is possibly caused by irrelevant events, which are the most significant for the case of AIDS. Due to the pitch of search-frequency time series

occurring on the World AIDS Day (Fig. 3), the correlation between the search frequency and the occurrence of AIDS was low (correlation equaled 0.468 for the infection data and 0.621 for the mortality data), so was the situation of news count series (correlation equaled 0.234 for the infection data and 0.337 for the mortality data). Though the estimation results had higher correlation with the occurrence of AIDS, compared with scarlet fever and dysentery, the correlation was still lower.

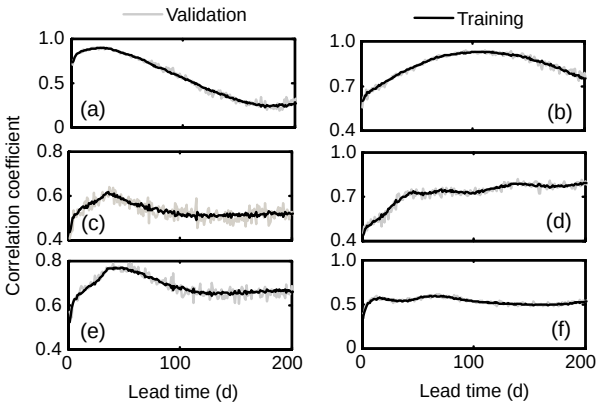

Fig. 5 Change of correlation coefficients between model estimates and disease occurrence measurements with respect to different values of lead time, including results over both training and validation points

(a) Scarlet fever infection; (b) Dysentery infection; (c) AIDS infection; (d) Tuberculosis infection; (e) AIDS morbidity; (f) Tuberculosis morbidity

The model can also be trained online, where only past data were accessible. In this experiment, the coefficients of Eq. (3) (u_1, u_2, and ε) were updated every month, when new data were published by Chinese CDC. Since the recent-search-frequency data and the recent-news-count data can be updated every day, up-to-date estimates of epidemic trends can be achieved. Here we used scarlet fever as an example to illustrate the temporal lag between model estimates and the data published (Fig. 6). We can detect the increase and the decrease of the epidemic trend 10–40 days before Chinese CDC reports are released. Timely infectious disease estimates may enable public health officials and health professionals to respond better to emerging epidemics. If the system detected a sharp increase at an early stage, it may be possible to focus additional resources on providing extra vaccine capacity or raising media awareness as necessary.

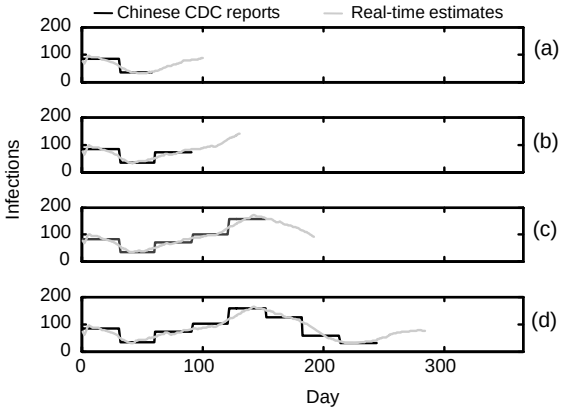

Fig. 6 Comparison of real-time estimates and the disease occurrence measurement of scarlet fever infection extracted from Chinese CDC reports, showing data available at (a) April 1, 2008, (b) May 1, 2008, (c) July 1, 2008, and (d) October 1, 2008

During April we detected a sharp increase; similarly, on July 1 our model indicated that the peak of epidemic trend had been reached during May, with sharp declines in July and August. Both results were later confirmed by Chinese CDC reports

5 Mortality measurement estimation

AIDS and tuberculosis are the two most lethal illnesses among all notifiable infectious diseases. They are attributable to over 50% of deaths reported by Chinese CDC each year. In this section, we summarize the mortality estimation results of AIDS and tuberculosis. In this experiment, all data were used and no temporal lag was considered. Similar to infection estimation, 219 points out of 365 points were used to fit the model for parameters u_1 and u_2, which was later evaluated over the remaining 73 points. The correlation coefficient between the estimates and the disease mortality measurement data were computed repeatedly over both training ($\rho_t^{(l)}$) and validation points ($\rho_v^{(l)}$). The resulting $\rho_t^{(l)}$ and $\rho_v^{(l)}$ varied with l (Fig. 5). To determine an appropriate l, we examined 200 possible values and compared the correlation for each model. The parameters resulting in the highest correlation over validation points were selected and listed in Table 3. The mortality estimates with the optimal l are presented in Fig. 7.

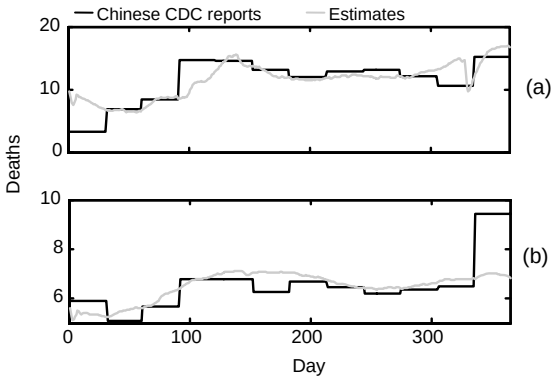

Fig. 7 A comparison of mortality trends estimates for (a) AIDS (*l*=41) and (b) tuberculosis (*l*=49) against the disease occurrence measurement of deaths extracted from Chinese CDC reports, including points over which the model was trained and validated

6 Conclusion

The timing and size of epidemic outbreaks vary from disease to disease, complicating efforts to produce reliable and timely surveillance results of different infectious diseases; however, we found that both disease-related search frequency and disease-related news count have a distinct temporal association with disease activity. With the Baidu search database and the Baidu News database, we constructed a timely epidemic monitoring system for detecting notifiable infectious diseases, including scarlet fever, dysentery, AIDS, and tuberculosis. Traditional disease surveillance networks publish the disease occurrence data on a weekly or monthly basis, usually with weeks' reporting lag. In contrast, our system can publish the up-to-date surveillance results, which are about 10–40 days ahead of the release of Chinese CDC reports. With a lead time of a few weeks, public health officials could mount a more effective early response.

If future results were consistent with our findings, data collected by a search engine may provide an important and cost-effective supplement to traditional disease-surveillance systems. Search terms classified by different geographic regions may provide even more useful information. For example, we fit linear models with search query data from three provinces of China and found that disease-related search terms are statistically and significantly related to disease infection. Models performed better for some regions than for others, suggesting that events in some regions may increase searches in other regions. Additional work is needed to examine the spatial relationship between Internet searches and the geographic spread of different infectious diseases. However, because infection and mortality data are not uniformly reported at the province level, local surveillance results are not included in this article.

In addition to data collected by search engines, data gleaned from website hits or searches on specific websites may also provide useful information about disease activity. For example, the number of articles retrieved on 'Healthlink', a consumer health information website maintained at the Medical College of Wisconsin, was found to correlate with influenza activity (Johnson *et al*., 2004). Therefore, searches for specific diseases on high-traffic websites may provide important time-series data, because such data capture the number of people who are investigating the activity of a specific disease.

References

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D., 2008. Surveillance sans frontiers: Internet-based emerging infectious disease intelligence and the Healthmap project. *PLoS Med.*, **5**(7):e151. [doi:10.1371/journal.pmed.0050151]

Bundorf, M.K., Wagner, T.H., Singer, S.J., Baker, L.C., 2006. Who searches the Internet for health information? *Health Serv. Res.*, **41**:819-836. [doi:10.1111/j.1475-6773.2006.00510.x]

Cooper, C.P., Mallon, K.P., Leadbetter, S., Pollack, L.A., Peipins, L.A., 2005. Cancer Internet search activity on a major search engine, United States 2001-2003. *J. Med. Internet Res.*, **7**(3):e36. [doi:10.2196/jmir.7.3.e36]

Diaz, J.A., Griffith, R.A., Ng, J.J., Reinert, S.E., Friedmann, P.D., Moulton, A.W., 2002. Patients' use of the Internet for medical information. *J. Gener. Intern. Med.*, **17**(3):180-185. [doi:10.1046/j.1525-1497.2002.10603.x]

Ettredge, M., Gerdes, J., Karuga, G., 2005. Using Web-based search data to predict macroeconomic statistics. *Commun. ACM*, **48**:87-92. [doi:10.1145/1096000.1096010]

Fox, S., 2006. Pew Internet and American Life Project. Online Health Search. Available from http://www.pewinternet.org/PPF/r/190/report display.asp [Accessed on Apr. 25, 2008].

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. *Nature*, **457**(7232):1012-1014. [doi:10.1038/nature07634]

Johnson, H.A., Wagner, M.M., Hogan, W.R., Chapman, W., Olszewski, R.T., Dowling, J., Barnas, G., 2004. Analysis of Web access logs for surveillance of influenza. *Stud. Health Technol. Inform.*, **107**:1202-1208.

Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, D., 2008. Using Internet searches for influenza surveillance. *Clin. Infect. Dis.*, **47**(11):1443-1448. [doi:10.1086/593098]

Wilson, K., Brownstein, J.S., 2009. Early detection of disease outbreaks using the Internet. *CMAJ*, **180**(8). [doi:10.1503/cmaj.090215]

Ybarra, M.L., Suman, M., 2006. Help seeking behavior and the Internet: a national survey. *Int. J. Med. Inform.*, **75**(1):29-41. [doi:10.1016/j.ijmedinf.2005.07.029]