CHAPTER 12

Early Warning Practice Using Internet-Based Data

Wenxiao Tu, Lianmei Jin, Daxin Ni
Chinese Center for Disease Control and Prevention, Beijing, China

Traditional surveillance is defined as systematic and continuous collection, analysis, and interpretation of the predefined and representative data (e.g., notifiable report, diagnosis/treatment registration, and death registration, etc.), closely integrated with the timely and coherent dissemination of the results and assessment to those who have the right to know so that action can be taken. On the Internet, there is a lot of open access information related to infectious disease, i.e., official announcement, media news, bulletin boards, blogs, and other social media, which could be an important data source of nontraditional surveillance. With different, surveillance using Internet-based data has three characteristics: (1) The data amount is so huge and scattered, which requires proper data retrieving and screening methods to improve the timeliness of early detection for infectious disease. And the data sources generally have poor stability and representation. (2) When establishing an Internet-based surveillance system, it should be an event-based surveillance system, as it is nearly impossible to apply an indicator-based surveillance. (3) Its purpose would focus on the early detection or getting the whole situation awareness of a particular event, which might be an important supplement to the traditional surveillance. For example, when an infectious disease outbreak or epidemic takes place, in addition to notifiable infectious disease case surveillance or event-based surveillance and other traditional surveillance approaches that allow us to understand the situation, we can also identify epidemic situation that is missed or lagged by traditional surveillance system, through media reports on the Internet, official announcement, agency or personal social communication media, forums, etc., and learn about the people's risk perception and response to the event, and the response of international peers and other agencies, in order to gain a more comprehensive understanding of the situations. These characteristics of surveillance using Internet-based data have important implications for early warning of infectious diseases.

Since the 2003 SARS epidemic, China became aware of the importance of information from nondedicated report sources, particularly surveillance using Internet-based data, in addition to the strengthening infectious disease

Early Warning for Infectious Disease Outbreak. http://dx.doi.org/10.1016/B978-0-12-812343-0.00012-6

surveillance system, the Public Health Emergency Event Surveillance System, and other notifiable report surveillance systems. Explorations were made gradually. Following 2008, this has become a routine part of work for CDCs at all levels. Generally speaking, the working model can be briefly described as: retrieve and filter Internet-based information on infectious diseases and public health emergency events; professional staff screen and assess the information; propose recommendations; and take necessary actions. This chapter mainly describes the practices of China CDC in utilizing Internet-based information to facilitate infectious disease surveillance and early warning.

12.1 INTRODUCTION OF SURVEILLANCE USING INTERNET-BASED DATA

Surveillance using Internet-based data was practiced as early as the 1990s. For example, ProMED (the Program for Monitoring Emerging Diseases) is an Internet-based reporting system dedicated to rapid global dissemination of information on outbreaks of infectious diseases and acute exposures to toxins that affect human health (Madoff, 2004), including those in animals and in plants grown for food or animal feed. ProMED started in 1994 and currently (as of February 2017) reaches over 70,000 subscribers in at least 185 countries. It was one of the earliest electronic surveillance systems using Internet-based data to detect the early signals of SARS at the beginning of the outbreak, before the official report.

The Global Public Health Intelligence Network (GPHIN), developed by Health Canada in collaboration with World Health Organization, is an Internet-based multilingual early-warning tool that continuously searches global media sources such as news wires and websites to identify information about disease outbreaks and other events of potential international public health concern (Public Health Agency of Canada, 2004). The GPHIN is one of the most important sources of informal information related to outbreaks. More than 60% (until July 2015) of the initial outbreak reports come from unofficial informal sources, including sources other than the electronic media, which require verification (WHO CSR, 2000).

Surveillance systems using Internet-based data with the working principle similar to ProMED or GHPIN include Threat Tracking Tool (TTT) of European Centre for Disease Prevention and Control (ECDC), HealthMap, etc. Some other systems utilize big data on the Internet to analyze and detect infectious diseases signals on possible aberrations and issue early warning. For example, Google Flu Trends (GFT) leverages search data of Google users on influenza-related symptoms, analyzes the influenza surveillance epidemic, develops models, and triggers early warning according to the signals on possible rise of influenza epidemic as revealed by search data from Internet users (Ginsberg et al., 2009).

12.2 NECESSITY OF SURVEILLANCE USING INTERNET-BASED DATA IN CHINA

In 2002–2003 when China did not identify the SARS epidemic, some other international public health agencies had detected clues on SARS through surveillance using Internet-based data. This low cost and high effective surveillance model was later acknowledged by some Chinese public health professionals, and they attempted to apply the model in practice. In the commonly applied early model, surveillance staff manually searched through media reports on mainstream newspapers, radio and TV on a weekly (or monthly) basis, detected possible abnormal signals, and drafted media surveillance reports which were disseminated to public health professionals and agencies. The 2003 SARS epidemic revealed the overall weakness of Chinese public health system. After that event improvement of traditional surveillance systems were well done; however, few efforts were made to explore and develop infectious disease/emergency public health event-based surveillance by leveraging Internet-based information. The early surveillance using Internet-based data model featured manual searches that were not systematic and long cycles, and could not handle the huge amount of Internet-based information and frequent updating, and therefore tended to miss huge quantity of abnormal signals.

Before the incident of milk powder contaminated by melamine in 2008 was officially confirmed and reported, there were already clues on the event on the Internet. For example, in the incident of poisonous pet foods that killed cats and dogs in the United States in 2007, the FDA of the United States found that melamine was detected in some wheat gluten and rice powder imported from two companies in Jiangsu and Shandong, China, and preliminarily concluded that melamine was the cause of disease. On May 20 and 21, 2008, an Internet user disclosed the problem with Sanlu milk powder he purchased from a superstore in Taishun County, Jiangsu Province in November 2007, with the title: “Can this milk powder be used for emergency relief?!” The milk powder resulted in abnormal urine of his daughter. In June 2008, on an official website of Chinese governmental authority, there were complaints about suspected contamination of milk powder. The above information was present on the Internet, but was beyond the coverage of Chinese traditional notifiable surveillance systems, and was not detected by additional media surveillance carried out by early public health professional agencies.

The incident of melamine-contaminated milk powder in 2008 further strengthened the determination of Chinese public health sector in fully developing surveillance based on Internet information. After the 2008 Olympic Games, China CDC arranged public health surveillance staff to access the Internet for information on public health on a daily basis, and incorporate valuable information into daily report through rapid screening; the risk assessment group consisting of more professional staff members would then analyze and make conclusions, aiming to identify public health event signals and bridge the gaps in traditional surveillance systems. In December 2008, the Chinese Ministry of Health (now the National Health and Family Planning Commission) released the Standards for Disease Prevention and Control Performance Evaluation, in which provincial, prefectural and county level CDCs were required to search international and domestic disease information. From then on, public health surveillance using Internet-based information has become an essential task of CDCs at all levels in China.

12.3 BASIS OF SURVEILLANCE USING INTERNET-BASED DATA IN CHINA

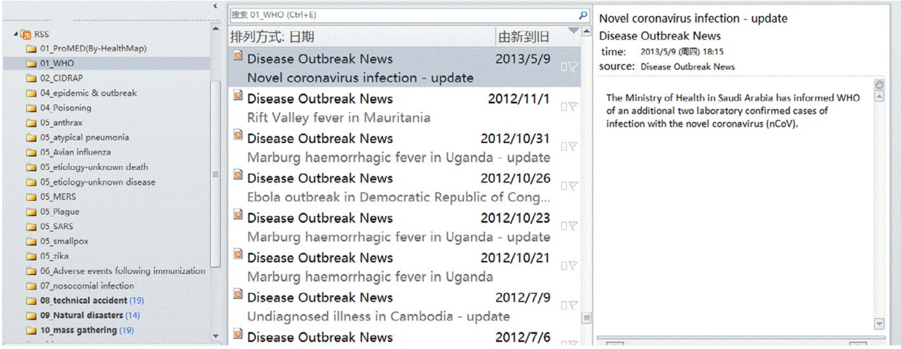
According to the *36th China Internet Development Statistical Report* (China Internet Information Center, 2015), in the past years, the number of Internet users in China has been on the rise. By June 2015, the number hit 668 million, and the Internet penetration rate reached 48.8%; Chinese domain names reached 22.31 million, and the number of websites reached 3,570,000. There is sound basis for the sources of information for public health surveillance using Internet-based information in China.

Of course, the Internet coverage in China is still limited to some extent: By June 2015, rural Internet users in China (those who primarily reside in rural areas in the last 6 months) amounted to 186 million, accounting for 27.9%; over the past years; however, the increase in Chinese Internet users continued to shrink, and it became harder to transform non-Internet users. Besides, the disparity between developed provinces and under-developed provinces is still significant, which means that Internet development in less developed provinces should be promoted for a long period of time. This also suggests that Chinese public health surveillance professionals should take into account the regional and urban-rural disparities while developing surveillance using Internet-based data.

12.4 IMPLEMENTATION OF SURVEILLANCE USING INTERNET-BASED DATA IN CHINA

Working model: Take the work of China CDC as an example, the current surveillance using Internet-based data is based on the "7/365" surveillance model, i.e., designated surveillance staff carry out search and analysis of Internet-based information 7 days a week for 365 days per year.

Sources of information: Including information from our independent development information searching and data mining system, and subscription to Internet-based data surveillance reports issued by other public health agencies or authorities (Fig. 12.1). Major public health information at international level is mainly collected from reports or websites of operational international surveillance systems using Internet-based data, e.g., ProMED, CIDRAP, GPHIN (access via subscription service), etc. Major sources of important Chinese public health



Disease Outbreak News 2013/5/9
Novel coronavirus infection - update

Disease Outbreak News 2012/11/1
Rift Valley fever in Mauritania

Disease Outbreak News 2012/10/31
Marburg haemorrhagic fever in Uganda - update

Disease Outbreak News 2012/10/26
Ebola outbreak in Democratic Republic of Cong...

Disease Outbreak News 2012/10/23
Marburg haemorrhagic fever in Uganda - update

Disease Outbreak News 2012/10/21
Marburg haemorrhagic fever in Uganda

Disease Outbreak News 2012/7/9
Undiagnosed illness in Cambodia - update

Disease Outbreak News 2012/7/6

Novel coronavirus infection - update
Disease Outbreak News
time: 2013/5/9 (周四) 18:15
source: Disease Outbreak News

The Ministry of Health in Saudi Arabia has informed WHO of an additional two laboratory confirmed cases of infection with the novel coronavirus (nCoV).

FIG. 12.1

Really Simple Syndication subscriptions of China CDC for developing surveillance using Internet-based data.

information include: the network monitoring system of public opinion in public health developed by China CDC; keyword subscriptions enabled by Baidu (a search engine) Really Simple Syndication (RSS), such as "outbreak, epidemic, intoxication, disease or death of unknown cause, emerging or reemerging infectious diseases, adverse reactions of vaccines, nosocomial infection, environment pollution or radiation event, various natural disasters, and major mass activities" (in Chinese); and media surveillance daily newsletters issued by other Chinese agencies, e.g., the ad hoc daily report on health edited and released by Health Daily, daily report on food safety edited and issued by China National Center for Food Safety Risk Assessment, and media surveillance report on disease prevention and control released by Guangdong CDC, etc.

Specifically, the network monitoring system of information and public opinion about public health developed by China CDC (Guo et al., 2011) can define metadata criteria for retrieving information from official announcements, media news, forums, blogs, and other social media sources, and collect data via spider directional grabbing; perform text conversion with the collected data (exclude links, advertisements, and other garbage information), search for duplication via text comparison and classify the data, and then store them in database. Then, analysis program is used for automatic classification, grouping, correlations analysis, hotspot analysis and sensitive data analysis and other intelligent processing, and news or messages with potential public health implications will be initially extracted; preliminary statistical and geographical distribution diagrams/tables are then generated, and subscription, push and briefing services are provided to facilitate further manual screening and utilization.

An important precondition for computer intelligent information retrieval and text analysis is to create a robust key word database. Education background, linguistic ability, age and region, and other factors can affect the habit of Internet information distributors in using texts. To avoid major missing of information, while preliminarily incorporating and screening certain type of events, all descriptive words for the type of events should be included whenever possible. Take infectious atypical pneumonia (SARS) as an example: Different people may use different words to describe it with online texts of different types, such as "infectious atypical pneumonia, severe acute respiratory syndrome, SARS…" If not all of these words are searched for, important information on SARS presented on certain web pages may be missed. Nevertheless, if all these words are used, large amount of irrelevant information may be included following preliminary screening, resulting in many white noises and reduction in the efficiency of manual screening. Therefore, the key word database should not only include the complete list of words to be searched for, but also define the corresponding key words for exclusion, in order to minimize white noises. Of course, reduction in white noise relies on in-depth text analytical technique, such as accurately extracting metadata from each piece of information (covering the event type, number of cases and deaths, and the location, etc.), establishing comprehensive scoring system according to metadata, and defining the appropriate thresholds for inclusion and exclusion, thus enabling the ultimate system to automatically report information that comes with practical implications.

Information processing procedure: On a daily basis, the staff member responsible for surveillance daily reporting rapidly screen information preliminarily from various sources according to the predefined principles, incorporating information eligible for surveillance reporting using Internet-based data into daily report with the specified format. The daily report for each event covers the event title, source of information, location, onset-date and number of cases and deaths, and brief description of the event (e.g., the process of event's initiation and development, and current situation, etc.). On the basis of the daily report, the risk assessment team further rapidly screens and assesses the information, determining whether or not to follow up with the event, verify or response to it (Fig. 12.2).

Specifically, the daily surveillance information screening staff determines whether or not to incorporate the collected media information into daily report according to the criteria below. Attentions should also be paid to technical incidents/disasters, natural disasters, and mass gatherings, because it is necessary to check whether these events could lead to outbreaks of particular infectious diseases.

1. If the information is related to public health emergency event and meets any one of the following criteria, it can be incorporated into daily surveillance report:
   (a) Severity of disease: (i) high pathogen infectivity (e.g., measles, pertussis, diphtheria, poliomyelitis, etc.); (ii) high mortality (e.g., SARS, smallpox, avian H5N1/H7N9 influenza, plague, pulmonary anthrax, Ebola virus disease, Marburg fever, yellow fever, etc.); (iii) high proportion of hospitalization or serious cases; and (iv) diseases without effective treatment or prevention measures.
   (b) Unusual or unexpected disease mode or pattern: (i) the number of cases and deaths exceed historical level; (ii) abnormal distribution of affected

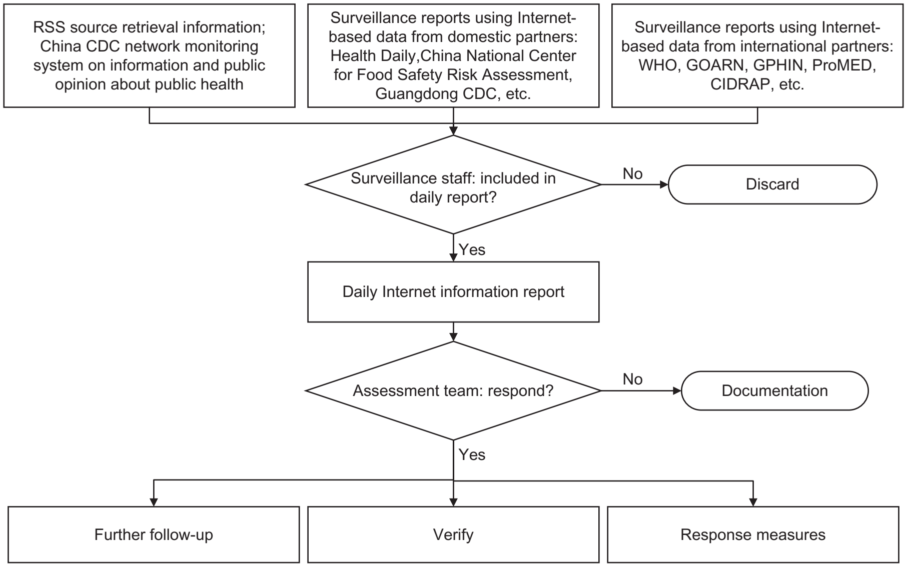

RSS source retrieval information; China CDC network monitoring system on information and public opinion about public health
Surveillance reports using Internet-based data from domestic partners: Health Daily,China National Center for Food Safety Risk Assessment, Guangdong CDC, etc.
Surveillance reports using Internet-based data from international partners: WHO, GOARN, GPHIN, ProMED, CIDRAP, etc.
Surveillance staff: included in daily report?
No
Discard
Yes
Daily Internet information report
Assessment team: respond?
No
Documentation
Yes
Further follow-up
Verify
Response measures

FIG. 12.2

Procedure of China CDC for screening, assessment, and response for surveillance using Internet-based data.

areas, seasons and populations; and (iii) abnormal infectivity and pathogenicity (e.g., abnormal increase in influenza mortality, etc.).

(c) Emerging or reemerging infectious diseases: (i) emerging infectious disease; or (ii) reemergence of infectious diseases which were assumed to have been eliminated or eradicated (e.g., smallpox, poliomyelitis, etc.).

(d) Restriction of travel or trade: (i) the event resulted in warning of tourism in local area; and/or (ii) the event resulted in constraint upon commodity trade.

(e) Special background of event: (i) the event took place in densely populated area; (ii) natural disaster, technical incident/disaster or other events affecting normal social functions took place immediately prior to this event; and (iii) the event took place during mass gatherings or activities.

(f) Media attentions: When any of the following conditions is met, it is deemed media attentions are paid to the event: (i) the event was first reported in internal report channel and there was no official release, but the media has begun to report the event; (ii) the event was released in publicly, while the mainstream media reported or reprinted the event in large number; or (iii) the event took place overseas, and there is already media report in Chinese.

(g) Likelihood of further spread or exaggeration: primarily consider the possibility of further expansion of affected area of the event, or increase in the number of cases or deaths.

(h) Progress of researches that bear significant implications for infectious disease prevention and control.

2. If the event coincides with natural disaster or technical incident/disaster, and meets any of the following conditions, it should be incorporated in daily report:
   (a) Number of people affected: the number of people requiring emergency care or immediate displacement as a result of the emergency event. While assessing this, agencies at different levels should define appropriate thresholds according to the volume of medical surge and displacement victims.

   Preliminarily recommended thresholds on medical surge: (1) at national level, medical surge > capacities of two Class 3 hospitals (the highest level hospital in China, with >500 beds); (2) at provincial level: medical surge > capacities of one Class 3 hospital; (3) at municipal level, medical surge > capacities of one Class 2 hospital; (4) at county level: medical surge > capacities of one Class 1 hospital.

   Preliminarily recommended thresholds on the volume of displacement victims: (1) at national level, >10,000 persons; (2) at provincial level, >1000 people; (3) at municipal level, >500 people; (4) at county level, > 100 people.

   (b) Level of impact: Whether the emergency event negatively affected the normal operation of local essential infrastructure and local normal social order.

   (c) Special background of event: Whether there was prevalence of certain infectious disease during the natural disaster, technical incident/disaster and social security incident, and whether there were multiple emergency events at the same time.

   (d) Likelihood of further exaggeration: Primarily consider whether the emergency event persists and whether the emergency response capacity of local area is adequate.
3. If the event coincides with mass gatherings, and meets any of the following conditions, it should be incorporated in daily report:
   (a) Scale of conference: (i) number of participants ≥1000 and (ii) participants come from multiple provinces of the country (≥10 provinces) or multiple countries (≥3 countries), or the duration of the conference ≥1 week.

   (b) Special background: (i) prevalence of certain infectious disease at the site of the conference or the source areas of the participants during the event; (ii) large-scale food poisoning at the site of the conference during the event.

   (c) Political concerns: Whether administrative authorities at higher levels expressed particular concerns.

The risk assessment team responsible for daily information screening and assessment should follow these criteria: whether the impact of the event upon public health is significant, or the event may probably further spread or exaggerate, resulting in major public health hazard. Meanwhile, the team should, depending on the reliability of the source of information, recommend further follow-up tracking, request the local CDCs in whose territory the event is allegedly occurring to verify the information, or propose China CDC to take actions as appropriate.

12.5 OUTCOME OF SURVEILLANCE USING INTERNET-BASED DATA IN CHINA

According to the preliminary self-evaluation by China CDC, the development of surveillance using Internet-based data (also including available reports from other partners) can play a critical role for bridging the gaps in traditional surveillance:

Firstly, more rapidly gain a more comprehensive understanding of the infectious disease epidemic around the world, and send early warning to health authorities, in order to enable preparedness against import of cases. During the human infected swine influenza epidemic in Mexico in 2009 [later evolved into influenza A (H1N1) pandemic], EHEC O104:H4 outbreak in Germany in 2011, Middle East novel coronavirus epidemic in 2012 (later renamed "MERS"), and West Africa Ebola virus disease in 2014, and other infectious diseases that aroused international concerns, surveillance using Internet-based data, compared to other sources of information shared by WHO and other international public health authorities, could more rapidly provide epidemic information. In addition, through initiative retrieval of relevant information from the Internet, more comprehensive information about the situation could be obtained, which were useful for assessing the risk of importation. Two case studies on MERS and Ebola virus disease are described below.

(1) *MERS*: On September 24, 2012, through routine surveillance using Internet-based data, China CDC found a piece of information as following: on September 23, 2012 the official website of the WHO reported a case of novel coronavirus (later named "MERS coronavirus") infected case identified by UK Department of Health; previously, Saudi Arabia also reported a case; the clinical presentations of the two cases primarily included fever, cough, polypnea and dyspnea and other acute serious respiratory tract conditions, both complicated by acute renal failure. The daily information screening and assessment team preliminarily concluded that the disease was a serious emerging infectious disease which might be international spreading, and therefore suggested China CDC to follow up.

On the same day, the Deputy Director General of China CDC organized experts to assess the risk, and then proposed the Chinese Ministry of Health to communicate with the WHO through the information exchange mechanism defined in the International Health Regulations (2005), in order to be informed of the latest development; and communicate promptly with other health agencies around the globe to obtain information on virus, epidemic and response; prepare necessary risk communication materials to launch risk communication in a timely manner as needed; the National

Institute for Viral Disease Control and Prevention of China CDC makes preparations for emergency laboratory testing; China CDC carries out dynamic risk assessment according to further information gained.

Subsequently, China CDC continuously tracked the progress of the epidemic according to Internet information. The official websites of the WHO, the Ministry of Health of Saudi Arabia, US CDC, ECDC and other health agencies, ProMED, CIDRAP and other third-party professional websites, and news media, etc. have become important sources of information on the disease and related studies for China CDC, and important evidence base to inform the decision of China Ministry of Health to launch joint prevention and control mechanism with the other departments.

(2) *West Africa Ebola Virus Disease Outbreak*: On March 20, 2014, China CDC received an early warning email from GPHIN:

> According to Reuter news, as indicated by official of Guinea's Health Ministry on March 19, 2014, since February 2014, an outbreak of hemorrhagic fever of unknown cause has killed at least 23 people among 35 cases in Guinea's southeastern forest region. Symptoms appear as diarrhea and vomiting, with a very high fever. Some cases showed relatively heavy bleeding. Most of the victims had been in contact with the deceased or had handled the bodies. Those infected had been isolated and samples had been sent to Senegal and France for further tests. Local staff presumes it was Lassa fever or another form of cholera, and meanwhile is looking for other possible causes, such as Ebola hemorrhagic fever. More than half of the country's 11.4 million people live on less than $1 a day and many lack access to basic medical facilities and qualified medical staff.

Upon receipt of the information, the daily information screening and assessment team concluded that the mortality of the condition was high, the level of prevalence was beyond expectation, and the cause was unknown and thus it could not be determined whether it was an emerging infectious disease; the event took place in poverty-stricken country with weak healthcare system (and special background), hence higher risks; nevertheless, the information was released from media, with lower reliability, and further follow-up was needed by referring to Internet-based information. On March 24, 2014, the WHO officially released on its website the information on confirmed Ebola virus disease outbreak in Guinea. On March 31, 2014, China CDC learned about the further progress of the epidemic from the Internet, and decided to discuss about the epidemic at the trend evaluation meeting to be held in April 2014 (a regular meeting held on the 4th working day of each month).

Secondly, the gaps in traditional surveillance in China have been bridged: China CDC used to analyze the events concerned in surveillance using Internet-based data in 2013 (Meng et al., 2015), and found that, among the events revealed on the Internet which conformed to the criteria for public health emergency event-based surveillance reporting in China, only 29% were actually reported by the notifiable reporting system; besides, among those events reported by both the Internet and the notifiable reporting system, 20% were first reported on the Internet instead of the regular surveillance system. This means that, for public health surveillance in China, surveillance using Internet-based data can fill in the gaps of the traditional surveillance to some extent, by reporting more events and more rapidly.

Weaknesses and prospect of surveillance using Internet-*based data in China*: To develop a more robust Internet-based information automated surveillance system, the first priority is to try to include as many known websites as possible, in order to avoid missing important information. Nowadays, the Internet is developing rapidly, with various types of websites emerging that present huge amount of information. The system development resources of CDCs at China are limited. Therefore, to date the system developed by China CDC covers only a tiny portion of mainstream websites, hence missing substantial information. Text analysis is also limited, and is simply based on screening with simple key words, text comparison and correlation, etc. Therefore, the information automatically presented by the system usually comes with many white noises, resulting in lower efficiency of routine daily manual information screening. To address these two weaknesses, it is advisable to pool the Internet-based information surveillance resources of various departments (e.g., public security, health, food and drug administration) at different administrative levels (national, provincial, city, and county levels) in China, develop an uniform information retrieval platform and text analysis techniques, and then develop subsystems appropriate for the different departments and administrative levels as needed.

References

China Internet Information Center, 2015. 36th China Internet Development Statistical Report (Online). Available from: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/t20150722_52624.htm (accessed 22.07.15).

Ginsberg, J., Mohebbi, M.H., Patel, R.S., et al., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.

Guo, Y., Wan, M., Zhu, D., et al., 2011. Design and implementation of network public opinion monitoring system on public health. J. Med. Inform. 32 (8), 6–9.

Madoff, L.C., 2004. ProMED-mail: an early warning system for emerging diseases. Clin. Infect. Dis. 39 (2), 227–232.

Meng, L., Lv, Y., Cao, Y., et al., 2015. Information obtained through Internet-based media surveillance regarding domestic public health emergencies in 2013. Chin. J. Epidemiol. 36 (6), 607–611.

Public Health Agency of Canada, 2004. Information: Global Public Health Intelligence Network (GPHIN) (Online). Available from: http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk-eng.php (accessed 22.07.15).

WHO CSR, 2000. A Framework for Global Outbreak Alert and Response. WHO, Geneva.