Key Technologies for Symptom Surveillance System: Enlightenment on the Construction of Multi-trigger Mechanism for Intelligent Early Warning of Infectious Disease

Rong TANG\textsuperscript{1,a,*} and Xiao-ying YE\textsuperscript{2,b}

\textsuperscript{1}Chongqing Center for Disease Control and Prevention, Chongqing, Yuzhong, China
\textsuperscript{2}Neusoft Institute of Guangdong, Guangdong, Foshan, China

\textsuperscript{a}20496921@qq.com, \textsuperscript{b}29217830@qq.com

*Corresponding author

**Keywords:** Symptom Surveillance, Infectious Diseases, Multiple Triggers.

**Abstract.** The most prominent feature of symptom surveillance is to pay attention to the collection and use of data from multiple sources, which gives full play to the ability of different data to support early detection of infectious diseases and the detection and early warning of major infectious disease outbreaks. This feature is in line with General Secretary Xi Jinping’s "Establishment of Intelligent Early Warning with Multi-trigger mechanism". With the construction of new infrastructure such as 5G base stations, data centers, and artificial intelligence, technical support and application scenarios are provided for the construction of an intelligent symptom surveillance system. This paper studies and analyzes the key technologies and applications of existing symptom surveillance systems in China and abroad, and aims to provide theoretical basis and practical reference for reforming and improving the disease surveillance system in China and constructing an intelligent early warning multi-trigger symptom surveillance system.

1. Introduction

On June 2, 2020, General Secretary Xi Jinping emphasized at a symposium of experts and scholars that strengthening early surveillance and early warning capabilities should be the top priority for improving the public health system, i.e., improving the surveillance mechanism for unexplained diseases and abnormal health events, and establishing intelligent early warning multi-point triggers mechanism to enhance the ability of real-time analysis and centralized research and judgment \cite{1}. Symptom surveillance, as a method of early detection and early warning of disease outbreaks, has laid a certain applied research foundation in the fields of public health protection for major events, response to biological threats, early identification of emerging infectious diseases and outbreaks. In particular, the symptom surveillance system continuously collects and analyzes a variety of data at different stages of disease development, each of which provides a possibility for timely detection of infectious disease outbreaks.

2. Current Status of Symptom Surveillance Systems at Home and Abroad

2.1. Current Situation Abroad

The BioSense system of US CDC collects data on emergency room diagnosis, test results, over-the-counter drug sales, etc., and automatically screens and promptly detects outbreaks of infectious diseases \cite{2}. The real-time outbreak and disease surveillance (RODS) system established by the University of Pittsburgh uses Bayesian networks to divide chief complaints into 7 syndromes.
including the respiratory tract and the gastrointestinal tract, which using recursive least squares model (RLS) and What's Strange About Recent Events algorithm (WSARE), combined with time and space characteristics, plays an important role in the timely detection of abnormal epidemic patterns of diseases and symptoms and the investigation of suspicious events [3]. The Electronic Surveillance System For The Early Notification Of Community-based Epidemics (ESSENCE) developed by Johns Hopkins University in the United States, via automatically acquiring clinical and non-clinical data from the military and local institutions in real time, adopts the ICD-9 disease code or natural language processing algorithm to divide the surveillance data into several syndromes according to the use institutions and their surveillance purposes, which uses time, space and time-space data analysis methods, making up for such deficiencies as the underreporting and late reporting of legal infectious diseases. Therefore, it has been promoted and applied in the United States, and constantly iterated and update [4-6]. Google Flu Trends (GFT), calculates the number of searches for influenza-related events or symptoms in a certain period of time, to discover the flu trend, 1 to 2 weeks ahead of the US CDC flu surveillance system as a whole [7]. In 2011, Enterohemorrhagic Escherichia coli (ECEH) O04:H4 broke out in Germany, whose symptom data were collected manually for reporting [8]. In 2013, Japan used data from the legal symptom surveillance system (OSS), ambulance operation symptom surveillance system (ATSS), drug surveillance (PS), and school absenteeism surveillance system (N-SASSy) for risk assessment during major events, supplementing traditional surveillance with symptom surveillance [9].

2.2. Current Domestic Situation

The symptom surveillance system established during the 2008 Beijing Olympics used five symptoms such as fever, diarrhea, and jaundice as the surveillance targets which were directly reported to reporting system via network by doctors in the venue medical centers and second-level and above hospitals, and then analyzed by SPSS statistical software to screen infectious diseases such as chickenpox, malaria, and bacillary dysentery, filling the gaps in the discovery and reporting of previous infectious disease outbreaks[10]. The "Pudong Syndromic Surveillance and Early Warning System" (PD-SEWS) established during the Shanghai World Expo 2010 was a symptom surveillance network consisting of medical institutions, schools, pharmacies, hotels and other surveillance sites, where medical institutions provided patient diagnosis and treatment information via HIS rebuilding; schools and childcare institutions reported information about absences due to illness; pharmacies reported over-the-counter drug (OTC) sales information; hotels reported abnormal health information of guests. Absolute value method and cumulative sum model (CUSUM) were used in the system for early warning.

3. Key Technologies of Symptom Surveillance System

The design and establishment of a symptom surveillance system need to comprehensively consider actual needs, available conditions and resources, etc., mainly including clarifying surveillance purpose, determining surveillance contents, data collection, data analysis, result interpretation and response. From the technical aspect of system construction, taking "data" collection, analysis and utilization as the main line, the symptom surveillance system selects appropriate information technologies for surveillance, early warning and response.

3.1. Data Collection Technology

The most prominent feature of symptom surveillance is the emphasis on collection and utilization of
data from multiple sources. Any information conducive to the early detection of infectious disease outbreaks can be included in the surveillance category. Not only can multi-source data increase the timeliness and sensitivity of detecting abnormalities, but they improve the credibility of outbreak detection since data corroborate each other compared with the direct reporting network of infectious diseases with a single data source. Figure 1 shows that, depending on the different stages of disease development, and based on the achievements of existing symptom surveillance systems at home and abroad, the article marked the collectable data on the time axis, the more left data collected, the better the sensitivity and timeliness of surveillance.

Figure 1. Collectable data at Different Stages of Disease Development.

At present, the main data sources used by symptom surveillance systems at home and abroad are search engine symptom search, news media and Weibo public opinion information and other internet data, as well as school absenteeism, factory absenteeism, non-prescription drug sales in pharmacies, outpatient and emergency patient complaints and preliminary diagnosis, hospitalization and discharge records and laboratory inspection information and other data. Relatively speaking, the specificity and accuracy of hospital visit data are relatively high, and easier to track disease case information, while other data are more sensitive and timely in detecting disease outbreaks from the perspective of the stage of disease development.

Since the collection methods are different depending on different data sources and original existence forms, the surveillance purpose and data accessibility should be comprehensively weighed to determine the data sources and collection methods which are roughly divided into web crawler, data sharing and exchange, system filling and reporting. Web crawler is mainly used for the collection of Internet data open to the society, which collects relatively simple contents, mainly collecting the information frequency associated with surveillance target syndromes. The fields covered by existing information systems, such as: non-prescription drug sales in pharmacies, absences from school or work due to illness, medical institution diagnosis and treatment, and laboratory test can directly collect from the database in real time, making full use of existing information infrastructure and ensuring the timeliness and completeness of data collection, which therefore should be used as the main method of data collection. However, subject to factors such as different data sources, insufficient integrity, inconsistent standards, and varying quality, there remain many problems to be solved in the construction of symptom surveillance system.
Meanwhile, the system filling & reporting method should be retained as a data collection tool for the fields not covered in information system or as public participation.

### 3.2. Classification Technology of Syndromes

Different data sources differ in the classification of syndromes due to differences in their collection methods, data structures and data quality. There are 11 syndromes in the BioSense system, and 7 syndromes in RODS system. PD-SEWS system divides hospital data into 9 syndromes, school and hotel data into 3 syndromes, and pharmacy data into 2 syndromes. In the early stage of system construction, it is necessary to study and determine the types and definitions of syndromes according to the surveillance targets and data characteristics of different sources, in a bid to classify each piece of collected information as a corresponding syndromes. The classification of main symptom surveillance system syndromes at home and abroad is shown in Table 1.

<table>
<thead>
<tr>
<th>Name of systems</th>
<th>Data sources</th>
<th>Classification of syndromes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioSense</td>
<td>Hospitals, pharmacies</td>
<td>Fever, respiratory tract, digestive tract, hemorrhagic, local skin injury, lymphadenitis, nerve, rash, serious illness and death, special infection, botulism.</td>
</tr>
<tr>
<td>RODS</td>
<td>Hospitals, pharmacies</td>
<td>Respiratory tract, digestive tract, skin rash, systemic disease, haemorrhage, neurosis, botulism.</td>
</tr>
<tr>
<td>ESSENCE</td>
<td>Hospitals, pharmacies and schools</td>
<td>Respiratory, gastrointestinal, fever, nervous system, botulism, rash, local injury, hemorrhagic disease, lymphadenopathy, sudden illness/death, others.</td>
</tr>
<tr>
<td>PD-SEWS</td>
<td>Hospitals</td>
<td>Febrile respiratory tract, severe febrile respiratory tract, diarrhea, rash, hemorrhage, central nervous system, severe central nervous system, botulism, jaundice.</td>
</tr>
<tr>
<td></td>
<td>pharmacies</td>
<td>Respiratory tract, digestive tract.</td>
</tr>
<tr>
<td></td>
<td>Schools</td>
<td>Respiratory tract, digestive tract, rash.</td>
</tr>
<tr>
<td></td>
<td>Hotels</td>
<td>Fever, diarrhea, emesis.</td>
</tr>
</tbody>
</table>

At present, there is no uniform standard for the definition of syndromes at home and abroad. Dynamically affected by multiple factors such as data sources, data quality, surveillance objectives and surveillance objects, the classification of symptoms cannot be generalized. In the construction of the symptom surveillance system, it is necessary to fully demonstrate the classification method of symptoms depending on data characteristics, in an effort to determine the definitions of symptoms and clarify the classification criteria. Aiming to adapt to the changes in disease surveillance objectives, the classification of symptoms should have a certain degree of flexibility and extensibility.

After the syndromes classification is determined, the collected data needs to be collected as accurately as possible into the corresponding syndromes database according to the definition and classification criteria of syndromes. For unstructured or semi-structured data such as the Internet, chief complaints during hospital visits, medical advice, etc., which need to be data pre-processed,
the natural language processing (NLP) information retrieval, information extraction, and text classification (text classification) technology can be used to find data related to syndromes and then automatically judge the predefined syndromes\textsuperscript{[11]}. Most hospital HIS systems use ICD-10 code whose data record is structured, and such data can be directly used for syndromes classification. With the mature application of artificial intelligence technology, machine learning algorithms such as Support Vector Machine (SVM), Decision Trees, and Bayesian network will play an increasingly important role in the automatic classification of syndromes.

### 3.3. Early Warning Model Construction Technology

Early detection of infectious disease outbreaks is the main purpose of symptom surveillance. At present, the main symptom surveillance systems at home and abroad carry out automatic early warning by constructing early warning models, which are divided into three early warning models based on time, space and time. The typical time models include regression model (RM), cumulative sum model (CUSUM) and exponential weight moving average model (EWMA). The spatial models include spatial scanning statistics model (SaTScan) and Bayesian statistical model (Bayes). The time-space models include anomaly detection and Bayesian network model (WSARE) and population-wide anomaly detection and assessment model (PAN-DA)\textsuperscript{[12]}. See Table 2.

<table>
<thead>
<tr>
<th>Name of systems</th>
<th>Early warning models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time models</td>
</tr>
<tr>
<td>BioSense</td>
<td>CUSUM, EWMA, SMART</td>
</tr>
<tr>
<td>RODS</td>
<td>CUSUM, EWMA, MA, RLS, Serfling</td>
</tr>
<tr>
<td>ESSENCE</td>
<td>CUSUM, RM, EWMA, EARS</td>
</tr>
<tr>
<td>PD-SEWS</td>
<td>CUSUM, absolute threshold</td>
</tr>
<tr>
<td>GFT</td>
<td>RM, SVM</td>
</tr>
</tbody>
</table>

RM is a mathematical model that studies how a variable changes with other variables and quantitatively describes the statistical relationship, which is suitable for detecting outbreaks of infectious diseases. CUSUM is a control chart-based method that uses limited benchmark data to monitor the average deviation of observed values. This method performs well for the rapid detection of small changes in the averages, whose disadvantage is lack of adaptability to seasonal or diurnal effects. EWMA, based on recent data, allows adopting different weighting values to adjust the sensitivity of movement, whose recent forecast has the highest smoothness index. SaTScan relies on scanning areas of interest using simple forming areas to provide early warning of outbreaks based on defined ratios. Bayes is used to calculate the probability of an outbreak, which is highly efficient and able to integrate prior knowledge to effectively predict the probabilities of outbreak and infection. WSARE uses specific features to group scan disease-related diagnoses, and compares with historical conditions to find abnormalities. The advantages of WSARE lie in the abilities to predict the probability of some situations that have not been encountered before based on the generalization ability of Bayesian networks, thus it has a good effect on the detection of
emerging infectious disease. PAN-DA is a king of probability distribution building and inferring the entire population based on a causal Bayesian network.

4. Enlightenment

In terms of different stages of disease development, all related data after the patient develops symptoms can be incorporated into the symptom surveillance system for abnormal detection. These various types of data with multiple sources serve as the foundation for establishing an intelligent early warning multi-point trigger mechanism for infectious diseases. However, the key to truly发挥 the role of these data relies on the choice of key technologies such as data collection, syndromes classification, and early warning model construction, since different technologies have their own limitations and advantages under different conditions.

Based on artificial intelligence technologies such as machine learning, deep learning, and unsupervised learning, the intelligent early warning technology with stronger data analysis capabilities can be used to analyze the relationship and change laws among different factors in multi-source data. However, the selection of intelligent algorithms requires long-term research and multi-party collaboration. Even so, the construction of an intelligent symptom surveillance system still shows bright application prospects. Based on the above research on the key technologies, the following aspects should be highlighted for further exploration and practice in the future.

4.1. Enhance Multi-department Business Collaboration and Data Integration

Symptom surveillance is a data-driven method of infectious disease surveillance, which involves all kinds of public information on the Internet, as well as business data of hospitals, schools, factories, pharmacies and other units. It is necessary to fully integrate existing internal and external information resources to achieve multi-source and multi-dimensional data integration with high degree, and gradually built a trinity business synergy application model of "departmental data sharing, disease control business management, and grassroots follow-up intervention".

4.2. Strengthen the Application of New-generation Information Technology

As new infrastructures develop rapidly such as 5G networks, data centers, artificial intelligence, and the Internet of Things, when building the symptom surveillance system, the new generation of information technologies such as big data, artificial intelligence and block chain should be utilized fully to reshape and optimize the existing business process and working mode of infectious disease surveillance and early warning, in terms of system architecture, database design, early warning model construction, and data mining and utilization.

4.3. Continuous Optimization and Iterative Updates

With the constantly changes in the spectrum of diseases, the replacement of information technology, and the changes in the form of information resources, it is necessary to constantly improve and optimize system functions, data collection methods and syndromes classification, and continuously conduct prospective research and application of surveillance and early warning models, thereby achieving in-depth mining and utilization of surveillance data.

5. Conclusion

The multi-source, multi-dimensional and multi-channel data used by the symptom surveillance
system effectively compensates such flaws as timeliness and sensitivity of traditional disease surveillance methods, and the combination of the two can provide more accurate and timely prediction and early warning. How to give full play to the unique advantages of timeliness and sensitivity in detecting infectious diseases and how to select surveillance indicators to achieve the surveillance targets are the current challenges to construct the symptom surveillance system and establish intelligent multi-point trigger mechanism for early warning. The diversification and integration of surveillance data, the intelligence of syndromes classification and early warning models, and the full integration of various surveillance technologies and intelligent analysis technologies will be the main research directions for constructing a symptom surveillance system.

Acknowledgement

[Fund Project: Characteristic innovation project of colleges and universities in Guangdong Province (Natural Science): Research on infectious disease monitoring and early warning technology based on Internet big data (2020KTSCX170); Jointly implemented Key R&D projects of Sichuan and Chongqing In 2020: Construction and application research of intelligent monitoring and early warning platform for key Infectious diseases based on big data (cstc2020jscx-cylh0019)]

Author’s profile: Tang Rong, female, master's degree, senior engineer, mainly engaged in disease control informatization and smart city research.

References


