Conception on Public Health Surveillance System Based on Big Data in Health and Medical Care

Xinran Xin

School of Management and Economics, The Chinese University of Hong Kong (Shenzhen), Shenzhen, 518172, China

Abstract. In recent years, public health events in varying degrees have occurred in different parts of China, these public health emergencies are characterized by volatility and complexity. To effectively respond to major challenges, we must establish a comprehensive public health surveillance system capable of efficiently collecting, analyzing, and processing data, while providing effective and timely information support. This paper focuses on how big data technologies can be used to achieve this goal. The theoretical basis is first introduced, including the definition, characteristics and application methods of big data; then the existing public health surveillance system is analyzed and summarized, its shortcomings are pointed out; finally, a design scheme for a public health monitoring system based on health care big data is proposed, detailing the design ideas, functional modules and database structure of the scheme. The scheme can help the authorities to monitor and manage information in real time so as to provide the basis for scientific-decision making. Through testing and verification, the scheme is found to have good performance and can meet the needs of large-scale public health surveillance. The findings of this project hold significant relevance in advancing the development of the public health surveillance system in China.

Keywords: Big Data in Health Care, Public Health Surveillance, System Concept.

1. Introduction

With technological advancements and global informationization, the development of big data in health and medical care and artificial intelligence technologies has become an important trend in the healthcare industry, which can provide patients with more convenient healthcare services. Effective data management serves as the cornerstone for realizing intelligent healthcare and unlocking the full potential of big data in healthcare and medical services, ultimately resulting in enhanced public welfare. . "Information Governance" is an approach to data management that aims to manage health information across the lifecycle of an individual and is defined in the *DAMA Data Management Body of Knowledge* [1] as an authoritative and controlling activity to ensure the effective use and sustainability of data. Data governance aims to manage data through efficient planning and control, including the development of effective planning, the implementation of effective oversight mechanisms and enforcement measures [2]. Big data in medical care is vital to the healthcare industry, the best clinical decisions can be made only by accessing the right medical information at the perfect time through a right way, no matter for doctors or patients.[3] Measures like optimizing the organizational structure by utilizing the concept of Data Lifecycle Governance and establishing a standard system can improve the effectiveness of hospital management and help doctors make the right clinical judgement. With the advancement of Internet technology, the availability of electronic health information has grown exponentially, resulting in improved manipulation and analysis of data. This has significantly boosted the value of these data.

2. Overview of Big Data in Health and Medical Care and Public Health Surveillance System

2.1. Big Data in Health and Medical Care

The analysis of big data in health and medical care can be divided into four categories: Treatment Support, Health Monitoring, Public Health and Targeted Biomedical. [4] The source of big data in

health and medical care can be traced back to real-world clinical data, research data, etc., which brings unprecedented development opportunities to the healthcare industry. The amount of healthcare data is expanding at a rapid pace due to the deep integration of the internet and technology. However, new challenges have arisen, such as the fact that web data may be concealed within complex layout designs, requiring valuable information to be extracted through data mining techniques. To fully leverage big data in healthcare, it is crucial for healthcare organizations and research institutes to address data generation at its source and adopt appropriate technology to maximize usefulness. Although current technology can address certain challenges, accurate comprehension of data is paramount for successful implementation. [5]. Through big data in health and medical care, doctors can collect information on patients suffering from similar conditions, including symptoms, side effects, hospital admissions, drug information, feedback from clinical reports and drug efficacy, thereby providing more accurate treatment options for patients, as shown in Figure 1. This data can be generated from a variety of different sources such as wearable devices, mobile apps, smartphone apps, smart-home apps and others [6-7]. If traditional database management is compared to “pond fishing”, big data is “ocean fishing”. The "fish" refers to the data to be processed, and the variation in how the “fishing” is done is a direct result of difference in the “fishing” environment. [8] Given the evolving dispersal and origins of information and emergence of non-structured data, the healthcare industry is confronting both exceptional prospects for big data advancement and new impediments. In order to maximize the value of healthcare, it is crucial to seamlessly integrate the diverse sources and structures of big data in health and medical care, enabling the provision of thorough and inclusive information to facilitate its advancement.

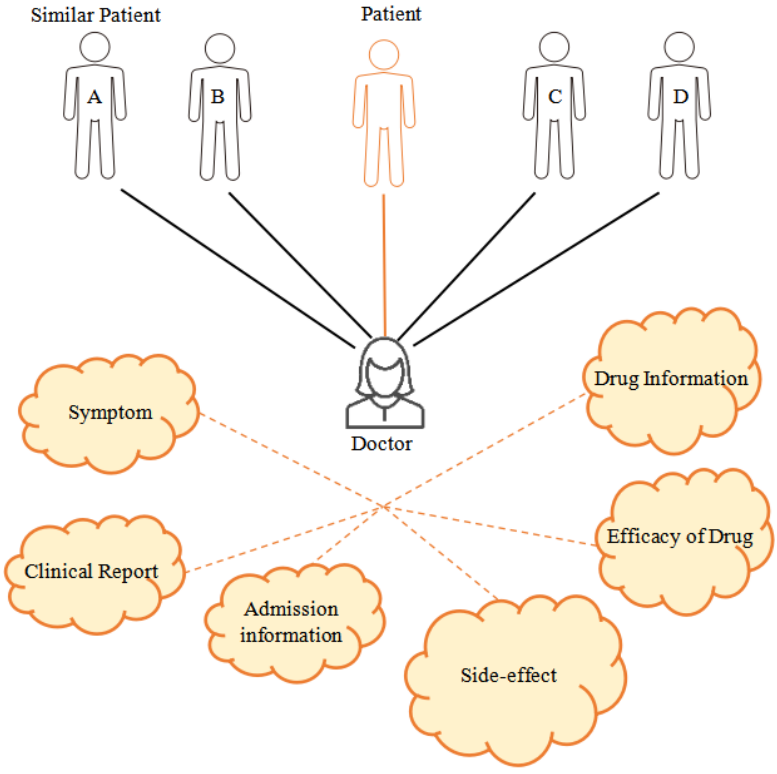


Figure 1. Big data in health and medical care to inform doctors' treatment plans

2.2. Public Health Surveillance System

At the 21st World Health Assembly in 1968, it was emphasized that public health surveillance is a critical tool for promoting global health and a fundamental responsibility of public health practice. [9] Both the World Health Organization (WHO) and the Chinese government have defined the scope of public health surveillance, as outlined in Table 1. In order to better monitor public health, public health surveillance is no longer limited to merely observing and recording epidemic events, but is conducted in a more comprehensive and systematic manner. [10] Through the establishment of a

theoretical system and a sound mechanism as well as the development of monitoring methods and technologies, growing demands can be satisfied [11].

Table 1. Definitions of public health surveillance by the World Health Organization and Chinese authorities

Organization or Country	Definition
World Health Organization	The ongoing and systematic gathering, analysis, and comprehension of health-related data needed to plan, implement and evaluate public health practice
China	The long-term, continuous and systematic collection of information on health events and health problems. The process involves the acquisition of crucial public health data through scientific analysis and interpretation, followed by timely communication of feedback to individuals or institutions who require such information to guide the development, refinement and evaluation of public health interventions and strategies.

The functionality of public health surveillance systems is continuously improving, allowing for more effective monitoring of individuals' health and more precise measures to prevent public health crises, manage behavioral risk factors, mitigate adverse drug effects, safeguard our environment, and enhance public health protection.[11] The United States, for example, has a "national statutory infectious disease program". For example, the United States has already established the National Notifiable Disease Surveillance System (NNDSS), the Behavioral Risk Factor Surveillance System (BRSS) , the Behavioral Risk Factor Surveillance System (BRFSS), the Surveillance for Foodborne Disease Outbreaks, the Surveillance, Epidemiology, and End Results (SEER)., etc. Our country has launched the National Injury Surveillance System (NISS) [11], the National Cause-of-Death Surveillance System, the National Adverse Events Following Immunization Surveillance System, the Chronic Diseases and its Risk Surveillance System, the Statutory Infectious Diseases Surveillance System, and some specialized infectious disease surveillance systems (such as influenza), etc. With the implementation of the initial concept of the National Cause-of-Death Surveillance System in 1978, the coverage of this public health surveillance system had rapidly expanded in 2013, with the number of monitoring sites increased to 605 and the population covered exceeding 300 million, or 24% of the country's total population.

2.3. Feasibility Analysis of Big Data in Health and Medical Care Applied in Public Health Surveillance

Health surveillance is a complex process of processing data that involves three main steps: firstly, determining the type, scope, source, and methods of data collection; secondly, conducting a detailed analysis, processing, and interpretation based on the information gathered in the first step; and finally, promptly publishing and implementing the research's results. Given the growing demographic shift towards an ageing population and the expanding variety of diseases, among other factors, the significance of public health surveillance is increasing in today's context. Therefore, how to optimize public health surveillance system using big data in health and medical care has become an urgent issue. Currently, the application of big data in health and medical care has gradually become a trend. By collecting a large amount of clinical data, accurate diagnosis and treatment plans on diseases can be realized. At the same time, because of its high visibility and instantaneity, big data in health and medical care can also provide powerful support for public health surveillance. To address this issue, this paper presents a conception of a public health surveillance system based on big data in health and medical care.

2.4. The Current State of Big Data in Health and Medical Care Applied in Public Health Surveillance

Currently, China has successfully developed a national health information platform and is exploring the creation of a standardized national-level system for Electronic Health Records (EHR),

Electronic Medical Records (EMR), drugs and medical instruments, public health, healthcare services and medical insurance, to enable seamless interoperability, information sharing, and business synergies. As of 2021, the number of medical and health institutions in China has reached 1,031,000. In order to accelerate the development of big data in health and medical care, nearly 200 policy documents related to big data were issued in 2019 by cities and regions across the country, such as the National Health Commission's overall plan to implement "1+7+X" for the development and application of big data in health and medical care (That is, the construction of 1 national data center, 7 regional centers, and the development of X (several) application development center specifications). In terms of healthcare industry regulation, data centers are being established in different regions to facilitate public health, disease prevention, health screenings, and healthcare oversight. Moreover, healthcare institutions have shifted to a data-driven decision-making model; in the area of patient health, medical and health data can be provided to medical institutions using portable medical devices. The proficient utilization of information technology within epidemic prevention and control has played a crucial role. Scholars suggest that advanced technologies such as Cloud Computing, big data, the Internet of Things, and 5G should continue to be employed extensively to cater to the requirements of internet hospitals and telemedicine, providing robust backing for public health surveillance.

3. Architecture of a Public Health Surveillance System based on big data in Health and Medical Care

3.1. General framework of the system

After analyzing the functional requirements of the system, the architecture of the public health surveillance system was designed, which is illustrated in Figure 2. The foundational layer consists of the data source. The second layer involves the standardization and integration of dissimilar data from the data source. Finally, the processed data is stored in the Hive database. The intermediate layer functions as the stage for data processing, utilizing the Spark platform to optimize data integration, manage data quality, fuse data, and perform other related operations. Ultimately, the user interface layer serves as the application interface for system management.

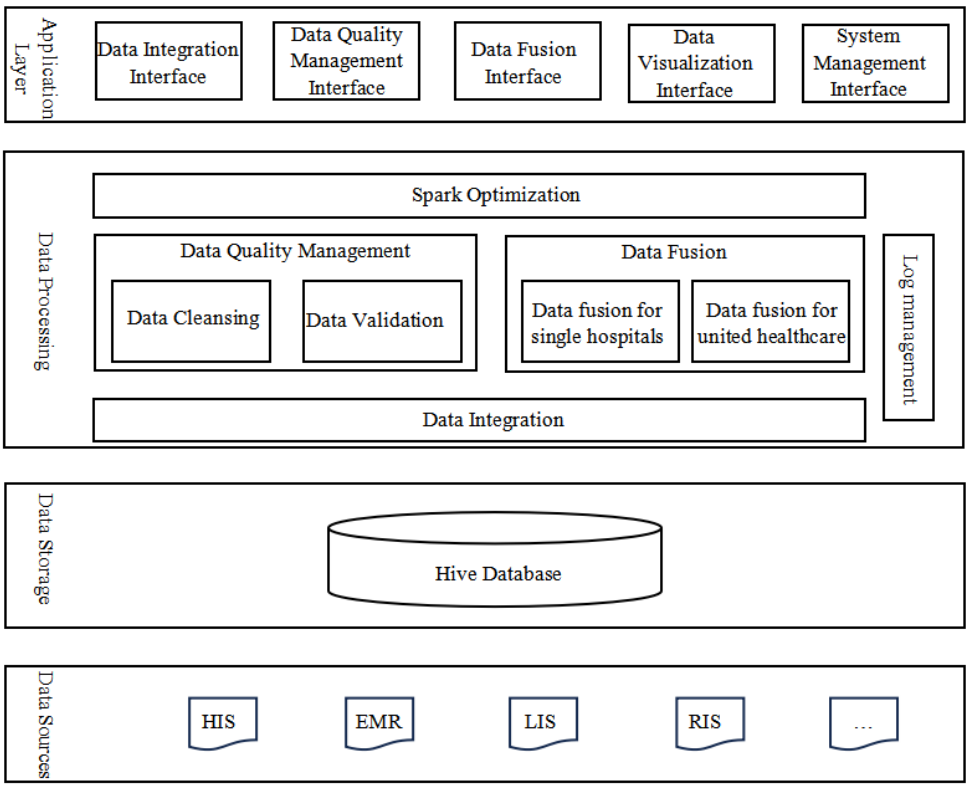


Figure 2. Overall framework of the public health surveillance system

3.2. System functional module division

The foundational layer comprises of the data source. The next layer involves the integration and harmonization of diverse data from the source, which is then deposited in the Hive database. The intermediate layer encompasses the various data processing activities, including data consolidation, data integrity management, and data synthesis, executed on the Spark platform. Finally, the user application layer is the interface for managing the system.

3.3. System database design

Due to the specific characteristic of big data in health and medical care, data sources from different medical institutions need to be addressed. The system supports the following data source types: HDFS, HBase, Hive, SQL Server, MySQL, Kudu, 0racle, FTP, Teradata, DB2, SFTP, and the data warehouse for the system is the Hive data warehouse. The data integration function mainly completes the preparation of data sources such as adding, deleting, changing and checking, as well as performing information matching for medical data source (medical data base). If the match is successful, the next step is to perform connection tests for data extraction and information integration. The configuration information includes telephone number similarity, name similarity, date of birth similarity, native place similarity, occupation similarity, ethnicity similarity, gender similarity and address similarity. The similarity code for each piece of information is defined as follows: a value of "1" denotes consistency, "0" denotes inconsistency, and a missing value is represented by "0.5". The next step involves training a machine learning model using the Random Forest algorithm.

4. Functional Modules for Public Health Surveillance Systems based on Big Data in Health and Medical Care

4.1. Data Acquisition Module

Before conducting data analysis in the public health surveillance system, data integration is performed. This entails aligning varied and dissimilar data using standardized coding techniques to optimize their storage in the Hive database. The next step involves analyzing the data using a standard structure. Figure 3 illustrates the data integration process, which involves verifying configuration information and extracting data from the source. This facilitates task progress logging and viewing.

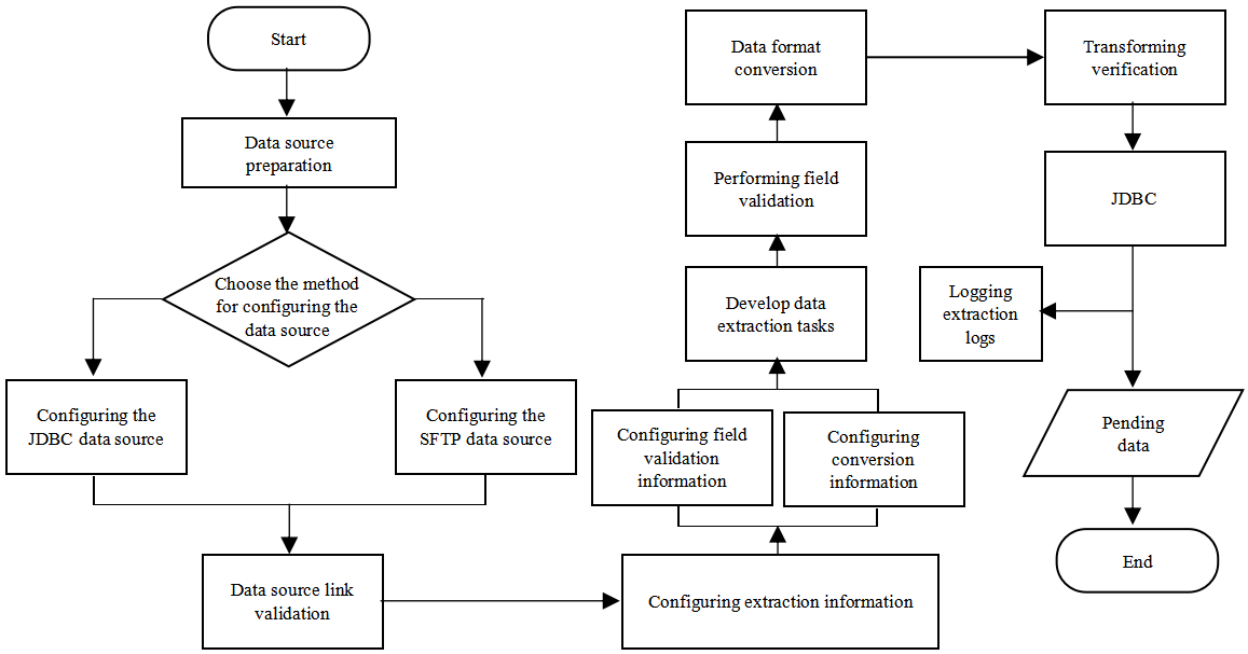


Figure 3. Public health surveillance data integration flowchart

Data analysis and processing is one of the core functions of this system, as it categorizes patients' medical information using patient IDs. However, the distinct medical institutions utilize disparate IDs, prompting this system to incorporate a rule-matching algorithm based on machine learning for compatibility. It firstly pre-processes the source data, then the qualified data tested by matching algorithm is filed into the index table, and the unsuccessful matching data can be reviewed manually again, the detailed process is shown in Figure 4.

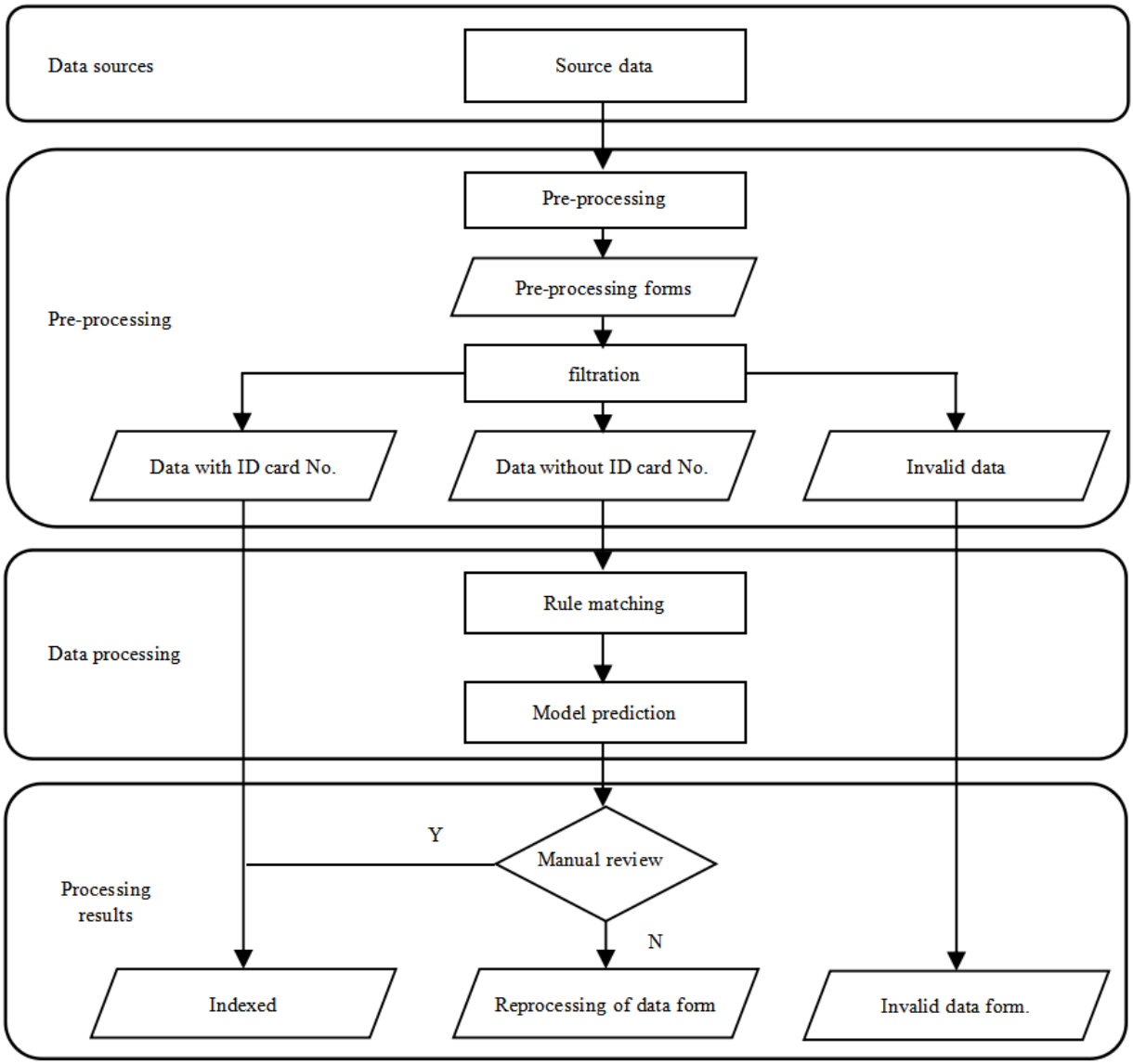


Figure 4. Public health surveillance data fusion flow chart

4.2. Data Analysis Module

There are various data analysis methods that employ Bayesian theory-based fusion models, including the Kalman filter model, the Markov model, and others. There are analysis models based on epistemology theories, but these methods have some limitations. Some may be sensitive to noise in data, others may not handle high-dimensional data properly, and some may struggle with processing missing data. Based on data analysis models learned by machine, this system designs a data analysis method based on deep learning Random Forest algorithm to construct unique patient master index ID for correlated data analysis, accelerate model training and optimization, and solving the problem of data sources from different hospitals. The process of data fusion is stored in a Hive data warehouse, which facilitates the integration and fusion of diverse data originating from multiple sources. When the training value of the model reaches the standard in the requirement analysis design,

the verifying and matching process is also carried out, as shown in Figure 4. If the data contains missing entries, duplicated values, incorrect information, null values and other errors, the system will utilize the appropriate validation and audit function modules to address these issues. Log reports or data quality reports for monitoring data will be generated, thus improving the data quality of health care intelligent governance.

4.3. Data Storage Module

The data warehouse utilized by this system is Hive. The data integration component primarily handles the preparation of data sources, including additions, deletions, modifications, and verification, as well as testing connections for information extraction and integration after configuring specific settings for medical data sources. The configuration information includes telephone number similarity, name similarity, date of birth similarity, native place similarity, occupation similarity, ethnicity similarity, gender similarity and address similarity. Each item of information's similarity code is expressed as: a value of "1" is consistent, "0" is inconsistent, and a missing value is "0.5".

4.4. Data Visualization Module

The primary objective of this module is to convert vast quantities of medical and health information into comprehensible infographics and charts for enhanced analysis and informed decision-making. Firstly, the module needs to clean and collate the collected data to ensure their accuracy and reliability. Secondly, it also needs to provide a variety of visualization tools to present this data, including bar charts, line graphs, scatter charts and many other forms. Furthermore, the module ought to enable users to personalize the presentation and parameter configurations by means of an interactive interface to satisfy the requirements of diverse user demographics. Finally, the module should include functionality for real-time updates and query history, enabling the user to access the most current data trends and changes at any given time.

5. Conclusion

This paper conceptualizes a data analysis method of Random Forest algorithm for the monitoring characteristics of current big data in health and medical care, to realize the special feature engineering of big data in health and medical care, solve the integration and analysis of multi-source heterogeneous data, improve the efficiency of information matching, and realize the data monitoring of multiple regions for medical data quality management. The research findings demonstrate that the implementation of big data within the health and medical sectors provides numerous advantages for utilization within public health surveillance:

Firstly, the timeliness and completeness of monitoring information has been improved. addressing two significant limitations in traditional monitoring methods: delayed information delivery and inaccurate geographic location. By harnessing big data from diverse sources, it is possible to collect vast quantities of data on individual subjects, which is both comprehensive and effective in its analysis.

Secondly, the development of big data can help to reveal information about “hidden” populations that are not captured by current surveillance systems. Social media and internet activity, for instance, can provide valuable data on the behavior and risk factors of populations that are traditionally “Hard-to-reach”.

Thirdly, it can help improve the sensitivity of early warning for emerging infectious diseases. In response to the global outbreak of various new epidemics, nations have invested heavily in prevention systems. The most important of which is the early prediction of emerging diseases. Big data analytics tools can effectively deliver insights on secondary data sources, offering more timely and geographically specific information that may precede diagnoses made at primary healthcare facilities. This heightened sensitivity enhances predictive capabilities and aids in early detection.

Fourthly, in contrast to conventional mathematical and statistical techniques, artificial intelligence (AI) algorithms leveraging big data are capable of swiftly processing information and precisely predicting disease transmission modes and occurrence probabilities via models. As a result, disease prevention and control systems can be enhanced.

References

[1] Liu Luyao, Cao Hang. Analysis of the development direction of electronic health records data governance [J]. Beijing Archives, 2021 (6): 14 - 19.

[2] Ruan Tong, Qiu Jiahui, Zhang Zhixing, et al. medical data governance - building a data foundation for high-quality medical big data intelligent analysis [J]. Big Data, 2019, 5 (1): 12 - 24.

[3] Long S. Tentative, Lin S. Yu, and Zhang W. Jun. The practice of whole life cycle governance model for hospital performance appraisal data [J]. Chinese Journal of Health Information Management, 2020, 17 (3): 347 - 351.

[4] Chang Chao-Die, Chen Min, Research on health care data governance methods in the era of big data [J]. China Digital Medicine, 2016, 11 (9): 2 - 5.

[5] Fei X.L., Li J., Huang Y., et al. Data governance practices in medical big data applications [J]. Chinese Journal of Health Information Management, 2018,15 (5):555 - 556.

[6] Mei Tian, Zhang Yang, Hu Shan, et al. The construction of precision medicine system and the challenges it faces [J]. China Digital Medicine, 2016, 11 (1): 44 - 48.

[7] Wang Liya, Qiu Hang, Chen Ruoya, A method of health care big data governance and visualization presentation based on metadata traceability [J]. Chinese Journal of Health Information Management, 2019, 16 (6): 661 - 666.

[8] Zhang Zhen, Zhou Yi, Du Shouhong, et al. Medical big data and the opportunities and challenges it faces [J]. Journal of Medical Informatics, 2014, 35 (6); 2 - 8.

[9] World Health Onganization. the surveillance of communicahlediseases [J]. WHO Chron, 1968, 22 (10): 439 - 444.

[10] Cai Zhiqiang, Li Liping, Bai Yunping. The past, present and future of public health surveillance: (I) the past [J]. Disease Surveillance, 2015, 30 (9): 706 - 716.

[11] Wang Xinyi, Wu Chen. Main techniques and methods of public health surveillance [J]. Zhejiang preventive medicine, 2014, 26 (1): 49 - 53.