Tuberculosis in Epidemic Prevention and Control Based on Big Data Technology

Jiao Tan, Yonghong Ma*, Ke Men, Jing Lei, Hairui Zhang and Yancheng Feng
School of Public Health, Xi’an Medical University, Xi’an 710021, Shaanxi, China

*Corresponding author e-mail: mayonghong2015@sph.edu.cn

Abstract. Tuberculosis has been a serious threat to the health of people in all countries for many years. This study mainly explores the application of tuberculosis in the prevention and control of epidemics based on big data technology. First, collect data on tuberculosis epidemics. Tuberculosis epidemic information according to the statutory infectious disease reporting system in the China Disease Prevention and Control Information System, case data of reported cases of tuberculosis in mainland China in the past three years was collected according to the review date. Next, collect map data and statistics of demographic data. Finally, clean and organize the data. Check whether there are outliers, duplications, missing and logical errors in the case data, and unify the units of the variables. The reported incidence rate was the highest in the third year and the lowest in the first year, which were 104.90 cases per 100,000 and 71.89 cases per 100,000, respectively. The number of reported cases in the three years was basically similar to the reported incidence rate. This research will effectively improve the accuracy of tuberculosis diagnosis and provide decision support for tuberculosis diagnosis.

Keywords: Big Data; Epidemic Prevention Control; Tuberculosis Disease; Data Cleaning

1. Introduction
Today tuberculosis is still one of the deadliest infectious diseases in the world. According to the WHO's global tuberculosis medical expert system, through the thinking process of medical experts to diagnose and treat the disease, it provides auxiliary tools for doctors to diagnose, treat and prevent. The product of the combination of intelligence and science. How to analyze medical clinical data, seek medical rules, and obtain valuable medical potential information is the difficulty of medical data mining research. Data mining of medical data can obtain a large number of hidden rules and knowledge in medical databases, which can provide basis for disease prediction and scientific diagnosis, and provide very valuable applications for hospital information construction.

At present, the diagnosis of typical tuberculosis is mainly judged by observing clinical manifestations, sputum examination for tuberculosis, chest imaging, and bronchoscopy [1-2]. The mathematical model can be used to analyze the correlation between more genes and the environment, but due to the complexity of the organism itself, the mathematical model will be extremely complex and requires a lot of computing resources [3-4]. Researchers have built a "big data" system for this purpose, using the network to obtain more computing power and solving mathematical model operations that could not be completed in the past [5-6]. With the continuous improvement of the Human Genome Project and the continuous growth of genome data, we may discover the evolution and changes of human susceptibility to tuberculosis in different regions, different races,
and different periods, and construct a systematic evolution of genes related to tuberculosis tree atlas [7-8]. Based on the genes of a large number of populations, search for accurate associations between individual genetic polymorphisms and susceptibility to tuberculosis, and discover the true susceptible population of tuberculosis [9-10].

The population, region, and time distribution of the high incidence of the epidemic during the pandemic and seasonal tuberculosis in each city are still unclear, and the epidemiological characteristics of each subtype of tuberculosis virus are also unclear. This information is very important for the research and judgment of the annual tuberculosis epidemic level and development trend. Furthermore, it provides an important reference for the planned and systematic preparation of seasonal tuberculosis epidemic prevention and control and pandemic response, and helps formulate tuberculosis prevention and control strategies and policies that meet the conditions of various provinces and cities.

2. Tuberculosis Prevention and Control  

2.1. Tuberculosis Consultation Model  
With the gradual decline in the number of new cases of tuberculosis, while continuing to strengthen and improve the existing treatment effects for active tuberculosis patients, it is necessary to gradually increase the protection of tuberculosis-susceptible populations. Only in this way can the effect of tuberculosis prevention and control reach a bottleneck. After this period, a new breakthrough will be formed. This strategy has been confirmed by some infectious diseases that are currently successfully controlled, such as hepatitis B and polio. Therefore, protecting susceptible people will become the focus of future tuberculosis prevention and control strategies.

2.2. Big Data  
The process of human beings fighting against tuberculosis has never stopped. In the past few decades, tuberculosis diagnosis and treatment methods have been continuously improved, and new methods and new technologies have continuously emerged. However, the global burden of tuberculosis is still very heavy, and traditional methods cannot meet today’s needs for early diagnosis of tuberculosis. Big medical data provides us with new ways to find early diagnosis of tuberculosis.

Medical big data has a wide range of sources, including a wide variety of data, complex structure, and huge scale. The development of computer networks and data science has enabled the data collected from these patients to be fully collected and stored. These data include various information about the patient. These data include life habits, work, hobbies, medical history, etc., as well as medical images, genetic status, biomarkers, and metabolism of the patient. The significance of medical big data for early diagnosis of tuberculosis is not only in the amount of data itself, but the hidden value behind it is immeasurable. At present, relevant standards have been formulated to establish and improve the medical big data system.

2.3. Kernel Density Estimation  
The kernel density estimation method can transform discrete variables into continuous variables through nuclear density conversion through kernel function, thus transforming the sudden changes of the attributes of different administrative divisions into slow transition processes, and simulating the continuous changes of attributes in real space.

Let \( x_1, x_2, \ldots, x_n \) be a sample of univariate \( x \), then:

\[
\hat{f}(x) = \frac{1}{nh\sigma} \sum_{i=1}^{n} K\left(\frac{x-x_i}{\sigma h}\right)
\]

(1)

Among them, \( k(\bullet) \) is the kernel function. When the research object is \( m \)-dimensional, it can be extended to multi-dimensional kernel density estimation:

\[
\hat{f}(x) = \frac{1}{nh \det(S)^{1/2}} \sum_{i=1}^{n} K\left(\frac{(x-x_i)^{T}S^{-1}(x-x_i)}{h^2}\right)
\]

(2)

Among them, \( m \) is the dimension of \( x \).
3. Tuberculosis Prevention and Control Experiment

(1) Collection of tuberculosis epidemic data. Tuberculosis Epidemic Information According to the statutory communicable disease reporting system in the China Disease Prevention and Control Information System, case data of reported cases of tuberculosis in mainland China in the past three years is collected according to the review date.

(2) Collection of map data. The map layer data of China comes from the Chinese Center for Disease Control and Prevention.

(3) Demographic data. The national population statistics data in the past three years are from the website of the National Bureau of Statistics.

(4) Data cleaning and sorting. Check whether there are outliers, duplications, missing and logical errors in the case data, and unify the units of the variables.

Encoding can effectively reduce the amount of data in database calculations. Before further data mining, the values and fields in the information data table need to be encoded to complete the digitization of tuberculosis disease. The measurement standards of tuberculosis disease are shown in Table 1.

| Attributes       | Data discretization and coding | Attributes       | Data discretization and coding |
|------------------|--------------------------------|------------------|--------------------------------|
| Gender           | (1)Male (2)Female              | Cough, expectoration | (1) No (2) Yes               |
| Grouping         | (1)Initial treatment (2) Retreatment | Cough, expectoration < 2 weeks | (1) No (2) Yes               |

4. Discussion

4.1. Tuberculosis Summary Results

The population distribution of reported tuberculosis cases nationwide in the past three years is shown in Table 1. In the gender distribution, the ratio of the number of reported tuberculosis cases between men and women is 2.28, 2.23, and 2.25, respectively. The number of reported cases of males is significantly higher than that of females, but the number of reported cases of males and females The ratio has basically not changed in ten years; in the distribution of age groups, the number of reported cases of tuberculosis in the 20-29 age group has always been the first in the past 3 years, and its proportions in the total number of reported cases are -18.69%, 18.50%, 17.87%, the change is not obvious, and there is also a peak in the number of reported cases in the 50-69 age group. The number of reported cases in the 0-9 age group is decreasing year by year, and the number of reported cases in the 30-39 age group is also decreasing. Obviously, its proportion in the total number of reported cases decreased from 16.88% in the past three years to 11.87% in 2014; in the occupational distribution, the number of reported cases of tuberculosis in the farmer population during the three-year period was always the first, and it accounted for the total number of reported cases in three years. The proportions of the number of cases were 61.21%, 60.43%, and 60.99%, respectively. The proportions increased slightly. The other top 5 occupations were workers, housekeeping, housework and unemployed, students, and retired people. The proportion of reported cases of household chores and unemployed people in the total number of reported cases is increasing year by year, and the changes are large, from 5.64% in the past three years to 11.96% in the past two years, and the number of reported cases from the student population accounts for the total number of reported cases the proportion of the population showed a downward trend, from 5.96% in the past three years to 3.92% in the past two years.

| Years | Number of reported cases | Male | Female |
|-------|--------------------------|------|--------|
|       |                          |      |        |
| Nearly 3 years | 678199 | 310253 |
| Nearly 2 years | 69.05  | 69.19 |
| Nearly 1 years | 680328 | 302995 |

4.2. Characteristics of Tuberculosis

The number of reported cases and reported incidence of tuberculosis nationwide in the past three years is shown
in Figure 1. The number of reported cases was the highest in the third year and the lowest in the second year, which were 1386021 and 982128, respectively. The first year was the lowest, 104.90 cases per 100,000 and 71.89 cases per 100,000. The number of reported cases in three years and the reported incidence rate were basically similar, with a slight increase in the third year, reaching the highest value in the second year, and then reaching 1 during the year, the decline was significant, and the decline has tended to be flat in the past two years.

![Graph showing number of reported cases and incidence of tuberculosis nationwide in the past three years](image1)

**Figure 1.** Number of reported cases and reported incidence of tuberculosis nationwide in the past three years

The seasonal distribution of reported cases of tuberculosis nationwide in the past three years is shown in Figure 2. In general, the number of reported tuberculosis cases from March to June is relatively high, and the number of reported cases from January to February is relatively low. The number of reported cases from January to February during the three years With a slight increase, the number of reported cases in other months has decreased. The difference between the highest and lowest months of reported cases gradually decreases, and the number of reported cases in each month tends to average.

![Graph showing seasonal distribution of reported cases of tuberculosis nationwide in the past 3 years](image2)

**Figure 2.** Seasonal distribution of reported cases of tuberculosis nationwide in the past 3 years

4.3. *Tuberculosis Prevention and Control Measures*

There were statistical differences in gender, age, and occupation distribution between the cases in the non-migrating group and the cases in the mobile group during the three-year period. In the cases of the mobile group, the proportion of women, the 20-40 age group, housekeeping, and unemployed was higher than that of the non-migrating group. This may be related to the fact that most of the floating population is young and middle-aged and mainly engaged in service industries. Therefore, these populations should be paid attention to in the
prevention and treatment of tuberculosis in mobile cases. In the distribution of occupations, the number of reported cases of farmers is significantly higher than that of other occupations. It should also be a population that needs to be focused on prevention and control of tuberculosis. In addition, the number of reported cases of households, housework, and unemployed people accounted for the proportion of the total number of reported cases in ten years. The large increase indicates that there may be a higher risk of tuberculosis occurrence and spread in these occupations, and management and control need to be strengthened. The number of reported cases of the student population has been declining as a proportion of the total number of reported cases. It is possible that the management and prevention of tuberculosis in schools The work is more complete and the results are better.

5. Conclusion
With the gradual decline in the number of new cases of tuberculosis, while continuing to strengthen and improve the existing treatment effects for active tuberculosis patients, it is necessary to gradually strengthen the protection of tuberculosis-susceptible populations. This study comprehensively analyzes the epidemiological characteristics of tuberculosis pandemic, seasonal tuberculosis, and tuberculosis, clarifies the population, region, and time distribution of the epidemic during the pandemic and seasonal tuberculosis in various provinces and cities, understands the epidemic trend of each subtype of tuberculosis virus, and explores the potential of tuberculosis virus Epidemic risk.

Combining past tuberculosis prediction research, formulate real-time prediction strategies for tuberculosis activity levels and epidemic trends in various provinces and cities, that is, using multi-source data including health system tuberculosis epidemic monitoring data, weather data, and tuberculosis-related Internet public opinion, using new big data the model constructs a predictive model of tuberculosis activity levels in various provinces and cities. The results of the research will help to transform a large amount of monitoring data into a more efficient public health policy in a timely and effective manner, and provide new theoretical basis and implementation strategies for the provinces and cities to achieve more precise tuberculosis prevention and control.

Acknowledgments
This study was supported by the project of the Shaanxi Provincial Department of Education in 2018, project name: clinical diagnosis and treatment of gastric cancer gene mutations based on high-throughput gene sequencing technology (grant no. 18JK0657).

References
[1] Wang X, Zhang Y, Leung V C M, et al. D2D Big Data: Content Deliveries over Wireless Device-to-Device Sharing in Large Scale Mobile Networks[J]. IEEE Wireless Communications, 2018, 25(1):32-38.
[2] Yudong C, Yuejie C. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation[J]. IEEE Signal Processing Magazine, 2018, 35(4):14-31.
[3] Gu K, Tao D, Qiao J F, et al. Learning a No-Reference Quality Assessment Model of Enhanced Images With Big Data[J]. IEEE Trans Neural Netw Learn Syst, 2018, 29(4):1301-1313.
[4] Al-Ali A R, Zualkerman I A, Rashid M, et al. A smart home energy management system using IoT and big data analytics approach[J]. IEEE Transactions on Consumer Electronics, 2018, 63(4):426-434.
[5] Zhang N, Yang P, Ren J, et al. Synergy of Big Data and 5G Wireless Networks: Opportunities, Approaches, and Challenges[J]. IEEE Wireless Communications, 2018, 25(1):12-18.
[6] Mohammadi M, Al-Fuqaha A. Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges[J]. IEEE Communications Magazine, 2018, 56(2):94-101.
[7] Wu C, Zapevalova E, Chen Y, et al. Time Optimization of Multiple Knowledge Transfers in the Big Data Environment[J]. Computers Materials & Continua, 2018, 54(3):269-285.
[8] Andre J C, Antoniu G, Asch M, et al. Big Data and Extreme-Scale Computing: Pathways to Convergence - Toward a Shaping Strategy for a Future Software and Data Ecosystem for
Scientific Inquiry[J]. International Journal of High Performance Computing Applications, 2018, 32(4):435-479.

[9] Xiao-Li M. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election[J]. The Annals of Applied Statistics, 2018, 12(2):685-726.

[10] Mann L. Left to Other Peoples’ Devices? A Political Economy Perspective on the Big Data Revolution in Development[J]. Development and Change, 2018, 49(1):3-36.