Progress in methodological research of infectious disease prediction

Ruidong Lv[1,a], Feng Liu[2,*]

[1]Shaanxi University of Chinese Medicine, Xianyang, Shaanxi, 712046, China
[2]Shaanxi Provincial Center for Disease Control and Prevention, Xi'an, Shaanxi, 710054, China
[a]dd27775251@163.com
[*]Corresponding author

Abstract: *Infectious diseases occur all year round in China, and often cause outbreaks and epidemics, which have made a great effect on people's health and economic development. The prevention and control of infectious diseases becoming an important public health issue. As a result, based on the previous surveillance data of infectious diseases, a prediction model suitable for the disease is built through the epidemic characteristics of various infectious diseases, and the prediction results are used to give early warning of the occurrence of infectious diseases and formulate corresponding prevention and control strategies. This paper summarizes the models used by scholars to predict infectious diseases in recent years, and provides ideas for establishing suitable models for different infectious diseases.*

Keywords: *Infectious diseases; prediction model*

1. Introduction

Infectious diseases are a type of disease caused by various pathogens that can be transmitted between humans, animals, or humans and animals[1]. Infectious diseases are mainly affected by social, economic, demographic, population health literacy, health conditions and other factors[2], these factors will lead to the accumulation, outbreak, spread and prevalence of infectious diseases. In the process of human social development, infectious diseases continue to occur, seriously affecting population health, social stability, and economic development, bringing great disease and economic burden to the country[3]. Prevention and control of infectious diseases has always been a major public health problem in the world at the present stage. With the advancement of globalization, the spread of infectious diseases has accelerated significantly. New and re-emerging infectious diseases have emerged one after another on a global scale[4]. Surveillance and early warning of infectious diseases is an important means for the prevention and control of infectious diseases. Based on the collection, analysis, evaluation and prediction of previous data, the epidemic characteristics and development trends in infectious diseases can be mastered[5]. As a result, many analysis methods and prediction models of disease incidence trend emerged. The prediction model for the characteristics of each infectious disease can provide certain technical support for the high-precision prediction of infectious diseases, and the results obtained from the analysis have important implications for the prevention and control of infectious diseases and government decision-making.

2. Overview of predictive model on infectious disease

The predictive model model of the trend of infectious disease is mainly used to simulate the epidemic transmission process of infectious diseases[6], which provides reference for the prevention and control of diseases and the formulation of prevention and control strategies. The principle of prediction model is to combine the theory of epidemiology with the scientific theory of mathematics, statistics, computer science, system science, etc. Based on the epidemic characteristics of different infectious diseases and the desired research results, appropriate tools or algorithms are selected to build suitable prediction models[7],therefore the effect of the treatment measures on the spread of the disease can be digitalized, the incidence of the disease in the future and the impact of other factors such as population flow on the spread of the disease were analyzed, so as to predict the trend of the epidemic, evaluate the effect of different prevention and measures, and timely adjust the prevention and control policies[8].

In the field of public health, the predictive model of infectious diseases is mainly applied to the prediction of the epidemic trend, the effect of different treatment measures, the early warning of infectious disease risk and monitoring[9]. The prediction model provides theoretical support for determining the etiology, source of infection, transmission route, susceptible population and risk factors of the disease, and formulating specific prevention and control measures, and providing powerful scientific support for the health administration departments in the formulation of prevention strategy of each diseases[10], and reducing the threat caused by diseases of human health and socio-economic development. At present, the mainstream infectious disease prediction models are mainly divided into three categories: the first is the infectious disease dynamics prediction model, the second is the time series prediction model, and the third is the predictive models generated by machine learning theory[11].

3. Predictive model of infectious disease dynamics

3.1 Macroscopic dynamic model

The dynamic model can characterize the macroscopic characteristics and internal laws of the spread of infectious diseases, but the prediction results are more sensitive to the value of model parameters. When the factors considered are increased, the model parameters tend to be numerous, and the prediction value will decrease significantly, but it still has a strong retrospective study value and can evaluate the effect of different intervention measures[12]. The foundation of the dynamic model is the SIR compartment model, which was put forward by Kermack and McKendrick in 1927, which divides the population within the scope of infectious disease prevalence into three categories: Susceptible(S), Infectious (I), and Removal(R). The model is suitable for diseases that are transmitted by the virus and the patient is immune to the original virus after recovery. On the basis of this prediction model, subsequent scholars proposed more refined and targeted models for disease characteristics, such as SIRS model considering the characteristics of immunity failure, SEIR model considering the incubation period of infectious diseases[13]. Xu Baochun analyzed the spread of SARS epidemic with SIR Model based on SARS epidemic data in Hong Kong[14]. Ding Huangyan used COVID-19 epidemic data in India built SIR Model to analyze the relationship between infection number and vaccine coverage rate[15]. Liu Xiaoyang used the COVID-19 epidemic data of some cities in China built the SIR Model to discuss wide-epidemic period, the final cumulative number of patients, and the hidden transmission period of the epidemic[16].

3.2 Based on individual dynamics model

With the development of model research, the individual-Based Model(IBM) have emerged, which are characterized by being able to simulate the contact and transmission process of infectious diseases in different populations more realistically and randomly[17]. For example, Kleczkowsi established an interpersonal interaction network, which simulates individual behavioral characteristics through cellular automata, to analyze the impact and magnitude of each influencing factor and treatment measure on disease transmission. Colizza divides local populations within relatively independent geographical regions into a composite population, and established a network among each local population through individual migration, while the evolution within the subpopulation was simulated with a dynamic model. The advantage of this kind of microdynamic model is that the transmission scenario of infectious disease can be added to the model to study the transmission mechanism of the disease. Due to the large number of parameters required for modeling and the large amount of calculation, there are few reports on the use of this model in infectious disease research in recent years.

4. Time series prediction model

The typical representative is the differential autoregressive moving average model (ARIMA), with full names of autoregressive, integrated and moving average models[18]. The ARIMA model is mainly used for predicting and modeling time series data. Its characteristic is to predict future trend by analyzing and fitting historical data Its basic principle is ARIMA(p,d,q) =AR(p)+I(d) +MA(q); AR(p) represents the autoregressive model, I(d) represents the difference model, and MA(q) represents the moving average model. p represents the number of lagged terms in the autoregressive model, q represents the number of lagged terms in the moving model, and d represents the number of differences in the time series data. ARIMA model can estimate the appropriate model parameters by analyzing and fitting the time series data, so as to carry out data prediction and modeling. Autoregression (AR): The value related to the current time point and several previous time points. Difference (I): In order to achieve stationarity in time series data, ARIMA models usually need to make a difference, that is, to convert raw data into differential data, eliminating influences such as trends and seasonality. Moving Average (MA): The ARIMA model is based on moving average, which refers to the error between the current time point value and several previous time points[19]. Shi Zhaohua predicted the peak incidence of respiratory infectious diseases in the next year through ARIMA model with good effect, and issued early warning information through relevant information systems[20]. Zhai Mengmeng found in the study of influenza data in Shanxi Province that the combinatorial model based on decomposition has higher prediction performance than the single model based on decomposition and the single model without decomposition when dealing with more complex temporal characteristic diseases. Wang Yanan's prediction data of four infectious diseases including tuberculosis, viral hepatitis, hand, foot and mouth disease and influenza based on ARIMAX model has been verified, proving that ARIMAX model can effectively predict the future data of the number of infectious diseases[21].

5. Machine learning prediction model

Machine learning is a multidisciplinary discipline involving statistics, mathematics, and computer science. It is the core of artificial intelligence, through the computer to simulate the real real-time human learning way, and the existing content of knowledge structure division to improve the learning efficiency. Simply put, machine learning is the process of how computer programs automatically improve their performance as they gain experience, making the system perfect itself. Depending on the learning style, machine learning theoretically includes three categories: unsupervised learning, semi-supervised learning, and supervised learning[22]. This research mainly introduces the research of artificial neural network, Bayesian learning and so on.

Dating back to the 1940s, Artificial Neural Networks (ANN) is an algorithm based on nonlinear adaptive information processing capabilities. Compared with traditional artificial intelligence methods, it has the advantage of overcoming defects in pattern recognition, speech recognition, and unstructured information processing[23]. BP neural network is a kind of multi-layer feedforward neural network which corrects error by error backpropagation algorithm. Yu Fei predicted the incidence trend of measles in Urumqi based on ARIMA model and BP neural network model, and the results showed that BP neural network model had better prediction effect than ARIMA model on the incidence of measles in each month. Gong Hao based on the ARIMA model, BP neural network model and the combination of the two models to predict the incidence trend of tuberculosis in China, the results show that the combination of ARIMA and BPNN model has the best effect, superior to BPNN model and ARIMA model[24].

Bayesian learning is an important part of statistics,it originated early in English mathematician Thomas, who proved Bayesian theory with a special case in 1763. Most of the researches based on Bayesian model are combined with the spatial and temporal distribution characteristics to comprehensively analyze the temporal and spatial information contained in the spatial and temporal data of infectious diseases and identify the association between social factors or environmental factors and diseases[25]. Xiang Yuanyuan established four models based on the Bayesian model, namely independent model, spatial independent model, spatio-temporal independent model and spatio-temporal interaction model, to judge whether the epidemic trend of scarlet fever in China from 2008 to 2018 was affected by social and environmental factors (including economic development factors, traffic pollution factors and population factors) and the size of the impact[26]. Based on the Bayesian model, Bao Kai constructed three models of time effect, space effect and spatio-temporal interaction effect to explore the influence of social economy on the incidence of hepatitis C in Lanzhou City[27].

6. Conclusions

At present, there are many forecasting models for various infectious diseases. Through the analysis and fitting of infectious disease monitoring data, most of the basic principles of various forecasting methods are to bring practical problems into mathematical models in order to better predict the future epidemic trend of infectious diseases. The spread of infectious diseases is affected by many factors such as social and environmental factors, emergencies, health conditions, etc. Therefore, before establishing the prediction model of infectious diseases, it is necessary to understand the application and characteristics of various prediction models and fully consider the role of various influencing factors to achieve the expected prediction results. The principle of infectious disease dynamic model is to study the pathways and methods of disease transmission. It is suitable for the prediction of emerging infectious diseases and influenza outbreaks. ARIMA as a one of the classical models for forecasting infectious diseases. According to the randomness, stationarity and seasonality of infectious diseases time series, appropriate models are selected to predict the epidemic trend of infectious diseases. When applying the machine learning-related model, factors such as environmental meteorological factors, health economic factors and public health measures can be incorporated into the prediction model to improve the accuracy of the model. When a single forecasting model cannot reflect the epidemic trend of infectious diseases, a combined model can be used, which means that the advantages of different forecasting models are combined to build a combined model.

This review discusses the advantages and disadvantages of various infectious disease prediction models in practical application, and the application value in the prevention and control of infectious diseases. Selecting a suitable prediction model according to the characteristics of the decomposition sequence will provide some theoretical support for the high-precision prediction of infectious diseases. Reasonable allocation of resources and timely improvement of prevention and control plans will be an important step to achieve accurate prevention and control of infectious diseases.

References

[1] Yanan Wang. "Analysis and prediction of influencing factors of infectious disease incidence in Shandong Province".Guangxi Normal University,2023.

[2] Tian Liu,Jing Zhao,Qinwen Xu. ,et al. "Incidence trends of notifiable infectious diseases in Jingzhou, Hubei Province,2005-2022", Disease surveillance, 2023.

[3] Zhongkai Wang., et al. "Epidemiological trends of notifiable infectious diseases in China from 2008 to 2020",Journal of Zhengzhou University (Medical Science Edition),vol.57,no. 03,pp. 350-356, 2022.

[4] Lili Ji,Jinyu Yu,Yumei Mao.,et al. "Analysis of Influenza like Cases and Pathogenic Surveillance in Huairou District of Beijing from 2014 to 2022",Disease Surveillance,pp.1-7,2023.

[5] Hao R,Liu Y,Shen W., et al. "Surveillance of emerging infectious diseases for biosecurity",Science China(Life Sciences) ,vol.65,no.08,pp.1504-1516,2022.

[6] HethcoteH W. "The mathematics of infectious diseases",Siam Review,vol.42,no.04,pp. 599-653, 2000.

[7] Wanqin Chen,Rongshou Zhen,Hongmei Zeng,.et al. "Analysis of the incidence trend of malignant tumors in China from 1989 to 2008" ,Chinese Journal of Oncology,vol.7,no.01,pp.517-524,2012.

[8] Xinhua Wu,Jun Wu,Renmei Xu., et al."Analysis on prevalence trend of schistosomiasis in Poyang County, Jiangxi Province from 2004 to 2020 based on Joinpoint regression model",Chinese Journal of Schistosomiasis Control, vol.34, no.01, pp.7-15, 2022.

[9] Siqin Zeng. "Joinpoint regression model and its application in epidemic trend analysis of infectious diseases", China Health Statistics, vol.36, no.05, pp.787-791,2019.

[10] Hui Li, Lina, Jiang,Weide Zeng.,et al. "Trend analysis of joinpoint regression model for the epidemiology of hand-foot and mouth disease in Guangxi Zhuang Autonomous region from 2008 to 2022",Disease surveillance , pp.1-12,2023.

[11] Meng Zhang,Dan Wu,Yixing Li.,et al. "Joinpoint regression analysis of pertussis incidence trend in some regions of China from 2004 to 2021",Chinese Journal of Vaccines and Immunization, vol.29, no.01,pp.25-30,2023.

[12] Shijing Shen,Xiaoming Cui,Wuchun Cao. "Application and research progress of prediction model of infectious diseases" ,Chinese Journal of Nosocomial Infectiology, vol.33,no.16,pp.2550-2554,2023.

[13] Baochun Xu. "Research on SARS infectious disease based on SIR Model",Shandong University, 2019.

[14] Huangyan Ding, Xiyue Tie. "Impact analysis of COVID-19 in India based on SIR Epidemic model", Journal of Chongqing Technology and Business University (Natural Science Edition), vol.39,no. 05,pp.70-77,2022.

[15] Xiaoyang Liu. "Application of SIR Model in COVID-19 analysis" , Shandong University,2023.

[16] Grimm V. "Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future", Ecological Modelling, vol.115,no.02,pp.129-148,1999.

[17] Kleczkowski A, Grenfell B T. "Mean-field-type equations for spread of epidemics: the 'small world' model" , PHYSICA A, vol.274,no.01,pp.355-360,1999.

[18] Fisman D N, Hauck T S, Tuite A R., et al. "An IDEA for short term outbreak projection:

nearcasting using the basic reproduction number",2013.

[19] Tuite A R, Fisman D N. "The IDEA model: A single equation approach to the Ebola forecasting challenge", Epidemics,pp.71-77,2018.

[20] Zhanggong Dong,Bo Song,Youxin MENG."Prediction of COVID-19 based on SEIR-ARIMA hybrid model", Computer and Modernization,pp.1-6,2022.

[21] Zhaohua Shi, Hong Su,Fengyun Qin.,et al. "Application of ARIMA model in predicting common respiratory infectious diseases", Journal of Anhui Medical University,vol.48,no.07,pp.783-786,2013.

[22] Haihong Chen, et al. "Principles and applications of machine learning". Chengdu:University of Electronic Science and Technology Pres,2017,pp.2-19.

[23] Yunkai Zhou. "Introduction of Machine Learning and its Related Algorithms", Science and Technology Communication,vol.11,no.06,pp.153-154+165,2019.

[24] Hao Gong. "Spatial and temporal distribution characteristics and prediction of pulmonary tuberculosis in China based on spatial autocorrelation and SARMIA-BPNN combined model", Yangzhou University,2023.

[25] Tian Li, Yuzeng Dai, Jiawei Li., et al. "Bayesian model of time and space in infectious disease research progress in data analysis of time and space" , Modern preventive medicine, vol.49,no. 16,pp. 2908-2912+2917,2022.

[26] Yuanyuan Xiang. "Study on spatial distribution characteristics of scarlet fever incidence and Bayes spatio-temporal modeling in China", Lanzhou University,2023.

[27] Kai Bao. "Analysis of temporal and spatial characteristics and socio-economic influencing factors of hepatitis C incidence in Lanzhou City based on Bayesian model", Lanzhou University,2023.