Proceedings of the 3rd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/21/20230799

# Forecasting the number of new crown infections in China based on machine learning methods

**Nieming Li**

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guang Zhou, 511400, China

20213801049@m.scnu.edu.cn

**Abstract.** Based on the current state and the evolution of the new crown epidemic in China, this paper uses machine learning models and historical data to predict the changes in the number of new crown infections in China in the next four months. First, analyzing the background and current situation of the new crown epidemic, and identified the research question by collecting relevant historical data, including indicators such as the number of infected people, the number of cured people, and the number of deaths. Second, employing machine learning models and MIR model to predict the trend and scale of the number of new crown infections in China over the next four months. Finally, coming to a forecast conclusion: in the next four months, the number of new crown infections in China will drop very slightly (almost remain unchanged) every month, and the monthly infection rate will remain at a low level. At the same time, discussing and summarizing the application value of the conclusions. The research results of this paper can provide useful references and guidance for government policymakers and the public, helping them better deal with the epidemic and formulate corresponding measures. In addition, the research methods and models in this paper also have a certain degree of versatility, which can provide a certain reference for other countries and regions to predict the trend and scale of the new crown epidemic.

**Keywords:** prediction, infections, model.

## 1. Introduction

China was one of the first countries to be infected by 2019-nCoV, and since December 2019, more than 1 million cases of 2019-nCoV infection have been confirmed in mainland China [1]. In China, the spread of COVID-19 has caused many social problems such as shortage of medical resources, shortage of medicines, and logistics bottlenecks. These problems have seriously affected the development of my country's economy, education and culture [2]. Take education: Many schools and universities have closed their campuses in favor of distance learning or hybrid models. This brings additional burdens and challenges to those districts and students with limited educational resources [3].

Now this study predicting the number of new crown infections has many implications: First, it helps the government and public health institutions guide resource allocation, improve medical efficiency, maximize the use of existing resources, and protect public health. Second, predicting the number of future infections can make the public more aware of the severity of the epidemic and further strengthen

DOI: 10.54254/2753-8818/21/20230799

self-protection and social responsibility. Furthermore, enterprises can adjust their production and operation strategies based on the forecast results, including taking measures to reduce the impact of the epidemic on the enterprise, adjusting product production and sales strategies, and improving the enterprise's ability to adapt and respond.

Having said that, to predict the number of new crown infections in the future, the paper need to use machine learning methods. Machine learning methods have produced impressive outcomes in many areas of forecasting. Such as image recognition and computer vision: Machine learning algorithms have shown amazing results in picture categorization, object detection, face recognition, etc., such as convolutional neural network (CNN) and recurrent neural network (RNN) in deep learning models; natural language Processing: Machine learning technology has made significant progress in speech recognition, text classification, machine translation, language generation, etc., such as recurrent neural network (RNN) and transformer model (Transformer); Finance: machine learning algorithms in stock forecasting, risk management Remarkable results have been achieved in, credit evaluation, market forecasting, etc., such as Random Forest (Random Forest), Support Vector Machines (Support Vector Machines) and Neural Networks (Neural Networks) [4-7]. Medicine and biology: Machine learning technology has achieved remarkable results in bioinformatics, medical image analysis, disease prediction, etc., such as convolutional neural networks (CNN) and recurrent neural networks (RNN) in deep learning models [8].

This is just a small sample of the applications of machine learning in different fields. Given the ongoing development of technology and the ongoing expansion of data, the scope of application of machine learning will become more and more extensive.

Furthermore, the steps of forecast research in this paper are as follows:

Data collection, cleaning and processing are carried out first. Then choose a machine learning model (when choosing a model, pay attention to choosing a model with high accuracy and generalization performance). Then build the model. Finally, the most important step is to perform model training, prediction and result visualization.

And in the following, the framework arrangement will be discussed separately: model, data, results, conclusion**.** It is predicted that in the next four months, the number of people infected with COVID-19 in China may decline slightly compared with the previous months, and the number of individuals who have COVID-19 will remain at a low level. So, government can continue to carry out epidemic monitoring and early warning; individuals can maintain personal hygiene habits, actively vaccinate, avoid gathering and cross infection and cultivate a good lifestyle.

## 2. Model

### *2.1. ML model*

The main idea of machine learning (ML) is to allow computers to automatically complete tasks by learning patterns and regularities in data sets, without explicitly writing specific programs. The process involves using mathematical algorithms to extract features and patterns from data, and then training a model to predict outcomes on new data.

ML model has many advantages. Firstly, adaptability: ML models can adaptively change their behavior to adapt to new data inputs and changes in the environment. Secondly, high accuracy: ML models can improve the accuracy of predictions by learning from large-scale datasets, especially in complex tasks. Thirdly, automation: ML models can automate tasks, reducing the need for human intervention and increasing efficiency. It is usually applied to various tasks, such as image recognition, speech recognition, natural language processing, recommendation system, etc.

The calculation formula of an ML model depends on the algorithm and model type used. For example, in a linear regression model, the calculation formula is:

$$y = mx + b \tag{1}$$

DOI: 10.54254/2753-8818/21/20230799

y represents the target variable (the variable to be predicted), x represents the independent variable (the variable used to predict the target variable), m represents the slope, and b represents the intercept distance. Therefore, the calculation formula of the ML model will vary with different algorithms and models.

### *2.2. SIR model*

The SIR model is a mathematical model used to describe the spread of infectious diseases. It divides the population into three transitional states: Susceptible, Infectious, and Recovered. In the COVID-19 epidemic, the SIR model is widely used in epidemic prediction and epidemic prevention decision-making. The main idea of this model is to study the spread and control of infectious diseases in the crowd by modeling the flow and interaction of the crowd. It's simple, clear and easy to implement and understand.

SIR Model is commonly used to predict the transmission trend of infectious diseases, develop effective measures for preventing and controlling outbreaks, and evaluate interventions such as vaccination and isolation. Or it can be used to study other types of dynamic processes, such as social networks, epidemiology, and economics.

Calculation formula:

$$\frac{dS}{dt} = -\beta SI \tag{2}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{3}$$

$$\frac{dR}{dt} = \gamma I \tag{4}$$

S means the number of susceptible people, I means the number of infected people, R means the number of recovered people, β means how many susceptible people each infected person can infect every day, γ means each infected person can recover or die every day The probability. These parameters can be adjusted according to disease characteristics and transmission [9, 10].

### *2.3. Linear regression model*

The main idea is to build a linear model to describe the relationship between independent and dependent variables, and use this model to predict new dependent variable values. It has many advantages: easy to understand and implement, high computational efficiency, good explainability, etc.

Calculation formula [11, 12]:

$$y = bx + a \tag{5}$$

$$b = \frac{n\sum_{i=1}^{n} XiYi - (\sum_{i=1}^{n} Xi)(\sum_{i=1}^{n} Yi)}{\sum_{i=1}^{n} Xi^2 - \left(\sum_{i=1}^{n} [\![Xi)]\!]\right.} \tag{6}$$

$$a = y - bx \tag{7}$$

### *2.4. Prediction model*

*2.4.1. MSE.* Mean square error is a common index for evaluating the accuracy of regression model forecasts. It calculates the square of the average difference between the predicted value and the actual value. Specifically, the smaller the mean square error is, the closer the predicted results of the model are to the actual results. The calculation method of mean square error is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{8}$$

Where n means sample size, yi represents the actual value.

DOI: 10.54254/2753-8818/21/20230799

*2.4.2. Accuracy.* Accuracy is an indicator for assessing the accuracy of predictions in classification models. It calculates the ratio of the correctly forecast sample count to the total sample count. Specifically, the higher the accuracy, the closer the model prediction results are to the actual results. The accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

TP represents the number of true cases, TN represents the number of true negative cases, FP represents the number of false positive cases, and FN represents the number of false negative cases.

*2.4.3. R-square.* R-square is a widely used indicator for estimating the prediction accuracy of regression models. It calculates the proportion of the model prediction results that can explain the actual data variance. Specifically, the higher the R-square, the more integrated the data can be, and the closer the prediction result is to the actual result. The calculation method of R square is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2} \tag{10}$$

n represents the number of samples," Yi" represents the actual value of the ith sample.

*2.4.4. RMSE.* RMSE is the abbreviation of Root Mean Squared Error and the square root of MSE. The calculation formula of RMSE is:

$$RMSE = sqrt\,(MSE) \tag{11}$$

*2.4.5. MAE.* MAE is the abbreviation of Mean Absolute Error. It is an indicator used to measure the prediction error of the model, usually used for regression problems. MAE represents the average of the absolute value of the difference between the predicted value and the true value. The smaller the MAE, the smaller the difference between predictive model outcomes and actual outcomes.

The calculation formula of MAE:

$$MAE = \left(\frac{1}{n}\right) * \Sigma |yi - \hat{y}i| \tag{12}$$

n is the sample count, yi represents true value, ŷ represents the anticipated model value.

## 3. Data

This article uses the statistics of the number of people infected with the new crown epidemic from the Official website of the World Health Organization, and first selects the number of COVID-19-infected people in China from international data centers in countries, then Select the data after November 2022, and only analyze the data after November 2022.

### *3.1. Descriptive statistics of data*

**Table 1.** Descriptive statistics.

| Number of samples | Mean | Maximum | Minimum | Median | Cumulative case |
|---|---|---|---|---|---|
| 133 | 796329 | 6966046 | 0 | 27606 | 98932687 |

Table 1 proves that there are enough samples in this study, and the values of Median and Cumulative case are also very real and reliable, which can be used for rigorous research and demonstration.

DOI: 10.54254/2753-8818/21/20230799

### *3.2. Data visualization with SIR model*

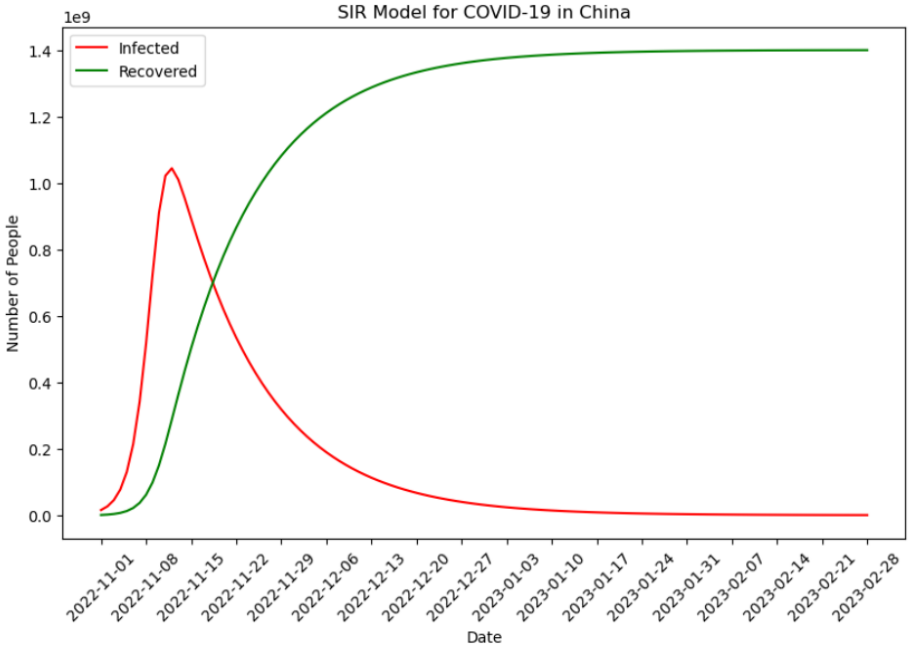


**Figure 1.** SIR Model for COVID-19 in China.

SIR is an infectious disease model, which is used to describe the mode of disease transmission in the population. The name of the SIR model comes from its classification of population into three categories: Susceptible, Infected and Recovered.

From Figure 1, it can be seen that the number of new cases of new crown infections in China rose sharply in early November, then dropped sharply in the second half of November, and then dropped back to a normal low value. Many of those infected with the new crown recovered quickly.

## 4. Result

### *4.1. Analysis of previous data*

Previous data show that since the end of 2020, COVID-19 has spread rapidly around the world. Through strong measures and extensive publicity campaigns, China gradually liberalized epidemic management in early November 2022, making positive contributions to economic recovery and social normalization. According to the data, the number of COVID-19 infections increased rapidly when China just opened the epidemic control, but after the peak of the first round of infection, the number of new infections has been declining The data shows that China has achieved some success in controlling the epidemic, and the growth rate of the cumulative number of cases has slowed down. Although new cases continue to appear, the overall trend is to gradually improve.

### *4.2. Analysis of forecast data*

The forecast data shows that in the coming months, China's COVID-19 epidemic may drop slightly or remain almost at the same value, which also shows that the Chinese government's decision to gradually liberalize epidemic management in early November 2022 is wise. The forecast data shows a relatively flat curve, indicating that the epidemic prevention measures taken by the Chinese government are effective and have mitigated the impact of the epidemic to a certain extent.

DOI: 10.54254/2753-8818/21/20230799

### 4.3. *Data prediction with SIR model*

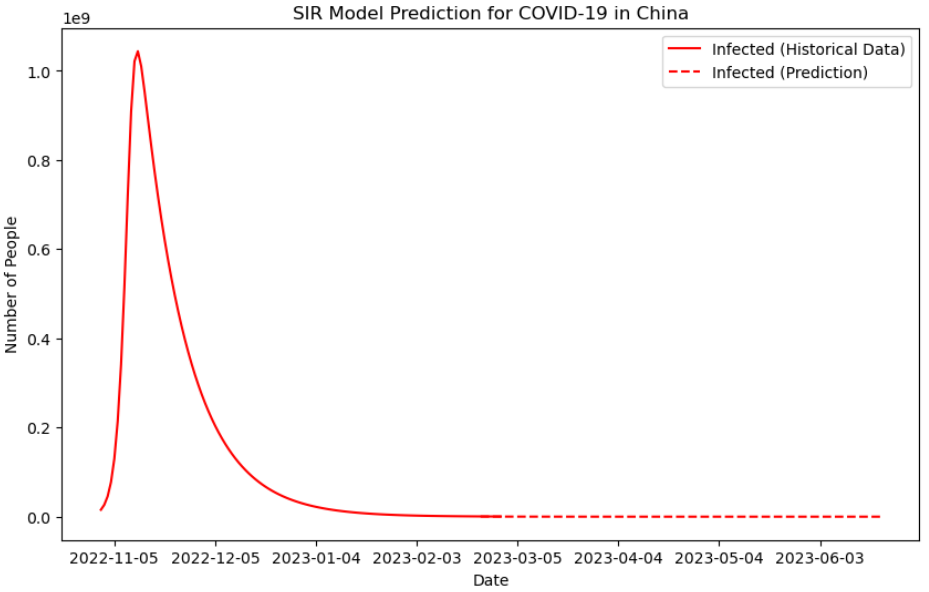


**Figure 2.** SIR Model Prediction for COVID-19 in China.

The solid line in the first half of Figure 2 is the same statistical data as in Figure 1 in the previous months, and the dotted line at the end is the forecast data for the number of new crown infections from March to June 2023. What is shown here is that the number of new crown infections will continue to be kept at a low value in the future.

### 4.4. *Data Prediction with Linear Regression Model*

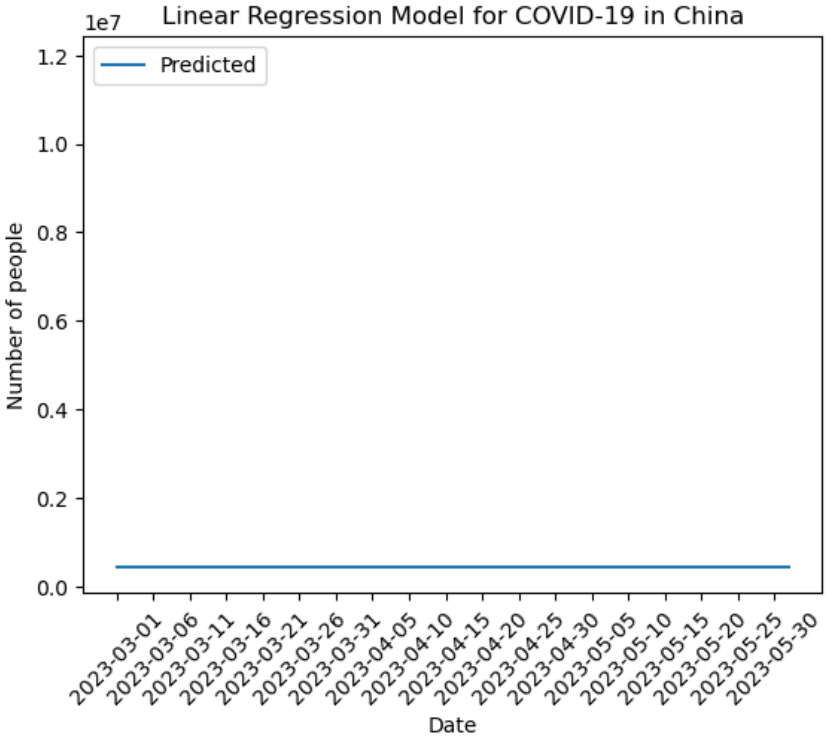


**Figure 3.** Linear Regression Model Prediction for COVID-19 in China.

DOI: 10.54254/2753-8818/21/20230799

Figure 3 shows the number of COVID-19 infections predicted from March to June 2023 through machine learning and linear regression models. Since the units of the ordinate are large, the forecasted direction of the solid line looks almost constant. It can be seen from this figure that the number of new crown infections will remain at a low value in the future.

### *4.5. Compare forecasts*

The Accuracy and R-squared values of Table 2 are very high, indicating that the prediction results of the model are very close to the actual results. Accuracy is a common performance indicator for classification models, and R-squared is a common performance indicator for regression models. The RMSE and MAE values of Table 3 are smaller, indicating that the prediction error of the model is smaller. Both RMSE and MAE have commonly used performance indicators of regression models, among which RMSE is more sensitive than MAE, because it squares the errors and then sums them, so the impact of larger errors on RMSE is more obvious. Furthermore, the predicted data and trends of the two graphs are very similar. are very slight drops and remain almost at the same value all the time.

**Table 2.** Prediction accuracy of the SIR model (Figure 2).

| index | numerical value |
|---|---|
| MSE | 1317.0000 |
| Accuracy | 0.9999 |
| R-squared | 0.9997 |

**Table 3.** Prediction accuracy of the ML regression model(Figure 3).

| index | numerical value |
|---|---|
| RMSE | 0.7831 |
| MAE | 0.7825 |

### *4.6. Discussion*

It is very important to predict the number of COVID-19 infections in China in the coming months. This is because prediction can help us better formulate prevention and control measures and effectively control the spread of the epidemic. At the same time, prediction can also help us better arrange medical resources and reduce the impact of the epidemic on society and the economy.

In terms of drug procurement, predicting the number of COVID-19 infections in China in the coming months is of great significance for drug procurement. According to a study in international literature, predicting the development trend of the epidemic can help medical institutions prepare necessary drugs and medical equipment in advance to ensure the treatment and care of patients [13]. In China, a study pointed out that at the beginning of the epidemic, due to the limited drug procurement channels, some drugs were in short supply, which seriously affected the treatment of patients [14]. Therefore, forecasting the number of COVID-19 infections in China in the coming months can help medical institutions to formulate more scientific drug procurement plans, replenish inventory in time, and ensure the treatment needs of patients.

In terms of hospital bed management and distribution, predicting the number of COVID-19 infections in China in the coming months is also of great significance for hospital bed management. An international document pointed out that forecasting the trend of epidemic development can help medical institutions optimize the allocation and use of hospital beds, increase the utilization rate of hospital beds, alleviate shortages of medical resources [15]. In China, a study also pointed out that at the peak of the epidemic, the hospital bed management of medical institutions is very critical and plays a vital role in treating patients and controlling the epidemic [16]. Therefore, forecasting the number of COVID-19 infections in China in the coming months can help medical institutions to reasonably plan hospital bed

DOI: 10.54254/2753-8818/21/20230799

resources, timely adjust the layout of wards and staffing, and improve the epidemic prevention and control capacity of medical institutions. In addition to the health sector.

In terms of logistics management, predicting the number of COVID-19 infections in China in the coming months can help logistics management. International literature pointed out that COVID-19 had a major impact on the logistics supply chain, including the shortage of raw materials and parts, the rise of logistics costs and the decline of logistics efficiency [17]. Therefore, forecasting the number of COVID-19 infections in China over the next few months can help logistics enterprises adjust their logistics plans in time, improve logistics efficiency and reduce costs. For example, when it is predicted that the epidemic will further intensify, logistics enterprises can take response measures in advance, such as increasing inventory, adjusting routes, etc., to ensure the timely delivery of materials.

In terms of corporate management, predicting the number of COVID-19 infections in China over the next few months is also of great significance for company management. International literature pointed out that COVID-19 had a profound impact on the company's operation and management, such as employee health, enterprise revenue, market demand, etc [18]. Therefore, forecasting the number of individuals infected with COVID-19 in China in the months ahead can help enterprises develop coping strategies, protect the health of employees, and stabilize revenue and market demand. For example, when it is predicted that the epidemic will worsen further, enterprises can make flexible work arrangements in advance, such as telecommuting, flexible working hours, etc., to ensure the health and safety of employees.

## 5. Conclusion

The problem of this paper is to predict the number of COVID-19 infections in China in the next four months. The research data comes from the official website of the World Health Organization. The models used in the study include the SIR model and the linear regression model. Statistical and predictive analysis of data through machine learning, modeling, and data visualization.

The main conclusion of this study is that the peak period of the first wave of COVID-19 infection has passed, and the number of COVID-19 infections in China will remain relatively low in the next four months.

However, this study also has some shortcomings: it does not consider too many external factors, such as holidays, residents' travel, weather changes, and so on. Moreover, the variation of COVID-19 was not fully considered. These issues can be predicted using a more powerful model by using more complex predictions, which can add and consider factors that affect the predicted data, thereby achieving more accurate data prediction; Find and collect more information about changes in a novel coronavirus, relevant policies and holiday arrangements to make the data more accurate.

## References

[1] National Health Commission of the People's Republic of China.. Update on the novel coronavirus outbreak. Retrieved March 8, 2023, from http://www.nhc.gov.cn/xcs/yqtb/202303/5c04e95eb16f446e969918e299e722a6.shtm, (2023)

[2] Huang, Y., Liu, Y., Li, L., Wu, Y.: Analysis on the impact of COVID-19 on China's cultural industry. Journal of Shenzhen University (Humanities & Social Sciences), 38(2), 1-10 (2021).

[3] Fang, L., Karim, S, A., Yan, H., Cao, W.: Impact of COVID-19 on education and its implications for educational technology in China. Journal of Educational Technology Development and Exchange, 14(2), 1-16 (2021).

[4] Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105 (2012).

[5] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587 (2014)

[6] Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27, 3104-3112 (2014).

[7] Kim, K. J., Han, I.: Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Systems with Applications, 19(2), 125-132 (2000).
[8] Wang, G., Li, W,: Zou, X.: Deep learning in drug discovery. Quantitative Biology, 6(3), 195-209 (2018).
[9] Kermack W., McKendrick A.: A contribution to the mathematical theory of epidemics. Proc R Soc Lond A. 115(772):700–721 (1927).
[10] Brauer, F., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology. New York: Springer, (2012).
[11] Zhou, Z.: Machine Learning. Tsinghua University Press, (2016).
[12] James G., Witten D., Hastie T., Tibshirani R.: An Introduction to Statistical Learning: with Applications in R. Springer, (2017).
[13] Carasco, L. R., Lee, V. J., Chen, M. I., Chua, H. Y., Li, Y., Oon, L. L., … Cook, A. R.: Strategies for global and regional supply of personal protective equipment amid the COVID-19 pandemic. East Asian Journal of Public Health, 30(1), 13-17 (2020).
[14] Chen, T., Zhang, W., Li Y., Tian, W.: Difficulties and countermeasures of the drug supply chain during the COVID-19 epidemic. Chinese Pharmacy, 31(23), 2759-2762 (2020).
[15] Lima, M. G., Santos, M. R. F., Nascimento, K. R. T.: Bed management optimization in hospital units using system dynamics: A case study in a COVID-19 hospital. Health Care Management Science, 24(1), 123-138 (2021).
[16] Liu Y., Li F., Wang Y., Shi D.: Discussion on hospital bed management strategies during the COVID-19 epidemic. Medicine and Philosophy (A), 41(04), 50-53 (2020).
[17] Huang, Y., Lim, Y. J., Tang, C. S.: Practical insights for managing through COVID-19 disruption with a focus on the Chinese supply chain and logistics context. Transportation Research Part E: Logistics and Transportation Review, 136, 101922 (2020).
[18] Kaplan, A. M., Haenlein, M.: Coronavirus pandemic: A global crisis of a different order. Business Horizons, 63(5), 591-595 (2020).