Modeling and Forecast of COVID-19 Vaccinations Population Size in China: A Non-Clinical Study

Wenbin Du (✉ duwb@swufe.edu.cn)  
Southwestern University of Finance and Economics  https://orcid.org/0000-0001-7604-3902

Fengrui Hua  
Southwestern University of Finance and Economics

Shengyuan Xu  
Tohoku University: Tohoku Daigaku

You Wu  
Southwestern University of Finance and Economics

Research

**Keywords:** COVID-19 vaccine, time series analysis, Python language, autoregressive integrated moving average (ARIMA) model, forecast

**DOI:** https://doi.org/10.21203/rs.3.rs-729132/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Modeling and Forecast of COVID-19 Vaccinations population size in China: a non-clinical study

Wenbin Du\textsuperscript{1,*}, Fengrui Hua\textsuperscript{2}, Shengyuan Xu\textsuperscript{3}, You Wu\textsuperscript{4}

\textsuperscript{1} Research Institute of Social Development, Southwestern University of Finance and Economics, No. 555, Liutai Road, Wenjiang District, Chengdu 611130, China. Tel: +86-13669357615; Fax: +86-28-87092257; Email: duwb@swufe.edu.cn

\textsuperscript{2} Research Institute of Social Development, Southwestern University of Finance and Economics, No. 555, Liutai Road, Wenjiang District, Chengdu 611130, China. Tel: +86-15034084982; Fax: +86-28-87092257; Email: huafengrui@mail.swufe.edu.cn

\textsuperscript{3} Graduate School of Information Sciences, Tohoku University, 6-3-09 Aoba, Aramaki Aobaku, Sendai, 980-8579, Japan. Tel: +81-08033363888; Fax: +81-22-795-5815; Email: kyos163@163.com

\textsuperscript{4} School of Accounting, Southwestern University of Finance and Economics, No. 555, Liutai Road, Wenjiang District, Chengdu 611130, China. Tel: +86-13717647930; Fax: +86-28-87092257; Email: waiola1999@sina.com

Corresponding author:
Wenbin Du, Research Institute of Social Development, Southwestern University of Finance and Economics, No. 555, Liutai Road, Wenjiang District, Chengdu 611130, China.
\textit{Tel:} +86-13669357615;
\textit{Fax:} +86-28-87092257;
\textit{Email:} duwb@swufe.edu.cn
Abstract

BACKGROUND

Since its outbreak in December 2019, severe acute respiratory syndrome coronavirus-2, the virus responsible for the COVID-19 pandemic, has considerably affected the worldwide population. Health authorities and the medical community identify vaccines as an effective tool for managing public health.

METHODS

In this study, the autoregressive integrated moving average (ARIMA) model built-in Python was adopted to establish the COVID-19 vaccination forecast model. In this study, the sample data were selected from the Our World in Data website. COVID-19 vaccinations administered daily in China from December 16, 2020 to March 21, 2021 were analyzed to establish an autoregressive integrated moving average (ARIMA) model.

RESULTS

The built-in ARIMA module function of Python was used, and the optimum model was ARIMA (3, 2, 3) according to the established time series analysis. The analysis showed that the predicted COVID-19 vaccination uptake supplemented well with the actual values with a small relative error.

CONCLUSIONS

This indicated that the ARIMA(3, 2, 3) model could be used to forecast the number of COVID-19 vaccinations in China.

CONTRIBUTION
In this study, the ARIMA model has demonstrated high goodness of fit for forecasting the COVID-19 vaccinations, which can be used in the short-term prediction of the monitored COVID-19 vaccination data sequence, providing a reference for the establishment of immune barrier and the supply of vaccines in China.

Keywords:
COVID-19 vaccine, time series analysis, Python language, autoregressive integrated moving average (ARIMA) model, forecast
Background

The COVID-19 pandemic is causing an unprecedented effect on global health and economics. When safe and efficient vaccines and treatments were unavailable, nonpharmaceutical interventions were used to reduce transmission and the burden of the pandemic. However, most of these measures involved huge economic costs. Therefore, effective COVID-19 vaccines have been urgently needed to lower the morbidity and mortality of COVID-19 (Li et al., 2020). The COVID-19 pneumonia outbreak at the beginning of 2020 involved the first COVID-19 cases caused by the novel β-coronavirus severe acute respiratory syndrome CoV-2 (SARS-CoV-2) (Tregoning et al., 2020). The genetic information was disclosed on January 10, 2020, 54 days after the first case was reported. On March 13, 2020, 63 days after the SARS-CoV-2 virus was sequenced, the news was released that the first dose of the human vaccine was under testing. The Strategic Advisory Group of Experts on Immunization of the World Health Organization has outlined the value framework for COVID-19 vaccine allocation and priority, revealing the core principles of vaccine distribution (World Health Organization, 2020). These guidelines require further specifications and should be targeted to each country and their local conditions, including the intensity of the pandemic, target of pandemic control, supply of vaccines, and population eligible for vaccination. As the first epicenter country of COVID-19 pneumonia, China has invested heavily in its vaccination, being the main participator in the COVID-19 vaccine development for controlling the pandemic, wherein its vaccines were provided by the government, vaccine makers, and
nongovernmental organizations (World Health Organization, 2020). On January 24, 2020, the Chinese Center for Disease Control and Prevention successfully separated the first COVID-19 virus strain. On June 19, 2020, the first mRNA vaccine was approved for a clinical test. As of February 25, 2021, there had been four types of COVID-19 vaccines, three inactivated vaccines, and one adenovirus vector vaccine. At 24 o’clock of March 20, 2021, the reported number of administered COVID-19 vaccinations reached 74.956 million in China (CCTV News, 2021). In this study, the autoregressive integrated moving average (ARIMA) model built-in Python was adopted to establish the COVID-19 vaccination forecast model. Time series analysis and short-term forecast were performed for daily data of COVID-19 vaccinations administered in China from December 16, 2020, to March 21, 2021. By forecasting the number of COVID-19 vaccinations and capturing the variation rule of the time series of vaccinations using the ARIMA model, the evolution rule of the time series could be extended, thereby accurately predicting the short-term COVID-19 vaccine demand and providing a reference for vaccine supply in China.

Data source and modeling principle

Data source

In this study, the sample data were selected from the Our World in Data website (https://ourworldindata.org/coronavirus-testing#china), established by the Oxford University, on the basis of the decades-long data involving the human living standards of various countries. The website provides the daily updated number of COVID-19 vaccinations and the complete COVID-19 dataset on its COVID-19 Explorer. The
data were collected by browsing the official public information, and specifically, the
data for China were sourced from the information published by the National Health
Commission of the People’s Republic of China. In this study, the daily vaccination
uptake in China from December 16, 2020, to March 21, 2021, updated by the Our
World in Data website, was used for modeling and the short-term forecast.

Method

ARIMA model

The ARIMA model is mostly used to analyze the nonstationary, nonseasonal time
series (Box, Jenkins & Reinsel, 2010), which is the most commonly used model for
fitting nonstationary series. It can be further divided into the autoregressive (AR),
moving average (MA), and autoregressive moving average (ARMA) models
(Kantelhardt, et al, 2002). They are explained below in detail:

AR model

An AR model with the following structure is called the $p$-order AR model and
denoted as AR($p$). The mathematical formula is as follows:

$$x_t = \varnothing_0 + \varnothing_1 x_{t-1} + \varnothing_2 x_{t-2} + \cdots + \varnothing_p x_{t-p} + \varepsilon_t \quad (1)$$

Here, the value $x_t$ of the random variable $X_t$ at moment $t$ is the multiple linear
regression of the first $p$ terms $x_{t-1}$, $x_{t-2}$ ..., and $x_{t-p}$. The error term is the
current random interference $\varepsilon$, which is a zero-mean white noise sequence. When
$\varnothing_p = 0$, the model is called the centralized AR($p$) model.

MA model
A MA model with the following structure is called the $q$th-order AR model and denoted as MA($q$). The mathematical formula is as follows:

$$x_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}$$  \hspace{1cm} (2)

Here, the value $x_t$ of the random variable $X_t$ at moment $t$ is the multiple linear function of the first $q$ random interferences $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$. The error term is the current random interference $\epsilon$, which is a zero-mean white noise sequence. Specifically, when $\mu = 0$, the model is called as the centralized MA($q$) model.

**ARMA model**

An ARMA model with the following structure is denoted as ARMA($p$, $q$) and its mathematical formula is as follows:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}$$  \hspace{1cm} (3)

The ARMA model is used for stationary time series, which is a combination of the AR and MA models. It is supposed that $x_t$ is the observed time series, $\epsilon_t$ is the white noise sequence, and $x_t$ is the mixed process of autoregression and moving average. When the orders of AR and MA are $p$ and $q$, respectively, the model is denoted as ARMA($p$, $q$).

**ARIMA model**

If the time series is nonstationary, its difference should be considered to transform it into a stationary one, which is then followed by the forecast using the ARMA($p$, $q$) model. These processes constitute the ARIMA($p$, $d$, $q$) model wherein $p$ represents the number of autoregressive terms, $d$ the times of difference of the time series, and $q$ the
number of moving average terms. The structure of the ARIMA\((p, d, q)\) model is as follows:

\[ \Phi(B)^d x_t = \Theta(B)\epsilon_t \]  \hspace{1cm} (4)

\[ E(\epsilon_t) = 0, Var(\epsilon_t) = \sigma^2, E(\epsilon_t \epsilon_s) = 0, s \neq t \]  \hspace{1cm} (5)

\[ E_{\epsilon_t} x_t = 0, \forall s < t \]  \hspace{1cm} (6)

\[ \nabla^d = (1 - B)^d \]  \hspace{1cm} (7)

\[ \Phi(B) = 1 - \phi_1B - \cdots - \phi_pB^p \]  \hspace{1cm} (8)

\[ \Theta(B) = 1 - \theta_1B - \cdots - \theta_qB^q \]  \hspace{1cm} (9)

**ARIMA modeling steps**

For time series analysis, the ARIMA model forecasts through the following four steps: time series stationarity test, model recognition and order determination, model validation, and model forecast. These specific processes are explained below:

**Series establishment and stationarity test**

The time series ARIMA model must be established on the basis of a stationary time series as it cannot capture nonstationary ones. Therefore, the stationarity test is needed at first. The typical methods of stationarity tests include the sequence diagram, autocorrelation, and unit root tests. In the sequence diagram test, the mean and variance curves of the time sequence data are plotted to determine if the time series is stationary by observing its trend. In the autocorrelation test, the autocorrelation function (ACF) figure is used. For nonstationary data, the ACF figure will approach 0 slowly or not at all. For the unit root test, the solved augmented Dickey-Fuller (ADF) value will be used to determine whether the sequence data are significant in the given
confidence interval. If so, the sequence data are stationary. The ADF test assumes that the time series has a unit root, i.e., the time series is nonstationary with periodic fluctuations. If the test result indicates that the absolute value of T is lower than the observation level of 1%, 5%, or 10% and the significance probability ρ value is larger than 0.05, the sequence is deemed as nonstationary, and the original hypothesis cannot be rejected. Similar to Eq. (10), the mean of the random series is expressed as follows:

\[ \mu_t = EX_t = \int_{-\infty}^{\infty} x dF_t(X) \]  

(10)

where \( EX_t \) represents the expectation of the random series \( X_t \) at moment \( t \).

If the time series is determined to be nonstationary, i.e., periodic or with a marked trend, the difference of the original sequence data will be determined. The following Eq. (11) gives the \( d \) difference equation:

\[ \nabla^d x_t = (1 - B)^d x_t = \sum_{i=0}^{d} (-1)^i c_d^i x_{t-i} \]  

(11)

Here, \( x_t \) indicates the stationary sequence after difference; \( B \) is the delay operator, indicating that the times of time index difference is \( d \) for the series.

Model recognition and order determination

After \( d \) times of difference, the ACF and partial autocorrelation function (PACF) tests on the stationary time series are commonly performed to obtain the number of layer \( p \) and order \( q \) for ARIMA. By observing the truncation and trailing of the ACF and PACF tests, the model is fitted. The fitting results are compared and adjusted accordingly for the preliminary construction of one or more suitable ARIMA models. If both the ACF and PACF tests of the time series are trailing, it can be determined
that the forecast model is ARMA($p$, $q$). The values of $p$ and $q$ are investigated gradually from the lower orders, and the smallest $p$ and $q$ are selected that can minimize the values of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Eq. (12) shows the mathematical expression of the ACF:

$$\rho_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})}$$

(12)

where $y_t$ represents the value of the random series at moment $t$, and $\bar{y}$ represents the mean of the random series.

Eq. (13) displays the mathematical expression of the PACF:

$$\varphi_{kk} = \rho_k - \sum_{j=1}^{k-1} \varphi_{k-1,j} \frac{\rho_{k-1,j}}{1} - 1 \sum_{t=1}^{n} \varphi_{k-1,t}$$

(13)

where $\varphi_{kk} = \varphi_{k-1}, \varphi_{kk} \varphi_{k-j}$, whereas $\rho_k$ indicates the ACF of the random series.

**Model validation**

Before establishing the ARIMA model to forecast the number of administered COVID-19 vaccinations, the reasonability of the model should be evaluated. The purpose of model evaluation is to avoid $p$ and $q$ deviations determined by figure reading and ensure the precision of the forthcoming vaccination forecast. The white noise and normality tests are typically used to examine the residual sequence. If $p$ values for the fitting test statistics are all considerably larger than the significance level of 0.05, the residual sequence can be observed as a white noise sequence. If the fitted model cannot pass the test, the model will be reselected for fitting. In the normality test, the obtained data are, however, assessed to see whether they satisfy the normal distribution. The standardized residual Q-Q plot of the ARIMA model is
drawn to determine the reasonability of the established forecast model.

**Model forecast**

The main function of the time series analysis is to explain the autocorrelation of a time series with mathematical models, thereby predicting the future variation rule of the series. In this study, the daily data of the COVID-19 vaccinations administered in China from December 16, 2020 to March 21, 2021 will be selected to find the optimum forecast model and make five-day projections. By comparing the predicted value with the actual data, the forecast performance will be assessed.

**Result**

**Time series establishment and stationarity test**

The data of the daily COVID-19 vaccinations administered in China from December 16, 2020 to March 21, 2021 were first used to create the time series plot. Figure 1(a) shows the date on the horizontal axis and the number of new COVID-19 vaccinations each day on the vertical axis. The vaccination data clearly showed sustainable growth, which is typical of a nonstationary series.

[Figure 1 about here] Time series plots of the COVID-19 vaccination data in China. (a) Time series plot of the original COVID-19 vaccination data, (b) Time series plot of the COVID-19 vaccination data after first-order difference, (c) Time series plot of the COVID-19 vaccination data after second-order difference
The original time series was nonstationary, which required the operation of the difference to be stationary. It was found that after trial, the first- and second-order differences could remove the trend of the original series, making it stationary. The time series plots after the difference are shown in Figures 2(b) and 2(c). In addition, after $k=3$, the autocorrelation and partial autocorrelation coefficients gradually decreased and fell into the confidence interval, indicating that the current time series turned stationary (Figure 2(c)).

[Figure 2 about here] Plots of autocorrelation and partial autocorrelation before and after difference. (a) Plots of autocorrelation and partial autocorrelation for the original series, (b) Plots of autocorrelation and partial autocorrelation after the first-order difference, (c) Plots of autocorrelation and partial autocorrelation after the second-order difference.

The unit root (ADF) method was then adopted for the stationarity test. As shown in Table 1, the $t$ statistic of the ADF test targeting the time series data of administered COVID-19 vaccinations in China was $0.519$; $p$ was $0.985$; and 1%, 5%, and 10% thresholds were $-3.503$, $-2.893$, and $-2.584$, respectively. Moreover, $p = 0.985 > 0.1$, so the original hypothesis could not be rejected, and the series was not stationary. The first-order difference of the series was then performed, followed by the ADF test, resulting in $p = 0.845 > 0.1$. Therefore, the original hypothesis still could not be rejected, indicating a nonstationary series. The second-order difference was then
performed, followed by ADF test, resulting in $p = 0.000 < 0.01$, indicating 99% or higher chance of rejecting the original hypothesis. At this moment, the series was stationary. Therefore, the parameter $d$ of the ARIMA model was 2.

[Table 1 about here] Table 1. ADF test results

Model recognition and order determination

According to the ACF and PACF plots of $x_t$ given in Figure 2(c), the autocorrelation plot after the second-order difference showed a significantly nonzero autocorrelation number of 2 or 3 with the trailing property, so $q = 2$ or 3. In addition, the partial autocorrelation plot showed a PACF of 2 or 3; therefore, $p = 2$ or 3. From the ADF test results, when the times of difference was 2, the series was stationary, so $d$ value was 2. The preliminary models were thus ARIMA(2, 2, 2), ARIMA(2, 2, 3), ARIMA(3, 2, 2), and ARIMA(3, 2, 3). Parameter estimation and model validation were then performed to further determine the optimum model. The AIC, BIC, and Hannan-Quinn information criterion (HQIC) were used to select the best $p$ and $q$ values. The equation of AIC and BIC is as follows:

$$AIC(n, m) = \ln \hat{\sigma}_a^2 + 2(n + m + 1)/N$$

(14)

If $AIC(p, q) = \min_{0 \leq n, m \leq 1} AIC(n, m)$, the order of the ARMA model is decided as $(p, q)$, and $\hat{\sigma}_\epsilon^2$ is the maximum likelihood estimate of the corresponding sequence.

Therefore, when selecting the parameters for ARMA($p, q$), AIC and BIC were combined, and the relatively optimum model was searched according to the minimization principle of AIC, BIC, and HQIC (Burnham & Anderson, 2004). From the lowest order, the AIC, BIC, and HQIC values of the preliminary models were
calculated, and the AIC thermodynamic diagram was created so as to determine the orders.

Table 2. AIC, BIC, and HQIC values of the ARMA model

As indicated in Table 2, the parameters of \( p \) and \( q \) were selected, which corresponded to the minimum AIC, BIC, and HQIC values. The research results showed that ARMA(3, 3) model had the lowest AIC, BIC, and HQIC values, which were 2316.015, 2336.361, and 2324.233, respectively. Similarly, from the order-determining AIC thermodynamic diagram (Figure 3), we noted that the optimum model occurred when \((p, q)\) values were (3, 3). Therefore, when \( p = 3 \) and \( q = 3 \), the AIC, BIC, and HQIC values reached the minima. Considering that the times of difference was \( d = 2 \), ARIMA(3, 2, 3) was determined to be a better model for forecasting the number of COVID-19 vaccinations, which can be expressed as follows:

\[
y_t = 2829.663 + 0.374 \ast y_{t-1} - 0.217 \ast y_{t-2} + 0.388 \ast y_{t-3} - 0.504 \ast \varepsilon_{t-1} + 0.585 \ast \varepsilon_{t-2} - 0.944 \ast \varepsilon_{t-3}
\]

Model validation

After model establishment, the residual sequence needs to be checked to determine if it only contains white noise. If not, it indicates that there still exists useful information in the residual, demanding further revision of the model until the residual
sequence is a white noise sequence. The validation process was as follows. First, when the size of the fitting model was determined, the significance of the model was assessed ($\alpha = 0.05$). It was $p = 0.14$, much larger than 0.05. Therefore, the original hypothesis was acknowledged, assuming that the time series was a white noise sequence, i.e., a randomly generated series with no temporal correlation. Second, the autocorrelation plot of the residual was created, and the pure randomness of the residual was tested. The results of the residual autocorrelation test of ARIMA(3, 2, 3) are shown in Figure 4(b), where the ACF is basically in the 95% confidence interval, and the residual is white noise. Finally, the Q-Q plot of the ARIMA model was drawn (Figure 4(c)). The points in the middle are almost overlapping with the straight line, proving that the residual of the forecast model ARIMA(3, 2, 3) followed the normal distribution, and thus, the established ARIMA(3, 2, 3) was reasonable.

[Figure 4 about here] Results of the ARIMA model validation. (a) Plot of residual sequence, (b) Autocorrelation and partial autocorrelation plots of the residual sequence, (c) Q-Q plot of the standardized residual for the ARIMA model.

Model forecast

The forecast ( ) function from the StatsModels library of Python was used for constructing the model combining the values of $p$, $d$, and $q$. Specifically, the ARIMA(3, 2, 3) model for forecasting COVID-19 vaccinations was established and five-day projections were made. The modeling code is as follows:
data=data.astype('float64')

model=ARIMA(data,order=(p,d,q)).fit()

predict_data=model.predict()

forecast=pd.Series([int(each) for each in model.forecast(5)[0]])

dates=pd.date_range(final_data['date'].values[-1],periods=6)[1:]

forecast.index=dates

The predicted values, actual values, absolute errors, relative errors, and variation coefficients are given in Table 3. The predicted trend was observed to match with the actual trend with a good fitting. The average absolute error was $-148175.2$, the average relative error was 10.06%, and the average variation coefficient was 5.34%.

The forecast of the model was within an acceptable range. The formula for calculating the average relative error is as follows:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}$$  \quad (16)

The formula of the variation coefficient is as follows:

$$c_v = \frac{\sigma}{\mu}$$ \quad (17)

In Eq. (17), $\sigma$ represents the standard deviation and $\mu$ represents the mean.

[Table 3 about here] Table 3. Predicted results by the ARIMA(3, 2, 3) model

Figure 5 shows that the five-day projections (March 22, 2021 to March 26, 2021) match well with the actual data of COVID-19 vaccination administration, showing a high accuracy of the forecast. In addition, the number of administered vaccinations shows a steady upward trend.

[Figure 5 about here] Performance of COVID-19 vaccination forecast.
Discussion

The data of COVID-19 vaccinations daily administered in China from December 16, 2020 to March 21, 2021 were selected for a time series analysis and fitting of the optimal model of ARIMA(3, 2, 3). The model was then used to predict the number of five-day vaccinations from March 22 to March 26, 2021, the results of which were compared with the actual data. The average relative error was shown to be 10.06%, whereas the average variation coefficient was 5.34%. The analysis showed that the predicted COVID-19 vaccination uptake supplemented well with the actual values with a small relative error. This indicated that the ARIMA(3, 2, 3) model could be used to forecast the number of COVID-19 vaccinations in China. The ARIMA model performs statistical analysis based on historic data. However, the vaccination speed and number were influenced by other factors such as the relevant policies and vaccine production scales. Historically, vaccine effectiveness often drops when the vaccines are ultimately rolled out into the general population because clinical trials occur under controlled and ideal conditions and may not capture all heterogeneity within the population (Fedson, 1998). This would raise the question of whether to vaccinate with currently available vaccines or wait for one with a higher effectiveness against that particular variant. Therefore, the ARIMA model was only applicable to short-term vaccination forecasts (Baden et al., 2021; Polack et al., 2021). For more accurate prediction, the inclusion of new data and adjustment of the model parameters are needed to adapt to the actual vaccination scenario.
Conclusion

In summary, the ARIMA(3, 2, 3) model has demonstrated high goodness of fit for forecasting the COVID-19 vaccinations in China, which can be used in the short-term prediction of the monitored COVID-19 vaccination data sequence, providing a reference for the establishment of immune barrier and the supply of vaccines in China.

List of abbreviations

autocorrelation function (ACF)
augmented Dickey-Fuller (ADF)
Akaike information criterion (AIC)
autoregressive (AR)
autoregressive integrated moving average (ARIMA)
Bayesian information criterion (BIC)
Hannan-Quinn information criterion (HQIC)
moving average (MA)
partial autocorrelation function (PACF)
Declarations
Ethics approval and consent to participate
Not applicable
Consent for publication
Not applicable
Availability of data and materials
The data that support the findings of this study are available from Our World in Data but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Our World in Data.
Competing interests
The authors declare that they have no competing interest.
Funding
This study was supported by the Introducing Talents Initiated Project of Southwestern University of Finance and Economics (Grant No. 230600001002020015).
Authors' contributions
WD, FH, SX conceived and designed the experiments. WD analyzed the data. WD wrote the paper. FH contributed to the revised manuscript. SX and YW collected the data. All authors read and approved the final manuscript.
Acknowledgements
Not applicable
References

Baden LR, Sahly H, Essink B, Kotloff K, Zaks T. Efficacy and safety of the mrna-1273 sars-cov-2 vaccine NEW ENGL J MED. 2021;384:403-416

Box G, Jenkins GM, Reinsel GC. Time series analysis. Forecasting and control. 3rd ed. Journal of Time. 2010.

Burnham KP, Anderson DR. Multimodel Inference Understanding AIC and BIC in Model Selection. SOCIOL METHOD RES. 2004;33:261-304.

China's State Council joint prevention and control mechanism held a press conference material. 2021. https://news.cctv.com/. Accessed 2 Jun 2021.

Fedson DS. Measuring protection: Efficacy versus effectiveness. Dev Biol Stand. 1998;95:195-201.

Kantelhardt JW, Zschiegner SA, Koscielnny-Bunde E, Bunde A, Stanley HE. Multifractal detrended fluctuation analysis of nonstationary time series. PHYSICA A. 2002;316:87-114.

Li Z, Chen Q, Feng L, Rodewald L, Li S. Active case finding with case management: the key to tackling the COVID-19 pandemic. Lancet. 2020;396:63-70.

Polack FP, Thomas SJ, Kitchin N, Absalon J, Gruber WC, et al. Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. NEW ENGL J MED. 2020;383:2603-2615.

Tregoning JS, Brown ES, Cheeseman HM, et al. Vaccines for COVID-19. CLIN EXP IMMUNOL. 2020;202:162-192.

WHO SAGE values framework for the allocation and prioritization of COVID-19 vaccination document. In: World Health Organization. 2020.
https://www.who.int/publications/i/item/who-sage-values-framework-for-the-allocation-and-prioritization-of-covid-19-vaccination. Accessed 14 Sep 2020.

Draft landscape and tracker of COVID-19 candidate vaccines document. In: World Health Organization. 2020.

https://www.who.int/docs/default-source/a-future-for-children/novel-coronavirus_landscape_covid-19.pdf?sfvrsn=4d8bd201_1. Accessed 10 Dec 2020.

Our World in Data website. https://ourworldindata.org/coronavirus-testing#china. Accessed 2 Jun 2021.
### Table 1. ADF test results

| Difference order | $t$       | $p$      | 1%    | 5%    | 10%   |
|------------------|-----------|----------|-------|-------|-------|
|                  |           |          |       |       |       |
| 0                | 0.519     | 0.985    | -3.503| -2.893| -2.584|
| 1                | -0.706    | 0.845    | -3.503| -2.893| -2.584|
| 2                | -5.54     | 0        | -3.503| -2.893| -2.584|

### Table 2. AIC, BIC, and HQIC values of the ARMA model

| Model     | AIC      | BIC      | HQIC    |
|-----------|----------|----------|---------|
| ARMA(3,3) | 2316.015 | 2336.361 | 2324.233|
| ARMA(2,2) | 2319.739 | 2334.999 | 2325.903|
| ARMA(2,3) | 2320.167 | 2337.97  | 2327.358|
| ARMA(3,2) | 2325.573 | 2343.377 | 2332.765|

### Table 3. Predicted results by the ARIMA(3, 2, 3) model

| Date       | Predicted value | Actual value | Absolute error | Relative error% | Variation coefficient% |
|------------|-----------------|--------------|----------------|-----------------|------------------------|
| 20210322   | 2010823         | 1973905      | 36918          | 1.87            | 0.93                   |
20210323  2248285  2077238  171047    8.23    3.95
20210324  2418990  2270243  148747    6.55    3.17
20210325  2527634  2816476  -288842   10.25    5.40
20210326  2645064  3453810  -808746   23.41    13.26

Note: Absolute error = Actual value – Predict value, whereas Relative error = Absolute error/Actual value

**Figure 1**

**Time series plots of the COVID-19 vaccination data in China.** (a) Time series plot of the original COVID-19 vaccination data, (b) Time series plot of the COVID-19 vaccination data after first-order difference, (c) Time series plot of the COVID-19 vaccination data after second-order difference
Figure 2

Plots of autocorrelation and partial autocorrelation before and after difference.

(a) Plots of autocorrelation and partial autocorrelation for the original series, (b) Plots of autocorrelation and partial autocorrelation after the first-order difference, (c) Plots of autocorrelation and partial autocorrelation after the second-order difference.
Figure 3

AIC thermodynamic diagram.
Figure 4

Results of the ARIMA model validation. (a) Plot of residual sequence, (b) Autocorrelation and partial autocorrelation plots of the residual sequence, (c) Q-Q plot of the standardized residual for the ARIMA model.
Figure 5

Performance of COVID-19 vaccination forecast.