RESEARCH

Forecasting COVID-19 cases using time series modeling and association rule mining

Rachasak Somyanonthanakul1, Kritsasith Warin2*, Watchara Amasiri3, Karicha Mairiang4, Chatchai Mingmalairak5, Wararit Panichkitkosolkul5, Krittin Silanun4, Thanaruk Theeramunkong6,7, Surapon Nitikraipot8 and Siriwan Suebnukarn9

Abstracts

Background: The aim of this study was to evaluate the most effective combination of autoregressive integrated moving average (ARIMA), a time series model, and association rule mining (ARM) techniques to identify meaningful prognostic factors and predict the number of cases for efficient COVID-19 crisis management.

Methods: The 3685 COVID-19 patients admitted at Thailand’s first university field hospital following the four waves of infections from March 2020 to August 2021 were analyzed using the autoregressive integrated moving average (ARIMA), its derivative to exogenous variables (ARIMAX), and association rule mining (ARM).

Results: The ARIMA (2, 2, 2) model with an optimized parameter set predicted the number of the COVID-19 cases admitted at the hospital with acceptable error scores ($R^2 = 0.5695$, $RMSE = 29.7605$, $MAE = 27.5102$). Key features from ARM (symptoms, age, and underlying diseases) were selected to build an ARIMAX (1, 1, 1) model, which yielded better performance in predicting the number of admitted cases ($R^2 = 0.5695$, $RMSE = 27.7508$, $MAE = 23.4642$). The association analysis revealed that hospital stays of more than 14 days were related to the healthcare worker patients and the patients presented with underlying diseases. The worsening cases that required referral to the hospital ward were associated with the patients admitted with symptoms, pregnancy, metabolic syndrome, and age greater than 65 years old.

Conclusions: This study demonstrated that the ARIMAX model has the potential to predict the number of COVID-19 cases by incorporating the most associated prognostic factors identified by ARM technique to the ARIMA model, which could be used for preparation and optimal management of hospital resources during pandemics.

Keywords: COVID 19, Pandemic, Data mining, Time series analysis, Association rule mining

Background

The crisis outbreak of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) started in Wuhan, Hubei Province, China in December 2019 [1]. The COVID-19 pandemic has required governments around the world to implement new policies under pressure from vulnerable people and communities [2]. Since the first outbreak, COVID-19 has mutated into many variants including the alpha, beta and delta SARS-COV-2 variants, which have been associated with new waves of infection [3]. The catastrophic effect across the entire world resulted in more than six million deaths worldwide in 2022 [4]. In addition, COVID-19 has caused a rapid deterioration in the condition of the disease, and the number of patients requiring hospitalization has increased significantly, resulting in a high demand for hospital resources [1].
Data mining is an efficient analytical methodology to recognize and investigate a huge data set to acquire meaningful information [5]. In the medical field, the large numbers of medical records (including demographic information, diagnoses, clinical notes, etc.) in the healthcare information systems are ideal targets for the use of data mining in improving the analysis and prognosis prediction of various diseases [6–8]. Examples include using an Artificial Neural Network (ANN) and Support Vector Machine (SVM) algorithm to predict cardiovascular disease [9], using data mining classification algorithms, Decision Tree and Naive Bayes algorithms to identify liver disease [10] and predict the recovery outcome of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [11]. With the unprecedented increase in COVID-19 cases worldwide, there is a need for effective prediction models to identify the associated prognostic factors and forecast the number of COVID-19 cases to optimally organize the hospital resources.

Time series analysis and association rule mining (ARM) models have been widely used to predict trends, structural breaks, cycles, and unobserved values, and have proven to be useful in the medical field [12–14]. The auto regressive integrated moving average (ARIMA), a time series analysis model, was shown to have a promising accuracy for forecasting of infectious diseases in medical fields [15, 16]. ARIMA was used to forecast the number of new COVID-19 cases, deaths, and recoveries based on the daily reported data from different countries for assessment of the future outbreak [17–20]. ARM was originally presented by Agrawal et al. as an algorithm for marketing data analysis [21]. ARM has been used to extract medical health information, which is currently being applied for the development of classification and prediction models to identify and forecast the possibility of development and progression of a disease by considering the rules of the disease [22]. ARM was demonstrated to be an effective model for mining the frequent symptom pattern for COVID-19 patients, which could assist clinicians in decision making [23]. Another study used ARM to analyze the patterns of different non-pharmaceutical interventions to manage the infection growth rate in the United States [24]. Even though there are many advanced data-driven time series methods used to predict the future number of COVID-19 patients, a new and more accurate prediction model is important in the pandemic crisis. The associated contributing factors should be considered to improve model performance. Therefore, the combination of ARM and ARIMA models by selecting the most associated prognostic rules and integrating with ARIMA models could increase the accuracy of predicting new cases to better understand the current situation and the progression of COVID-19, which can be easily used by society, organizations, or governments to assess and manage the crisis during the future outbreak.

The aim of this study was to evaluate the most effective combination of ARM techniques and ARIMA models to identify prognostic factors and predict the number of COVID-19 patients. These models are expected to allow for better preparation, organizing hospital resources of further such units and more optimal use of medical personnel and equipment to enhance healthcare decision-making to manage COVID-19 patients in this crisis situation.

**Methods**

**Administration protocol and data collection**

The study was conducted at Thailand’s first university-based field hospital. The field hospital was transformed from the service apartment style 14-story building of the university dormitory into a 494-bed facility for non-critical COVID-19 patients [25]. The field hospital was managed by the main university hospital and included the patients referred from the project’s five university hospitals and hospitals in the central area of Thailand. Sources of funding come mainly from the donations of university alumni, community groups and non-governmental organizations. Upon admission, a nurse records patient data in the COVID-19 screening of the field hospital information system; the patient undergoes a chest x-ray, blood tests for complete blood count (CBC), liver function tests (LFTs), electrolyte, balance urine nitrogen (BUN), and Creatinine (Cr). The doctor interprets the labs and chest x-ray, and records the results in the admission note. The patients are only admitted to the field hospital if they meet all of the following criteria: 1) asymptomatic, mild or moderate symptoms; 2) normal activities of daily living; 3) no important organ dysfunction; 4) no psychiatric history; and 5) resting pulse oxygen saturation (SpO2) > 95%. To avoid unnecessary contact between patients and medical personnel, the patient reports signs and symptoms, wants and needs via an internal field hospital application. Any consultation with the attending physician is done through a notification form. If the attending physician wishes to speak to the patient, the patient’s telephone number is obtained from the respective patient’s floor. All prescriptions must be made using a prescription form which will then be processed by the attending nurse and recorded in the progress note in the field hospital information system and in the university hospital electronic medical record system. In this field hospital system, the laboratory and radiographic examination would be performed on symptomatic COVID-19 patients with a history of taking Favipiravir and for severity assessment of symptomatic COVID-19 patients.
For Favipiravir-naïve patients: 1) A follow-up chest x-ray may be considered in patients with worsening signs and symptoms (body temperature (BT) > 38.0°C, cough, fatigue, SpO2 < 96%, or decreased SpO2 > 3% after a stress test); and 2) if the chest x-ray infers pneumonia with respiratory signs and symptoms (as mentioned in 1), refer the patient to the originating hospital for continued treatment with Favipiravir.

For patients previously treated with Favipiravir: 1) Follow-up by chest x-ray, LFTs; 2) if LFTs increase, consider consulting an ID specialist to terminate/adjust medication use; and 3) if the chest x-ray infers a progression of the infiltration accompanied by respiratory signs and symptoms (cough, fatigue, SpO2 < 96% and SpO2 drop > 3% after a stress test), consider referring the patient to the hospital of origin.

Asymptomatic patients who have been hospitalized for at least 14 days after a positive COVID-19 testing will be discharged home. The patients who received Favipiravir should fulfill all the following criteria: 1) The patients signs and symptoms have improved without progression of infiltration on chest x-ray; 2) BT < 37.8°C continuously for 24–48 hours; 3) respiratory rate (RR) < 20/min; and SpO2 > 96% at rest. In the event of a patient’s condition deteriorating, they are quickly transferred to the designated higher-level hospitals.

The criteria for transfer are 1) meeting the criterion of severe or critical, and 2) lung imaging showing a greater than 50% progression of lesions. Patients do not need Real-time Polymerase Chain Reaction (RT-PCR) or Antigen/Antibody detection for COVID-19 prior to discharge. One day before discharge, the attending nurse informs the attending physician of the number of potential discharges, so that the physician can prepare medical certificates and insurance documents according to the patient’s needs. Upon discharge, the attending physician updates the patient’s progress and discharge summary in the electronic medical record system of the university hospital.

A total number of 3685 patient records were retrieved from the electronic hospital information systems of the referral hospitals and the field hospital information system. In this study, we included all patients confirmed with asymptomatic and mild-to-moderate COVID-19 conditions from March 2020 to August 2021 (four waves of COVID-19 in Thailand). Collected data included patient demographics, comorbidities, body mass index (BMI), job, place of exposure to coronavirus, symptom before field hospital admission, sign of pneumonia in chest x-ray, field hospital length of stay, and the field hospital discharge destination. Table 1 shows the preliminary analysis of the dataset, including attributes, values, and frequency of each attribute-value pair.

Time-series analysis and association analysis
In this work, we present a study to combine time series analysis and association analysis to forecast the COVID-19 admitted cases as well as to analyze their potential factors and characteristics. To estimate the number of new cases and to predict the prognosis for better understanding of the current situation and progression of COVID-19, we exploited the autoregressive integrated moving average (ARIMA) model and its subclasses (i.e., AR, MA, ARMA) [12, 17, 26], and association rule mining (ARM) [21, 24] as tools for investigation (Fig. 1).

The autoregressive (AR) model
In the AR model, the predictive value at the time period t is modeled by the observed values at various time slots t − 1, t − 2, . . . , t − k. The impact of the value at each previous time period on the value at the current time is determined by the coefficient factor at that particular period of time. With this assumption, the model performs the regression of past time series and then calculates the present or future values in the series, commonly known as an auto regression (AR) model. It can be modeled as follows.

\[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \epsilon_t \]

Here, \( y_t \) is the value at the current time t, and \( y_{t-1}, y_{t-2}, \ldots, y_{t-p} \) are the observed values at the previous p time spots with their corresponding coefficients \( \beta_1, \beta_2, \ldots, \beta_p \), respectively, \( \beta_0 \) is the intercept, and \( \epsilon_t \) is the residual error at the time t. Therefore, \( y_t - \epsilon_t \) is the expected value at the current time t. In this work, the value \( y_t \) can be modeled as the number of inpatients, incoming patients, or outgoing patients at the time period t.

The moving-average (MA) model
Since the value of the time period t may be impacted by unexpected external factors, i.e., noises, we can alleviate such impact by means of the moving average method. Analogous to AR, the predicted value at the time period t can be modeled by the previous q lagged forecast errors \( \epsilon_t \) as follows.

\[ y_t = \phi_0 + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t \]

Here, \( y_t \) is the value at the current time t and the lagged errors \( \epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q} \) are residual errors of the q autoregressive models at time \( t-1 \) to \( t-q \) with \( \phi_1, \phi_2, \ldots, \phi_q \) as their corresponding coefficients, \( \phi_0 \) is the intercept, and \( \epsilon_t \) is the residual error at the time t.
The residual error at the time points after $t-1$ can be derived by the auto-regressive (AR) model as follows.

$$
\epsilon_{t-1} = y_{t-1} - (\beta_0 + \beta_1 y_{t-2} + \cdots + \beta_p y_{t-p-1}) \\
\epsilon_{t-2} = y_{t-2} - (\beta_0 + \beta_1 y_{t-3} + \cdots + \beta_p y_{t-p-2}) \\
\quad \quad \quad \quad \quad \quad \quad \cdots \\
\epsilon_{t-q} = y_{t-q} - (\beta_0 + \beta_1 y_{t-q-1} + \cdots + \beta_p y_{t-p-q})
$$

Although the standard AR and MA may use the auto-correlation function (ACF), which takes into account all of the points, it is possible to apply the partial auto-correlation function (PACF), which accounts for the values of the intervals between.

### Table 1  Preliminary analysis of the dataset: attributes, values, and frequency of each attribute-value pair

| No | Attribute name | Attribute value | Attribute code | Frequency |
|----|----------------|-----------------|----------------|-----------|
| 1  | Gender         | Male            | sex_male       | 1711      |
|    |                | Female          | sex_female     | 1974      |
| 2  | Age (year)     | Less than 24    | age_24         | 1148      |
|    |                | 25–44           | age_45_44      | 1838      |
|    |                | 45–64           | age_45_64      | 625       |
|    |                | More than 65    | age_65         | 74        |
| 3  | Body Mass Index| Less than 25    | bmi_25         | 2309      |
|    |                | 25–29           | bmi_25_29      | 931       |
|    |                | More than 30    | bmi_30         | 445       |
| 4  | Underlying     | None            | ud_none        | 3392      |
|    | Diseases       | Respiratory     | ud_repp        | 82        |
|    |                | Hypertension    | ud_ht          | 39        |
|    |                | Metabolic       | ud_meta        | 53        |
|    |                | Dyslipidemia    | ud_dlp         | 14        |
|    |                | Other           | ud_oth         | 64        |
|    |                | Diabetes mellitus| ud_dm        | 18        |
|    |                | Pregnant        | ud_preg        | 23        |
| 5  | Job            | General worker  | job_gen        | 3592      |
|    |                | Healthcare worker| job_health    | 93        |
| 6  | Source of infection | Community | source_com   | 3119      |
|    |                | Family          | source_fam     | 475       |
|    |                | Hospital        | source_hosp    | 91        |
| 7  | Symptom        | Asymptomatic    | symp_ast       | 2295      |
|    |                | Mild            | sym_mild       | 1371      |
|    |                | Moderate        | sym_mode       | 19        |
| 8  | Chest X-ray    | No lesion       | cxr_no         | 3213      |
|    |                | Pneumonia       | cxr_pneu       | 472       |
| 7  | Length of stay (Day) | Less than 14 | los_1_14      | 3625      |
|    |                | More than 14    | los_15         | 60        |
| 8  | Patient Discharge | Home discharge | dc_home       | 3600      |
|    |                | Refer to general hospital | dc_hosp    | 85        |
| 9  | Current Incidence | Wave 1 (MAR-MAY 2020) | wave_1       | 55        |
|    |                | Wave 2 (JAN-MAR 2021) | wave_2       | 311       |
|    |                | Wave 3 (APR-MAY 2021) | wave_3       | 1779      |
|    |                | Wave 4 (JUN-JUL 2021) | wave_4       | 1540      |

The autoregressive moving average (ARMA) model

The Auto Regressive Moving Average Model (ARMA) combines the AR and MA models. In ARMA, the impact of previous lags along with the residuals is considered for forecasting the future values of the time series as follows.

$$
y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} \\
+ \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t
$$

Here, $\beta_i$ represents the coefficients of the AR model, $\phi_i$ represents the coefficients of the MA model, and $\epsilon_t$ is the residual error at the time $t$. We assume only one
significant value from the AR model and one significant value from the MA model, so the ARMA model will be obtained from the combined values of these two models, denoted as the order of ARMA (1,1).

The autoregressive integrated moving average (ARIMA) model
As a generalization of AR, MA, and ARMA, the ARIMA model introduced differencing (integration) into the ARMA model to make the series stationary exploit to forecast future values under the factor of previous lag value and residuals errors. Besides manipulating the time lag and alleviating noise by smoothing, it is also possible to decompose a series into trend, seasonal, and residual components, by assuming an additive model. With this addition, the series can be transformed to a stationary time series. To achieve the transformation, the differencing method is applied. For example, we can subtract the \( t - 1 \) value from \( t \) values of time series. After applying the first differetntiation, if we are still unable to get the stationary time series, we can again apply the second-order differetntiation. The ARIMA model is an extension of the ARMA model by the fact that it includes one more factor known as integrated (i.e., differetntiation) which stands for \( I \) in the ARIMA model. The ARIMA model, denoted by ARIMA \((p,d,q)\), can be formulated as follows:

\[
y'_t = \beta_0 + \beta_1 y'_t + \beta_2 y''_t + \cdots + \beta_p y'_{t-p} + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t
\]

Here, \( p \) is the order of the autoregressive process, \( d \) (set to 1 in this case) is the degree of differetntiation (the number of times the series was differenced), and \( q \) is the order of the moving average component. In this model, the first-order difference \((d=1)\) between consecutive observations \( y'_t \) was computed and used, instead of the original observed value \( y_t \) as shown below.

\[
y'_t = y'_t - y'_{t-1}
\]

Differencing removes the changes in the level of a time series, eliminating trend and seasonality and, consequently, stabilizing the mean of the time series.

In some situations, we may need to difference the series data a second time \((d=2)\) to obtain a stationary time series, which is referred to as second order differencing as follows:

\[
y''_t = y''_t - y''_{t-1}
\]

A higher-order differencing can be pursued analogously in the same manner.
The autoregressive integrated moving average with exogenous covariables (ARIMAX) model

When an ARIMA model includes other time series as input variables, the model is referred to as an Autoregressive Integrated Moving Average with Exogenous Covariates (ARIMAX) model. An ARIMAX model can be viewed as a multiple regression model that takes the impact of covariates on the forecasting into account, improving the comprehensiveness and accuracy of the prediction. The ARIMAX(p,d,q) extends the ARIMA(p,d,q) model by including the linear effect that one or more exogenous series has on the stationary response series $y_t$. This method is suitable for forecasting when data is stationary/non-stationary, and multi-variate with any type of data pattern, i.e., level/trend/seasonality/cyclicity. The ARIMAX(p,d,q) model can be formulated as follows:

$$y'_t = \beta_0 + \beta_1 y'_{t-1} + \beta_2 y'_{t-2} + \cdots + \beta_p y'_{t-p} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q} + \varepsilon_t + \theta_1(X_1) + \theta_2(X_2) + \cdots + \theta_m(X_m) + \varepsilon_t$$

Here, $d$ is set to 1, $(X_i)$ is the value at the time $t$ of the i-th exogenous covariable $(X_i)$, $\theta_i$ is the corresponding coefficient for the covariable $X_i$, and $m$ is the number of exogenous covariables to be considered, while $p$, $d$, and $q$ indicate the same parameters as in the ARIMA model.

Association rule mining

Besides the time-series analysis, association rule mining (ARM) can be used as a multivariate analysis to help us understand the correlation among factors [24]. Given a dataset containing a collection of records or transactions, each record comprises a set of categorical attributes. An association rule can be denoted by $A \rightarrow B$, where $A$ (the antecedent or LHS) and $B$ (the consequent or RHS) are sets of various attribute-value pairs (also called itemsets), and are disjoint. The rule represents the hypothesis that when variables in $A$ occur in the dataset, the variables in $B$ also occur. Association mining generates a large number of rules from a given dataset. In a dataset with $m$ attributes $n-1$ antecedents and one consequent, each with $n$ values, each can generate a maximum of $nm^{n-1} - 1$ rules. However, not all rules are significant. The goal of this approach is to find rules that have high practical significance. To eliminate spurious rules, we use three measures: support, confidence, and lift. In addition, we also use the chi-squared test to measure the statistical significance of the association between the antecedent and the consequent. Given two disjoint sets of attribute-value pairs $A$ and $B$, and an association rule $A \rightarrow B$; support of the rule refers to the number of records where the attribute-value pairs in either set $A$ or $B$ appear in the dataset relative to the total number of records (transactions or instances). This denotes the prevalence of the rule in the dataset. By definition, the support value is symmetric, that is Support $(A \rightarrow B) = \text{Support} (B \rightarrow A)$, and it equals the total numbers of records containing both $A$ and $B$ to the total number of records in the dataset. The confidence of the rule $A \rightarrow B$ measures the conditional probability of $B$, given $A$. Thus, the confidence measure for a given rule is asymmetric, that is Confidence $(A \rightarrow B) \neq \text{Confidence} (B \rightarrow A)$. The lift measure is the ratio between the observed support and the expected support between the independent variables $A$ and $B$. Implicitly, lift $>1$ means a greater degree of dependence, lift $<1$ specifies negative dependence, and lift $=1$ indicates independence between $A$ and $B$. Lift is also a symmetric measure between the itemsets $A$ and $B$, that is Lift $(A \rightarrow B) = \text{Lift} (B \rightarrow A)$.

$$\text{Support}(A \rightarrow B) = \frac{|A \cap B|}{N}$$

$$\text{Confidence}(A \rightarrow B) = \frac{|A \cap B|}{|A|}$$

$$\text{Lift}(A \rightarrow B) = \frac{|A \cap B| \times N}{|A||B|}$$

Here, $|A|$ and $|B|$ are the numbers of records that include $A$ and $B$, respectively, while $|A \cap B|$ is the number of records that contain both $A$ and $B$. In this paper, the antecedent $A$ can be either patient demo-graphics (either male or female), age ($<25$, $25-44$, $45-64$, and $>65$), body mass index or BMI ($<25$, $25-29$, and $>29$), underlying diseases (none, respiratory, hypertension, metabolic, dyslipidemia, diabetes mellitus, pregnant, or others), job (healthcare or non-healthcare patient), infection source (community infection, family infection, or hospital infection), symptoms before field hospital admission (asymptomatic, mild, or moderate), sign of pneumonia in chest x-ray (no lesion or pneumonia) or length of stay in the field hospital ($14 \text{ or} >14$), and patient discharge (home discharge or refer to general hospital), as the contributing factors. On the other hand, for the consequent $B$ we focus on (1) the length of stay (either $1-14$ or $>14$), (2) the patient discharge (either home discharge or hospital discharge), (3) the chest x-ray result, and (4) current incidence (wave 1, 2, 3 or 4). Since one assumption for ARM is that all the values of attributes are discrete, we translate the numerical data used in the study into discrete labels, as well as split the continuous data of infection growth curve into four phases.
**Experiment settings**

**Data collection and parameter settings**

The dataset includes 3685 records registered with the electronic hospital information systems of the field hospital during March 2020 to August 2021. It displays characteristics of the dataset, including attributes, values, and frequency of each attribute-value pair. Each of the nine attributes contains 2–8 possible values. Most attributes have imbalanced numbers in their values, except gender (Table 1). In our time series analysis, the target of prediction is the number of patients in the field hospital for each day during the observation period, that is March 2020 to August 2021. We have explored the value of the three ARIMA parameters as \( p \in \{1, 2, 3\} \), \( d \in \{1, 2\} \), \( q \in \{1, 2, 3\} \) due to our preliminary test. In addition, we applied association rule mining to find the most influential factors among the eleven factors, that is patient demographics, age, body mass index, underlying diseases, job, inflection source, symptom before field hospital admission, sign of pneumonia in chest x-ray, length of stay in the field hospital, patient discharge, and current incidence. As an ARIMAX model, we extend the ARIMA(\( p, d, q \)) model to include the parameters as a series that are the most influential to the prediction of the number of patients in the hospital. The parameters included are known as exogenous series that are expected to trigger the stationary response on the series that we are predicting.

**Performance metrics and evaluation**

Given a data set has \( n \) values, denoted by \( y_1, \ldots, y_n \) each associated with a predicted value \( f_1, \ldots, f_n \), the following three metrics can be formulated. Coefficient of determination \((R^2)\) is the proportion of the variation in the dependent variable that is predictable from the independent variable(s) as follows:

\[
R^2 = 1 - \frac{SS_r}{SS_t} \quad (1)
\]

\[
SS_r = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \quad (2)
\]

\[
SS_t = \sum_i (y_i - \bar{y})^2 \quad (3)
\]

\[
\bar{y} = \frac{1}{n} \sum_i y_i \quad (4)
\]

Here, \( SS_r \) is the sum of squares of residuals, \( SS_t \) is the total sum of squares, proportional to the variance of the data, and \( \bar{y} \) is the mean of the observed data. Ranging from 0 to 1, it provides a measure of how well observed outcomes are replicated by the model. The higher the coefficient value is, the closer the dependent variable and independent variable are.

Root mean square error (RMSE) the standard deviation of the prediction errors [27], which are a measure of the distance of the data from the regression line, indicating the concentration of the data around the line of best fit as follows:

\[
RMSE = \sqrt{SS_r} = \sqrt{\frac{1}{n} \sum_i (y_i - f_i)^2} \quad (5)
\]

It expresses the dispersion of these errors.

Mean absolute error (MAE) allows measurement of the average magnitude of the errors for a set of predictions, regardless of their direction.

\[
MAE = \frac{1}{n} \sum_i |y_i - f_i| \quad (6)
\]

It represents the mean of the absolute difference in the sample between the prediction and the actual observation, taking into account that all individual differences are of equal significance. Therefore, compared to RMSE, MAE is less sensitive to outliers.

**Results**

**Time series analysis**

This section presents a time series analysis to forecast the number of patients admitted to the field hospital. Figure 2 shows the number of patients from 26 March 2020 to 22 July 2020. Three time series represent the relationships among a number of residents patients that are equal to a cumulative difference between admitted and discharged patients living in the hospital. The graph presents four waves of pandemic following the number of patients in hospital. The four waves are as follows: The first wave (Wave 1), the emergence of SAR-CoV-2, is the smallest period (34 days) from 26 March 2020 to 16 May 2020. The second wave (Wave 2) was from 11 January 2021 to 14 March 2020 (44 days). After that, the third wave (Wave 3) and fourth wave (Wave 4) were the continuous periods from 11 April 2021 to 31 May 2021 (51 days) and 1 June 2021 to 22 July 2021 (52 days), respectively. Finally, the forecasting models are validated by a test dataset from 1 August 2021 to 30 August 2021(30 days).

In this study, the time series models were trained using six training datasets. The first training set (All Wave) covers all datasets Wave 1 to Wave 4 of 228 days; the second training set, Wave 1 of 34 days; the third training set, Wave 2 of 45 days; the fourth training set, Wave 3 of 51 days; the fifth training set, Wave 4 of 52 days; the sixth training set, Wave 3 and Wave 4 of 103 days.
In this work, we tested the estimated model using an autocorrelation function (ACF) and a partial autocorrelation function (PACF) plots to ensure that the model fits the data [17]. Figure 3 presents the steady-state prediction of time-series models. An estimation of the model explored the coefficient (Coef.), the standard error (Std err.) and z. An estimate of the first model was the AR model which gave a coefficiency of 0.3808, standard error of 0.243 and z of 1.565. The second model was an MA model which gave coefficiency of −0.5287, standard error of 6.841 and z of −0.077. The sigma value or constant value was coefficiency of −0.5287, standard error of 6.841 and z of −0.077. Moreover, we further estimated the model with Jarque-Bera of 7.70, heteroskedasticity of 0.57 and skew of 0.68.

For the data set, the time series method was applied using Python (PyFlux library) for time series analysis and prediction to compare the criteria of each setting. The ARIMAX \((p,d,q)+X\) models were parameterized with \(X \in \{\phi, x_1, x_2\}, p \in \{0, 1, 2, 3\}, q \in \{0, 1, 2, 3\}, d \in \{0, 1, 2\}\), where \(X\) is additional exogenous variables, with 51 combinations. Moreover, we select key features from association rule mining such as symptoms, age, and underlying diseases, etc. \(X=\phi\) specifies no additional exogenous variable used. \(X=x_1\) indicates additional exogenous variables. There are 15 variables, composed of three attributes in the symptom feature, four attributes in the age feature, and eight attributes in the underlying diseases feature. \(X=x_2\) represents four variables of the selected attributes, that is the ‘moderate’ symptom, the ‘more-than-65’ age, and the underlying diseases of ‘diabetes mellitus’ and ‘pregnant’.

The forecasting-accuracy metrics of the 51 models summarized on the six datasets and the evaluation of models with the measures of RMSE and MAE are shown in Table 2. The forecasts for the admitted patients with prediction confidential intervals (CI) between 5 and 95% are presented in Fig. 4 for ARIMA (2,2,2) and Fig. 5 for ARIMAX (1,1,1)+ \(x_2\). Overall, the most accurate estimation was obtained by improving from ARIMA (2, 2, 2) to ARIMAX (1, 1, 1) + \(x_2\) for the training set in Wave 4, covering from 11 April 2021 to 31 May 2021. For the first setting (All-Wave), the best model is ARIMA (1,2,1) with the RMSE of 22.8141 and MAE of 19.4133, which was closer to the actual data. For Wave-1, ARIMAX (2,2,2) + \(x_2\) performs the best with the RMSE of 277.9974 and MAE of 273.4644, which was the highest to the actual data of all models. For Wave-2, AR(1) + \(X1\) model is the best with the smallest RMSE and MAE. Based on RMSE and MAE, the value of ARIMA (1,1,1) + \(X1\) was the closest to the actual data in Wave-3. The RMSE and MAE of ARIMAX (1,1,1) + \(X2\) appeared to be the best predictive models.

The comparisons among forecasting models are shown in Tables 3, 4 and 5. The models numbered 12–17 in Table 2 are defined to be the baseline models. The models with \(x_1\) are the models numbered 29–34 while the models with \(x_2\) are the models numbered 46–51. The compared

![Fig. 2 The number of daily data of patients in the field hospital; New patients; Admitted Patients; Discharged Patients in four waves of COVID-19 pandemics in Thailand](image-url)
pairs were (baseline vs $x_1$), ($x_1$ vs $x_2$), and (baseline vs $x_2$). The comparison was done under the same parameter setting. The result of $R^2$, RMSE and MAE (Tables 3, 4 and 5) yielded a good result indicating that time forecasting models could improve correlation of determination when we added exogenous variables.

The predicted values, CI 5% (lower confidence interval) and CI 95% (upper confidence interval), and actual data of the models are shown in Table 6 and Fig. 4. In addition, the improved predictive values of the models by adding exogenous variables are shown in Table 7 and Fig. 5. For example, ARIMA (2, 2, 2) predicted that the number of cumulative confirmed cases for the next 30 days could be 291 to 334 cases. ARIMAX (1, 1, 1) + $x_2$ predicted that the number of cumulative confirmed cases for the next 30 days could be 293–330 cases.
| No | Model                  | RMSE  | MAE  | RMSE  | MAE  | RMSE  | MAE  | RMSE  | MAE  | RMSE  | MAE  | RMSE  | MAE  |
|---|------------------------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|
| 1 | (1)                    | 0.898 | 0.199| 0.994 | 0.199| 0.777 | 0.188| 0.898 | 0.199| 0.994 | 0.199| 0.777 | 0.188|
| 2 | (2)                    | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 3 | AR(1)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 4 | AR(2)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 5 | AR(3)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 6 | MA(1)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 7 | MA(2)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 8 | MA(3)                  | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 9 | ARMA(1,1)              | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 10| ARMA(2,2)              | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 11| ARIMA(1,2,1)           | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 12| ARIMA(2,2,2)           | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 13| ARIMA(3,3,3)           | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 14| MAX(1,1,1)             | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 15| MAX(2,1,2)             | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
| 16| MAX(3,1,3)             | 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127| 0.335 | 0.134| 0.434 | 0.134| 0.306 | 0.127|
Table 2 (continued)

| No | Model | All Wave | Wave 1 | Wave 2 | Wave 3 | Wave 4 | Wave 3–4 |
|----|-------|----------|--------|--------|--------|--------|---------|
|    |       | R2 | RMSE | MAE | R2 | RMSE | MAE | R2 | RMSE | MAE | R2 | RMSE | MAE | R2 | RMSE | MAE |
| 33 | ARI-MAX(2,2,2)+X1 | 0.2168 | 95.7906 | 88.9073 | 0.3083 | 316.2992 | 315.1248 | 0.0490 | 287.2771 | 283.2256 | 0.5580 | 297.0652 | 27.6102 | 0.5853 | 29.7605 | 27.6102 |
| 34 | ARI-MAX(3,2,3)+X1 | 0.4787 | 63.5331 | 57.3825 | 0.1452 | 336.3347 | 332.9383 | 0.0884 | 278.0144 | 273.4693 | 0.0000 | 99.6750 | 75.6992 | 0.0016 | 144.7672 | 139.9775 |
| 35 | I(1)+X2 | 0.4787 | 63.5331 | 57.3825 | 0.1452 | 336.3347 | 332.9383 | 0.0884 | 278.0144 | 273.4693 | 0.0000 | 99.6750 | 75.6992 | 0.0016 | 144.7672 | 139.9775 |
| 36 | I(2)+X2 | 0.0105 | 249.4904 | 231.4904 | 0.0421 | 326.0146 | 323.9672 | 0.0192 | 284.7981 | 279.4427 | 0.0192 | 284.7981 | 279.4427 | 0.0192 | 284.7981 | 279.4427 |
| 37 | AR(1)+X2 | 0.0447 | 289.1389 | 281.2524 | 0.0041 | 326.0347 | 323.9383 | 0.0884 | 278.0144 | 273.4693 | 0.0000 | 99.6750 | 75.6992 | 0.0016 | 144.7672 | 139.9775 |
| 38 | AR(2)+X2 | 0.0105 | 249.4904 | 231.4904 | 0.0421 | 326.0146 | 323.9672 | 0.0192 | 284.7981 | 279.4427 | 0.0192 | 284.7981 | 279.4427 | 0.0192 | 284.7981 | 279.4427 |
| 39 | AR(3)+X2 | 0.0584 | 64.0891 | 52.3809 | 0.5545 | 338.2531 | 336.1401 | 0.0007 | 293.3307 | 290.4134 | 0.0131 | 70.7792 | 58.5222 | 0.3373 | 82.2578 | 62.6506 |
| 40 | MA(1)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 41 | MA(2)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 42 | MA(3)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 43 | ARMA(1,1)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 44 | ARMA(2,2)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 45 | ARMA(3,3)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 46 | ARMA(1,1)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 47 | ARMA(2,2)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |
| 48 | ARMA(3,3)+X2 | 0.0670 | 183.8616 | 168.7598 | 0.0392 | 329.9775 | 327.7278 | 0.0091 | 289.6989 | 286.3389 | 0.0011 | 72.7548 | 60.1999 | 0.0091 | 72.7548 | 60.1999 |

* the best time series analysis model performance
Fig. 4 The ARIMA (2,2,2) forecasting value of the admitted patients with prediction confidential intervals (CI) between 5 and 95%

Fig. 5 The ARIMAX (1, 1, 1)+ X2 forecasting value of the admitted patients with prediction confidential intervals (CI) between 5 and 95%

Table 3 The comparison of Coefficient of determination ($R^2$)

| No | $R^2$ Comparison | All Wave | Wave 1 | Wave 2 | Wave 3 | Wave 4 | Wave 3–4 |
|----|-----------------|---------|--------|--------|--------|--------|---------|
|    |                 | Win     | Loss   | Win    | Loss   | Win    | Loss    | Win     | Loss   |
| 1  | baseline vs $x_1$ | 3       | 3      | 6      | 0      | 1      | 5       | 5       | 1      | 0      | 6      | 6      | 0      |
| 2  | $x_1$ vs $x_2$   | 2       | 4      | 3      | 3      | 2      | 4       | 2       | 4      | 4      | 2      | 3      | 3      |
| 3  | baseline vs $x_2$| 4       | 2      | 6      | 0      | 0      | 6       | 4       | 2      | 4      | 2      | 4      | 2      |
|    | SUM              | 9       | 9      | 15     | 3      | 3      | 15      | 11      | 7      | 8      | 10     | 13     | 5      |
Association rule mining
This section explores the association analysis when association rule mining is applied. We present significant rules for the data that included four attributes’ values in the dataset. Table 1 shows preliminary analysis of dataset that was extracted for a total of 3685 patients. The patient data consist of eleven attributes and 35 attribute values. In addition, an attribute code is defined for item set name and frequency of each attribute code. We extract 595 significant rules for the data.

The association rules grouped by four attributes related to managing hospital resources are shown in Table 8. Length of stay more than 14 days is related to healthcare workers and three underlying diseases other, pregnant, and dyslipidemia that have the same value of 1.017. Length of stay less than 14 give the interesting result on symptom mode (Lift of 6.464), three underlying diseases, and age more than 65 years old.

The interesting rule of discharge had two value attributes. The result showed that referral to hospitals was strongly related to symptom of Mode (Lift of 9.127). In addition, four features in this attribute showed high Lift values; underlying diseases (5.655), metabolic syndrome (4.098), length of stay more than 14 days (3.613), and age more than 65 years old (5.515). Chest x-ray with no lesion presented the same level of Lift. However, two features which showed high numbers of patients were age less than 24 years old (1148) and symptom asymptomatic (2295). Moreover, chest x-ray with pneumonia showed all high interesting value Symptom of Mode (3.287), age more than 65 (3.271), underlying diseases diabetes mellitus (2.169), and underlying diseases Metabolic (2.062). In current incident, Wave 1 showed high interest on Length of stay more than 14-days and source of infection from hospital and healthcare worker patients. Wave 2 was also related to healthcare worker, asymptomatic and source of infection from hospital, as was Wave 3. In Wave 4, underlying diseases, age more than 65 and symptom mode showed strong relationships. Association rules selected key attributes of the data set to be exogenous variables of a time series analysis.

Discussion
The first wave of SARS-CoV-2 occurred in early 2020, and the second, third and fourth waves rapidly spread from early to mid-2021, representing an unprecedented phenomenon in medical services, society and the economy of Thailand. The number of COVID-19 patients shown in this study increased from the first wave of just 55 patients to 311, 1779 and 1540 in the second, third and fourth waves, respectively, which evolved more than 30 times of the total number of patients admitted at the field hospital. Most of patients were at least 44 years old and were predominantly female. Patients included in this study were mostly asymptomatic and had no sign of pneumonia in the chest x-ray due to the field hospital system’s focus on patients who did not require advanced treatment. But during the third and fourth waves, the number of mild to moderate symptoms with pneumonia of COVID-19 patients significantly increased because of the greater severity of the delta variant of SARS-COV-2. The huge number of patients was a burden on the limited resources of Thailand’s healthcare system. Therefore, this study presented the use of time series modeling and
Association rule mining to forecast the COVID-19 pandemic outbreak as well as to analyze its associated prognostic factors. The method presented a data-oriented approach that applies time-series analysis and association analysis to reveal meaningful hidden patterns for efficient handling of another pandemic crisis.

ARIMA models have been successfully applied for predicting the disease outbreak. Several studies have utilized the ARIMA model to forecast the spread of COVID-19 in many countries including the US, Brazil, India, Russia and Spain [28, 29]. The studies using ARIMA models to predict COVID-19 cases relative to total confirmed cases presented an average RMSE of 144.81 across 6 geographic regions [28], MAE of 787 to 1506 in USA and 82 to 570 in Italy [18], and MAE of 2967 in Indonesia [20].

In this work, ARIMA (2, 2, 2) was selected as the most accurate ARIMA model for predicting the number of admitted COVID-19 cases in the field hospital, which achieved a $R^2 = 0.5695$, RMSE = 29.7605, MAE = 27.5102 (Fig. 4). The forecast results of admitted cases on August 15 and August 30, 2021 were 335 and 294, respectively. In comparison with the actual values reported on the same dates, the forecasted values of our selected ARIMA model were within the upper and lower bounds at 95% confidence intervals. This signified an acceptable accuracy of this model for estimating admitted cases in the field hospital.

ARM is a structured method of discovering frequent patterns in a data set and forming noticeable rules among regular patterns. In the COVID-19 crisis, many nations,

**Table 6** The number of patient prediction for time-series model ARIMA (2, 2, 2) + X2 Training from May 1 to July 22, 2021, Prediction from August 1 to August 30, 2021

| Date         | Actual data | Prediction | Lower CI | Upper CI |
|--------------|-------------|------------|----------|----------|
| August 1, 2021 | 334        | 361        | 327      | 394      |
| August 2, 2021 | 339        | 313        | 279      | 347      |
| August 3, 2021 | 347        | 326        | 292      | 361      |
| August 4, 2021 | 361        | 346        | 311      | 380      |
| August 5, 2021 | 387        | 364        | 330      | 398      |
| August 6, 2021 | 404        | 395        | 361      | 430      |
| August 7, 2021 | 393        | 411        | 377      | 445      |
| August 8, 2021 | 384        | 386        | 351      | 420      |
| August 9, 2021 | 381        | 371        | 337      | 405      |
| August 10, 2021 | 391      | 372        | 338      | 406      |
| August 11, 2021 | 386       | 390        | 356      | 424      |
| August 12, 2021 | 382       | 381        | 348      | 415      |
| August 13, 2021 | 376       | 375        | 342      | 408      |
| August 14, 2021 | 350       | 368        | 335      | 401      |
| August 15, 2021 | 340       | 335        | 302      | 368      |
| August 16, 2021 | 328       | 328        | 295      | 361      |
| August 17, 2021 | 296       | 319        | 286      | 352      |
| August 18, 2021 | 301       | 280        | 247      | 313      |
| August 19, 2021 | 301       | 295        | 262      | 328      |
| August 20, 2021 | 331       | 301        | 268      | 334      |
| August 21, 2021 | 327       | 342        | 309      | 375      |
| August 22, 2021 | 313       | 331        | 297      | 364      |
| August 23, 2021 | 306       | 305        | 272      | 338      |
| August 24, 2021 | 301       | 299        | 266      | 332      |
| August 25, 2021 | 301       | 297        | 264      | 330      |
| August 26, 2021 | 294       | 300        | 267      | 332      |
| August 27, 2021 | 297       | 291        | 259      | 323      |
| August 28, 2021 | 303       | 296        | 264      | 328      |
| August 29, 2021 | 296       | 305        | 273      | 337      |
| August 30, 2021 | 291       | 294        | 262      | 326      |

CI confidence interval

**Table 7** The number of patient prediction for time-series model ARIMAX (1,1,1) + X2 Training from May 1 to July 22, 2021, Prediction from August 1 to August 30, 2021

| Date         | Actual data | Prediction | Lower CI | Upper CI |
|--------------|-------------|------------|----------|----------|
| August 1, 2021 | 334        | 330        | 293      | 365      |
| August 2, 2021 | 339        | 333        | 296      | 368      |
| August 3, 2021 | 347        | 342        | 305      | 380      |
| August 4, 2021 | 361        | 345        | 307      | 382      |
| August 5, 2021 | 387        | 362        | 325      | 399      |
| August 6, 2021 | 404        | 391        | 353      | 428      |
| August 7, 2021 | 393        | 404        | 366      | 441      |
| August 8, 2021 | 384        | 385        | 348      | 422      |
| August 9, 2021 | 391        | 379        | 343      | 416      |
| August 10, 2021 | 391      | 377        | 341      | 414      |
| August 11, 2021 | 386       | 392        | 355      | 428      |
| August 12, 2021 | 382       | 380        | 344      | 416      |
| August 13, 2021 | 376       | 379        | 343      | 415      |
| August 14, 2021 | 350       | 371        | 336      | 407      |
| August 15, 2021 | 340       | 340        | 304      | 375      |
| August 16, 2021 | 328       | 338        | 302      | 373      |
| August 17, 2021 | 296       | 322        | 287      | 358      |
| August 18, 2021 | 301       | 290        | 250      | 321      |
| August 19, 2021 | 301       | 305        | 270      | 341      |
| August 20, 2021 | 331       | 298        | 263      | 333      |
| August 21, 2021 | 327       | 335        | 304      | 376      |
| August 22, 2021 | 313       | 321        | 285      | 356      |
| August 23, 2021 | 306       | 309        | 273      | 344      |
| August 24, 2021 | 301       | 304        | 269      | 339      |
| August 25, 2021 | 301       | 299        | 264      | 334      |
| August 26, 2021 | 294       | 301        | 266      | 335      |
| August 27, 2021 | 297       | 291        | 256      | 325      |
| August 28, 2021 | 303       | 298        | 264      | 332      |
| August 29, 2021 | 296       | 303        | 269      | 338      |
| August 30, 2021 | 291       | 293        | 259      | 327      |

CI confidence interval
### Table 8: Top 5 association rules for different combinations of particular consequence, their Support, Average-confidence, Confidence (LHS ➔ RHS), Confidence (RHS ➔ LHS) and Lift measures

| No | LHS                  | RHS                  | N(A) | N(B) | N(A,B) | Sup_{LR} | Conf_{A} | Conf_{LR} | Conf_{RL} | Lift_{LR} |
|----|----------------------|----------------------|------|------|--------|----------|----------|----------|----------|-----------|
| 1  | job_health           | los_1_14             | 93   | 3625 | 93     | 2.524    | 51.283   | 100.000  | 2.566    | 1.017     |
| 2  | ud_oth               | los_1_14             | 64   | 3625 | 64     | 1.737    | 50.883   | 100.000  | 1.766    | 1.017     |
| 3  | ud_preg              | los_1_14             | 23   | 3625 | 23     | 0.624    | 50.317   | 100.000  | 0.634    | 1.017     |
| 4  | ud_dlp               | los_1_14             | 14   | 3625 | 14     | 0.380    | 50.193   | 100.000  | 0.386    | 1.017     |
| 5  | cxr_pneu             | los_1_14             | 472  | 3625 | 470    | 12.754   | 56.271   | 99.576   | 12.966   | 1.012     |

|   |                      |                      |      |      |        |          |          |          |          |           |
|---|----------------------|----------------------|------|------|--------|----------|----------|----------|----------|-----------|
| 1  | Length of Stay less than or equal to 14 days |                      |      |      |        |          |          |          |          |           |
| 2  | Length of Stay more than or equal to 15 days |                      |      |      |        |          |          |          |          |           |
| 3  | Home Discharge       |                      |      |      |        |          |          |          |          |           |
| 4  | Refer to General hospital |                  |      |      |        |          |          |          |          |           |
| 5  | Chest X-ray is No lesion |                |      |      |        |          |          |          |          |           |
| 6  | Chest X-ray is Pneumonia |                 |      |      |        |          |          |          |          |           |
| 7  | Current incidence in Wave 1 |               |      |      |        |          |          |          |          |           |
| 8  | Current incidence in Wave 2 |             |      |      |        |          |          |          |          |           |
including Thailand, have a highest priority to save lives and protect their economies. A previous study using ARM for mining COVID-19 data to analyze factors related to COVID-19 situation management showed that face mask mandates combined with mobility reduction through moderate stay-at-home orders were most effective in reducing the number of COVID-19 cases in United State [24]. In this study, the ARM technique was used to analyze and identify factors related to the length of stay and prognosis of COVID-19 patients and found that the top five factors related to hospital stays longer than 14 days consisted of healthcare workers uncommon underlying diseases such as thalassemia, thyroid diseases, gout and G6PD deficiency, pregnant patients, dyslipidemia and signs of pneumonia in chest x-rays. This study also identified a clinical factor rule related to the worsening condition of the inpatient. Among those who needed more advanced medical treatment, the rules included mild to moderate COVID-19 symptoms, pregnant patients, metabolic syndrome, length of hospital stay more than 14 days, and patients older than 65 years old. These factors are consistent with those in a previous study, which reported similar conditions among patients who had a poor prognosis in COVID-19 infections [1, 30].

In any prediction tasks, more data is needed to achieve better performance from the models. This study developed the combination of the ARM technique and the ARIMA model, as the ARIMAX model. This model worked by selecting the rules related to COVID-19 prognosis from the ARM technique, including mild to moderate COVID-19 symptoms, patients with metabolic syndrome and patients older than 65 years old, and integrating them to the ARIMA model. Experimental results showed that the ARIMAX model (1, 1, 1) improved the accuracy of forecasting the number of admitted COVID-19 cases, which achieved a $R^2 = 0.5695$, RMSE = 27.7508, MAE = 23.4642 (Fig. 5). The forecast value of this model for August 30, 2021 was estimated to be 259 to 327 cases. The actual number of cases on the same date was 291 cases. The actual value also was within the lower and upper prediction bounds for both 95% confidence intervals. To the best of our knowledge, this is the first study to combine the ARM technique with the ARIMA model for forecasting the COVID-19 cases by integrating the optimal exogenous variables from the ARM rules to form a predictive model. This ARIMAX model had the potential to predict the number of COVID-19 patients, which could be one of the reliable forecasting-based models for the future outbreak. These predictive models are intended to help better decision-making to plan an effective management system if the virus outbreak has not subsided.

**Limitations**
The limitation of this study is that the dataset was based on retrospective data from a single COVID-19 field hospital in Thailand with a limited number of cases and clinical variables of COVID-19 patients.

**Future directions**
In future work, the collaboration between multi-medical centers for a larger number and different variables of COVID-19 cases, including the medical records of clinical, laboratory and treatment data from various COVID-19 centers, would upgrade the forecasting performance of this AI model to predict the COVID-19 event more accurately. Additionally, geographic data related to the pandemic area could be used as a variable for alternative time series models such as space-time ARIMA models [31], which could be more reliable in predicting future COVID-19 outbreaks.
Conclusion
This study demonstrated that the ARIMAX model has the potential to increase the accuracy for predicting the number of COVID-19 cases by incorporating the most associated prognostic factors identified by ARM technique to the ARIMA model. The result of this study proved to be an effective AI model to predict the number of and to identify prognostic factors of admitted COVID-19 patients. This work is expected to be a novel AI-based decision-making model for preparation, organizing hospital resources and more optimal use of medical personnel and equipment to enhance healthcare decision-making, and to manage the COVID-19 pandemic but as well as other epidemic crises.

Abbreviations
COVID-19: Coronavirus disease 2019; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus-2; MERS-CoV: Middle East Respiratory Syndrome Coronavirus; CBC: Complete blood count; LFTs: Liver function tests; BUN: Balance urine nitrogen; Cr: Creatinine; SpO2: Pulse oxygen saturation; BT: Body temperature; BMI: Body mass index; G6PD: Glucose-6-Phosphate Dehydrogenase; ANN: Artificial Neural Network; SVM: Support Vector Machine; ARM: Association Rule Mining; ARIMA: Auto Regressive Integrated Moving Average; ARIMAX: Autoregressive Integrated Moving Average with Exogenous Covariates; R²: Coefficient of determination; RMSE: Root mean square error; MAE: Mean absolute error; CI: Confidence intervals.

Acknowledgements
The authors would like to thank Supasek Sanmano from Thammasat Field Hospital and Kunch Ringrod from Thai Network for Disaster Resilience (TNDR) for data preparation. We thank Mr. Terrance J. Downey, English Editor for Thammasat University Office of Research and Innovation for English language editing.

Authors’ contributions
Conceptualization: K.W., S.N., S.S.; Methodology: R.S., K.W., T.T.; Formal analysis and investigation: R.S., K.W., W.A., W.P., T.T., S.S.; Fund acquisition: W.A., C.M., K.M., K.S.; Supervision: S.S., S.N. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by the Thammasat University Research Fund (CovidTU-03/2564), Center of Excellence in Intelligent Informatics and Service Innovation (CILS), and Intelligent Informatics and Service Innovation (IIS).

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable requests.

Declarations
Ethics approval and consent to participate
The study protocol and the exempt from the need to obtain informed consent was approved by the Ethics Committee of the Thammasat University (COE 008/2564) in accordance with the 1964 Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
1 College of Digital Innovation Technology, Rangsit University, Pathum Thani 12000, Thailand. 2 Faculty of Dentistry, Thammasat University, Pathum Thani 12121, Thailand. 3 Faculty of Engineering, Thammasat University, Pathum Thani 12121, Thailand. 4 Faculty of Medicine, Thammasat University, Pathum Thani 12121, Thailand. 5 Faculty of Science and Technology, Thammasat University, Pathum Thani 12121, Thailand. 6 Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12121, Thailand. 7 Academy of Science, Royal Society of Thailand, Sanam Sueapa, Khet Dusit, Bangkok 10300, Thailand 8 Thammasat University Hospital, Pathum Thani 12121, Thailand. 9 Research and Innovation Division, Thammasat University, Pathum Thani 12121, Thailand.

Received: 2 December 2021 Accepted: 14 October 2022
Published online: 01 November 2022

References
1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497–506.
2. Wolkeiwitz M, Pulkaj L. Methodological challenges of analysing COVID-19 data during the pandemic. BMC Med Res Methodol. 2020;20(1):181.
3. Tao K, Tzou PL, Noushi J, Gupta RK, de Oliveira T, Kosakovsky Pond S., et al. The biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet. 2021;22(12):757–73.
4. World Health Organization: COVID-19 Weekly Epidemiological Update, Edition 95, 2022.
5. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012;36(4):2431–48.
6. Huang F, Wang S, Chan C. Predicting disease by using data mining based on healthcare information system. In: 2012 IEEE International Conference on Granular Computing: 11–13 Aug. 2012. vol. 2012; 2012. p. 191–4.
7. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag. 2005;19(2):64–72.
8. Kriston L. Predictive accuracy of a hierarchical logistic model of cumulative SARS-CoV-2 case growth until May 2020. BMC Med Res Methodol. 2020;20(1):278.
9. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health. 2019;19(1):448.
10. Alfsahin SNN, Mantoro T. Data Mining Techniques for Optimization of Liver Disease Classification. In: 2013 International Conference on Advanced Computer Science Applications and Technologies: 23–24 Dec. 2013, vol. 2013; 2013. p. 379–84.
11. Al-Turaiki I, Alshahrani M, Almutairi T. Building predictive models for MERS-CoV infections using data mining techniques. J Infect Public Health. 2016;9(6):744–8.
12. Abonzuel M, Ibrahim A. Forecasting Egyptian GDP using ARIMA models. Rep Econ Finance. 2019;5(5):35–47.
13. Cryer JD, Chan K-S. Time series analysis with applications in R, 2nd 2008. Edn. New York: Springer New York; 2008.
14. Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
15. Heisterkamp SH, Dekkers AL, Heijne JC. Automated detection of infectious disease outbreaks: hierarchical time series models. Stat Med. 2006;25(24):4179–96.
16. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159–75.
17. Abonzuel M, Darwisc N. Forecasting confirmed and recovered Covid-19 cases and deaths in Egypt after the genetic mutation of the virus: ARIMA box-Jenkins approach. Commun Math Biol Neurosci. 2022;2022:17.
18. Gecili E, Ziad M, Szczesniak RD. Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy. PLoS One. 2021;16(1):e0244173.
19. Singh S, Parmar KS, Makkhan SJ, Kaur J, Peshoria S, Kumar J. Study of ARIMA and least square support vector machine (LS-SVM) models for the
prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos, Solitons Fractals. 2020;139:110086.

20. Aditya Satrio CB, Darmawan W, Nadia BU, Hanafiah N. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Proc Comput Sci. 2021;179:524–32.

21. Agrawal R, Imeliriski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data. Washington, D.C.: Association for Computing Machinery; 1993. p. 207–16.

22. K S L, G DV: Extracting association rules from medical health records using multi-criteria decision analysis. Proc Comput Sci 2017, 115:290–295.

23. Tandan M, Acharya Y, Pokharel S, Timilsina M. Discovering symptom patterns of COVID-19 patients using association rule mining. Comput Biol Med. 2021;131:104249.

24. Katragadda S, Gottumukkala R, Bhupatiraju RT, Kamal AM, Raghavan V, Chu H, et al. Association mining based approach to analyze COVID-19 response and case growth in the United States. Sci Rep. 2021;11(1):18635.

25. Amasiri R, Warin K, Mairiang K, Mingmalarik C, Panichkirkosollik W, Silanun K, et al. Analysis of characteristics and clinical outcomes for crisis management during the four waves of the COVID-19 pandemic. Int J Environ Res Public Health. 2021;18(23):12633.

26. Time Series Models AR, MA, ARMA, ARIMA; 2020 [cited 2021 7 December] Available from: https://towardsdatascience.com/time-series-models-d9266d8ac7b0.

27. Barnston AG. Correspondence among the correlation, RMSE, and Heidke forecast verification measures, refinement of the Heidke score. Weather Forecast. 1992;7(4):699–709.

28. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Appl Soft Comput. 2020;96:106610.

29. Darapaneni N, Reddy D, Paduri AR, Acharya P, Nithin HS. Forecasting of COVID-19 in India Using ARIMA Model. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON): 28–31 Oct. 2020, vol. 2020, p. 0894–9.

30. Noor FM, Islam MM. Prevalence and associated risk factors of mortality among COVID-19 patients: a Meta-analysis. J Community Health. 2020;45(6):1270–82.

31. Awwad FA, Mohamoud MA, Abonazel MR. Estimating COVID-19 cases in Makkah region of Saudi Arabia: space-time ARIMA modeling. PLoS One. 2021;16(4):e0250149.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.