Original Article

Forecasting the Number of Patients in the Intensive Care Unit Due To Covid-19 in Turkey by Multiple Linear Regression and Holt Time-series Methods

Sema ÜZÜLMEZ [1*], Zafer ASLAN [1]

Abstract

Objective: A novel coronavirus disease (COVID-19) pandemic originated in Wuhan, China, but has now spread around the globe. Data mining has been extensively employed throughout the pandemic's process concerning huge public health and many major consequences in the social, cultural, economic, political, legal, and military spheres. Using data mining methods, provinces, regions, and nations determined cases, mortality rates, and epidemic spread rates. Patients with severe disease suffered substantial respiratory distress and were treated in intensive care units. Hospital demand for intensive care surged due to the increasing number of intensive care units.

Methods: Due to the COVID-19 pandemic in Turkey, this study tried to anticipate the number of patients admitted to the intensive care unit between March 27, 2020, and June 5, 2020, using data from the Turkish Ministry of Health. The study compared methods using multiple linear regression and Holt Time Series analysis.

Results: The R2 value of the multiple regression method was calculated as 0.93. Since the value of 0.93 for the R2 result gives a value close to 1, the model fit is significant. According to the analysis results, the RMSE values taken from the established models are as follows. The multiple Linear Regression RMSE value is 22.221, Holt Time Series RMSE value is 39.815. Accordingly, since the Regression model produces realistic and meaningful data in the 71-day intensive care patient number in a 7-day study, the established model fit produces better forecast results than the Holt time series method.

[1] *Istanbul Aydin University, Faculty of Engineering, Department of Computer Engineering, Istanbul, Turkey*

** Corresponding author*
E-mail: semauzulmez@gmail.com; ORCID: 0000-0003-3181-3226
Zafer Aslan: zaferaslan@aydin.edu.tr; ORCID: 0000-0001-7707-7370

Received: 15 November 2021 Revised: 29 November 2021 Accepted: 3 December 2021
DOI: 10.17932/EJOH.2020.022/ejoh_v02i2005

Conclusion: The results of the present study highlighted the use of a variety of non-pharmacological behavior management techniques among dental specialists, although few acknowledged having adequate skills to apply the techniques. The choice of the technique was mainly influenced by the children`s factors.

Keywords: *COVID-19, Intensive Care, Regression Analysis, Holt Time Series, Pandemic*

Introduction

Pandemics caused by viruses have appeared with different names many times in human history and have caused high mortality rates. After being reported to the World Health Organization (WHO) on December 31, 2019, the coronavirus disease, which caused a worldwide crisis, spread rapidly all over the world under the name of the COVID-19 epidemic. The World Health Organization (WHO) declared COVID-19 a pandemic in March 2020, as the epidemic spread to different continents (1). With the pandemic, data analysis methods gained importance. The number of cases and death rates in the countries were reported due to these analyses. All health centers in Turkey started analyzing the daily number of cases, the number of intensive care patients, and mortality rates and sent the data to the Turkish Ministry of Health (2). The needs were determined with the data analysis, and quick action was taken. As in other countries, many scientific studies arose to cope with the pandemic in Turkey. The data used in the study are real data published by the Turkish Ministry of Health due to the COVID-19 pandemic in Turkey. In the study, 71-day intensive care inpatient data were used to make predictions, and predictions were produced for 7 days and analyzed. The analysis was made using data mining methods, and the best forecast method was chosen according to the analysis results. Python was preferred as the software language and analyzes were made with two data mining methods; libraries used in Python: Pandas, Matplotlib, Numpy, and Sklearn.

This study will help determine how present healthcare services should increase their capacity in the coming days to accommodate the expected volume of cases. Given the expected numbers, this study aims to provide communities and the government a sense of how swiftly this epidemic is growing and alert them to the initiatives that need to be taken. The number of COVID-19 epidemic cases in the G8 countries, Germany, the United Kingdom, France, specific countries, and Turkey was computed and forecasted in this study utilizing multiple curve Forecasting models, Box-Jenkins and Brown Holt linear, exponential smoothing techniques (3). Because the epidemic's beginning date differs by nation, the models are studied and applied separately for each country. The forecasts depict how the epidemic will proceed in the next days, based on the present high rate of cases.Time series forecasting involves examining earlier observations of a random variable to construct a model that best reflects the underlying relationship and its patterns.

The model then forecasts the future values of this random variable. This method is especially beneficial when there is little or no information about the underlying data-generating distribution. There is no explanatory model capable of correctly linking the prediction variable to other explanatory variables (4).

Methods

Data Review

The data used in the study are real data obtained from the web page of the Turkish Ministry of Health. The data are 71-day intensive care real inpatient data as of March 27, 2020, and analyses were carried out by producing forecasts with data mining methods. Multiple Linear Regression and Holt Time Series methods were used in the analysis, and the analysis results were compared. Python was chosen as the programming language. The intensive care patient data set used is as in Table 1.

Table 1. Sample dataset used for analysis (Turkish Ministry of Health) *

Date	Intensive care Number of patients	Intubated Number of patients	Recovering Number of patients
27.03.2020	344	241	42
28.03.2020	445	309	70
29.03.2020	568	394	105
30.03.2020	725	523	162
31.03.2020	847	622	243
01.04.2020	979	692	692
02.04.2020	1.101	783	415
03.04.2020	1.251	867	484
04.04.2020	1.311	909	786
…	…	…	…
30.05.2020	649	308	126984
31.05.2020	648	287	127973
01.06.2020	651	283	128947
02.06.2020	633	271	129921
03.06.2020	612	261	130852
04.06.2020	602	265	131778
05.06.2020	592	269	133400

** Table 1 includes the official intensive care patient data, sampled for analysis. On January 10, 2020, the Coronavirus Scientific Advisory Board was established under the Ministry of Health to combat the COVID-19 disease in Turkey (2). The Board took the necessary precautions and decisions based on these data.*

Multiple Linear Regression

Multiple Linear Regression is a sub-extension of Simple Linear Regression. It is used to predict the value of a variable based on the value of two or more variables. The variable we want to predict is called the Dependent Variable. The variables we use to forecast the value of the dependent variable are called "Independent Variables." Multiple Linear Regression also allows us to determine the model's overall fit and the relative contribution of each predictor to the total variance explained, as seen in Eq.

Multiple Linear Regression Formula (5):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Y: The dependent Variable,

X: The independent Variable,

$\boldsymbol{\beta_0, \beta_1}$ **:** B represents the coefficients.

RMSE (Root Mean Square Error)

To measure the performance of the models used in the predictions made by regression, the root of the sum of squares of mean error is calculated as seen in Eq.

RMSE (Root Mean Square Error) Formula (5):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

R SQUARE (R squared regression)

In regression analysis, after the regression Forecasting model is established, the regression coefficients, which are the coefficient of determination and called R2, are calculated. In this way, according to the result of the coefficients, the suitability of the established model is observed in Eq. (6)

R2 Formula: $$\mathbf{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}}$$

$0 \leq R2 \leq 1$ takes a value between R2, when it takes a value close to 1, it can be assumed that the established model is compatible (7).

Holt time-series models

The theoretical foundations of the exponential smoothing method were first introduced in 1958 by C.C. Dropped by Holt. Holt's simple form of exponential smoothing method is applied for time series that do not contain seasonal and trend elements. The exponential smoothing method is a method that can be used in all series with a variable trend and is constantly updated by considering the latest changes and increases in the data. It is double exponential smoothing.

Holt for this Forecasting method produced one prediction and two-level equations (6).

Forecasting Equation: $\hat{y}t+h|t = lt + hbt$ (8)

Level Equation: $lt = \alpha yt + (1 - \alpha)\ (lt-1 + bt-1)$ (9)

Trend Equation: $bt = \beta\ (lt + lt-1) + (1 - \beta)\ bt-1$ (10)

Holt Model Selection Criteria:

There are criteria calculated within the model to make the best choice among Holt models. AIC and BIC, which constitute these criteria, are explained as follows.

AIC: AIC, known as the Akaike information criterion, is a technique for forecasting the probability of a model predicting future values. A good model is the one with the lowest AIC among other benchmarks. AIC aims to make the best choice among Holt-Winters models. Accordingly, a lower AIC value indicates a better fit with the model (11).

BIC: BIC, known as Bayesian information criterion, is a criterion for measuring the balance between model selection and model fit. A lower BIC value indicates a better fit see Eq. (11). The following equations are used when forecasting the AIC and BIC of a model (11, 12):

AIC = -2*In(L)+2*k

BIC = -2*In(L)+2*In(N)*k

Results

The future 7-day forecast results of the Multiple Linear Regression model created in Python and the actual number of intensive care patients received from the Turkish Ministry of Health are given in Table 3. In Python, which is used as a programming language, Pandas, Numpy, matplotlib. pyplot libraries and Sklearn were used. To make a Multiple Linear Regression forecast within the program, R2 value and β coefficient values were calculated. In addition, the RMSE value was calculated for the Multiple Linear Regression model.

Since the R2 value gives a value which is very close to 1, the model fit is significant. As seen from table 2, Multiple linear regression formula calculation and dependent and independent variable information, model outputs are shown in Table 2. Multiple Linear Regression gave the results RMSE: 22.221, R2: 0.937.

Y: The dependent Variable, **X:** The independent Variable and $\boldsymbol{\beta_0, \beta_1}$: B represents the coefficients (13).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$Y =$ -300.270+1.985* +0.002* ...

Table 2. Calculated Multiple Regression model

Index ['IntensiveCare', 'Intubated', 'Intubated', 'Recovering']	
Forecast	R2 = 0.937
[[667.472]	
[[667.472]	B coefficients = [[1.985 0.0028]]
[623.346]	
[602.254]	RMSE = 22.221
[585.012]	
[595.551]	B0 coefficient = [[-300.270]]
[608.043]]	

Table 3. Actual and forecasted number of intensive care patients, using Multiple Linear Regression method.

Date	Intensive Care Actual	Intensive Care Forecast with Multiple Linear Regression
30.05.2020	649	667,47
31.05.2020	648	628,55
01.06.2020	651	623,35
02.06.2020	633	602,25
03.06.2020	612	585,01
04.06.2020	602	595,55
05.06.2020	592	608,04

The 7-day forecast results of the Holt Time Series model made in Python and the actual number of intensive care patients received from the Turkish Ministry of Health are given in Table 6. In addition, the Holt Time Series RMSE value was calculated. Holt time series AIC and BIC values are shown in Table 4. The RMSE for the Holt model was calculated as 39.815. The details are given in Table 5.

Table 4. Holt Model results

Dep. variable	endog	**No. observations**	63
Model	Holt	**SSE**	308207.685
Optimized	False	**AIC**	543.210
Trend	Multiplicative	**BIC**	551.782
Seasonal	None	AICC	544.710
Seasonal Periods	None	Box-Cox	False
Box-Cox-Coeff	None		

As it is seen from equation (11, 12)

AIC = -2*In(L)+2*k

BIC = -*In(L)+2*In(N)*k

where AIC is the Akaike information criterion. 2K – 2 is the AIC function (loglikelihood). Lower AIC values show a better-fit model, and a model with a delta-AIC (the difference of the two AIC values being compared) greater than -2 is deemed considerably better than the model to which it is compared (11). Furthermore, BIC stands for Bayesian information criterion where BIC is calculated as BIC = -2 * loglikelihood + d * log(N), where N is the training set random sample and d is the overall parameters. A lower BIC score indicates a better model (14).

Table 4 shows the optimized False AIC is 543.210 while Multiplicative BIC is 551.782 (11, 12).

Table 5. Coeff code optimized results

Coeff	Code	Optimized	
smoothing_level	0.8000000	Alpha	False
smoothing_slope	0.2000000	Beta	False
initial_leve	1445.0000	1.0	False
initial_slope	1.2764045	b.0	
RMSE = 39.815			

According to Table 6, it is seen that the actual number of intensive care patients with the forecasts made with the Holt time series is close to each other on some days.

Table 6. Actual and forecasted number of intensive care patients, using Holt time series method.

Date	Intensive Care Actual	Intensive Care Forecast with Holt time series
30.05.2020	649	643,91
31.05.2020	648	625,07
01.06.2020	651	606,77
02.06.2020	633	589,01
03.06.2020	612	571,78
04.06.2020	602	555,04
05.06.2020	592	538,80

According to the Holt Time Series analysis, the estimated number of intensive care patients in Turkey between the dates specified in the study and the next 7 days is given in Figure 1.

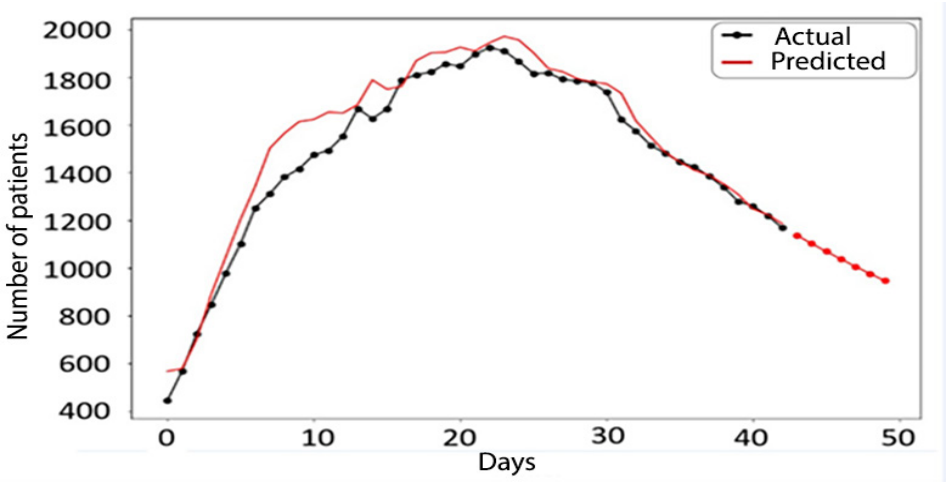

Figure 1. *Intensive care patient forecasting by Holt Time Series Model*

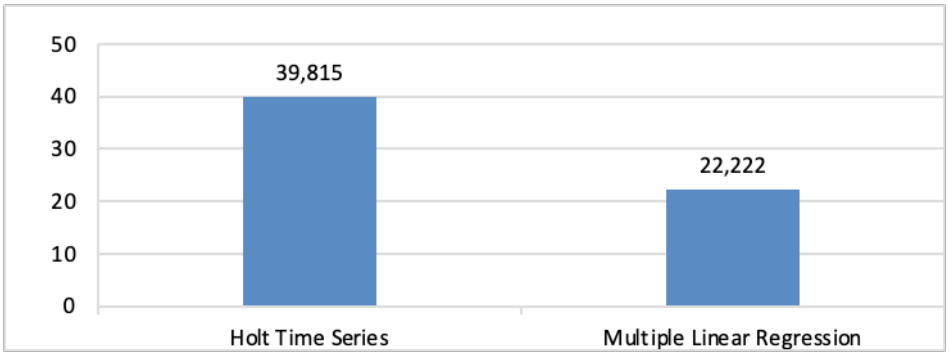

Figure 2. *Root Mean Square Error (RMSE) comparison of Holt Time Series and Multiple Linear Regression Methods*

Table 7. The actual number of intensive care patients and the forecasted results with regression analysis and Holt time series methods.

Date	Actual	Regression Forecast	Holt Forecast
30.05.2020	649	667,47	643,91
31.05.2020	648	628,55	625,07
01.06.2020	651	623,35	606,77
02.06.2020	633	602,25	589,01
03.06.2020	612	585,01	571,78
04.06.2020	602	595,55	555,04
05.06.2020	592	608,04	538,80

Table 8. COVID-19 patient outcome data*

	Number	Mean	Std	Min	25%	50%	75%	Max
Total Cases	36	9.62	1.44	768	8.47	4.76	1.01	6.63
Active	36	2.75	7.63	0.00	1.23	3.10	2.15	4.44
Discharged	36	9.46	1.41	755	8.36	4.65	1.00	6.48
Deaths	36	1.31	2.46	4.00	8.23	5.62	1.38	1.41
Active Ratio	36	3.04	4.87	0.00	3.75	1.35	3.60	2.61
Discharge Ratio	36	9.84	6.22		9.80	9.83	9.86	9.99
Death Ratio	36	1.29	5.63	0.04	9.62	1.31	1.66	2.75
Population	36	3.97	5.05	660	1.69	2.41	6.97	2.31

**The mean value and standard deviations (std) and the lowest (min) and maximum (max) quantiles are all in close correlation with one another. Mean, and standard deviation are often used metrics of central tendency and variability in data from scale variables. If data are not normally distributed, some researchers prefer reporting median and quartiles instead.*

Discussion

As a result of the analyses examined in the literature review studies, important predictions were made about the intensive care units. It was ensured that the intensive care unit needs that may arise in the future were determined in advance. Analysis by Meares and Jones (15) calculated the R2 value and a summary of results of 0.96. The R2 value of the multiple regression method in our study is 0.93. Our model fit and the success rate in the estimations are very close to each other and have values close to the predictable truth.

Ankaralı (7) stated that the time series model was less successful than the second method, as in our study. Although the prediction sizes of the two models are close, the time series analysis by Du et al. (16) was made in monthly periods due to the incubation period, to obtain 0.95 impressions in the forecasting. The authors aimed also to model the diurnal variation of infection criteria. The results showed that Germany and South Korea were at the top of the countries that control the process best and that the process in our country was similar to the countries that spread rapidly in the first 10 days. Pandey et al. (17) applied the SEIR (Susceptible-Exposed-Infected-Recovered) model and regression model in India for forecasts based on data collected from the John Hopkins University repository in January and March. The performance of the models was evaluated using the Root Mean Squared Log Error (RMSLE), obtaining 1.52 and 1.75 for the SEIR model. Weekly estimates were used in our study.

Meares and Jones (15) made forecasts for the number of intensive care beds and ventilators using different statistical models using daily data announced by official sources. In case of possible increases in the number of intensive care patients, future forecasts were made with the Gompertz and Time Series model for the number of beds and ventilators needed. SPSS and MINITAB programs were used in the calculations.

While the pandemic has been rising since February 2020 in Italy, useful estimations and forecasts have been made in the resource facility for the intensive care staff, intensive care unit, and intensive care beds needed (7).

Use of thermal camera for controlling the epidemic at airport entrances and exits, was evaluated by data mining methods, and 46% (95% confidence interval) of infected passengers could not be detected (13). Another study analyzed the probability of the virus spread to 369 cities in China from Wuhan, where the coronavirus emerged, using time series methods (14). The data and instructions that are constantly updated and instantly shared on the official website of the World Health Organization (WHO) were examined by the document analysis method. Therefore, between December 31, 2019, and March 10, 2020, the data related to the travel restrictions applied by the governments as a precautionary measure, the curfews applied in the regions where the epidemic was detected, the special measures that directly affect the tourism sector, such as the canceled international sports and arts events were evaluated within the scope of the study (18).

A study aimed to prepare the Asian intensive care unit community for the negativities in the epidemic's future. A joint study was conducted for the intensive care needs of patients with COVID-19 regarding critical patient management, and they analyzed it with the Post-hoc Analysis method (19). Despite claims by Özşahin et al. (1) that the virus is unstoppable, efforts to overcome are carried out mainly by artificial intelligence and medical advancements.

Study by Kayış (20) evaluated the 2019-nCoV process in our country, and the applicability of the growth model was examined between March 11 and April 27 in Turkey, using the 2019-nCoV data officially published by the Ministry of Health. The author proposed a Gompertz growth model to forecast the need for the number of intensive care patient beds, intensive care patients and doctors in our country (5). Another study aimed to analyze the clinical determinants of COVID-19 in patients living in Wuhan, China. Using retrospective data concerning patients with laboratory-confirmed SARS infection from the Jin Yin-tan Hospital and Tongji Hospital database, a multicenter study was conducted to forecast COVID-19 cases of death and discharge. For statistical analysis, continuous measurements were shown as mean (SD) or median (IQR) compared to Student's t-test or Mann-Whitney-Wilcoxon test (18). Using Natural Language Processing, Text Mining, and Network Analysis to analyze the community of tweets about the COVID-19 outbreak, they identified overall responses to the pandemic and how those responses differed over time worldwide until January 22, 2020.

Accurate and rapid diagnosis of suspected cases of COVID-19 plays a crucial role in quarantine measures and medical treatments. Zheng et al. (10) developed a weakly supervised deep learning-based software system for automatic COVID-19 detection in chest CTs.
Ahamed and Samad (21) explain that a graph-based model was developed using summaries of 10,683 scientific papers to find key information on three topics: genome research on transmission, drug types, and coronavirus. To obtain more topic-oriented information, they created a footer for each of the three topics.

Conclusion

In our study, it was seen that the prediction success was good in the two models used to forecast the number of intensive care patients but the forecasted numbers with the Multiple Linear Regression method were closer to the daily actual values. Therefore, the Multiple Linear Regression method should be preferred, with a lower error in line with the specified values, to help determine the deficiencies in terms of occupancy rates in intensive care units, review the intensive care bed capacities and the number of ventilators, as well as to meet the needs without delay. Based on the prediction of conditions in a probable severe pandemic such as COVID-19 in future, new data input to the Long Short-Term Memory (LSTM) techniques can also be studied, resulting in a smaller forecasting error than that in Holt Time Series.

References

1. Özşahin O, Akgül Ö, Çalışkan R, Sapmaz B, Öner YA. Investigation of the Awareness Levels of COVID-19 in University Students. EURAS Journal of Health. 2020;1(1):73-81.
2. Türkiye Cumhuriyeti Sağlık Bakanlığı. Available at: https://www.saglik.gov.tr. Accessed 30.05.2021.
3. Yonar H, Yonar A, Tekindal MA, Tekindal M. Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Forecasting Models, the Box-Jenkins and Exponential Smoothing Methods. EJMO. 2020;4(2):160-165.
4. Hair Jr JF, Sarstedt M. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice. 2021;29(1):65-77.
5. Ankarali H, Ankarali S, Erarslan N. COVID-19, SARS-CoV2, Infection: Current Epidemiological Analysis and Modeling of Disease Course. Anatolian Clinic the Journal of Medical Sciences. 2020;25:1-22.
6.Phyton Methodology. Available at: https://towardsdatascience.com/simple-and-multiple-linear-regression-with-python-c9ab422ec29c.Accessed 15.05.2020.
7. Ankarali H. Direct Forecasting of the Number of Intensive Care Beds and Respirators to be Needed in the COVID-19 Epidemic Process in Turkey. Anatolian Clinic the Journal of Medical Sciences. 2020;25:59-62.
8. Chatfield C. The Analysis of the Time Series an Introduction. Chapman Hall/CRC Texts in Statistical Science. 2003.

9. Lopez CE, Vasu M, Gallemore C. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. arXiv.org. arXiv:2003.10359v1
10. Zheng C, Deng X, Fu Q, Zhou Q. Deep Learning-based Detection for COVID-19 from Chest CT using Weak Label. Available at: https://www.medrxiv.org/content/medrxiv/early/2020/03/26/2020.03.12.20027185.full.pdf.Accessed 25.05.2020.
11. Cifci MA. Deep learning model for diagnosis of corona virus disease from CT images. Int. J. Sci. Eng. Res. 2020;11(4):273-278.
12. Cifci MA, Aslan Z. Deep Learning Algorithms for Diagnosis of Breast Cancer with Maximum Likelihood Estimation. In International Conference on Computational Science and Its Applications. Springer, Cham. 2020: 486-520.
13. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next?. The Lancet Journal. 2020;395:1225–1228.
14. Phua J, Weng L, Ling L, Egi M, Lim CM, Divatia JV, Shrestha BR, Arabi YM, Ng J, Gomersall CD, Nishimura M, Koh Y, Du B, for the Asian Critical Care Clinical Trials Group. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. The Lancet Respiratory Medicine. 2020;8(5):506-517.
15. Meares HD, Jones MP. When a system breaks: queueing theory model of intensive care bed needs during the COVID-19 pandemic. Med J Aust. 2020;212(10):470-471.
16. Du Z, Wang L, Cauchemez S, et al. Risk for Transportation of Coronavirus Disease from Wuhan to Other Cities in China. Emerg Infect Dis. 2020;26(5):1049-1052.
17. Pandey G, Chaudhary P, Gupta R, Pal S. SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv:2004.00958v1
18. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med. 2020;46(5):846-848.
19. Quilty Billy J, Clifford Sam, CMMID nCoV working group2, Flasche Stefan, Eggo Rosalind M. Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV). Euro Surveill. 2020;25(5):pii=2000080.
20. Kayıs S. A. Türkiye 2019-nCoV Salgın Süreci ve Büyüme Modeli Destekli Salgın Süreci Yönetimi. J Biotechinol & Strategic Health Res. 2020;4:152-157.
21. Ahamed S, Samad M. Information Mining for COVID-19 Research from a Large Volume of Scientific Literature. Cornell University. arxiv.org/abs/2004.02085.