VIA MEDICA

ORIGINAL PAPER

Shahir Asfahan, Maya Gopalakrishnan, Naveen Dutt, Ram Niwas, Gopal Chawla, Mehul Agarwal, Mahendera Kumar Garg

All India Institute of Medical Sciences, Rajasthan, Jodhpur, India

Using a simple open-source automated machine learning algorithm to forecast COVID-19 spread: A modelling study

Abstract

Introduction: Machine learning algorithms have been used to develop prediction models in various infectious and non-infectious settings including interpretation of images in predicting the outcome of diseases. We demonstrate the application of one such simple automated machine learning algorithm to a dataset obtained about COVID-19 spread in South Korea to better understand the disease dynamics.

Material and methods: Data from 20th January 2020 (when the first case of COVID-19 was detected in South Korea) to 4th March 2020 was accessed from Korea's centre for disease control (KCDC). A future time-series of specified length (taken as 7 days in our study) starting from 5th March 2020 to 11th March 2020 was generated and fed to the model to generate predictions with upper and lower trend bounds of 95% confidence intervals. The model was assessed for its ability to reliably forecast using mean absolute percentage error (MAPE) as the metric.

Results: As on 4th March 2020, 145,541 patients were tested for COVID-19 (in 45 days) in South Korea of which 5166 patients tested positive. The predicted values approximated well with the actual numbers. The difference between predicted and observed values ranged from 4.08% to 12.77% . On average, our predictions differed from actual values by 7.42% (MAPE) over the same period.

Conclusion: Open source and automated machine learning tools like Prophet can be applied and are effective in the context of COVID-19 for forecasting spread in naïve communities. It may help countries to efficiently allocate healthcare resources to contain this pandemic.

Key words: machine learning, COVID-19, coronavirus, pandemic, South Korea

Adv Respir Med. 2020; 88: 400–405

Introduction

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) originated in the Wuhan province of China in December 2019, has since spread to many countries, prompting WHO to declare it as a pandemic [1]. As of 11th March 2020, 118 326 cases have been confirmed and 4 292 deaths reported worldwide with the pathogen rapidly spreading to newer areas [2]. Having never encountered this virus before, health care systems are grappling with multiple unknowns about this pandemic and navigating uncharted territories [3]. One of the important parameters to be addressed is the rate and scale of spread in disease naïve communities. Multiple epidemiologic studies are attempting to determine the epidemiology and estimated basic reproduction number (R_0) of the disease to help assess the spread pattern [4]. Predicting the scale of spread will help prepare fragile health care systems, especially in the developing world to get ready and allocate resources accordingly.

Machine learning algorithms have been used to develop prediction models in various infectious and non-infectious settings, including interpretation of images in predicting the outcome of diseases [5]. The ease with which these can be used at a low cost and minimum learning curve is an advantage, especially in epidemic situations. We

Address for correspondence: Gopal Chawla, All India Institute of Medical Sciences, Rajasthan, Jodhpur, India; e-mail: dr.gopalchawla@gmail.com
DOI: 10.5603/ARM.a2020.0156
Received: 24.06.2020

ISSN 2451–4934

demonstrate the application of one such simple automated machine learning algorithm to a dataset obtained about COVID-19 spread in South Korea to better understand the disease dynamics. The aim of this study is to demonstrate the ease with which simple algorithms can be used effectively as disease prediction models to help health care systems anticipate new cases and prepare accordingly, especially in resource-limited settings.

Material and methods

Korea's centre for disease control (KCDC) has released in the public domain the database of patients who are being tested for COVID-19 and those who are diagnosed with the same. This dataset is being updated daily and is licensed for research purposes by Creative Commons licence (CC BY-NC-SA 4.0) [6].

"Prophet" is an open-source automated machine learning actuarial modelling system that has been used in insurance and financial services for improving risk management available in the public domain since 2017 [7, 8]. It uses linear and non-linear regression techniques and takes into account seasonality in the final analysis. This approach is a more accurate reflection of patterns of human activities as well as biological variables. Python 3.6 was used as the programming language given its many supporting libraries and ease of use.

Data from 20th January 2020 (when the first case of COVID-19 was detected in South Korea) to 4th March 2020 was accessed from the above-mentioned source. The columns of date and cumulative sum of COVID-19 cases of the corresponding dates were selected and included in the analysis.

A future time series of specified length (taken as 7 days in our study) starting from 5th March 2020 to 11th March 2020 was generated and fed to the model to generate predictions with upper and lower trend bounds of 95% confidence intervals. We used the future time series of 7 days as the disease is rapidly evolving and trends are changing in real time. The model was assessed for its ability to reliably forecast using mean absolute percentage error (MAPE) as the metric.

Results

As of 4th March 2020, 145 541 patients were tested for COVID-19 (in 45 days) in South Korea of which 5 166 persons tested positive (Figure 1).

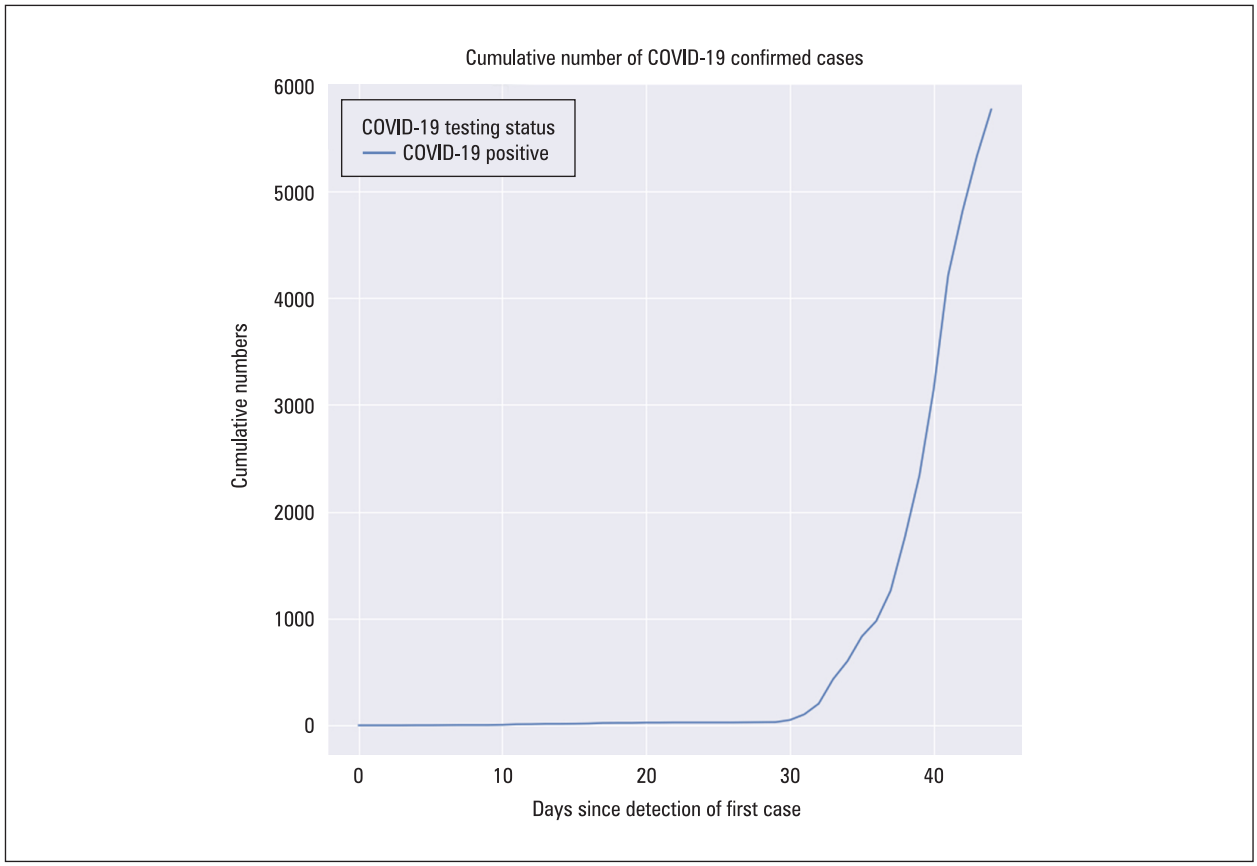

Figure 1. Total number of COVID-19 positive patients in South Korea as of 4th March 2020 plotted over 45 days since the detection of the first case

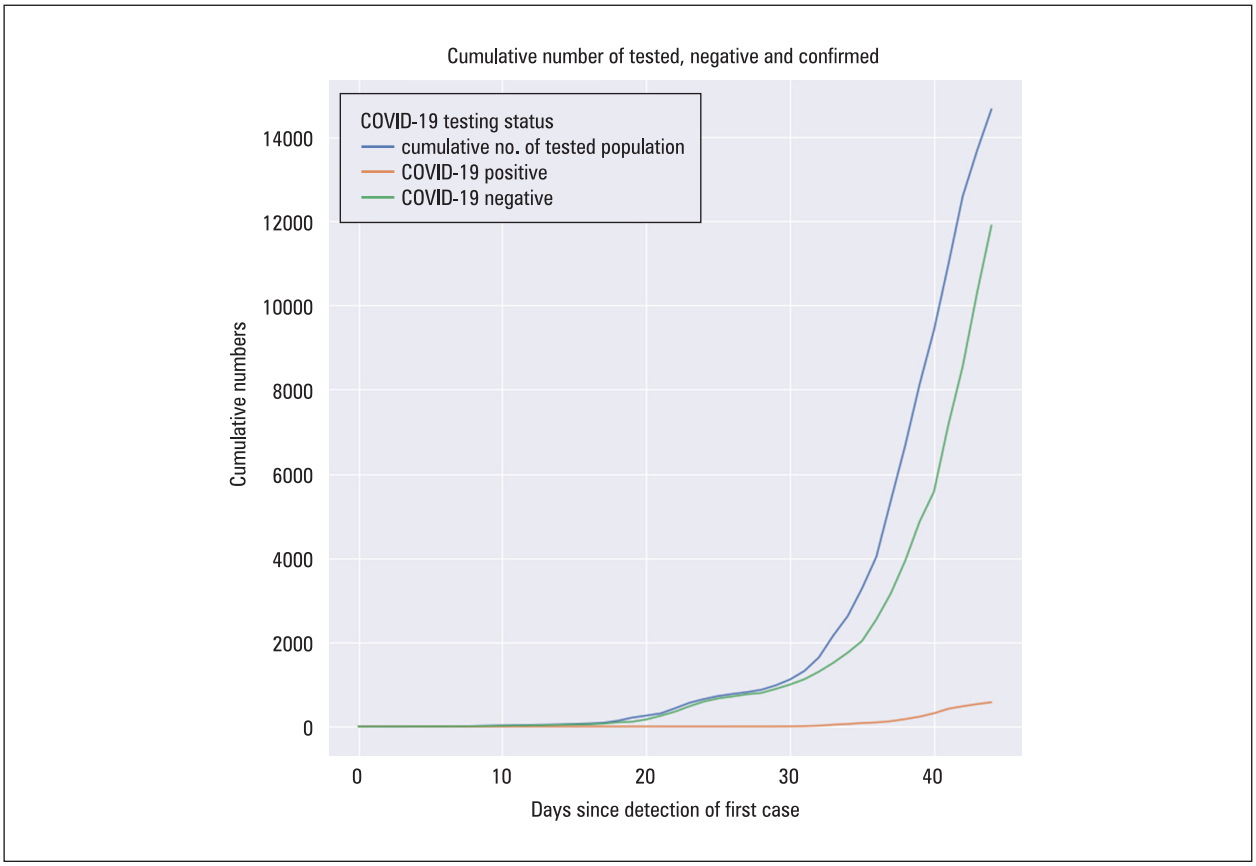

Figure 2. Total number tested as of 4th March 2020 in South Korea were 146 541, of which 5 766 tested positive and the rest were negative

The trend line of the positive and negative cases plotted since the day of detection of the first case are shown in Figure 2. Forecasts estimates were drawn for the next 7 days as shown in Figure 3. The forecasts were made with a confidence interval of 95% represented by the upper and lower range bounds for each prediction as shown in Table 1.

We followed up the actual numbers till 11th March 2020. The predicted values approximated well with the actual numbers. The difference between predicted and observed values ranged from 4.08% to 12.77% (Table 1).

The total increase in the number of actual confirmed cases over the predicted period of 7 days was 1 989, which is an increase of 35.5% over the baseline value on 4th March 2020. On average, our predictions differed from actual values by 7.42% (MAPE) over the same period (Figure 4).

Discussion

COVID-19 is a rapidly evolving pandemic with limited data regarding its spreading potential. Many countries are reporting new cases daily. Several large modelling studies have focussed on nowcasting and forecasting potential domestic and international spread of COVID-19 as well as health care system preparedness for handling disease spread in Africa [9, 10]. A modelling study into COVID-19 evaluating the usefulness and feasibility of isolation has also been recently published [11].

We employed public domain data from South Korea for demonstrating the use of simple open-source automated machine learning algorithm which has been hitherto used for non-medical purposes to help model the current spread of disease in South Korea. Although we are analysing the data of another country, in this digital era where the world has been reduced to a global village, we expect to replicate the performance of a neutral machine learning tool if applied in any other context.

MAPE index of our model for one week was 7.42%, which is indicative of a highly accurate forecasting model [12]. Part of this success was how machine learning model learnt from the training data provided to it and automatically detecting the trend changes and estimating the predictions (Figure 4). The inaccuracy decreased to a minimum over the middle of predictions and

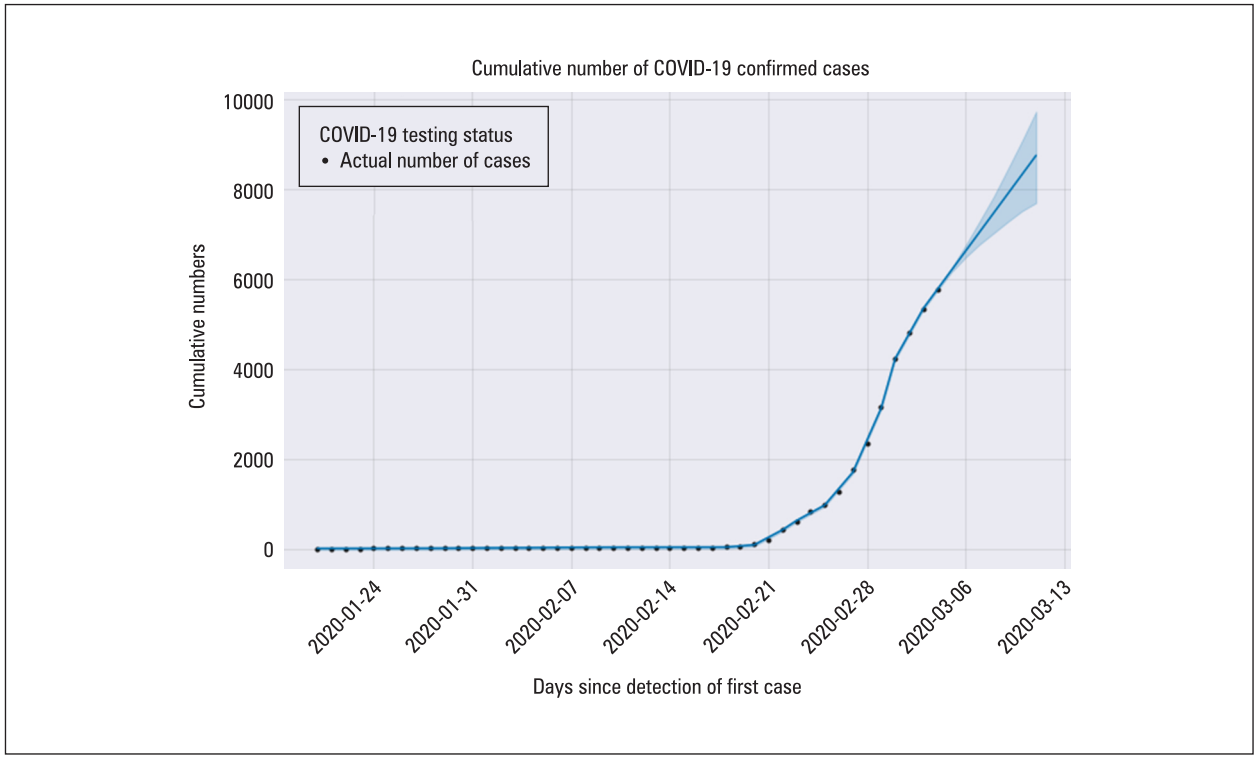

Figure 3. Black dots represent actual values till 4th March 2020 and the blue line indicates the predictions for the next 7 days with the shaded blue designating 95% confidence intervals

Table 1. Actual numbers versus the predicted numbers till 11th March 2020 (as of the day of writing this manuscript, with upper and lower 95% confidence intervals) with the difference and percentage change of predicted values

Date	Actual numbers	Predicted numbers	Predicted lower limit	Predicted upper limit	Difference between actual and predicted numbers	Percent difference between actual and predicted numbers
05.03.2020	5766	6192.34	6144.153	6236.549	-426.3397742	-7.39
06.03.2020	6284	6617.772	6486.003	6744.628	-333.7715647	-5.31
07.03.2020	6767	7043.203	6788.055	7273.402	-276.2033551	-4.08
08.03.2020	7134	7467.634	7064.978	7839.158	-334.6351456	-4.69
09.03.2020	7382	7894.067	7333.481	8455.096	-512.066936	-6.94
10.03.2020	7513	8319.499	7555.599	9061.543	-806.4987264	-10.73
11.03.2020	7755	8744.931	7749.495	9722.656	-989.9305169	-12.77

then continued to increase towards the end of the predictive period. Machine learning algorithms are data hungry and need larger data for more accurate predictions over a longer period. In a fast moving and novel pandemic we do not have this luxury, hence restricting ourselves to a limited period as of the date of this analysis, with larger data in the future, longer range predictions would be possible [13].

We have learnt from China and Italy on how intensive care units can get easily overwhelmed with patients suffering from COVID-19 [14, 15]. This outbreak seems to test the capacity of health care systems like never before, even in developed countries. Hence the application of machine learning to anticipate and arrange health care resources a week in advance might make a big difference in managing the pandemic.

The advantage of this approach of forecasting is that it takes local factors into account rather than global aspects. This machine learning tool is able to factor in changes and figure out any pattern in the rise or fall of cases, including seasonality and can be used during the initial slow

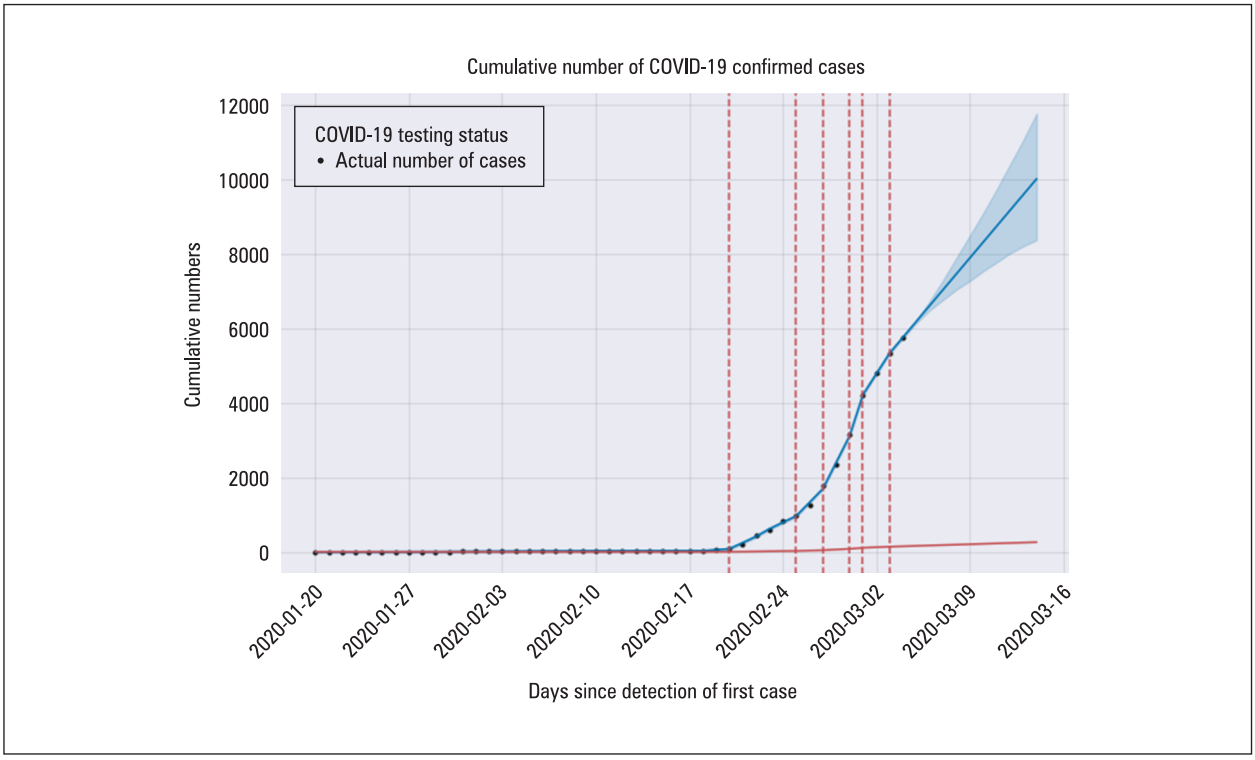

Figure 4. Dotted red lines indicate the automatic trend changes detected by our machine learning model

rise, followed by steep rise, eventual fall in cases or even any seasonal recurring trends [7]. The approach and threshold of one country's testing strategy might differ from another, so one-size-fits-all attitude may not work best. Using the data from a particular region and forecasting based on that may accurately reflect the pattern that is likely to arise in that particular region. Although we agree that in an ideal world, epidemiology should be consistent between countries but on the practical plane, the difference between health care systems between countries might be too stark to make reasonable parallels. As the case fatality ratios become clearer, models like these might be useful to calculate in advance the number of intensive care beds/ECMO units that might be required and also to assess whether public health strategies are effective.

Our study has several limitations involving the assumptions used in the preparation of this model. We employed a linear model, as in the short term, we don't anticipate saturation in the total number of the vulnerable population. This limits the applicability in the long term. Also, calculations of replication potential and spread potential were not performed as the aim of the study was to demonstrate the ease of use and applicability of the machine learning algorithm rather than a large-scale forecasting project. Also, since the intent of the study was to evaluate the performance of new technologies into disease modelling, we did not go into the aspect of the effect of national responses to the pandemic if any in the preceding period of forecast, although that is certainly a possibility and may need an appropriate selection of response measures and the predictive period where the difference between predictive and actual numbers may demonstrate the effect of measures like social distancing and 'lockdown' on the rate and extent of the spread of the pandemic.

Conclusions

Open source and automated machine learning tools like Prophet can be applied and are effective in the context of COVID-19 for forecasting spread in naïve communities. It may help countries to efficiently allocate health care resources to contain this pandemic.

Conflict of interest

None declared.

References:

1. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet. 2020; 395(10229): 1054–1062, doi: 10.1016/s0140-6736(20)30566-3.
2. World Healt Organization. Coronavirus disease (COVID-19) pandemic. Available online: www.who.int/emergencies/diseases/novel-coronavirus-2019. [Last accessed at: 05.10.2020].
3. Lipsitch M, Swerdlow DL, Finelli L. Defining the epidemiology of COVID-19 — studies needed. N Engl J Med. 2020; 382(13): 1194–1196, doi: 10.1056/NEJMp2002125, indexed in Pubmed: 32074416.
4. Fauci AS, Lane HC, Redfield RR. COVID-19 — navigating the uncharted. N Engl J Med. 2020; 382(13): 1268–1269, doi: 10.1056/NEJMe2002387, indexed in Pubmed: 32109011.
5. Skrede OJ, Raedt SDe, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet. 2020; 395(10221): 350–360, doi: 10.1016/s0140-6736(19)32998-8.
6. KCDC. Available online: http://www.cdc.go.kr. [Last accessed at: 09.05.2020].
7. Taylor SJ, Letham B. Prophet: forecasting at scale. Facebook Research. Available online: https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/. [Last accessed at: 05.10.2020].
8. Fang WX, Lan PC, Lin WR, et al. Combine Facebook prophet and LSTM with BPNN forecasting financial markets: the Morgan Taiwan Index. 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). 2019, doi: 10.1109/ispacs48206.2019.8986377.
9. Gilbert M, Pullano G, Pinotti F, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. The Lancet. 2020; 395(10227): 871–877, doi: 10.1016/s0140-6736(20)30411-6.
10. Wu J, Leung K, Leung G. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet. 2020; 395(10225): 689–697, doi: 10.1016/s0140-6736(20)30260-9.
11. Hellewell J, Abbott S, Gimma A, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020; 8(4): e488–e496, doi: 10.1016/s2214-109x(20)30074-7.
12. Meade N. Industrial and business forecasting methods. J Forecast. 1983; 2(2): 194–196, doi: 10.1002/for.3980020210.
13. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014; 14: 137, doi: 10.1186/1471-2288-14-137, indexed in Pubmed: 25532820.
14. Kuo L. Wuhan nurses' plea for international medics to help fight coronavirus. The Guardian. 26.02.2020. Available online: www.theguardian.com/world/2020/feb/26/wuhan-nurses-plea-international-medics-help-fight-coronavirus. [Last accessed: 05.10.2020].
15. Orecchio-Egresitz H. Faced with tough choices, Italy is prioritizing young COVID-19 patients over the elderly. That likely "won't fly" in the US. Buisness Insider. 11.03.2020. Available online: www.businessinsider.in/science/news/faced-with-tough-choices-italy-is-prioritizing-young-covid-19-patients-over-the-elderly-that-likely-wont-fly-in-the-us-/articleshow/74567872.cms. [Last accessed at: 05.10.2020].