

# Prognostic machine learning models for COVID-19 to facilitate decision making

## Abstract

An increasing number of COVID-19 cases worldwide has overwhelmed the healthcare system. Physicians are struggling to allocate resources and to focus their attention on high-risk patients, partly because early identification of high-risk individuals is difficult. This can be attributed to the fact that COVID-19 is a novel disease and its pathogenesis is still partially understood. However, machine learning algorithms have the capability to analyse a large number of parameters within a short period of time to identify the predictors of disease outcome. Implementing such an algorithm to predict high-risk individuals during the early stages of infection would be helpful in decision making for clinicians such that irreversible damage could be prevented. Here, we propose recommendations to develop prognostic machine learning models using electronic health records so that a real-time risk score can be developed for COVID-19.

The current surge in COVID-19 patients has created unprecedented stress on healthcare infrastructure. Early identification of high-risk patients can allow healthcare workers to allocate their efforts and resources early during the clinical course to maximise their impact on patient health. Early critical care management in certain clinical settings has demonstrated improvement in mortality.<sup>1</sup> However, the identification of patients at high risk of progressive and severe disease remains a challenge. Although patient characteristics, such as chest radiograph findings, lymphopenia, age, gender and viral load, are associated with severe COVID-19 disease, there are no currently available biomarkers that can reliably distinguish patients that need immediate medical attention.<sup>2</sup> Here we lay out recommendations to implement a machine learning algorithm that would facilitate early clinical decision making during outbreaks like COVID-19.

## 1 | RATIONALE FOR MACHINE LEARNING

In the case of the COVID-19 outbreak, there have been more than 2.3 million cases in the United States and more than 9 million cases

worldwide as of June 23, 2020.<sup>3</sup> Given the number of cases, an analogue approach to reviewing cases to identify patterns that indicate poor prognosis is not feasible. A large number of cases have particularly stressed intensive care unit (ICU) settings with an increase in the need for ICU beds. With this increase in ICU beds, there is an immense need for ventilators and continuous renal replacement machines given high rates of pulmonary and renal failure.<sup>2</sup> A prediction model, which can identify patients more likely to deteriorate and require ICU care at an early stage, will allow physicians to allocate manpower and resources in an expeditious and informed manner.

In certain cases, it has been demonstrated that machine learning models might outperform traditional clinical scoring systems or regression methods.<sup>4,5</sup> Machine learning models which utilise decision tree or neural networks can detect non-linear relationships and interactions between variables, which could explain their better performance.

Prediction models can also estimate risk for specific diseases, eg, distinguishing those who will develop respiratory failure and require ventilators from those who will develop renal failure and require renal replacement therapy, as well as identifying patients at risk of requiring both life-supporting treatments. The integration of the prediction model with the electronic health record (EHR) can give physicians immediate information about the expected patient course and predicted response to treatments.

## 2 | OUTCOME OF INTEREST AND APPLICABILITY

Machine learning models could be trained to learn and detect patterns in a large number of records in a fraction of time. Supervised machine learning is a type of machine learning where the model trains itself using patient traits as input and disease outcome as output. Early clinical, radiological and laboratory data could be considered as input, while disease severity by a variety of metrics could be the output to train a predictive model for COVID-19. By providing input data from the electronic health records, certain characteristics or lab values that have yet to be associated with disease severity could be found to be strong predictors in specific situations giving clinicians information they had otherwise not had time to investigate in a novel disease such as COVID-19.

### 3 | LESSONS FROM THE PAST

Multiple examples of machine learning in predicting clinical outcomes currently exist. Using a longitudinal dataset of electronic health records (EHR) from more than 700 000 patients, a machine learning model was able to predict future acute kidney injury.<sup>6</sup> Another similar machine learning model based on hospital data from a Portuguese and American hospital was able to predict the risk of ICU admission.<sup>7</sup> A study from Denmark using a machine learning model was able to predict 90-day mortality for intensive care unit patients using time series data.<sup>8</sup> The key findings of this model were that the predictive performance significantly improved over the timecourse of ICU stay and input features of patients interacted and compensated for one another to pull the patient towards survival at one timepoint and towards mortality at another. Static prognostic scoring systems usually fail to adapt to such patient dynamics. In the field of cardiology, machine learning techniques have been employed for in-hospital monitoring, precision medicine, imaging and electrocardiography and have shown to improve patient outcomes.<sup>9</sup> These examples underscore the capabilities of machine learning.

### 4 | MODEL DEVELOPMENT AND CLINICAL IMPLEMENTATION

Building a machine learning model for COVID-19 would require early stage clinical, radiological and laboratory data from a large cohort (Figure 1). The training dataset must also include information about the patient outcomes one is looking to predict, which forms the primary basis of machine learning training. To improve the robustness

of the model, we suggest removing patients with unknown outcomes during model development. The training dataset should include patient data, such as demographic data, co-morbidities, outpatient medications, vital signs and laboratory values, which could be obtained from the EHR database. In this situation, supervised machine learning algorithms for classification, such as random forest, gradient boost, support vector machines, naïve bayes, k-nearest neighbour and logistical regression, could be tested. While developing the model, a fraction of patients should be kept out during the training process (if separate dataset is available that should be used for testing and would result in a more robust model), to serve as a testing cohort and help validate the accuracy of the model, which can be performed using confusion matrix analysis. Once the accuracy of the model significantly improves as compared to no-information-rate, the model could be deployed to new patients. Before deployment, sensitivity analysis would ensure that the variables of importance are relatable with the actual clinical scenario. For the implementation of this model, the risk score could be provided to clinicians at the earlier stages of disease for new patients, which would then help them guide treatment or allocate resources based on the risk of particular outcomes and plan for future treatment needs. Another area where such an algorithm can be used would be in clinical trials, where it could help in forecasting the probability of success of the trial.

Machine learning approaches were implemented on COVID-19 patient data in China.<sup>10,11</sup> The aim was to predict the severity of disease based on initial presentation data. One of the models was accurately able to predict disease outcome in 90% of the cases.<sup>11</sup> In this model, the most important features used for prediction were lactate dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP). Similar implementations of the machine learning

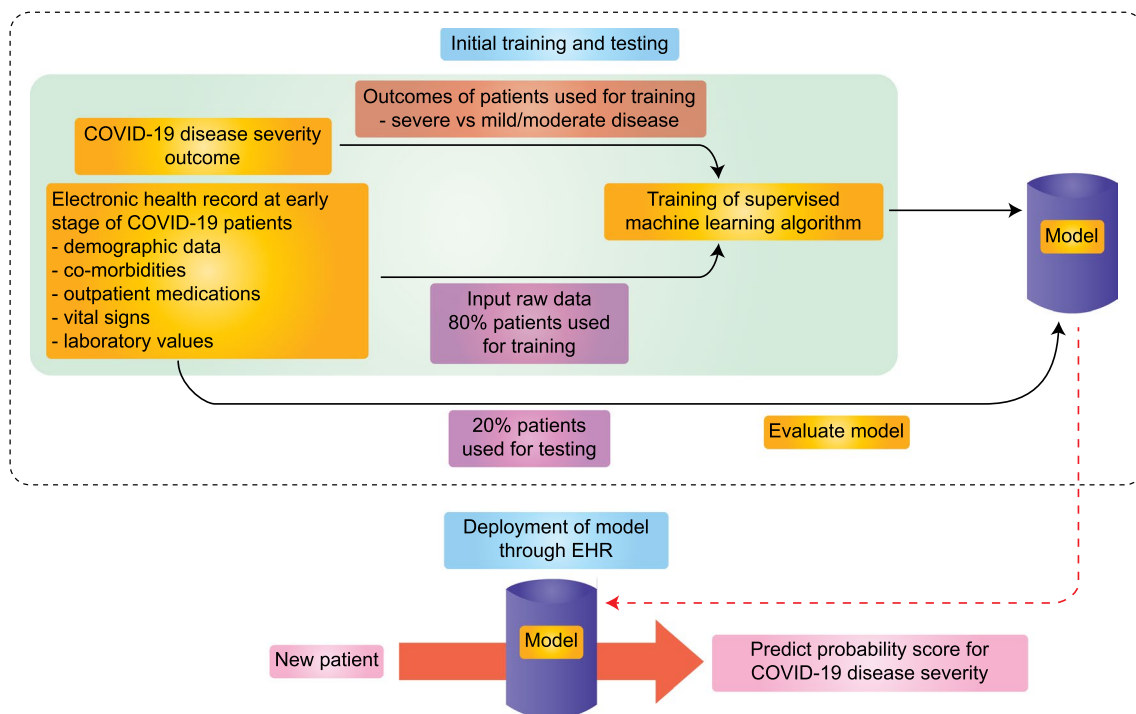


FIGURE 1 Implementing machine learning algorithm to predict COVID-19 disease outcome

approach in larger cohorts from other countries can provide more specific models to understand local factors as predictors of disease.

## 5 | ADVANTAGES AND CHALLENGES

Machine learning models benefit from larger sample sizes, which in most cases, improve the accuracy of models.<sup>12</sup> Also, machine learning algorithms can be used to detect novel biomarkers which have non-linear interactions with each other. An added advantage of this model is that the risk score can be generated in real-time by integrating the model with electronic health records. As vitals or laboratory values change, the model can predict a new score based on the most recent data, allowing real-time monitoring.

Another key challenge of clinical data is the missing data in variables. Although there are different imputation methods available for substituting missing data, we recommend not using it for variable where high levels of missingness exist. But for a small percentage of missingness, imputation using the k-nearest neighbour algorithm could be used, which are more accurate than using mean/median values.<sup>13</sup> With the introduction of newer medications, the model performance might be affected. This limitation needs to be assessed and necessary changes in covariates should be updated to ensure a good performance of the model.

The existing machine learning models are usually single-centred, retrospective studies.<sup>10,11,14</sup> Although they identified LDH, hs-CRP and lymphocytes as key factors, this data needs to be evaluated at a multicentre level with a larger dataset, which would also remove any bias that might have been introduced in a single-centre study.<sup>11</sup> These models have also not discussed how different co-morbidities alter COVID-19 severity. For example, patients with diabetes and cancer might not have the same risk factors. Splitting patient data into subgroups and generating models would allow us to identify important covariates for specific co-morbidities.

Some complex machine learning models such as gradient boost and neural network models are black box models that are difficult to interpret but generally perform better than white box models. While selecting the best model, this trade-off between accuracy and interpretability should be considered and based on risk-benefit assessment, appropriate decisions should be taken. It is important to be cautious of the model overfitting the data which can be compensated by increasing the number of patients used in training the model. As the COVID-19 outbreak expands, the accuracy of the model should improve. For the clinician to remember that machine learning provides you with prediction. However, blind reliance on predictive models may lead to automation bias and must be carefully monitored during the implementation of predictive models.


## 6 | CONCLUSION

The overall goal of this approach would be to provide a system to identify high-risk individuals at an early stage to help in allocating

adequate resources, such as ICU beds and to provide necessary interventions before irreversible clinical damage occurs.

## DISCLOSURE

The authors have declared no conflicts of interest for this article.

Sonu Subudhi<sup>1</sup>   
Ashish Verma<sup>2</sup>  
Ankit B. Patel<sup>2</sup>

<sup>1</sup>Gastroenterology Unit, Department of Medicine,  
Massachusetts General Hospital, Harvard Medical School,  
Boston, MA, USA

<sup>2</sup>Renal Division, Department of Medicine, Brigham and Women's  
Hospital, Harvard Medical School, Boston, MA, USA

## Correspondence

Ashish Verma, Renal Medicine, Department of Medicine, 75  
Francis Street, Boston, MA 02115, USA.  
Email: averma8@bwh.harvard.edu

## ORCID

Sonu Subudhi  <https://orcid.org/0000-0002-5937-1880>

## REFERENCES

1. Sun Q, Qiu H, Huang M, Yang Y. Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. *Ann Intensive Care*. 2020;10:33.
2. Phua J, Weng L, Ling L, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *Lancet Respir Med*. 2020;8:506-517.
3. World Health Organization. Coronavirus disease (COVID-2019) situation reports; 2020. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>. Accessed June 23, 2020.
4. Ong ME, Lee Ng CH, Goh K, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Crit Care*. 2012;16:R108.
5. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44:368-374.
6. Tomasev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572:116-119.
7. Fernandes M, Mendes R, Vieira SM, et al. Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PLoS One*. 2020;15:e0229331.
8. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digital Health*. 2020;2:e179-e191.
9. Sevakula RK, Au-Yeung WM, Singh JP, Heist EK, Isselbacher EM, Aroundas AA. State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. *J Am Heart Assoc*. 2020;9:e013924.

10. Jiang X, Coffee M, Bari A, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Comput Mater Continua*. 2020;62:537-551.
11. Yan L, Zhang H-T, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell*. 2020;2:283-288.
12. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Mach Intell*. 1991;13:252-264.
13. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell*. 2019;33:913-933.
14. Cohen JP, Dao L, Morrison P, et al. Predicting COVID-19 Pneumonia Severity on Chest X-ray with Deep Learning. arXiv e-prints. 2020:arXiv:2005.11856. <https://ui.adsabs.harvard.edu/abs/2020arXiv200511856C>. Accessed May 01, 2020.