Machine learning to predict COVID-19 outcomes to facilitate decision making

Sonu Subudhi[1], Ashish Verma[2], and Ankit Patel[2]

[1]Massachusetts General Hospital
[2]Brigham and Women's Hospital

June 8, 2020

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 correlate 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. Here, we propose recommendations to integrate machine learning model with electronic health records so that a real-time risk score can be developed for COVID-19.

Machine learning to predict COVID-19 outcomes to facilitate decision making.

Sonu Subudhi, M.B.,B.S, PhD [1], Ashish Verma M.B.,B.S[2], Ankit B.Patel, MD,PhD[2]

[1]Gastroenterology Unit, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

[2] Renal Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

[*]Corresponding author:

Ashish Verma MBBS

Renal Medicine,

Department of Medicine,

75 Francis Street

Boston, MA, 02115

Email: averma8@bwh.harvard.edu

Conflict of Interest: None

Financial disclosure: None

Keywords: COVID-19, SARS-CoV-2, Machine learning, Artificial Intelligence

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 correlate 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. Here, we propose recommendations to integrate machine learning model with 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 an unprecedented stress on health care infrastructure. Early identification of high-risk patients can allow healthcare workers to allocate their efforts and resources during early clinical course to maximize their impact on patient health. Early critical care management in certain clinical settings has demonstrated improvement in mortality[1]. However, identification of patient's at high risk of progressive and severe disease remains a challenge. Previous methods, such as scoring systems based on clinical signs, perform poorly when novel diseases emerge. Clinical characteristics such as Chest CT findings and lymphopenia are helpful for diagnosis but these predictors fail to show up at early stages of COVID-19. Other characterisitics such as age, gender, and viral load have been associated with COVID-19 severity but have not yet proven to predict disease severity with accuracy[2]. Here we lay out recommendations to implement a machine learning algorithm which would facilitate clinical decision making during outbreaks like COVID-19.

Rationale for machine learning: In the case of the COVID-19 outbreak, there have been more than 800,000 cases in the United States and more than 2.6 million cases worldwide as of April 22, 2020. Given the number of cases, an analog approach to reviewing cases to identify patterns that indicated poor prognosis is not feasible. A large number of cases has particularly stressed the intensive care unit (ICU) settings with increasing needs 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. A prediction model, which can identify patients more likely to deteriorate and require ICU care will allow physicians to allocate manpower and resources in an expeditious and informed manner. Prediction models can also hone in on specific disease and identify the subset of patients that will develop respiratory failure and require ventilators from patients that will develop renal failure and require renal replacement therapy as well as patients that are at risk of requiring both life-supporting treatments. The integration of prediction model with the electronic health record can give physicians immediate information about the expected patient course and predicted response to treatments.

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 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.

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[3]. Another similar machine learning model based on hospital data from a Portuguese and American hospital was able to predict the risk of ICU admission[4]. A study from Denmark using the machine learning model was able to predict 90-day mortality for intensive care unit patients[5]. One key finding of this model was that patient features can interact and compensate for one another and could 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. These examples underscore the capability of machine learning.

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 you are looking to predict, which forms the primary basis of machine learning training. While developing the model, a fraction of patients should be kept out during the training process, to serve as a testing cohort and help validate the accuracy of the model. Once the accuracy of the model significantly improves as compared to no-information-rate, the model could be deployed to new patients. This model would primarily be able to predict a risk score based on new input data from a patient, which would then help clinicians guide treatment based on risk of particular outcomes and plan for future treatment needs.

Machine learning approach was implemented on COVID-19 patient data in China[6,7]. The aim was to predict the severity of disease based on initial presentation data. One of the model was accurately able to predict disease outcome in 90% of the cases[7]. In this model, the most important features used for prediction were lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP). Similar implementations of the machine learning approach in larger cohorts from other countries, can provide more specific models to understand local factors as predictors of disease.

Advantages and challenges: Machine learning models benefit out from larger sample sizes, which in most cases, improve the accuracy of models[8]. For evolving outbreaks, such as COVID-19, as the number of patients increases and more data becomes available for training, the model would likely evolve and become more accurate in predicting disease severity from initial presentation data. Current scoring systems lack this sort of evolving scoring criteria which make them less accurate particularly in novel disease entities were limited data exists.

An added advantage of deploying such a model would be improving patient care by aiding clinicians obtaining data that is most relevant for understanding risk of disease progression. This process can be performed in real-time when an integrated electronic health record can alert the clinican about key data in regards to demographics, clinical characteristics, or laboratory data that would be helpful in predicting patient outcomes.

A key challenge is providing high-quality data for training the predictive model. Variable data or noise could hamper the performance of such a model. It is important to be cautious of the model overfitting the data which can be compensated by increase the number of patients used in training the model. As the COVID-19 outbreak expands, the accuracy of the model should improve. It is important for clinician to remember that machine learning provides you with a prediction. Blind reliance on predictive models leads to automation bias and should be monitored for with implementation of a predictive models.

Conclusion: The overall goal of this approach would be to provide an early clue to future predictions concerning COVID-19. However, such a system in place could act as a model for such future outbreaks as well.

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(1):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.

3. Tomasev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.*2019;572(7767):116-119.

4. 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(3):e0229331.

5. 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. *The Lancet Digital Health.*2020;2(4):e179-e191.

6. Jiang X, Coffee M, Bari A, et al. Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity.*Computers, Materials & Continua.* 2020;62(3):537-551.

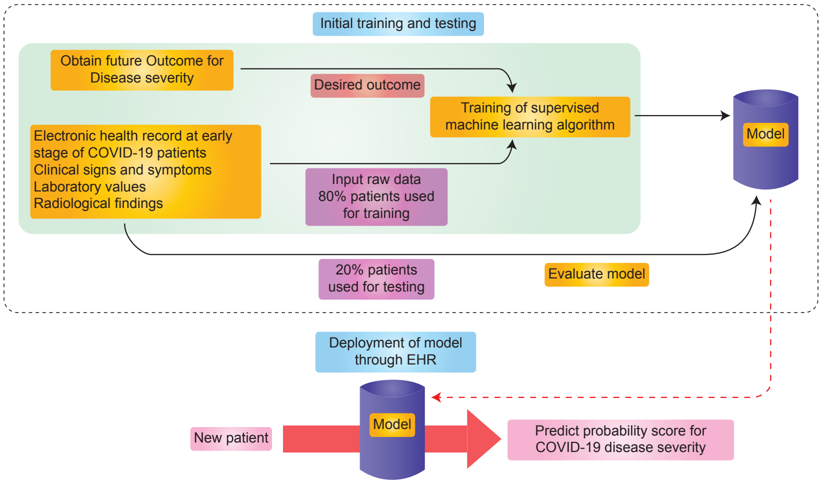
7. Yan L, Zhang H-T, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence.* 2020;2(5):283-288.

8. S.J. Raudys AKJ. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1991;13:252-264.

Figure legend.

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

Figure 1


Initial training and testing
Obtain future Outcome for Disease severity
Desired outcome
Training of supervised machine learning algorithm
Electronic health record at early stage of COVID-19 patients
Clinical signs and symptoms
Laboratory values
Radiological findings
Input raw data 80% patients used for training
Model
20% patients used for testing
Evaluate model
Deployment of model through EHR
New patient
Model
Predict probability score for COVID-19 disease severity