Israel Medical Association Journal (IMAJ), in press

Separating the Wheat from the Chaff: On the Importance of Machine Learning Models in the Fight Against COVID-19 and on the Necessity to Scrutinize Them

Zvia Agur, PhD

Institute for Medical BioMathematics (IMBM), 6099100 Hate'ena Str., 10, Bene Ataroth, Israel

The COVID-19 pandemic has had a profound impact on our world and has cost millions their lives. It has disrupted economies and education systems and has taken away means of support from masses of people around the world. No wonder this pandemic is like a black hole, drawing in all resources and all expertise. In the scientific arena, the pandemic has created a tremendous opportunity for new and exciting synergies between different disciplines. One of the most prominent synergies in the fight against the COVID-19 pandemic uses machine learning to diagnose and prognosticate the disease.

Machine learning is responsible for some of the most sensational technological advancements in modern times, self-driving vehicles, for example, or the discovery of hundreds of exoplanets - planets that orbit stars other than the sun. Machine learning algorithms automatically build a computational model that uses sample data – also known as “training data” – to make decisions without being explicitly programmed to make those decisions. This property renders machine learning especially attractive when medicine faces a global outbreak of a fast-spreading new disease, caused by an unfamiliar virus, which threatens to inflict damage of biblical dimensions. The enormous gap between the almost non-existent knowledge about the disease, on the one hand, and the urgency in finding efficient solutions to it, on the other hand, underscores the potential value of a

method enabling the prediction of processes, such as personal disease progression, with no prior knowledge on the driving forces underlying these processes. Indeed, since the disease outbreak, machine learning was the backbone of thousands of publications suggesting models for the diagnosis or prognosis of people with COVID-19.

Machine learning traces its roots to the 1950s when Arthur Samuel of IBM developed a computer program for playing checkers, coining the term "Machine Learning" for mechanisms he designed, which allowed his program to improve [1]. But machine learning remained a niche area for decades, taking off only in the 21st century when increasing computing power and gigantic amounts of data converged to finally take full advantage of machine learning algorithms, which require massive data and fast processing speed to be useful. Yet, until recently, the contribution of this field to healthcare was limited. The COVID-19 pandemic has changed this, providing the impetus for the increasing willingness of physicians to join forces with data scientists in the quest for solutions for the long list of unknowns of the current crisis.

The downside of this exciting development is the need to materialize the new synergy straightaway, whereas fruitful collaboration depends on thorough interdisciplinary understanding, which demands time and effort: the data scientists should understand the crucial needs of the physicians, and their practical limitations, while the physicians should be able to evaluate the quality and the feasibility of applying the proposed machine learning tools. Unfortunately, most of the machine learning-based prediction models for COVID-19, published thus far, are fraught with faults in both the methodology itself, the suitability of the data used for model development, the validation of model accuracy, and the applicability to the clinic [2-5].

Take, for example, the work by Arjun S Yadaw and colleagues from the Icahn School of Medicine at Mount Sinai, New York, USA, in *The Lancet Digital Health* [6]. Yadaw and colleagues present machine learning models predicting mortality during medical encounters of unspecified duration, in patients with COVID-19, admitted to the Mount Sinai Health System in the New York City area. The researchers highlight one of the models they developed, which is based on three features: patient's age, minimum oxygen saturation throughout their medical encounter, and type of patient encounter (inpatient, outpatient, or telehealth visits). They use a relatively large patient dataset for model development (n=3841), the number of patients who died (n=313) seems appropriate for the statistical analysis [7], and high accuracy is achieved in model validation (AUC of 0.91). The authors propose to use this model in clinical settings to guide the management and prognostication of patients affected by the COVID-19 disease.

But the experienced reader is not convinced by the proposition of Yadaw and colleagues. In their paper [6], the authors mention some of the caveats hampering the clinical use of the model, notably, insufficient external validation of its accuracy. But the unmentioned methodological problems in the work seem to be insurmountable. Essentially, the highlighted model predicts death using measurements collected throughout the entire encounter of the patient with the health system, with no specific moment at which the prediction is generated and tested. This raises questions about the actual prognostic value of the only time-varying model parameter - the minimum oxygen saturation, and about when and how the model should be used. As the predictive value of time-varying clinical parameters tends to increase when measured closer to the outcome - in this case, death of the patient - it remains unclear how to interpret the reported performance measurement,

i.e., the area under the curve of 91%, vis-à-vis the time of measurement of this time-varying predictor [3]. The mere definition of the minimum saturation as the lowest value of oxygen saturation over the entire encounter [6], implies that the prediction itself becomes immaterial at the time it is created – when the patient is already dying or discharged. Furthermore, in [6], patients who did not die by the end of the study were considered as remaining alive. But since the death of these patients might have occurred after the study ended, the actual incidence of mortality could be underestimated, putting in doubt the value of the minimum oxygen saturation as a sole time-varying predictor [3]. A possible solution for such a conundrum may be to fix a short-term prediction scope, such as, "predict death in the coming twenty-four hours". But from the applicability point of view, the fixed follow-up window should be carefully determined, to allow sufficient time for efficacious relief of the predicted fatal outcome, e.g., by corticosteroids [8].

The result of Yadaw and colleagues that the minimum oxygen saturation is responsible for the high predictive capacity of the model is striking also from another point of view: even though the leading cause of death of critically ill patients with COVID-19 is a refractory respiratory failure (45%), more than half of the deceased patients suffer from other failures, such as cardiac arrest or hemorrhagic and ischemic strokes [9]. Therefore, it is not clear how minimum oxygen saturation represents almost all the potentially deceased patients in [6]. Overestimation of the model accuracy is conceivable in this case, due to potential correlations between the consecutive measurements over time in the same patients.

The analysis of Yadaw and colleagues' work surfaces some of the prerequisites for prediction models to become more helpful in the clinic. Better collaboration is necessary among researchers from different backgrounds, clinicians, and institutes for determining

the clinical need and for sharing patient data from COVID-19 studies and registries. Another issue is the requirement for external model validation, currently much complicated by the incompatibility of the recording in different hospitals. Consensual representation of the patient's follow-up and treatments is required for allowing external validation of prediction models and their subsequent generalization. Most important, in this context, is the necessity to adhere to unified sets of criteria for evaluating prediction models, e.g., the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) set of recommendations [10], or the Prediction Model Risk Of Bias ASsessment Tool criteria (PROBAST), which enable accurate evaluation of the risk of bias and applicability of a prediction model [2]. Another important way to refine the prognostic model landscape is by a critical analysis of the diverse modeling efforts, and by recommendations for their improvement.

At present, there is an urgent need to separate the wheat from the chaff and underline those predictive models which can become useful in the clinic. But how can one do this? The pandemic has created huge amounts of information, which the traditional method of academic reporting cannot encompass. As a result, atlases, and catalogs, covering extensive disease-related data, acquire a special status these days. An example is the multi-omics blood atlas of immune profiles of patients with varying COVID-19 severity, back-to-back with the immune profiles of patients with influenza or sepsis, and with healthy volunteers. This massive work, by more than two hundred scientists from many research centers, could aid future drug developers and designers of precision medicine modalities [11]. Another example is the COVID-19-related mortality dataset by Karlinsky and Kobak [12], which the authors use to compute the excess mortality in each country during the

COVID-19 pandemic and identify the countries which have been substantially underreporting their COVID-19 deaths.

The review article by Shapiro and colleagues [13], from Tel Aviv Sourasky Medical Center, joins this new class of publications. The paper aims at separating the wheat from the chaff in the multitude of prognostic models for COVID-19, by cataloging and scrutinizing the major models for classifying patients at risk of deterioration. The authors discuss the tools at our disposal for critical model assessment and evaluate the clinical adequacy of the analyzed models. First, Shapiro and colleagues discuss scoring systems, both established scores, and scores designed specifically for COVID-19 patients. Then, they list and analyze models that use machine learning to predict risk in COVID-19 patients. Shapiro and colleagues provide a comprehensive table of models and their main attributes, including their point of view on the highlights and difficulties in each of the models. Upon regular update, this table can serve as a concise navigation map in the turbulent water of machine learning risk predictors for COVID-19.

Conclusions

Ultimately, one should test the prediction models in prospective clinical trials and evaluate how they objectively improve the clinical outcomes. Thereupon, these models may be used to better triage patients to an appropriate level of care, streamline resource allocation, improve care in times of hospital overload, and optimize the timing of disease-modifying treatment. Well-validated prediction models can empower care teams and healthcare administrators to make the right decisions under stress.

1. Keith DF. A brief history of machine learning: Dataversity (2021). Available from: https://www.dataversity.net/a-brief-history-of-machine-learning/.
2. Miller JL, Tada M, Goto M, Chen H, Dang E, Mohr NM, et al. Prediction models for severe manifestations and mortality due to COVID-19: A systematic review. *Acad Emerg Med* (2022) 29(2):206-16.
3. Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. *Lancet Digit Health* (2020) 2(10):496-7.
4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* (2020) 369:1328.
5. Bachtiger P, Peters NS, Walsh SL. Machine learning for COVID-19-asking the right questions. *Lancet Digit Health* (2020) 2(8):391-2.
6. Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health* (2020) 2(10):516-25.
7. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* (2020) 368:441.
8. Bartoletti M, Marconi L, Scudeller L, Pancaldi L, Tedeschi S, Giannella M, et al. Efficacy of corticosteroid treatment for hospitalized patients with severe COVID-19: a multicentre study. *Clin Microbiol Infect* (2021) 27(1):105-11.
9. Contou D, Cally R, Sarfati F, Desaint P, Fraisse M, Plantefeve G. Causes and timing of death in critically ill COVID-19 patients. *Crit Care* (2021) 25(1):79.
10. Goodacre S, Thomas B, Sutton L, Burnsall M, Lee E, Bradburn M, et al. Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: The PRIEST observational cohort study. *PLoS One* (2021) 16(1):e0245840.
11. Consortium C-M-oBA. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* (2022) 185(5):916-38 e58.
12. Karlinsky A, Kobak D. The World Mortality Dataset: Tracking excess mortality across countries during the COVID-19 pandemic. *medRxiv* (2021) (10):e69336.
13. Shapiro M, Yavne Y, Shepshelovich D. Predicting which patients are at risk for clinical deterioration in COVID-19 – an extensive review of the current models in use. *IMAJ* (In Press).