Prediction models for COVID-19 clinical decision making

As of Sept 2, 2020, more than 25 million cases of COVID-19 have been reported, with more than 850 000 associated deaths worldwide. Patients infected with severe acute respiratory syndrome coronavirus 2, the virus that causes COVID-19, could require treatment in the intensive care unit for up to 4 weeks. As such, this disease is a major burden on health-care systems, leading to difficult decisions about who to treat and who not to.[1] Prediction models that combine patient and disease characteristics to estimate the risk of a poor outcome from COVID-19 can provide helpful assistance in clinical decision making.[2]

In a living systematic review by Wynants and colleagues,[3] 145 models were reviewed, of which 50 were for prognosis of patients with COVID-19, including 23 predicting mortality. Critical appraisal of these models showed a high risk of bias for all models (eg, because of a high risk of model overfitting and unclear reporting on intended use of the models, or because of no reporting of the models' calibration performance). Moreover, external validation of these models, deemed essential before application can even be considered, was rarely done. Therefore, use of any of these reported prediction models was not recommended in current practice.

In *The Lancet Digital Health*, Arjun S Yadaw and colleagues present two models to predict mortality in patients with COVID-19 admitted to the Mount Sinai Health System in the New York city area.[4] These researchers have addressed many of the issues encountered by Wynants and colleagues[3] and provide extensive information about the modelling in the appendix. The dataset used for model development (n=3841) is larger than in most currently published models, and the accompanying number of patients who died (n=313) seems appropriate according to the prediction model risk of bias assessment tool (PROBAST)[5] and guidance on sample size requirements for prediction model development.[6] The calibration performance of the models is reported, which (although essential) is often missing, particularly in studies reporting on machine-learning algorithms,[7] and external validations of the models was done. Yadaw and colleagues acknowledge that additional external validation will be necessary[4] because external validation was done in a random subset of the initial patient population and another set of recent patients from the same health system, and because the number of events in the validation sets were below the 100 suggested for reliable external validity assessment.[8]

For other researchers to apply and externally validate models, adherence to transparent reporting of a multi-variable prediction model for individual prognosis or diagnosis (TRIPOD) criteria[9] is advised to present the full models, accompanied by code in case of complex machine-learning models. Yadaw and colleagues reported many items in TRIPOD, however, the models themselves are not reported in the Article or appendix (item 15a of TRIPOD) so it is not possible for a reader to make predictions for new individuals (eg, to validate the developed models in their own data or investigate the contribution of the individual predictors).

The moment for risk estimation defines which values of predictors will be available and is especially important for time-varying predictors (eg, temperature). The models reported by Yadaw and colleagues predict risk using measurements collected throughout the entire encounter of the patient with the health system, with no specific moment of prediction defined.[4] This raises questions about the actual prognostic value of the time-varying predictors (eg, the minimum oxygen saturation) and, hence, how and when the model should be used as the predictive value of time-varying predictors will likely increase when measured closer to the outcome. Consequently, it remains unclear how to interpret the reported area under the curve of approximately 90% in relation to the moment of measurement of these time-varying predictors.

Two suggestions can be made regarding modelling. First, the current machine-learning models were constructed using the default hyperparameter values provided by the respective software packages. These often provide reasonable starting values, but important hyperparameters should be carefully tuned to the specific use case.[10] Second, as acknowledged by Yadaw and colleagues,[4] patients who had not developed the outcome by the end of the study were considered not to have the outcome. Since the outcome for these patients might occur after the study ended, the actual incidence of mortality could have been underestimated. Alternatively, a fixed follow-up period per patient could

have been defined to allow sufficient follow-up time to measure the outcome in each patient.

The study by Yadaw and colleagues ticks a lot of boxes,[4] but it still struggles somewhat to break away from the overall negative picture painted by Wynants and colleagues.[3] Improvements can be achieved by more and better collaboration among researchers from different backgrounds, clinicians, and institutes and sharing of patient data from COVID-19 studies and registries. Then, and with improved reporting (by adherence to TRIPOD criteria), validity, and quality (according to PROBAST), prediction models can provide the decision support that is needed when COVID-19 cases and hospital admissions will again test the limits of the health-care system.

We declare no competing interests.

*Artuur M Leeuwenberg, *Ewoud Schuit*
e.schuit@umcutrecht.nl

Cochrane Netherlands (ES), Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht 3508 GA, Netherlands (AML, ES)

1   Emanuel EJ, Persad G, Upshur R, et al. Fair allocation of scarce medical resources in the time of Covid-19. *N Engl J Med* 2020; **382:** 2049–55.
2   Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009; **338:** b375.
3   Wynants L, Van Calster B, Bonten MM, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.
4   Yadaw AS, Li Y-c, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digital Health* 2020; **2:** e516–25.
5   Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; **170:** 51–58.
6   Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368:** m441.
7   Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17:** 1–7.
8   Vergouwe Y, Steyerberg E, Eijkemans R, Habbema D. Sample size considerations for the performance assessment of predictive models: a simulation study. *Control Clin Trials* 2003; **24:** 43s–44s.
9   Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Circulation* 2015; **131:** 211–19.
10   Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. *J Mach Learn Res* 2019; **20:** 1–32.