BMJ Health & Care Informatics

Validation of parsimonious prognostic models for patients infected with COVID-19

Keerthi Harish ,[1] Ben Zhang,[2] Peter Stella,[3] Kevin Hauck,[4] Marwa M Moussa,[4] Nicole M Adler,[4] Leora I Horwitz,[1,4] Yindalon Aphinyanaphongs[1]

To cite: Harish K, Zhang B, Stella P, *et al*. Validation of parsimonious prognostic models for patients infected with COVID-19. *BMJ Health Care Inform* 2021;**28**:e100267. doi:10.1136/bmjhci-2020-100267

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2020-100267).

KH, BZ and PS are joint first authors.

Received 20 October 2020
Accepted 19 July 2021

[1]Department of Population Health, New York University School of Medicine, New York, New York, USA
[2]Department of Radiology, New York University School of Medicine, New York, New York, USA
[3]Department of Pediatrics, New York University School of Medicine, New York, New York, USA
[4]Department of Medicine, New York University School of Medicine, New York, New York, USA

Correspondence to
Dr Yindalon Aphinyanaphongs; yin.a@nyulangone.org

ABSTRACT

Objectives Predictive studies play important roles in the development of models informing care for patients with COVID-19. Our concern is that studies producing ill-performing models may lead to inappropriate clinical decision-making. Thus, our objective is to summarise and characterise performance of prognostic models for COVID-19 on external data.

Methods We performed a validation of parsimonious prognostic models for patients with COVID-19 from a literature search for published and preprint articles. Ten models meeting inclusion criteria were either (a) externally validated with our data against the model variables and weights or (b) rebuilt using original features if no weights were provided. Nine studies had internally or externally validated models on cohorts of between 18 and 320 inpatients with COVID-19. One model used cross-validation. Our external validation cohort consisted of 4444 patients with COVID-19 hospitalised between 1 March and 27 May 2020.

Results Most models failed validation when applied to our institution's data. Included studies reported an average validation area under the receiver–operator curve (AUROC) of 0.828. Models applied with reported features averaged an AUROC of 0.66 when validated on our data. Models rebuilt with the same features averaged an AUROC of 0.755 when validated on our data. In both cases, models did not validate against their studies' reported AUROC values.

Discussion Published and preprint prognostic models for patients infected with COVID-19 performed substantially worse when applied to external data. Further inquiry is required to elucidate mechanisms underlying performance deviations.

Conclusions Clinicians should employ caution when applying models for clinical prediction without careful validation on local data.

Summary

What is already known?

- The novelty of COVID-19 resulted in a knowledge gap regarding the clinical trajectory of hospitalized patients. In an effort to address this knowledge gap, researchers have developed and published models to estimate the prognosis of hospitalised patients. These models have performed well on data from populations similar to those used to construct them. In general, however, models are known to perform poorer on populations different from those used to train them.

What does this paper add?

- The ability of models to predict patients' clinical courses is substantially impaired when such models are applied to real-world data. As such, published external models are unlikely to be appropriate as significant, reliable inputs for clinical decision making. This study serves as a reminder that predictive models should be carefully applied in new settings only after local validation.

INTRODUCTION

COVID-19 is a rapidly growing threat to public health. As of 4 October 2020, over 35 million positive cases and over 1 million deaths have been reported.[1] While most of these deaths have occurred in older patients and those with chronic disease, outcomes even within these strata are highly variable.[2] Given the large number of cases and limited healthcare resources, there exists substantial need for predictive models that allow healthcare providers and policymakers to estimate prognoses for individual patients.

Several such models have been published or made available in preprint. Many have been derived through machine learning techniques to identify a reasonably small set of features that are predictive of poor outcomes in order to make their application in other settings feasible. While these models have generally performed well when applied to their own 'held-out' data, it is well known that such models are often biased and rarely perform as well on 'real-world' data. A systematic review and critical appraisal by Wynants *et al* found that prognostic models examined were at a high risk of bias and postulated that real-world performance on these models would likely be worse than that reported.[3]

A secondary concern is the use of these prognostic models in a clinical setting without validation. A paper that reports specific prognostic factors may misinform providers about trends, relationships and associations and inadvertently drive faulty decision-making regarding prognosis and treatment decisions.

In order to evaluate applicability to data from American patients, we report the performance of 10 such prognostic models on data from New York University (NYU) Langone Health, a multisite hospital system in New York City.

METHODS

Literature review

We searched PubMed, arXiv, medRxiv and bioRxiv for papers reporting prognostic predictive models between 1 January 2020 and 3 May 2020. Queries were constructed by combining COVID-19 illness with terms denoting predictive or parsimonious models (online supplemental table A). Results were supplemented with individual hits from Google Scholar searches using the same queries. Both peer-reviewed articles and preprint manuscripts were considered.

Search results were subjected to six inclusion criteria:

1. The model was developed using patients with COVID-19. Models approximating COVID-19 using other types of viral pneumonia were excluded.
2. The model predicted prognosis of individual cases. Various targets were considered, including mortality, intensive care unit transfer and WHO definitions of severe and critical illness.[2] Models seeking to predict diagnostic test results or epidemiological trends were excluded.
3. The model used only clinical and/or demographical factors. The American College of Radiology has outlined contamination-related and technical challenges associated with the use of imaging for patients with COVID-19.[4] Given these challenges, models requiring the use of chest radiographs or CT scans were excluded.
4. The model was parsimonious, involving fewer than 20 features. Models with large numbers of features require collection of more information from patients and are difficult to reliably apply to other settings.
5. The model was validated on a held-out test set (internal validation), on an outside dataset (external validation) or via cross-validation. Reporting training performance alone, without one of these three forms of author-facilitated validation (online supplemental table B), was not sufficient.
6. The model is reportedly applicable as a prediction model. Classification models, which report on a snapshot in time, were not included.

In order to effectively rebuild and assess each model, we further subjected search results to two exclusion criteria:

1. The model used features assessed within standard of care protocols at our institution. Features outside standard of care include unique laboratory values such as T cell subtyping and epidemiological factors such as travel history.
2. The model used features and targets with characterisable definitions. The feature 'other precondition', for example, cannot be succinctly and reliably characterised.

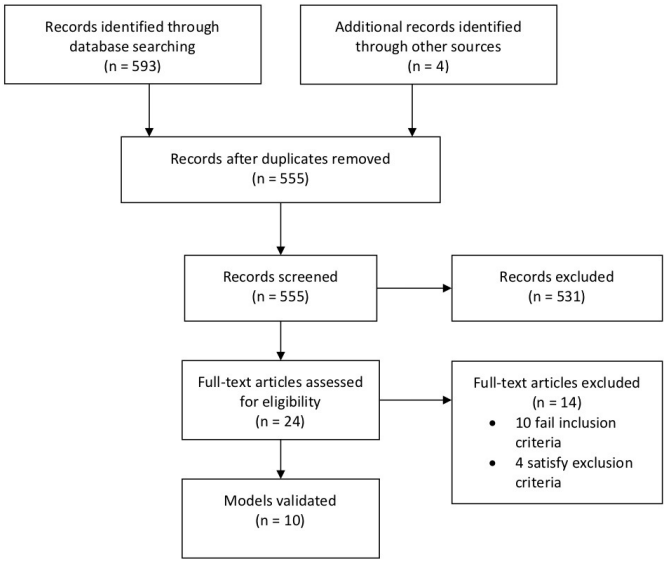

Figure 1 Literature search screening and selection.

Our selection process (figure 1) yielded ten studies, as summarised in table 1. Each study's model parameters are detailed in table 2. These models were subsequently applied on our own NYU validation dataset, shown in table 3.

Evaluation methods

Evaluating models requires four types of information: features, feature weights, population inclusion and exclusion criteria and targets. Studies were categorised based on the degree of information reported. Five papers reported feature weights,[5–9] and the remaining five did not report feature weights.[10–14] Those that reported feature weights, whether explicitly or through another elucidating form, such as a nomogram, were applied directly to our external validation cohort (applied models). For those papers that lacked feature weights, we rebuilt models by using reported features (rebuilt models). When discussed, we replicated construction of those models, including the train-to-test split and cross-validation. Where construction was not discussed, we performed a default 8:2 train-to-test split and threefold cross-validation to choose hyperparameters. CIs were estimated using the DeLong method.

All studies reported population inclusion and exclusion criteria and targets. We were able to run four models to these reported specifications (models without deviations).[6 7 10 11] For the remaining six models, we deviated from the reported specifications (models with deviations) for one of two reasons.[5 8 9 12–14] First, the models defined criteria using data that were not collected at our institution. For example, Gong *et al*[5] defined a partial pressure of oxygen to fraction

Table 1 Summary of studies selected

Study	Task	Training cohort—reported from study	Validation cohort—reported from study
Gong *et al*[15]	Predict progression to severe pneumonia in 15 days	189 inpatients with COVID-19	External 1: 165 inpatients with COVID-19 External 2: 18 inpatients with COVID-19
Zhou *et al*[8]	Predict no progression to severe pneumonia in 1 day	250 inpatients with COVID-19	Internal: 127 inpatients with COVID-19
Zou *et al*[9]	Predict 7.5-day survival rate	445 inpatients with COVID-19	Internal: 224 inpatients with COVID-19
Xie *et al*[5]	Predict in-hospital mortality	299 inpatients with COVID-19	External: 145 inpatients with COVID-19
Yan *et al*[6]	Predict in-hospital mortality	375 inpatients with severe COVID-19	Internal: 110 inpatients with severe COVID-19
Levy *et al*[10]	Predict 14-day survival for hospitalised patients with COVID-19	5233 inpatients with COVID-19	Cross-validation
Zhang *et al*[11]	Task 1: predict in-hospital mortality Task 2: predict in-hospital deterioration	775 inpatients with COVID-19	External: 220 inpatients with COVID-19
Guo *et al*[12]	Predict in-hospital deterioration	818 inpatients with mild or moderate COVID-19	External: 320 inpatients with mild or moderate COVID-19
Hu *et al*[13]	Predict in-hospital mortality	182 inpatients with COVID-19	External: 64 inpatients with COVID-19
Carr *et al*[14]	Predict 14-day deterioration	452 inpatients with COVID-19	External: 256 inpatients with COVID-19

of inspired oxygen ratio(PaO_2/FiO_2)threshold target. We had to exclude this target because PaO_2 was not commonly recorded in our dataset. Second, the models defined criteria using labels that are not characterisable. For example, Zhou *et al*[8] used severe respiratory distress. Severe respiratory distress is a subjective measure of acuity and not defined explicitly in the study. Thus, this target was excluded. In general, the features used in selected models represented results from clinical tests used commonly across facilities. Commonly used tests include complete blood counts and metabolic panels.

Therefore, the 10 studies were split into four designations: (1) models applied without deviations (table 4), (2) models applied with deviations (table 5), (3) models rebuilt without deviations (table 6) and (4) models rebuilt with deviations (table 7). Area under the receiver–operator curve (AUROC) was used as our main measure of model performance, with F1 score used as a secondary measure of model performance, if used in the original study.

Validation cohort

The 4444 inpatients with COVID-19 in our validation cohort were admitted after 1 March 2020 and were followed until either discharge or the occurrence of an outcome on or before 27 May 2020. Outcomes include any of those listed in table 2. Some papers did not specify the prediction time. If prediction time was not specified, we used the earliest data points available. We excluded patients with missing features on a case-by-case basis, as determined by the range of features required by each model. If the minimum set of features was not available for a patient, this patient was excluded from the evaluation. An overview of the NYU validation cohort used in each study is shown in table 3. For reference, a comparison of cohort demographics is available in the supplement (online supplemental table C).

RESULTS

We summarise our results in multiple tables.

Table 4 shows the performance of models applied without deviations. In these studies, we applied the model as reported in the respective paper, as is. The study reported mean AUROC dropped from 0.98 to 0.67 when applied to our dataset, with a mean AUROC difference of 0.31. When we retrained against our own data, the mean AUROC dropped from 0.98 to 0.82, with a mean difference of 0.21. In this cohort of models, the models do not validate optimally.

Yan *et al* reported performance metrics using the most recent laboratory values taken from patients.[7] However, the study claims that the published model can be used to predict outcomes several days in advance.[7] For this reason, we have evaluated the model using both patients' earliest and most recent laboratory values. We consider the earliest laboratory values as the preferable model to evaluate. This model gives the longest lead time towards patient prognosis.

Table 5 shows the performance of models applied with deviations. In these studies, we applied the model as reported

Table 2 Model parameters

Study	Model construction	Features	Predicted outcome	Data processing
Gong *et al*[5]	Nomogram built from LASSO algorithm and logistic regression model	Age, direct bilirubin, BUN, CRP, LDH, albumin, RDW	Shortness of breath, respiratory rate ≥30/min, SpO_2 ≤93% or PaO_2/FiO_2 ≤300 mm Hg	Variables with >7% values missing excluded. Other missing variables imputed by expectation maximisation
Zhou *et al*[8]	Formula derived from logistic regression	Neutrophil count, lymphocyte count, D-dimer	Respiratory ≥30/min, severe respiratory distress, SpO_2 <90% on room air, ARDS, sepsis or septic shock	Not discussed
Zou *et al*[9]	Nomogram derived from logistic regression and Cox regression model	Age, disturbance of consciousness (GCS under 15,), LDH, CRP, chronic heart disease, chronic renal insufficiency, septic shock	In-hospital mortality	Not discussed
Xie *et al*[5]	Nomogram derived from logistic regression	Age, LDH, log (lymphocyte count), SpO_2	In-hospital mortality	Not discussed
Yan *et al*[6]	Multitree XGBoost	LDH, lymphocyte percent, CRP	In-hospital mortality	Pregnant, lactating women, minors, cases with data <80% complete excluded. Missing data '–1' padded
Levy *et al*[10]	LASSO regression, multivariate logistic regression	Age, BUN, Emergency Severity Index, RDW, neutrophil count, serum bicarbonate, serum glucose	In-hospital mortality	Not discussed
Zhang *et al*[11]	Logistic regression	Age, sex, neutrophil count, lymphocyte count, platelet count, CRP, D-dimer, creatinine, ALT	Task 1: ARDS, intubation, ECMO, ICU admission or in-hospital mortality Task 2: ARDS, intubation, ECMO, ICU admission or in-hospital mortality	Not discussed
Guo *et al*[12]	Cox regression	Age, underlying chronic disease, neutrophil-to-lymphocyte ratio, CRP, D-dimer	Shortness of breath, respiratory rate ≥30/min, SpO_2 ≤93%, PaO_2/FiO_2 ≤300 mm Hg, respiratory failure with mechanical ventilation, circulatory shock, multiple organ or failure with ICU admission	Cases with missing data excluded
Hu *et al*[13]	Logistic regression	Age, CRP, lymphocyte count, D-dimer	In-hospital mortality	Missing values imputed using bagging trees
Carr *et al*[14]	Logistic regression augmented by XGBoost	NEWS2 (respiratory rate, SpO_2, systolic BP, heart rate, GCS<15, temperature, supplemental oxygen (binary)), CRP, neutrophil count, eGFR, albumin, age	In-hospital mortality or ICU admission	Not discussed

ALT, alanine aminotransferase; ARDS, acute respiratory distress syndrome; BP, blood pressure; BUN, blood urea nitrogen ; CRP, C-reactive protein; ECMO, extracorporeal membrane oxygenation; eGFR, estimated glomerular filtration rate; FiO2, fraction of inspired oxygen; GCS, Glascow coma score ; ICU, intensive care unit; LASSO, least absolute shrinkage and selection operator; LDH, lactate dehydrogenase ; NEWS2, national early warning score 2; PaO2, partial pressure of oxygen; RDW, erythrocyte distribution width; SpO2, oxygen saturation.

Table 3 Summary of validation cohorts

Author	Numbers of participants with outcome	Numbers of participants without outcome	Percent cohort excluded for missing values	Follow-up time
Gong *et al*[5]	912	1107	55	1 March to 27 May
Zhou *et al*[8]	1900	1397	26	
Zou *et al*[9]	418	4026	0	
Xie *et al*[6]	885	3333	5	
Yan *et al*[7]	848	2676	21	
Levy *et al*[10]	616	2868	22	
Zhang *et al*[12]	Task 1: 814	Task 1: 2819	30	
	Task 2: 881	Task 2: 2327		
Guo *et al*[13]	508	1751	49	
Hu *et al*[14]	834	2869	35	
Carr *et al*[11]	1341	3072	1	

in the respective paper with deviations as outlined in Methods section. The study reported mean AUROC dropped from 0.83 to 0.66 when applied to our dataset, with a mean AUROC difference of 0.19. When we retrained against our own data, the mean AUROC dropped from 0.83 to 0.71, with a mean difference of 0.13. In this cohort of models, the models do not validate optimally.

Table 6 shows the performance of models rebuilt without deviations. In these studies, we rebuilt the model with our data using the features outlined in the respective paper. After retraining, the mean AUROC increased slightly, from 0.73 to 0.76, with a mean difference of 0.02. Levy *et al* did not report a testing AUROC. However, we rebuilt the model and made the comparison.[10] We note a small increase in performance; however, Levy *et al* do not report a validation performance, and the bump may be a statistical artefact.[10]

Table 7 shows the performance of models rebuilt with deviations. In these studies, we rebuilt the model with our data with deviations as outlined in the table. After retraining, the mean AUROC dropped slightly, from 0.78 to 0.75, with a mean difference of 0.03.

Table 8 summarises the bottom-line results from table 4 to table 7. Studies are stratified by each of the four study types: studies applied without deviation, studies applied with deviation, studies rebuilt without deviation and studies rebuilt with deviation. Because not all studies reported AUROC values for study validation performances, not all studies are represented where mean values are given. N is shown for all mean values.

We make a few observations. First, the applied models perform more poorly than the rebuilt models. This poor generalisation is expected as models are transferred from one setting to another. However, such a large difference is not expected, and likely, there are methodological errors in model construction in the original papers, or the sample cohorts are significantly different. We believe though that the sample cohorts are quite similar. The rebuilt models perform close to the reported studies. This result implies that the cohorts and the features that define them are similar.

Table 9 summarises results from table 4 to table 7. Studies are stratified by three types of tasks: models predicting only clinical deterioration, models predicting only clinical mortality or models that

Table 4 Performance of models applied without deviations

Study	Validation performance—reported from study	Validation performance—NYU data	Performance difference between study validation performance and NYU original validation performance	Validation performance—NYU retrained (95% CI)	Performance difference between study validation performance and NYU retrained validation performance
Xie *et al*[6]	AUROC=0.98	AUROC=0.67	AUROC difference=0.31	AUROC=0.76 (0.72 to 0.80)	AUROC difference=−0.22
Yan *et al*[7]	Avg F1=0.97	Most recent values: Avg F1=0.63	F1 difference=0.34	Most recent values: AUROC=0.95 (0.93 to 0.96)	*
Yan *et al*[7]	*	Earliest values: Avg F1=0.51	*	Earliest values: AUROC=0.70 (0.65 to 0.75)	*
	Mean AUROC=0.98	**Mean AUROC**=0.67	**Mean AUROC difference**=0.31	**Mean AUROC**=0.82	**Mean AUROC difference**=−0.22

*Value unavailable because authors did not provide an AUROC value when reporting validation performance.
AUROC, area under the receiver–operator curve; NYU, New York University.

Table 5 Performance of models applied with deviations

Study	Validation performance—reported from study	Validation performance—NYU data	Performance difference between study validation performance and NYU original validation performance	Validation performance—NYU retrained (95% CI)	Performance difference between study validation performance and NYU retrained validation performance	NYU deviations from original
Gong *et al*[5]	Cohort 1: AUROC=0.85	AUROC=0.59	AUROC difference=0.26	AUROC=0.68 (0.64 to 0.74)	AUROC difference=0.17	PaO_2 data not available. Target excluded
Gong *et al*[5]	Cohort 2: AUROC=0.75	AUROC=0.60	AUROC difference=0.16	AUROC=0.68 (0.64 to 0.74)	AUROC difference=0.07	PaO_2 data not available. Target excluded
Zhou *et al*[8]	AUROC=0.88	AUROC=0.74	AUROC difference=0.14	AUROC=0.75 (0.71 to 0.78)	AUROC difference=0.14	'Severe respiratory distress' target not characterisable. Target excluded
Zou *et al*[9]	*	AUROC=0.70	*	AUROC=0.72 (0.67 to 0.77)	*	Altered mental status measured as Glasgow Coma Score <15
	Mean AUROC=0.83	**Mean AUROC**=0.66	**Mean AUROC difference**=0.19	**Mean AUROC**=0.71	**Mean AUROC difference**=0.13	

*Value unavailable because authors did not provide an AUROC value when reporting validation performance.
AUROC, area under the receiver–operator curve; NYU, New York University; PaO2, partial pressure of oxygen.

predict the occurrence of either clinical deterioration or mortality. Because not all studies reported AUROC values for study validation performances and not all studies provided feature weights, not all studies are represented where mean values are given. N is shown for all mean values. In general, predicting mortality is easier than deterioration. Both deterioration and mortality tasks do not generalise against the reported results. The mean AUROC differences for predicting deterioration and mortality respectively are 0.10 and 0.15. Finally, predicting either deterioration or mortality is consistent but poor. We also note that in the mean AUROC differences for those studies with compound tasks, we were unable to apply them and verify that the study weights are clinically useful.

By rebuilding each model using its features, we were able to elucidate positive predictive value–sensitivity relationships. We show these results to further make the justification of clinical applicability and the potential false positives that the various models may produce. Table 10 shows the positive predictive values of rebuilt models given a sensitivity threshold. Only two models achieved average positive predictive value scores over 0.75: Yan *et al* (given the most recently taken laboratory values as features) and Guo *et al*[7 13]. We note that, in the case of Yan *et al*, using the most recent values effectively renders the model a classifier rather than a predictor.[7]

DISCUSSION

Principal findings

Prognostic models for COVID-19 may be able to provide important decision support to policymakers and clinicians attempting to make treatment and resource allocation decisions under adverse circumstances. Several such models have been developed and have reported excellent performance on held-out data from their own sources. Unfortunately, when applied to data from a large

Table 6 Performance of models rebuilt without deviations

Study	Validation performance—reported from study	Validation performance—NYU retrained (95% CI)	Performance difference between study validation performance and NYU retrained validation performance
Levy *et al*[10]	*	AUROC=0.80 (0.75 to 0.84)	*
Carr *et al*[11]	AUROC=0.73	AUROC=0.74 (0.71 to 0.78)	AUROC difference=0.01
	Mean AUROC=0.73	**Mean AUROC**=0.77	**Mean AUROC difference**=0.01

*Value unavailable because authors did not provide an AUROC value when reporting validation performance.
AUROC, area under the receiver–operator curve; NYU, New York University.

Table 7 Performance of models rebuilt with deviations

Study	Validation performance—reported from study	Validation performance—NYU retrained (95%CI)	Performance difference between study validation performance and NYU retrained validation performance	NYU deviations from original
Zhang *et al*[12]	Task 1: AUROC=0.74	Task 1: AUROC=0.78 (0.69 to 0.86)	AUROC Difference=+0.04	ECMO, ARDS and intubation targets excluded. These targets were excluded during original validation
Zhang *et al*[12]	Task 2: AUROC=0.72	Task 2: AUROC=0.69 (0.64 to 0.73)	AUROC difference=−0.03	ECMO, ARDS and intubation targets excluded. These targets were excluded during original validation
Guo *et al*[13]	AUROC=0.78	AUROC=0.79 (0.74 to 0.84)	AUROC difference=+0.01	PaO_2 and radiographic progression data not available. Circulatory shock and multiorgan dysfunction not characterisable. ICU admission used as surrogate target
Hu *et al*[14]	AUROC=0.88	AUROC=0.74 (0.69 to 0.78)	AUROC difference=−0.14	PaO_2 data not available. Target excluded
	Mean AUROC=0.78	**Mean AUROC=**0.75	**Mean AUROC difference=**0.03	

ARDS, acute respiratory distress syndrome; AUROC, area under the receiver–operator curve; ECMO, extracorporeal membrane oxygenation; ICU, intensive care unit; NYU, New York University; PaO_2, partial pressure of oxygen.

American healthcare system, the predictive power of these models is substantially impaired.

While disappointing, this loss of performance is not surprising. Multiple differences may account for the performance loss, such as different populations, different viral strains, different clinical workflows and treatments, lab variations, small sample model construction, poor experimental design and general overfitting.

Meaning of the study

Translating any model to apply to data from another source inevitably introduces error caused by divergence in how data are defined and represented. While strictly 'unfair' to the models under consideration, this issue directly results from attempting to implement any algorithm in a new setting. The phenomenon is widely known as the curly braces problem in medical informatics, so named after the curly braces used in Arden Syntax to identify a piece of clinical information that may be stored or structured differently between electronic health record (EHR) systems.[15] As such, studies like ours provide a sensible estimate of how well these prognostic algorithms will perform should they be applied to an urban American population.

Our finding of markedly decreased performance has significant implications, suggesting that these models are **unlikely** to be useful as a major, reliable input for clinical decision-making or for institutional resource allocation planning. Our results should serve as a reminder that predictive models should only be applied in new settings with local validation and that inferences from identified features about prognostic value should be carefully considered.

Ultimately, in clinical settings, users must choose a point on the receiver–operator curve. This point

Table 8 Model performance summary

Study type	Mean validation performance—reported from study	Mean validation performance—NYU data	Mean performance difference between study validation performance and NYU original validation performance	Mean validation performance—NYU retrained	Mean performance difference between study validation performance and NYU retrained validation performance
Applied without Deviation (n=3)	Mean AUROC=0.98 (n=1)	Mean AUROC=0.67 (n=1)	Mean AUROC difference=0.31 (n=1)	Mean AUROC=0.82 (n=3; 0.75–0.93)	Mean AUROC difference=0.21 (n=1)
Applied with deviation (n=4)	Mean AUROC=0.83 (n=3; 0.75–0.88)	Mean AUROC=0.66 (n=4; 0.59–0.74)	Mean AUROC difference=0.19 (n=3; 0.14–0.26)	Mean AUROC=0.71 (n=4; 0.68–0.74)	Mean AUROC difference=0.13 (n=3; 0.07–0.17)
Rebuilt without deviation (n=2)	Mean AUROC=0.73 (n=1)	–	–	Mean AUROC=0.77 (n=2; 0.71–0.80)	Mean AUROC difference=0.01(n=1)
Rebuilt with deviation (n=4)	Mean AUROC=0.78 (n=4; 0.72–0.88)	–	–	Mean AUROC=0.75 (n=4; 0.72–0.79)	Mean AUROC difference=0.03 (n=4; −0.01–0.09)

–=Value unavailable because authors did not provide feature weights when reporting model development.
AUROC, area under the receiver–operator curve; NYU, New York University.

Table 9 Model performance by task type

Outcome type	Mean validation performance—reported from study	Mean validation performance—NYU data	Mean performance difference between study validation performance and NYU original validation performance	Mean validation performance—NYU retrained	Mean performance difference between study validation performance and NYU retrained validation performance
Predicting deterioration (n=4)	Mean AUROC=0.82 (n=4; 0.75–0.88)	Mean AUROC=0.64 (n=3; 0.59–0.74)	Mean AUROC difference=0.18 (n=3; 0.14–0.26)	Mean AUROC=0.72 (n=4; 0.68–0.77)	Mean AUROC difference=0.10 (n=4; 0.01–0.17)
Predicting mortality (n=5)	Mean AUROC=0.93 (n=2; 0.88–0.98)	Mean AUROC=0.72 (n=2; 0.67–0.79)	Mean AUROC difference=0.31 (n=1)	Mean AUROC=0.77 (n=5; 0.74–0.93)	Mean AUROC difference=0.15 (n=2; 0.09–0.21)
Predicting either deterioration or mortality (n=3)	Mean AUROC=0.73 (n=3; 0.72–0.74)	–	–	Mean AUROC=0.72 (n=3; 0.71–0.73)	Mean AUROC difference=0.01 (n=2; −0.01–0.02)

–=Value unavailable because authors did not provide feature weights when reporting model development.
AUROC, area under the receiver–operator curve; NYU, New York University.

generates calculable positive predictive value and sensitivity (table 10). The consideration of potential clinical workflows and integration must be driven by the desired sensitivities and positive predictive values. In general, the values are low except for bolded instances.

Strengths and weaknesses of the study

The primary strength of this study is its dataset, which is made up of over four thousand patients with COVID-19 infection, most of whom presented as infected during the peak of the epidemic in the New York City metropolitan area. As such, this dataset is likely to be reasonably representative in one of the scenarios under which these prognostic algorithms might be used to guide decision-making: a severe epidemic in a major metropolitan centre.

Several caveats should be applied. The most obvious is that like all of the models themselves, we report here on retrospective data, rather than performing a prospective validation, which is the true standard by which predictive models should be judged. It should also be noted that in several cases, we deviated from exact reproductions of previously reported models in order to facilitate their application to our data, which is likely to explain at least some of their decreased performance.

Strengths and weaknesses in relation to other studies

As far as we know, there are no other studies validating multiple COVID-19 prognostic models with which to compare as of the time of writing.

Unanswered questions and future research

Multiple mechanisms may account for performance differences. More analysis would be required in order to elucidate these mechanisms.

Table 10 Positive predictive value of rebuilt models given sensitivity

Study	Average positive predictive value score	50% sensitivity	70% sensitivity	90% sensitivity
Gong *et al*[7]	0.58	0.64	0.60	0.50
Zhou *et al*[8]	0.72	0.73	0.69	0.65
Zou *et al*[9]	0.20	0.21	0.19	0.17
Xie *et al*[5]	0.49	0.45	0.39	0.31
Yan *et al*[6]	**Most recent values: 0.85**	**Most recent values: 0.93**	**Most recent values: 0.84**	**Most recent values: 0.82**
Yan *et al*[6]	Earliest values: 0.38	Earliest values: 0.4	Earliest values: 0.337	Earliest values: 0.31
Levy *et al*[10]	0.45	0.50	0.40	0.31
Zhang *et al*[11]	Task 1: 0.44	Task 1: 0.44	Task 1: 0.39	Task 1: 0.32
Zhang *et al*[11]	Task 2: 0.50	Task 2: 0.53	Task 2: 0.46	Task 2: 0.37
Guo *et al*[12]	**0.90**	**0.92**	**0.88**	**0.86**
Hu *et al*[13]	0.50	0.49	0.40	0.38
Carr *et al*[14]	0.59	0.54	0.47	0.41

First, most obvious are geographical and demographical differences. It is noteworthy that models derived from Chinese data showed the greatest decrement when applied to our data, which was also seen when Zhang *et al* performed an external validation of their model using data from the UK.[12] Differences in access to care, healthcare facility policies and patient demographics between countries may make generalisation difficult for prognostic models derived in one setting to another.

Second are differences in care practices over time. Many of the models we report on here were derived from an earlier phase of the epidemic, which may further have changed the characteristics of the patients in the training sets from which the models were built.[16] Altered clinical practice, trialled therapeutics or shifting demographics over time might endanger the utility of models built towards the beginning of the pandemic.

Third, it is possible the virus itself has changed. Though there is evidence of viral mutation, the clinical effects of which have not been fully characterised.[17] These changes may not be reflected in this validation analysis.

Regarding future research, additional models are being produced, and rigorous validations should be done and encouraged to establish potential clinical use cases.

Twitter Yindalon Aphinyanaphongs @yinformatics

Contributors KHar, BZ and PS performed the literature search and extracted information. BZ and PS collected validation data and assessed models. KHar, BZ, PS and YA drafted the manuscript. KHar, BZ, PS, KHau, MMM, NMA, LIH and YA critically revised the work. MMM, NMA and LIH provided clinical guidance. YA supervised the work and is the guarantor.

Funding YA was partially supported by NIH grant 3UL1TR001445-05 and National Science Foundation award number 1928614.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study met the criteria for quality improvement established by the NYULH IRB and was exempt from institutional review board review.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Due to specific institutional requirements governing privacy protection, data used in this study will not be available.

ORCID iD

Keerthi Harish http://orcid.org/0000-0001-9244-7253

REFERENCES

1 COVID-19 map. Johns Hopkins coronavirus Resour. *Cent* https://coronavirus.jhu.edu/map.html
2 Clinical management of severe acute respiratory infection when COVID-19 is suspected. Available: https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-(ncov)-infection-is-suspected [Accessed 25 Apr 2020].
3 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
4 ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. Available: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection [Accessed 18 Apr 2020].
5 Gong J, Ou J, Qiu X, *et al*. A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. *Clin Infect Dis* 2020;71:833–40.
6 Xie J, Hungerford D, Chen H. *Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19*. Rochester, NY: Social Science Research Network, 2020.
7 Yan L, Zhang H-T, Goncalves J, *et al*. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020;2:283–8.
8 Zhou Y, Yang Z, Guo Y. A new predictor of disease severity in patients with COVID-19 in Wuhan, China. *Respiratory Medicine* 2020.
9 Zou H, Tao J, Yang Q, *et al*. Accurate classification system for patients with COVID-19 based on prognostic nomograms. *SSRN Journal*.
10 Levy T, Richardson S, Coppa K. *Estimating survival of hospitalized COVID-19 patients from admission information*, 2020.
11 Carr E, Bendayan R, Bean D. Supplementing the National early warning score (NEWS2) for Anticipating early deterioration among patients with COVID-19 infection. *medRxiv*2020:2020.04.24.20078006.
12 Zhang H, Shi T, Wu X. *Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK*, 2020.
13 Guo Y, Liu Y, Lu J. Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. *Infectious Diseases* 2020.
14 Hu C, Liu Z, Jiang Y. Early prediction of mortality risk among severe COVID-19 patients using machine learning. *Epidemiology*2020.
15 Hripcsak G, Ludemann P, Pryor TA, *et al*. Rationale for the Arden SYNTAX. *Comput Biomed Res* 1994;27:291–324.
16 COVID-19 hospitalizations. Available: https://gis.cdc.gov/grasp/COVIDNet/COVID19_5.html [Accessed 2 Jul 2020].
17 Korber B, Fischer WM, Gnanakaran S, *et al*. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020;182:812–27.