

RESEARCH

Open Access



# Predicting clinical outcomes among hospitalized COVID-19 patients using both local and published models

William Galanter<sup>1\*</sup>, Jorge Mario Rodríguez-Fernández<sup>2</sup>, Kevin Chow<sup>3</sup>, Samuel Harford<sup>4</sup>, Karl M. Kochendorfer<sup>5</sup>, Maryam Pishgar<sup>4</sup>, Julian Theis<sup>4</sup>, John Zulueta<sup>6</sup> and Houshang Darabi<sup>4</sup>

## Abstract

**Background:** Many models are published which predict outcomes in hospitalized COVID-19 patients. The generalizability of many is unknown. We evaluated the performance of selected models from the literature and our own models to predict outcomes in patients at our institution.

**Methods:** We searched the literature for models predicting outcomes in inpatients with COVID-19. We produced models of mortality or criticality (mortality or ICU admission) in a development cohort. We tested external models which provided sufficient information and our models using a test cohort of our most recent patients. The performance of models was compared using the area under the receiver operator curve (AUC).

**Results:** Our literature review yielded 41 papers. Of those, 8 were found to have sufficient documentation and concordance with features available in our cohort to implement in our test cohort. All models were from Chinese patients. One model predicted criticality and seven mortality. Tested against the test cohort, internal models had an AUC of 0.84 (0.74–0.94) for mortality and 0.83 (0.76–0.90) for criticality. The best external model had an AUC of 0.89 (0.82–0.96) using three variables, another an AUC of 0.84 (0.78–0.91) using ten variables. AUC's ranged from 0.68 to 0.89. On average, models tested were unable to produce predictions in 27% of patients due to missing lab data.

**Conclusion:** Despite differences in pandemic timeline, race, and socio-cultural healthcare context some models derived in China performed well. For healthcare organizations considering implementation of an external model, concordance between the features used in the model and features available in their own patients may be important. Analysis of both local and external models should be done to help decide on what prediction method is used to provide clinical decision support to clinicians treating COVID-19 patients as well as what lab tests should be included in order sets.

**Keywords:** Mortality, Hospitalization, COVID-19, Statistical model, Prediction, Model generalizability

## Background

The coronavirus disease 2019 (COVID-19) caused by the SARS-CoV-2 has been devastating compared to other viruses (seasonal, avian and swine influenza), in regard to both the morbidity and mortality and its economic impact, despite advancements in medical care since the Spanish Flu of 1918 [1]. COVID-19 has had a dramatic impact on health systems globally and the

\*Correspondence: BillG@uic.edu

<sup>1</sup> Departments of Medicine and Pharmacy Systems, Outcomes and Policy, University of Illinois At Chicago (UIC), Chicago, USA

Full list of author information is available at the end of the article



US economy despite assistance from the US Federal government, via the CARES Act [2] and other funding programs.

The COVID-19 pandemic occurred quickly and was rapidly followed by a massive production of academic output, including prediction models for a variety of clinical outcomes; the initial models for hospital outcomes came from the city of Wuhan in the Hubei province of China, where the initial cases were discovered. From there, models around the globe surged and were likely integrated into many hospital guidelines. However, it is unclear if those models could be applied to local cohorts. Having a rapidly available and accurate prediction model for COVID-19 patients being admitted from the emergency department (ED) would be useful for making accurate triage and prognostic assessments to inform decisions regarding treatment and resource allocation. While knowledge of the likelihood of death in those sent home from the ED would also be of interest, this requires longitudinal data which is often not as readily available. The value of appropriate triage decisions is important, especially in time when resources are stretched.

The growth in the volume of readily available health-care data has facilitated the development of artificial intelligence-based models; however, a significant factor limiting the utility of dissemination of such models is the issue of generalizability. For example, the earliest computer-aided decision models evaluating abdominal pain were not able to be replicated in different institutions [3]. A mortality prediction tool in acute alcoholic pancreatitis (Ranson's criteria) [4] developed in a small cohort has a wide acceptance compared to superior scoring tools [5].

One of the most popular predictions tools in clinical use today is the 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk [6]. This risk tool uniformly overestimated risk in non-diabetic patients in a large, multi-ethnic, socioeconomically group of patients in California [7].

We performed an analysis of how well published and self-developed models would predict clinical outcomes after admission on a cohort of diverse urban patients in Chicago. Our self-developed models were trained using data from our local patient cohort. Published, external models were not re-trained with our cohort's data. We aim to close the gap in the understanding if COVID-19 prediction models on mortality and criticality could be potentially used in local cohorts despite ethnic, geographic and timeline differences. We postulate that due to our incomplete understanding of the pathophysiology, ethnic, racial and socioeconomic differences by location, and improving treatment over time, that models may not predict well in a cohort different than their validation and development cohorts.

## Methods

### University of Illinois Hospital (UIH) Cohort

UIH is a tertiary, academic teaching hospital in Chicago. The UIC Institutional Review Board approved this study. All admissions to UIH for COVID-19 positive patients were reviewed for the time of the first COVID-19 positive test and the date of admission. If the first positive COVID-19 test was performed greater than 14 days prior to admission or greater than 48 h after admission, the patient was excluded. Patients transferred from another institution were reviewed for prior COVID-19 testing. If the COVID-19 test was greater than 14 days before transfer, the patient was excluded. If the transfer was not related to any possible COVID-19 symptoms, the patient was excluded. If the patient was discharged and then readmitted less than 14 days after the first positive COVID-19 test, the encounter was included. Patients were discharged or expired prior to 8/18/20. Pregnant patients were included.

Since our goal was to assess the predictive power of our own prediction model as well as some of those in the literature, we partitioned our data into a training cohort consisting of the first 60% of patients admitted prior to 5/9/20 and a test cohort consisting of patients admitted and discharged from 5/9/20 through 8/18/20.

Variable selection was based on a review of the extant literature and expert opinion. The variables selected are shown in Table 1. Admission vital signs, laboratory values and clinical and radiological features were assessed. The results were the first available up to 24 h after admission. Two outcomes were evaluated, mortality (death during hospitalization), and "criticality", defined as mortality or admission to an ICU.

### Literature search

We searched for articles published in PubMed, Embase, Arxiv and medRxiv using the search string: [Prediction] AND [Human] AND [COVID-19] OR [SARS-COV2] AND [Clinical Trial] OR [Observational Trial] which were published before 8/27/2020. Articles were reviewed to determine whether the models described predicted our outcomes of interest and whether there was sufficient concordance and detail provided to implement the model using our cohort's data.

### Model development

The objective of our model development was to accurately predict patient outcomes using a reduced number of key input features. A variety of popular machine learning algorithms were evaluated to classify mortality and criticality. These algorithms include Linear Regression [8], Decision Tree [9], Random Forest [10], XGBoost [11], LightGBM [12], and CatBoost [13]. The training process

**Table 1** Characteristics of the development and test cohorts

| Characteristics                             | Development cohort (N = 309) |        | Test cohort (N = 207) |          | P       |        |
|---|------------------------------|--------|-----------------------|----------|---------|--------|
|   | Missing Data                 |        | Missing Data          |          |         |        |
| <i>Outcome variables(N, (%))</i>            |                              |        |                       |          |         |        |
| Mortality                                   | 38                           | (12.3) | 21                    | (10.1)   | 0.45    |        |
| Criticality                                 | 80                           | (25.9) | 46                    | (22.2)   | 0.34    |        |
| <i>Demographics</i>                         |                              |        |                       |          |         |        |
| Age (Mean, (SD))                            | 56.5                         | (16.0) | 53.3                  | (18.5)   | 0.008*  |        |
| Female (%)                                  | 49.8                         |        | 48.3                  |          | 0.73    |        |
| Race (N (%))                                |                              | 1%     |                       | 1.4%     | 0.22    |        |
| African American                            | 152                          | (49.2) | 86                    | (41.5)   |         |        |
| Hispanic                                    | 37                           | (12)   | 25                    | (12.1)   |         |        |
| Other, Non- Hispanic                        | 94                           | (30.4) | 82                    | (39.6)   |         |        |
| White                                       | 23                           | (7.4)  | 11                    | (5.3)    |         |        |
| <i>Vital signs on admission (mean (SD))</i> |                              |        |                       |          |         |        |
| Systolic blood pressure                     | 135                          | (25)   | 134                   | (24.7)   | 0.80    |        |
| Diastolic blood pressure                    | 78.3                         | (15)   | 77.9                  | (14.3)   | 0.92    |        |
| Hearth rate                                 | 102                          | (21)   | 97.1                  | (20.1)   | 0.70    |        |
| Respiratory rate                            | 23.6                         | (6.9)  | 22.7                  | (6.6)    | 0.52    |        |
| Temperature                                 | 37.5                         | (1.1)  | 37.2                  | (1.0)    | 0.095   |        |
| Oxygen saturation                           | 93.4                         | (7.5)  | 97.7                  | (62.4)   | 0.16    |        |
| <i>Clinical and radiological features</i>   |                              |        |                       |          |         |        |
| BMI, mean (SD)                              | 32.3                         | (10.5) | 32.0                  | (9.6)    | 0.56    |        |
| GCS, mean (SD)                              | 14.9                         | (0.8)  | 14.8                  | (1.2)    | 1%      | 0.28   |
| Dyspnea (N (%))                             | 125                          | (40.5) | 90                    | (43.5)   | 0.49    |        |
| Coma (N (%))                                | 3                            | (1)    | 1                     | (0.5)    | 0.54    |        |
| Pregnant (N (%))                            | 10                           | (3.2)  | 14                    | (6.8)    | 0.062   |        |
| Abnormal chest X-ray (N (%))                | 228                          | (75)   | 136                   | (74)     | 10.6%** | 0.67   |
| <i>Laboratory findings (mean, (SD))</i>     |                              |        |                       |          |         |        |
| White blood cells                           | 6.8                          | (3.1)  | 7.7                   | (3.9)    | 1%      | 0.001* |
| Neutrophils                                 | 5.2                          | (3.6)  | 5.6                   | (3.7)    | 3.4%    | 0.038* |
| Lymphocytes                                 | 1.1                          | (0.7)  | 1.3                   | (1)      | 3.4%    | 0.024* |
| Hemoglobin                                  | 13.0                         | (2.2)  | 12.7                  | (2.3)    | 1%      | 0.57   |
| Hematocrit                                  | 39.4                         | (6.3)  | 37.9                  | (6.8)    | 1%      | 0.36   |
| RDW   | 14.9                         | (2)    | 15.1                  | (2.2)    | 1%      | 0.68   |
| Platelets                                   | 215                          | (91)   | 236                   | (105)    | 1%      | 0.15   |
| Creatinine                                  | 1.8                          | (3)    | 1.9                   | (2.6)    | 2.4%    | 0.47   |
| Lactic acid                                 | 1.6                          | (1.6)  | 1.8                   | (1.9)    | 30.4%   | 0.11   |
| Lactate dehydrogenase                       | 353                          | (227)  | 386                   | (521)    | 27.5%** | 0.14   |
| Pro-calcitonin                              | 1.4                          | (6.7)  | 2.2                   | (10.4)   | 32.9%** | 0.12   |
| Troponin I                                  | 0.11                         | (0.8)  | 0.04                  | (0.1)    | 31.4%   | 0.076  |
| B-type natriuretic peptide                  | 527                          | (1939) | 369                   | (664)    | 63.8%   | 0.10   |
| Albumin                                     | 3.7                          | (0.5)  | 3.6                   | (0.6)    | 7.2%    | 0.009* |
| ALT   | 38.8                         | (44.4) | 37.8                  | (49.9)   | 7.2%    | 0.88   |
| AST   | 48.2                         | (58.3) | 52.3                  | (81.3)   | 7.2%    | 0.20   |
| Total bilirubin                             | 0.7                          | (0.9)  | 0.8                   | (0.7)    | 7.2%    | 0.76   |
| Direct bilirubin                            | 0.2                          | (0.5)  | 0.2                   | (0.3)    | 7.2%    | 0.57   |
| Creatine kinase                             | 281                          | (471)  | 3071 <sup>+</sup>     | (22,782) | 67.6%   | 0.012* |
| C-reactive protein                          | 101                          | (84)   | 98.6                  | (86.7)   | 19.8%   | 0.46   |
| Interleukin 6                               | 24.9                         | (33.9) | 28.1                  | (40.2)   | 87.9%** | 0.17   |
| D-dimer                                     | 1.9                          | (2.6)  | 2.2                   | (3)      | 27.5%** | 0.46   |

**Table 1** (continued)

| Characteristics                                  | Development cohort (N = 309) |        |      | Test cohort (N = 207) |        |         | P      |
|--|------------------------------|--------|------|-----------------------|--------|---------|--------|
|  | Missing Data                 |        |      | Missing Data          |        |         |        |
| Ferritin   | 884                          | (1562) | 7.1% | 930                   | (1360) | 17.4%** | 0.45   |
| <i>Medical condition</i>                         |                              |        |      |                       |        |         |        |
| Hypertension                                     | 178                          | (57.6) |      | 111                   | (53.6) |         | 0.37   |
| Heart disease                                    | 94                           | (30)   |      | 56                    | (27)   |         | 0.41   |
| Stroke   | 23                           | (7.4)  |      | 11                    | (5.3)  |         | 0.34   |
| Diabetes   | 161                          | (52.1) |      | 96                    | (46)   |         | 0.20   |
| Asthma   | 65                           | (21)   |      | 45                    | (22)   |         | 0.85   |
| COPD   | 24                           | (7.8)  |      | 13                    | (6.3)  |         | 0.52   |
| Chronic kidney disease                           | 52                           | (17)   |      | 38                    | (18)   |         | 0.65   |
| End-stage renal disease                          | 30                           | (9.7)  |      | 24                    | (12)   |         | 0.49   |
| Cancer   | 33                           | (11)   |      | 23                    | (11)   |         | 0.88   |
| Transplant                                       | 1                            | (0.3)  |      | 1                     | (0.5)  |         | 0.78   |
| Human immunodeficiency virus                     | 7                            | (2.3)  |      | 1                     | (0.5)  |         | 0.11   |
| Immunosuppression                                | 2                            | (0.6)  |      | 9                     | (4.3)  |         | 0.004* |
| Sickle cell disease                              | 6                            | (2)    |      | 6                     | (2.9)  |         | 0.48   |
| Nicotine use                                     | 34                           | (11)   |      | 32                    | (16)   |         | 0.14   |
| Alcohol use                                      | 59                           | (19)   |      | 45                    | (22)   |         | 0.46   |
| Substance use                                    | 16                           | (5)    |      | 11                    | (5.3)  |         | 0.95   |
| <i>Variables needed for external models only</i> |                              |        |      |                       |        |         |        |
| Blood urea nitrogen, mean (SD)                   |                              |        |      | 23.5                  | (20.7) | 1.9%    |        |
| eGFR, mean (SD)                                  |                              |        |      | 70.9                  | (40.6) | 2.4%    |        |
| Partial thromboplastin time, mean (SD)           |                              |        |      | 34.4                  | (6.5)  | 42.5%   |        |

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; GCS, Glasgow coma scale; RDW, red blood cell distribution width; SD, standard deviation

\* Continuous variables were compared using a t-test and categorical variables, including missingness, were compared using a Chi-square test

\*\*  $P < 0.05$  using a chi-square test, development versus test cohort

Significance was set at 0.05

† A single very high, but clinically consistent creatine kinase accounted for the very large mean in this group

uses a combination of step forward feature selection and parametric grid search. Step forward feature selection is the process of starting with a single feature and iteratively adding one additional feature until there is no increase in model performance. For each step in the feature selection, a parametric grid search is performed to determine the optimal parameter set for each model. We use the area under the receiver operating characteristic curve (AUC) as the evaluation metric.

### Statistical analysis of models

No missing data were imputed in our test cohort. External models were included in our analyses if predictions could be generated for greater than 60% of the patients based on this missingness. If odds or a point scale was available, a receiver operator curve was developed and the area under the curve (AUC) was calculated.

Confidence interval and comparison of ROCs were performed using DeLong's method [14]. The training and

test cohorts were compared using Chi-Square tests for categorical variables and two-sided t-tests for continuous variables using a significance level of  $P < 0.05$ . The fraction of missingness for each variable was compared between the cohorts using the Bonferroni correction to control the family-wise error rate.

Descriptive statistics were performed using Stata 12 SE version (StataCorp, TX). Model development was conducted using the Python libraries sklearn (v.0.23.1), TensorFlow (v.2.2.0), XGBoost (v0.90), LightGBM (v.2.3.1) and Catboost (v.0.23.1). Statistical analysis was performed using the pROC package in R. This study was approved by the UIC Institutional Review Board.

## Results

### UIH cohort characteristics model compilation

A description of the UIH cohorts is shown in Table 1. There was a total of 516 patients. The training cohort included the first 309 patients (60%), and the test cohort

was the subsequent 207 patients (40%). The test cohort was slightly younger, 53.3 vs 56.5 years [ $P=0.008$ ]. Though the whole racial distribution was not significantly different between the cohorts, the proportion of self-declared black patients was 49% in the training cohort and 42% in the test cohort. The lymphocyte, white blood cell and neutrophil counts were significantly higher in the test cohort.

Though some lab tests were performed on almost all patients, many tests were performed in a more discretionary fashion. The missingness of some of the more discretionary tests was higher in the test cohort than in the training cohort: ferritin 7.1–17.4%, Lactate Dehydrogenase (LDH) 16.5–27.5%, Procalcitonin 16.8–32.9%, Interleukin 6 (IL-6) 75.1–87.9%. D-dimer was missing less frequently in the test cohort, 40.8–27.5%.

### Model compilation summary

Ninety-one abstracts were reviewed. After applying our inclusion criteria, 41 articles remained. The models and references are shown in Table 2.

Over 60% of the models ( $n=26$ ) were derived in China, 11 in Europe, 3 in the US and 2 were multinational. The most common methods were logistic regression ( $n=25$ ) and Cox Regression ( $n=12$ ). A small number of models used neural networks and decision trees. Among models which published an AUC, the AUC's ranged from 0.74 to 0.98.

### UI health internal model development

Multiple methods of machine learning were assessed to develop the best prediction model of the training (60%) cohort. The best models for both mortality and criticality were random forest models, based on the AUC values. Table 3 lists the key modeling parameters and covariates for the mortality and criticality models. The covariates are listed in the order of importance generated by the step forward regression. The key parameters for the random forest models were determined during the grid search of the development data set. The AUC for the mortality model in the training cohort was 0.98, and for criticality it was 0.97.

If model coefficients in the papers in Table 2 were sufficiently described and the model variables were available for more than 60% of admissions, the model was used to predict outcomes in the UIH test cohort. Results are shown in Table 4.

A total of 10 models were assessed using the test cohort, 8 from the literature and 2 internal. Seven of the external models used logistic regression and one used a decision tree. One external model predicted criticality; the remainder predicted mortality. The most common variables used in the models were the age (7 models),

lymphocyte count or lymphocytes/WBC ratio (6 models), C-reactive protein (CRP) and LDH (4 models), D-dimer (3 models) and BUN (2 models). The number of features used in each model ranged from 2 to 11, with a median of 3.5. These models assessed clinical features and laboratory testing upon admission. In addition, 1 model explicitly included pregnant patients [19], 2 excluded pregnant patients [28, 42], and 5 were undetermined [20, 26, 44, 46, 51].

Three of the models, B, G and H [19, 46, 51], had open access web-based calculators to predict outcomes for individual patients. One model used a decision tree of only three variables which is easy for a clinician to use (A) [42]. Two models used a nomogram to try to simplify use (D and F) [26, 44].

All external models were trained using cohorts of Chinese patients. Though there were non-Chinese cohort models in Table 2, none of them provided sufficient description of their models to be implemented on our test cohort without retraining.

Common reasons why models were not used were the lack of availability of the coefficients needed to calculate a prediction score, lack of concordance between the features used in the model and features available in our test cohort, and outcome data not available in our test cohort (e.g., mortality).

Figure 1 shows the confidence intervals of the AUC's obtained on the test cohort. Table 4 and Fig. 1 show that the best estimate for the AUC ranges from 0.68 for model G to 0.89 for model C. The internal models have an AUC of 0.84 for mortality and 0.83 for criticality. The mortality model with the highest AUC, C, was not statistically different than the UIH mortality model, 0.89 (0.82–0.96) vs AUC 0.84 (0.74–0.94), [ $P>0.5$ ].

The confidence intervals range from 0.13 to 0.30. The difference in performance between the published fit and that of its performance on our test set varied significantly. For model B this difference in AUC was only 0.04 and for model E it was 0.26. The UI Health models were in the middle with a 0.14 AUC difference.

For all 8 models, the mean values for lab results and those of the UI Health test cohort are shown in Table 5. The variables shown were used in at least one model and were available in five or more of the model cohorts. Age and CRP were reported in all papers. The creatinine was reported in seven papers. Though rigorous statistical testing cannot be performed due to the inability to obtain the raw data, some of the variables are clinically significantly different between the cohorts from China and UIH. The mean CRP at UIH is more than three times higher than in the external model average, the creatinine is two-fold higher and the LDH is roughly 1/3 higher.

**Table 2** Prediction models of inpatients matching outcomes of interest

| # | Title  | Location               | Population timepoint/<br>inclusion criteria          | N      | Outcome                               | Method <sup>a</sup>          | Features  | Performance <sup>b</sup> |
|---|--|------------------------|--|--------|---------------------------------------|------------------------------|---|--------------------------|
| 1 | ADL-dependency, D-Dimers, LDH and absence of anticoagulation are independently associated with one-month mortality in older inpatients with Covid-19 [15]                          | France                 | Inpatients, >65-year-old                             | 108    | Mortality                             | Cox regression               | ADL-dependency, D-Dimer, LDH, Anticoagulation   | AUC 0.83                 |
| 2 | Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in COVID-19 patients: a meta-regression and decision tree analysis [16] | Brazil, China, UK, USA | Inpatients   | 17,364 | Mortality, ICU, mixed                 | Meta-analysis, decision tree | Age, Troponin I, AST  | Precision 74% recall 86% |
| 3 | Prediction for progression risk in patients with COVID-19 pneumonia: the CALL Score [17]   | China                  | Inpatients, excluding non-COVID-19 primary infection | 248    | Deterioration or worsening of CT lung | Cox regression, nomogram     | Age, ALC, comorbidities, LDH  | AUC 0.91                 |
| 4 | Clinical and laboratory predictors of in-hospital mortality in patients with COVID-19: a cohort study in Wuhan, China [18]   | China                  | Inpatients excluding pregnancy & multorgan failure   | 296    | Mortality                             | Logistic regression          | Age, ALC, SAT, ANC, CRP, D-Dimer, AST, eGFR   | AUC 0.88                 |
| 5 | Development and validation of a clinical risk score to predict the occurrence of critical illness of hospitalized patients with COVID-19 [19]                                      | China                  | Inpatients   | 710    | Criticality                           | Logistic regression          | Age, CXR, Cancer, LDH, Hemoptysis, Dyspnea, Unconsciousness, NLR, comorbidities, Direct bilirubin | AUC 0.88                 |
| 6 | Development and validation of prognosis model of mortality risk in patients with COVID-19 [20]   | China                  | Inpatients   | 305    | Mortality                             | Logistic regression          | Age, CRP, LDH   | AUC 0.95                 |
| 7 | Laboratory predictors of death from coronavirus disease 2019 (COVID-19) in the area of Valcamonica, Italy [21]   | Italy                  | Inpatients   | 144    | Mortality                             | Logistic regression          | Age, LDH, CRP, WBC, ANC, ALC, albumin, PTT  | 80% of variance          |

**Table 2** (continued)

| #  | Title  | Location               | Population timepoint/<br>inclusion criteria | N     | Outcome          | Method <sup>a</sup> | Features   | Performance <sup>b</sup> |
|----|--|------------------------|---|-------|------------------|---------------------|--|--------------------------|
| 8  | Prediction model and risk scores of ICU admission and mortality in COVID-19 [22]   | USA, New York          | Inpatients                                  | 641   | Mortality        | Logistic regression | LDH, PC, smoking, SAT, ALC   | AUC 0.82                 |
| 9  | Risk factors associated with clinical outcomes in 323 COVID-19 hospitalized patients in Wuhan, China [23]  | China                  | Inpatients                                  | 323   | Mortality        | Logistic regression | Age, smoking, ALC, ANC, critical disease, DM, Troponin I   | Not Reported             |
| 10 | Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China [24]  | China                  | Inpatients                                  | 1,590 | Mortality        | Cox regression      | Age, CHD, dyspnea, CVA, PC, AST  | AUC 0.91                 |
| 11 | Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multi-center study [25] | China, Italy & Belgium | Inpatients, without other severe illness    | 725   | "Severe disease" | Logistic regression | Age, ALC/WBC, CRP, LDH, CK, urea, calcium  | AUC 0.88                 |
| 12 | Diagnostic performance of initial blood urea nitrogen combined with D-dimer levels for predicting in-hospital mortality in COVID-19 patients [26]                              | China                  | Inpatients                                  | 305   | Mortality        | Cox regression      | D-Dimer, BUN   | AUC 0.94                 |
| 13 | IL-6-based mortality risk model for hospitalized patients with COVID-19 [27]   | Spain                  | Inpatients                                  | 501   | Mortality        | Logistic regression | ALT, IL6, CRP, LDH, ferritin, ANC, NLR, ALC, albumin, platelets, monocytes, SAT/FiO <sub>2</sub> ratio | AUC: 0.87                |
| 14 | Laboratory findings and a combined multifactorial approach to predict death in critically ill patients with COVID-19: a retrospective study [28]                               | China                  | Inpatients, severe                          | 336   | Mortality        | Logistic regression | D-Dimer, ALC/WBC, BUN  | AUC 0.99                 |

**Table 2** (continued)

| #  | Title  | Location      | Population timepoint/<br>inclusion criteria | N    | Outcome     | Method <sup>a</sup>       | Features  | Performance <sup>b</sup> |
|----|--|---------------|---|------|-------------|---------------------------|---|--------------------------|
| 15 | Predictive values of blood Urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients [29]    | Turkey        | Inpatients                                  | 139  | Mortality   | Cox hazard regression     | BUN/Cr, NLR   | AUC 0.95                 |
| 16 | Prognostic modelling of COVID-19 using artificial intelligence in a UK population [30]   | UK            | Inpatients                                  | 398  | Mortality   | Artificial neural network | Age, altered mentation, HTN, CLD, collapse, Sex, cough, fever, CKD, DM, CHD, CVA, myalgia, smoking, symptom onset, BMI, diarrhea, vomiting, anosmia, ageusia, cirrhosis, abdominal pain | AUC 0.90                 |
| 17 | Redefining cardiac biomarkers in predicting mortality of inpatients with COVID-19 [31]   | China         | Inpatients                                  | 3219 | Mortality   | Mixed effects cox model   | Troponin I, CK-MB, BNP, CK, Myoglobin   | AUC 0.83                 |
| 18 | Risk factors for severe illness in hospitalized Covid-19 patients at a regional hospital [32]  | USA, Maryland | Inpatients                                  | 117  | Criticality | Logistic regression       | Oxygen requirement, Sputum production, DM, CKD  | AUC 0.88                 |
| 19 | Scoring systems for predicting mortality for severe patients with COVID-19 [33]  | China         | Inpatients                                  | 452  | Mortality   | Lasso/regression          | Age, CHD, D-Dimer, PC, ALC  | AUC 0.94                 |
| 20 | Simple nomogram based on initial laboratory data for predicting the probability of ICU transfer of COVID-19 patients: Multicenter retrospective study [34] | China         | Inpatients                                  | 461  | ICU         | Cox regression            | Age, HTN, ANC, PC, PT, D-Dimer, ALC, albumin  | AUC 0.85                 |
| 21 | Clinical characteristics, associated factors, and predicting COVID-19 mortality risk: a retrospective study in Wuhan, China [35]                           | China         | Inpatients                                  | 1633 | Mortality   | Logistic regression       | Age, Sex, DM, ALC, PC   | AUC 0.76                 |

**Table 2** (continued)

| #  | Title   | Location       | Population timepoint/<br>inclusion criteria | N    | Outcome           | Method <sup>a</sup> | Features   | Performance <sup>b</sup> |
|----|---|----------------|---|------|-------------------|---------------------|--|--------------------------|
| 22 | Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19 [36]                               | China          | Inpatients                                  | 336  | Mortality         | Cox regression      | Age, GSH, CD3 ratio, total protein   | AUC 0.98                 |
| 23 | Identification and validation of a novel clinical signature to predict the prognosis in confirmed COVID-19 patients [37]                      | China          | Inpatients                                  | 270  | Mortality         | Cox regression      | Age, CRP, ALC, ANC, PC   | AUC 0.95                 |
| 24 | Myocardial injury determination improves risk stratification and predicts mortality in COVID-19 patients [38]                                 | Spain          | Inpatients excluding cardiac primary        | 707  | Mortality         | Cox regression      | Age, sex, CRP, myocardial injury, HTN, RAAS inhibitor, hematocrit, Cr, D-Dimer, CCI  | AUC 0.79                 |
| 25 | Neutrophil-to-lymphocyte ratio and outcomes in Louisiana Covid-19 patients [39]   | USA, Louisiana | Inpatients                                  | 125  | Mortality         | Cox Regression      | NLR (day 2), NLR (day 5)   | AUC 0.78                 |
| 26 | Prediction of the severity of Corona Virus Disease 2019 and its adverse clinical outcomes [40]  | China          | Inpatients                                  | 88   | Mortality         | Logistic regression | Age, ALC, IL 6   | AUC 0.97                 |
| 27 | A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: An observational cohort study [41] | UK             | Inpatients                                  | 1157 | Mortality and ICU | Lasso/regression    | Age, sex, Cr, CKD, CLD, Ethnicity, "index of multiple deprivation", O2 requirement, SAT, respiratory rate, CXR, ALC, ANC, CRP, albumin, cancer, DM, HTN, CHD | AUC 0.76                 |
| 28 | An interpretable mortality prediction model for COVID-19 patients [42]  | China          | Inpatients                                  | 375  | Mortality         | Decision tree       | LDH, CRP, ALC/WBC  | AUC 0.99                 |
| 29 | Clinical prediction model for mortality of adult Diabetes inpatients with COVID-19 in Wuhan, China: a retrospective pilot study [43]          | China          | Inpatients with diabetes                    | 78   | Mortality         | Logistic regression | PTT, BUN, WBC, LDH   | AUC 0.84                 |

**Table 2** (continued)

| #  | Title   | Location | Population timepoint/<br>inclusion criteria  | N   | Outcome                      | Method <sup>a</sup> | Features  | Performance <sup>b</sup> |
|----|---|----------|--|-----|------------------------------|---------------------|---|--------------------------|
| 30 | Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19 [44]   | China    | Inpatients                                   | 299 | Mortality                    | Logistic regression | Age, LDH, ALC, SAT  | AUC 0.98                 |
| 31 | Development and validation of a risk factor-based system to predict short-term survival in adult hospitalized patients with COVID-19: a multicenter, retrospective, cohort study [45] | China    | Inpatients admitted w/o other severe illness | 828 | Mortality                    | Cox regression      | Age, LDH, NLR, Direct bilirubin                           | AUC 0.88                 |
| 32 | Early prediction of mortality risk among severe COVID-19 patients using machine learning [46]   | China    | Inpatients                                   | 183 | Mortality                    | Logistic regression | Age, CRP, D Dimer, ALC                                    | AUC 0.88                 |
| 33 | Estimation of risk factors for COVID-19 mortality—preliminary results [47]  | China    | Unclear                                      | N/A | Mortality                    | Logistic regression | Age, CHD, CLD, Sex  | Not reported             |
| 34 | Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study [48]  | UK       | Inpatients                                   | 439 | Criticality                  | Lasso/regression    | Age, BUN, SAT, CRP, eGFR, ANC, NLR, NEWS2, O2 requirement | AUC 0.74                 |
| 35 | Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan [49]   | China    | Inpatients                                   | 487 | Mortality and "severe cases" | Logistic regression | Age, sex, HTN   | Not reported             |

**Table 2** (continued)

| #  | Title  | Location    | Population timepoint/<br>inclusion criteria | N   | Outcome             | Method <sup>a</sup>       | Features   | Performance <sup>b</sup> |
|----|--|-------------|---|-----|---------------------|---------------------------|--|--------------------------|
| 36 | Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis [50]            | China       | Inpatients                                  | 125 | "Severe" pneumonia  | Logistic regression       | Underlying disease, respiratory rate, CRP, LDH                                     | AUC 0.98                 |
| 37 | Risk prediction for poor outcome and death in hospital in-patients with COVID19: derivation in Wuhan, China and external validation in London, UK [51]   | China       | Inpatients                                  | 775 | Mortality           | Logistic regression       | Age, sex, ALC, ANC, platelets, CRP, Cr   | AUC 0.91                 |
| 38 | Predicting severe COVID-19 at presentation, introducing the COVID Severity Score [52]  | Netherlands | Inpatients                                  | 261 | Respiratory failure | Logistic regression       | Age, CRP, ALC, NLR, BUN, LDH, RDW, SAT   | AUC 0.79                 |
| 39 | Comorbidity and prognostic factors on admission of a covid-19 cohort: in a general hospital [53]   | Spain       | Inpatients                                  | 96  | Mortality           | Regression                | Age, LDH, Cardiomyopathy   | Not reported             |
| 40 | Epidemiology, risk factors and clinical course of SARS-CoV-2 infected patients in a Swiss university hospital: an observational retrospective study [54] | Switzerland | Inpatients                                  | 200 | Ventilation         | Regression                | Sex, qSOFA score, CXR, CRP   | Not reported             |
| 41 | Comparison of deep learning with regression analysis in creating predictive models for SARS-CoV-2 outcomes [55]  | UK          | Inpatients                                  | 398 | Mortality           | Artificial neural network | Confusion, collapse, dyspnea, cough, CKD, heart failure, CVA, fever, sex, CHD, HTN | AUC 0.93                 |

ADL, activities of daily living; ALC, absolute lymphocyte count; ANC, absolute neutrophil count; ALT, alanine aminotransferase; AST, aspartate transaminase; BUN, blood urea nitrogen; CCI, Charlson comorbidity index; CD, cluster of differentiation; CHD, coronary heart disease; CK, creatinine kinase; CK-MB, creatinine kinase-MB; CKD, chronic kidney disease; Cr, creatinine; CRP, C reactive peptide; CLD, chronic lung disease; CXR, chest X-ray; CVA, cerebrovascular accident; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; FIO2, fraction of inspired oxygen; GSH, glutathione reductase; HTN, hypertension; IL6, interleukin 6; LDH, lactate dehydrogenase; NEWS2, national early warning score 2; [48] NLR, neutrophil to lymphocyte ratio; O2, oxygen; PC, procalcitonin; PTT, partial thromboplastin time; qSOFA, quick sequential organ failure assessment; RAAS, renin-angiotensin-aldosterone system; RDW, red blood cell distribution width; SAT, oxygen saturation; WBC, white blood cell count

<sup>a</sup> Many studies used multiple methods to help select variables or as trials. The method listed corresponds to the method used to produce the final performance listed

<sup>b</sup> When available the area under the curve (AUC) or c-statistic is shown, if multiple, best is shown. Although other measures may have been performed, they are not shown

<sup>c</sup> "Criticality" is mortality or ICU stay or ventilation

**Table 3** Internal Model Fit on first 60% of admissions for mortality and criticality

| Model       | Method        | Key parameter  | Covariates  |
|-------------|---------------|--|---|
| Mortality   | Random forest | Number of estimators: 100<br>Max depth: 5<br>Minimum sample Split: 3 | Age, diastolic pressure, O2 Sat, BMI, AST, creatinine, CRP, ferritin, platelet, RDW, WBC          |
| Criticality | Random forest | Number of estimators: 100<br>Max depth: 5<br>Minimum sample Split: 2 | Age, O2 Sat, ALT, AST, creatinine, CRP, ferritin, platelet, RDW, WBC, neutrophil/lymphocyte ratio |

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; CRP, C-reactive protein; O2 Sat, oxygen saturation; RDW, red blood cell distribution width; WBC, white blood cell count

## Discussion

All the models in Table 1 could not be used to make predictions on our test cohort for multiple reasons. Without chart review, symptomology and its duration are difficult to obtain, excluding some models. Unusual imaging grading schemes or mandatory CT scans were not available in our cohort. Some studies used labs that were not ordered frequently in our hospital. Lack of longitudinal follow up limited the use of timed mortality, i.e., 30-day etc. These issues, along with the lack of well described coefficients of models produced the inability to use models except for the 8 models in Table 1.

The features used in the models were surprisingly diverse. The number of variables in each model ranged from 2 through 11, with 19 different variables across the studies. The most common variables used were age, lymphocyte count, CRP and LDH. It is surprising that only 7 of the 10 models used age as a predicting variable, and the 3 models that did not use it did not perform well. In large multi-site cohorts examined in Britain [56], the US [57] and internationally [58], age was a strong predictor of mortality.

Three of the external models performed very well, with AUC's of 0.84–0.89. This demonstrates that although the patients were geographically distant, ethnically different, in different health systems and cultures, and at different times during the pandemic, reasonable prediction was possible. Our initial hypothesis was that these models would not work well, but this was not the case in all the models.

It is likely that some of the models may have had better performance if retrained using our local cohort, but this was not done as the purpose was to see how they worked “out of the box”. This appeared to be the intent of many of the authors of the published models as evidenced by the publishing of web calculators, nomograms and decision trees. One of the issues which may cause worse or better performance in a model is that the outcomes have been found to be a function of time during the pandemic,

not just patient factors, with improving outcomes more recently [59, 60].

Models A, F, G and H were also evaluated in a review and cohort prediction comparison by Gupta et al. [61] using their cohort of 440 patients from London with a mortality rate of around 28%. For Models F, G and H, the AUCs in our cohort were slightly different than in the London Cohort [61] respectively, model F, 0.84 vs. 0.76, model G, 0.68 vs. 0.74 and model H, 0.72 vs. 0.69.

Review of the characteristics of the cohorts in Table 5 is instructive in understanding why some of the models did not perform well. Model A is a decision tree based on only 3 features, CRP, LDH and the percentage of lymphocytes. The first decision node suggests mortality if the LDH is greater than 365 U/L. In their cohort, the average LDH was 274 U/L. The average LDH in our test cohort was 386 however, thus a large portion were predicted to die at the first node, causing a poor positive predictive value (PPV). In the London cohort the average LDH was about the same as ours, 395 U/L, and this model performed poorly in that cohort also [61].

The average LDH was roughly 1/3 higher in our test cohort than in the average of the cohorts from China. It is not clear what the reason for this is. In a healthy multi-ethnic cohort from Hawaii [62], there were at most minor differences between black, Hispanic, White and Asian patients in their LDH, suggesting that the differences in LDH are not likely due to racial factors. It is possible that a difference in the time of infection to presentation might explain the difference. The other models which used LDH predicted well, but this might be in part related to use of a logistic regression instead of a decision tree.

The average CRP in our cohort is roughly 350% of the average in the external models, 99 mg/L vs. 27 mg/L. Four models used the CRP and only one model performed well, model C. The creatinine was significantly higher in our cohort than in any of the derivation cohorts and as well as the average of the studies, 0.84 mg/dL. Only one model used the creatinine, model H. Its derivation cohort average creatinine was 0.72 mg/dL. Thus, model

**Table 4** Prediction models fit to UH test cohort

| Study   | Cohort origin | N (develop) | Outcome       | Method                   | Covariates   | Performance            | Test cohort # evaluable | Test cohort performance |
|---|---------------|-------------|---------------|--------------------------|--|------------------------|-------------------------|-------------------------|
| A) An interpretable mortality prediction model for COVID-19 patients [42]   | Wuhan, China  | 351         | Mortality*    | Decision tree            | (3) CRP, LDH, lymphocyte percentage  | PPV 96.9%<br>NPV 98.4% | 145 (70%)*              | AUC 0.69 (0.60–0.79)    |
| B) Development and validation of a clinical risk score to predict the occurrence of critical illness of hospitalized patients with COVID-19 [19] <sup>a</sup> | China         | 1590        | Criticality** | Logistic regression (LR) | (10) Age, cancer, direct bilirubin, comorbidity, dyspnea, hemoptysis, LDH, neutrophils/lymphocytes, unconscious, CXR | AUC 0.88 (0.84–0.93)   | 144 (70%)               | AUC 0.84 (0.78–0.91)    |
| C) Development and validation of prognosis model of mortality risk in patients with COVID-19 [20]   | Wuhan, China  | 292         | Mortality     | LR                       | (3) Age, LDH, CRP  | AUC 0.95               | 141 (69%)               | AUC 0.89 (0.82–0.96)    |
| D) Diagnostic performance of initial blood urea nitrogen combined with D-dimer levels for predicting in-hospital mortality in COVID-19 patients [26]          | Wuhan, China  | 305         | Mortality     | LR                       | (2) BUN, D-dimer   | AUC 0.94 (0.90–0.97)   | 150 (72%)               | AUC 0.73 (0.62–0.83)    |
| E) Laboratory findings and a combined multifactorial approach to predict death in critically ill patients with COVID-19: a retrospective study [28]           | Wuhan, China  | 336         | Mortality     | LR                       | (3) BUN, D-dimer, lymphocyte percentage  | AUC 0.99 (0.98–1.0)    | 148 (71%)*              | AUC 0.72 (0.61–0.82)    |
| F) Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19 [44]                          | Wuhan, China  | 299         | Mortality     | LR                       | (4) Age, LDH, lymphocyte count, O2 Saturation  | AUC 0.98 (0.96–1.0)    | 150 (72%)               | AUC 0.84 (0.73–0.94)    |
| G) Early prediction of mortality risk among severe COVID-19 patients using machine learning [46] <sup>b</sup>   | Wuhan, China  | 183         | Mortality     | LR                       | (4) Age, CRP, D-dimer, lymphocyte count  | AUC 0.90               | 142 (69%)               | AUC 0.68 (0.51–0.81)    |

**Table 4** (continued)

| Study  | Cohort origin | N (develop) | Outcome     | Method        | Covariates  | Performance         | Test cohort # evaluable | Test cohort performance |
|--|---------------|-------------|-------------|---------------|---|---------------------|-------------------------|-------------------------|
| HJ Risk prediction for poor outcome and death in hospital in-patients with COVID19: derivation in Wuhan, China and external validation in London, UK [(51)] <sup>c</sup> | Wuhan, China  | 775         | Mortality   | LR            | (7) Age, CRP, sex, Cr, lymphocytes, neutrophils, platelets count  | AUC 0.91            | 165 (80%)               | AUC 0.72 (0.58–0.86)    |
| UJH mortality model  | Chicago, USA  | 309         | Mortality   | Random forest | (11) Age, AST, BMI, Cr, CRP, diastolic BP, ferritin, O2 saturation, platelet count, RDW, WBC            | AUC 0.98 (0.96–1.0) | 152 (73%)               | AUC 0.84 (0.74–0.94)    |
| UJH criticality model  | Chicago, USA  | 309         | Criticality | Random forest | (11) Age, ALT, AST, Cr, CRP, ferritin, RDW, neutrophils/lymphocytes, O2 saturation, platelet count, WBC | AUC 0.97 (0.94–1.0) | 152 (73%)               | AUC 0.83 (0.76–0.90)    |

ALT, alanine aminotransferase; AST, aspartate transaminase; BMI, body mass index; BP, blood pressure; BUN, blood urea nitrogen; Cr, creatinine; CRP, C reactive peptide; CXR, chest X-ray; LDH, lactate dehydrogenase; O2 oxygen; PC, procalcitonin; RDW, red blood cell distribution width; SAT, oxygen saturation; WBC, white blood cell count

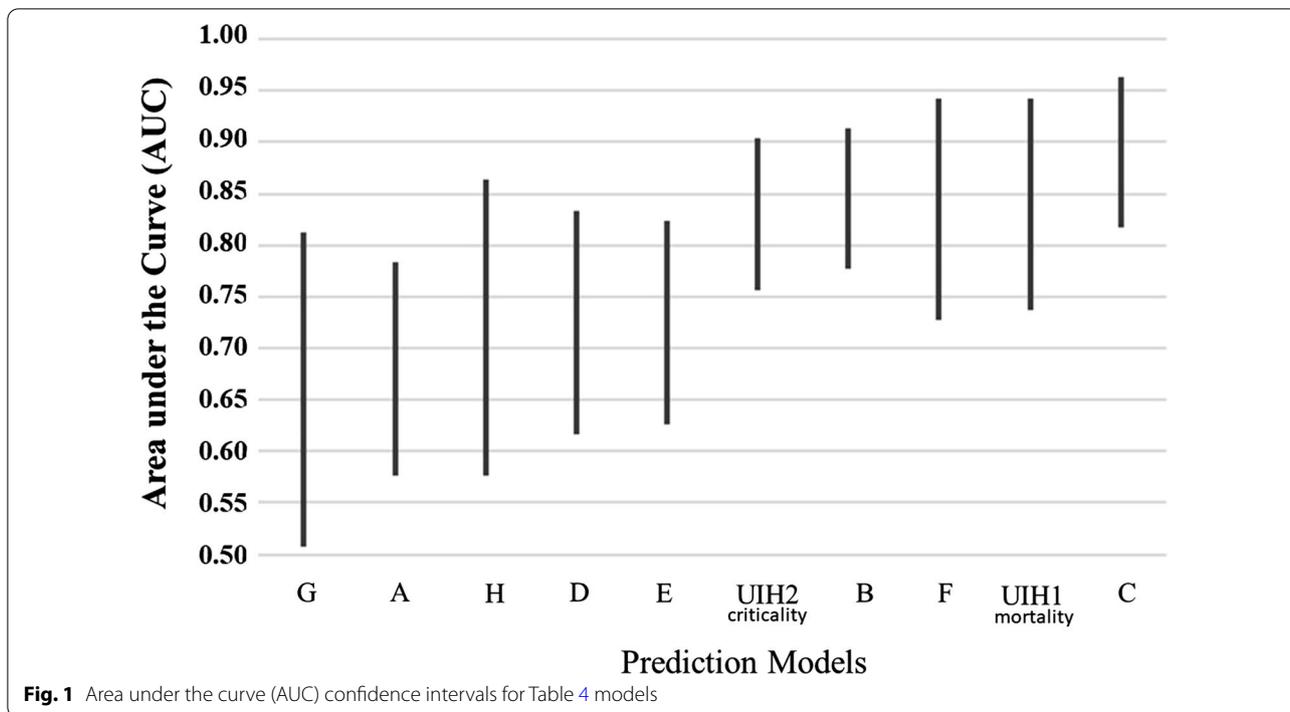
\* Mortality defined as death prior to discharge\*\*Criticality defined as mortality or intensive care unit stay

\*\*\* For these models only non-pregnant patients were used. The other models either included or did not specify inclusion of pregnant patients

<sup>a</sup> <https://118.126.104.170/>

<sup>b</sup> [https://phenomics.fudan.edu.cn/risk\\_scores/](https://phenomics.fudan.edu.cn/risk_scores/)

<sup>c</sup> <https://covid.datahelps.life/prediction/>



**Table 5** Values of the most common variables in the 8 external models and the test cohort

| Characteristics     | Age (N=8)                 | CRP (mg/L) (N=8)          | Cr (mg/dl) (N=7)          | LDH (U/L) (N=6)   | Lymph (1000/uL) (N=6) | Lymph/WBC (N=6)        | Neut (1000/uL) (N=4) |
|---------------------|---------------------------|---------------------------|---------------------------|-------------------|-----------------------|------------------------|----------------------|
| <i>Model cohort</i> |                           |                           |                           |                   |                       |                        |                      |
| A                   | 58.8                      | <b><u>26.3*</u></b>       | N/A                       | <b><u>274</u></b> | <i>N/A</i>            | <b><u>0.14</u></b>     | <i>N/A</i>           |
| B                   | <b><u>48.9</u></b>        | 34.8                      | 0.86                      | <b><u>314</u></b> | <b><u>1.4</u></b>     | <i>N/A</i>             | <b><u>4.1</u></b>    |
| C                   | <b><u>56.5</u></b>        | <b><u>21.1</u></b>        | 1.08                      | <b><u>148</u></b> | <i>N/A</i>            | <b><u>0.24</u></b>     | <i>N/A</i>           |
| D                   | 65.0                      | 22.5                      | 0.78                      | 272               | 1.1                   | 0.19                   | 4                    |
| E                   | 65.0                      | 6.2                       | 0.76                      | N/A               | 1.3                   | 0.23                   | 3.9                  |
| F                   | <b><u>62.0</u></b>        | 64.5                      | 0.83                      | <b><u>362</u></b> | <b><u>0.8</u></b>     | <b><u>0.11</u></b>     | <i>N/A</i>           |
| G                   | <b><u>63.5</u></b>        | <b><u>41.3</u></b>        | 0.84                      | 345               | 0.9                   | 0.14                   | N/A                  |
| H                   | <b><u>61.0</u></b>        | <b><u>2.7</u></b>         | <b><u>0.72</u></b>        | N/A               | 1.4                   | N/A                    | 3.5                  |
| Total (Mean, (SD))  | 60.1 (5.4)                | 27.4 (19.8)               | 0.84 (0.12)               | 286 (77)          | 1.1 (0.3)             | 18% (5)                | 3.9 (0.3)            |
| <i>UIH cohort**</i> |                           |                           |                           |                   |                       |                        |                      |
| Mean (SD)           | <b><u>53.3</u></b> (18.5) | <b><u>98.6</u></b> (86.7) | <b><u>1.93</u></b> (2.63) | 386 (521)         | 1.3 (1)               | <b><u>18%</u></b> (11) | 5.8 (3.6)            |
| Median (IQR)        | 55 (40–67)                | 75.2 (32–146)             | 1.02 (0.8–1.6)            | 297 (230–417)     | 1.1 (0.7–1.5)         | 16% (10–24)            | 4.7 (3.3–7.3)        |

\* Bolded, italicized, underlined values represent variables used in the final models

\*\* These values are for the entire cohort, validate and test, N=516

Cr creatinine, CRP C reactive peptide, IQR interquartile range, LDH lactate dehydrogenase, Neut neutrophile count, Lymph lymphocyte count, Lymph/WBC lymphocyte to white blood cell ratio, SD standard deviation, UIH University of Illinois Hospital

H used both the CRP and creatinine, helping explain its poor performance. For creatinine, there are studies showing socioeconomic and ethnic variations in chronic kidney disease [63] with one systematic review showing the prevalence of chronic kidney disease in China was less than a fourth of the rate in the US [64]. The higher

creatinine in the test cohort may not be related only to differences in illness at presentation, but rather differences in the prevalence of CKD.

It is not fully clear why the models produced at UIH using our training cohort did not perform better on our test cohort, though there are some likely factors.

The AUC for mortality decreased from 0.98 to 0.84 and for criticality, from 0.97 to 0.83. In analysis of the entire cohort, we were able to determine that the mortality and criticality were associated with the admission date. This is consistent with publications showing an improved mortality rate over time [59, 60]. The WBC, lymphocytes and neutrophils were not all used in each model and all went up in the test cohort. Thus, it is possible that some variable not in the models changed over time, producing a worse fit compared to the first 60% of patients.

The number of cases for which a model is unable to generate a prediction due to missing data is an important practical consideration for model implementation. The fraction of the test cohort for which predictions could not be generated due to missingness ranged between 17 and 31% for external models. The UI Health models could not generate predictions in 27% of the patients. Though retrospectively missing data can be imputed, this is not so easy in real time by clinicians during patient care, so was not done. This demonstrates non standardized test ordering, which is not surprising as our understanding of what is useful and necessary for testing in suspected COVID patients has evolved.

It is interesting to note that many of the tests which have been used commonly in these and other models were missed more frequently in the test cohort than the earlier development cohort. Ferritin 17.4% from 7.1%, LDH 27.5% from 16.5%. It is not clear why these tests were ordered less over time, particularly LDH with many publications demonstrated its prognostic power [15, 16, 21, 22, 25, 27, 42–45, 50, 52, 53]. It is possible that the ordering of these inflammatory prognostic markers [65] decreased as clinicians' confidence with clinical prognosis improved.

D-dimer on the other hand was missing less frequently in the test cohort, 27.5% from 40.8%. This difference may be due to an increased concern for venous thromboembolism in COVID 19 infections [66] which developed over time.

An important question is what model to use to provide prognostic information to clinicians. Using your own data to inform future care is consistent with a learning health system [67]. The ideal situation is that clinical decision support (CDS) could supply the best prediction for a patient based on the most recent trends at the time. Another reason to use your own data, especially with COVID-19, is that the disease, treatment and outcomes are likely to change over time [59, 60], while the models in the literature are static. An additional benefit of using your own data and predictive models is the ability to see which diagnostic tests are

most useful prognostically, but are not ordered enough, leading to more evidenced based order sets.

Our literature search has limitations due to the inability to ensure that all possible synonyms were used along with other reasons that the search strategy may have missed articles. As related to COVID-19, the rate of discovery and publication is so rapid that many models were likely published between the time of study completion and study publication.

Limitations related to our cohort and analysis are first that this is a single site study, and these models may have performed differently at other sites. The size of the test cohort contributed to the relatively large confidence intervals of the AUC's, making statistical significance difficult to prove. We were unable to follow patients consistently after discharge, thus could not measure timed outcomes like 30-day mortality. Lastly, we could not control for changes in treatment which have occurred over time.

## Conclusions

Both internal and some external models were found to work well at predicting mortality in our test cohort. The 3 best external models used at least age, LDH and lymphocytes. Inconsistent ordering of lab tests led to the inability to generate predictions for 27–31% of our cohort using the 3 best external models and the 2 UIH models.

As not all the external models worked well, it would be difficult to know which model to use for future admissions at a particular time during the pandemic as treatment and patient mix can change. As an institution's own prior patients are most similar to their next group of patients, using models from local data should be considered.

## Abbreviations

ALC: Absolute lymphocyte count; ANC: Absolute neutrophil count; AST: Aspartate transaminase; BUN: Blood urea nitrogen; CHD: Coronary heart disease; CK: Creatinine kinase; CKD: Chronic kidney disease; Cr: Creatinine; CRP: C reactive peptide; CLD: Chronic lung disease; CXR: Chest Xray; DM: Diabetes mellitus; eGFR: Estimated glomerular filtration rate; HTN: Hypertension; IL 6: Interleukin 6; LDH: Lactate dehydrogenase; NEWS2: National Early Warning Score 2; NLR: Neutrophil to lymphocyte ratio; PC: Procalcitonin; PT: Prothrombin time; PTT: Partial thromboplastin time; RDW: Red cell distribution width; SAT: Oxygen saturation; Stan Dev: Standard deviation; UIH: University of Illinois Hospital; WBC: White blood cell count.

## Acknowledgements

Amulya Boppana and Yufu Zhang assisted with data extraction and data validation.

## Authors' contributions

WG: Involved in all aspects of this study. JMRF: Data acquisition, data interpretation, drafting and revision of the manuscript. KC: Data acquisition, data interpretation and drafting of the initial manuscript. SH: Data interpretation and drafting of the initial manuscript. KMK: Drafting of the initial manuscript and revision of the manuscript. MP: Data interpretation and drafting of the

initial manuscript. JT: Data interpretation and drafting of the initial manuscript. JZ: Data interpretation and revision of the manuscript. DH: Design, Data interpretation, drafting of the initial manuscript. All authors read and approved the final manuscript.

#### Funding

This research has been funded by the University of Illinois at Chicago Center for Clinical and Translational Science (CCTS) Award UL1TR002003. The funding body did not take part in the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

#### Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to privacy but are available from the corresponding author on reasonable request.

#### Declarations

##### Ethics approval and consent to participate.

This study was approved by University of Illinois at Chicago Internal Review Board. Permission from University of Illinois at Chicago Privacy Board and Internal Review Board were required to access the data used in this study. All the experiment protocols involving human data were in accordance with the University of Illinois at Chicago Privacy Board and Internal Review Board guidelines.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Departments of Medicine and Pharmacy Systems, Outcomes and Policy, University of Illinois At Chicago (UIC), Chicago, USA. <sup>2</sup>Department of Neurology, Clinical Informatics Fellowship, UIC, Chicago, USA. <sup>3</sup>College of Medicine, UIC, Chicago, USA. <sup>4</sup>Department of Mechanical and Industrial Engineering, UIC, Chicago, USA. <sup>5</sup>Department of Family and Community Medicine, UIC, Chicago, USA. <sup>6</sup>Department of Psychiatry, UIC, Chicago, USA.

Received: 27 March 2021 Accepted: 29 June 2021

Published online: 24 July 2021

#### References

- Taubenberger JK, Morens DM. 1918 Influenza: the mother of all pandemics. *Rev Biomed*. 2006;17(1):69–79.
- Bhutta N, Blair J, Dettling LJ, Moore KB. COVID-19, the CARES Act, and families' financial security. SSRN 2020. <https://doi.org/10.2139/ssrn.3631903>.
- De Dombal F, Leaper D, Staniland JR, McCann A, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. *Br Med J*. 1972;2(5804):9–13.
- Ranson J, Rifkind K, Roses D, Fink S, Spencer F. Prognostic signs and the role of operative management in acute pancreatitis. *Surg Gynecol Obstet*. 1974;139(1):69–81.
- Kuo DC, Rider AC, Estrada P, Kim D, Pillow MT. Acute pancreatitis: what's the score? *J Emerg Med*. 2015;48(6):762–70.
- Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'donnell CJ. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63(25 Pt B):2935–59.
- Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, Ballantyne CM, Go AS. Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population. *J Am Coll Cardiol*. 2016;67(18):2118–30.
- Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*, vol. 398. Hoboken: Wiley; 2013.
- Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern Syst*. 1991;21(3):660–74.
- Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002;2(3):18–22.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; San Francisco, CA. 2016. p. 785–94.
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. In: *Presented at Advances in Neural Information Processing Systems, NIPS, Long Beach, CA, vol. 30*. 2017. p. 3146–54.
- Prokhorenkova L, Gusev G, Vorobev A, Drogosh AV, Gulina A. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*; 2018. p. 6638–48.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*. 1988;44(3):837–45.
- Bousquet G, Falgarone G, Deutsch D, Derolez S, Lopez-Sublet M, Goudot F-X, Amari K, Uzunhan Y, Bouchaud O, Pamoukdjian F. ADL-dependency, D-Dimers, LDH and absence of anticoagulation are independently associated with one-month mortality in older inpatients with Covid-19. *Aging (Albany NY)*. 2020;12(12):11306–13.
- Toraih EA, Elshazli RM, Hussein MH, Elgaml A, Amin MN, El-Mowafy M, El-Mesery M, Ellythy A, Duchesne J, Killackey MT. Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in COVID-19 patients: a meta-regression and decision tree analysis. *J Med Virol*. 2020;92(11):2473–88.
- Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, Chen G, Cheng G, Wang Y, Bi J. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL Score. *Clin Infect Dis*. 2020;71(6):1393–9.
- Wang K, Zuo P, Liu Y, Zhang M, Zhao X, Xie S, Zhang H, Chen X, Liu C. Clinical and laboratory predictors of in-hospital mortality in patients with COVID-19: a cohort study in Wuhan, China. *Clin Infect Dis*. 2020;71(16):2079–88.
- Liang W, Liang H, Ou L, Chen B, Chen A, Li C, Li Y, Guan W, Sang L, Lu J. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med*. 2020;180(8):1–9.
- Ma X, Ng M, Xu S, Xu Z, Qiu H, Liu Y, Lyu J, You J, Zhao P, Wang S. Development and validation of prognosis model of mortality risk in patients with COVID-19. *Epidemiol Infect*. 2020;148:e168.
- Bonetti G, Manelli F, Patroni A, Bettinardi A, Borrelli G, Fiordalisi G, Marino A, Menolfi A, Saggini S, Volpi R, et al. Laboratory predictors of death from coronavirus disease 2019 (COVID-19) in the area of Valcamonica, Italy. *Clin Chem Lab Med*. 2020;58(7):1100–5.
- Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, Thode HC, Singer AJ, Duong TQ. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PLoS ONE*. 2020;15(7):e0236618.
- Hu L, Chen S, Fu Y, Gao Z, Long H, Ren H, Zuo Y, Li H, Wang J, Xu Q. Risk factors associated with clinical outcomes in 323 coronavirus disease 2019 (COVID-19) Hospitalized Patients in Wuhan, China. *Clin Infect Dis*. 2020;71(16):2089–98.
- Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, Tang C, Sang L, Liu J, Ni Z. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. *Chest*. 2020;158(1):97–105.
- Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, Frix A-N, Louis R, Moutschen M, Li J. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J*. 2020;56(2):2001104.
- Cheng A, Hu L, Wang Y, Huang L, Zhao L, Zhang C, Liu X, Xu R, Liu F, Li J. Diagnostic performance of initial blood urea nitrogen combined with D-Dimer levels for predicting in-hospital mortality in COVID-19 patients. *Int J Antimicrob Agents*. 2020;56(3):106110.
- Laguna-Goya R, Utrero-Rico A, Talayero P, Lasa-Lazaro M, Ramirez-Fernandez A, Naranjo L, Segura-Tudela A, Cabrera-Marante O, de Frias ER, Garcia-Garcia R. IL-6-based mortality risk model for hospitalized patients with COVID-19. *J Allergy Clin Immunol*. 2020;146(4):799–807.
- Liu Q, Song N, Zheng Z, Li J, Li S. Laboratory findings and a combined multifactorial approach to predict death in critically ill patients with COVID-19: a retrospective study. *Epidemiol Infect*. 2020;148:e129.

29. Ok F, Erdogan O, Durmus E, Carkci S, Canik A. Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients. *J Med Virol*. 2021;93(2):786–93.
30. Abdulaal A, Patel A, Charani E, Denny S, Mughal N, Moore L. Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: model development and validation. *J Med Internet Res*. 2020;22(8):e20259.
31. Qin J-J, Cheng X, Zhou F, Lei F, Akolkar G, Cai J, Zhang X-J, Blet A, Xie J, Zhang P. Redefining cardiac biomarkers in predicting mortality of inpatients with COVID-19. *Hypertension*. 2020;76(4):1104–12.
32. Turcotte JJ, Meisenberg BR, MacDonald JH, Menon N, Fowler MB, West M, Rhule J, Qureshi SS, MacDonald EB. Risk factors for severe illness in hospitalized Covid-19 patients at a regional hospital. *PLoS ONE*. 2020;15(8):e0237558.
33. Shang Y, Liu T, Wei Y, Li J, Shao L, Liu M, Zhang Y, Zhao Z, Xu H, Peng Z. Scoring systems for predicting mortality for severe patients with COVID-19. *EClinicalMedicine*. 2020;24:100426.
34. Zeng Z, Ma Y, Zeng H, Huang P, Liu W, Jiang M, Xiang X, Deng D, Liao X, Chen P. Simple nomogram based on initial laboratory data for predicting the probability of ICU transfer of COVID-19 patients: multicenter retrospective study. *J Med Virol*. 2020. <https://doi.org/10.1002/jmv.26244>.
35. Yu C, Lei Q, Li W, Wang X, Liu W, Fan X, Li W. Clinical characteristics, associated factors, and predicting COVID-19 mortality risk: a retrospective study in Wuhan, China. *Am J Prev Med*. 2020;59(2):168–75.
36. Sun L, Liu G, Song F, Shi N, Liu F, Li S, Li P, Zhang W, Jiang X, Zhang Y. Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. *J Clin Virol*. 2020;128:104431.
37. Wu S, Du Z, Shen S, Zhang B, Yang H, Li X, Cui W, Chen F, Huang J. Identification and validation of a novel clinical signature to predict the prognosis in confirmed COVID-19 patients. *Clin Infect Dis*. 2020;71:3154–62.
38. Lorente-Ros A, Ruiz JMM, Rincón LM, Pérez RO, Rivas S, Martínez-Moya R, Sanromán MA, Manzano L, Alonso GL, Ibáñez B. Myocardial injury determination improves risk stratification and predicts mortality in COVID-19 patients. *Cardiol J*. 2020;27(5):489–96.
39. Tatum D, Taghavi S, Houghton A, Stover J, Toraih E, Duchesne J. Neutrophil-to-lymphocyte ratio and outcomes in Louisiana COVID-19 patients. *Shock*. 2020;54(5):652–8.
40. Liu X, Shi S, Xiao J, Wang H, Chen L, Li J, Han K. Prediction of the severity of Corona Virus Disease 2019 and its adverse clinical outcomes. *Jpn J Infect Dis*. 2020;73(6):404–10.
41. Galloway JB, Norton S, Barker RD, Brookes A, Carey I, Clarke BD, Jina R, Reid C, Russell MD, Snee R. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: an observational cohort study. *J Infect*. 2020;81(2):282–8.
42. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell*. 2020;2(5):283–8.
43. Su M, Yuan J, Peng J, Wu M, Yang Y, Peng YG. Clinical prediction model for mortality of adult diabetes inpatients with COVID-19 in Wuhan, China: a retrospective pilot study. *J Clin Anesth*. 2020;66:109927.
44. Xie J, Hungerford D, Chen H, Abrams ST, Li S, Wang G, Wang Y, Kang H, Bonnett L, Zheng R. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. <https://doi.org/10.1101/2020.03.28.20045997>.
45. Zhang S, Guo M, Duan L, Wu F, Hu G, Wang Z, Huang Q, Liao T, Xu J, Ma Y. Development and validation of a risk factor-based system to predict short-term survival in adult hospitalized patients with COVID-19: a multicenter, retrospective, cohort study. *Crit Care*. 2020;24(1):438–51.
46. Hun C, Liu Z, Jiang Y, Shi O, Zhang X, Xu K, Suo C, Wang Q, Song Y, Yu K, et al. Early prediction of mortality risk among severe COVID-19 patients using machine learning. *Int J Epidemiol*. 2020;69:343.
47. Caramelo F, Ferreira N, Oliveiros B. Estimation of risk factors for COVID-19 mortality-preliminary results. *medRxiv*. 2020;382:727.
48. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, Searle T, Kraljevic Z, Shek A, Phan HT. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *medRxiv*. 2020;46:357.
49. Shi Y, Yu X, Zhao H, Wang H, Zhao R, Sheng J. Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. *Crit Care*. 2020;24(1):1–4.
50. Huang H, Cai S, Li Y, Li Y, Fan Y, Li L, Lei C, Tang X, Hu F, Li F, et al. Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis. *medRxiv*. 2020;2:113.
51. Zhang H, Shi T, Wu X, Zhang X, Wang K, Bean D, Dobson R, Teo JT, Sun J, Zhao P. Risk prediction for poor outcome and death in hospital inpatients with COVID-19: derivation in Wuhan, China and external validation in London, UK. 2020. <https://doi.org/10.1101/2020.04.28.20082222>.
52. Jans M, Kuijper T, den Hollander J, Bisoendial R, Pogary K, van den Dorpel M, Zirkzee E, Kok M, Waverijn G, Ruiter R. Predicting severe COVID-19 at presentation, introducing the COVID severity score. *SSRN 2020*; <https://doi.org/10.2139/ssrn.3627260>.
53. Pérez FM, Del Pino JL, García NJ, Ruiz EM, Méndez CA, Jiménez JG, Romero FN, Rodríguez MN. Comorbidity and prognostic factors on admission in a COVID-19 cohort of a general hospital. *Rev Clin Esp*. 2020. <https://doi.org/10.1016/j.rceng.2020.05.010>.
54. Regina J, Papadimitriou-Olivgeris M, Burger R, Le Pogam M-A, Niemi T, Filippidis P, Tschopp J, Desgranges F, Viala B, Kampouri E. Epidemiology, risk factors and clinical course of SARS-CoV-2 infected patients in a Swiss university hospital: an observational retrospective study. *PLoS ONE*. 2020;15(11):e0240781.
55. Abdulaal A, Patel A, Charani E, Denny S, Alqahtani SA, Davies GW, Mughal N, Moore LS. Comparison of deep learning with regression analysis in creating predictive models for SARS-CoV-2 outcomes. *BMC Med Inform Decis Mak*. 2020;20(1):1–11.
56. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D, Inglesby P. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430–6.
57. Ioannou GN, Locke E, Green P, Berry K, O'Hare AM, Shah JA, Crothers K, Eastment MC, Dominitz JA, Fan VS. Risk factors for hospitalization, mechanical ventilation, or death among 10 131 US veterans with SARS-CoV-2 infection. *JAMA Netw Open*. 2020;3(9):e2022310.
58. O'Driscoll M, Dos Santos GR, Wang L, Cummings DA, Azman AS, Paireau J, Fontanet A, Cauchemez S, Salje H. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*. 2020;552:1708.
59. Dennis JM, McGovern AP, Vollmer SJ, Mateen BA. Improving survival of critical care patients with coronavirus disease 2019 in England: a national cohort study, March to June 2020. *Crit Care Med*. 2020;49(2):209–14.
60. Horvitz L, Jones SA, Cerfolio RJ, Francois F, Greco J, Rudy B, Petrilli CM. Trends in Covid-19 risk-adjusted mortality rates in a single health system. *J Hosp Med*. 2020;16(2):90–2.
61. Gupta RK, Marks M, Samuels TH, Luintel A, Rampling T, Chowdhury H, Quartagno M, Nair A, Lipman M, Abubakar I. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. *Eur Respir J*. 2020;56(6):2003498.
62. Lim E, Miyamura J, Chen JJ. Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and White. *Hawaii J Med Public Health*. 2015;74(9):302–10.
63. Brück K, Stel VS, Gambaro G, Hallan S, Völzke H, Ärnlöv J, Kastarinen M, Guessous I, Vinhas J, Stengel B. CKD prevalence varies across the European general population. *J Am Soc Nephrol*. 2016;27(7):2135–47.
64. McCullough K, Sharma P, Ali T, Khan I, Smith WC, MacLeod A, Black C. Measuring the population burden of chronic kidney disease: a systematic literature review of the estimated prevalence of impaired kidney function. *Nephrol Dial Transplant*. 2012;27(5):1812–21.
65. Zeng F, Huang Y, Guo Y, Yin M, Chen X, Xiao L, Deng G. Association of inflammatory markers with the severity of COVID-19: a meta-analysis. *Int J Infect Dis*. 2020;96:467–74.
66. Schulman S, Hu Y, Konstantinides S. Venous Thromboembolism in COVID-19. *Thromb Haemost*. 2020;120(12):1642–53.
67. Borsky AE, Savitz LA, Bindman AB, Mossburg S, Thompson L. AHRQ series on improving translation of evidence: perceived value of translational products by the AHRQ EPC Learning Health Systems Panel. *Jt Comm J Qual Patient Saf*. 2019;45(11):772–8.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.