Clinical features of COVID-19 mortality: development and validation of a clinical prediction model

Arjun S Yadaw, Yan-chak Li, Sonali Bose, Ravi Iyengar, Supinda Bunyavanich*, Gaurav Pandey**

Lancet Digital Health 2020; 2: e516–25

See **Comment** page e496

*Joint senior authors

Department of Pharmacological Sciences (A S Yadaw PhD, Prof R Iyengar PhD), **Department of Genetics and Genomic Sciences** (Y-c Li MPhil, Prof S Bunyavanich MD, G Pandey PhD), **and Department of Medicine** (S Bose MD), **Icahn School of Medicine at Mount Sinai, New York, NY, USA**

Correspondence to:
Dr Gaurav Pandey, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
gaurav.pandey@mssm.edu

Summary

Background The COVID-19 pandemic has affected millions of individuals and caused hundreds of thousands of deaths worldwide. Predicting mortality among patients with COVID-19 who present with a spectrum of complications is very difficult, hindering the prognostication and management of the disease. We aimed to develop an accurate prediction model of COVID-19 mortality using unbiased computational methods, and identify the clinical features most predictive of this outcome.

Methods In this prediction model development and validation study, we applied machine learning techniques to clinical data from a large cohort of patients with COVID-19 treated at the Mount Sinai Health System in New York City, NY, USA, to predict mortality. We analysed patient-level data captured in the Mount Sinai Data Warehouse database for individuals with a confirmed diagnosis of COVID-19 who had a health system encounter between March 9 and April 6, 2020. For initial analyses, we used patient data from March 9 to April 5, and randomly assigned (80:20) the patients to the development dataset or test dataset 1 (retrospective). Patient data for those with encounters on April 6, 2020, were used in test dataset 2 (prospective). We designed prediction models based on clinical features and patient characteristics during health system encounters to predict mortality using the development dataset. We assessed the resultant models in terms of the area under the receiver operating characteristic curve (AUC) score in the test datasets.

Findings Using the development dataset (n=3841) and a systematic machine learning framework, we developed a COVID-19 mortality prediction model that showed high accuracy (AUC=0·91) when applied to test datasets of retrospective (n=961) and prospective (n=249) patients. This model was based on three clinical features: patient's age, minimum oxygen saturation over the course of their medical encounter, and type of patient encounter (inpatient *vs* outpatient and telehealth visits).

Interpretation An accurate and parsimonious COVID-19 mortality prediction model based on three features might have utility in clinical settings to guide the management and prognostication of patients affected by this disease. External validation of this prediction model in other populations is needed.

Funding National Institutes of Health.

Introduction

The COVID-19 pandemic has affected more than 18 million individuals, and caused almost 700 000 deaths worldwide as of Aug 3, 2020.[1] Although the causative virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) primarily targets the respiratory system,[2,3] complications in other organ systems (eg, cardiovascular, neurological, and renal) can also contribute to death from the disease. Clinical experience thus far has shown substantial heterogeneity in the trajectory of SARS-CoV-2 infection, spanning from patients who are asymptomatic to those with mild, moderate, and severe disease forms with low survival rates.[2,3] Notably, accurate prediction of clinical outcomes for patients across this spectrum of clinical presentations can be difficult. This problem presents an enormous challenge to the prognostication and management of patients with COVID-19, especially within disease epicentres that need to triage a high volume of patients. Therefore, accurate prediction of COVID-19 mortality and the identification of contributing factors would allow for targeted strategies in patients with the highest risk of death.

Towards this aim, we analysed clinical data from 5051 patients who had laboratory confirmed SARS-CoV-2 infection and were treated in multiple hospitals and ambulatory locations of the Mount Sinai Health System in New York City, NY, USA, spanning different boroughs of the city. We aimed to use multiple machine learning-based classification algorithms[4] to develop models that can accurately predict mortality from COVID-19. We aimed to identify clinical features that contributed the most to this prediction. An improved understanding of predictive factors for COVID-19 is crucial for the development of support systems for clinical decision making that can better identify those with higher risk of mortality, and inform interventions to reduce the risk of death.

Research in context

Evidence before this study

We searched PubMed and its associated LitCovid repository for publications in English from database inception until May 10, 2020, using the terms “coronavirus”, “COVID-19”, “death”, “mortality”, and “prediction”. The studies we identified generally focused on a small set of clinical features, risk factors, or small cohorts to study or predict mortality due to COVID-19, which might not sufficiently capture the novelty and complexity of the disease. These studies also used relatively simple analytical methods, which might not adequately model the inherent difficulties of the data (eg, collinear or irrelevant features, non-linear relationships between features and mortality, noise, and missing data). These factors likely restrict the ability of existing work to accurately predict mortality.

Added value of this study

We analysed clinical data from a large cohort of patients with COVID-19 treated in a major health system serving a global epicentre of COVID-19 (New York City, NY, USA) to identify prediction models of mortality due to COVID-19. The novel model we developed was based on three clinical features (age, minimum oxygen saturation during encounter, and health-care setting of patient encounter) and accurately predicted mortality risk in two validation cohorts (area under the receiver operating characteristic curve of >0·90).

Implications of all the available evidence

We present here a highly robust COVID-19 mortality prediction analysis, derived from working with both the largest number of patients and clinical features to date. Our large cohort and the rigorous analytical methods we used lend our study two major advantages over previous studies: our prediction model performs better than those proposed previously, and since we started with a large set of clinical features, the ones that we identified to be most strongly associated with mortality are more objective and accurate. After validation in other health system populations, our model could be implemented in clinical settings to enable improved prognostication and management of COVID-19.

Methods

Study design and population

In this prediction model development and validation study, we used anonymised electronic medical record (EMR) data from patients with a confirmed diagnosis of COVID-19 who had been treated in the Mount Sinai Health System, between March 9 and April 6, 2020. The Mount Sinai Health System is a network of eight hospitals and over 400 ambulatory practices spanning the New York metropolitan area (appendix p 4). A diagnosis of COVID-19 was determined by positive PCR-based clinical laboratory testing for SARS-CoV-2.

See Online for appendix

Data were internally stored and managed by the Mount Sinai Data Warehouse. After anonymisation and removal of protected health information, the data were released in a text-delimited format for research purposes.

Patient-level data were collected for the initial analyses in our study. All patients with a confirmed diagnosis of COVID-19 and an inpatient or outpatient (including telehealth) visit to the Mount Sinai Health System during the study period were included. All collected data and events occurring during the time that medical attention was provided to the patient during the visit were defined as an encounter. An encounter included all collected data and events occurring over the time that medical attention was provided to the patient during the visit or stay in hospital. The initial clinical data element collected during the patient encounter was considered as the data at presentation. These data included demographic variables, such as age, sex, and ethnicity, and comorbidities, such as diabetes and asthma, defined by the presence of corresponding International Classification of Diseases tenth revision codes that were active on the patient’s EMR problem list at the beginning of the encounter. These data were self-reported by patients or recorded in the medical chart by health-care providers during current and previous medical encounters.

Patients also reported their smoking status as current, never, past, or passive (passive was included as a category given the potential health effects of second-hand smoke[5]). Variables pertaining to the highest or lowest value recorded of its kind (ie, highest body temperature and lowest oxygen saturation level) during the encounter were designated as maximum or minimum. Notably, datasets downloaded on a set date (eg, April 6) only included data collected up until the day before that date (ie, April 5), and therefore did not include any follow-up data of patients in those cohorts beyond those respective dates. The final set of features and variables, which were harmonised across the various Mount Sinai Health System medical centres as well as possible, are listed and defined in the appendix (p 5). The mortality outcome predicted in this work was defined as positive if a patient with COVID-19 died during their encounter with the health system by the current date (ie, April 5, 2020). If a patient had not reached the outcome (death) by the time the data were obtained, their outcome was marked as negative (alive).

Under an agreement with the Institutional Review Board (IRB), which is a precursor and separate from our study, the Mount Sinai Data Warehouse released anonymised clinical data of all patients with COVID-19 who had or were having treatment in the Mount Sinai Health System to the Mount Sinai Health System (not to

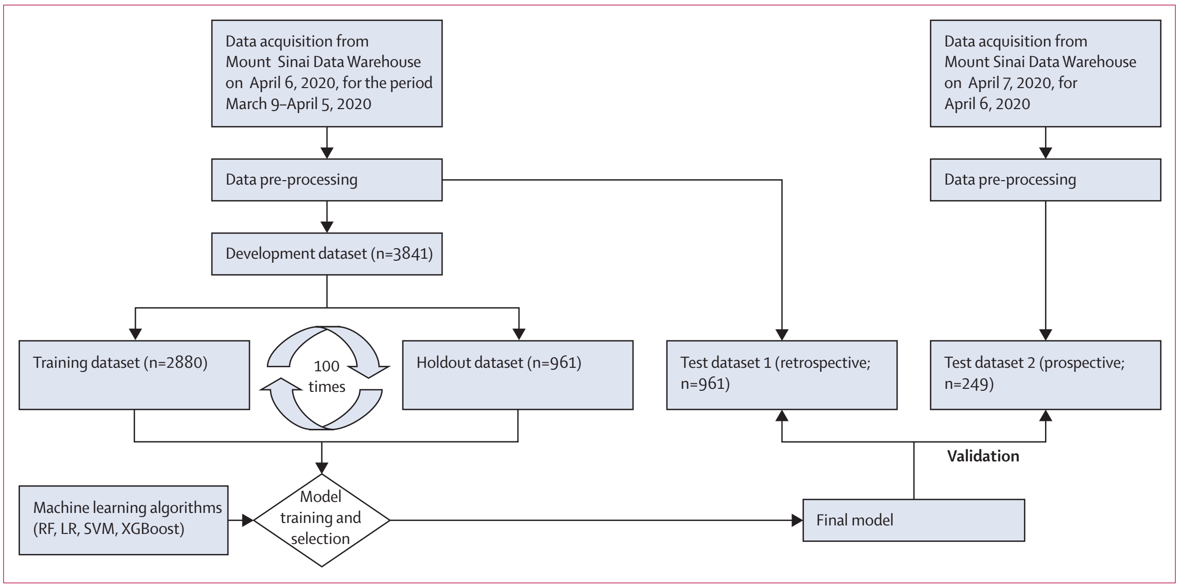
Data acquisition from Mount Sinai Data Warehouse on April 6, 2020, for the period March 9–April 5, 2020
Data pre-processing
Development dataset (n=3841)
Training dataset (n=2880)
100 times
Holdout dataset (n=961)
Machine learning algorithms (RF, LR, SVM, XGBoost)
Model training and selection
Data acquisition from Mount Sinai Data Warehouse on April 7, 2020, for April 6, 2020
Data pre-processing
Test dataset 1 (retrospective; n=961)
Test dataset 2 (prospective; n=249)
Validation
Final model

Figure 1: **Workflow for data management and COVID-19 mortality prediction model development**
Data were obtained from the Mount Sinai Data Warehouse. After pre-processing, data for patients with COVID-19 (n=4802) were randomly divided in an 80:20 ratio into a prediction model development dataset (n=3841) and an independent retrospective validation dataset (test dataset 1; n=961). For prediction model training and selection, the development dataset was further randomly split into a 75% training dataset (n=2880) and a 25% holdout dataset (n=961). Four classification algorithms were assessed. The final predictive model was validated on test dataset 1 and another independent prospective validation dataset (test dataset 2; n=249). LR=logistic regression. RF=random forest. SVM=support vector machine. XGBoost=eXtreme Gradient Boosting.

the public). Since anonymised data were used, patient consent was not sought.

Datasets

We downloaded patient-level data on April 6 for the period March 9–April 5, 2020, and this dataset was randomly split, without replacement based, on a uniform distribution, into two groups of independent patients comprising 80% of the sample for development of the mortality prediction model (ie, development dataset), and 20% for a retrospective test dataset (referred to as test dataset 1). The date of data acquisition was chosen to include as large a cohort as possible in the prediction model development dataset. These data were pre-processed to address quality issues, such as repeated entries for any patients, excessive missing values in any features, and the absence of normalisation of continuous features (figure 1; details are in the appendix [p 2]).

Furthermore, a prospective validation set of patients, independent of the other datasets, referred to as test dataset 2, included new patients with COVID-19 included in the Mount Sinai Health System database on April 7, 2020. Based on the definitions above, this prospective test dataset comprised patients with COVID-19 who were treated on April 6, 2020, specifically, and whose data, including their outcome, were recorded on the same day. The demographic and clinical data (ie, feature data) recorded for these patients were consistent with those of the patients in the development dataset and test dataset 1. Test dataset 2 was pre-processed using the same process as the development dataset and test dataset 1 (appendix p 2).

We did univariate significance analyses of the differences in continuous features between alive and deceased patients using Student's *t* test and of categorical features using the χ^2 test in all the resultant datasets. We calculated effect sizes as odds ratios (ORs) for values of categorical features using their respective counts in the data, and for continuous features using a logistic regression model. We used a p value threshold of 0·05 to determine significance in these and other analyses in this study.

Identification and validation of the prediction model

We implemented a systematic machine learning-based framework to construct the mortality prediction model from the development dataset using missing value imputation,[6] feature selection,[7] classification,[4] and statistical[8] techniques. Specifically, we used the recursive feature elimination method[7] for feature selection, and logistic regression, support vector machine, random forest, and eXtreme Gradient Boosting (XGBoost) algorithms[4] for prediction. We adopted these multivariate algorithms because they attempt to find the best combination of individual features that can constitute as accurate a prediction model as possible.

We aimed to build a model that could classify a patient with COVID-19 as likely to survive or die from the disease—ie, “alive” or “deceased”. The resultant prediction model was then validated in test datasets 1 and 2 in terms of the area under the receiver operating characteristic curve (AUC) score.[9] On the basis of missing value imputation, feature selection, and prediction methods listed here and

	Development dataset					Test dataset 1 (retrospective)					Test dataset 2 (prospective)				
	Total (n=3841)	Deceased (n=313)	Alive (n=3528)	Odds ratio (95% CI)	p value	Total (n=961)	Deceased (n=78)	Alive (n=883)	Odds ratio (95% CI)	p value	Total (n=249)	Deceased (n=25)	Alive (n=224)	Odds ratio (95% CI)	p value
Age, years	56·2 (19·0)	73·4 (12·7)	54·7 (18·7)	1·07 (1·06–1·08)	<0·0001	56·2 (18·5)	72·9 (13·2)	54·7 (18·2)	1·07 (1·05–1·09)	<0·0001	56·0 (18·4)	69·3 (11·6)	54·5 (18·4)	1·05 (1·02–1·08)	<0·0001
Sex															
Male	2125 (55%)	192 (61%)	1933 (55%)	1·29 (1·02–1·64)	0·092	534 (56%)	48 (62%)	486 (55%)	1·29 (0·80–2·08)	0·34	139 (56%)	16 (64%)	123 (55%)	1·45 (0·61–3·41)	0·67
Female	1695 (45%)	121 (39%)	1575 (45%)	0·78 (0·61–0·98)	..	427 (44%)	30 (38%)	397 (45%)	0·77 (0·48–1·24)	..	110 (44%)	9 (36%)	101 (45%)	0·70 (0·30–1·66)	..
Ethnicity															
White	1008 (26%)	93 (30%)	915 (26%)	1·17 (0·91–1·52)	0·66	247 (26%)	20 (26%)	227 (26%)	0·98 (0·57–1·67)	0·014	68 (27%)	2 (8%)	66 (29%)	0·19 (0·04–0·82)	0·0083
African American	973 (25%)	83 (27%)	890 (25%)	1·04 (0·80–1·35)	..	242 (25%)	18 (23%)	224 (25%)	0·86 (0·50–1·50)	..	63 (25%)	10 (40%)	53 (24%)	2·06 (0·85–4·98)	..
Asian	162 (4%)	11 (4%)	151 (4%)	0·79 (0·43–1·48)	..	38 (4%)	8 (10%)	30 (3%)	3·21 (1·42–7·28)	..	8 (3%)	0	8 (4%)	0·51 (0·03–9·14)	..
Latino	932 (24%)	74 (24%)	858 (24%)	0·93 (0·71–1·23)	..	227 (24%)	12 (15%)	215 (24%)	0·55 (0·29–1·04)	..	57 (23%)	5 (20%)	52 (23%)	0·76 (0·27–2·16)	..
Other	528 (14%)	39 (12%)	489 (14%)	0·86 (0·61–1·22)	..	158 (16%)	17 (22%)	141 (16%)	1·45 (0·82–2·56)	..	22 (9%)	6 (24%)	16 (7%)	3·95 (1·37–11·42)	..
Encounter type															
Inpatient	3011 (78%)	307 (98%)	2704 (77%)	15·59 (6·92–35·11)	<0·0001	749 (78%)	77 (99%)	672 (76%)	24·18 (3·34–174·88)	<0·0001	197 (79%)	25 (100%)	172 (77%)	15·12 (0·90–252·72)	0·014
Other	830 (22%)	6 (2%)	824 (23%)	0·06 (0·03–0·14)	..	212 (22%)	1 (1%)	211 (24%)	0·04 (0·01–0·30)	..	52 (21%)	0	52 (23%)	0·07 (0·00–1·11)	..
Temperature, °C															
Mean during encounter	37·3 (1·4)	37·1 (2·1)	37·3 (1·2)	0·93 (0·88–0·99)	0·098	37·3 (0·9)	37·2 (1·1)	37·3 (0·9)	0·88 (0·67–1·14)	0·41	37·3 (0·9)	37·1 (1·3)	37·3 (0·9)	0·73 (0·46–1·158)	0·33
Maximum during encounter	38·0 (1·1)	38·4 (1·1)	37·9 (1·0)	1·46 (1·31–1·63)	<0·0001	37·9 (1·2)	38·4 (1·1)	37·9 (1·2)	1·39 (1·12–1·72)	<0·0001	38·0 (1·0)	38·3 (0·9)	37·9 (1·0)	1·54 (0·99–2·40)	0·057
Diastolic blood pressure at presentation, mm Hg	75·7 (13·6)	71·0 (16·6)	76·2 (13·2)	0·97 (0·96–0·98)	<0·0001	76·4 (13·6)	73·8 (19·4)	76·7 (12·8)	0·98 (0·97–1·00)	0·21	75·2 (13·2)	73·6 (17·2)	75·4 (12·6)	0·99 (0·96–1·02)	0·62
Oxygen saturation, %															
At presentation	95·2 (6·2)	90·8 (11·9)	95·6 (5·0)	0·92 (0·91–0·94)	<0·0001	95·2 (6·4)	91·4 (10·8)	95·6 (5·5)	0·94 (0·91–0·96)	0·0001	95·4 (4·6)	93·5 (6·6)	95·6 (4·1)	0·92 (0·85–0·99)	0·13
Minimum during encounter	89·4 (14·5)	69·3 (26·3)	91·6 (10·4)	0·93 (0·92–0·94)	<0·0001	90·2 (12·7)	71·3 (22·9)	92·2 (8·9)	0·91 (0·89–0·93)	<0·0001	90·1 (12·5)	76·9 (21·5)	92·0 (9·3)	0·92 (0·87–0·96)	0·0018
Smoking															
Current	134 (3%)	13 (4%)	121 (3%)	1·17 (0·65–2·11)	0·048	41 (4%)	3 (4%)	38 (4%)	0·75 (0·23–2·50)	0·11	7 (3%)	1 (4%)	6 (3%)	1·27 (0·15–11·14)	0·87
Never	1960 (51%)	148 (47%)	1812 (51%)	0·68 (0·51–0·90)	..	497 (52%)	40 (51%)	457 (52%)	0·59 (0·34–1·00)	..	137 (55%)	15 (60%)	122 (54%)	0·76 (0·27–2·09)	..
Past	625 (16%)	69 (22%)	556 (16%)	1·49 (1·11–2·01)	..	143 (15%)	21 (27%)	122 (14%)	1·99 (1·14–3·47)	..	36 (14%)	5 (20%)	31 (14%)	1·29 (0·44–3·79)	..
Passive	2 (<1%)	0	2 (<1%)	2·71 (0·12–60·17)	..	1 (<1%)	0	1 (<1%)	4·82 (0·16–145·09)	..	0	0	0	NA	..

(Table continues on next page)

this score, we followed a sampling-based training and validation process to evaluate several prediction models. Specifically, we randomly split the development dataset into training and holdout datasets in a 3:1 ratio using sampling without replacement from a uniform distribution (figure 1). Candidate prediction models were trained on the training split using the prediction algorithms mentioned, and assessed on the holdout split in terms of the AUC score. This process was repeated 100 times, and the performance results collected for all the prediction algorithms and compared. We selected the final prediction algorithm, features, and model on the basis of these performance results. Full details of all the methods are provided in the appendix (pp 2–3).

All the analyses and figure generation were done using the Python programming language (version 3.7.3) in this study.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data management, data interpretation, or writing of the report. ASY, Y-cL, RI, and GP had full access to all the data in the study and had responsibility for the decision to submit for publication.

Results

The overall workflow of the model development process is shown in figure 1. The demographic and clinical characteristics of patients with COVID-19 included in the development dataset (n=3841; of whom 313 were deceased and 3528 were alive), test dataset 1 (n=961; of whom 78 were deceased and 883 were alive), and test dataset 2 (n=249; of whom 25 were deceased and 224 were alive) are shown in the table. 2125 (55%) of 3841 patients in the development dataset were male, and 192 (61%) of 313 deceased patients were male. Patients were mostly white (1008 [26%]), African American (973 [25%]) and Latino (932 [24%]), with 162 (4%) identifying as Asian. Hypertension and diabetes were the most common comorbidities in this dataset, and few patients had obesity or cancer, and even fewer had asthma, chronic obstructive pulmonary disease (COPD), or currently smoked.

Univariate analyses of patient characteristics in the development dataset (table) showed that those who died were significantly older, with a mean age of 73·4 years (SD 12·7) compared with 54·7 years (18·7) years in those who were alive (p<0·0001). Patients who were alive were more likely to have had their initial encounter at a hospital than at an outpatient or telehealth setting within our hospital system than were patients who died (OR 15·59, 95% CI 6·92–35·11; p<0·0001). Those who died had lower oxygen saturation at initial presentation, and their minimum oxygen saturation over the duration of their encounter was also lower (p<0·0001 for both). Patients who died were more likely to be current smokers (p=0·048) and have COPD (p<0·0001), hypertension (p<0·0001), and diabetes (p<0·0001) than were those who were alive.

	Development dataset					Test dataset 1 (retrospective)					Test dataset 2 (prospective)				
	Total (n=3841)	Deceased (n=313)	Alive (n=3528)	Odds ratio (95% CI)	p value	Total (n=961)	Deceased (n=78)	Alive (n=883)	Odds ratio (95% CI)	p value	Total (n=249)	Deceased (n=25)	Alive (n=224)	Odds ratio (95% CI)	p value
(Continued from previous page)															
Asthma	160 (4%)	15 (5%)	145 (4%)	1·17 (0·68–2·02)	0·67	43 (4%)	4 (5%)	39 (4%)	1·17 (0·41–3·36)	>0·99	10 (4%)	4 (16%)	6 (3%)	6·92 (1·81–26·49)	0·0073
COPD	89 (2%)	19 (6%)	70 (2%)	3·19 (1·90–5·37)	<0·0001	19 (2%)	4 (5%)	15 (2%)	3·13 (1·01–9·67)	0·097	4 (2%)	1 (4%)	3 (1%)	3·07 (0·31–30·68)	0·87
Hypertension	869 (23%)	132 (42%)	737 (21%)	2·76 (2·18–3·51)	<0·0001	224 (23%)	29 (37%)	195 (22%)	2·09 (1·28–3·39)	0·0039	65 (26%)	11 (44%)	54 (24%)	2·47 (1·06–5·77)	0·06
Obesity	229 (6%)	23 (7%)	206 (6%)	1·28 (0·82–2·00)	0·34	73 (8%)	8 (10%)	65 (7%)	1·44 (0·66–3·12)	0·48	23 (9%)	2 (8%)	21 (9%)	0·84 (0·19–3·82)	0·89
Diabetes	608 (16%)	90 (29%)	518 (15%)	2·35 (1·80–3·05)	<0·0001	173 (18%)	25 (32%)	148 (17%)	2·34 (1·41–3·89)	0·0013	56 (22%)	10 (40%)	46 (21%)	2·58 (1·09–6·12)	0·050
HIV	62 (2%)	6 (2%)	56 (2%)	1·21 (0·52–2·83)	0·83	16 (2%)	0	16 (2%)	0·35 (0·02–5·85)	0·46	2 (1%)	0	2 (1%)	0·56 (0·02–12·42)	0·48
Cancer	209 (5%)	24 (8%)	185 (5%)	1·50 (0·96–2·33)	0·093	49 (5%)	5 (6%)	44 (5%)	1·31 (0·50–3·40)	0·76	15 (6%)	2 (8%)	13 (6%)	1·41 (0·30–6·65)	>0·99

Data are mean (SD), n (%), odds ratios with 95% CIs in parentheses, or p values. Odds ratios and 95% CIs for categorical variables were calculated using the counts in the table, and for continuous variables using a logistic regression model. p values were calculated via Student's *t* test for continuous features and χ^2 test for categorical features. COPD=chronic obstructive pulmonary disease. NA=not applicable.

***Table:* Characteristics of patients in the development and test datasets**

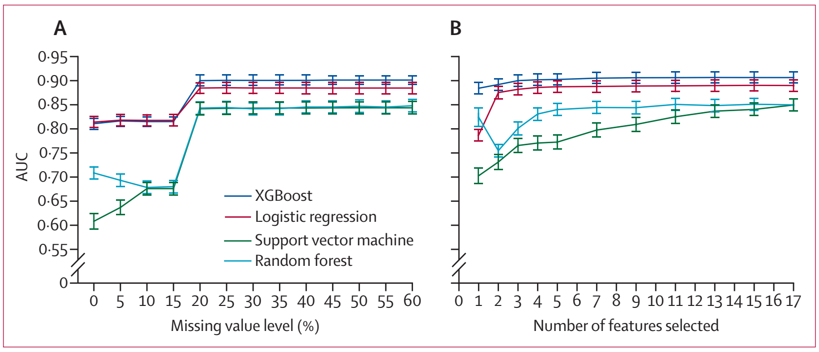

Figure 2: **Results from missing value imputation (A) and feature selection (B) during prediction model training and selection**

(A) Datapoints show the average AUC score for each candidate algorithm and missing value level, with error bars shown by whiskers. (B) Datapoints show the average AUC score for each subset of number of features with error bars shown by whiskers. The details of the computational methods underlying these analyses are provided in the appendix (pp 2–3). AUC=area under the receiver operating characteristic curve. XGBoost=eXtreme Gradient Boosting.

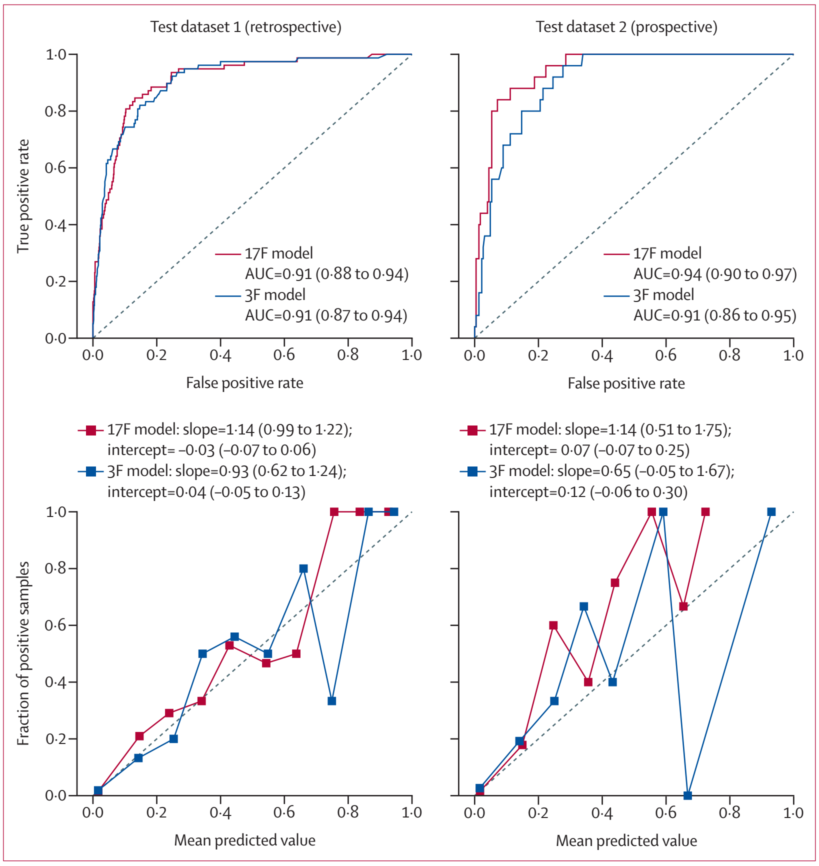

Figure 3: **Performance of the mortality prediction models on two validation datasets**

Evaluation results for test datasets 1 (A) and 2 (B) are shown here in terms of the ROC curves obtained, as well as their AUC scores, with 95% CIs in parentheses. Calibration curves of the 3F and 17F models on test datasets 1 (C) and 2 (D), with the slopes and intercepts of all the curves, along with their 95% CIs in parentheses. AUC=area under the ROC curve. ROC=receiver operating characteristic.

The characteristics of test datasets 1 and 2 were largely similar to those of the development dataset, except for some differences in the relative proportions of ethnicity (table). Although minimum oxygen saturation during encounter was consistently lower among the deceased patients versus alive patients in both test datasets, oxygen saturation levels at presentation were significantly lower among the deceased patients in test dataset 1 only. The prevalence of asthma was significantly higher in the deceased group than in the alive group in test dataset 2 and no difference was seen for the other datasets. The prevalence of diabetes was higher in deceased patients in both datasets.

Using our development dataset, we first attempted to find the optimal proportion of missing values in each variable across the patients (missing value level) that could be imputed and lead to more accurate prediction models. For this step, we took incremental steps of 5% in missing value levels in the range of 0% to 60%, and used mean imputation for continuous features and mode imputation for categorical features. At each level, the four candidate algorithms (logistic regression, random forest, support vector machine, and XGBoost) were trained and assessed on the corresponding holdout dataset in terms of the AUC score as the measure. This process was repeated 100 times and the average AUCs for each candidate prediction model were calculated to find the optimum missing value level. We identified 17 distinct clinical features with less than 20% missing values among the patients in the development dataset that improved prediction performance (figure 2). Compared with the other classification algorithms, XGBoost was significantly better at 20% and higher levels of missing values (figure 2). Therefore, we used the imputed version of the development dataset with 17 features and XGBoost to develop the first COVID-19 mortality prediction model in this study, referred to as the 17F model.

We also tested if a smaller subset of the 17 features could yield an even more accurate prediction model, since such a subset would be easier to study and implement in a clinical setting. Using a setup analogous to the imputation method, we used the recursive feature elimination algorithm and assessed the performance of the four classification algorithms with different numbers of features selected from the full set of 17. We repeated the process 100 times to get an average AUC score for each number of features included in the corresponding candidate prediction models. We found that for the XGBoost algorithm, the AUC became saturated at as few as three features (figure 2). This observation validated our hypothesis that fewer than 17 features could yield an accurate prediction model. The three features identified from the development dataset were minimum oxygen saturation recorded during the encounter, patient age, and type of encounter. We trained this second COVID-19 mortality prediction model, referred to as the 3F model, by applying XGBoost to these three features in the imputed development dataset.

Validation of the 17F and 3F models on test dataset 1 (retrospective data) and test dataset 2 (prospective data) both yielded good performance (AUC score of >0·9; figure 3). Calibration curves[10] of the 17F model's and 3F model's performances on the two test datasets also showed that the models did reasonably well at predicting patient mortality (figure 3). This interpretation is based on the observation that the slopes of all the curves were relatively close to one and the intercepts were close to zero, although the results were better on test dataset 1 than on test dataset 2. The prediction models' strong performance in both test datasets suggests the possibility that COVID-19 mortality predictors constructed from data on a given day can be applied retrospectively and prospectively.

Similar to the features that the 3F model was based on, we identified the three most predictive features for the other classification algorithms we tested (figure 4). For this analysis, we identified the three most predictive features selected in all the 100 training-holdout splits of the development dataset for the four prediction algorithms assessed in figure 2B, and ranked them in terms of their frequency of selection across the 100 runs. Although there was variability among these features due to the inherent differences among the algorithms, the age of the patient and their minimum oxygen saturation level during the clinical encounter were consistent features across the algorithms. The values of minimum oxygen saturation and age were significantly different between the deceased and alive groups in all datasets (table, figure 4), supporting their predictive power. The top three features for each algorithm were consistent when the feature selection and prediction model development process was repeated three times on the development dataset (appendix p 6).

Discussion

We applied machine learning algorithms to clinical and demographic data from patients with COVID-19 from a major New York metropolitan area health system to develop and test a mortality prediction model that showed high accuracy (AUC score of 0·91) when applied to test datasets of retrospective and prospective patient data. This 3F mortality prediction model was based on three clinical features: age, minimum oxygen saturation, and type of patient encounter (inpatient *vs* outpatient and telehealth encounters). Given the heterogeneity in clinical presentation and disease course observed among patients with

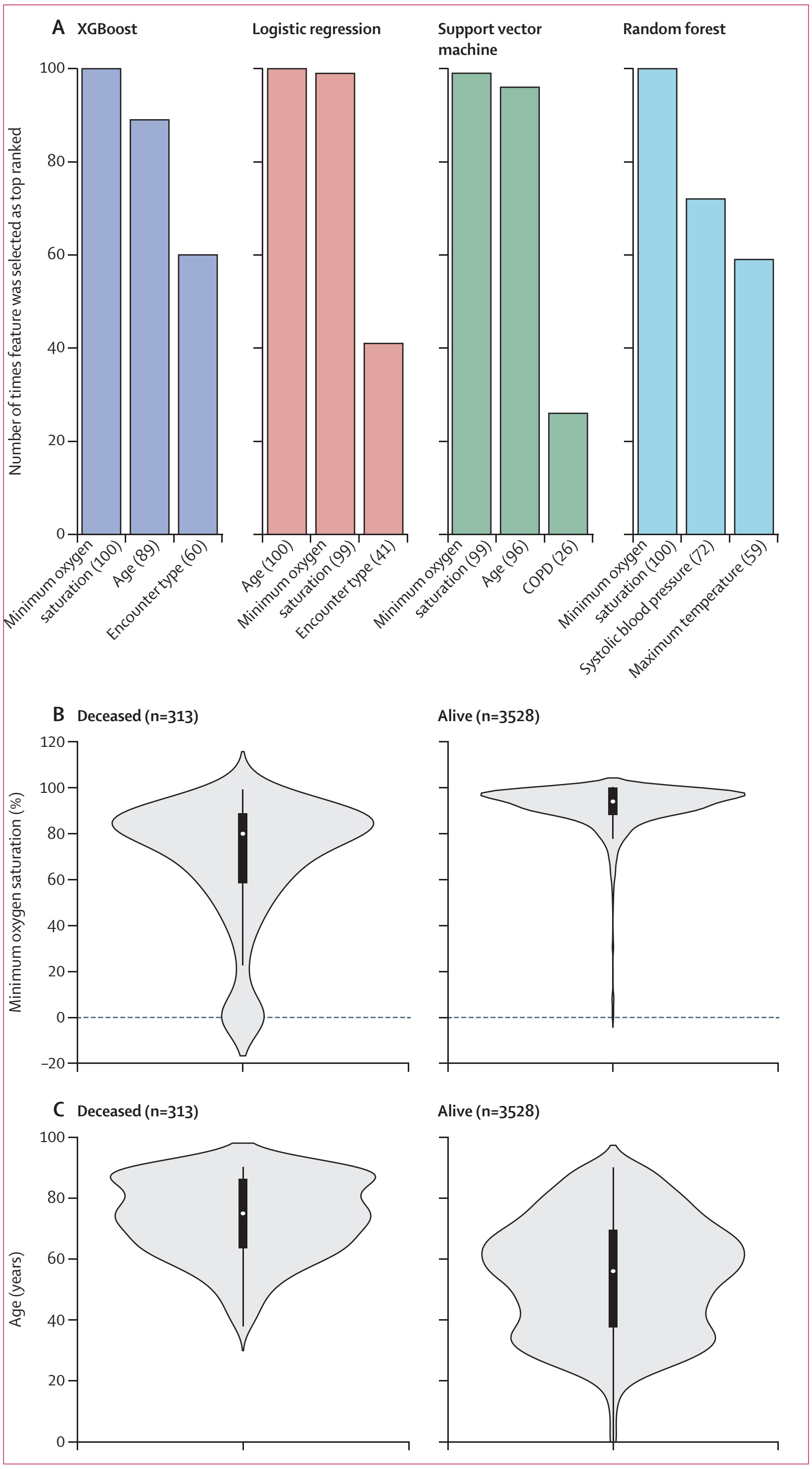

Figure 4: **Top predictive features selected for the four classification algorithms**
(A) Top three predictive features identified using the recursive feature elimination method for the four classification algorithms across the 100 runs used to select the most discriminative features and train the corresponding candidate prediction models; the values in parentheses indicate the number of times the feature was selected as top ranked in the development dataset. Minimum oxygen saturation (B) and age (C) features, which were selected as top predictive features for all the four algorithms, are presented as violin plots showing the distributions of the values in the development dataset. In panels B and C, the black boxplots in the middle show the distribution of the values on the y axis, with the white dot indicating the median value; the width of the grey shape at a given value on the y axis indicates the probability of occurrence of that value in the population shown. The plots in panel B show that the median value (79%) of minimum oxygen saturation for the deceased group was significantly lower (Student's *t* test p<0·0001) than the median value (92%) for the alive group. Similarly, the plots in panel C show that the median age (75 years) in the deceased group is higher (Student's *t* test p<0·0001) than that in the alive group (56 years). COPD=chronic obstructive pulmonary disease.

COVID-19,[2,3] factors that contribute most to mortality are not always readily apparent, rendering care and management of these patients difficult in settings of finite health-care resources. Our work shows that input of three highly accessible clinical parameters for a patient—age, minimum oxygen saturation, and type of patient encounter—into an automatable XGBoost algorithm has the potential to accurately classify patients as likely to live or die. We recognise that external validation of our prediction model in other populations is the next step in model development. Should such validation show that our model performs well in multiple populations, we envision that incorporation of our automatable mortality prediction model into the clinical care workflow of a patient with COVID-19 could yield an additional vital sign that is assessed regularly during a patient's encounter. Clinical teams could use results from the prediction model throughout patients' encounter to flag individuals at high risk of death so that they can promptly focus treatment and attention on such individuals to prevent deaths.

A major strength of this study is that it was based on recent data from thousands of patients with COVID-19 in a global epicentre of the pandemic (New York City), resulting in findings that are highly relevant to the current pandemic. The results are based on rigorous machine learning analyses powered by a robust sample of patients with laboratory confirmed SARS-CoV-2 infection and show the potential of these methods to identify factors predicting mortality in clinical settings. Application of machine learning enabled the identification of prediction models based on the XGBoost algorithm.[11] These prediction models worked with high accuracy (AUC scores of 0·91–0·94) in two independent validation datasets of patients with COVID-19. Furthermore, the 3F model, which was based on only the three features identified, did almost as well as the 17F model, which was based on all the features, that had a level of missing values that was useful for prediction. This finding indicates that accurate mortality predictions can be obtained from a more parsimonious model, facilitating more efficient implementation in clinical environments after extensive validation in other datasets and health systems.

Age and minimum oxygen saturation during encounter were the most predictive features not only for the XGBoost algorithm, but for all four mortality prediction models tested, emphasising these features' epidemiological and clinical relevance. Since the beginning of the COVID-19 pandemic, older age has been recognised as a risk factor for worse outcomes.[12,13] In New York State, USA, patients aged 60 years and older represent nearly 85% of all deaths due to COVID-19 as of Sept 2, 2020,[14] and similarly, higher rates of mortality among those in older age groups than in younger age groups have been noted in other COVID-19 hotspots across the USA.[15] Additionally, the fundamental clinical presentation of patients with COVID-19 during the pandemic has been respiratory symptoms associated with hypoxia, often leading to subsequent respiratory failure and requiring ventilator support, extracorporeal membrane oxygenation, or both.[16] Our finding that a patient's minimum oxygen saturation value during their encounter was the strongest predictive feature of mortality is in line with global epidemiological observations that respiratory failure is the most common feature of critical illness and death in patients with COVID-19.[17,18]

In addition to age and oxygen saturation, health-care encounter type (inpatient *vs* outpatient and telehealth), was identified as a highly indicative feature in the 3F mortality prediction model. This finding reflects the fact that patients with COVID-19 with more severe symptoms are more likely to have their initial encounter in the hospital than in an outpatient setting as their first point of contact. Also, although not the most predictive feature, maximum body temperature during encounter was a top-ranked feature identified by the random forest-based mortality predictor. Although fever is a common symptom and sign of COVID-19, patients might not always present with an increased body temperature, and fever might develop later during the disease course.[2,19] Consistent with this observation, these mortality predictors identified maximum body temperature during encounter, rather than body temperature at presentation, as a top classifying feature.

Several other studies have been published on the investigation of factors affecting mortality due to COVID-19. Some investigators have done statistical association analyses of individual patient characteristics and risk factors with mortality, albeit on small cohorts (<200 patients).[20–23] Another small cohort study used linear feature selection and prediction model development methods to identify severe cases of COVID-19, with an AUC of 0·853 in a validation cohort of 165 patients.[24] Some other studies have started leveraging clinical data from larger cohorts of several hundred patients to predict mortality and other COVID-19 outcomes.[25] A relative strength of our study is that we used a large patient cohort and systematic combinations of machine learning methods to develop a more accurate and informative mortality prediction model. In particular, the comparatively larger set of predictor variables compared with previous studies and the recursive feature elimination selection method we used in our study provided an opportunity to automatically identify accurate and parsimonious sets of variables and prediction models. Due to the size of the cohort and prediction methods used, these datasets and models were likely to perform better than previously proposed methods for this problem.

Machine learning-based methods are designed to sift through large amounts of structured or unstructured data to discover actionable knowledge without bias from biomedical hypotheses.[4,26] In this study, we used this power of machine learning, especially for feature selection[7] and classification,[4] to develop accurate and parsimonious prediction models of mortality from COVID-19 from structured clinical and demographic data. In particular, we found that the XGBoost[11] algorithm produced the most accurate prediction models. XGBoost is a sophisticated prediction algorithm that builds an ensemble of decision trees by iteratively focusing on harder to predict subsets of the training data. Due to its systematic optimisation-based design, this algorithm has shown superior performance in predictive modelling applications involving structured data,[27,28] which is consistent with our observations.

Our study had several limitations. Although our datasets are probably the largest that have been used to predict COVID-19 mortality, the clinical features available to us were limited to those routinely collected during hospital encounters. Although we were able to develop accurate prediction models from these limited data using our machine learning framework, development of even better prediction models should be possible using a richer set of features. In the future, development of more accurate prediction models for COVID-19 mortality and other outcomes should be possible via integration of multimodal data collected from patients. These data include demographics, comorbidities, laboratory test measurements, vital signs, chest imaging, clinical notes, and omic data, and can be integrated into prediction models using techniques like heterogeneous ensembles[29] and deep learning.[30] Our study data were also limited in several other aspects. First, although our development and validation datasets were larger than previous studies, some of them were small in size. Specifically, test dataset 2 included only 249 patients, with only 25 patients who died. Second, our datasets only represent a snapshot in time, and mortality outcomes might change in different timeframes. For instance, test datasets 1 and 2 contained data from patients with COVID-19 who had encounters in our health system during the period March 9–April 5, 2020, and on April 6, 2020. Changing these date ranges could have changed their respective mortality rates. Similar changes could also occur for the values of time-varying features like minimum oxygen saturation. Factors like these could have affected our prediction results. Finally, although our development dataset, test dataset 1, and test dataset 2 were generated as systematically and without bias as possible, significant differences existed between them in terms of feature values and mortality rates.

A key limitation of the clinical indices included in the datasets include the uniformity of EMR-derived data. For example, although minimum oxygen saturation during encounters was identified as a key predictor for mortality, limitations inherent in the interpretation of these data must be noted, such as the unavailability of the amount of supplemental oxygen being administered at the time of recording and acquisition-related limitations, such as readings below the threshold of accuracy of the monitoring device (eg, <70%). Nonetheless, we found a clearly lower distribution of minimum oxygen saturation levels in patients who died from COVID-19 compared with those who were alive, highlighting this clinical feature as central to predicting mortality. Furthermore, since EMR systems are not fully synchronised across health systems yet, another health system's data might not be consistent with ours. Thus, appropriate modifications to another health system's EMR data might be needed to apply our approach. We expect that the details in the Methods and appendix will assist in these modifications and applications.

Applying machine learning approaches to data from a large cohort of patients with COVID-19 resulted in the identification of accurate and parsimonious prediction models of mortality. After extensive validation in other datasets and health systems, these data-driven findings might help clinicians better recognise and prioritise the care of patients at greatest risk of death.

Contributors

GP, SBu, RI, and ASY conceived the study. ASY and Y-cL managed and analysed the data. SBo provided clinical expertise. GP, SBu, and RI supervised the work. All authors interpreted the results. GP, SBu, RI, and SBo wrote the manuscript, which was approved by all the authors.

Declaration of interests

SBo, RI, SBu, and GP have received research funding from the US National Institutes of Health (NIH). GP has also received funding from the Intelligence Advanced Research Projects Activity Agency. SBo has also received funding from Mount Sinai Genomics. RI has also received research funding from the US Food and Drug Administration and the New York State Department of Health. All other authors declare no competing interests.

Data sharing

Due to IRB restrictions, we are unable to publicly share the study data. The analysis source code for this study is available online.

For the **source code** see https://github.com/SBCNY/Clinical-predictors-of-COVID-19-mortality

Acknowledgments

ASY and RI receive funding from NIH (U54 HG008098 and P50 GM071558). SBu receives funding from NIH (R01 AI118833, R01 AI147028, and U19 AI136053). SBo receives funding from NIH (R01 HL147328 and UG3 OD023337). This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai (New York City, NY, USA). We thank Sharon Nirenberg and Patricia Kovatch, both of the Icahn School of Medicine at Mount Sinai, for their clarifications regarding the EMR data.

References

1 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; **20:** 533–34.

2 Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med* 2020; **382:** 1708–20.

3 Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* 2020; **323:** 2052–59.

4 Alpaydin E. Introduction to machine learning, 3rd edn. Cambridge, MA: MIT Press, 2014.

5 US Centers for Disease Control and Prevention. Health effects of second-hand smoke. Atlanta GA: US Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/secondhand_smoke/health_effects/index.htm (accessed Aug 17, 2020).

6 Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)* 2013; **1:** 1035.

7 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002; **46:** 389–422.

8 Demsar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006; **7:** 1–30.

9 Lever J, Krzywinski M, Altman N. Classification evaluation. *Nat Methods* 2016; **13:** 603–04.

10 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17:** 230.

11 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; San Francisco, CA: Association for Computing Machinery; 2016: 785–94. https://dl.acm.org/doi/10.1145/2939672.2939785 (accessed Sept 4, 2020).

12 Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in china: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* 2020; **323:** 1239–42.

13 Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular disease, drug therapy, and mortality in COVID-19. *N Engl J Med* 2020; **382:** 2582.

14 COVID-19 tracker: fatalities. Albany, NY: New York State Department of Health, 2020. https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-Fatalities?%3Aembed=yes&%3Atoolbar=no&%3Atabs=n (accessed Sept 2, 2020).

15 Bhatraju PK, Ghassemieh BJ, Nichols M, et al. COVID-19 in critically ill patients in the Seattle region—case series. *N Engl J Med* 2020; **382:** 2012–22.

16 Prekker ME, Brunsvold ME, Bohman JK, et al. Regional planning for extracorporeal membrane oxygenation allocation during COVID-19. *Chest* 2020; **158:** 603–07.

17 Chen T, Wu D, Chen H, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ* 2020; **368:** m1091.

18 Grasselli G, Zangrillo A, Zanella A, et al. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy region, Italy. *JAMA* 2020; **323:** 1574–81.

19 Zavascki AP, Falci DR. Clinical characteristics of COVID-19 in China. *N Engl J Med* 2020; **382:** 1859–62.

20 Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med* 2020; **46:** 846–48.

21 Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; **395:** 1054–62.

22 Gao L, Jiang D, Wen XS, et al. Prognostic value of NT-proBNP in patients with severe COVID-19. *Respir Res* 2020; **21:** 83.

23 Du R-H, Liang L-R, Yang C-Q, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: A Prospective Cohort Study. *Eur Respir J* 2020; **55:** 2000524.

24 Gong J, Ou J, Qiu X, et al. A tool to early predict severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. *Clin Infect Dis* 2020 published online April 16. https://doi.org/10.1093/cid/ciaa443.

25 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.

26 Cleophas TJ, Zwinderman AH. Machine learning in medicine—a complete overview. Springer, 2016.

27 Morde V, Setty VA. XGBoost algorithm: long may she reign! Towards Data Science, April 8, 2019. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d (accessed Sept 2, 2020).

28 Reinstein I. XGBoost, a top machine learning method on Kaggle, explained. KD Nuggets, Oct 4, 2017. https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html (accessed Sept 2, 2020).

29 Wang L, Law J, Murali TMN, Pandey G. Data integration through heterogeneous ensembles for protein function prediction. *bioRxiv* 2020; published online May 31. https://doi.org/10.1101/2020.05.29.123497v1 (preprint).

30 Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; **15:** 20170387.