An ensemble prediction model for COVID-19 mortality risk

Jie Li1,*, Xin Li1, John Hutchinson2, Mohammad Asad Shoaib2, Yinghui Liu1, Yadong Wang1 and Edwin Wang2,3,*

1School of Computer Science and Technology, Harbin Institute of Technology, 92 Xidazhi Street, Nangang District, Harbin, Heilongjiang 150006, China
2Department of Medical Genetics, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada
3Department of Medicine, McGill University, 845 Sherbrooke Street West, Montreal, Quebec H3A 0G4, Canada

*Correspondence address: (J.L.) School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. E-mail: jieli@hit.edu.cn. (E.W.) Department of Medical Genetics, University of Calgary, Calgary, Canada. Tel: 1-403-220-6609; E-mail: edwin.wang@ucalgary.ca

Abstract

Background: It’s critical to identify COVID-19 patients with a higher death risk at early stage to give them better hospitalization or intensive care. However, thus far, none of the machine learning models has been shown to be successful in an independent cohort. We aim to develop a machine learning model which could accurately predict death risk of COVID-19 patients at an early stage in other independent cohorts.

Methods: We used a cohort containing 4711 patients whose clinical features associated with patient physiological conditions or lab test data associated with inflammation, hepatorenal function, cardiovascular function, and so on to identify key features. To do so, we first developed a novel data preprocessing approach to clean up clinical features and then developed an ensemble machine learning method to identify key features.

Results: Finally, we identified 14 key clinical features whose combination reached a good predictive performance of area under the receiver operating characteristic curve 0.907. Most importantly, we successfully validated these key features in a large independent cohort containing 15 790 patients.

Conclusions: Our study shows that 14 key features are robust and useful in predicting the risk of death in patients confirmed SARS-CoV-2 infection at an early stage, and potentially useful in clinical settings to help in making clinical decisions.

Keywords: COVID-19; SARS-CoV-2; mortality prediction; prognosis; cohort studies

Background

The global COVID-19 pandemic is putting high pressure on healthcare systems around the world [1–3]. Most of people infected with the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) have mild disease and self-limiting, but there is still a significant proportion of patients who develop severe disease that may result in death [4, 5]. In epidemic areas, the shortage of medical resources may lead to an increase in mortality [6]. Therefore, it is important to distinguish patients at high risk of severe illness or mortality from others in the early stages of disease development.

There have been several researchers contributing to the areas of mortality risk prediction for patients. In May 2020, Yan et al. [7] selected three features that predict the mortality of individual patients more than 10 days in advance through machine learning tools. However, Yan et al. gathered samples from a cohort of only 485 patients with confirmed SARS-CoV-2 infection may not be sufficient, and their mortality predictive model may not perform well in other cohorts [8]. Comorbidity, features for kidney disease, and other clinical characteristics have also been associated with the severity of a patient’s disease, according to previous studies [9–12]. Liang et al. [13] proposed a deep learning-based survival model that can predict the risk of COVID-19 patients developing critical illness based on clinical characteristics at admission. The deficiency of this study is that the features selected may not sufficient to reflect the patient’s condition, which might be the reason for the differences in the performance of different validation sets [13]. Based on clinical information, Altschul et al. [14] proposed a novel severity score to assess the severity of patients infected with the SARS-CoV-2. Patients were classified into low (0–3), moderate (4–6), and high (7–10) COVID-19 severity scores (CSSs). A receiver operating characteristic (ROC) curve analysis showed that the area under the ROC curve (AUC) of the derivation cohort was 0.824 and the AUC of the validation cohort was 0.798 [14].

Except these examples, thus far, lots of similar efforts have been made by others. Recently, Wynants et al. [15] conducted a comprehensive and systematic review of 145 prediction models from 107 studies, with a brief summary of the features (predictors) used by these models. The key message from this analysis was that none of the models can be validated independently (i.e. their predictions failed when validating in an independent...
cohort), in another word, none of the predictive models developed in the COVID-19 domain could be used in clinics for decision marking. The prediction power of most of the models was similar to that of flipping a coin [15]. The situation has changed though with the introduction of ensemble models (EMs). These models extract data features through a variety of prediction methods, and then use a variety of machine learning algorithms to combine weak prediction results of individual methods and integrate these results with various voting mechanisms. When the number of samples is large, the EMs usually have higher predictive accuracy than individual single models [16] because the complementary information of each model is effectively utilized. Schapire [17] confirmed the feasibility by incorporating multiple weak learning models into a high-precision model in early years. In this study, we develop an EM that accurately predicts mortality risk of COVID-19 patients at an early stage of infection just based on several medical tests performed during their admission to a hospital.

Materials and methods

The flow chart of our prediction method is shown in Fig. 1. We first group the features and preprocess them, including processing extreme values and imputing missing values. Then, feature selection is carried out for constructing ensemble model (EM). Five base models: gradient-boosted decision tree (GBDT), extreme gradient boosting (XGBoost) [18], random forest (RF), logistic regression (LR), and support vector machine (SVM) are used to select features. Features selected by more than half (three or more) of the base models are used to construct the EM. Finally, performance of the EM and base models is compared and validated on independent datasets.

Datasets

Two datasets are used in the study: Cohort 1 and Cohort 2. Cohort 1 with 4711 COVID-19 patients (1148 deaths) is from a recent study [14], which was collected from 1 March 2020 to 16 April 2020. The mortality rate is 24.3% (1148/4711) in Cohort 1. All patients in Cohort 1 were hospitalized and their clinical features were obtained at admission [14]. These clinical features must be entered later. Clinical features include patient’s age, mean arterial pressure (MAP), oxygen saturation (OsSats), etc., and the details and statistical information of these clinical features are shown in Table 1 and Supplementary Table S1. According to the types and meanings of clinical features (Table 1), numerical features that can directly reflect the physiological conditions of patients are selected features of models. These clinical features include age, OsSats, temperature (Temp), MAP, D-dimer (Ddimer), platelets (Plts), international normalized ratio (INR), blood urea nitrogen (BUN), creatinine, sodium, glucose, aspartate aminotransferase (AST), alanine aminotransferase (ALT), white blood cells (WBCs), lymphocytes (Lympho), interleukin-6 (IL-6), ferritin, C-reactive protein (CrtProtein), procalcitonin, and troponin. If COVID-19 patient dies, his or her label is set to 1, otherwise it will be set to 0. We select features and trained our models on Cohort 1. Cohort 2 is an independent validation data containing 15 790 COVID-19 patients from UK Biobank [19, 20]. The statistical results of clinical features and population structure of this data are shown in Supplementary Table S2. The mortality rate is 4.21% (664/15 790) in Cohort 2. We selected hundreds of features (which are identical to or functionally related to the features selected in Cohort 1) from Cohort 2, and divided them into 55 functionally related features (Supplementary Table S3). These features included age, blood pressure-related features (such as hypertension), kidney function-related features (such as creatinine), inflammation-related features (such as monocyte), and so on.

Feature grouping and feature preprocessing

The presence of missing/error data will reduce the performance of the predictive model. Therefore, we developed a novel feature grouping and preprocessing method to deal with missing/error data. We hypothesized that features closely related to patient’s physiological conditions are better indicators associated with death risk. Thus, a key concept in this study is to select features and group them based on a particular aspect of the patient’s
Table 1: Clinical features of patients infected with SARS-CoV-2 in Cohort 1

| Features                          | Description of features | Types of feature attributes | Feature groups | Impute method                |
|-----------------------------------|-------------------------|----------------------------|----------------|------------------------------|
| LOS                               | Length of hospital stay | Numerical variables        | No group       | Leave untreated              |
| Black                             | Race information        | Binary variables           | No group       | Leave untreated              |
| White                             | Race information        | Binary variables           | No group       | Leave untreated              |
| Asian                             | Race information        | Binary variables           | No group       | Leave untreated              |
| Latino                            | Race information        | Binary variables           | No group       | Leave untreated              |
| MI                                | Myocardial infarction   | Binary variables           | No group       | Leave untreated              |
| PVD                               | Peripheral vascular disease | No group   | Leave untreated              |
| CHF                               | Congestive heart failure | No group                  | Leave untreated |
| CVD                               | Cardiovascular disease  | No group                  | Leave untreated |
| DEMENT                            | Dementia                | No group                  | Leave untreated |
| COPD                              | Chronic obstructive pulmonary disease | No group | Leave untreated |
| DM complicated                    | Diabetes mellitus complicated | No group | Leave untreated |
| DM simple                         | Diabetes mellitus simple | No group                  | Leave untreated |
| Renal Disease                     | Renal disease           | No group                  | Leave untreated |
| Stroke                            | Stroke                  | No group                  | Leave untreated |
| Seizure                           | Seizure                 | No group                  | Leave untreated |
| Age                               | Age                     | Numerical variables        | Independent feature group 1 | Leave untreated |
| OsSats                            | Oxygen saturation       | Numerical variables        | Independent feature group 2 | Imputed by the mean |
| Temp                              | Temperature             | Numerical variables        | Independent feature group 3 | Imputed by the mean |
| MAP                               | Mean arterial pressure  | Numerical variables        | Independent feature group 4 | KNN imputing in the same group |
| Ddimer                            | D-dimer                 | Numerical variables        | Cardiovascular group | KNN imputing in the same group |
| Plts                              | Platelets               | Numerical variables        | Cardiovascular group | KNN imputing in the same group |
| INR                               | International normalized ratio | Numerical variables | Cardiovascular group | KNN imputing in the same group |
| Troponin                          | Troponin                | Numerical variables        | Cardiovascular group | KNN imputing in the same group |
| BUN                               | Blood urea nitrogen     | Numerical variables        | Hepatorenal group | KNN imputing in the same group |
| Creatinine                        | Creatinine              | Numerical variables        | Hepatorenal group | KNN imputing in the same group |
| Sodium                            | Sodium                  | Numerical variables        | Hepatorenal group | KNN imputing in the same group |
| Glucose                           | Glucose                 | Numerical variables        | Hepatorenal group | KNN imputing in the same group |
| Ferritin                          | Ferritin                | Numerical variables        | Hepatorenal group | KNN imputing in the same group |
| AST                               | Aspartate aminotransferase | Numerical variables | Hepatorenal group and cardiovascular group | KNN imputing in the same group |
| ALT                               | Alanine aminotransferase | Numerical variables        | Hepatorenal group and cardiovascular group | KNN imputing in the same group |
| WBC                               | While blood cells       | Numerical variables        | Inflammatory group | KNN imputing in the same group |
| Lymphoto                          | Lymphocytes             | Numerical variables        | Inflammatory group | KNN imputing in the same group |
| IL6                               | Interleukin-6           | Numerical variables        | Inflammatory group | KNN imputing in the same group |
| Creactive protein                 | C-reactive protein      | Numerical variables        | Inflammatory group | KNN imputing in the same group |
| Procalcitonin                     | Procalcitonin           | Numerical variables        | Inflammatory group | KNN imputing in the same group |
| All CNS                           | No introduction found   | Binary variables           | No group       | Leave untreated              |
| Pure CNS                          | No introduction found   | Binary variables           | No group       | Leave untreated              |
| OldSyncpe                         | No introduction found   | Binary variables           | No group       | Leave untreated              |
| OldOtherNeuro                     | No introduction found   | Binary variables           | No group       | Leave untreated              |
| OtherBrnLsn                       | No introduction found   | Binary variables           | No group       | Leave untreated              |
| Derivation cohort                 | Grouping in the original literature [14] | No group | Leave untreated |
| Death                             | Whether the patient died or not | Binary variables | No group | Leave untreated |
| Severity                          | COVID-19 severity (score given in the original literature) [14] | Binary variables | No group | Leave untreated |

* No group means that the feature is not used in this study.
KNN is an old and simple data imputation method, it has shown better imputation quality among above various data imputation algorithms [32]. Pujianto et al.'s experimental results shown handling missing data with KNN-based imputation can reach the accuracy of complete data [33]. Mean, median, constant 0, and chained equation methods are also widely used to impute missing data. In this study, we compared mean, median, constant 0, chained equation, and KNN, among which the KNN has the best performance (see Supplementary Table S5). Therefore, KNN is used in the study. For KNN, theoretically, if there is no noise in the data, the smaller the value of k, the more accurate the prediction result for an unknown point. Because the nearest two points (i.e. two points are nearest in the feature space) may have the most similar features, namely, the nearest point is the most accurate to impute the missing value. However, the data are subject to various disturbances (such as noise, extremes, measurement errors), and the point closest to the noise point may be inaccurate and deviate from the actual value. To reduce these errors, we selected k as 1, 2, 3, … and test the performance of models and found that when k is 3, the model performed better, therefore k is set as 3 in the study. We take WBC in inflammatory group as an example to illustrate how to impute the missing WBC value. First, patients were divided into two groups (death and survival groups) according to whether the patient died or not. Then, patients in the same group were clustered using KNN according to four features: Lympho, IL6, CrctProtein, and Procalcitonin. If WBC value of certain patient is missing, it is imputed by the average WBC value of three nearest patients to the patient. For AST and ALT, we imputed their missing values using hepatorenal group. Although both AST and ALT are associated with both hepatorenal conditions and cardiovascular conditions, they are the most commonly used indicator of liver function, often part of a routine blood screening to check the health of patient’s live. A meta-analysis with 12 882 confirmed COVID-19 patients suggests that AST and ALT are significantly associated with COVID-19 severity (P < 0.00001) [34]. For these reasons, AST and ALT are more closely related to hepatorenal status compared with cardiovascular status; therefore, the missing values of AST and ALT are based on imputation results of hepatorenal group in the study.

### Base model parameter settings

In order to select valuable features and develop an EM which could take advantages of several base models such as GBDT, XGBoost [18], RF, LR, and SVM, we first conducted experiments in Cohort 1 and set the best parameters for five base models, respectively. For RF, GBDT, and XGBoost, we adjusted the number of decision trees (n_estimators), and for XGBoost, we also adjusted the maximum depth (max_depth) and the subsample ratio of features (colsample_bytree) to control over-fitting or under-fitting when constructing each tree. For the SVM model, we chose the radial basis function (RBF) (kernel) as the kernel function, and the regularization parameter C (C) was set to 0.7 to reduce overfitting. Since SVM favors the majority class on unbalanced datasets, we adjusted the weights of the two classes inversely proportional to the frequency of the classes (class_weight) in the dataset. In addition, z-score was used to standardize the data before input into LR and SVM models due to the characteristics of the algorithm. The detail of parameters of five base models is shown in Table 1.

### Feature selection for EM

Redundant features could be detrimental to predictive models to make correct predictions. Therefore, we screened out most valuable features from the clinical features to improve the

---

Table 2: The maximum in the original data and the selected cutoff95

| Feature      | Maximum  | Cutoff95 |
|--------------|----------|----------|
| Ddimer       | 20.00001 | 20.00001 |
| Plts         | 1226     | 433      |
| INR          | 17.0001  | 1.7      |
| BUN          | 301      | 97       |
| Creatinine   | 31.65    | 7.35     |
| Sodium       | 170.001  | 153      |
| Glucose      | 1000.001 | 423.6    |
| AST          | 10.000   | 159      |
| ALT          | 3228     | 116      |
| WBC          | 219.7    | 16.9     |
| Lympho       | 209.1    | 2.4      |
| IL-6         | 111.040  | 372.74   |
| Ferritin     | 100.000  | 4508.55  |
| CrctProtein  | 100.0001 | 34.3     |
| Procalcitonin| 50.0001  | 12.52    |
| Troponin     | 9.56     | 0.21     |

---

physiology, but not strictly based on medical definitions. By doing so, age, OsSats, Temp, and MAP were divided into independent feature groups. Independent features are those that are not significantly associated with other. Other clinical features such as BUN, creatinine, glucose, sodium, ferritin, AST, and ALT could indicate the conditions of liver or kidney [21–25]; therefore, such as BUN, creatinine, glucose, sodium, ferritin, AST, and ALT, we imputed their missing values using hepatorenal group in the study. Troponin 9.56 0.21 Procalcitonin 50.0001 12.52 CrctProtein 100.0001 34.3 Ferritin 100 000 4508.55 Lympho 209.1 2.4 WBC 219.7 16.9 Sodium 170.001 153 Creatinine 31.66 7.35 BUN 301.97 INR 17.0001 1.7
performance of our predictive models. The feature selection process is divided into two steps. In the first step, we select high performance feature set for five base models, respectively, from the 20 features in Table 1. In the second step, we combine feature sets of five base models to form the final selected feature set.

Genetic algorithm (GA) is used to select feature set [35], which is a heuristic search algorithm that simulates the process of natural selection. In the feature selection process, the AUC of each base model is taken as the objective function, each individual in GA represents a set of features, consisting of a binary string called a chromosome, and multiple individuals constitute a population. In each generation, a subset of individuals with the highest fitness (maximizing the objective function) goes into the next generation. In this way, we finally select a feature set that makes the predictive model get the best performance.

Taking the feature selection for GBDT as an example, the population size is set to 40, and each chromosome is encoded into a binary string of length 20. Each position of the chromosome represents whether the corresponding feature is selected or not. We used the elite-tournament method [36] as a selection operator to select the chromosomes with the highest fitness in the population. The single-point crossover operator is chosen as the offspring chromosome recombination method, the crossover probability is set to 0.7 and the probability of offspring mutation is the reciprocal of chromosome length. According to the above settings, after running 200 generations, the high performance feature set is selected for GBDT.

We also select high performance feature set for other base models. Thus, each base model has a high performance feature set. We combine these feature sets and select features that appear in more than half (three or more) of the feature sets to form the final selected feature set.

### EM construction

In order to construct an EM, which could take advantages of the prediction results of several models, we chose the above five models as base models to construct our EM. Similar to the feature selection method mentioned above, we also used GA to find a set of coefficients $C$ as the weight of the prediction results of the five base models. The prediction results of EM ($\text{prob}_{em}$) for patients are the weighted average of the prediction results of each base model ($\text{prob}$), as defined below:

$$\text{prob}_{em} = \frac{\sum C \ast \text{prob}}{\sum C},$$  \hspace{1cm} (1)

We used 0.5 as the threshold, and patient whose prediction result is higher than the threshold is predicted as dead. To obtain this set of coefficients $C$ using the GA, fitness score is calculated using the AUC of the EM. We code each set of coefficients for five base models as a binary string (i.e. a chromosome), which length is 70. The number of individuals (chromosomes) in the population is set to 40. Other parameters of GA, such as selection operators and crossover operators, are the same as those used in feature selection process.

### Performance evaluation of predictive models

In this study, we evaluate the prediction performance of different models using accuracy, AUC, precision, and recall. The definitions of accuracy, precision, and recall are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$  \hspace{1cm} (2)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$  \hspace{1cm} (3)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$  \hspace{1cm} (4)

Here, TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative, respectively. In this study, the death patient is positive sample. The predictive model gives the probability of death for each patient, and we set a threshold of 0.5, above which the patient is considered to be dead. Using clinical data composed of selected features as inputs and patient status (alive or death) as labels, we performed half-half cross-validations 100 times for each model (including EM and base models) on the entire cohort (Cohort 1, $n=4711$). Specifically, we first randomly chose half of the samples (2355 patients) as training set and the other half (2356 patients) as test set from Cohort 1. Then, we trained EM and base models on training set and tested the performance of these models on test set. Subsequently, training set and test set were exchanged (i.e. the former training set was the test set and the former test set was the training set), and these models were retrained and tested. At the same time, prediction results of these models were saved. Above process repeats 100 times, average values and standard deviations (SD) of accuracy, AUC, precision and recall of EM, and five base models were calculated.

### Validation on independent dataset

Cohort 2 is an independent cohort containing 15 790 patients (664 patients died and 15 126 patients survived). Since Cohort 2 did not fully contain the features in Cohort 1, we selected 55 features in Cohort 2 that were identical or functionally related to features selected from Cohort 1 (as shown in Supplementary Table S3 in this study) for further analysis. The missing data in Cohort 2 are imputed according to the method used in Cohort 1. In order to validate the performance of different models in Cohort 2, we first used GA to select a feature set that was most related to the mortality risk of patients from 55 features. Cohort 2 and Cohort 1 are different in several aspects, such as mortality rate, age distribution, and so on (Supplementary Fig. S1 and Supplementary Tables S1 and S2). For comparison purposes, next, we selected a subset of patients aged 50–84 years from Cohort 1 (Subset 1), since the patients in Cohort 2 ranged in age.

---

**Table 3: Parameter settings of five base models**

| GBDT          | XGBoost       | RF            | LR            | SVM          |
|---------------|---------------|---------------|---------------|--------------|
| random_state  | random_state  | random_state  | random_state  | random_state |
| learning_rate | learning_rate | n_estimators  | n_estimators  | n_estimators |
| n_estimators  | 10            | 10            | 10            | 10           |
| 0.1           | 0.1           | 150           | 200           | 10           |
| 12            | 0.3           | 12            | 0.7           | 0.1          |
| class_weight  | "balanced"   | "rbf"        | "rbf"        | "rbf"       |
| C             | 0.7           | 0.1           | 0.1           |              |
| probability   | True          | True          | True          |              |

---

Li et al. | 5
from 50 to 84 years. As with Cohort 1, the data in subset 1 were also half-and-half cross-validation for 100 times. Subsequently, we randomly sampled a subset of patients (Subset 2) with a similar percentage of patients surviving in Subset 1 from Cohort 2 and run a half-and-half cross-validation on Subset 2. The process also repeats 100 times. Finally, we test different models on Subset 1 and Subset 2 and calculated their accuracy, AUC, precision, and recall.

**Results**

We first selected feature set for each base model in Cohort 1 according to our method. Feature sets of different base model were shown in Supplementary Table S4. Then, 14 key features were chose as the final feature set, which are Age, OsSats, MAP, Ddimer, Glucose, WBC, Lympho, IL-6, CrtctProtein, Procalcitonin, Troponin, Plts, INR, and ALT. Finally, EM was constructed and the coefficients of base models of the EM calculated by GA were $0.39620338$ (GBDT), $0.9574559$ (XGBoost), $0.26222304$ (RF), $0.0315571$ (LR), and $0.24549838$ (SVM), respectively. Mean values and SD of accuracy, AUC, precision and recall of EM, and five base models were shown in Table 4. Experimental results indicated that feature preprocessing and selection significantly improved the performance of the predictive models. In addition, the EM reached the best performance in unprocessed and preprocessed data, which showed the robustness of the EM. We also conducted an experiment in Cohort 1 using different imputation methods and test the performance of different models. Experimental results on Cohort 1 suggested that KNN imputation method was best and improved the performance of the predictive models than simply replacing missing data with 0 in the original data (as shown in Supplementary Table S5).

We calculated the mean value of the prediction results of the EM for each patient in 100 rounds to study the changes in precision and recall of EM when the threshold changed from 0 to 1 (Fig. 2). With the increasing of the threshold, the precision had a trend of rapidly increasing at first and then slowly increasing, and correspondingly, the recall had a trend of slowly decreasing at first and then rapidly declining. Our goal was to find a reasonable range of thresholds in which the precision and recall can have a practical value. We selected the threshold (0.24) when the recall reached 0.8 and the threshold (0.46) when the precision reached 0.8, and marked it with a dashed line in Fig. 2. The precision and recall under these thresholds were $0.656998$ and $0.63676$, respectively.

To assess the impact of each feature on mortality in patients, we calculated the mean value of the importance of each feature

| Table 4: Performance results of different models on Cohort 1 |
|----------------------------------------|
| Unprocessed data in Cohort 1 | GBDT | XGBoost | RF | LR | SVM | EM |
|----------------------------------------|
| Accuracy (SD) | 0.834(0.005) | 0.830(0.004) | 0.832(0.004) | 0.811(0.005) | 0.813(0.006) | 0.837(0.004) |
| AUC (SD) | 0.847(0.007) | 0.844(0.007) | 0.848(0.007) | 0.803(0.007) | 0.826(0.007) | 0.854(0.006) |
| Precision (SD) | 0.736(0.020) | 0.750(0.019) | 0.754(0.020) | 0.707(0.020) | 0.688(0.024) | 0.772(0.020) |
| Recall (SD) | 0.495(0.020) | 0.434(0.018) | 0.460(0.020) | 0.385(0.019) | 0.424(0.021) | 0.471(0.019) |

| Data in Cohort 1 after preprocessing and feature selection | GBDT | XGBoost | RF | LR | SVM | EM |
|----------------------------------------|
| Accuracy (SD) | 0.864(0.005) | 0.864(0.005) | 0.862(0.005) | 0.847(0.004) | 0.855(0.006) | 0.868(0.005) |
| AUC (SD) | 0.900(0.005) | 0.904(0.005) | 0.900(0.005) | 0.870(0.006) | 0.890(0.005) | **0.907(0.005)** |
| Precision (SD) | 0.774(0.018) | 0.805(0.019) | 0.791(0.019) | 0.764(0.018) | 0.738(0.020) | 0.804(0.019) |
| Recall (SD) | 0.625(0.016) | 0.582(0.017) | 0.588(0.017) | 0.542(0.018) | 0.629(0.021) | 0.605(0.016) |

Bold values indicate best performance.
in the XGBoost over 100 rounds of predictions (Fig. 3). MAP, IL-6, and Procalcitonin contributed the most to the decision of XGBoost predictive model. Other features such as Ddimer, age, and CrctProtein also played an important role in the prediction model.

We further explored whether a single clinical feature could be used to stratify patients for mortality risk. To do so, we selected the first six clinical features which have a higher importance. Patients were divided into 10 groups according to the value range of each clinical feature and the number of patients in each group was approximately equal. Patients for which procalcitonin is from 0.099 to 0.1 were one group because the interval value is too small. The mortality rate in Cohort 1 was 0.244 (1148/4711). The mortality rate of each group for six clinical features is shown in Fig. 4.

From Fig. 4, we can see that when these clinical features: MAP < 79.67 mmHg, IL-6 > 72.8 pg/ml, procalcitonin > 0.5 ng/ml, Ddimer > 2.4 mg/ml, age > 69 years, and Glucose > 174.0 mg/dL, patients have higher death risk.

Experimental results on an independent dataset

Finally, we further validate our predictive models on an independent cohort (Cohort 2). First, we selected a set of features (Supplementary Table S6), and then compared the performance of the predictive models in the corresponding subsets of the two cohorts, as described in the “Materials and methods” section. The results of our predictive model on two subsets are shown in Table 5. In general, these predictive models still performed well in a large dataset with similar features. EM performs best in three of the four performance metrics, which indicates that EM has a good robustness.

Comparison with CSSs

Altschul et al. [14] collected COVID-19 patients’ clinical information (Cohort 1) and also developed a predictive model (i.e. CSS). Therefore, it is possible to directly compare EM and the CSS. In CSS, patients were classified into low risk (0–3 points), moderate...
Table 5: Performance of different models on an independent dataset

| Subset 1 | GBDT | XGBoost | RF | LR | SVM | EM |
|----------|------|---------|----|----|-----|----|
| Accuracy (SD) | 0.850(0.006) | 0.849(0.006) | 0.850(0.006) | 0.828(0.006) | 0.838(0.006) | 0.854(0.005) |
| AUC (SD)     | 0.884(0.007) | 0.888(0.007) | 0.887(0.007) | 0.846(0.008) | 0.875(0.007) | 0.893(0.007) |
| Precision (SD) | 0.764(0.020) | 0.797(0.021) | 0.788(0.021) | 0.749(0.020) | 0.723(0.021) | 0.799(0.020) |
| Recall (SD)  | 0.616(0.022) | 0.565(0.023) | 0.583(0.023) | 0.515(0.020) | 0.616(0.025) | 0.588(0.022) |

| Subset 2 | GBDT | XGBoost | RF | LR | SVM | EM |
|----------|------|---------|----|----|-----|----|
| Accuracy (SD) | 0.803(0.009) | 0.804(0.009) | 0.803(0.009) | 0.804(0.009) | 0.785(0.009) | 0.810(0.009) |
| AUC (SD)     | 0.858(0.009) | 0.863(0.010) | 0.860(0.009) | 0.862(0.009) | 0.847(0.009) | 0.873(0.008) |
| Precision (SD) | 0.644(0.023) | 0.667(0.029) | 0.662(0.025) | 0.654(0.024) | 0.612(0.024) | 0.684(0.026) |
| Recall (SD)  | 0.533(0.030) | 0.461(0.029) | 0.507(0.031) | 0.533(0.027) | 0.500(0.033) | 0.512(0.028) |

risk (4–7 points), and high risk (>7 points) groups. For the sake of comparison, we also classified patients as low risk group (<4), moderate risk group (0.4–0.7), and high risk group (>0.7) according to the average prediction probability of each patient. The comparison results are shown in Table 6. The analysis showed that EM was much better than the CSS. EM was able to assign a higher proportion of patients who died to the high-risk group and a higher proportion of patients who survived to the low-risk group. These results indicated that feature preprocessing, feature selection, and EM are helpful in improving predictive performance.

Discussion

In this study, we developed a novel data preprocessing method to deal with complex clinical data, and an EM to predict high-risk COVID-19 patients. Most importantly, we trained and tested the models in a large cohort and successfully validated the predictive models in a large independent cohort. This is the first study to show that high-risk COVID-19 predictive models and key features are able to reproduce the predictions in a large independent cohort, suggesting that they could be potentially useful in clinical settings.

By comparing our model with CSSs [14], we showed that our missing clinical features imputation method and the EM more accurately help physicians predict patients’ mortality risk. In addition, we used GA to select the most appropriate 14 features from the 20 clinical features, and demonstrated that a removing of redundant features significantly improved the performance of the predictive models. The feature importance analysis showed that MAP, IL-6, procalcitonin, Ddimer, age, and glucose were the most important features affecting the mortality risk of patients.

We used GA to find the optimal combination coefficient of comprehensive usage of five predictive models to construct the EM. In 100 rounds of half-to-half cross-validation, the EM achieved the best performance in multiple evaluation indicators. Moreover, we analyzed the precision and recall of each model under different thresholds to help clinicians in making choices according to the availability of clinical information. If physiological indicators, especially clinical features that reflect inflammation, hepatorenal function, and cardiovascular function can be obtained during the patient’s stay in hospital, our models could be easily used to predict high-risk patients timely.

Our study showed that clinical features, such as age, MAP, and features that are associated with physiological status of the patient, can contribute to the predictive model of mortality stratification for COVID-19 in patients. The physiological status of coagulation function (related feature: Ddimer), hepatorenal function (related feature: glucose), and cardiac function (related feature: troponin) also had a noteworthy effect on mortality, which is consistent with previous findings [9–12]. In addition, we provided reference ranges for clinical features to help physicians quickly stratify patients using our models.

The experiments on the Cohort 2 demonstrated the correctness of our feature selection and the robustness of the predictive model. Despite the differences in population characteristics such as age distribution, ethnic proportion between Cohort 1 and Cohort 2 (Supplementary Tables S1 and S2), and the inconsistent clinical features adopted (Supplementary Table S3), our prediction model still achieved good performance. These results further confirm that age, MAP, and clinical features related to inflammation, coagulation, hepatorenal function, and cardiovascular function can be used to predict the risk of death in patients with COVID-19.

Finally, we compared our model with those published by others [7, 13, 14, 37–40] (Supplementary Table S7). We first compared features of different models. Overall, age, features associated with inflammation, kidney function, cardiovascular function, and lung function were selected for multiple studies, suggesting that the features we selected were more reasonable. Moreover, we employ more efficient feature selection methods to improve model prediction performance. Then, we compare the frameworks used by different studies. Three studies adopted the gradient boosting framework [7, 38, 39], another three adopted the deep learning framework [13, 37, 40], and one invented a scoring method [14]. Our model (EM) takes advantage of the gradient boosting frameworks (XGBoost and GBDT) with proven predictive performance, as well as the RF model, LR model, and SVM. Finally, we compared performance of different models. Three of the models were not completely consistent with our model in the selection of predicted clinical outcomes, such as severity of the disease [13, 40] and distinguishing COVID-19 patients from other pneumonia patients [37]. Furthermore, our model still achieved a
good performance from the perspective of the comparison of the models’ discriminative ability. For the remaining models, our model had the best discriminative performance compared with Rechtman et al.’s [38] model and the CSS [14]. Compared with Bara et al.’s [39] study (Supplementary Table S8), our subjects had a higher percentage of deaths and the AUC of our model was slightly lower than their model, but we achieved a higher precision when we achieved the same recall. Compared with the research of Yan et al. [7], their “single-tree XGBoost” model has an outstanding predictive performance (AUC: 0.9506), but they chose only three features and their study cohort consisted of only 485 patients, making their model unreliable and not performing well on the tests of others [8]. In general, our predictive model (EM) is effective in predicting COVID-19 mortality risk.

There are also some limitations in this study. First of all, for Cohort 1 (the training set), the patient population we studied was mainly hospitalized patients, and they generally exhibited more severe symptoms and therefore had a higher mortality rate than the general population, which may have caused some bias in our predictive model in the general population. Second, the characteristics of the cohort may change the performance of models and its ability to be validated. For example, the model’s performance was slightly lower in Cohort 2 than Cohort 1, because the structure of the two cohorts, such as age distribution, sex ratio, mortality rate, etc., is different. In addition, although we adopt functionally similar features, the differences between these features may also be responsible for the difference in model performance between cohorts. Moreover, since most of the clinical features adopted in this study were missing to varying degrees, the imputed data were affected by other data, which may affect the accuracy of the predictive model. Finally, COVID-19 pandemics are often accompanied by surges in patient numbers, resulting in difficulties in collecting all the required clinical features data, which will limit the application of our predictive model.

Conclusions
In summary, we selected 14 clinical features from 20 clinical features, and comprehensively utilized five predictive models to construct our predictive model: the EM, which had the best performance on multiple predictive evaluation indicators for COVID-19 mortality risk. Most importantly, EM was successfully validated in an independent cohort containing a large number of patients. We also studied the changes of precision and recall of each model under different thresholds, so as to provide reference for doctors to select appropriate thresholds according to medical resources. In addition, feature importance analysis showed that clinical features related to inflammation, hepatorenal function, and cardiovascular function were good predictors for COVID-19 mortality risk, which was consistent with previous studies.

Declarations

Ethics approval and consent to participate
Informed consent was waived due to the nature of study being retrospective.

Consent for publication
Not applicable.

Supplementary data
Supplementary data are available at Biology Methods and Protocols online.

Data availability
All data after de-identification will be made available with publication upon request to the corresponding author. The source code for data analysis is available. The data underlying this article are available at https://doi.org/10.1383/41598-020-73962-9 and in UK biobank at https://doi.org/10.1371/journal.pmed.1001779.

Acknowledgements
The authors thank Prof. Andrzej Stasiak for his valuable suggestions to improve the quality of the manuscript.

Authors’ contributions
J.L. and E.W. were responsible for the conception and design of the study. Y.W. provided support. J.L., X.L., and E.W. were responsible for the implementation and analysis of the algorithm. J.H., M.A., J.L., X.L., and E.W. were responsible for data collection. J.H. and M.A. were responsible for model validation on Cohort 2. X.L., J.L., Y.L., and E.W. were responsible for manuscript writing.

Funding
This work was supported by the Key R&D Plan of Heilongjiang Province through “Research on Key Technologies of Artificial Intelligence Assisted Drug Discovery and Development” and by the Alberta Innovates for Health.

Conflict of interest statement. There are no conflicts of interest.

References
1. Arabi YM, Murthy S, Webb S. COVID-19: A novel coronavirus and a novel challenge for critical care. Intensive Care Med 2020;46:833–6.
2. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. JAMA 2020;323:1545–6.
3. Xie J, Tong Z, Guan X et al. Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med 2020;46:837–40.
4. Guan W-J, Ni Z-Y, Hu Y et al.; China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382:1708–20.
5. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72314 cases from the Chinese center for disease control and prevention. JAMA 2020;323:1239–42.
6. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. Lancet Glob Health 2020;8:e480.
7. Yan L, Zhang H-T, Goncalves J et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020;2:283–8.
8. Barish M, Bolourani S, Lau LF et al. External validation demonstrates limited clinical utility of the interpretable mortality
