*JAMIA Open*, 5(2), 2022, 1–9
https://doi.org/10.1093/jamiaopen/ooac036
Research and Applications

Research and Applications

# Generalizable prediction of COVID-19 mortality on worldwide patient data

**Maxim Edelson**[1] **and Tsung-Ting Kuo**[2]

[1]UCSD Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA and [2]UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, California, USA

Corresponding Author: Tsung-Ting Kuo, PhD, UCSD Health Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA; tskuo@ucsd.edu

Received 14 March 2022; Revised 30 April 2022; Editorial Decision 4 May 2022; Accepted 5 May 2022

## ABSTRACT

**Objective:** Predicting Coronavirus disease 2019 (COVID-19) mortality for patients is critical for early-stage care and intervention. Existing studies mainly built models on datasets with limited geographical range or size. In this study, we developed COVID-19 mortality prediction models on worldwide, large-scale "sparse" data and on a "dense" subset of the data.
**Materials and Methods:** We evaluated 6 classifiers, including logistic regression (LR), support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), AdaBoost (AB), and Naive Bayes (NB). We also conducted temporal analysis and calibrated our models using Isotonic Regression.
**Results:** The results showed that AB outperformed the other classifiers for the sparse dataset, while LR provided the highest-performing results for the dense dataset (with area under the receiver operating characteristic curve, or AUC $\approx$ 0.7 for the sparse dataset and AUC = 0.963 for the dense one). We also identified impactful features such as symptoms, countries, age, and the date of death/discharge. All our models are well-calibrated ($P > .1$).
**Discussion:** Our results highlight the tradeoff of using sparse training data to increase generalizability versus training on denser data, which produces higher discrimination results. We found that covariates such as patient information on symptoms, countries (where the case was reported), age, and the date of discharge from the hospital or death were the most important for mortality prediction.
**Conclusion:** This study is a stepping-stone towards improving healthcare quality during the COVID-19 era and potentially other pandemics. Our code is publicly available at: https://doi.org/10.5281/zenodo.6336231.

**Key words:** COVID-19, coronavirus, predictive modeling, machine learning, data mining

**LAY SUMMARY**
Our study aims to develop a globally generalizable Coronavirus disease 2019 (COVID-19) death prediction tool. To achieve this, we used a large quantity of publicly available COVID-19 patient data collected from across the globe. We also examined the effects of data quality on our results by forming a high missing-value dataset and a low missing-value 1, and then comparing their respective results. We used a variety of classification models, and found that patient information on symptoms, countries (where the case was reported), age, and the date of discharge from the hospital or death were the most important for deciding patients' COVID-19 mortality outcomes. Our models provide a reference for improving the healthcare quality that patients receive during the COVID-19 pandemic era.

## INTRODUCTION

Coronavirus disease 2019 (COVID-19) has resulted in more than 5.2 million confirmed deaths[1] and spans across almost every country in the world. The World Health Organization (WHO) has declared that the infection fatality ratio (aka the mortality rate) among all infected individuals of COVID-19 converges at 0.5–1.0%.[2] Thousands of people worldwide continue to be deceased due to COVID-19[3] and this trend is likely to continue for the foreseeable future as cases continue to spike sporadically, vaccine mandates are fiercely resisted, and new mutations emerge. It is therefore imperative to identify patients with higher risk of fatality, so that healthcare institutions can provide adequate early-stage care and interventions to reduce the risk of COVID-19 mortality.

The Centers for Disease Control and Prevention (CDC) has recognized older age, kidney disease, lung disease, and certain neurological and developmental conditions as factors that can increase a patient's risk for COVID-19 mortality.[4] Based on these factors, several existing studies[5–11] have proposed pipelines aiming to leverage artificial intelligence/machine learning (AI/ML) into predicting mortality using patients' data. Most of these studies were performed on smaller datasets collected from 1 city[5,6] or on a moderate cohort size (<5000 patients).[6–10] These datasets contain detailed/curated clinical information on each patient, contain a low missing value ratio, and are specific to 1 geographic location. However, in a real, clinical COVID-19 setting, overwhelmed hospitals, or intensive care units may not have the resources or time to contact the patients' primary care providers to complete the missing medical history information, and thus the missing data ratio tends to be high.[12,13]

Also, a model built on data from a particular hospital or city may be less pertinent to COVID-19 patients outside of that region. In addition, there are some studies that use a relatively large dataset with the assumption that the dataset is balanced. For example, a recent study[11] used a dataset containing >~110 000 patients and adopted a preprocessing step to balance the dataset between deceased and discharged patients; this effectively creates a mortality rate of 50%, which may limit the application for real clinical use. Although these studies showed the effectiveness of adopting AI/ML methods to predict patient fatality, the data assumptions of (1) a low missing value ratio, (2) a single region, and (3) a balanced mortality rate may hinder the generalizability for real-world clinical applications.

## OBJECTIVE

Our goal is to create a model that is generalizable to the world in retrospectively predicting COVID-19 patient mortality using real-world data (1) with a medium to high missing value ratio, (2) with multiple regions encompassed, and (3) without manual balancing of the discharged and deceased patients' relative ratios.

## MATERIALS AND METHODS

### Data

To address the overall goal of model generalizability, we utilized an open-source COVID-19 dataset[14,15] collected from government sources, scientific papers, and news websites, which contained 2 676 403 COVID-19-confirmed patients from around the world as of March 31, 2021. The Institutional Review Board at University of California San Diego (UCSD) approved this study (no. 190385). Although the earlier versions of this dataset were also used by previous predictive modeling studies,[9–11] we used a more recent version, which is therefore more complete.

We kept 2 567 823 patients with a known COVID-19 confirmation date (Figure 1A) and discarded those without it. The average age was 45 (SD = 20) and the gender was 47.6% female. The countries that are represented among ≥1% of the total number of patients are the following: India = 11.3% (positive = 5.1%), United States = 4.5% (positive = 87.5%), France = 4.1% (positive = 42.9%), and China = 1.6% (positive = 20.0%). Demographic statistics were obtained from nonunknown data.

We used 2 subsets of the same dataset: a *sparse* dataset and a *dense* dataset. For the sparse dataset, our inclusion criteria for the dataset included (1) patients with a known a COVID-19 confirmation date (ie, COVID-19+ patients), and (2) patients with known outcomes (ie, either "deceased" or "discharged"). We manually reviewed the outcome values to combine semantically equivalent ones (eg, "death" was considered the same as "deceased"); this process was executed by extracting all unique outcomes, and then manually separating them into "deceased," "discharged," or "ambiguous."

All observations with "ambiguous" outcomes (eg, "undertreatment") were then discarded. We did not exclude patient data using any other criteria. Based on these inclusion criteria, the sparse dataset contains 104 047 patients, with a deceased (or positive) rate of 5.73% (5958 positive patients) and a discharged (or negative) rate of 94.27% (98 089 negative patients), as shown in Figure 1B.

To examine the effects of data sparsity (ie, various levels of missing data) and to cross-examine the results with the sparse dataset, we created the dense dataset (Figure 1C), which is a subset of the sparse one. The major difference in the dense dataset is that the fraction of deceased patients is 21.1% or 1452 patients out of 5441; just like with the sparse dataset, we left the positive ratio as-is without balancing them. Each observation in the dense dataset was extracted from the sparse one and the basis of inclusion was whether they reported demographic data for age, sex, symptoms, chronic diseases, or optional dates (optional dates are all date features excluding the confirmation date), as shown in Figure 1D. That is, an observation only needed to include one of those fields to merit being placed in the dense dataset. The sparse dataset is a superset of the dense set, meaning that every patient in the dense dataset is also in the sparse dataset.

### Method overview

A high-level overview of our methodology is illustrated in Figure 2. The following section will introduce our data preprocessing steps in "Data preprocessing" section. The 6 ML classifiers that we used will be explored in "Classifiers" section. Last, our validation, calibration, and evaluation framework will be described in "Validation, calibration, and evaluation" section.

### Data preprocessing

Both the sparse and the dense datasets contained 33 fields. After manual review, we kept 12 relevant fields (Table 1). The original dataset contained 33 fields, and after manual review, we kept 12 relevant fields. The manual review process included removing the following fields that are potentially irrelevant, redundant, or too specific: ID, City, Province, Latitude, Longitude, Geographic Resolution, Lives in Wuhan, Travel History Location, Reported Market Exposure, Additional Information, Source, Sequence Available, Notes for Discussion, Location, Admin 3, Admin 2, Admin 1, New Country, Admin ID, Data Moderator Initials, and Travel History

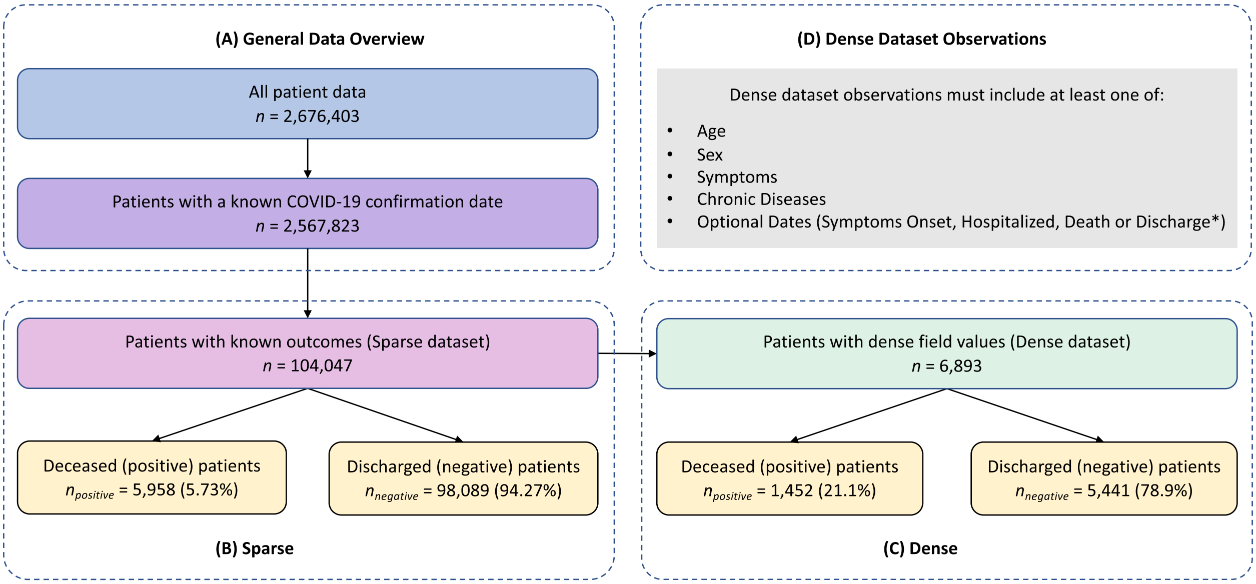
(A) General Data Overview
All patient data
n = 2,676,403
Patients with a known COVID-19 confirmation date
n = 2,567,823
(D) Dense Dataset Observations
Dense dataset observations must include at least one of:
Age
Sex
Symptoms
Chronic Diseases
Optional Dates (Symptoms Onset, Hospitalized, Death or Discharge*)
Patients with known outcomes (Sparse dataset)
n = 104,047
Deceased (positive) patients
n_positive = 5,958 (5.73%)
Discharged (negative) patients
n_negative = 98,089 (94.27%)
(B) Sparse
Patients with dense field values (Dense dataset)
n = 6,893
Deceased (positive) patients
n_positive = 1,452 (21.1%)
Discharged (negative) patients
n_negative = 5,441 (78.9%)
(C) Dense

**Figure 1**. COVID-19 patient data included in this study. (A) The original dataset contained $n = 2\,676\,403$ patients. We kept $n = 2\,567\,823$ patients after discarding all observations without a valid COVID-19 confirmation date. (B) The data breakdown of the "sparse" dataset with $n = 104\,047$ patients. (C) The data breakdown for the "dense" dataset with $n = 6893$. (D) The inclusion requirements for the dense dataset. The "Death or Discharge Date" field (*) has no death or discharge indication and is just a date.

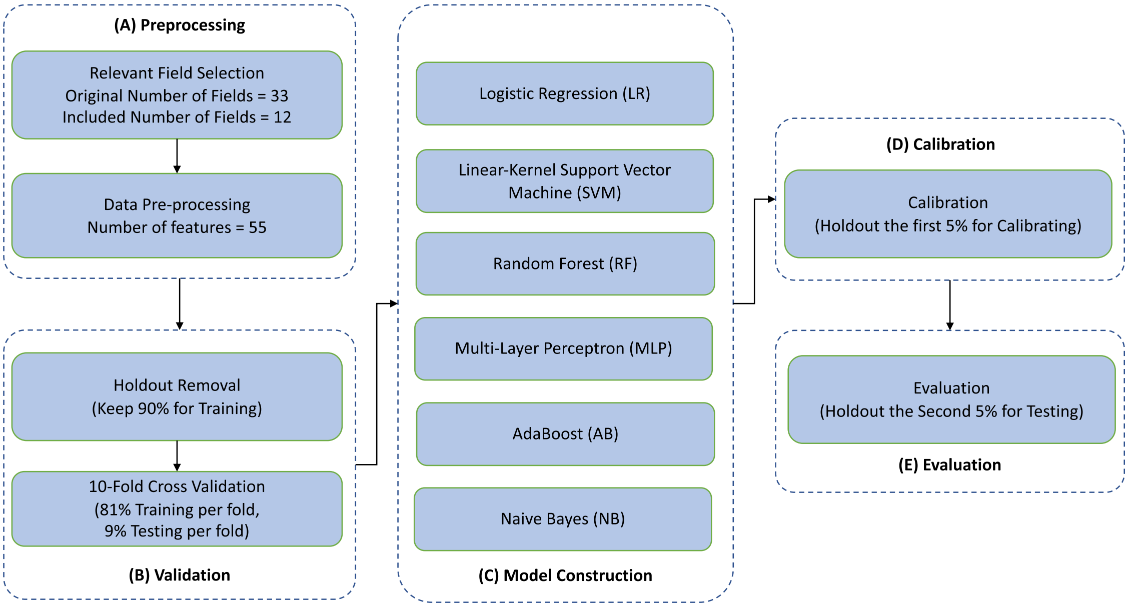

**Figure 2**. Overview of our study's workflow. (A) We started by preprocessing the original dataset from 33 fields down to the 12 most important and relevant fields. From these 12 remaining fields, we extracted 55 features. (B) We then split the dataset to obtain 90% training data. (C) Next, we performed 10-fold cross validation with the training data by feeding our data to our 6 classifiers. (D) We calibrated our models using the first 5% of the holdout data. (E) Finally, we evaluated our calibrated models using the second 5% of the holdout data.

Binary. The nondiscarded fields and their statistics are summarized in Table 1. The missing value ratio for most of the included fields is high to reflect the fact that a real-world COVID-19 dataset must be flexible in its assumptions to be generalizable for countries around the entire globe. We preprocessed the 12 data fields, to extract 1 binary outcome label (ie, whether the patient was deceased/positive or discharged/negative, no. 1 in Table 1) and 55 features. Specifically, we extracted the features from the following fields:

- *Age* (no. 2 in Table 1). We split this field into *Age Lower* and *Age Upper* because certain ages were given as ranges. For ages given as a single value, we assign both *Age Lower* and *Age Upper* to the same age value.
- *Sex and chronic disease flag* (nos. 3 and 4 in Table 1). We converted the sex field into 2 features, the first to indicate male and the second to indicate female (if both are zero, then the sex was considered unreported). The chronic disease flag field was made into a single binary feature (one if a patient has chronic diseases,

**Table 1.** The 12 relevant fields and statistics of our data for both the sparse and dense datasets

| Nos. | Field | Description | Data type | No. of possible values (NOM), or range of values (NUM/DAT) | Missing value percentage (%) | |
|---|---|---|---|---|---|---|
| | | | | | Sparse | Dense |
| 1 | Outcome | Patient outcome from COVID-19 (deceased = 1 or discharged = 0) | NOM | 2 | 0.0 | 0.0 |
| 2 | Age | Age of the patient in years | NUM | 0–101 | 94.5 | 18.9 |
| 3 | Sex | Sex of the patient (male, female, unreported) | NOM | 3 | 93.4 | 0.2 |
| 4 | Chronic disease flag | Binary flag for whether the patient has chronic diseases (true, false) | NOM | 2 | 0.0 | 0.0 |
| 5 | Chronic diseases | List of reported chronic diseases (asthma, chronic kidney disease, diabetes, and hypertension) | NOM | 4 | 99.9 | 98.5 |
| 6 | Symptoms | List of symptoms of the patient experienced | NOM | 10 | 99.8 | 97.4 |
| 7 | Country | Name of country in which the case was reported | NOM | 20 | 0.0 | 0.0 |
| 8 | Date confirmation | Date when patient was confirmed to have COVID-19 | DAT | January 2, 2020–June 3, 2020 | 0.0 | 0.0 |
| 9 | Date of onset symptoms | Date when patient began reporting symptoms | DAT | January 2, 2020–May 27, 2020 | 96.6 | 49.4 |
| 10 | Date of admission hospital | Date when patient was recorded to be hospitalized | DAT | January 2, 2020–April 5, 2020 | 99.8 | 96.5 |
| 11 | Date of death or discharge | Date when death or discharge of the patient was reported (only contains a date without revealing outcome information) | DAT | January 2, 2020–June 4, 2020 | 98.9 | 83.6 |
| 12 | Travel history dates | Recorded travel dates to a location | DAT | January 3, 2020–April 3, 2020 | 99.8 | 97.0 |

*Notes*: The field names and descriptions are adapted from the original dataset.[14,15] We only enumerate possible values of the nominal field with a total number of values <10.

NUM: Numerical, NOM: Nominal, DAT: Date. Dates are given in YYYY/MM/DD format.

otherwise zero). The chronic disease binary flag does not necessarily align with the chronic disease field; that is, even if the chronic disease flag is one (meaning a patient suffers from chronic diseases), the chronic disease field may still contain no data.

- *Chronic diseases, symptoms,* and *country* (nos. 5–7 in Table 1, respectively). We manually reviewed the values of these fields to combine equivalent values.
- *Date confirmation* (no. 8 in Table 1). To enable comparison between dates, we converted this date into an "absolute" day with a reference to the earliest confirmation date available in our entire dataset (ie, January 6, 2020) inclusive of the last day. For example, if the current patient's COVID-19 confirmation date is June 3, 2020, the absolute days for date confirmation for this patient would be 150. In addition, we also use this field as the "base date" for other types of dates to compute "relative" days (details in the next bullet).
- *Date of onset symptoms, date of admission hospital,* and *date of death or discharge* (nos. 9–11 in Table 1, respectively). For these types of dates, we convert each of them into both "absolute" and "relative" days. The process of absolute date conversion is the same as the one for *date confirmation* (ie, computing the difference between a specific date and the earliest value for that type of date, inclusive of that specific day). On the other hand, each relative date value is the difference between the patient's *date confirmation* and the date in question. For example, if the patient's *date confirmation* was March 21, 2020 and their *date of death or discharge* was May 4, 2020, their relative days for *date of death or discharge* would be 45. Note that the *date of death or discharge* only contains a date without revealing outcome information.
- *Travel history dates* (no. 12 in Table 1). Many of this specific type of date were given as ranges. Therefore, we first split this field into *Travel History Dates Begin* and *Travel History Dates End*, (similarly to age, we give these 2 dates the same value if the original data field contains only 1 value). Then, for each of the "begin" and "end" fields, we further extracted both absolute day and relative day features, resulting in 4 features in total.

We created dummy variable features for categorical fields. Then, we normalized numerical features to [0, 1] using the equation of (*current value—minimum value*)/(*maximum value—minimum value*). For fields with missing values, we added a missing indicator feature. For the dense dataset, we further removed all features that were not represented among ≥5 unique observations, to ensure that one feature would not be unrealistically predictive due to only one observation having that feature. As many of the fields have high missing value ratios, this allowed us to use a much denser subset of the sparse dataset to examine the effects of sparsity.

## Classifiers

We adopted 6 classifiers for our COVID-19 mortality prediction (binary classification) task: logistic regression (LR), support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), AdaBoost (AB), and Naive Bayes (NB). All hyper-parameter combinations are shown in Supplementary Appendix Table SA1. For SVM, we used a linear version.[16–18] For MLP, we set the learning rate to 0.1, number of hidden layers = 1, the number of hidden neurons = 110, the learning rate decay = false, and the threshold for consecutive errors = 20. Appropriate hyper-parameter options were discovered through previous studies[16,17,19–23] using similar implementations for the classifiers. While many of the previous studies

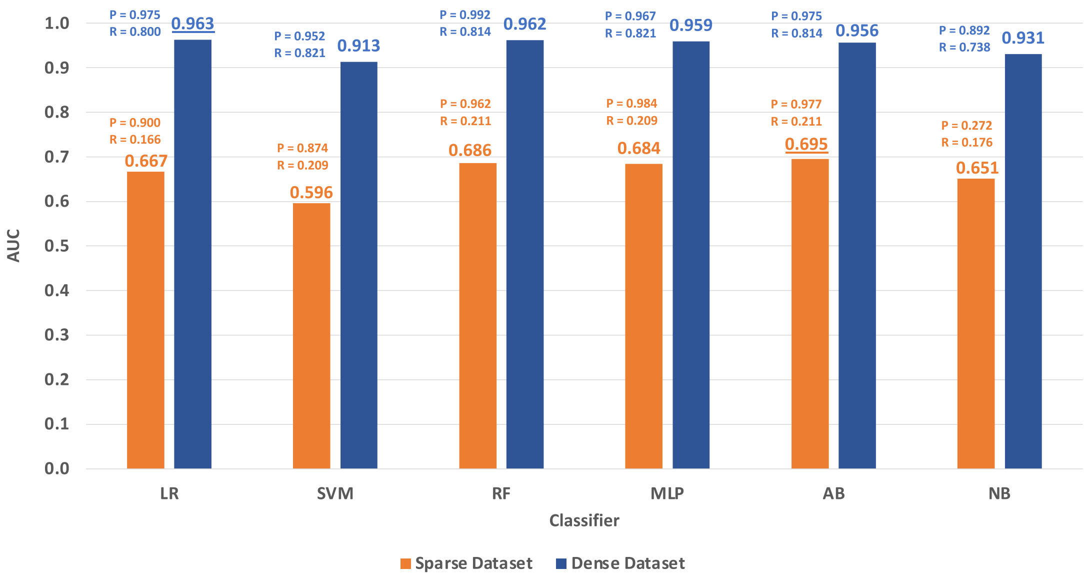

**Figure 3.** The performance of our 6 classifiers with AUC as the evaluation metric. AB outperformed the other 5 classifiers for the sparse dataset and LR was the best performer when trained on the dense dataset. The precision and recall for each result are provided near each respective AUC result, with "P" being the precision and "R" being the recall. We used the default decision threshold of 0.5 when computing the precision and recall values. The classifier abbreviations are as follows: LR: logistic regression; SVM: support vector machine; RF: random forest; MLP: multi-layer perceptron; AB: AdaBoost; NB: Naive Bayes.

**Table 2.** The best hyper-parameter combinations for each of the 6 classifiers on both the sparse and dense datasets

| Classifier | Hyper-parameters | Best sparse data combination | Best dense data combination |
|---|---|---|---|
| LR | • Ridge | • $10^3$ | • $10^2$ |
| SVM | • Cost | • $2^{-5}$ | • $2^{-5}$ |
| RF | • Number of attributes | • $m^{1/3}$ | • $m^{1/2}$ |
|  | • Sample size | • 50% | • 50% |
|  | • Number of trees | • 100 | • 175 |
| MLP | • Momentum | • 0.1 | • 0.3 |
|  | • Number of epochs | • 750 | • 500 |
| AB | • Weight threshold | • 100 | • 100 |
|  | • Number of iterations | • 70 | • 20 |
|  | • Resampling for boosting | • True | • True |
|  | • Base classifier | • J48 | • J48 |
| NB | • Kernel estimator | • True | • False |
|  | • Supervised discretization | • False | • True |

*Note*: Notation: $m$ is the number of attributes.

differed in their datasets and application, we adopted a grid search hyper-parameter tuning approach. We selected the initial values of the grid search based on the previous studies' explored hyper-parameter combinations to optimize the performance of our models. We expanded our grid search hyper-parameter values as necessary; we determined necessity based on whether the highest-performing hyper-parameter combination was an edge case in the grid search. We implemented the classifiers using the WEKA library.[24,25] SVM was implemented using the LibLINEAR API[16,17,26] (also WEKA).

## Validation, calibration, and evaluation

Our validation, calibration and evaluation processes are shown above in Figure 2. The data were split into 3 parts: 90% for training/validation (Figure 2B), the first 5% for calibration (Figure 2D), and the second 5% for evaluation (Figure 2E). We used the full area under the receiver operating characteristic curve (AUC) as our evaluation metric for the classifiers. We built and tested our models on an Amazon Web Services virtual machine with 2 vCPUs, 8 GB RAM, and 100 GB SSD.

1. For *training/validation*, we performed 10-fold cross-validation for each classifier on the 90% training data to tune the hyper-parameters, averaged the validation results in AUC over 10 folds, and calculated the 95% confidence intervals (CIs) of AUC using the best-performing hyper-parameter combinations.
2. For *calibration*, the best hyper-parameter combination for each classifier was trained on the validation data, first 5%, and then tested on the testing data, second 5%, which then provided the input for Isotonic Regression.[18,25]
3. For *evaluation*, we tested the AUC and computed the Hosmer–Lemeshow (H-Statistic[27]) Test on the evaluation data. Given the change in the COVID-19 viral variants, it is imperative to further show our model's ability to predict mortality in different epochs

**Table 3.** The top 10 most important features using both (a) sparse and (b) dense datasets

| Dataset | Nos. | Feature name | Description | Weight |
|---|---|---|---|---|
| (a) Sparse | 1 | Date of death or discharge (absolute) | The number of days that passed between the first recorded date of death or discharge and this patient's date of death or discharge | −4.439 |
| | 2 | Malaysia | Whether the case was reported in Malaysia | 3.567 |
| | 3 | Algeria | Whether the case was reported in Algeria | −3.162 |
| | 4 | Singapore | Whether the case was reported in Singapore | 3.006 |
| | 5 | South Korea | Whether the case was reported in South Korea | 2.712 |
| | 6 | Australia | Whether the case was reported in Australia | 2.633 |
| | 7 | Vietnam | Whether the case was reported in Vietnam | 2.424 |
| | 8 | Date of death or discharge (missing) | Whether the date of the patient's death or discharge was reported (binary) | 1.814 |
| | 9 | United States | Whether the case was reported in the United States | −1.760 |
| | 10 | Chills (symptom) | Whether the patient reported suffering from chills because of COVID-19 | 1.708 |
| (b) Dense | 1 | Date of death or discharge (absolute) | The number of days that passed between the first recorded date of death or discharge and this patient's date of death or discharge | 4.026 |
| | 2 | Algeria | Whether the case was reported in Algeria | 3.535 |
| | 3 | United States | Whether the case was reported in the United States | 2.376 |
| | 4 | India | Whether the COVID-19 case was reported in India | 2.015 |
| | 5 | Age (lower) | The lower age in a patient's age range | 2.003 |
| | 6 | Age (upper) | The upper age in a patient's age range | 1.973 |
| | 7 | Date of death or discharge (missing) | Whether the date of the patient's death or discharge was missing (binary) | 1.918 |
| | 8 | Singapore | Whether the case was reported in Singapore | 1.898 |
| | 9 | Malaysia | Whether the case was reported in Malaysia | 1.873 |
| | 10 | Headache (symptom) | Whether the patient reported suffering from headaches because of COVID-19 | 1.601 |

*Notes*: These features were results of the LR classifier with a ridge-parameter of $10^3$ for the sparse dataset and $10^2$ for the dense dataset. The date of death or discharge only contains a date without outcome information. The features are ordered by descending absolute weight. Negative weights are indicative of discharge and positive weights are indicative of death

of time. Thus, following the CDC's COVID-19 timeline,[28] we split the evaluation data into 2 parts separated by May 2, 2020 (ie, when the WHO declared that COVID-19 was a global health crisis). The first part contains all the data before May 2, 2020 and the second part contains all other data (inclusive of May 2, 2020). We split the evaluation data for both the sparse and dense datasets in this manner.

## RESULTS

The discrimination results of each classifier on the full evaluation data for both datasets are demonstrated in Figure 3. For the sparse dataset, AB resulted in the highest AUC values (AUC ≈ 0.7), followed by RF and MLP (AUC ≈ 0.685). For the dense dataset, LR performed the best (AUC = 0.963), whereas RF, MLP, and AB followed behind closely (AUC ≈ 0.96). SVM and NB provided less competitive results for both datasets. The precision and recall results (computed using a decision threshold of 0.5) for each classifier can also be seen in Figure 3. In general, all models provided good precision (>0.89) and recall (>0.73) for the dense dataset; for the sparse dataset, the precision is still high (>0.87 except for NB), whereas the recall is relatively low (~0.2). The best-performing hyper-parameter combinations for the sparse and the dense datasets are shown in Table 2. We also analyzed the top 10 most important features derived from LR for both the sparse and the dense datasets (Table 3) and ordered them by their decreasing absolute value of their trained weights. Symptoms, countries, ages, and dates of death or discharge were found to be among the most predictive factors. The full LR models for the spares and the dense datasets are shown in the Supplementary Appendix Table SA2 and SA3, respectively.

The temporal and calibration results are shown in Table 4. RF outperformed the other classifiers for the "before May 2, 2020" time period, whereas MLP performed best for the "on-and-after May 2, 2020" epoch for both datasets. All models (including the full evaluation data, evaluation data from before May 2, 2020, and evaluation data from on-and-after May 2, 2020 for both the sparse and the dense datasets) are well-calibrated ($P > .1$). The training time measurements on the full 90% training/validation data for each classifier are shown in Figure 4, and MLP took by far the longest time to train with both datasets (sparse and dense).

## DISCUSSION

### Findings

For the full evaluation data, AB provided higher best-case AUC results for the sparse dataset and LR performed the best on the dense one. Based on the results of all hyper-parameter combinations, we observed that altering the base classifier for AB resulted in the greatest direct change in discrimination results. For LR, SVM, RF, MLP, and NB, none of the hyper-parameter combinations changed the results significantly. Our results also highlight the tradeoff we've embraced when compared with the previous studies[5–11]: models trained using sparse data to increase generalizability (and a poorly performing AUC as a result) versus models trained using dense data with higher discrimination results yet decreased generalizability. The relatively high precision and recall values for our models run on the dense dataset indicate that a decision threshold of 0.5 may be appropriate. On the other hand, the low recall on the sparse dataset was expected due to the imbalance of data (only ~6% positive, while the dense dataset has ~21% of positive examples). Finally, the low recall and precision for NB suggest that a threshold of 0.5 might not be appropriate for the sparse data.

For temporal analysis, all our models performed even better for the patients whose confirmation dates were *before* May 2, 2020

**Table 4.** Temporal and calibration test results for the 6 classifiers

| Dataset | Setting | | Metric | LR | SVM | RF | MLP | AB | NB |
|---|---|---|---|---|---|---|---|---|---|
| (a) Sparse | Training/Validation | | AUC Average | 0.665 | 0.604 | 0.699 | 0.676 | 0.697 | 0.665 |
| | | | AUC 95% CI Low | 0.656 | 0.597 | 0.690 | 0.668 | 0.675 | 0.656 |
| | | | AUC 95% CI High | 0.674 | 0.610 | 0.708 | 0.685 | 0.720 | 0.675 |
| | Evaluation | All | AUC | 0.667 | 0.596 | 0.686 | 0.684 | 0.695 | 0.651 |
| | | | H-L Test *P*-value | 0.975 | 0.856 | 0.375 | 0.207 | 0.296 | 0.381 |
| | | Before May 2, 2020 | AUC | 0.832 | 0.812 | 0.912 | 0.885 | 0.895 | 0.710 |
| | | | H-L Test *P*-value | 0.267 | 1.000 | 1.000 | 0.999 | 0.997 | 0.357 |
| | | On-and-after May 2, 2020 | AUC | 0.615 | 0.530 | 0.630 | 0.640 | 0.631 | 0.625 |
| | | | H-L Test *P*-value | 0.981 | 0.962 | 0.999 | 1.000 | 1.000 | 0.122 |
| (b) Dense | Training/Validation | | AUC Average | 0.961 | 0.910 | 0.968 | 0.960 | 0.959 | 0.925 |
| | | | AUC 95% CI Low | 0.952 | 0.894 | 0.960 | 0.948 | 0.939 | 0.913 |
| | | | AUC 95% CI High | 0.971 | 0.926 | 0.976 | 0.972 | 0.979 | 0.938 |
| | Evaluation | All | AUC | 0.963 | 0.913 | 0.962 | 0.959 | 0.956 | 0.931 |
| | | | H-L Test *P*-value | 0.998 | 1.000 | 0.763 | 1.000 | 0.999 | 0.883 |
| | | Before May 2, 2020 | AUC | 0.982 | 0.960 | 0.997 | 0.993 | 0.992 | 0.974 |
| | | | H-L Test *P*-value | 0.644 | 0.462 | 0.874 | 0.968 | 0.890 | 1.000 |
| | | On-and-after May 2, 2020 | AUC | 0.919 | 0.900 | 0.924 | 0.959 | 0.945 | 0.838 |
| | | | H-L Test *P*-value | 0.830 | 0.717 | 0.917 | 0.839 | 0.996 | 0.848 |

*Notes*: The (**a**) sparse and (**b**) dense evaluation data were split into 2 parts, the first containing all data before May 2, 2020 and the second part contains the rest of the instances (inclusive) because the CDC's COVID-19 timeline[28] depicts May 2, 2020 as the date when the WHO declared that COVID-19 was a global health crisis. The H-L test *P* values show that all models are well-calibrated ($P > .1$) by using isotonic regression calibration.

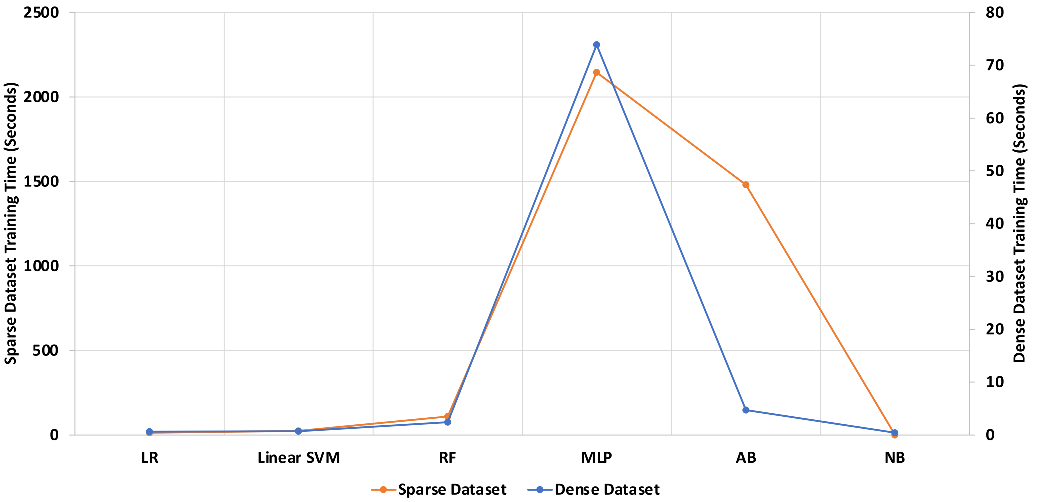

**Figure 4.** The time taken for the training on the full 90% training/validation data for each model. The vertical axis on the left correlates with the sparse dataset and the vertical axis on the right is for the dense dataset.

(with RF's AUC being 0.912 as the best result for the sparse dataset, and 0.997 for the dense one). This is significantly higher than the evaluation on the full evaluation dataset (of which the best models only reached AUC ≈ 0.7 for the sparse dataset and 0.963 for the dense one). On the other hand, for the patients whose confirmation dates were *on-and-after* May 2, 2020, all the AUC values are less than that of the results for the full evaluation data on both sparse and dense datasets. This may be a result of the first part of the dataset (data from before May 2, 2020) containing more nonmissing data, while later instances have a higher rate of missing values. Another possibility is that before May 2, 2020, COVID-19 mortality may be easier to predict due to the case counts increasing, but not yet surging, while after May 2, 2020, the number of cases began to surge making it more difficult to predict. Furthermore, our models can provide reliable prediction scores after calibration.

## Limitations

There are several limitations for this study:

1. Using a longitudinal dataset would allow us to showcase the relative dangers that each COVID-19 variants (eg, alpha, delta, and omicron) and their mixes (eg, percent delta/alpha or delta/omicron) poses. Although exploring such a dataset may have the potential to reveal certain symptoms or risk factors that are correlated with specific variants, we are yet to extend our study to model such type of data.
2. Calculating the "optimal" decision threshold is often desirable because the threshold is usually problem-specific, and the real threshold value may be more biased towards one outcome over the other.[29] Moreover, this "optimal" decision threshold can potentially affect our precision and recall results, especially for

the dense dataset. Additionally, performing calibration near the estimated “optimal” decision threshold can be important because even a minute change in prediction scores near the decision threshold can flip the predicted class. We are yet to consult with clinical experts to estimate such “optimal” decision threshold, as well as performing subsequent calibration around the estimated threshold and recompute the precision and recall results.

3. Our data contains a skewed geographical distribution (ie, most of the observations were from just a handful of countries and the rest of the countries are only represented in this dataset by a small number of cases). This leads to the less represented countries being more influential to the overall feature prediction. For example, if one of the country features, like Gabon, reported only 5 observations, and all 5 patients died due to COVID-19, then this feature may be receiving a higher absolute weight (importance) in determining patient outcome than it should be. Additionally, the dataset is lacking data from certain countries, which may result in a model that does not directly represent a global sample. Therefore, further investigation of the potential geographical biases in the dataset may be required.
4. Our hyper-parameter exploration process involved iterating through a grid search algorithm, which is a computationally intensive process. Therefore, alternative hyper-parameter tuning techniques (eg, random search[30]) may allow us to search hyper-parameter combinations more efficiently, and thus warrant further studies.
5. The information on what treatments patients received was not present in our dataset, and therefore the effects of certain treatments on mortality were not compared. We are yet to include such additional information to examine if certain treatments might directly affect the probability of survival. Moreover, we are yet to consult with clinical experts to perform “blind assessment” of the features and mortality used in our data, as well as to create risk groups for stratifying which specific groups of patients are more susceptible to high risk of death.
6. We adopted more traditional classification methods, prioritizing the simplicity and explainability of the models. On the other hand, advanced techniques such as Deep Learning could also be considered. For example, our tabular datasets can potentially be converted to sequential representations[31] for recurrent neural network,[32] or to 2D representations[33] for convolutional neural network.[34] These advanced Deep Learning methodologies for predicting patient mortality warrant further exploration.
7. There have been several COVID-19 clinical prediction instruments developed since the start of the pandemic including the AIFELL[35] and the 4C[36] scores. The AIFELL score was designed to differentiate between severe and less severe COVID-19 cases in emergency room environments. The 4C score was developed to directly inform clinicians in their decision-making process, as well as to separate COVID-19 hospital admittees into different risk management groups. We have yet to compare our prediction results with those of these existing tools, or to combine various models to introduce a mortality prediction tool with potentially better predictive capability.

## CONCLUSIONS

In this study, we demonstrated the feasibility to build generalizable COVID-19 mortality predictive models. To do this, we used a worldwide dataset that contained high missing value ratios for the most of our included fields. We evaluated 6 classifiers on a COVID-19 dataset featuring patients from around the world and reached an AUC $\approx 0.7$ for the sparse dataset and AUC$=0.963$ for the dense dataset. This study is a stepping-stone to creating highly generalizable models that can predict mortality for COVID-19 patients with the goal of improving healthcare quality during the COVID-19 era and other future pandemics.

## FUNDING

The author T-TK was funded by the US National Institutes of Health (NIH) R00HG009680, R01HL136835, R01GM118609, R01HG011066, and U24LM013755. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

## AUTHOR CONTRIBUTIONS

ME contributed to the conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft), and visualization. T-TK contributed to conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (review and editing), visualization, supervision, project administration, and funding acquisition.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

The data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.6336231. The datasets were derived from sources in the public domain: https://github.com/beoutbreakprepared/nCoV2019.

## REFERENCES

1. CDC Covid Data tracker [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker/#global-counts-rates. Accessed May 12, 2022.
2. Estimating Mortality from Covid-19 [Internet]. World Health Organization. World Health Organization; 2020. https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19. Accessed May 12, 2022
3. Ritchie H, Mathieu E, Rodés-Guirao L, *et al.* Coronavirus (COVID-19) Deaths [Internet]. Our World in Data; 2020. https://ourworldindata.org/covid-deaths. Accessed May 12, 2022.
4. People With Certain Medical Conditions [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2022. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html. Accessed May 12, 2022.
5. Jamshidi E, Asgary A, Tavakoli N, *et al.* Symptom prediction and mortality risk calculation for COVID-19 using machine learning. *Front Artif Intell* 2021; 4: 673527.
6. Shanbehzadeh M, Orooji A, Kazemi-Arpanahi H. Comparing of data mining techniques for predicting in-hospital mortality among patients with covid-19. *J Biostat Epidemiol* 2021; 7 (2): 154–73.

7. Broberg CS, Kovacs AH, Sadeghi S, *et al.* COVID-19 in adults with congenital heart disease. *J Am Coll Cardiol* 2021; 77 (13): 1644–55.
8. Di Castelnuovo A, Bonaccio M, Costanzo S, *et al.* Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. *Nutr Metab Cardiovasc Dis* 2020; 30 (11): 1899–913.
9. Albitar O, Ballouze R, Ooi JP, *et al.* Risk factors for mortality among COVID-19 patients. *Diabetes Res Clin Pract* 2020; 166: 108293.
10. Mohammed M, Muhammad S, Mohammed FZ, *et al.* Risk factors associated with mortality among patients with novel coronavirus disease (COVID-19) in Africa. *J Racial Ethnic Health Disparities* 2020; 8 (5): 1267–72.
11. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. *Smart Health* 2020; 20: 100178.
12. Huq F, Manners E, O'Callaghan D, *et al.* Patient outcomes following transfer between intensive care units during the COVID-19 pandemic. *Anaesthesia* 2022; 77 (4): 398–404.
13. Key Considerations for Transferring Patients to Relief Healthcare Facilities When Responding to Community Transmission of COVID-19 in the United States [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2020. https://www.cdc.gov/coronavirus/2019-ncov/hcp/relief-healthcare-facilities.html. Accessed May 12, 2022.
14. Xu B, Gutierrez B, Mekaru S, *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 2020; 7 (1): 1–6.
15. Open COVID-19 Data Working Group. Detailed Epidemiological Data from the COVID-19 Outbreak [Internet]. 2020. http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337. Accessed May 12, 2022.
16. Mantovani RG, Rossi AL, Vanschoren J, *et al.* To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning. In 2015 *International Joint Conference on Neural Networks (IJCNN)*; July 12, 2015: 1–8. IEEE; Killarney, Ireland.
17. Wang H, Khoshgoftaar TM, Napolitano A. An empirical study of software metrics selection using support vector machine. In: Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011), Eden Roc Renaissance; Miami Beach: Knowledge Systems Institute Graduate School; 2011: 83–8.
18. De Leeuw J, Hornik K, Mair P. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software* 2010; 32: 1–24.
19. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *J R Stat Soc Ser C (Appl Stat)* 1992; 41 (1): 191–201.
20. Hoang TB, Mothe J. Location extraction from tweets. *Inform Process Manage* 2018; 54 (2): 129–44.
21. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; 9 (3): e1301.
22. Martínez-Muñoz G, Suárez A. Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recogn* 2010; 43 (1): 143–52.
23. Kang K, Michalak J. Enhanced version of AdaBoostM1 with J48 Tree learning method. *arXiv* preprint arXiv:1802.03522. 2018.
24. Witten IH, Frank E, Hall MA. The WEKA workbench. In: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". 4th ed. Burlington, MA: Morgan Kaufmann; 2016.
25. Hall M, Frank E, Holmes G, *et al.* The WEKA data mining software: an update. *ACM SIGKDD Explor Newsl* 2009; 11 (1): 10–8.
26. Fan RE, Chang KW, Hsieh CJ, *et al.* LIBLINEAR: A library for large linear classification. *J Mach Learn Res* 2008; 9: 1871–4.
27. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27 (4): 621–33.
28. CDC Museum Covid-19 TimeLine [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2022. https://www.cdc.gov/museum/timeline/covid19.html. Accessed May 12, 2022.
29. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *J Biomed Inform* 2017; 76: 9–18.
30. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012; 13 (2): 281–305.
31. Chopra C, Sinha S, Jaroli S, Shukla A, Maheshwari S. Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients. In: proceedings of the 2017 International Conference on Computational Biology and Bioinformatics; October 18, 2017: 18–23.
32. Medsker L, Jain LC, eds. *Recurrent Neural Networks: Design and Applications*. Boca Raton, FL: CRC Press; 1999.
33. Zhu Y, Brettin T, Xia F, *et al.* Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep* 2021; 11 (1): 1.
34. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET); August 21, 2017: 1–6. IEEE.
35. Levenfus I, Ullmann E, Petrowski K, *et al.* The AIFELL Score as a Predictor of Coronavirus Disease 2019 (COVID-19) severity and progression in hospitalized patients. *Diagnostics* 2022; 12 (3): 604.
36. Knight SR, Ho A, Pius R, *et al.* Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ* 2020; 370: m3339.