COVID-19 mortality prediction from deep learning in a large multistate EHR and LIS dataset: algorithm development and validation

Saranya Sankaranarayanan, Jagadheshwar Balan, Jesse R. Walsh, Yanhong Wu, Sara J. Minnich, Amy L. Piazza, Collin Osborne, Gavin R. Oliver, Jessica L. Lesko, Kathy L. Bates, Kia Khezeli, Darci R. Block, Margaret A. DiGuardo, Justin Kreuter, John C. O'Horo, Iman J. Kalantari, Eric W. Klee, Mohamed E. Salama, Benjamin R. Kipp, William G. Morice II, Garrett Jenkinson

Submitted to: Journal of Medical Internet Research
on: May 03, 2021

Table of Contents

COVID-19 mortality prediction from deep learning in a large multistate EHR and LIS dataset: algorithm development and validation

Saranya Sankaranarayanan[1*]; Jagadheshwar Balan[1*]; Jesse R. Walsh[1]; Yanhong Wu[1]; Sara J. Minnich[1]; Amy L. Piazza[1]; Collin Osborne[1]; Gavin R. Oliver[1]; Jessica L. Lesko[1]; Kathy L. Bates[1]; Kia Khezeli[1]; Darci R. Block[1]; Margaret A. DiGuardo[1]; Justin Kreuter[1]; John C. O'Horo[1]; Iman J. Kalantari[1]; Eric W. Klee[1]; Mohamed E. Salama[1]; Benjamin R. Kipp[1]; William G. Morice II[1]; Garrett Jenkinson[1]

[1]Mayo Clinic Rochester US
[*]these authors contributed equally

Corresponding Author:
Garrett Jenkinson
Mayo Clinic
200 1st St SW
Rochester
US

Abstract

Background: COVID-19 is caused by the SARS-CoV-2 virus and has strikingly heterogeneous clinical manifestations with most individuals contracting mild disease but a substantial minority experiencing fulminant cardiopulmonary symptoms or death. The clinical covariates and the lab tests performed on a patient provides robust statistics to guide clinical treatment. Deep learning approaches on a dataset of this nature enables patient stratification and provide methods to guide clinical treatment.

Objective: Here we report on the development and prospective validation of a state-of-the-art machine learning model to provide mortality prediction shortly after confirmation of SARS-CoV-2 infection in the Mayo Clinic patient population.

Methods: We constructed one of the largest reported and most geographically diverse laboratory information system (LIS) and electronic health record (EHR) COVID-19 datasets in the published literature, which included 11,808 patients with residence in 41 states, treated at medical sites across five states in three time zones. This data was split by date into an 80/20 training and prospective testing cohort. In the training data, model selection and evaluation were performed using stratified 10-fold cross-validation. Traditional machine learning models were evaluated independently as well as in a stacked learner approach using Autogluon, and various recurrent neural network architectures were considered. We trained these models to operate solely using routine laboratory measurements and clinical covariates available within 72 hours of a patient's first positive COVID-19 nucleic acid test.

Results: The GRU-D recurrent neural network achieved peak cross-validation performance with 0.938±0.004 AUROC. In cross-validation, this model provides accuracy of 89% (95% CI: [88,90]), a recall of 80% (95% CI: [74,85]), a precision of 17% (95% CI: [15,19]), a negative predictive value (NPV) of 99% (95% CI: [99,100]), and statistically significant stratification in our Cox proportional hazards survival model (risk 18.9, P<.001). The model retained strong performance when reducing the follow-up time down to 12 hours (0.916±0.005 AUROC), and leave-one-out feature importance analysis indicates the most independently valuable features were: age, Charlson score, minimum oxygen saturation, fibrinogen and serum iron level. In the prospective testing cohort this model provides AUROC of 0.901, an accuracy of 78% (95% CI: [76,79]), a recall of 85% (95% CI: [77,91]), a precision of 14% (95% CI: [12,17]), a negative predictive value (NPV) of 99% (95% CI: [99,100]), and statistically significant difference in survival (P<.001, hazard ratio for those predicted to survive: 95% CI [0.043,0.106]).

Conclusions: Our deep learning approach using GRU-D provides an alert system to flag mortality on COVID-19 positive patients, using clinical covariates and lab values within a 72-hour window after the first positive nucleic acid test.

(JMIR Preprints 03/05/2021:30157)
DOI: https://doi.org/10.2196/preprints.30157

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Saranya Sankaranarayanan*, Jagadheshwar Balan*, Jesse R. Walsh, Yanhong Wu, Sara J. Minnich, Amy L. Piazza, Collin Osborne, Gavin R. Oliver, Jessica L. Lesko, Kathy L. Bates, Kia Khezeli, Darci R. Block, Margaret A. DiGuardo, Justin Kreuter, John C. O'Horo, Iman J. Kalantari, Eric W. Klee, Mohamed E. Salama, Benjamin R. Kipp, William G. Morice II, Garrett Jenkinson†

Mayo Clinic, Rochester, Minnesota, USA

Corresponding Author:
Garrett Jenkinson, Ph.D.
Assistant Professor of Biomedical Informatics
Mayo Clinic, Rochester, Minnesota
Email: Jenkinson.William@mayo.edu

COVID-19 mortality prediction from deep learning in a large multistate EHR and LIS dataset: algorithm development and validation

Abstract

Background

COVID-19 is caused by the SARS-CoV-2 virus and has strikingly heterogeneous clinical manifestations with most individuals contracting mild disease but a substantial minority experiencing fulminant cardiopulmonary symptoms or death. The clinical covariates and the lab tests performed on a patient provide robust statistics to guide clinical treatment. Deep learning approaches on a dataset of this nature enable patient stratification and provide methods to guide clinical treatment.

Objective

Here we report on the development and prospective validation of a state-of-the-art machine learning model to provide mortality prediction shortly after confirmation of SARS-CoV-2 infection in the Mayo Clinic patient population.

Methods

We retrospectively constructed one of the largest reported and most geographically diverse laboratory information system (LIS) and electronic health record (EHR) COVID-19 datasets in the published literature, which included 11,807 patients with residence in 41 states, treated at medical sites across five states in three time zones. Traditional machine learning models were evaluated independently as well as in a stacked learner approach using AutoGluon, and various recurrent neural network (RNN) architectures were considered. The traditional machine learning models were implemented using the AutoGluon-Tabular framework, whereas the RNNs utilized the tensorflow keras framework. We trained these models to operate solely using routine laboratory measurements and clinical covariates available within 72 hours of a patient's first positive COVID-19 nucleic acid test.

Results

The GRU-D recurrent neural network achieved peak cross-validation performance with 0.938±0.004 AUROC. The model retained strong performance when reducing the follow-up time to 12 hours (0.916±0.005 AUROC), and leave-one-out feature importance analysis indicated the most independently valuable features were: age, Charlson score, minimum oxygen saturation, fibrinogen and serum iron level. In the prospective testing cohort this model provides an AUROC of 0.901 and statistically significant difference in survival (P<.001, hazard ratio for those predicted to survive: 95% CI [0.043,0.106]).

Conclusions

Our deep learning approach using GRU-D provides an alert system to flag mortality on COVID-19 positive patients, using clinical covariates and lab values within a 72-hour window after the first positive nucleic acid test.

Keywords

COVID-19, mortality prediction, recurrent neural networks, missing data, time series

Introduction

Coronavirus Disease 2019 (COVID-19) is caused by the SARS-CoV-2 virus and is suspected to be of zoonotic origin with spillover from bats or pangolins into humans in Wuhan, China [1, 2]. COVID-19 has become one of the largest public health emergencies of the past century with over 203 million confirmed cases and 4.3 million deaths as of August 2021 according to the World Health Organization [3]. The pandemic has overwhelmed global medical supply chains, hospitals, and economies, which has led governments to respond with varying policies including mask mandates and travel restrictions [4, 5]. At times, hospitals and healthcare workers have become so overcapacity with COVID-19 patients that they have been forced to ration care, raising logistical and ethical concerns [6].

The clinical course of COVID-19 is diverse with most individuals experiencing mild or asymptomatic disease, but many patients develop life-threatening disease including features such as cytokine storms, thrombotic complications, or severe acute respiratory syndrome requiring mechanical ventilation or extracorporeal membrane
oxygenation (ECMO) [7]. A major medical challenge is therefore to reliably triage patients according to their risk for severe disease. Age is consistently observed to be a predominant risk factor for severe disease [7], but deaths are not limited to the elderly and the majority of elderly patients survive COVID-19 [7]. Other comorbidities and laboratory test values are expected to be capable of further individualizing and enhancing mortality prediction. Recent studies investigating statistical and machine learning models for mortality prediction have confirmed that detailed evaluation of medical records can further stratify patients [8-12].

A systematic review of 147 published or preprint prediction models found consistent problems with inherent biases in the datasets investigated or created in all such studies, ultimately concluding “we do not recommend any of these reported prediction models for use in current practice” [12]. Clinical practices differ in the nature of their observational electronic health record (EHR) dataset, patient population, clinical practices, and electronic record or laboratory ordering practices.

Correspondingly, the literature review conducted at the outset of this study indicated that existing prediction models were likely unsuited to our clinical setting without essentially starting afresh by retraining, validating and testing predictions.

We describe Mayo Clinic's experience assembling what is to our knowledge the largest reported COVID-19 database for mortality prediction and using this to create a system for COVID-19 mortality prediction, tailored to a unique patient population. Despite the biases inherent to it, because this large and growing database represents a healthcare system spanning five states and three time zones over a study window greater than eleven months, our model is expected to be the least confounded and most generalized COVID-19 mortality predictor published to-date. We report the successful development and validation of a state-of-the-art machine learning model to provide mortality prediction shortly after confirmation of SARS-CoV-2 infection in this Mayo Clinic patient population, and discuss in detail the various logistical and scientific challenges involved in the early deployment of such a system in fast-moving pandemic environment.

Methods

This work required both the development of a dataset and the subsequent modeling of the resultant cohort. After data collection and cleaning, two broad classes of algorithms were considered to model this data. The first approach ignores the time series nature of the underlying data and applies traditional machine learning classifiers. The second approach explicitly models the time series data while dealing with the missing-not-at-random (MNAR) values using specialized recurrent neural networks (RNNs). Both types of modeling methods were run independently and compared using cross-validation and a single winning model was selected for prospective performance validation.

EHR and LIS Observational Cohort Data Collection

This study adheres to a research protocol approved by the Mayo Clinic Institutional Review Board. Data was retrospectively collected after March 1, 2020 on COVID-19 positive individuals presenting to a Mayo Clinic site or health system, while excluding patients without research consent or from EU countries covered by the GDPR law. We restricted our focus to the 11,807 patients with a positive COVID-19 nucleic acid test on or before January 27, 2021 and at least one non-COVID test result. Although the data collection system is deployed and ongoing, the January cutoff was selected for this manuscript to provide sufficient cohort size while allowing a minimum of three weeks of follow-up to accurately establish survival status.

Mayo Clinic's EHR and laboratory information system (LIS) contain data from each of its three campuses (Rochester, Minnesota; Jacksonville, Florida; Scottsdale, Arizona) as well as the surrounding health system sites spanning five states (MN, IA, WI, FL, AZ). Although the EHR contains clinically reportable laboratory results, many of these can only be reported within defined ranges, which can result in qualitative text values rather than the raw numeric measurements. Because many machine learning algorithms typically work better with quantitative rather than qualitative results, we used the LIS to gather such laboratory testing measurements and the EHR to gather the remaining variables. The EHR data was queried from an underlying IBM DB2 database, and the LIS data has been queried from a Microsoft SQL database.

Multivariate time-series data with missingness

The clinical covariates collected were age, sex, height, weight, Charlson comorbidity score, temperature, blood pressure, respiratory rate, SpO2 levels, and diagnoses of chronic kidney disease (CKD) or Diabetes Mellitus (DM). Furthermore, we included lab test values from a basic metabolic panel (BMP), complete blood counts (CBC), and some less routine test results of relevance to COVID-19 as determined by scientific literature and physician collaborators. (Table 1) details the features collated into our database. In (Multimedia Appendix 1), we provide a detailed breakdown of these clinical covariates and laboratory values in our cohort (Table 5) as well as the cohort's geographic distribution (Figure 5).

Differentiating between missing data and absence of a condition is not possible from EHR diagnostic codes, particularly for the patients treated in an outpatient setting. Therefore, we focused mainly on the Charlson Comorbidity Index [13], which is populated in our EHR when there is a recorded medical history during a "patient encounter" in the EHR. Thus, this variable is available and can be assigned a value corresponding to no comorbidities, which is distinct from missingness in the case of no recorded medical history in the EHR. However, due to their emphasis within the literature, we also included CKD [9] and DM [14] as independent comorbidity variables using their ICD10 codes, while acknowledging that these variables conflate missingness with lack of a condition.

Table 1: Feature measurements collected

Abbreviation	**Description (units or levels)**
sex	Sex (male or female)
age	Age at time of PCR positive test (years)
weight	Weight (kg)
height	Height (cm)
PCR	SARS-CoV-2 nucleic acid test (+ or -)
SERO	SARS-CoV-2 serology antibody test (+ or -)
BASAA	Basophil count test (10^9/L)
EOSAA	Eosinophil count test (10^9/L)
HCT	Hematocrit test (%)
HGB	Hemoglobin test (g/dL)
LYMAA	Lymphocyte count test (10^9/L)
MCV	Mean corpuscular volume test (fL)
MONAA	Monocyte count test (10^9/L)
NEUAA	Neutrophil count test (10^9/L)
PLTC	Platelet count test (10^9/L)
RBC	Red blood cell count test (10^{12}/L)
RDW	Red cell distribution width test (%)
WBC	White blood cell count test (10^9/L)

CRP	C-reactive protein test (mg/L)
D-DIMER	D-dimer test (ng/mL)
FERR	Ferritin test (μg/L)
IL6	Interleukin-6 test (pg/mL)
TRPS	Troponin T test (ng/L)
FIBTP	Fibrinogen test (mg/dL)
LD	Lactate Dehydrogenase test (U/L)
IRON	Serum iron test (μg/dL)
TIBC	Total iron binding capacity test (μg/dL)
SAT	Percent iron saturation test (%)
TRSFC	Transferrin test (μg/dL)
BUN	Blood urea nitrogen test (mg/dL)
CHL	Chloride test (mmol/L)
GLU	Glucose test (mg/dL)
CALC	Calcium test (mg/dL)
CREA	Creatinine test (mg/dL)
POTA	Potassium test (mmol/L)
ALB	Albumin test (g/dL)
BICA	Bicarbonate test (mmol/L)
SODI	Sodium test (mmol/L)
BILI	Bilirubin test (mg/dL)
BPsystole	Blood pressure systole (mmHg)
BPdiastole	Blood pressure diastole (mmHg)
Temp	Temperature (C)
Pulse	Heart rate (1/min)
Resp	Respiratory rate (1/min)
SpO2	SpO2 oxygen saturation (%)
Charlson	Charlson Comorbidity Index (10-year survival probability)
CKD	Chronic kidney disease (+ or -)
DM	Diabetes Mellitus (+ or -)

Clinical covariates such as pre-existing conditions, height and weight are sampled infrequently, whereas heart rate and SpO2 are recorded every fifteen minutes for inpatients in our EHR, and other laboratory tests are intermediate in terms of frequency. Therefore, to deal with these multiscale timeseries measurements, we use the laboratory measurements as the starting point to define our sampling time points. For the variables of sex, age, weight, height, DM, CKD, and Charlson, we encoded these variables to exist at the first time point only; in our top performing RNN models, we observed no difference in performance using this strategy when compared to repeating the observations at each time point. For the frequently observed variables of BPsystole, BPdiastole, Temp, Pulse, Resp, and SpO2, we computed the minimum and maximum measurement for each calendar day and appended these to each laboratory time point during those dates; if no laboratory time point existed on a given day, we created a new one at noon using these minimum and maximum values. We considered time points within ±72 hours of each patient's first positive PCR result, and perform a sensitivity analysis on the length of patient follow up after this positive test result.

Time-flattened machine learning models

Time series data were flattened/encoded to a fixed length list of features by carry forward imputation (i.e., selection of the most recently observed covariate values), ensuring compatibility with traditional ML models. Specifically, after the data is flattened in this fashion, it forms a tabular prediction task suitable for any canonical supervised classification algorithm. The recently published [15] python-based autoML tool AutoGluon-Tabular (v0.2.0) was utilized to enable standardized and reproducible ensemble stacking of many model classes (e.g., deep neural networks, LightGBM boosted trees, CatBoost boosted trees, Random Forests, Extremely Randomized Trees, XGBoost, and kNearest Neighbors).

AutoGluon-Tabular models were fit to our tabular data frames using the "AutoGluon.TabularPrediction.fit" function using all the default parameters except eval_metric='roc_auc'. After running the fit function, access to each individual model created by AutoGluon was achieved by the "get_model_names" method on the resulting prediction object. This then allowed us to pass the specified model to the "predict_proba" method's optional "model" argument for each of the following model types: KNeighborsUnif, KNeighborsDist, NeuralNetFastAI, LightGBMLarge, NeuralNetMXNet, RandomForestGini, ExtraTreesGini, RandomForestEntr, ExtraTreesEntr, LightGBM, XGBoost, LightGBMXT, CatBoost, WeightedEnsemble_L2. We refer to WeightedEnsemble_L2 as simply the "AutoGluon" model in the subsequent since this was the output of the "predict_proba" method when no single model type was specified.

For relatively static features such as height, weight, or Charlson, we would expect the time flattened models to be at no disadvantage, whereas the more frequently measured data such as lab values or blood pressure will lose information, particularly about trends in the covariates. For instance, two individuals with a fever of 39C recorded in the most recent observation would be treated the same even if one had a sustained high fever and the other had a brief downward trending spike. Of course, there are many potential degrees of freedom to capture more information in the flattened data; one could define a fixed number of most recent observations, or fit a line through the observations over time and pass the slope and intercept as features to the classifier. But ultimately, the choice to flatten the time series is a choice of convenience and one that attempts to leverage the extensive research efforts devoted to tabular prediction, and so we study here only last observation carried forward modeling, since proper modeling efforts should account for the time series structure in the EHR data. We next look at models of this form.

RNN time series models

As a second approach, we implemented the modified Gated Recurrent Unit (GRU) binary classification models proposed by Che et al. [16] that are capable of accounting for the MNAR patterns within EHR data, and we adopt their notation. Namely, for a given patient, we have $D = 54$ variables and a given time series of T time points can be represented as a $T \times D$ matrix X whose rows $x_t \in \mathbb{R}^D$, $t = 1, \ldots, T$ represent the t-th observation with D variables x_t^d, $d = 1, \ldots, D$. Accompanying each observation x_t is a time stamp $s_t \in \mathbb{R}$, which starts at time zero $s_1 = 0$, and a binary masking vector $m_t \in \{0, 1\}^D$ with m_t^d taking value one when x_t^d is observed and zero otherwise. From these values, we can compute the time intervals.

$$\delta_t^d = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^d, & \text{if } t > 1, m_{t-1}^d = 0 \\ s_t - s_{t-1}, & \text{if } t > 1, m_{t-1}^d = 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

With these definitions, we can look at various modifications to the standard GRU architecture whose j-th hidden unit has a reset gate r_t^j and update gate r_t^j with hidden state h_t^j at time t and update equations

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (3)$$

$$\tilde{h}_t = \tanh(W x_t + U(r \odot h_{t-1}) + b) \quad (4)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

with matrices W_z, W_r, W, U_z, U_r, U and vectors b_z, b_r, b composed of model parameters, $\odot$ is the Hadamard product, $(\cdot)$ is the elementwise sigmoid function. Before modifying the architecture, there are three methods to use the GRU above to handle missing data: in "GRU-Mean" missing values are imputed by their means in the training data, in "GRU-Forward" missing values are imputed by their last observed value, and in "GRU-Simple" we simply concatenate the x_t ,m_t , $_t$ variables into a single observation vector $x't$ and pass this through the GRU equations above.
The GRU-D method uses trainable decay weights

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\} \quad (6)$$

with W_γ, b_γ being trainable model parameters. The observations are then replaced by the update.

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)(\gamma_{x_t}^d x_{t'}^d + (1 - \gamma_{x_t}^d)\bar{x}^d \quad (7)$$

where $x_{t'}^d$ is the last observed value of the d-th variable and $\bar{x}^d$ is the empirical mean of the d-th variable in the training data. The modified GRU update equations for GRU-D become the following.

$$r_t = \sigma(W_r \hat{x}_t + U_r \hat{h}_{t-1} + V_r m_t + b_r) \quad (8)$$

$$z_t = \sigma(W_z \hat{x}_t + U_z \hat{h}_{t-1} + V_z m_t + b_z) \quad (9)$$

$$\tilde{h}_t = \tanh(W \hat{x}_t + U(r \odot \hat{h}_{t-1}) + V m_t + b) \quad (10)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (11)$$

$$h_{t-1} = \gamma_{h_t} \odot h_{t-1} \quad (12)$$

where V_z, V_r, V are new model parameters to directly handle the masking vector m_t in the model.

Our implementation of the above equations in python is a slightly modified version of the code available on the GRU-D paper's [16] GitHub repository. Namely, for the core RNN algorithms we only edited the original GRU-D code where required to be compatible with the more recent versions of tensorflow.keras (version 2.1.0) and numpy (version 1.19.2) used in our high-performance computing cluster environment. We selected the specific RNN algorithm by setting the "--model" argument to be "GRUforward", "GRU0", "GRUsimple", and "GRUD" for GRU-forward, GRU-mean, GRU-simple, and GRU-D, respectively. We utilized the default hyperparameters of the algorithm, however in our testing found that increasing the batch size from 32 to 256 facilitating faster training of the algorithms. Therefore, a batch size of 256 is the only non-default hyperparameter selection made in our implementation of the RNN algorithms.

Temporal cohort split

As depicted in the consort diagram of (Figure 1), patients who first tested positive for COVID-19 from March 1, 2020 through December 15, 2020 (N=9,435, 80%) were assigned to a model selection cohort, whereas patients who first tested positive for COVID-19 from December 16, 2020 through January 27, 2021 (N=2,372, 20%) were used as a prospective testing cohort for the final algorithm. All experiments in the model selection cohort were performed using an identical 10-fold stratified cross-validation using binary classification with the positive class defined as death within 21 days of first positive PCR test. Only the single best performing model was evaluated on the prospective cohort after being fit against the entire model selection cohort.

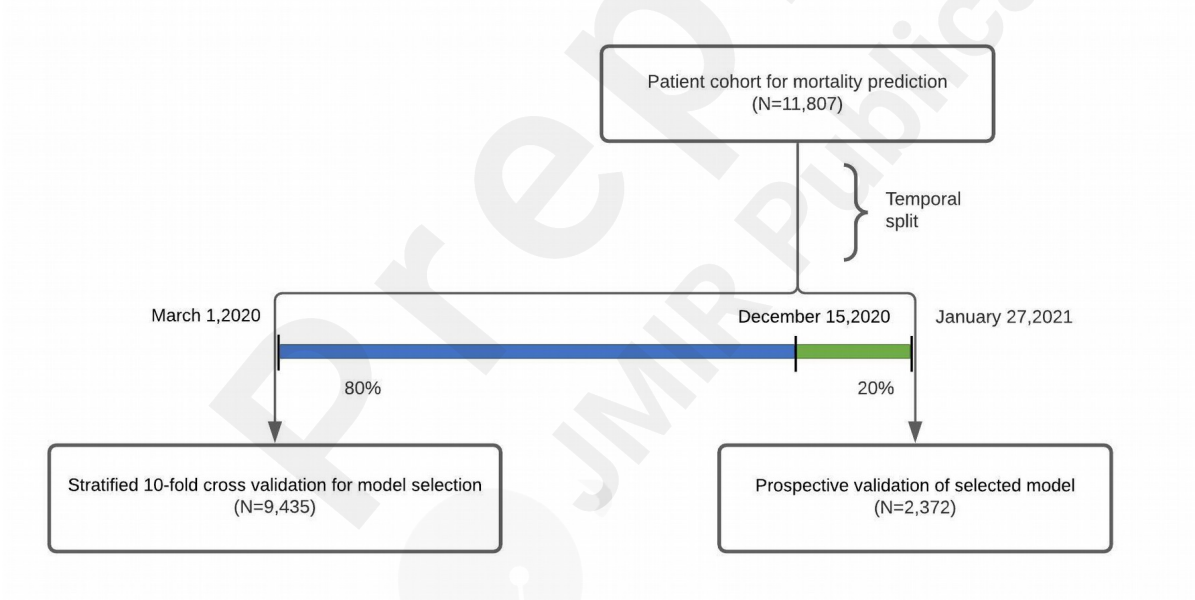


Figure 1: Consort diagram demonstrating the temporal split of our cohort for the purposes of model selection and prospective validation.

Results

Model selection

In (Figure 2) and (Table 2), we compare the results of our various models using cross-validation

AUROC in the training cohort. Although not in a statistically significant way, we recapitulate the findings of Che et al. [16], discovering the GRU-D model has the highest average cross-validation AUROC among all other standard variants of GRU modeling in time series with missing values. In addition, GRU-Simple has higher average cross-validation AUROC than the GRU-Forward and GRU-Mean and the most notable difference underlying these categories is the inclusion of missingness indicators as features to GRU-Simple, which could indicate the value of MNAR patterns in the classification task. GRU-D's biologically inspired architecture attempts to make even more efficient use of this information and exceeds the performance of all tested RNN methods.

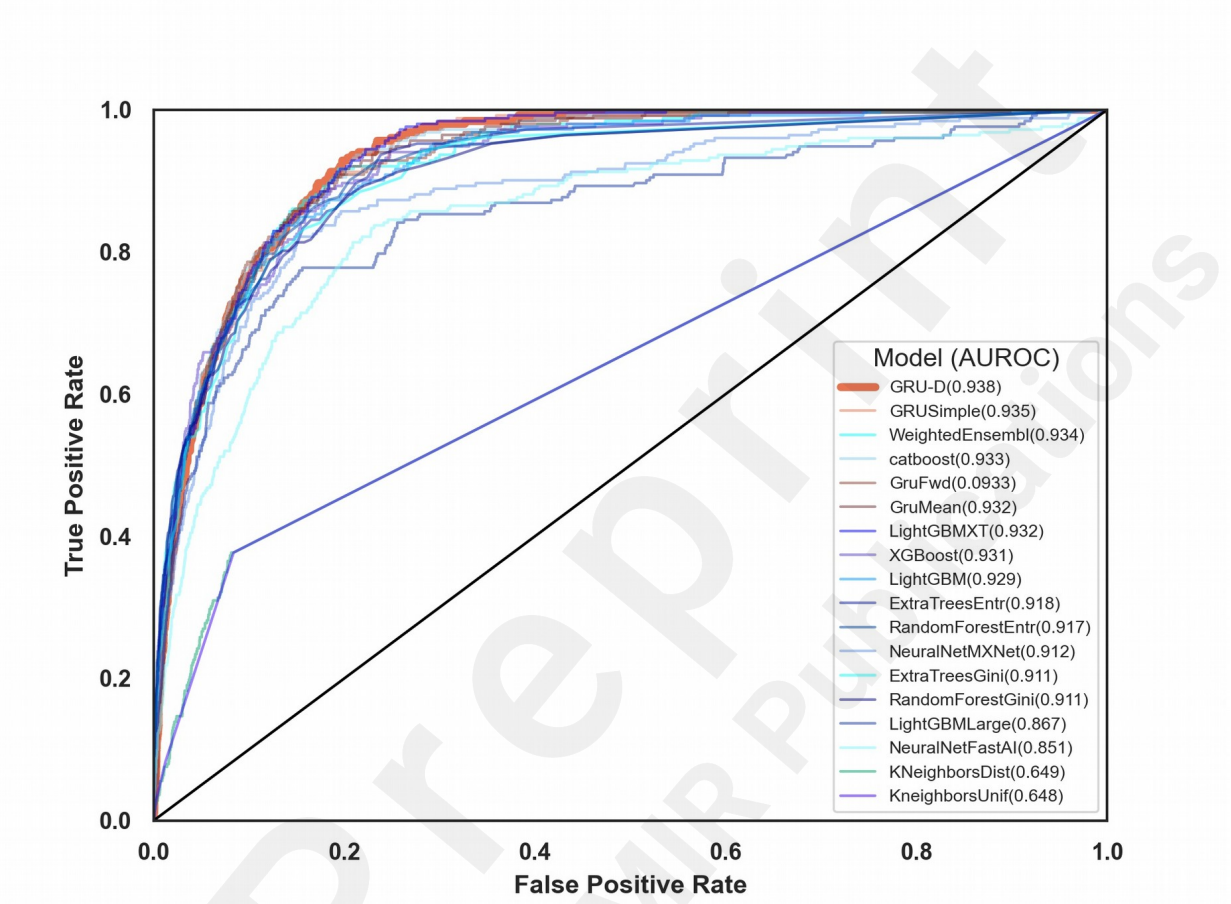


Figure 2: The receiver operating characteristic (ROC) curves for the eighteen models evaluated.

AutoGluon, which only had access to the last measurement of each variable, showed strong performance despite this limitation. In (Table 2), each individual AutoGluon model was also benchmarked (those with suffix "-AG"), along with the final ensemble estimate (labeled simply "AutoGluon"). While GRU-D ultimately outranked AutoGluon, each method's performance fell within the other's standard error intervals. AutoGluon's automated hyperparameter tuning and model stacking may indicate that GRU-D could benefit from the addition of hyperparameter search. However, this process may risk overfitting this cross-validation dataset, and thus we select GRU-D with the default settings rather than attempting to further improve the cross-validation AUROC via hyperparameter optimization.

Table 2. Modeling results sorted by performance

Model	AUROC (± standard error)
KNeighborsUnif-AG	0.648 (±0.011)
KNeighborsDist-AG	0.649 (±0.011)
NeuralNetFastAI-AG	0.858 (±0.013)
LightGBMLarge-AG	0.867 (±0.014)
NeuralNetMXNet-AG	0.907 (±0.008)
RandomForestGini-AG	0.911 (±0.007)
ExtraTreesGini-AG	0.911 (±0.009)
RandomForestEntr-AG	0.917 (±0.008)
ExtraTreesEntr-AG	0.918 (±0.007)
LightGBM-AG	0.929 (±0.007)
XGBoost-AG	0.931 (±0.006)
LightGBMXT-AG	0.931 (±0.005)
GRU-Mean	0.932 (±0.005)
GRU-Forward	0.933 (±0.006)
CatBoost-AG	0.933 (±0.005)
AutoGluon	0.934 (±0.005)
GRU-Simple	0.935 (±0.004)
GRU-D	0.938 (±0.004)

Length of time series

Clearly, we would expect availability of more time series data to result in improved model performance. To determine if predictions could be made utilizing data prior to 72 hours of a patient's first positive PCR test, we assessed the performance of GRU-D when we restrict the time series to 12, 24, 48, and 72 hours of follow up after the first positive PCR test. The results in (Table 3) demonstrate that although we lose performance when predicting earlier in the patient's disease, we are still able to provide accurate predictions even using data within the same day (12 hours of follow up) that a patient tests positive for COVID.

Table 3. GRU-D performance versus length of time series

Follow up after positive PCR	AUROC (± standard error)
12 hours	0.916 (±0.005)
24 hours	0.919 (±0.006)
48 hours	0.925 (±0.005)
72 hours	0.938 (±0.004)

MNAR as an asset and feature importance

To demonstrate the fact that MNAR data can improve model predictions by GRU-D, we generated a synthetic data set with lab test values replaced by Bernoulli coin flips. Therefore, the only valuable information contained within this dataset's laboratory values is the missing data patterns that can be viewed as encoding clinical suspicion or concern. For instance, the D-dimer lab value is ordered less

frequently than other tests, and so its presence alone can be informative of clinical concern for thrombotic events.

Our results found that randomizing the laboratory values resulted in an AUROC of 0.904 (±0.006), which indicates that the laboratory values in aggregate contributed 0.034 to the AUROC score (since this is the drop in performance compared to the model with actual laboratory values). We ran a further experiment omitting the laboratory values entirely, which produced the lower AUROC of 0.890 (±0.006). Therefore, the missing patterns alone contributed 0.014 to the AUROC. To contextualize this finding, we dropped each feature individually from the model, assessed the decrease in AUROC score and summarized the top ten features in decreasing order of the difference in the AUROC score (Figure 3). We note here that the drop due to missing patterns exceeds the drop due to removing any single variable from the analysis, making the MNAR pattern one of the most valuable pieces of information available to GRU-D.

In (Multimedia Appendix 1), we perform a detailed error analysis of our model using these top ten features. The fact that age and Charlson comorbidity index are the most significant contributors to mortality prediction are consistent with the well-known risk factors for COVID mortality. FIBTP, IRON and FERR were the three most important laboratory values in our models. Presence of chronic kidney disease, weight, serology test and SpO2 are clinical covariates that also rank in the top ten variables by importance. Interestingly, height has low importance indicating that BMI may not be as effective as weight itself in mortality prediction. However, a limitation of this drop-one-feature variable importance is that a low-ranking feature such as height cannot be said to be irrelevant, just that any information it carries is redundant within other features.

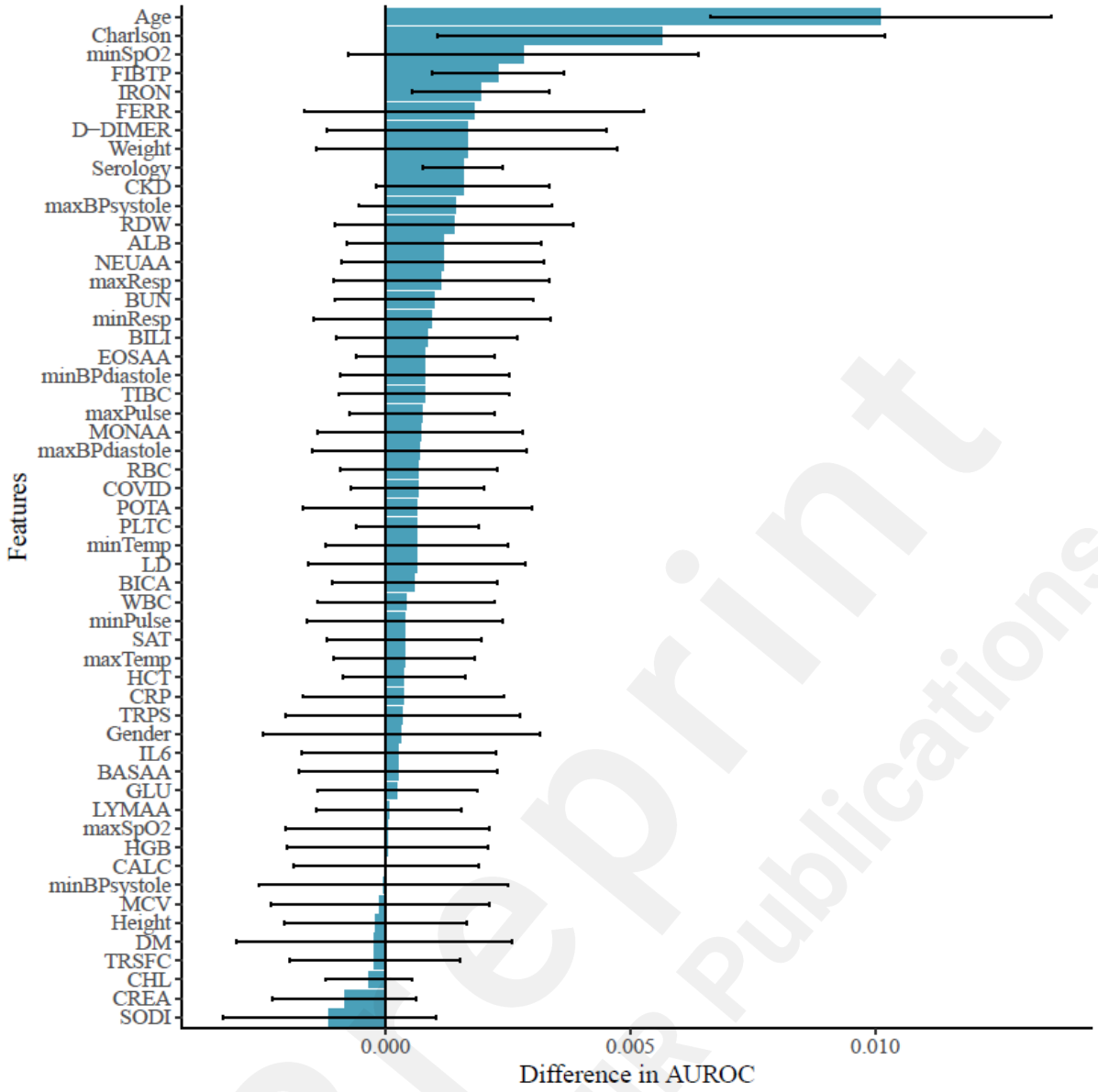


Figure 3: The feature importance in the GRU-D RNN model as defined by the average drop in AUROC (with 95% confidence intervals) when each feature is individually removed from the analysis. (Table 1) explains all the variable abbreviations, and the top five features are seen to be Age, Charlson comorbidity index, minSpO2, FIBTP and IRON.

Prospective Validation and Survival analysis

To demonstrate the efficacy of our proposed mortality prediction, we performed a Kaplan Meier analysis using the survival R library [17]. Specifically, we chose a decision boundary on the GRU-D model's ROC curve which provided a specific delineation of high-risk and low-risk groups of patients. In our cross-validation cohort binary classification provides an accuracy of 89% (95% confidence interval: [88,90]), a recall of 80% (95% confidence interval: [74,85]), a precision of 17% (95% confidence interval: [15,19]), and a negative predictive value (NPV) of 99% (95% confidence interval: [99,100]). Furthermore, although the precision is somewhat low with numerous false positives, we see among the survivors over twice the rate of mechanical ventilation or ECMO when they are predicted to die by GRU-D (Fisher's exact test $P<.001$, odds ratio 2.1, 95% confidence

interval [1.8, 2.5]). We validated this performance in our prospective testing cohort, finding an AUROC of 0.901, an accuracy of 78% (95% confidence interval: [76,79]), a recall of 85% (95% confidence interval: [77,91]), a precision of 14% (95% confidence interval: [12,17]), and a negative predictive value (NPV) of 99% (95% confidence interval: [99,100]).

Our Kaplan Meier analysis results in (Figure 4) demonstrate the statistically significant stratification provided by our machine learning model in both the cross-validation and prospective testing experiments. Building a Cox Proportional Hazards model for our prediction in the cross-validation cohort provides a statistically significant difference in survival between the two groups (P<.001 for the likelihood ratio, logrank, and Wald tests), with a prediction of survival having a substantially improved hazard ratio of 0.053 (95% confidence interval: [0.043,0.066]). We validated this finding in the prospective testing cohort with a statistically significant difference in survival between the two groups (P<.001 for the likelihood ratio, logrank, and Wald tests), with a prediction of survival having a substantially improved hazard ratio of 0.067 (95% confidence interval: [0.043,0.106]).

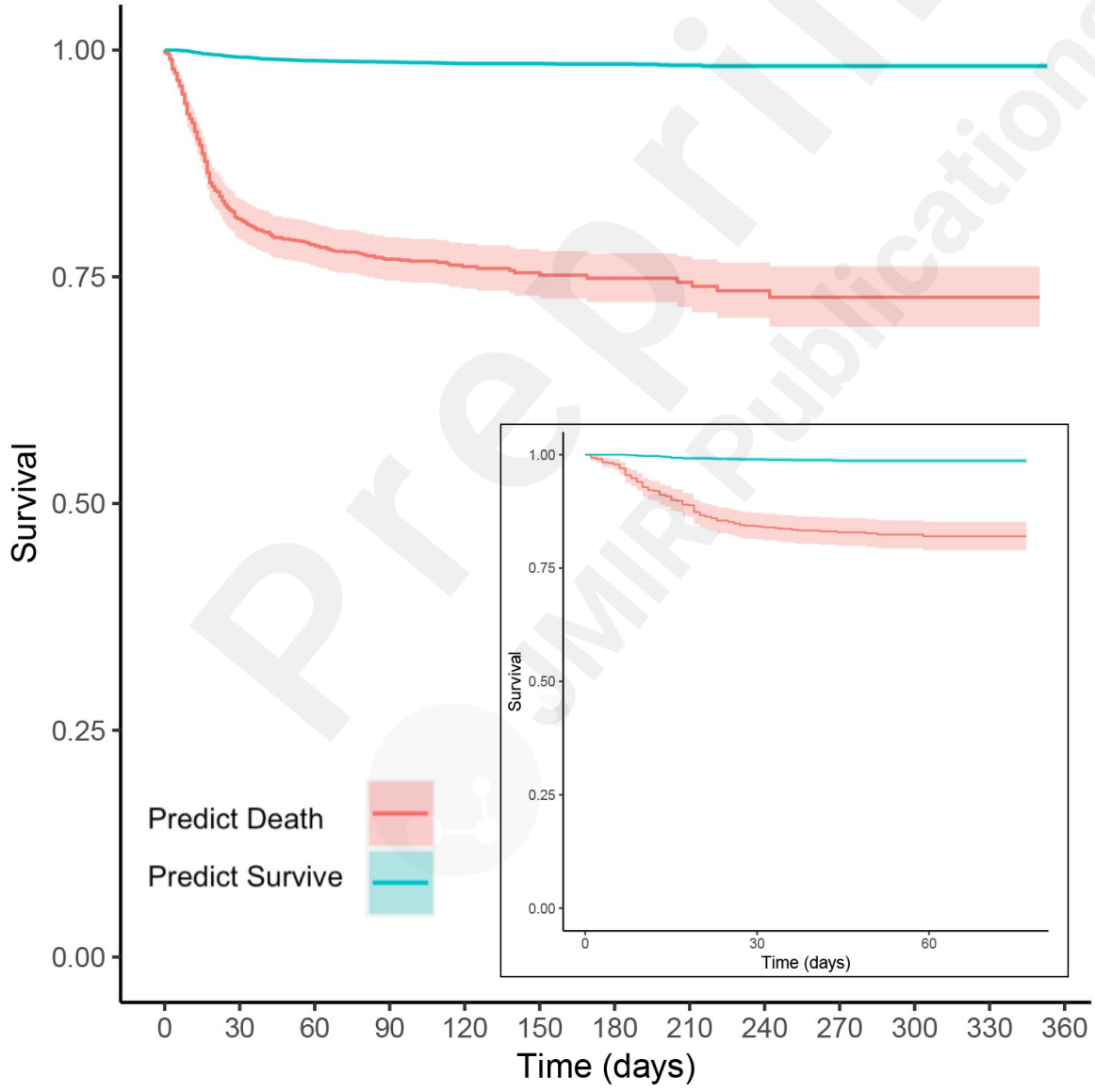


Figure 4. Kaplan Meier survival curves for the GRU-D stratified populations in the cross-validation cohort (main figure) and the prospective test cohort (inset), where teal is a prediction of low risk of death and red is a prediction of high risk. Both figures have 95% confidence bands visualized for the teal and red curves, although the teal confidence bands are tight due to our large sample sizes.

Discussion

In this study, we collected and processed over fifty laboratory and clinical covariates in a population of nearly twelve-thousand Mayo Clinic patients who tested positive for SARS-CoV-2 by PCR. In this large and geographically diverse dataset, we found that the GRU-D recurrent neural network could provide state-of-the-art mortality prediction. This performance remained strong even in a held-out test set that mimics how a deployed system would be trained retrospectively and then prospectively utilized in a clinically evolving pandemic setting.

Principal Results

Our cross-validation experiments summarized in (Table 2) indicated that the top performing model to predict mortality in our cohort was the GRU-D recurrent neural network. We thus selected the GRU-D method to predict mortality of COVID-19 patients, and prospectively found an AUROC of 0.901, an accuracy of 78% (95% confidence interval: [76,79]), a recall of 85% (95% confidence interval: [77,91]), a precision of 14% (95% confidence interval: [12,17]), a negative predictive value (NPV) of 99% (95% confidence interval: [99,100]), and statistically significant difference in survival (P<.001, hazard ratio for those predicted to survive: 95% confidence interval [0.043,0.106]). As can be expected in prospective validation, we observed a modest drop in AUROC although most of the performance characteristics were close to their original cross-validation estimates. Namely, the NPV was largely unchanged, while precision and accuracy showed minor decreases with the recall showing modest improvements.

We chose a prospective/retrospective split in time since this is the most realistic way to assess the potential performance of the system if launched clinically, because it would be trained on data up until its go-live date and then run prospectively in a potentially evolving pandemic environment. Notably, the cutoff date for the 80/20 split creating the prospective test set was December 15, 2020, which is the day after the first COVID vaccine received FDA approval, meaning that our prospective cohort represented a distinctly different clinical environment compared to the period in which the model was trained. The relatively minor loss of performance in prospective validation shows the robustness of the modeling herein, but the observed loss of performance also demonstrates the need for continued retraining/validation of such a model during a constantly evolving pandemic.

Limitations

The application and deployment of ML methods in clinical practice require concerted care and diligence. One may be inclined to interpret the high NPV of our prediction algorithm as an indication that the best use of the algorithm in practice is as a screening mechanism to discharge patients who are not at risk in order to conserve resources for higher risk individuals. However, such a conclusion illustrates a pitfall of using a correlative prediction algorithm to make causal conclusions. The algorithm is highly confident that under the current standards of care at Mayo Clinic these individuals are not likely to succumb to their illness; this is quite distinct from asserting that it is safe to reduce the care for these patients. Arriving at this latter conclusion would likely require a randomized control trial, and given the much lower survival rate published in the NYC dataset [11] where medical systems were overcapacity it seems unlikely that reducing care from those who survived in our cohort would have been a safe measure. Because the Mayo Clinic health systems have not been overcapacity, our mortality predictions should be viewed as representing patient

stratification when full clinical support is available.

Therefore, we conclude that the algorithm is better deployed as an alert system that flags only those patients it deems as high risk to provide the treating physician with an additional datapoint that aims to summarize the many covariates and laboratory values routinely available. In this context, the algorithm has had abundant experience in the provider's system, effectively "seeing" all COVID-19 patients that have attended Mayo Clinic and conveying these lessons to physicians who could not have gained such experience personally.

A web interface to this model may allow for widespread usage but given the complexity and error-prone nature of users providing the high dimensional time-series measurements with correct units, the system is better suited for integration within the EHR/LIS infrastructure. We are now exploring the details of deployment of such a GRU-D alert system, which involves discussions with physicians to assess numerous implementation details. For example, deciding whether the alerts would be passive EHR/chart-based flags or a direct page to the frontline clinical provider. Passive chart alerts are less intrusive to existing workflows (i.e., a direct page interrupts a physician while tending to other patients) but also provide less-immediate feedback. Additionally, active alerts could also be sent to a triage group to consider if evaluation is needed (for example, from the registered respiratory therapist) rather than interrupting bedside clinicians. Furthermore, for either type of alert there is the question of prescribing a bedside assessment or leaving it to provider discretion, which is again a matter of balancing disruption of workflow with likelihood of missing a critical event. There will not be a universally appropriate implementation for all hospital systems, due to staffing and procedural differences. However, since our algorithm predicts overall COVID-19 mortality and is not tailored to flag imminent events such as cardiopulmonary arrest, it may be appropriate to consider less intrusive chart alerts without prescribed bedside follow up.

We have also seen nuances in the challenges and opportunities presented by MNAR data. In the context of traditional statistical inference and imputation, MNAR data is a worst-case scenario so challenging that many practical applications effectively ignore the reality and proceed with algorithms designed for the missing completely at random (MCAR) or missing at random (MAR) settings. A diligent statistician making this decision may perform a sensitivity analysis under a very limited set of assumed MNAR mechanisms to provide some assurances regarding the robustness of the chosen imputation or analytical strategy [18].

However, here we have demonstrated that classification problems can be quite distinct in this regard. Specifically, if the missing data mechanism is tightly coupled to the ultimate prediction task, it is entirely possible for MNAR data to be an asset rather than an impediment. One can construct a context where the class label is so tightly linked to the missing data mechanism that the patterns of missingness provide more discriminative power than the underlying values themselves (see Multimedia Appendix 2). In LIS systems, the number of potential laboratory tests that could be ordered at any time is astronomical, and it is unlikely that a practicing physician will ever order a "complete observation" of every test available on a single patient at every point in time. Instead, tests are ordered based on reasonable clinical suspicion that a test might return an abnormal result. From a prognostication point of view, this clinical suspicion is an enormously valuable piece of information that will almost never be captured in a structured data field in the EHR. If an algorithm cannot build off of this clinical suspicion as a starting point, it is also likely that its conclusions may appear to be a "step behind" the ordering clinician. Instead, an algorithm should learn what it can from the MNAR data patterns (here partly encoding clinical suspicion) in addition to the final value returned by the laboratory test.

We also note some of the real-world challenges that are faced when attempting to deploy such an alert system into clinical practice. First, in the retrospective experimental design followed here and by other papers in the literature, the time series data are constructed using the time of sample collection since this is the most biologically accurate way to represent the data and build predictive models. However, in practice if there can be delays in turn around for certain tests, this will either result in delayed predictions (so that the deployed testing data matches its retrospective training counterpart) or result in biased predictions when delayed labs are treated as missing. Therefore, although 72 hours is early in the course of illness, it is crucial that we have demonstrated reasonable performance even when only considering data collected on the same day as the first positive PCR, because a real-world delay of 48 hours on certain lab values may occur during a global pandemic and thus it is critical that the system can still provide accurate and timely predictions even when labs are delayed. Additionally, with vaccines now being delivered the models presented herein should be considered as mortality predictions for an unvaccinated individual, and in practice a vaccinated individual will be expected to be at low risk for mortality based on the clinical trials data.

Another challenge in dealing with LIS data comes from non-standardization of test coding prior to reporting out to the EHR. In a multisite system the same laboratory test may have multiple test codes to account for the different ordering facilities or variability in local billing regulations. This creates the potential for discrepancies in values stored within the underlying database such as differing units of measure. Substantial effort is therefore devoted to linking the LIS results to the EHR to ensure consistency across test codes and complete coverage of results in the EHR. The COVID-19 pandemic has created added complexity due to the rapidly evolving and continuously updating availability of COVID-19 nucleic acid and antibody tests. Therefore, effective data collection and deployment of machine learning methodologies necessitates extensive team-based laboratory and medical expertise to ensure that data aggregation and modeling efforts can be rapidly modified to suit the changing nature of the underlying dataset. Scalability also presents practical challenges. This is illustrated by a scenario in which internal workflows began to fail due to limitations on number of query results being returned by Tableau, necessitating that SQL queries take place on a high-performance computing cluster using a python/pandas toolchain. While these logistical challenges may be of limited academic interest, they are important to document, as such barriers have been a greater impediment to rapid real world deployment than more traditional topics in the machine learning literature such as identification of appropriate classification algorithms.

Comparison with Prior Work

For context, in (Table 4) we summarize some of the largest published COVID-19 mortality studies and specifically the cohorts analyzed, and the most relevant features identified. When smaller cohorts see insufficient numbers of deaths for direct mortality prediction, studies tend to focus on prediction of severe outcomes.

For instance, in a cohort of 123 patients with COVID-19 of Vulcan Hill Hospital, China, by Pan et. al [19] the mortality classifier based on XGBoost yielded an AUC of 0.86-0.92. Likewise, in a cohort of 372 Chinese cases (99.7% cohort survival rate) Gong et al. [9] found that the following variables provided an AUROC of 0.85. Similarly, in a study of 375 COVID-19 patients conducted by Ko et al. [20], the mortality prediction model based on XGBoost had an accuracy of 92%. In a study of 398 COVID-19 positive patients by Abdulaal et. al [21], an accuracy of 86% was achieved (95% CI 75%-93%). In a larger study of 2160 cases over 54 days from three hospitals in Wuhan China with sufficient cases to assess mortality (88% cohort survival rate), Gao et al. [8] reported 0.92-0.98 AUROC using an ensemble classifier. Furthermore, Vaid et al. [11] used 4098 inpatient cases over 68

days in New York City (83% cohort survival rate) to achieve AUROC of 0.84-0.88 in mortality prediction. Distinct from mortality prediction, Kim et. al [22] studied 4787 patients and their XGBoost based classifier demonstrated an AUC of 0.88-0.89 (95 % CI 0.85-0.91) in predicting the need for intensive care. Also distinct from mortality prediction, Bolourani et. al [23] used 11,525 patients to achieve AUROC of 0.77 in predicting respiratory failure within 48 hours of admission based on data from emergency department, using a XGBoost model.

The dramatically different cohort mortality rates and associated predictive accuracies, may be in part due to the differing straining of the local healthcare systems at times of study (both Wuhan and NYC experienced waves of patients that at different times overwhelmed health infrastructure), and the relatively geographically narrow nature of each of these data sets underscores why it is unlikely that these mortality predictions would extend directly to our patient population in a health care system spanning three time zones and multiple locales unrepresented in the literature.

Table 4. Summary of related studies

Study	Number of patients	Model/ algorithm	Cohort survival	Prediction	AUROC	Feature Importance
Pan et. al [19]	123	XGBoost	52.8%	Mortality	0.86-0.92	Lymphocyte percentage, prothrombin time, lactate dehydrogenase, total bilirubin, eosinophil percentage, creatinine, neutrophil percentage, and albumin level
Gong et al. [9]	372	Nomogram	99.7%	Severity	0.85 (95% CI 0790-0.916)	Higher lactate dehydrogenase, C-reactive protein, red blood cell distribution width, direct bilirubin, and blood urea nitrogen; and lower albumin
Ko et al. [20]	375	XGBoost	98.1%	Mortality	Not available. Accuracy 92%	Not assessed
Abdulaal et. al [21]	398	Artificial Neural Network	Not available	Mortality	Not available. Accuracy 86% (95% CI 75%-93%)	Altered mentation, dyspnea, age, collapse, gender and cough
Shi et al. [10]	487	Custom risk score calculation	100%	Severity	Not available	Advanced age, presence of hypertension, and being male
Gao et al. [8]	2160	Ensemble model based on Logistic Regression, Gradient Boosted Decision Tree, Neural Network and Support Vector Machine	88%	Mortality	0.92-0.98	Consciousness, Chronic Kidney Disease (CKD), lymphocyte counts, sex, sputum, blood urea nitrogen, respiratory rate, SpO2 oxygen saturation, D-Dimer, number of comorbidities, albumin, age, fever, and platelet count
Vaid et al. [11]	4098	XGBoost	83%	Mortality	0.84-0.88	Age, anion gap, C-reactive protein, lactate Dehydrogenase, SpO2 oxygen saturation, blood urea nitrogen, ferritin, red cell distribution width (RDW), and diastolic blood pressure.
Kim et. al	4787	XGBoost	Not	Need for	0.88-0.89	ADL, age, dyspnea, body

[22]			available	intensive care		temperature, sex and underlying comorbidities
Bolourani et. al [23]	11525	XGBoost	Not available	Predicting respiratory failure	0.77	Invasive mode of oxygen delivery being a nonrebreather mask, Emergency Severity Index, (ESI) values of 1 and 3, maximum respiratory rate, maximum, oxygen saturation, Black race, age on admission, eosinophil percentage, serum sodium level, and serum lactate level.
This manuscript	11807	GRU-D	95.4%	Mortality	0.938 cross-validation; 0.901 prospectively	(Figure 2). Top 5: Age, Charlson comorbidity index, minSpO2, FIBTP and IRON

As indicated in (Table 4), the present paper represents the largest cohort collected for mortality prediction in COVID-19, and the GRU-D algorithm shows state-of-the-art performance. Notably, many papers selected models based on XGBoost, which also showed strong cross-validation performance in our data. However, (Table 2) demonstrates that it was not even in the top 5 algorithms that we assessed. Additionally in agreement with Gao et al. [8], we find that ensemble algorithms such as AutoGluon can provide stronger performance, although as noted previously the GRU-D algorithm ended up ranked most highly in our cross-validation experiments.

Conclusions

We have aggregated and analyzed one of the largest multistate COVID-19 EHR databases for mortality prediction. Using this database, a highly effective machine learning algorithm using the GRU-D neural network architecture has been trained and prospectively validated to predict mortality of COVID-19 patients shortly after their first positive PCR test.

Acknowledgements

We thank the Advanced Diagnostics Laboratory, Department of Laboratory Medicine and Pathology, and Center for Individualized Medicine at Mayo Clinic for funding this research. We thank Dr. Nicholas Chia for his insightful discussions and feedback on this manuscript. We are grateful to the thorough and constructive comments of the anonymous reviewers whose feedback greatly improved our paper.

Conflicts of Interest

None declared.

Abbreviations

AUROC (Area Under Receiver Operator Characteristic Curve), AutoML (Automated Machine Learning), BMP (Basic metabolic panel), CBC (Circulating Blood Cell), CI (Confidence Interval), CKD (Chronic Kidney Disease), DM (Diabetes Mellitus), GDPR (General Data Protection Regulation), GRU (Gated Recurrent Unit), EHR (Electronic Health Record), MCAR (Missing Completely At Random), MAR (Missing At Random), MNAR (Missing Not At Random), ML (Machine Learning), PCR (Polymerase Chain Reaction), RNN (Recurrent Neural Network), XGBoost (Extreme Gradient Boost).

Multimedia Appendix 1

Cohort Description

In (Figure 5) we visualize the geographic distribution of our cohort within the continental United States. And in (Table 5) we provide statistical summaries of each covariate broken down by survival versus death in our cohort.

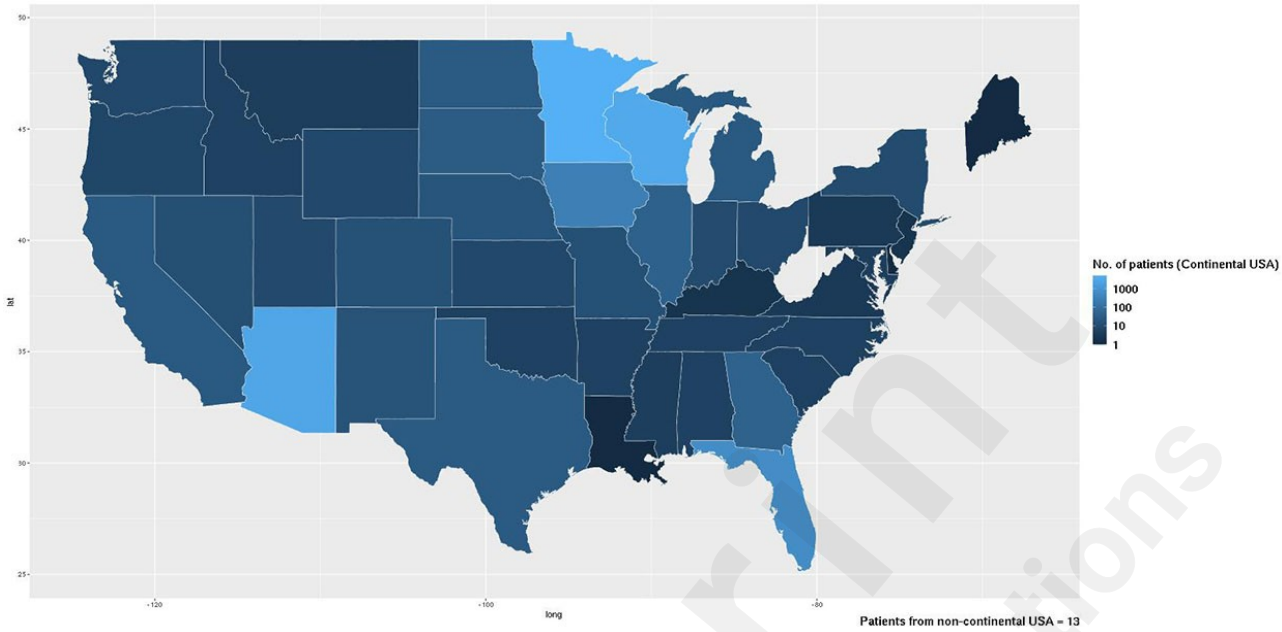


Figure 5: The geographic distribution of the Mayo Clinic cohort's home addresses. Thirteen patients are not represented in the map as they had addresses outside the continental United States.

Table 5. Cohort summary with categorical features summarized as number (percentage within group) and numerical features summarized as mean (standard deviation).

	TOTAL	SURVIVE	DEATH
	n = 11807	n = 11263	n = 544
FEMALE	6198 (52.5%)	5973 (53%)	225 (41.4%)
MALE	5609 (47.5%)	5290 (47%)	319 (58.6%)
NO CHRONIC KIDNEY DISEASE	9776 (82.8%)	9490 (84.3%)	286 (52.6%)
CHRONIC KIDNEY DISEASE	2031 (17.2%)	1773 (15.7%)	258 (47.4%)
NO DIABETES MELLITUS	8703 (73.7%)	8423 (74.8%)	280 (51.5%)
DIABETES MELLITUS	3104 (26.3%)	2840 (25.2%)	264 (48.5%)
AGE	55.19 (19.94)	54.20 (19.68)	75.66 (13.13)
BODY MASS INDEX	30.86 (7.96)	30.99 (8.02)	28.57 (6.39)
WEIGHT	89.17 (25.08)	89.49 (25.23)	82.69 (20.86)
HEIGHT	169.73 (11.17)	169.77 (11.22)	169.05 (10.10)
CHARLSON	66.18 (37.19)	68.53 (36.23)	28.69 (31.85)
BASPHIL COUNT TEST	0.03 (0.03)	0.03 (0.03)	0.02 (0.02)
EOSINOPHIL COUNT TEST	0.09 (0.08)	0.09 (0.08)	NA
HEMATOCRIT TEST	36.15 (6.04)	36.53 (5.88)	33.30 (6.54)
HEMOGLOBIN TEST	11.83 (2.18)	12.01 (2.11)	10.56 (2.28)
LYMPHOCYTE COUNT TEST	1.35 (4.98)	1.41 (5.28)	0.81 (0.62)
MEAN CORPUSCULAR VOLUME TEST	89.79 (7.25)	89.76 (6.98)	89.98 (9.15)
MONOCYTE COUNT TEST	0.54 (0.30)	0.55 (0.30)	0.49 (0.26)
NEUTROPHIL COUNT TEST	4.81 (3.37)	4.57 (3.14)	6.62 (4.37)
PLATELET COUNT TEST	205.54 (104.10)	210.35 (94.50)	171.33 (153.27)
RED CELL DISTRIBUTION COUNT TEST	4.05 (0.74)	4.09 (0.72)	3.74 (0.84)
RED CELL DISTRIBUTION WIDTH TEST	14.05 (2.07)	13.85 (1.91)	15.40 (2.56)
WHITE BLOOD CELL COUNT TEST	6.93 (5.39)	6.78 (5.46)	7.95 (4.77)
C-REACTIVE PROTEIN TEST	54.59 (56.23)	51.20 (53.34)	81.29 (69.74)
D-DIMER TEST	1642.09 (4231.08)	1388.55 (3182.19)	3690.41 (8728.90)
FERRITIN TEST	810.66 (1366.01)	748.53 (1116.00)	1240.68 (2442.42)
INTERLEUKIN-6 TEST	68.66 (152.18)	64.88 (149.61)	91.42 (166.51)
TROPONIN T TEST	42.14 (155.96)	35.73 (146.11)	95.69 (214.75)
FIBRINOGEN TEST	494.79 (173.78)	500.24 (171.15)	464.26 (185.68)
LACTATE DEHYDROGENASE TEST	329.31 (608.81)	301.21 (132.08)	519.10 (1650.79)
SERUM IRON TEST	52.36 (57.64)	53.09 (59.35)	46.27 (41.42)
TOTAL IRON BINDING CAPACITY TEST	218.20 (74.62)	220.29 (74.31)	200.86 (76.69)
PERCENTAGE IRON SATURATION TEST	24.25 (21.22)	24.09 (21.24)	25.59 (21.50)
TRANSFERRIN TEST	184.90 (63.25)	186.69 (62.99)	170.14 (64.95)
BILIRUBIN TEST	0.52 (0.70)	0.48 (0.33)	0.83 (1.81)
ALBUMIN TEST	3.45 (0.55)	3.50 (0.53)	3.07 (0.59)
BICARBONATE TEST	23.77 (3.36)	23.96 (3.23)	22.28 (4.00)
BLOOD UREA NITROGEN TEST	24.07 (17.11)	22.38 (15.37)	37.55 (23.36)
CREATININE TEST	1.38 (1.44)	1.33 (1.44)	1.77 (1.44)
POTASSIUM TEST	4.21 (0.51)	4.22 (0.50)	4.16 (0.53)
SODIUM TEST	138.09 (4.11)	138.08 (3.91)	138.15 (5.51)
CHLORIDE TEST	102.07 (4.81)	101.90 (4.54)	103.38 (6.42)

GLUCOSE TEST	134.75 (55.09)	134.20 (55.80)	138.93 (49.55)
CALCIUM TEST	8.52 (0.59)	8.55 (0.58)	8.28 (0.63)
MAXIMUM BLOOD PRESSURE SYSTOLE	135.35 (24.61)	135.73 (24.02)	130.84 (30.38)
MINIMUM BLOOD PRESSURE SYSTOLE	119.54 (15.99)	119.89 (15.90)	115.42 (16.56)
MAXIMUM BLOOD PRESSURE DIASTOLE	82.45 (11.16)	82.52 (11.02)	81.62 (12.64)
MINIMUM BLOOD PRESSURE DIASTOLE	71.47 (17.72)	71.80 (17.12)	67.57 (23.31)
MAXIMUM TEMPERATURE	98.58 (1.19)	98.57 (1.19)	98.71 (1.18)
MINIMUM TEMPERATURE	97.74 (2.00)	97.77 (2.02)	97.42 (1.61)
MAXIMUM PULSE	88.88 (18.34)	88.43 (18.01)	94.04 (21.19)
MINIMUM PULSE	70.34 (15.64)	70.77 (15.59)	65.33 (15.26)
MAXIMUM RESPIRATORY RATE	22.37 (7.43)	21.94 (7.08)	27.20 (9.31)
MINIMUM RESPIRATORY RATE	15.98 (3.41)	16.03 (3.22)	15.48 (5.05)
MAXIMUM OXYGEN SATURATION	97.67 (2.02)	97.67 (1.93)	97.59 (2.86)
MINIMUM OXYGEN SATURATION	92.37 (5.83)	92.86 (4.97)	86.80 (10.34)
NOT VENTILATED	8931 (75.6%)	8727 (77.5%)	204 (37.5%)
VENTILATED	2876 (24.4%)	2536 (22.5%)	340 (62.5%)
SURVIVE PAST 21DAYS	11454 (97%)	11263 (100%)	191 (35.1%)
DEATH WITHIN 21DAYS	353 (3%)	0 (0%)	353 (64.9%)

Error Analysis

To supplement the findings of our survival analysis, we embark here to study the false positives and negatives for the selected threshold on our ROC curve. To this end, we have examined the ten variables of highest importance to our model, stratified by the categories of our confusion matrix, i.e., true negative (TN), false positive (FP), false negative (FN), and true positive (TP). (Figures 6-14) represent the continuous variables whereas (Table 6) captures the data for CKD status. While CKD conflates missingness with lack of the condition, the continuous variables do not suffer from this limitation and therefore the percentage of patients with missing values in each confusion matrix category are included. When multiple time points were available for a given data type, only the most recent observation was included, in order to ensure only a single data point per patient.

Examination of the described results uncovers notable trends in the data. Firstly, the false negatives tend to have higher rates of missing data. Missing data not only hampers our model's ability to accurately gauge the outcome status of these patients, but it also indicates that–at least in the first 72 hours of confirmed infection the clinical suspicion on these patients was low. Compounding the issue of missing data is the fact that the more readily available metrics such as age trend in the direction of favorable prognoses for these false negative patients. For instance, the FN cohort is younger than the TP and FP predicted by GRU-D to be at high risk. Likewise, the Charlson estimated 10-year survival rate differs across our groups in a predictable manner, with predicted risk by GRU-D being anticorrelated with 10-year survival estimates, and our false negatives have correspondingly higher Charlson survival estimates. Interestingly, GRU-D is able to correctly classify both negatives and positives at the high and low end of the Charlson scale, indicating that GRU-D is not a simple recapitulation of Charlson predictions. Furthermore, the univariate trends between TP versus TN in the data presented are increased Age, FERR, D-DIMER; decreased Charlson 10-year survival rate, weight, FIBTP, IRON and MinSpO2. The FN and FP tended to be intermediate between these values, with FN more closely resembling TN and FP more closely

resembling TP as one would expect based on the behavior of a prediction algorithm.

Finally, we note that no univariate marker appears to cleanly discern the true labels (i.e., red versus green in the violin plots for death versus survival, respectively). Furthermore, the reductions in AUROC from dropping univariate markers in (Figure 3) were relatively small, indicating both redundancy in the information represented by these features and a general importance of multivariate context. (Figure 13), for instance shows no stratification of the individual groups, but this is likely due to the fact that covariates including sex, age and height will play a necessary role in con- textualizing weight from a clinical prognostication point of view. Ultimately however, the univariate trends displayed here provide a good first-order approximation to understanding the risk factors for mortality in COVID-19, which coincide with known biology.

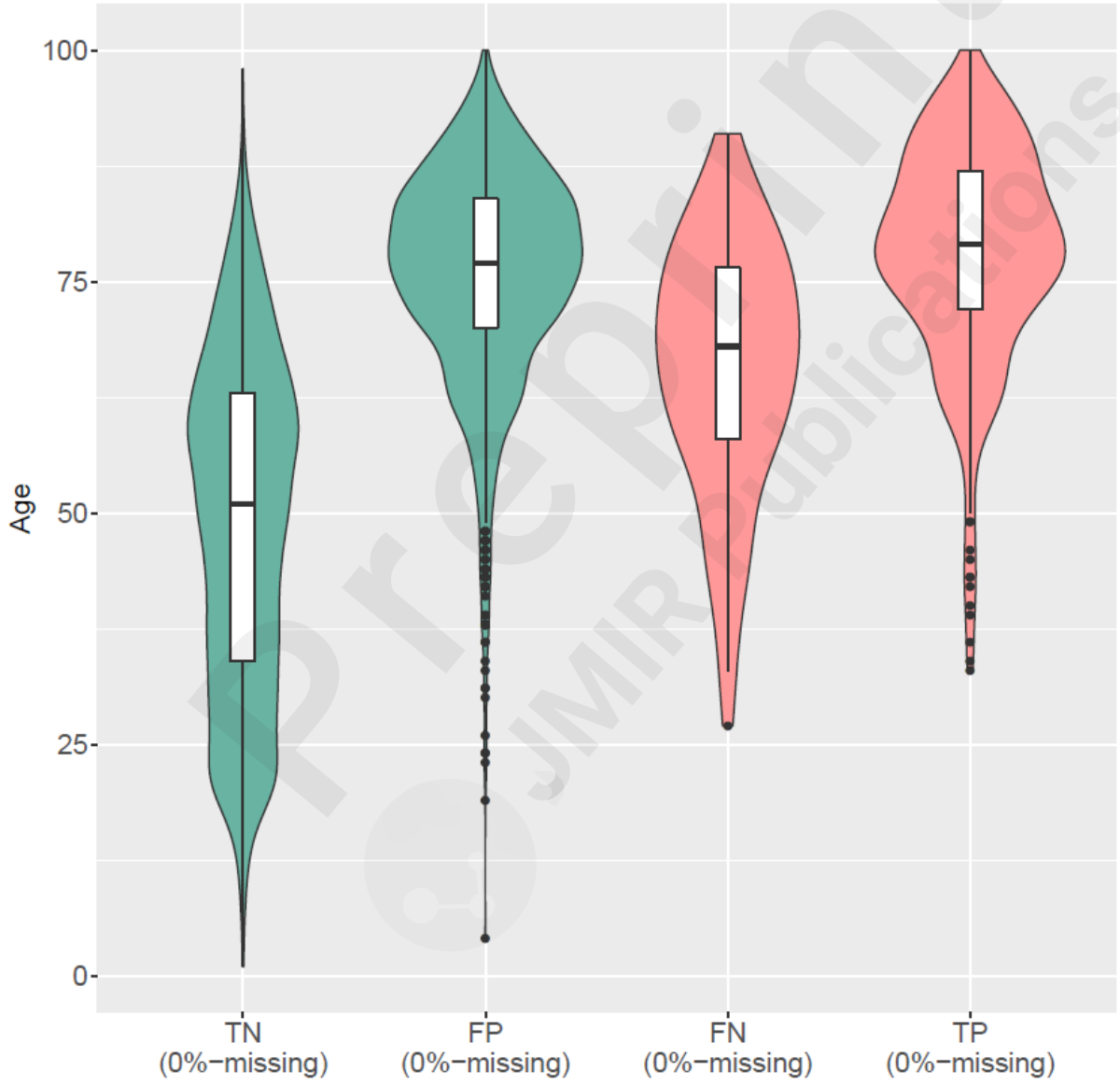


Figure 6: Violin plot of age (years) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at

the base of the plot.

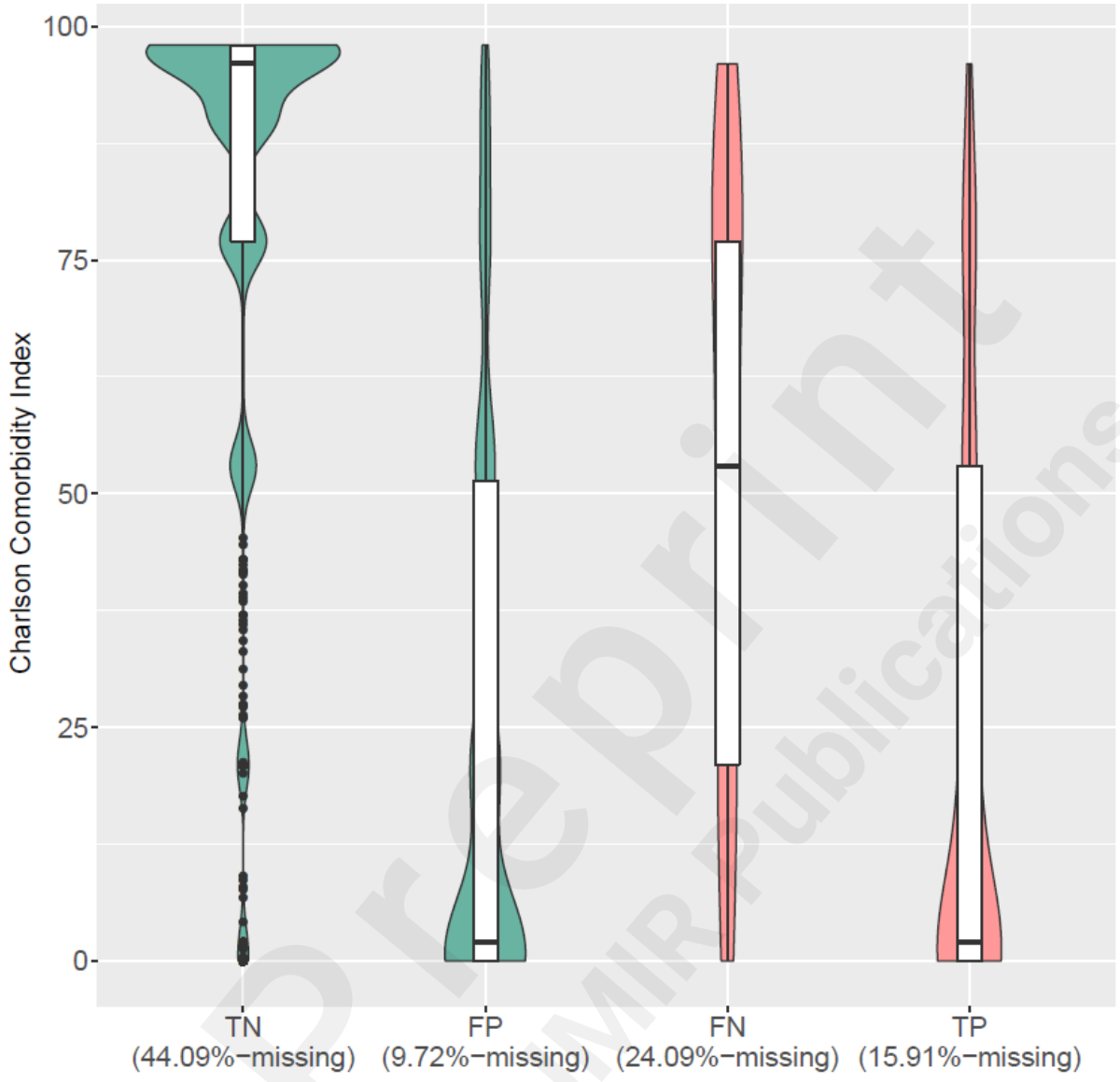


Figure 7: Violin plot of Charlson Comorbidity Index (10-year survival probability) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

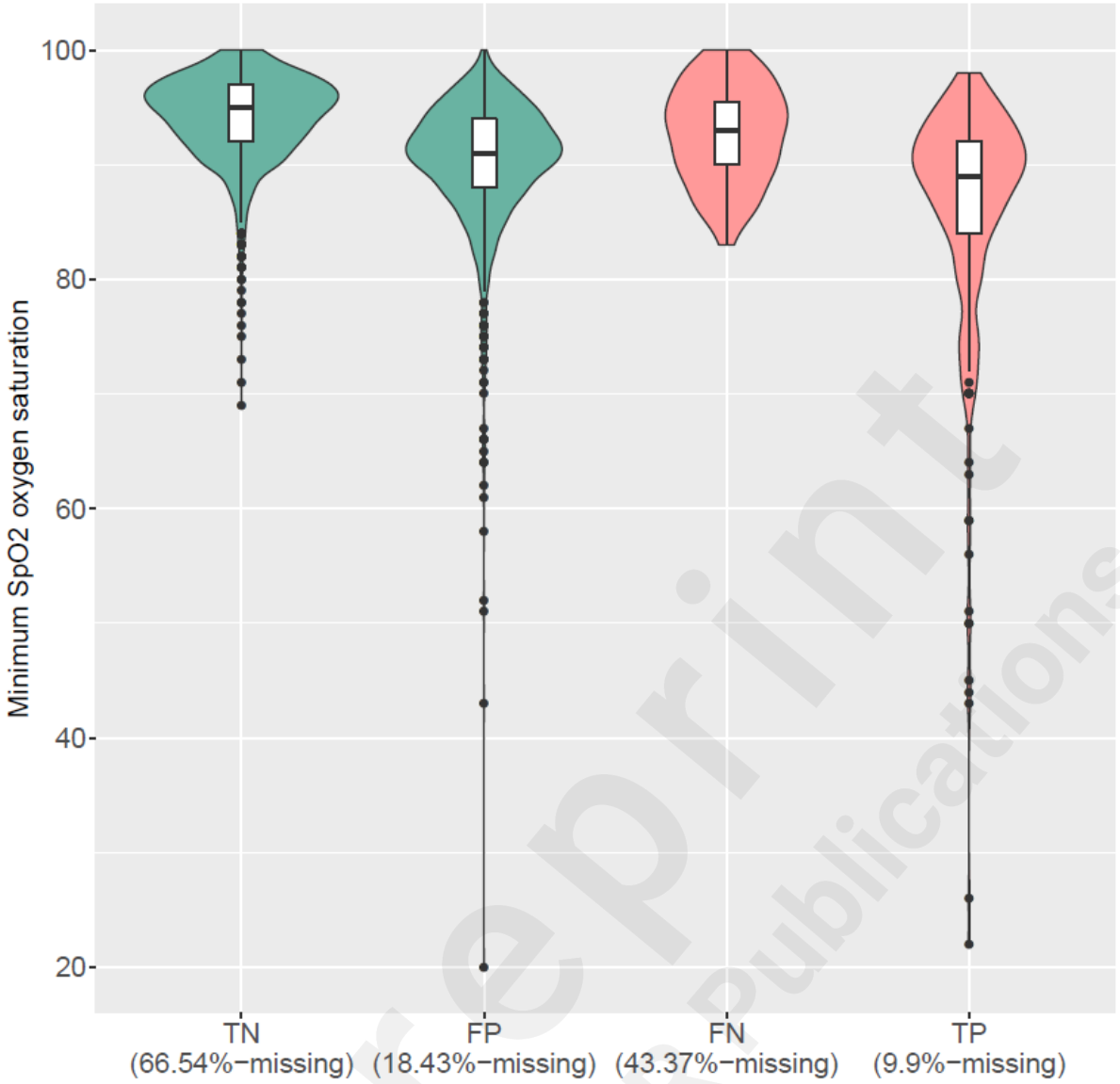


Figure 8: Violin plot of MinSpO2 (%) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot

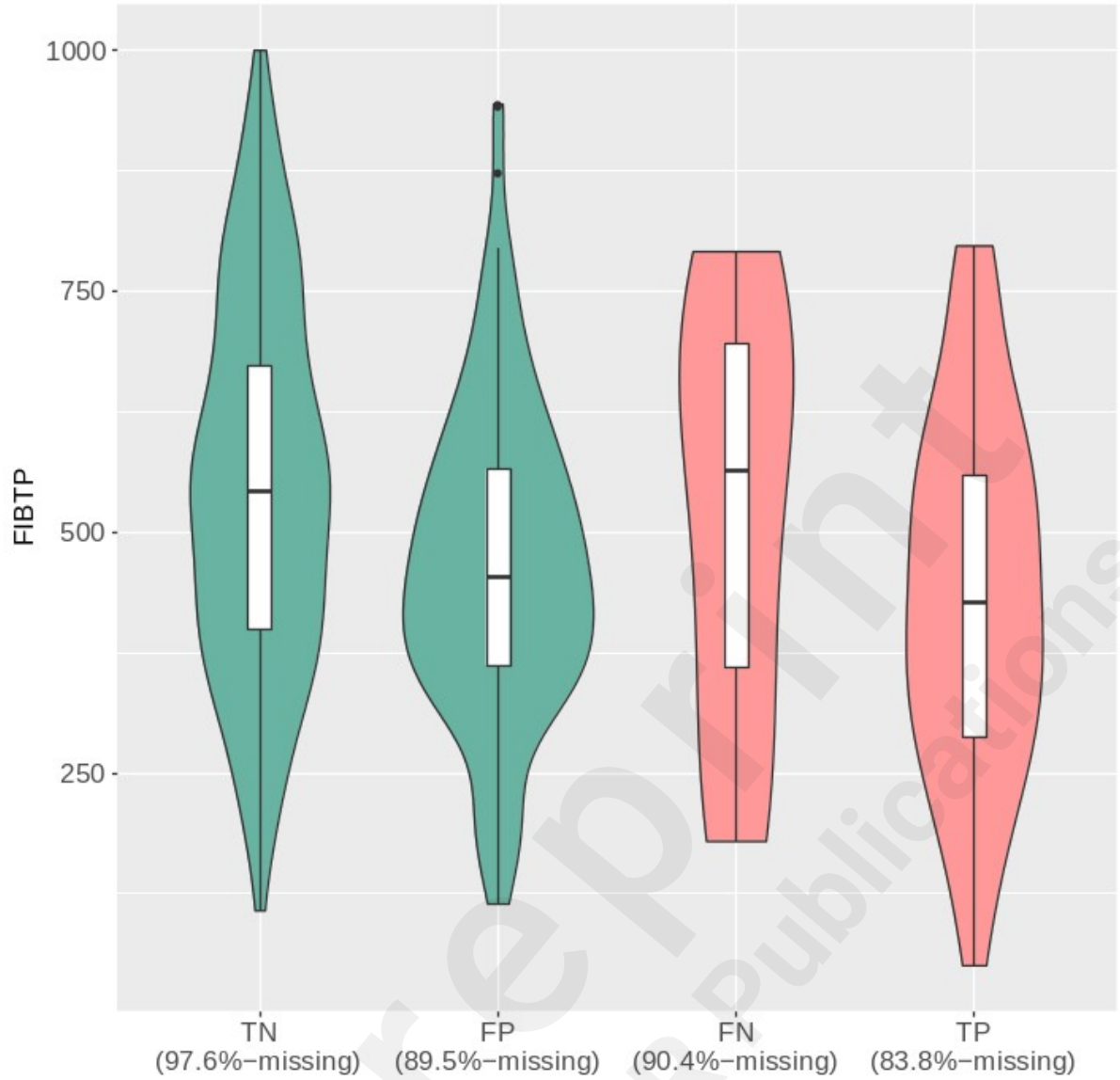


Figure 9: Violin plot of FIBTP (mg/dL) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

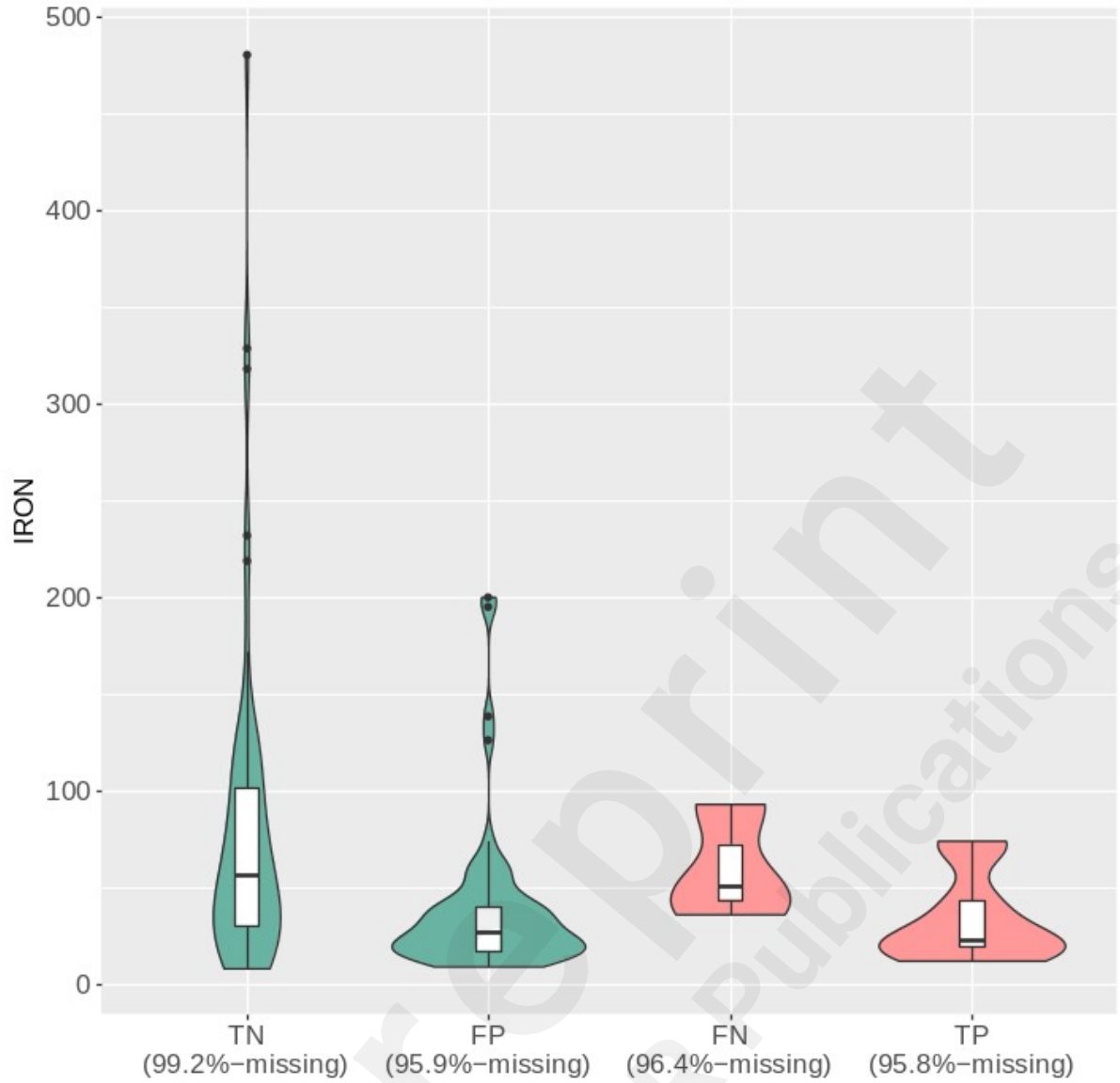


Figure 10: Violin plot of IRON (mg/dL) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

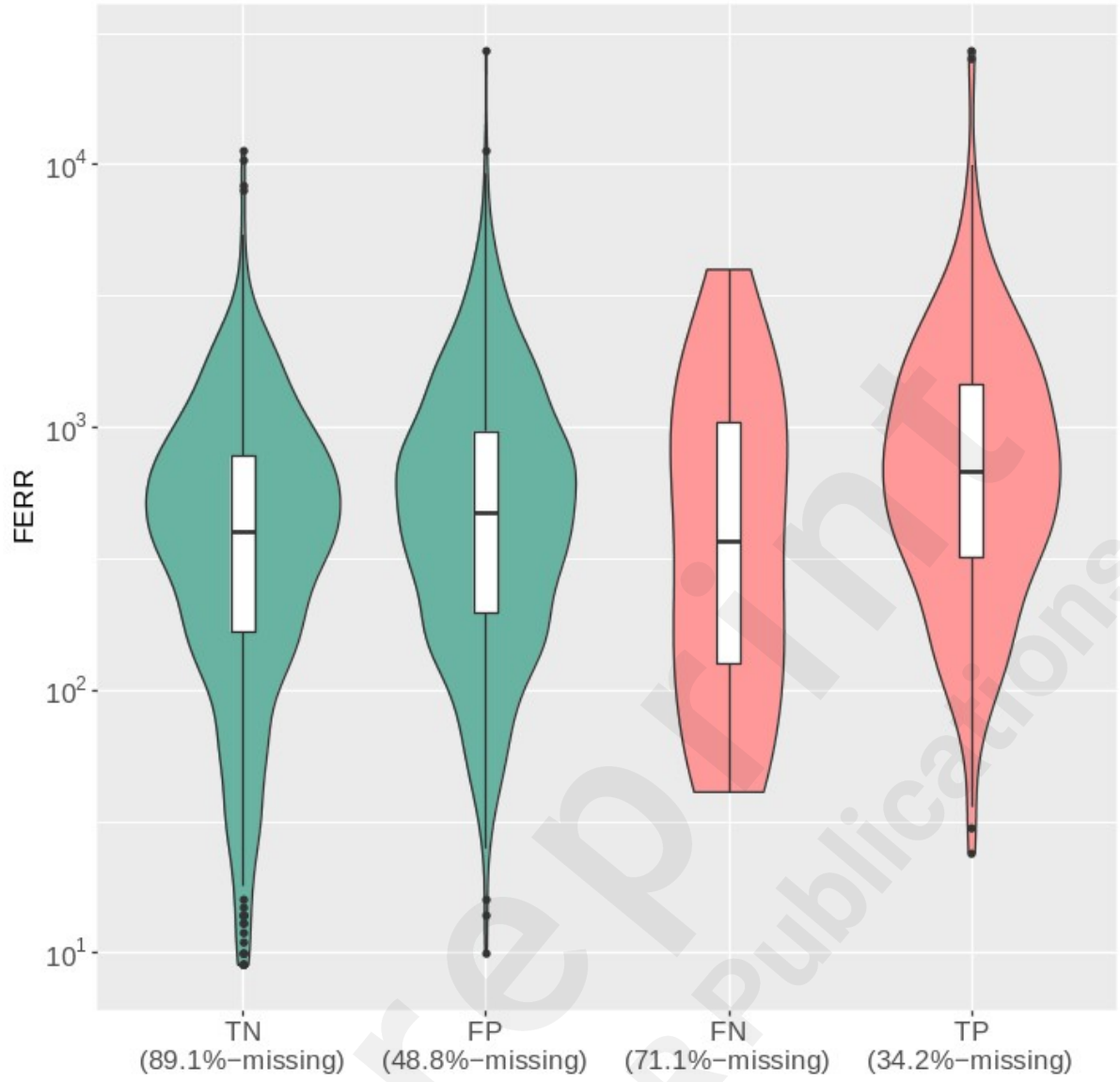


Figure 11: Violin plot of FERR (mg/L) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

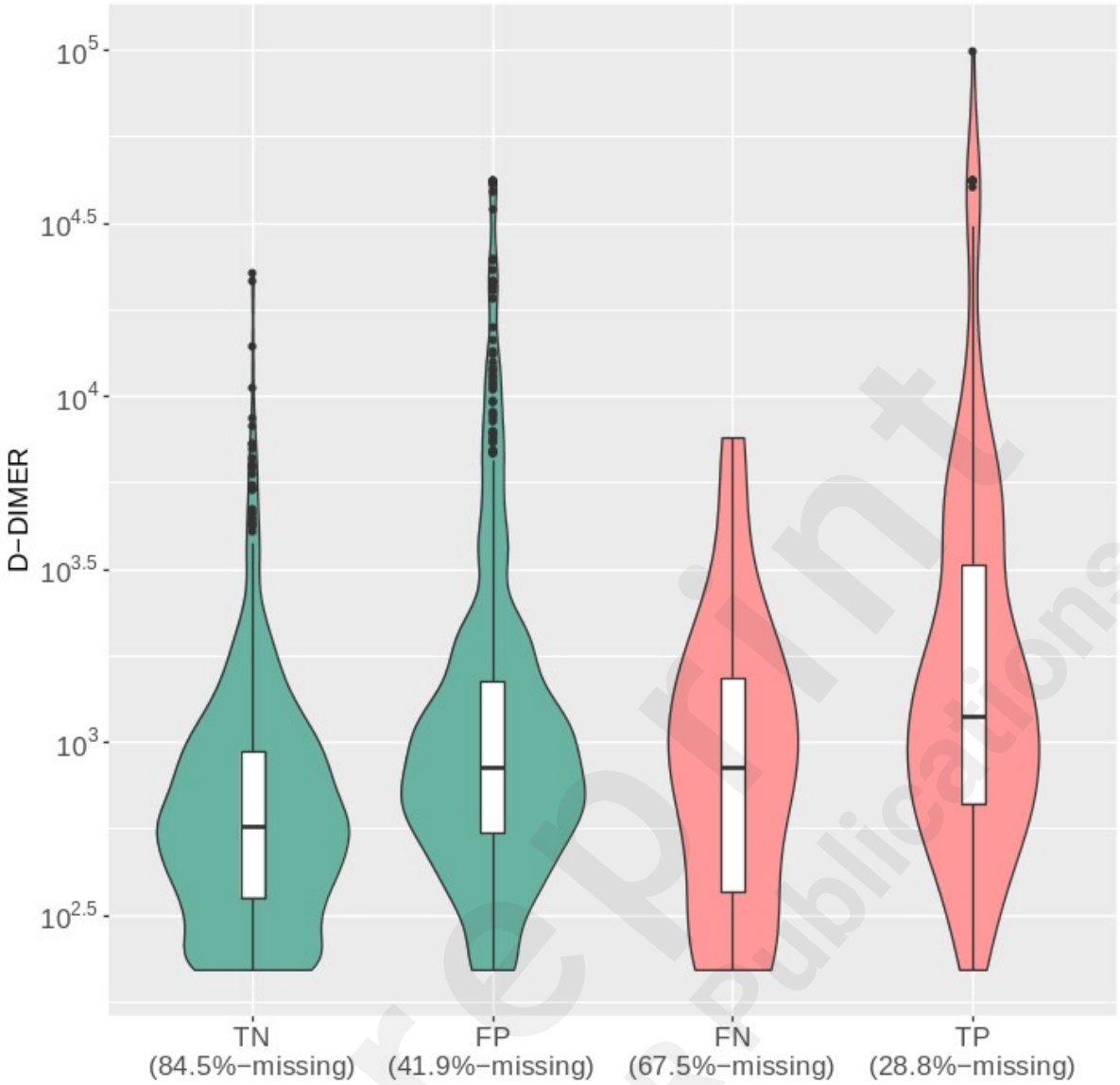


Figure 12: Violin plot of the D-DIMER (ng/mL) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

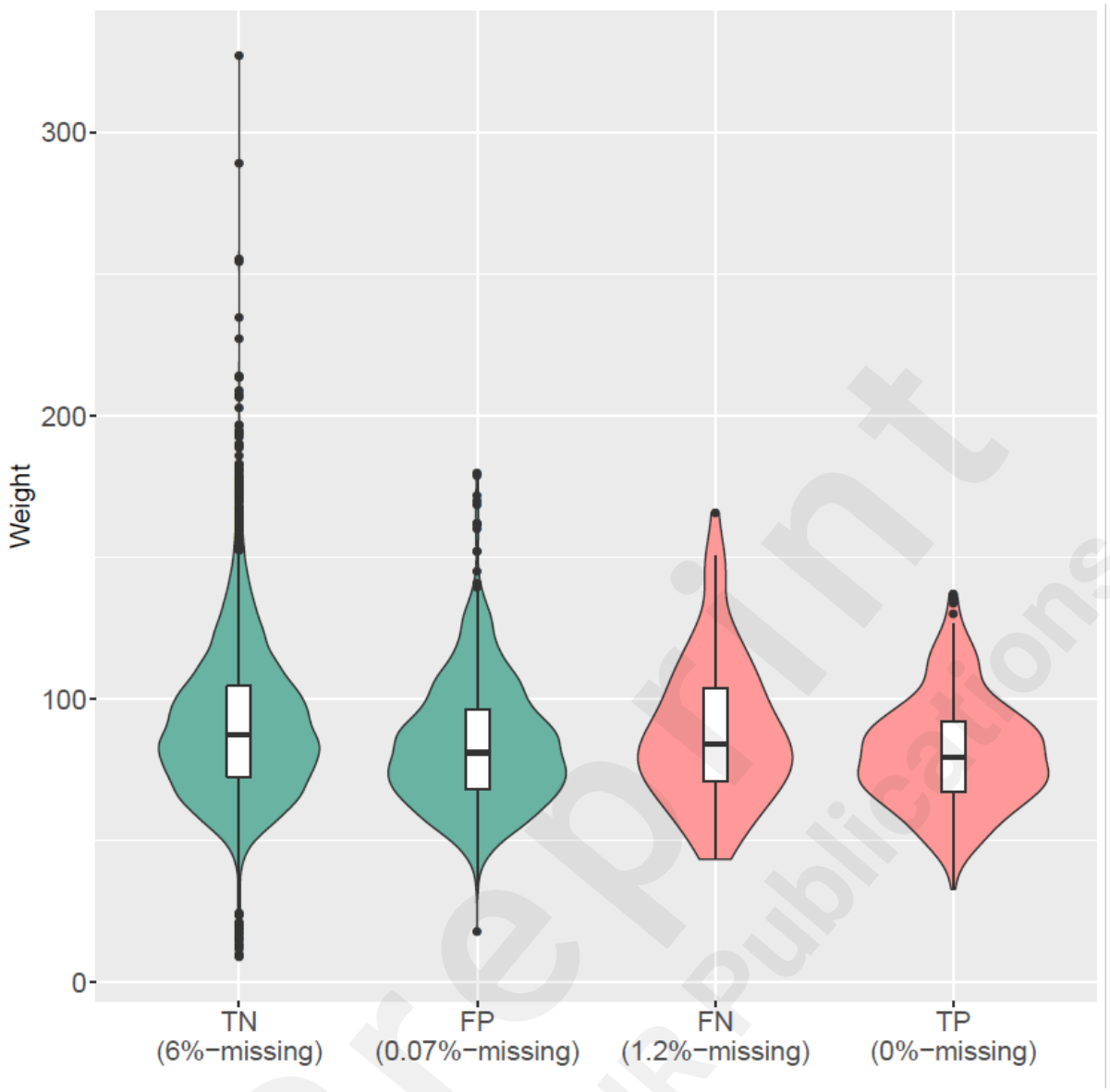


Figure 13: Violin plot of weight (kg) versus confusion matrix categorization: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Color indicates true state of the patient as survivor (green) or non-survivor (red). Percentage of missing values in each category is indicated at the base of the plot.

Table 6: Chronic Kidney disease Error Analysis

	TN	FP	FN	TP
No CKD	7073	663	68	146
CKD	568	715	15	187

Table 7: Serology Error Analysis

	TN	FP	FN	TP
Negative	305	91	0	34
Positive	155	33	4	8

Multimedia Appendix 2

Monte Carlo missing not at random simulation

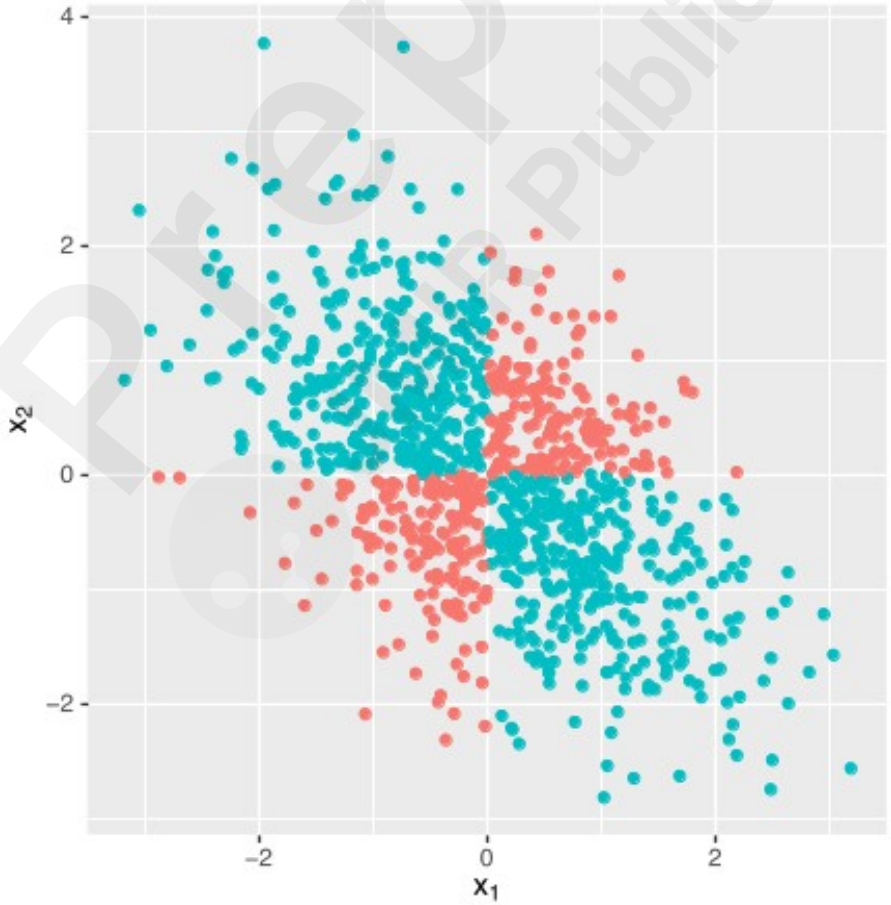


Figure 14: The results of 1000 Monte Carlo samples from the MNAR example. Red dots indicate aberrant labs and green indicate normal labs that will have one of their values censored by a

Bernoulli trial.

Here we construct a simple model to demonstrate the power of MNAR data in classification tasks, and the challenge they pose in imputation strategies. Consider the random variables X_1, X_2, Y, where we assume Y is the binary class label for severe $Y = 1$ or non-severe disease $Y = 0$, and X_1 and X_2 are two lab values that are zero-mean jointly Gaussian features with covariance matrix

$$\sum = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

Suppose normal labs have $X1 \cdot X2 < 0$ and patients with abnormal labs (i.e., $X1 \cdot X2 \geq 0$) are the only individuals going to have severe disease. Now suppose the physician is randomly (i.e., chosen by Bernoulli trial) provided a single lab value, but has perfect clinical judgement only orders the second missing clinical tests to confirm a patient is going to have severe disease; those without severe disease will thus only have a single lab value measured. These missing values will be MNAR. The class variable $Y = 1$ whenever the features are in the lower left or upper right quadrant of the feature space (i.e., $X1 \leq 0 \geq X2$ or $X2 \geq 0 \leq X1$), and $Y = 0$ otherwise; see (Figure 14) for an example Monte Carlo dataset. Clearly, in the fully observed dataset a simple decision tree can perfectly classify this data, but here whenever $Y = 0$ one of the features is randomly selected to be missing. The common missing completely at random (MCAR) strategy of “complete case analysis” wherein one simply removes all rows with missing data will clearly fail completely here since no $Y = 0$ data would make its way into the training set. Likewise, treating the data as MCAR and imputing based on the known mean of zero will result in all normal labs being in the abnormal ranges (and the more common strategy of using an empirical mean or median clearly only adds small noise to this imputation). More sophisticated missing at random (MAR) strategies such include using a KNN to find the nearest point in the reduced feature space to fill the missing values, or using the more sophisticated statistical imputation tools such as the multiple imputation by chained equations (MICE) method implemented in the mice library in R [24] to randomly fill in missing values while accounting for observed variables and to perform model averaging over the many randomly imputed datasets. In this classification context all such strategies will fail, and the only successful techniques will employ added features for missingness indicator variables to allow the classifier to explicitly model missing data patterns. We demonstrate these facts by employing the above techniques with a decision tree in our Monte Carlo samples, using an 80/20 train/test split.

Figure 15: One sample of MICE imputation on the MC samples from (Figure 14).

In the fully observed data, as expected the decision tree performs very well producing the confusion matrix (rows are true label, columns are predicted)

$$\begin{bmatrix} 144 & 1 \\ 1 & 54 \end{bmatrix}$$

In the complete case analysis, training is not even possible since all $Y = 0$ labels are thrown out from the analysis, and thus at best we are left with a dummy classifier always predicting class $Y = 1$, resulting in the confusion matrix

$$\begin{bmatrix} 0 & 145 \\ 0 & 55 \end{bmatrix}$$

If we consider the state-of-the-art MICE method, allowing it the benefit of using the class label in its multiple imputations modeling and filling in the entire test/train dataset (a data leakage problem giving undue advantage to the method), we still find the imputation fails and gives poor performance even with model averaging, resulting in the confusion matrix

$$\begin{bmatrix} 145 & 0 \\ 52 & 3 \end{bmatrix}$$

The failure of this method can be seen clearly when we plot one of the imputed datasets produced by this method in (Figure 15).

Finally, if we use a decision tree on only the missingness indicators, we achieve perfect performance (exceeding even the fully observed data)

$$\begin{bmatrix} 145 & 0 \\ 0 & 55 \end{bmatrix}$$

Clearly this example is overly simplified and taken to an extreme in terms of the coupling of missingness to class labels, but it illustrates the salient points surrounding MNAR in classification tasks when such coupling exists. In our real data experiments, we have verified that this coupling exists (albeit to a lesser extent than this toy example) with lab value ordering. GRU-D is a neural network architecture that has been designed around the notion of leveraging MNAR in multivariate time series classification tasks and the authors plan to invest further research effort in this direction.

References

1. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine 2020; 26(4):450–452. PMID: 32284615
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579(7798):270–273. PMID: 32015507
3. World Health Organization Coronavirus Disease (COVID-19) dashboard. Available from: https://covid19.who.int/ [accessed Aug 11, 2021]
4. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Pastore YPA, Mu K, Rossi L, Sun K, Viboud C, Xiong X, Yu H, Halloran ME, Longini IM, Jr., Vespignani A. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak.

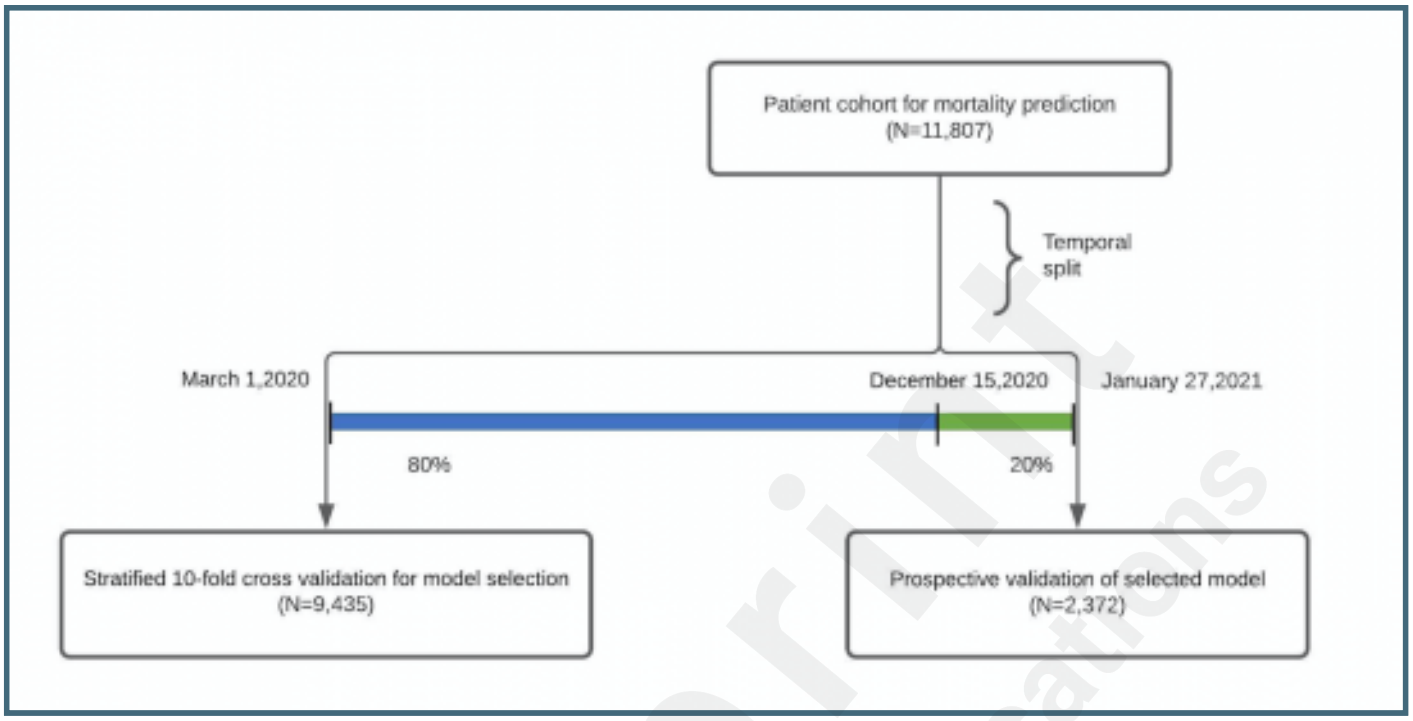
Science 2020; 368(6489):395—400. DOI: 10.1126/science.aba9757. PMID: 32144116
5. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? The Lancet 2020; 395(10231):1225–1228. PMID: 32178769
6. Rosenbaum L. Facing Covid-19 in Italy — ethics, logistics, and therapeutics on the epidemic's front line. New England Journal of Medicine 2020; 382(20):1873–1875. PMID: 32187459
7. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of Coronavirus Disease 2019 (COVID-19): A Review. JAMA 2020; 324(8):782–793. PMID: 32648899
8. Gao Y, Cai GY, Fang W, Li HY, Wang SY, Chen L, Yu Y, Liu D, Xu S, Cui PF, Zeng SQ, Feng XX, Yu RD, Wang Y, Yuan Y, Jiao XF, Chi JH, Liu JH, Li RY, Zheng X, Song CY, Jin N, Gong WJ, Liu XY, Huang L, Tian X, Li L, Xing H, Ma D, Li CR, Ye F, Gao QL. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nature Communications 2020;11(1):5033. PMID: 33024092
9. Gong J, Ou J, Qiu X, Jie Y, Chen Y, Yuan L, Cao J, Tan M, Xu W, Zheng F, Shi Y, Hu B. A tool for early prediction of severe Coronavirus Disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. Clinical Infectious Diseases 2020; 71(15):833–840. PMID: 32296824
10. Shi Y, Yu X, Zhao H, Wang H, Zhao R, Sheng J. Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. Critical Care 2020; 24(1):108. PMID: 32188484
11. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, Johnson KW, Lee SJ, Miotto R, Richter F, Zhao S, Beckmann ND, Naik N, Kia A, Timsina P, Lala A, Paranjpe M, Golden E, Danieletto M, Singh M, Meyer D, O'Reilly PF, Huckins L, Kovatch P, Finkelstein J, Freeman RM, Argulian E, Kasarskis A, Percha B, Aberg JA, Bagiella E, Horowitz CR, Murphy B, Nestler EJ, Schadt EE, Cho JH, Cordon-Cardo C, Fuster V, Charney DS, Reich DL, Bottinger EP, Levin MA, Narula J, Fayad ZA, Just AC, Charney AW, Nadkarni GN, Glicksberg BS. Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. J Med Internet Res 2020; 22(11):e24018. PMID: 33027032
12. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Dahly DL, Damen JAA, Debray TPA, de Jong VMT, De Vos M, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Heus P, Kammer M, Kreuzberger N, Lohmann A, Luijken K, Ma J, Martin GP, McLernon DJ, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Tzoulaki I, van Kuijk SMJ, van Bussel B, van der Horst ICC, van Royen FS, Verbakel JY, Wallisch C, Wilkinson J, Wolff R, Hooft L, Moons KGM, van Smeden M. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020; 369. PMID: 32265220
13. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987; 40(5):373–383. PMID: 3558716
14. Kumar A, Arora A, Sharma P, Anikhindi SA, Bansal N, Singla V, Khare S, Srivastava A. Is diabetes mellitus associated with mortality and severity of COVID-19? A meta-analysis. Diabetes Metab Syndr 2020; 14(4):535–545. PMID: 32408118
15. Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, Smola A. AutoGluon-Tabular: robust and accurate AutoML for structured data. 7th ICML Workshop on Automated Machine Learning; 2020 Jul 12-18, Virtual location. URL: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_7.pdf
16. Z Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports 2018; 8(1):6085. PMID: 29666385
17. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox Model. New York,

NY: Springer; 2000. ISBN 9781475732948
18. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. Statistics in Medicine 2020; 39(21):2815–2842. PMID: 32419182
19. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, Li Y, Zhang S, Jiang D, Chen X, Zhou F, Ma L, Bao P, Xie L. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. J Med Internet Res 2020;22(11):e23128. PMID: 33035175
20. Ko H, Chung H, Kang W, Park C, Kim D, Kim S, Chung C, Ko R, Lee H, Seo J, Choi T, Jaimes R, Kim K, Lee J. An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: development and validation of an ensemble model. J Med Internet Res 2020;22(12):e25442. PMID: 33301414
21. Abdulaal A, Patel A, Charani E, Denny S, Mughal N, Moore L. Prognostic modeling of COVID-19 using artificial intelligence in the united kingdom: model development and validation. J Med Internet Res 2020;22(8):e20259. PMID: 32735549
22. Kim H, Han D, Kim J, Kim D, Ha B, Seog W, Lee Y, Lim D, Hong S, Park M, Heo J. An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: retrospective cohort study. J Med Internet Res 2020;22(11):e24225. PMID: 33108316
23. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch J, Barnaby D, Zanos T, Northwell COVID-19 Research Consortium. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res 2021;23(2):e24246. PMID: 33476281
24. Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: multivariate imputation by chained equations in R. Journal of Statistical Software, Articles 2011; 45(3):1–67. DOI: 10.18637/jss.v045.i03

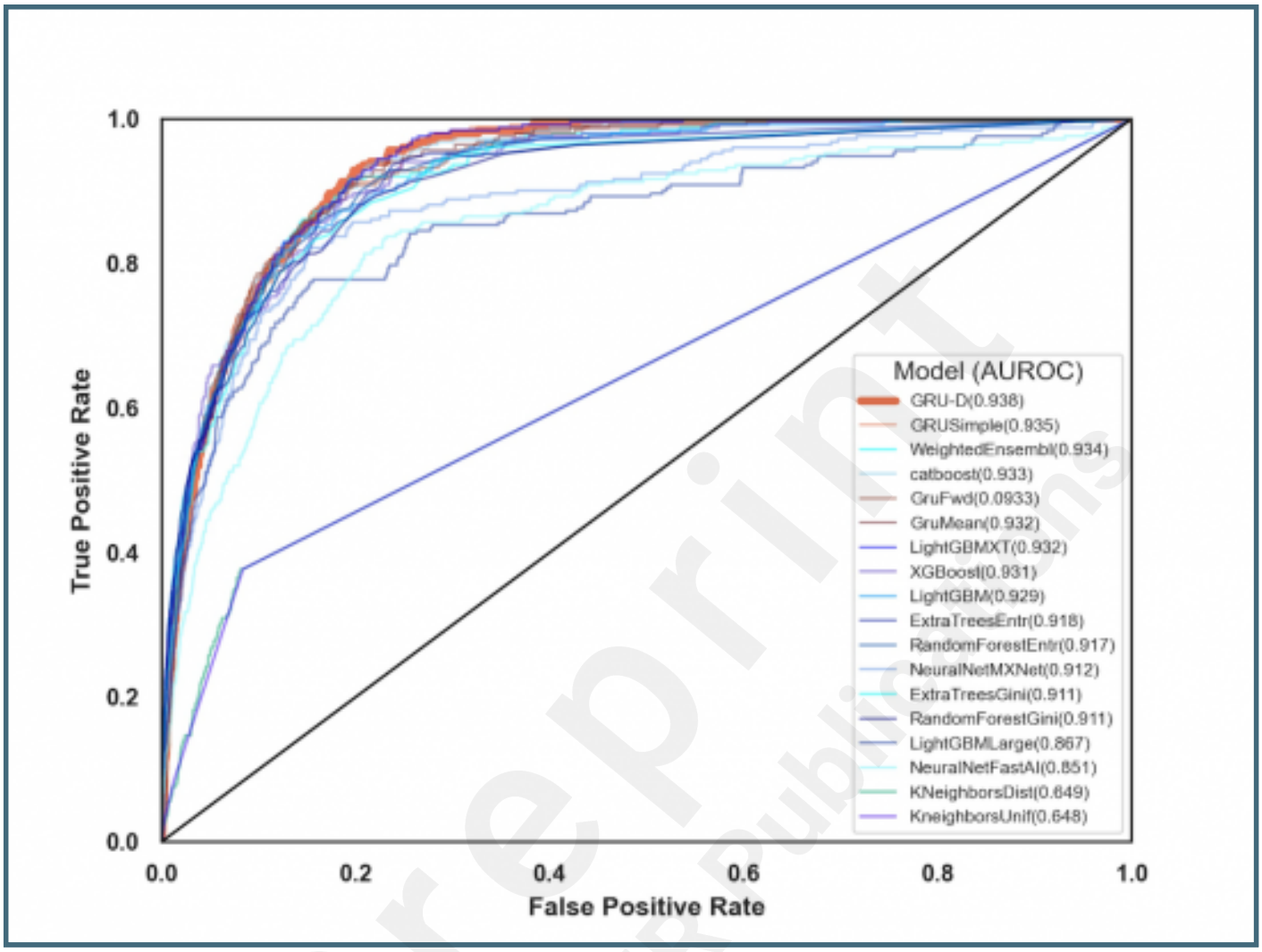
Supplementary Files

Figures

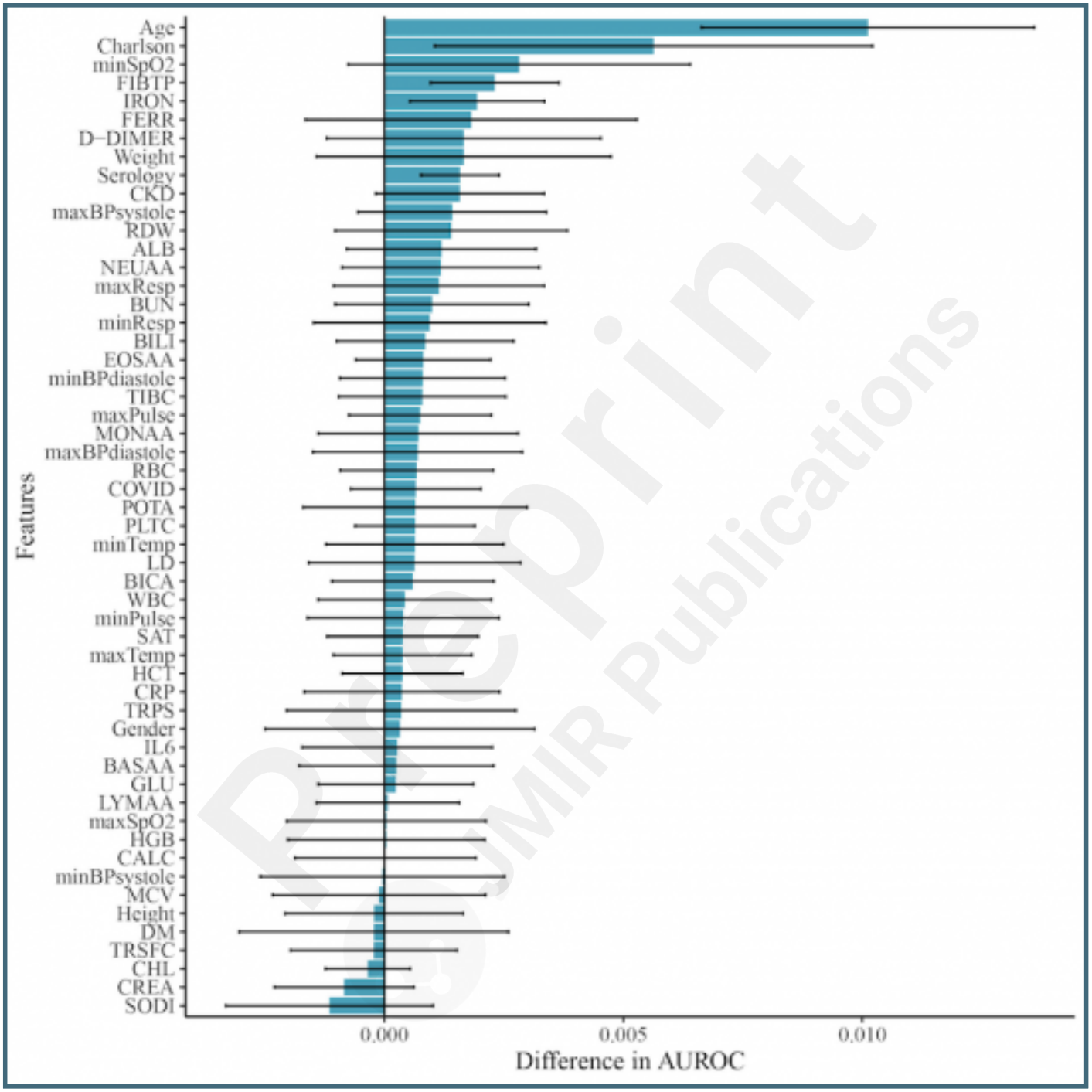
Consort diagram demonstrating the temporal split of our cohort for the purposes of model selection and prospective validation.



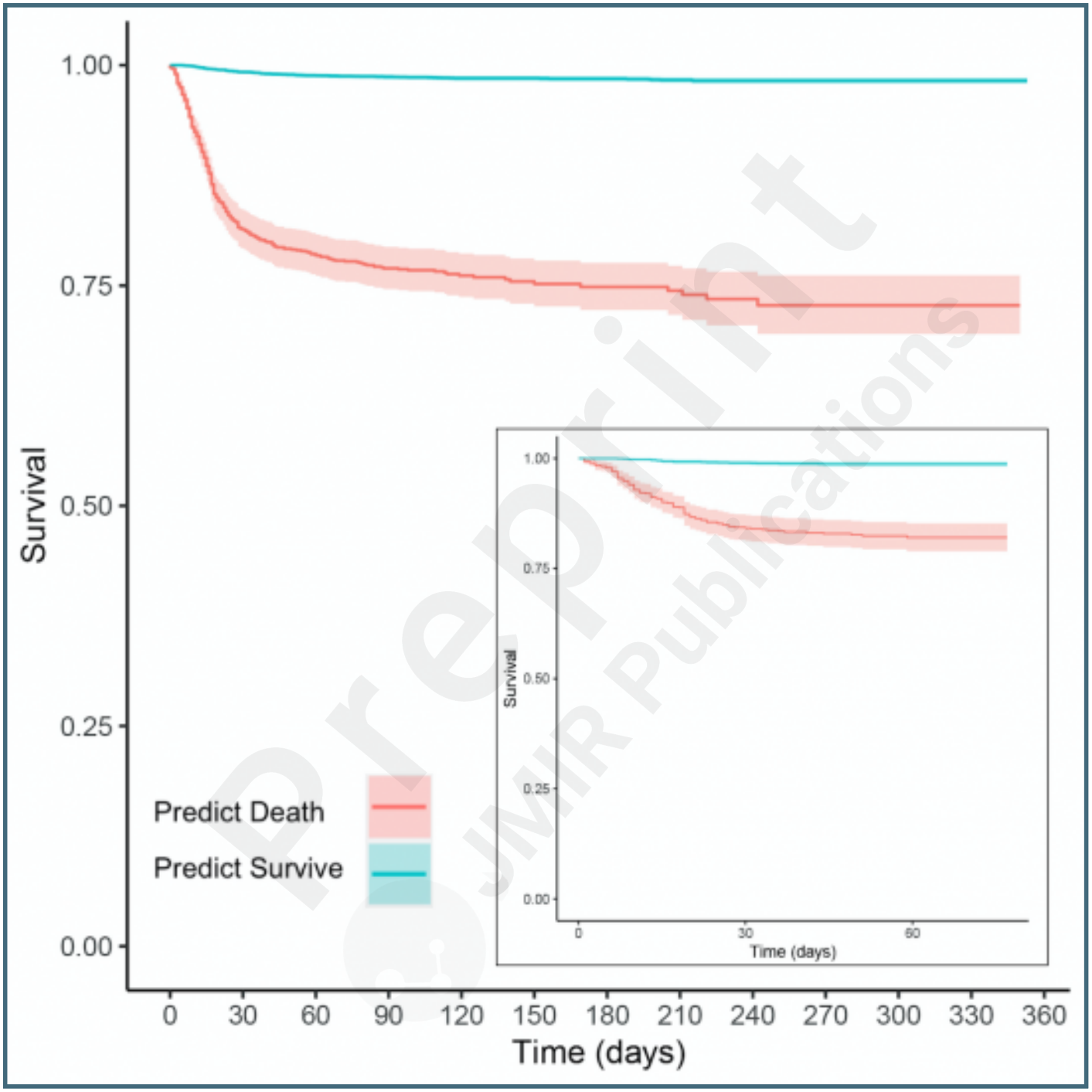
The receiver operating characteristic (ROC) curves for the eighteen models evaluated.



The feature importance in the GRU-D RNN model as defined by the average drop in AUROC (with 95% confidence intervals) when each feature is individually removed from the analysis. (Table 1) explains all the variable abbreviations, and the top five features are seen to be Age, Charlson comorbidity index, minSpO2, FIBTP and IRON.



Kaplan Meier survival curves for the GRU-D stratified populations in the cross-validation cohort (main figure) and the prospective test cohort (inset), where teal is a prediction of low risk of death and red is a prediction of high risk. Both figures have 95% confidence bands visualized for the teal and red curves, although the teal confidence bands are tight due to our large sample sizes.



Multimedia Appendixes

Error analysis.
URL: http://asset.jmir.pub/assets/4eea9d78bd02e0408a303dcaef3beb4e.docx

Monte Carlo missing not at random simulation.
URL: http://asset.jmir.pub/assets/dad138221d8a08f58cfd5d023f6ce6ab.docx