Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Clinical prediction system of complications among patients with COVID-19: A development and validation retrospective multicentre study during first wave of the pandemic

Ghadeer O. Ghosheh, Bana Alamad, Kai-Wen Yang, Faisil Syed, Nasir Hayat, Imran Iqbal, Fatima Al Kindi, Sara Al Junaibi, Maha Al Safi, Raghib Ali, Walid Zaher, Mariam Al Harbi, Farah E. Shamout

PII: S2666-5212(22)00018-7
DOI: https://doi.org/10.1016/j.ibmed.2022.100065
Reference: IBMED 100065

To appear in: Intelligence-Based Medicine

Received Date: 8 January 2022
Revised Date: 21 April 2022
Accepted Date: 1 June 2022

Please cite this article as: Ghosheh GO, Alamad B, Yang K-W, Syed F, Hayat N, Iqbal I, Al Kindi F, Al Junaibi S, Al Safi M, Ali R, Zaher W, Al Harbi M, Shamout FE, Clinical prediction system of complications among patients with COVID-19: A development and validation retrospective multicentre study during first wave of the pandemic, Intelligence-Based Medicine (2022), doi: https://doi.org/10.1016/j.ibmed.2022.100065.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier B.V.
Clinical prediction system of complications among patients with COVID-19: a development and validation retrospective multicentre study during first wave of the pandemic

Ghadeer O. Ghosheh¹, Bana Alamad¹, Kai-Wen Yang¹, Faisil Syed², Nasir Hayat¹, Imran Iqbal², Fatima Al Kindi², Sara Al Junaibi², Maha Al Safi², Raghib Ali¹, Walid Zaher³, Mariam Al Harbi²*, and Farah E. Shamout¹†

¹Engineering Division, NYU Abu Dhabi
²Abu Dhabi Health Services
³G42 Healthcare
*Joint supervision
†fs999@nyu.edu

Abstract
Clinical evidence suggests that some patients diagnosed with coronavirus disease 2019 (COVID-19) experience a variety of complications associated with significant morbidity, especially in severe cases during the initial spread of the pandemic. To support early interventions, we propose a machine learning system that predicts the risk of developing multiple complications. We processed data collected from 3,352 patient encounters admitted to 18 facilities between April 1 and April 30, 2020, in Abu Dhabi (AD), United Arab Emirates. Using data collected during the first 24 hours of admission, we trained machine learning models to predict the risk of developing any of three complications after 24 hours of admission. The complications include Secondary Bacterial Infection (SBI), Acute Kidney Injury (AKI), and Acute Respiratory Distress Syndrome (ARDS). The hospitals were grouped based on geographical proximity to assess the proposed system’s learning generalizability, AD Middle region and AD Western & Eastern regions, A and B, respectively. The overall system includes a data filtering criterion, hyperparameter tuning, and model selection. In test set A, consisting of 587 patient encounters (mean age: 45.5), the system achieved a good area under the receiver operating curve (AUROC) for the prediction of SBI (0.902 AUROC), AKI (0.906 AUROC), and ARDS (0.854 AUROC). Similarly, in test set B, consisting of 225 patient encounters (mean age: 42.7), the system performed well for the prediction of SBI (0.859 AUROC), AKI (0.891 AUROC), and ARDS (0.827 AUROC). The performance results and feature importance analysis highlight the system’s generalizability and interpretability. The findings illustrate how machine learning models can achieve a strong performance even when using a limited set of routine input variables. Since our proposed system is data-driven, we believe it can be easily repurposed for different outcomes considering the changes in COVID-19 variants over time.
1 Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has led to a global health emergency since the emergence of the coronavirus disease 2019 (COVID-19) [1]. Despite containment efforts, more than 491 million confirmed cases have been reported globally, including 892,170 cases in the United Arab Emirates (UAE) as of April 4, 2022 [1]. Due to unexpected burdens on healthcare systems, identifying high risk groups using prognostic models has become vital to support patient triage and resource allocation.

Most of the published prognostic models for patients with COVID-19 focus on predicting mortality or the need for intubation [2]. While the prediction of such adverse events is important for patient triage, clinical evidence suggests that patients with COVID-19 may also experience a variety of complications in organ systems that could lead to severe morbidity and mortality [3][4], especially amongst severe cases during the early waves of the pandemic. In this study, we identified three such complications associated with poor patient outcomes based on clinical evidence, prior to the emergence of the less severe variants [5]: Acute Respiratory Distress Syndrome (ARDS) [6], Acute Kidney Injury (AKI) [7], and Secondary Bacterial Infection (SBI) [8].

ARDS-related pneumonia has been reported as a major complication among patients with COVID-19 that have poor prognosis [9] and was a major cause of ventilator shortages worldwide [6][10][11]. In a Chinese study published in 2020, 31.0% of patients developed ARDS within a median of 12 days from the onset of COVID-19, and ARDS was the second most frequently observed complication after sepsis [10]. Additionally, only a few patients manifest clear clinical symptoms in the early stages of developing ARDS [6][12], so it is difficult to suspect ARDS unless it occurs. Hence, we identified early prediction of the risk of developing ARDS, prior to its onset, of high importance, since ARDS was considered as one of the main risk factors of death among hospitalized patients with COVID-19 [13].

Although COVID-19 primarily emerged as a respiratory disease, some patients with COVID-19 experience both respiratory and extra-respiratory complications including renal complications such as AKI [13][7]. Patients with AKI require special care and resources such as renal replacement therapy and dialysis [15]. It was estimated that AKI developed in 36.6% of patients admitted with COVID-19 in metropolitan New York in 2020, of which 35% had died [15]. Therefore, risk prediction of AKI can help in initiating preventive interventions in order to avoid quite poor patient prognosis.

Moreover, several studies reported alarming percentages of hospitalized patients with COVID-19 who develop SBI [10]. SBI is known for poor outcomes in several respiratory viral infections. Hence, it led to increased burdens on hospitals in the 1918 influenza pandemic, 2009 H1N1 influenza pandemic, and in seasonal flu [16][17][18]. Patients with COVID-19 who developed SBI have shown worse outcomes, including admission to the Intensive Care Unit (ICU) and mortality, compared to those who did not develop SBI [19]. Therefore, early prediction of SBI can potentially improve patient prognosis, such as by taking aseptic procedures especially when hospitals get crowded [8].

In recent years, machine learning gained popularity for the development of algorithms for clinical decision support tools [20][21][22]. In the context of COVID-19, most machine learning studies have focused either on diagnosis or prognosis based on adverse events, mostly mortality and intubation [23][24][2]. We summarize a few examples in Table 1. Since ARDS is considered a major manifestation of the COVID-19 disease, some studies focused on developing machine learning models to predict ARDS as an outcome [6][12], such as by using a large set of hematological and biochemical markers [12]. One limitation of such approaches is that they rely on laboratory-test results that may not be routinely measured. In another study, the authors used both statistical machine learning models and deep neural networks for the prediction of ARDS, by combining a large feature set of chest Computed Tomography (CT) findings, demographics, epidemiology, clinical symptoms, and laboratory-test results [6]. Similarly, for AKI prediction amongst patients with COVID-19, a multivariate logistic regression was developed using findings of CT imaging, laboratory-test results, vital-sign measurements, and patient demographics [23]. While recent work on SBI mainly focused on its clinical manifestations and occurrence [26][16][27], one study investigated sepsis risk prediction among patients with COVID-19 using hematological parameters and other biomarkers [25]. To summarize, existing work tends to predict a single complication at a time, which is less informative than predicting multiple complications known to be common among patients with COVID-19, use costly input features that may not be readily available, or rely on training deep neural networks that require high computational resources and

| Table 1: Examples of Machine Learning Approaches for COVID-19 Complication Prediction |
|---------------------------------|---------------------------------|---------------------------------|
| ARDS prediction                  | AKI prediction                  | SBI prediction                  |
| [6][12]                         | [23]                            | [25]                            |
| Statistical machine learning    | Deep neural networks            | Multivariate logistic regression |
| models                          |                                 |                                 |
Table 1: Examples of machine learning studies that aim to predict various outcomes for in-patients with confirmed COVID-19 diagnosis. We refer the readers to extensive published literature reviews [2, 23, 24].

| Reference | Outcome Description | Input Data Description | Models Description | Study Location |
|-----------|---------------------|------------------------|-------------------|----------------|
| [29]      | Deterioration (intubation or ICU admission or mortality) | Chest X-ray images and clinical data (patient demographics, seven vital-sign variables, and 24 laboratory-test results) | Convolutional neural network for chest X-ray images and gradient boosting model for clinical data | United States |
| [30]      | Mortality | Five laboratory-test results | Support vector machine | United States |
| [31]      | Severe progression (high oxygen flow rate, mechanical ventilation or mortality) | Chest CT scans, patient demographics, five vital-sign variables, symptoms, comorbidities, 14 laboratory-test results, and chest CT radiology report findings | Deep neural network and logistic regression | France |
| [32]      | Prognostication (intubation or hospital admission, or mortality) | Chest X-ray images, two vital-sign variables, and nine laboratory-test results | Convolutional neural network | United States |
| [28]      | Sepsis | Eight laboratory-test results | Gradient boosting model | China |
| [25]      | AKI | Findings of abdominal CT scans, demographics, vital signs, comorbidities, and three laboratory-test results | Logistic regression | United States |
| [33]      | ARDS | Demographics, interventions, comorbidities, 17 laboratory-test results, and eight vital signs | Gradient boosting model | United States |

Therefore, there is a pressing need for a low-cost predictive system that uses routine clinical data to predict complications and support patient management. In this work, we address this need by developing and evaluating a machine learning system that predicts the risk of ARDS, SBI, and AKI among patients with COVID-19 admitted to the Abu Dhabi Health Services (SEHA) facilities, UAE, from April 1st, 2020 to April 30th, 2020, during the first wave of the pandemic. While we focus on three complications only, namely because their occurrence could be identified retrospectively using clinical criteria, the system and proposed training framework can be scaled to incorporate predictions of other complications, and can be fine-tuned using datasets of other patient cohorts. An overview of the pipeline is shown in Figure 1. Next, we describe our methodology in Section 2 and the performance and explainability results in Section 3. We then discuss the limitations and strengths of the study in Section 4 and conclude by highlighting the potential of our system in clinical settings in Section 5. To allow for reproducibility and external validation, we made our code and one of the evaluation test sets publicly available at: [github.com/nyuad-cai/COVID19Complications](https://github.com/nyuad-cai/COVID19Complications)

2 Methods

This study is reported following the TRIPOD guidance [34].

2.1 Data source

This study is a retrospective multicentre study that includes anonymized data recorded within 3,493 COVID-19 hospital encounters at 18 Abu Dhabi Health Services (SEHA) healthcare facilities in Abu Dhabi, United Arab Emirates. The study received approval by the Institutional Review Board (IRB) from the Department of Health (Ref: DOH/CVDC/2020/1125) and New York University Abu Dhabi (Ref: HRPP-2020-70). Informed consent was not required for this study as it was determined as exempt. All methods were performed in accordance with the relevant guidelines and regulations. There were nine facilities in the Middle region,
Figure 1: Overview of our proposed model development approach and expected application in practice. As shown in the first row, we develop our complication-specific models by first preprocessing the data, identifying the occurrences of the complications based on the criteria shown in Table 2, training and selecting the best-performing models on the validation set, and then evaluating the performance on the test set, retrospectively. As for the application (second row), we expect our system to predict the risk of developing any of the three complications for any patient after 24 hours of admission.

which includes the capital city, and nine facilities in the Eastern and Western regions. Those regions are highlighted in Figure 2(a). Figure 2(b) shows the flowchart of how the exclusion criteria was applied to obtain the final data splits. We excluded 127 non-adult encounters and 14 pregnant encounters and split the dataset into training and test sets. The training sets were used for model training and selection, while the test sets were used for evaluation. Training set A consisted of 1,829 encounters recorded in the Middle region between April 1, 2020 and April 25, 2020. To evaluate for temporal generalizability, test set A included 587 encounters recorded in the Middle region between April 26, 2020 and April 30, 2020. Training set B included 711 encounters admitted to the Eastern and Western regions between April 1, 2020 and April 25, 2020 and test set B included 225 encounters admitted to the same hospitals between April 26, 2020, and April 30, 2020.

2.2 Outcomes

Based on clinical evidence and in collaboration with clinical experts, we focused on predicting three clinically diagnosed events, SBI, AKI [35] and ARDS [36] that are associated with poor patient prognosis. For each patient encounter in the training and test sets, we identified the first occurrence (i.e., date and time), if any, of each complication based on the criteria shown in Table 2. SBI is defined based on positive cultures within 24 hours of sample collection, AKI is defined based on the KDIGO classification criteria [35], and ARDS is defined based on the Berlin definition [36], which required the processing of free-text chest radiology reports. Further details on the processing of those reports is described in Supplementary Section A.

2.3 Input features

We considered data recorded within the first 24 hours of admission as input features for the predictive models. This data included continuous and categorical features related to the patient baseline information, demographics, and vital signs. Within the patient’s baseline and demographic information, age and Body Mass Index (BMI) were treated as continuous features, whereas pre-existing medical conditions (i.e., hypertension, diabetes, chronic kidney disease, and cancer), symptoms recorded at admission (i.e., cough, fever, shortness of breath, sore throat, and rash) and patient sex were treated as binary features. As for the vital signs, we included seven continuous features, including systolic blood pressure, diastolic blood pressure, respiratory rate,
Abu Dhabi Middle region
Abu Dhabi Western region
Abu Dhabi Eastern region

Key

(a) (b)

Figure 2: (a) The UAE map showcasing the location of the healthcare facilities included in this study. (b) Flowchart for the overall dataset showing how the inclusion and exclusion criteria were applied to obtain the final training and test sets, where \( n \) represents the number of patient encounters, and \( p \) represents the number of unique patients.

Table 2: Criteria used to define the occurrence of complications.

| Complication | Definition | Reference |
|--------------|------------|----------|
| SBI          | Positive blood, urine, throat or sputum cultures within 24 hours of sample collection | * |
| AKI          | Based on the Kidney Disease Improving Global Guidelines (KDIGO) classification, increase in Serum Creatinine by \( \geq 0.3 \text{mg/dl} \) within 48 hours OR Increase in Serum Creatinine by \( \geq 1.5 \) times OR Urine volume < 0.5ml/kg/hr for 6 hours† | [35] |
| ARDS         | Based on the Berlin definition, presence of bilateral opacity in radiology reports AND Oxygenation: \( \text{PaO}_2/\text{FiO}_2 \leq 300 \text{ mm Hg} \) AND Timing: \( \leq \) one week AND Origin: pulmonary | [36] |

* Based on SEHA’s clinical standards.
† Urine output was not measured in our dataset because it is collected in the intensive care unit.

peripheral pulse rate, oxygen saturation, auxiliary temperature, and the Glasgow Coma Score. We selected those features as they are commonly used in early warning score systems [37]. All vital signs measurements were processed into minimum, maximum, and mean statistics. We summarized patient demographics, prevalence of the complications, and the distributions of the input features across the training and test sets.

2.4 Predictive modeling

The proposed system predicts the risk of developing each of the three complications during the patient’s stay after 24 hours of admission. This is represented by a vector \( \mathbf{y} \) consisting of three predictions, where each prediction is computed by a complication-specific model, such that

\[
\mathbf{y} = [y^{\text{SBI}}, y^{\text{AKI}}, y^{\text{ARDS}}],
\]

Journal Pre-proof
where $y_{\text{complication}} \in [0, 1]$.

The overall workflow of the model development is depicted in Figure 1. For each complication-specific model, we excluded from its training and test sets patients who developed that complication prior to the time of prediction. For AKI, we also excluded patients with chronic kidney disease. Then for each complication, our system trains four model ensembles based on four types of base learners: logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM) and a light gradient boosting model (LGBM). Missing data was imputed using median imputation for all models except for LGBM, which can natively learn from missing data, and the data was further scaled using min-max scaling for LR and standard scaling for SVM and KNN.

For each type of base learner, the system performs a stratified k-folds cross-validation using the complication’s respective training set with $k = 3$. We performed random hyperparameter search for each base learner with 30 iterations, resulting in three trained models for each hyperparameter set selected per iteration. The choice of random search was motivated by its relative simplicity, and high efficiency and performance compared to other hyperparameter tuning methods. The hyperparameter search ranges are summarized in Supplementary Section B. The ranges were defined based on initial experiments with manually chosen hyperparameters.

We then selected the top two hyperparameter sets whose models achieved the highest average area under the receiving operator characteristic curve (AUROC) on the validation sets, resulting in six trained models. We created an ensemble of those six models, and each model within the ensemble was further calibrated using isotonic regression on its respective validation set to ensure non-harmful decision making, except for the LR models. Isotonic regression takes a trained model’s raw predictions as inputs, and computes well-calibrated output probabilities. This is done by grouping the raw predictions into bins associated with estimates of empirical probabilities. The final prediction of each complication consisted of an average of the calibrated predictions of all models within an ensemble. All analysis was performed using Python (version 3.7.3). The LR, KNN, and SVM models were implemented using the Python scikit-learn package and the LGBM models were implemented using the LightGBM package.

2.5 Model interpretability

We performed post-hoc feature importance analysis using the SHapley Additive exPlanations (SHAP). SHAP values are indicative of the relative importance of the input variables and their impact on the predictions. The analysis was conducted using the open-source SHAP package, where we obtained the mean absolute SHAP values of the features for the six models per ensemble. For each feature, the six SHAP values were averaged and then ranked to reveal the overall importance of the features with respect to the ensembled prediction. We present the four top ranked features per complication ensemble for each test set using bar plots.

2.6 Performance assessment

We evaluated each complication ensemble using the AUROC and the area under the precision-recall curve (AUPRC) on the test set. The AUROC is a measure of the model’s ability to discriminate between positive (complications) and negative cases (no complication), while the AUPRC is a measure of model robustness when dealing with imbalanced datasets, i.e. unequal distribution of positive and negative cases. The closer the AUROC and AUPRC are to 1, the better the performance of the model. Confidence intervals for all of the evaluation metrics were computed using bootstrapping with 1,000 iterations. We also assessed the calibration of the ensemble, after post-hoc calibration of its trained models, using reliability plots and reported calibration intercepts and slopes.

3 Results

A total of 3,352 encounters were included in the study and the statistics of the characteristics of the final data splits are presented in Table 3. Across all the data splits, the mean age ranges between 39.3 and 45.5
Table 3: Summary of the baseline characteristics of the patient cohort in the training sets and test sets and the prevalence of the predicted complications. Note that n represents the total number of patients while % is the proportion of patients within the respective dataset.

| Patient Cohort | Encounters, n | Age, mean (IQR) | Male, n (%) | Arab, n (%) | Non-Arab, n (%) | Mortality, n (%) |
|----------------|---------------|-----------------|-------------|-------------|----------------|-----------------|
| Training set A | 1829          | 41.7 (17.0)     | 1582 (86.5) | 1534 (83.9) | 36 (2.0)       | 36 (2.0)        |
| Test set A     | 587           | 45.5 (18.0)     | 522 (88.9)  | 498 (84.8)  | 22 (3.7)       | 17 (5.0)        |
| Training set B | 711           | 39.3 (17.0)     | 622 (87.5)  | 591 (83.1)  | 9 (1.3)        | 9 (1.3)         |
| Test set B     | 225           | 42.7 (20.0)     | 191 (84.8)  | 182 (80.9)  | 3 (1.3)        | 3 (1.3)         |

| Complications | SBI, n (%) | Developed within 24 hours from admission, n (%) | Developed after 24 hours from admission, n (%) |
|----------------|------------|-----------------------------------------------|-----------------------------------------------|
| Training set A | 92 (5.0)   | 91 (5.0)                                      | 92 (5.0)                                      |
| Test set A     | 45 (7.7)   | 39 (7.2)                                      | 42 (7.2)                                      |
| Training set B | 23 (3.2)   | 32 (4.5)                                      | 22 (3.1)                                      |
| Test set B     | 17 (7.6)   | 16 (7.1)                                      | 16 (7.1)                                      |

| Complications | AKI, n (%) | Developed within 24 hours from admission, n (%) | Developed after 24 hours from admission, n (%) |
|----------------|------------|-----------------------------------------------|-----------------------------------------------|
| Training set A | 126 (6.9)  | 126 (6.4)                                     | 126 (6.4)                                     |
| Test set A     | 52 (8.9)   | 57 (9.7)                                      | 52 (8.9)                                      |
| Training set B | 43 (7.3)   | 45 (6.3)                                      | 43 (7.3)                                      |
| Test set B     | 32 (4.5)   | 24 (10.7)                                     | 32 (4.5)                                      |

| Complications | ARDS, n (%) | Developed within 24 hours from admission, n (%) | Developed after 24 hours from admission, n (%) |
|----------------|------------|-----------------------------------------------|-----------------------------------------------|
| Training set A | 117 (6.4)  | 117 (6.4)                                     | 117 (6.4)                                     |
| Test set A     | 57 (9.7)   | 57 (9.7)                                      | 57 (9.7)                                      |
| Training set B | 18 (2.5)   | 24 (10.7)                                     | 18 (2.5)                                      |
| Test set B     | 13 (5.8)   | 13 (5.8)                                      | 13 (5.8)                                      |

Table 4: Characteristics of the variables that were used as input features to our models. The mean and interquartile ranges are shown for the demographic features, and vital-sign measurements. For the comorbidities and symptoms admission, n denotes the number of patients and % denotes the percentage of patients per the respective dataset.

| Variable, unit | Demographics, mean (IQR) | Comorbidities, n (%) | Symptoms at admission, n (%) | Vital-sign measurements, mean (IQR) |
|----------------|---------------------------|----------------------|-----------------------------|------------------------------------|
| Age            | 41.7 (17.0)               | 550 (30.1)           | 851 (46.5)                  | 126.3 (15.0)                       |
| BMI            | 26.9 (5.2)                | 427 (23.3)           | 28 (1.5)                    | 77.5 (9.8)                         |
| Male, n (%)    | 1582 (86.5)               | 68 (3.7)             | 30 (1.6)                    | 18.9 (1.0)                         |
| Hypertension   | 213 (36.3)                | 30.1                  | 12 (1.5)                    | Systolic blood pressure, mmHg      |
| Diabetes       | 221 (37.6)                | 3.7                   | 7 (1.2)                     | 126.8 (16.0)                       |
| Chronic kidney disease | 121 (17.0)            | 3.7                   | 7 (1.2)                     | Diastolic blood pressure, mmHg     |
| Cancer         | 168 (23.6)                | 3.7                   | 7 (1.2)                     | 20 (2.4)                           |
| Cough          | 338 (57.6)                | 3.7                   | 7 (1.2)                     | Respiratory rate, breaths per minute |
| Fever          | 20 (3.4)                  | 3.7                   | 7 (1.2)                     | 20.2 (2.5)                         |
| Shortness of breath | 99 (16.9)           | 3.7                   | 7 (1.2)                     | Peripheral pulse rate, beats per minute |
| Sore throat    | 89 (15.2)                 | 3.7                   | 7 (1.2)                     | 85.4 (11.6)                        |
| Rash           | 10 (1.7)                  | 3.7                   | 7 (1.2)                     | Oxygen saturation, %               |

years and the proportion of males ranges between 84.8% to 88.9%. The mortality rate was also less than 4% across all data splits, ranging between 1.3% and 3.7%. ARDS was the most prevalent complication developed in the first 24 hours of admission across all datasets. The incidence of the complications developed after 24
hours were higher in the test sets than in their respective training sets. The distributions of the vital signs and demographics in terms of the mean and interquartile ranges, are shown in Table 4.

The performance results of the models selected by our system across the two test sets in terms of the AUROC and AUPRC are shown in Table 5. The Receiver Operating Characteristic curve (ROC), Precision Recall Curve (PRC), and reliability plots are also visualized in Figures 3(a), 3(b), and 3(c), respectively. Across both test sets, our data-driven approach achieved good performance (>0.82 AUROC) for all of the complications. In test set A, AKI was the best discriminated endpoint at 24 hours from admission, with 0.906 AUROC. This is followed by SBI (0.902 AUROC), and SBI (0.854 AUROC). In test set B, AKI was the best discriminated endpoint with 0.891 AUROC, followed by SBI (0.859 AUROC), and ARDS (0.827 AUROC).

The prevalence of the predicted complications ranged between 3.2%-6.9% and 7.1%-10.7% in the training and test sets, respectively. This high class imbalance is reflected in the AUPRC results, since the AUPRC depends on the prevalence of the outcome and tends to have a low value when there is class imbalance [47].

We also observe that LR was selected as the best performing model on the validation sets for most complications, highlighting its predictive power despite its simplicity compared to the other machine learning models. LGBM was selected for ARDS in test set B, as shown in Supplementary Section C.

The top four important features for each complication are shown in Figure 4 across the two test sets. Age was among the top predictive features for all of the complications in both test sets. Similarly, systolic blood pressure was one of the top features for predicting SBI and AKI across both sets. Other features such as peripheral pulse rate and respiratory rate were among the top predictive features across both sets, for AKI and ARDS respectively.

The calibration results show that our ensemble models were adequately calibrated across all complications as the calibration slopes were approximately equal to 1, as shown in Table 5 and Figure 3(c). This is also reflected in the sample patient timelines visualized in Figure 5, where the predicted risks for the patient who experienced the complications were relatively higher than those predicted for the patient who did not experience any complications. In Figure 5(a), the patient shown developed all three complications during their hospital stay of 44 days. This highlights the importance of predicting all complications simultaneously, especially for patients who may develop more than one complication. In Figure 5(b), the patient did not develop any complications during their hospital stay of two days. To compare both patients, the system’s predictions for patient (a) were relatively higher than those for patient (b). For example, the AKI predictions were 0.73 and 0.002, respectively, despite the fact that patient (a) developed AKI at around 20 days from admission. This demonstrates the value of our system in predicting the risk of developing complications early during the patient’s stay.

Table 5: Performance evaluation of the best performing models on test sets A & B, which were selected based on the average AUROC performance on the validation sets, as shown in Supplementary Section C. Model type indicates the type of the base learners within the final selected ensemble. All the metrics were computed using bootstrapping with 1,000 iterations [46].
Figure 3: The (a) ROC curves, (b) PRC curves, and (c) calibration curves are shown for all model ensembles evaluated on test set A (top) and test set B (bottom). The color legend for all figures is shown on the right. The numerical values for the AUROC, AUPRC, calibration slopes and intercepts can be found in Table 5.

Figure 4: The four most important features are shown for each complication in (a) test set A and (b) test set B. Feature importance was computed using the average SHAP values of the six models per ensemble.
Figure 5: Timeline showing the development of complications with respect to number of days from admission (x-axis) for two sample patients. (a) For \([y^{\text{SBI}}, y^{\text{AKI}}, y^{\text{ARDS}}]\), our system predictions (multiplied by a 100 to obtain percentages) were [64%, 73%, 51%]. (b) This patient did not develop complications and our model predictions were [0.2%, 0.2%, 2%].

4 Discussion

In this study, we developed an automated prognostic system to support patient assessment and triage early on during the patient’s stay. We demonstrate that the system can predict the risk of multiple complications simultaneously and achieves a good performance across all complications across two geographically independent datasets. The feature importance analysis revealed that age, systolic blood pressure and respiratory rate are highly predictive of several complications across the two datasets. Since COVID-19 was predominantly a pulmonary illness especially in its early variants [48], it was not surprising that respiratory rate ranked among the highest predictive features. We also identified age and systolic blood pressure as markers for severity among patients with COVID-19, which is aligned with clinical literature [49, 50]. Specifically, systolic blood pressure has been determined as an important covariate of morbidity and mortality in patients with COVID-19 [51]. This analysis demonstrates that our system’s learning is clinically meaningful and relevant.

In addition, we assessed our models’ calibration through reporting the calibration slopes and intercepts and visualized the calibration curves. Sufficiently large datasets are usually needed to produce stable calibration curves at model validation stage [39]. Despite the size of our dataset, we found that reporting the calibration slopes and intercepts would provide a concise summary of potential problems with our system’s calibration, to avoid harmful decision-making [39].

One of the main strengths of this study is that we used multicentre data collected at 18 facilities across several regions in Abu Dhabi, UAE. COVID-19 treatment is free for all patients in the UAE, hence there were no obvious gaps in terms of access to healthcare services in our dataset. Across the training and test sets in regions A & B, 15.2%-19.1% of encounters were for Arab patients. This reflects the diversity of our dataset, since Abu Dhabi is residence for more than 200 nationalities, of which only 19.6% of the population is Emirati. This diversity makes our findings relevant to a global audience. While most previous studies have focused on European or Chinese patient cohorts [52, 53], our study is one of few studies with large sample sizes (3,352 COVID-19 patient encounters) that focus on the patient cohort in the UAE. Compared to other international patient cohorts, our cohort is relatively younger (39.3-45.5 years across training and test sets), with a lower overall mortality rate (1.3%-3.7% across the training and test sets), suggesting that our system needs to be further validated on populations with different demographic distributions [10, 55, 56]. Our data-driven approach and open-access code can be easily adapted for such purposes.

Another strength is that our system predicts three complications simultaneously that are indicative of patient severity, in order to avoid poor patient outcomes. From a clinical perspective, several studies reported worse prognosis among patients with COVID-19 who had multi-organ failure, and co-infections [8, 57, 6, 11]. Most of the existing COVID-19 prognostic studies focus on predicting mortality as an adverse
The low mortality rates in our dataset strongly discouraged the development of a mortality risk prediction score, as such small sample sizes may lead to biased models [2]. An important aspect of this study is that the labeling criteria of the complications rely on renowned clinical standards and hospital-acquired data to identify the exact time of the occurrence of such complications. In collaboration with the clinical experts, this approach was considered more reliable than using International Classification of Disease (ICD) codes [58, 59]. Despite the development of new ontologies [60], ICD codes are generally used for billing purposes and their derivation may vary across facilities, especially during a pandemic [61]. We also introduce new benchmark results that can be contested with other competing models on test set B. Future work should also investigate the use of multi-label deep learning classifiers for larger datasets, while accounting for the exclusion criteria during training.

Moreover, our system uses routinely collected data and does not incur high data collection costs. Other prognostic machine learning studies have also adopted this strategy to predict adverse outcomes [62, 63]. By using routinely collected data rather than hematologic, cardiac, or biochemical laboratory tests that are associated with high processing times, our system is suitable for low-cost deployment. Existing studies achieved comparable performance with our system. For example, an AKI prediction model achieved 0.78 AUROC using findings of abdominal CT scans, vital-sign measurements, comorbidities, and laboratory-test results [25]. Although the results are not directly comparable due to differences in study design, our system achieved 0.91 and 0.89 AUROC in test sets A & B, respectively, without needing any imaging or laboratory-test results. In another study, an ARDS prediction model achieved 0.89 AUROC using patient demographics, interventions, comorbidities, 17 laboratory-test results and eight vital signs [33]. In comparison, our system achieved 0.85 AUROC in test set A and 0.83 AUROC in test set B. This implies that we should consider including additional variables to improve the performance of the model, such as laboratory-test results. One other study highlighted the predictive ability of eight laboratory-test results for the prediction of sepsis, where it achieved 0.93 AUROC [28]. We avoided the use of laboratory-test results to ensure that there is no overlap between the set of input features and the variables used to define the output complications (i.e. label leakage), however this is an area of future work.

Our study also has several limitations. One limitation of the labeling procedure is that it could miss patients for whom the data used in identifying a particular complication was not collected. However, this issue is more closely related to data collection practices at institutions as clinical data is often not completely missing at random. Another limitation is that since we relied on a minimal feature set, our system does not account for possible effects of treatment on the predicted outcomes and feature interactions, which is an area of future study. Moreover, the models are not perfectly calibrated due to small dataset size, which could also be attributed to the fact that the final predictions are based on model ensembles, rather than an individually calibrated model. Future work should investigate how to further improve the calibration of ensemble models. Furthermore, we utilized a dataset collected during the first wave of the pandemic, which did not include any information indicating the type of variant. Hence, the results presented here may not be directly applicable to patients with new COVID-19 variants. However, the system can be easily reused, fine-tuned, and validated using new datasets.

5 Conclusion

Our data-driven approach and results highlight the promise of machine learning in risk prediction in general and COVID-19 complications in particular. The proposed approach performs well when applied to two independent multicentre training and test sets in the UAE. The system can be easily implemented in practice due to several factors. First, the input features that our system uses are routinely collected by hospitals that accommodate patients with COVID-19 as recommended by the World Health Organization. Second, training the machine learning models within our system does not require high computational resources. Finally, through feature importance analysis, our system can offer interpretability, and is also fully automated as it does not require any manual interventions. To conclude, we propose a clinically applicable system that predicts complications among patients with COVID-19. Our system can serve as a guide to anticipate the course of patients with COVID-19 and to help initiate more targeted and complication-specific decision-making on
treatment and triage.

Acknowledgments
We would like to thank Waqqas Zia and Benoit Marchand from the Dalma team at NYU Abu Dhabi for supporting access to computational resources and Philip P. Rodenbough from the NYU Abu Dhabi Writing Center for revising the manuscript. This study was supported through the data resources and staff expertise provided by Abu Dhabi Health Services.

Contributors
GOG, BA, and KWY managed and analyzed the data. FS and II extracted, anonymized, and provided the dataset for analysis. GOG, KWY, and NH developed and maintained the experimental codebase. FAK, SAJ, MAS, RA, and MAH provided clinical expertise. WZ, FS, MAH and FES designed the study. MAH and FES supervised the work. GOG, BA, KWY, and FES wrote the manuscript. All authors interpreted the results and revised and approved the final manuscript.

Funding
The work of FES, GOG, BA, KWY, RA and NH is funded by NYU Abu Dhabi, and the work of FS, II, FAK, SAJ, MAS, and MAH is funded by Abu Dhabi Health Services. This work was also supported by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001.

Patient and public involvement
No patient involvement.

Competing interests
The authors declare no competing interests.

Data availability statement
To allow for reproducibility and benchmarking on our dataset, we are sharing test set B (n=225) at [https://github.com/nyuad-cai/COVID19Complications](https://github.com/nyuad-cai/COVID19Complications) We are unable to share the full dataset used in this study due to restrictions by the data provider. The trained models and the source code of the pipeline are also included in the repository.

References
[1] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.

[2] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna AA Damen, Thomas PA Debray, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, 369, 2020.

[3] John C Marshall, Srinivas Murthy, Janet Diaz, Neil Adhikari, Derek C Angus, Yaseen M Arabi, Kenneth Baillie, Michael Bauer, Scott Berry, Bronagh Blackwood, et al. A minimal common outcome measure set for covid-19 clinical research. *The Lancet Infectious Diseases*, 20(8):e192–e197, 2020.
[4] Riccardo Cau, Gavino Faa, Valentina Nardi, Antonella Balestrieri, Josep Puig, Jasjit S Suri, Roberto SanFilippo, and Luca Saba. Long-covid diagnosis: From diagnostic to advanced ai-driven models. *European Journal of Radiology*, 148:110164, 2022.

[5] Adam S Lauring, Mark W Tenforde, James D Chappell, Manjusha Gaglani, Adit A Ginde, Tresa McNeal, Shekhar Ghamande, David J Douin, H Keipp Talbot, Jonathan D Casey, et al. Clinical severity of, and effectiveness of mrna vaccines against, covid-19 from omicron, delta, and alpha sars-cov-2 variants in the united states: prospective observational study. *bmj*, 376, 2022.

[6] Wan Xu, Nan-Nan Sun, Hai-Nv Gao, Zhi-Yuan Chen, Ya Yang, Bin Ju, and Ling-Ling Tang. Risk factors analysis of covid-19 patients with ards and prediction based on machine learning. *Scientific Reports*, 11(1):1–12, 2021.

[7] Cigdem Yildirim, Hasan Sercuk Ozger, Emran Tombul, Ozlem Gubalbar, Mehmet Yildiz, Gulendam Bozdayi, Ulver Derici, and Murat Dizbay. Early predictors of acute kidney injury in covid-19 patients. *Nephrology*, 26:513–521, 2021.

[8] Marco Ripa, Laura Galli, Andrea Poli, Chiara Oltolini, Vincenzo Spagnuolo, Andrea Mastrangelo, Camilla Muccini, Giacomo Monti, Giacomo De Luca, Giovanni Landoni, et al. Secondary infections in patients hospitalized with covid-19: incidence and predictive factors. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 27(3):451–457, 2021.

[9] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Biabin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, Yan Zhao, Yirong Li, Xinghuan Wang, and Zhiyong Peng. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA*, 323(11):1061–1069, 03 2020.

[10] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The Lancet*, 395(10229):1054 – 1062, 2020.

[11] Pankaj Kumar and Malay Kumar. Management of potential ventilator shortage in india in view of on-going covid-19 pandemic. *Indian Journal of Anaesthesia*, 64(Suppl 2):S151, 2020.

[12] Xiangao Jiang, Megan Coffee, Anasse Bari, Junzhang Wang, Xinyue Jiang, Jianping Huang, Jichan Shi, Jianyi Dai, Jing Cai, Tianxiao Zhang, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua*, 63(1):537–551, 2020.

[13] Xiaobo Yang, Yuan Yu, Jiqian Xu, Huaqing Shu, Hong Liu, Yongran Wu, Lu Zhang, Zhui Yu, Minghao Fang, Ting Yu, et al. Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, 8(5):475–481, 2020.

[14] Jamie S Hirsch, Jia H Ng, Daniel W Ross, Purva Sharma, Hitesh H Shah, Richard L Barnett, Azzour D Hazzan, Steven Fishbane, Kenar D Jhaveri, Mersema Abate, et al. Acute kidney injury in patients hospitalized with covid-19. *Kidney International*, 98(1):209–218, 2020.

[15] Chih-Cheng Lai, Wen-Chien Ko, Ping-Ing Lee, Shio-Shin Jean, and Po-Ren Hsueh. Extra-respiratory manifestations of covid-19. *International Journal of Antimicrobial Agents*, 56(2):106024, 2020.

[16] Mylene Vaillancourt and Peter Jorth. vaillancourt2020unrecognized. *MBio*, 11(4):e01806–20, 2020.

[17] Gustavo Palacios, Mady Hornig, Daniel Cisterna, Nazir Savji, Ana Valeria Bussetti, Vishal Kapoor, Jeffrey Hui, Rafal Tolarz, Thomas Briese, Elsa Baumeister, et al. Streptococcus pneumoniae coinfection is correlated with the severity of h1n1 pandemic influenza. *PLoS One*, 4(12):e8540, 2009.
[18] Daniel S Chertow and Matthew J Memoli. Bacterial coinfection in influenza: a grand rounds review. *JAMA*, 309(3):275–282, 2013.

[19] Charles Feldman and Ronald Anderson. The role of co-infections and secondary infections in patients with covid-19. *Pneumonia*, 13(1):1–15, 2021.

[20] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[21] Jay L Koyner, Kyle A Carey, Dana P Edelson, and Matthew M Churpek. The development of a machine learning inpatient acute kidney injury prediction model. *Critical Care Medicine*, 46(7):1070–1077, 2018.

[22] Md Mohaimenul Islam, Tahmina Nasrin, Bruno Andreas Walther, Chieh-Chen Wu, Hsuan-Chia Yang, and Yu-Chuan Li. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer methods and programs in biomedicine*, 170:1–9, 2019.

[23] Norah Alballa and Isra Al-Turaiki. Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*, 24:100564, 2021.

[24] Maladieh Montazeri, Roxana Zahedi-Nasab, Ali Farahani, Hadis Moliseni, Fahimeh Ghasemian, et al. Machine learning models for image-based diagnosis and prognosis of covid-19: Systematic review. *JMIR medical informatics*, 9(4):e25181, 2021.

[25] Stefanie J Hectors, Sadjad Riyahi, Hreedi Dev, Karthik Krishnan, Daniel JA Margolis, and Martin R Prince. Multivariate analysis of ct imaging, laboratory, and demographical features for prediction of acute kidney injury in covid-19 patients: a bi-centric analysis. *Abdominal Radiology*, 46(4):1651–1658, 2021.

[26] Bradley J Langford, Miranda So, Sumit Raybardhan, Valerie Leung, Duncan Westwood, Derek R MacFadden, Jean-Paul R Soucy, and Nick Daneman. Bacterial co-infection and secondary infection in patients with covid-19: a living rapid review and meta-analysis. *Clinical Microbiology and Infection*, 2020.

[27] Noa Shafran, Inbal Shafran, Haim Ben-Zvi, Summer Sofer, Liron Sheena, Ilan Krause, Amir Shlomai, Elad Goldberg, and Ella H Sklan. Secondary bacterial infection in covid-19 patients is a stronger predictor for death compared to influenza patients. *Scientific Reports*, 11(1):1–8, 2021.

[28] Guoxing Tang, Ying Luo, Feng Lu, Wei Li, Xiongcheng Liu, Yucen Nan, Yufei Ren, Xiaofei Liao, Song Wu, Hai Jin, et al. Prediction of sepsis in covid-19 using laboratory indicators. *Frontiers in Cellular and Infection Microbiology*, 10, 2020.

[29] Farah E Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Koji Makino, Stanislaw Jastrzebski, Jan Witowski, Duo Wang, Ben Zhang, et al. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine*, 4(1):1–11, 2021.

[30] Adam L Booth, Elizabeth Abels, and Peter McCaffrey. Development of a prognostic model for mortality in covid-19 infection using machine learning. *Modern Pathology*, 34(3):522–531, 2021.

[31] Nathalie Lassau, Samy Ammari, Emilie Chouzenoux, Hugo Gortais, Paul Herent, Matthieu Devilder, Samer Soliman, Olivier Meyrignac, Marie-Pauline Talabard, Jean-Philippe Lamarque, et al. Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients. *Nature communications*, 12(1):1–11, 2021.

[32] Young Joon Kwon, Danielle Toussie, Mark Finkelstein, Mario A Cedillo, Samuel Z Maron, Sayan Manna, Nicholas Voutsinas, Corey Eber, Adam Jacobi, Adam Bernheim, et al. Combining initial radiographs and clinical variables improves deep learning prognostication in patients with covid-19 from the emergency department. *Radiology: Artificial Intelligence*, 3(2):e200098, 2020.
[33] Lakshya Singhal, Yash Garg, Philip Yang, Azade Tabaei, A Ian Wong, Akram Mohammed, Lokesh Chinthala, Dipen Kadaria, Amik Sodhi, Andre L Holder, et al. eards: A multi-center validation of an interpretable machine learning algorithm of early onset acute respiratory distress syndrome (ards) among critically ill adults with covid-19. *PloS one*, 16(9):e0257056, 2021.

[34] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. *British Journal of Surgery*, 102:148–158, 2015.

[35] Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184, 2012.

[36] The ARDS Definition Task Force. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA*, 307(23):2526–2533, 06 2012.

[37] David R Prytherch, Gary B Smith, Paul E Schmidt, and Peter I Featherstone. Views—towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[38] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[39] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016.

[40] Otto Nyberg and Arto Klami. Reliably calibrated isotonic regression. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 578–589. Springer, 2021.

[41] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

[42] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

[43] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv:1705.07874 [preprint]*, 2017.

[44] A Cecile JW Janssens and Forike K Martens. Reflection on modern methods: revisiting the area under the roc curve. *International journal of epidemiology*, 49(4):1397–1403, 2020.

[45] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[46] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, pages 189–212, 1996.

[47] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8):855–859, 2015.

[48] Magdalini Alexandridi, Julija Mazej, Enrico Palermo, and John Hiscott. The coronavirus pandemic–2022: Viruses, variants & vaccines. *Cytokine & Growth Factor Reviews*, 2022.
[49] Bart G Pijls, Shahab Jolani, Anique Atherley, Raissa T Derckx, Janna IR Dijkstra, Gregor HL Franssen, Stevie Hendriks, Anke Richters, Annemarie Venmans-Jellem, Saurabh Zalpuri, et al. Demographic risk factors for covid-19 infection, severity, icu admission and death: a meta-analysis of 59 studies. *BMJ open*, 11(1):e044640, 2021.

[50] Jinjun Ran, Ying Song, Zian Zhuang, Lefei Han, Shi Zhao, Peihua Cao, Yan Geng, Lin Xu, Jing Qin, Daihai He, et al. Blood pressure control and adverse outcomes of covid-19 infection in patients with concomitant hypertension in wuhan, china. *Hypertension Research*, pages 1–10, 2020.

[51] Antoine Cuillon, Kaigongzhao, Kathleen Oros Klein, Celia MT Greenwood, Zhibing Lu, Pierre Paradis, and Ernesto L Schiffrin. High systolic blood pressure at hospital admission is an important risk factor in models predicting outcome of covid-19 patients. *American Journal of Hypertension*, 34(3):282–290, 2021.

[52] Xuedi Ma, Michael Ng, Shuang Xu, Zhouning Xu, Hui Qiu, Yuwei Liu, Jiayou Lyu, Jiwen You, Peng Zhao, Shihao Wang, and et al. Development and validation of prognosis model of mortality risk in patients with covid-19. *Epidemiology and Infection*, 148:e168, 2020.

[53] Zhaozhi Qian, Ahmed M Alaa, and Mihaela van der Schaar. Cpas: the uk’s national machine learning-based hospital capacity planning system for covid-19. *Machine Learning*, 110(1):15–35, 2021.

[54] Simin Li, Yulan Lin, Tong Zhu, Mengjie Fan, Shicheng Xu, Weihao Qiu, Can Chen, Linfeng Li, Yao Wang, Jun Yan, et al. Development and external evaluation of predictions models for mortality of covid-19 patients using machine learning method. *Neural Computing and Applications*, pages 1–10, 2021.

[55] Parag Goyal, Justin J. Choi, Laura C. Pinheiro, Edward J. Schenck, Rujun Chen, Assem Jabri, Michael J. Satlin, Thomas R. Campion, Musarrat Nahid, Joanna B. Ringel, Katherine L. Hoffman, Mark N. Alshak, Han A. Li, Graham T. Wehmeyer, Mangala Rajan, Evgeniya Reshetnyak, Nathaniel Hupert, Evelyn M. Horn, Fernando J. Martinez, Roy M. Gulick, and Monika M. Safford. Clinical characteristics of covid-19 in new york city. *New England Journal of Medicine*, 382(24):2372–2374, 2020.

[56] Kyung Soo Hong, Kwan Ho Lee, Jin Hong Chung, Kyeong-Cheol Shin, Eun Young Choi, Hyun Jung Jin, Jong Geol Jang, Wonhwa Lee, and June Hong Ahn. Clinical features and outcomes of 98 patients hospitalized with sars-cov-2 infection in daegu, south korea: a brief descriptive study. *Yonsei Medical Journal*, 61(5):431, 2020.

[57] Jamie S. Hirsch, Jia H. Ng, Daniel W. Ross, Purva Sharma, Hitesh H. Shah, Richard L. Barnett, Azzour D. Hazzan, Steven Fishbane, Kenar D. Jhaveri, Mersema Abate, and et al. Acute kidney injury in patients hospitalized with covid-19. *Kidney International*, 98(1):209–218, 2020.

[58] Thomas L Higgins, Abhishek Deshpande, Marya D Zilberberg, Peter K Lindenauer, Peter B Imrey, Pei-Chun Yu, Sarah D Haessler, Sandra S Richter, and Michael B Rothberg. Assessment of the accuracy of using icd-9 diagnosis codes to identify pneumonia etiology in patients hospitalized with pneumonia. *JAMA network open*, 3(7):e207750–e207750, 2020.

[59] Jason P Burnham, Jennie H Kwon, Hilary M Babcock, Margaret A Olsen, and Marin H Kollef. Icd-9-cm coding for multidrug resistant infection correlates poorly with microbiologically confirmed multidrug resistant infection. *Infection control & hospital epidemiology*, 38(11):1381–1383, 2017.

[60] Ling Wan, Justin Song, Virginia He, Jennifer Roman, Grace Whah, Suyuan Peng, Luxia Zhang, and Yongqun He. Development of the international classification of diseases ontology (icdo) and its application for covid-19 diagnostic data analysis. *BMC bioinformatics*, 22(6):1–19, 2021.

[61] Brendan T Crabb, Ann Lyons, Margaret Bale, Valerie Martin, Ben Berger, Sara Mann, William B West, Alyssa Brown, Jordan B Peacock, Daniel T Leung, et al. Comparison of international classification of diseases and related health problems, tenth revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. *JAMA network open*, 3(8):e2017703–e2017703, 2020.
[62] Christopher Barton, Uli Chettipally, Yifan Zhou, Zirui Jiang, Anna Lynn-Palevsky, Sidney Le, Jacob Calvert, and Ritankar Das. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, 109:79–84, 2019.

[63] Nehemiah T Liu, John B Holcomb, Charles E Wade, Mark I Durrah, and Jose Salinas. Utility of vital signs, heart rate variability and complexity, and machine learning for identifying the need for lifesaving interventions in trauma patients. *Shock*, 42(2):108–114, 2014.

[64] Vitaly Herasevich, Murat Yilmaz, Hasrat Khan, Rolf D Hubmayr, and Ognjen Gajic. Validation of an electronic surveillance system for acute lung injury. *Intensive Care Medicine*, 35(6):1018–1023, 2009.

[65] Helen C. Azzam, Satjeet S. Khalsa, Richard Urbani, Chirag V. Shah, Jason D. Christie, Paul N. Lanken, and Barry D. Fuchs. Validation Study of an Automated Electronic Acute Lung Injury Screening Tool. *Journal of the American Medical Informatics Association*, 16(4):503–508, 07 2009.
Supplementary Information

A  Details of data pre-processing for labeling the complications

We used the KDIGO criteria to classify AKI encounters \cite{35}. The definition has three components, and if any of them are satisfied, then the patient is assigned a diagnosis of AKI. The three criteria were either an increase in serum creatinine of 0.3 mg/dl within 48 hours, an increase of 1.5 times the baseline serum creatinine measurement, or urine output of less than 0.5 ml/kg/hr for 6 hours \cite{35}. We only assessed the first two definitions, since urine output was not available in our dataset as it is usually measured in the intensive setting only. The patient’s first record of serum creatinine was treated as the baseline for that patient. Patients with reported chronic kidney disease were excluded from the training and testing AKI subsets.

The Berlin definition was employed to identify the timing and incidence of ARDS \cite{36}. The full ARDS labeling process is illustrated by the flow diagram in Figure S1. Textual chest X-ray reports and CT scan reports were processed using natural language processing (NLP) techniques to identify three categorized key terms: opacity, bilaterality, and ARDS. The lexicon developed was in reference to the Herasevich \cite{64} and ASSIST \cite{65} sniffers, which was further refined and validated based on clinical expertise. To minimize the influence of uncertainty profiles, the negation expression “no” was searched 40 characters prior to the identification of opacity. The ARDS diagnosis was confirmed if either one of the two criteria is satisfied: (1) the ARDS term is present or (2) both terms of bilaterality and opacity are present in the report. We identified the first radiology observation of bilateral opacity, as subsequent reports usually refer to the ones previously conducted for the identical patient instead of repeating the full interpretation and findings. Manual inspection of portions of the reports was done to validate the efficacy of the algorithm.

For the oxygenation criteria, 13,862 arterial partial pressure of oxygen (\(\text{PaO}_2\)) measurements acquired through arterial blood gas tests (ABG) were recorded for 358 unique patients. We have confirmed with SEHA clinicians that such test is only conducted for patients suspected of ARDS or with severe symptoms, and therefore, patients without one can be ruled out of ARDS directly. Each \(\text{PaO}_2\) measurement was matched with the closest prior record of \(\text{FiO}_2\) (the fraction of inspired oxygen) for the given patient to obtain the P/F ratio. For patients with missing \(\text{FiO}_2\) measurements, we assumed that they were not on oxygen therapy and were assigned a value of 0.2095 (20.95% of oxygen in air). The patients were then labeled as potentially having ARDS if their P/F ratio \(\leq 300\ \text{mm Hg}\).

The earliest recorded time —either arrival time, admission time, or the first time the patient tested positive for COVID-19 —was utilized in lieu of the precise point of clinical insult of respiratory symptoms for the timing criteria of the Berlin definition. To rule out pulmonary edema of other origin, patients with cardiac edema prior to the onset of ARDS were identified from the vitals and excluded. With the criteria and steps delineated herein, 243 patients were identified as having ARDS across both training sets as well as test sets.
Figure S1: The ARDS labeling process in our dataset, in accordance with the four criteria of the Berlin definition [36]: imaging, oxygenation, timing, and origin. The lexicon developed for identifying bilateral opacity in radiology reports is also shown within the table on the left.

B Hyperparameter search

Our system performs random search for the hyperparameters of the machine learning models and then evaluates their performance on the validation sets. The searched hyperparameters for each of the models are shown in Table A.

Table A: Hyperparameter values considered during the random hyperparameter search. Ranges are indicated with a ‘-‘.

| Model                              | Hyperparameters            | Values                      |
|------------------------------------|-----------------------------|-----------------------------|
|                                    | Regularization parameter C  | [1.e-03 - 1.e+01]           |
| Logistic Regression                | Max iterations              | [50-300]                    |
| K-Nearest Neighbors                | Leaf size                   | [20-60]                     |
|                                    | Power parameter             | [2]                         |
|                                    | N neighbors                 | [5 - 20]                    |
| Support Vector Machine             | Regularization parameter C  | [1.e-04 - 1.e+01]           |
|                                    | Gamma                       | [1.e-2-1.e+05]              |
| Light Gradient Boosting Model      | Number of leaves            | [10-60]                     |
|                                    | Learning rate               | [1.e-04- 1.e0]              |
|                                    | Max depth                   | [10-60]                     |
|                                    | N estimators                | [100-500]                   |
C  Model comparison

After preprocessing the data, we compared the performance of 4 ensembles based on 4 types of base learners on the validation sets: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Light Gradient Boosting Model (LGBM). The models were compared using the AUROC and AUPRC and the results are shown in Table B. We selected the ensemble that achieved the highest AUROC on the validation set.

Table B: Performance comparison for the different ensembles on the validation sets. Best performance is shown in bold.

| Complication | Models | Validation Set A | Validation Set B |
|--------------|--------|------------------|------------------|
|              | AUROC  | AUPRC | AUROC | AUPRC |
| SBI          | LR     | 0.895 | 0.411 | 0.980 | 0.596 |
|              | KNN    | 0.813 | 0.285 | 0.861 | 0.471 |
|              | SVM    | 0.853 | 0.321 | 0.965 | 0.543 |
|              | LGBM   | 0.888 | 0.382 | 0.934 | 0.395 |
| AKI          | LR     | 0.888 | 0.411 | 0.892 | 0.272 |
|              | KNN    | 0.777 | 0.234 | 0.770 | 0.166 |
|              | SVM    | 0.814 | 0.138 | 0.830 | 0.221 |
|              | LGBM   | 0.854 | 0.404 | 0.873 | 0.166 |
| ARDS         | LR     | 0.917 | 0.379 | 0.942 | 0.396 |
|              | KNN    | 0.814 | 0.245 | 0.809 | 0.290 |
|              | SVM    | 0.868 | 0.217 | 0.921 | 0.412 |
|              | LGBM   | 0.912 | 0.352 | 0.949 | 0.417 |
**Declaration of interests**

☑ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: