CHEST

Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation

Check for updates

Supreeth P. Shashikumar, PhD; Gabriel Wardi, MD, MPH; Paulina Paul, MS; Morgan Carlile, MD; Laura N. Brenner, MD; Kathryn A. Hibbert, MD; Crystal M. North, MD, MPH; Shibani S. Mukerji, MD; Gregory K. Robbins, MD, MPH; Yu-Ping Shao, MS; M. Brandon Westover, MD, PhD; Shamim Nemati, PhD; and Atul Malhotra, MD

BACKGROUND: Objective and early identification of hospitalized patients, and particularly those with novel coronavirus disease 2019 (COVID-19), who may require mechanical ventilation (MV) may aid in delivering timely treatment.

RESEARCH QUESTION: Can a transparent deep learning (DL) model predict the need for MV in hospitalized patients and those with COVID-19 up to 24 h in advance?

STUDY DESIGN AND METHODS: We trained and externally validated a transparent DL algorithm to predict the future need for MV in hospitalized patients, including those with COVID-19, using commonly available data in electronic health records. Additionally, commonly used clinical criteria (heart rate, oxygen saturation, respiratory rate, F_{IO_2}, and pH) were used to assess future need for MV. Performance of the algorithm was evaluated using the area under receiver operating characteristic curve (AUC), sensitivity, specificity, and positive predictive value.

RESULTS: We obtained data from more than 30,000 ICU patients (including more than 700 patients with COVID-19) from two academic medical centers. The performance of the model with a 24-h prediction horizon at the development and validation sites was comparable (AUC, 0.895 vs 0.882, respectively), providing significant improvement over traditional clinical criteria ($P < .001$). Prospective validation of the algorithm among patients with COVID-19 yielded AUCs in the range of 0.918 to 0.943.

INTERPRETATION: A transparent deep learning algorithm improves on traditional clinical criteria to predict the need for MV in hospitalized patients, including in those with COVID-19. Such an algorithm may help clinicians to optimize timing of tracheal intubation, to allocate resources and staff better, and to improve patient care.

CHEST 2021; 159(6):2264-2273

KEY WORDS: artificial intelligence; artificial respiration; coronavirus; deep learning; lung

ABBREVIATIONS: AUC = area under the receiver operating characteristic curve; COVID-19 = coronavirus disease 2019; DL = deep learning; EHR = electronic health record; IQR = interquartile range; MGH = Massachusetts General Hospital; ML = machine learning; MV = mechanical ventilation; PEEP = positive end-expiratory pressure; ROX = ratio of pulse oximetry/F_{IO_2} to respiratory rate; UCSD = University of California San Diego Health

AFFILIATIONS: From the Department of Biomedical Informatics (S. P. Shashikumar, P. Paul, S. Nemati), the Department of Emergency Medicine (G. Wardi, M. Carlile), the Division of Pulmonary, Critical Care, and Sleep Medicine (G. Wardi, A. Malhotra), University of California, San Diego, La Jolla, CA; the Division of Pulmonary and Critical Care Medicine (L. N. Brenner, K. A. Hibbert, C. M. North), the Department of Neurology (S. S. Mukerji, Y.-P. Shao, M. B. Westover),

The novel coronavirus pandemic, caused by severe acute respiratory syndrome coronavirus 2, has strained global health care systems[1] and the supply of mechanical ventilators,[2] because approximately 3% to 79% of hospitalized patients require invasive mechanical ventilation (MV).[3-7] Major concern exists regarding whether the supply of mechanical ventilators is insufficient for certain regions.[8,9] Appropriate triage and identification of patients at high risk for respiratory failure may help hospital systems to guide resource allocation better and to triage patients into treatment cohorts.[9,10] Additionally, identification of patients who may need intubation allows health care providers to prepare for endotracheal intubation (eg, by moving the patient to a negative pressure room), thereby preventing an emergent procedure that is inherently high risk and aerosol generating.[11-14] Related to fears of contamination, many providers decided to intubate early on the assumption that patients eventually will need MV so as to avoid crash intubation.[15] Others have called for more judicious use of MV and to avoid high positive end-expiratory pressure (PEEP) in poorly recruitable lungs, which tends to result in severe hemodynamic impairment and fluid retention.[16] Both patient self-inflicted lung injury and ventilator-associated lung injury could exacerbate lung inflammation and biotrauma.[17] As such, objective and consistent methods to determine who and when to intubate,[18] how to optimize treatment parameters, and when to extubate patients safely are needed to lower the long-term complications and mortality rate in this very sick patient population.

The field of machine learning (ML) refers to a subset of artificial intelligence that automates analytical model building to identify patterns in data to predict outcomes. In particular, ML algorithms are powerful tools for the detection of complicated and nonlinear outcomes when traditional statistical methods (eg, linear regression or decision trees) are overrun by a large number of variables. Deep learning (DL) models, a branch of ML, use multiple layers of processing (known as *artificial neural networks*), which can capture nonlinearity and complex interactions among clinical variables. Prior studies using DL-based algorithms have been shown to improve diagnostic accuracy and to predict outcomes across a variety of clinical scenarios.[19-26] Such algorithms can interpret and make useful predictions from large and dynamic data available in the electronic health record (EHR). Recently, we have shown ML algorithms to be superior to traditional metrics in the prediction of sepsis.[27]

Current scoring systems that predict respiratory failure and need for MV are limited by small sample size and have low predictive power.[28] Frontline providers have called for urgent development of new warning systems for patients in whom conservative management is likely to fail and who will require MV.[29] No reliable models exist to predict the need for MV in patients with COVID-19; therefore, we sought to use dynamic EHR data at hourly resolution to determine if such an approach would provide value over traditional methods such as the ratio of pulse oximetry/F_{IO_2} to respiratory rate (ROX) ROX index or simple regression-based risk scores.[28] In this study, we trained and prospectively validated a DL algorithm that predicts the need for invasive MV in hospitalized patients and those with known or suspected coronavirus disease 2019 (COVID-19) up to 24 h in advance of tracheal intubation.

Methods

Development and reporting of the prediction model presented in this study was in accordance with the checklist provided by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Consortium.[30]

Patient Population and Outcomes

An observational, multicenter cohort consisting of all adult patients (≥ 18 years of age) admitted to the ICU between January 1, 2016, and January 15, 2020, at two large urban academic health centers, University of California San Diego Health (UCSD) and

and the Division of Infectious Diseases (G. K. Robbins), Massachusetts General Hospital, Boston, MA.

FUNDING/SUPPORT: S. N. is supported by the National Institutes of Health (NIH) National Library of Medicine [Grant R56LM013517] and by NIH [Grant K01ES025445], the Biomedical Advanced Research and Development Authority [Grant HHS0100201900015C], and the Gordon and Betty Moore Foundation [Grant GBMF9052]. M. B. W. is supported by the NIH [Grants 1R01NS102190, 1R01NS102574, 1R01NS107291, and 1RF1AG064312]. A. M. is supported by the NIH [Grants R01 HL085188, K24 HL132105, T32 HL134632, R01 AG063925, R01 HL148436, R21 HL121794, R01 HL 119201, and R01 HL081823]. S. S. M. is funded by the NIH [Grant K23MH115812] and the Harvard Medical School Eleanor and Miles Shore Foundation. M. B. W. is supported by the Glenn Foundation for Medical Research and the American Federation for Aging Research through a Breakthroughs in Gerontology Grant; the American Academy of Sleep Medicine through an AASM Foundation Strategic Research Award; the Department of Defense through a subcontract from Moberg ICU Solutions, Inc., and by the NIH [Grants 1R01NS102190, 1R01NS102574, 1R01NS107291, and 1RF1AG064312). G. W. is supported by the National Foundation of Emergency Medicine and funding from the Gordon and Betty Moore Foundation [Grant GBMF9052].

CORRESPONDENCE TO: Atul Malhotra, MD; e-mail: amalhotra@health.ucsd.edu

Copyright © 2020 American College of Chest Physicians. Published by Elsevier Inc. All rights reserved.

DOI: https://doi.org/10.1016/j.chest.2020.12.009

Massachusetts General Hospital (MGH), was considered in this study. Throughout the article, we refer to the respective hospital systems as the *development* and the *validation* sites. Additionally, both datasets included prospectively collected temporal validation cohorts, involving known or suspected patients with COVID-19 between February 1 and May 4, 2020 (because of expansion of ICU care to nontraditional floors, the MGH cohort included all hospitalized patients with COVID-19 independent of explicit indication of ICU level of care). Patients were excluded if (1) their length of stay was less than 4 h or more than 20 days, or (2) the start of invasive MV occurred before hour 4 of ICU admission (or hospitalization for the MGH COVID-19 cohort), or (3) if they received noninvasive MV. Institutional review board approval of the study was obtained at both sites with a waiver of informed consent (UCSD Identifier: 191098; MGH Identifier: 2013P001024).

Data from both sites were abstracted into a clinical data repository (Epic Clarity; Epic Systems) and included vital signs, laboratory values, PEEP, Sequential Organ Failure Assessment scores, Charlson comorbidity index scores, demographics, and length of stay. Data were available to the treating clinician at the time of entry into the electronic health record electronic health record (EHR) and input into VentNet. Specific inputs to the model were prespecified and included 40 clinical variables (34 dynamic and six demographic variables) that were selected based on their availability in EHRs across the two hospitals considered in our study, similar to our previous work in the PhysioNet 2019 Challenge.[31] These included vital signs measurements (heart rate, pulse oximetry, temperature, systolic BP, mean arterial pressure, diastolic BP, respiration rate, and end-tidal CO_2), laboratory measurements (bicarbonate, measure of excess bicarbonate, F_{IO_2}, pH, P_{CO_2}, P_{O_2}, aspartate transaminase, BUN, alkaline phosphatase, calcium, chloride, creatinine, bilirubin direct, serum glucose, lactic acid, magnesium, phosphate, potassium, total bilirubin, troponin, hematocrit, hemoglobin, partial thromboplastin time, leukocyte count, fibrinogen, and platelets), and demographic variables (e-Table 1). Additionally, for every vital sign and laboratory variable, the slope of change since its last measurement was included as an additional variable. All variables were organized into 1-h nonoverlapping time bins to accommodate different sampling frequencies of available data. All the variables with sampling frequencies higher than once per hour were resampled uniformly into 1-h time bins by taking the median values if multiple measurements were available. Variables were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). Mean imputation was used to replace all remaining missing values (mainly at the start of each record). To assist in model training, features in the development site training set first underwent normality transformations and then were standardized by subtracting the mean and dividing by the SD. All other datasets were normalized using the mean and SD computed from the development site training set.

Use of MV was defined as the first occurrence of simultaneous recording of F_{IO_2} and PEEP. For prediction purposes, we defined our outcome of interest as continuous MV for at least 24 h or MV followed by death. Patients who were placed on a mechanical ventilator within 3 h of admission were excluded because our model makes its first prediction at hour 4 of ICU admission (or hospitalization in the case of the MGH COVID-19 cohort); this allowed for the collection and processing of laboratory samples required by the algorithm to make accurate predictions.

Model Development and Statistical Analyses

VentNet (a two-layer feedforward neural network of size 40 and 25) was trained to predict the onset of MV 24 h in advance, starting from hour 4 into admission up to the time of MV or end of hospitalization. Additionally, the predictions from VentNet were calibrated using isotonic regression.[32] VentNet was implemented in TensorFlow version 1.12.0 (Google Brain) and machine learning frameworks for Python version 2.7 (Python Software Foundation). The parameters of VentNet were initialized randomly and optimized on the training data from the development site using the gradient descent algorithm with L1-L2 regularization to avoid overfitting.[33] Model interpretability was achieved by calculating the relevance score[22] of each input variable for every predicted risk score (e-Appendix 1).

The output of VentNet was a probability score between 0 and 1. The decision threshold was chosen corresponding to an 80% sensitivity level. Any score beyond this threshold (0.03) implied that in the given prediction window, the algorithm predicted that the patient would undergo tracheal intubation within the prespecified period. A score of less than the decision threshold meant that VentNet did not predict tracheal intubation within the prediction window.

Within the development cohort, 10-fold cross-validation (with an 80%-20% split within each fold) was used for training and testing purposes. We report median and interquartile values of the area under the receiver operating characteristic curve (AUC; and specificity at 80% sensitivity) for the held-out testing sets within the development cohort (details on precision-recall curves are presented in e-Appendix 1). AUCs are reported under an end-user clinical response policy in which the model is silenced for 6 h after an alarm is fired, and correct alarms that are fired up to 72 h before onset of MV are not penalized. The best performing model at the development site then was fixed and used for evaluation on the validation cohort and the prospectively collected cohort of COVID-19 patients. Comparison of receiver operating characteristic curves was performed using DeLong's method.[34] All continuous variables are reported as medians with 25% and 75% interquartile ranges (IQRs). Binary variables are reported as percentages.

Results

Patient Characteristics

After applying the exclusion criteria, a total of 18,528 and 3,888 ICU patients were included in the development and validation cohorts, respectively. Patient characteristics including the percentage of ventilated patients before and after application of exclusion criteria are presented in Table 1 and e-Table 2. Additionally, data from 26 COVID-19 patients from the development site (UCSD) and 402 patients from the validation site (MGH) were used for prospective validation (Table 2, e-Table 3).

Model Performance on General ICU Populations

The median 10-fold cross-validated AUC on the held-out development site testing set for prediction horizon of 24 h was 0.886 (IQR, 0.878-0.892), and the specificity when measured at the 80% sensitivity level was 0.824

TABLE 1] Demographic Comparisons of the UCSD and MGH General ICU Cohorts

Demographics	UCSD (Development Site)		MGH (Validation Site)	
	Nonventilated	Ventilated	Nonventilated	Ventilated
Patients	17,723 (95.6)	805 (4.4)	3,602 (92.6)	286 (7.4)
Age, y	61.3 (48.3-72.6)	61.2 (48.6-71.2)	62 (51-72)	64 (53-74)
Male sex	10,421	521	1,948	173
Race				
White	9,659	440	2,925	229
Black	1,330	60	191	19
Asian	1,081	43	119	8
ICU LOS, h	48.3 (26.7-95.9)	221.5 (113.8-386.9)	50.9 (27.2-98.0)	183.7 (92.2-309.9)
CCI	3 (2-7)	3 (1-6)	4 (2-6)	4 (2-6)
SOFA score	0.6 (0-1.8)	3.3 (1.9-5.1)	0.9 (0.3-2.1)	4.1 (2.5-6.3)
Inpatient mortality	869	329	223	109
Time from ICU admission to start of ventilation, h	N/A	20 (7.8-45)	N/A	13 (6-33)

Data are presented as No. (%), No., or median (interquartile range), unless otherwise indicated. CCI = Charlson comorbidity index; LOS = length of stay; MGH = Massachusetts General Hospital; N/A = not applicable; SOFA = Sequential Organ Failure Assessment; UCSD = University of California San Diego Health. Patients were excluded if (1) their LOS was less than 4 h or more than 20 d or (2) the start of mechanical ventilation was before hour 4 of ICU admission.

(IQR, 0.818-0.838). We observed a drop in AUC when the prediction horizon increased from 6 h to 48 h (from 0.950 [IQR, 0.948-0.952] to 0.845 [IQR, 0.838-0.869], respectively) (e-Fig 1). Comparisons of the VentNet algorithm against the ROX index[28] and a logistic regression model (baseline model 1) based on commonly used clinical variables (namely, heart rate, oxygen saturation, respiratory rate, and pH) are shown in Figure 1. VentNet significantly outperformed the baseline models ($P < .001$) on the development site testing set (AUC, 0.895 vs 0.738 and 0.769, respectively) (Fig 1A). Performance of the VentNet on the external validation cohort (Fig 1B) was comparable (AUC, 0.882 vs 0.782 and 0.773, respectively). See e-Figure 1A, 1B, and e-Figure 2A, 2B, for additional information, including precision-recall curves. Additionally, the

TABLE 2] Demographic Comparisons of the Prospective Validation Cohorts Consisting of COVID-19 Patients at UCSD and MGH

Demographics	UCSD COVID-19		MGH COVID-19	
	Nonventilated	Ventilated	Nonventilated	Ventilated
Patients	16 (61.5)	10 (38.5)	343 (85.3)	59 (14.7)
Age, y	57.6 (45.2-81.6)	52.8 (42.3-65.9)	65 (47-78)	61.5 (50-73)
Male sex	9	7	176	40
Race				
White	7	< 5	207	30
Black	< 5	< 5	46	10
Asian	< 5	< 5	13	< 5
ICU LOS, h	51.4 (37.7-128.4)	368.7 (247.0-430.0)	131 (87.5-230)	258.5 (141-396)
CCI	4 (2.8-5.3)	2 (1-4.3)	3 (1-6)	3 (1-5)
SOFA	1.3 (0-2.1)	2.5 (0-5.4)	0.1 (0-0.7)	3.0 (1.6-4.7)
Inpatient mortality	< 5	< 5	24	14
Time from ICU admission to start of ventilation, h	N/A	23 (10-63)	N/A	49.5 (20.6-143)

Data are presented as No. (%), No., or median (interquartile range), unless otherwise indicated. Patients were excluded if (1) their LOS was less than 4 h or more than 20 d or (2) the start of mechanical ventilation was before hour 4 of ICU admission. CCI = Charlson comorbidity index; COVID-19 = coronavirus disease 2019; LOS = length of stay; MGH = Massachusetts General Hospital; N/A = not applicable; SOFA = Sequential Organ Failure Assessment; UCSD = University of California San Diego Health.

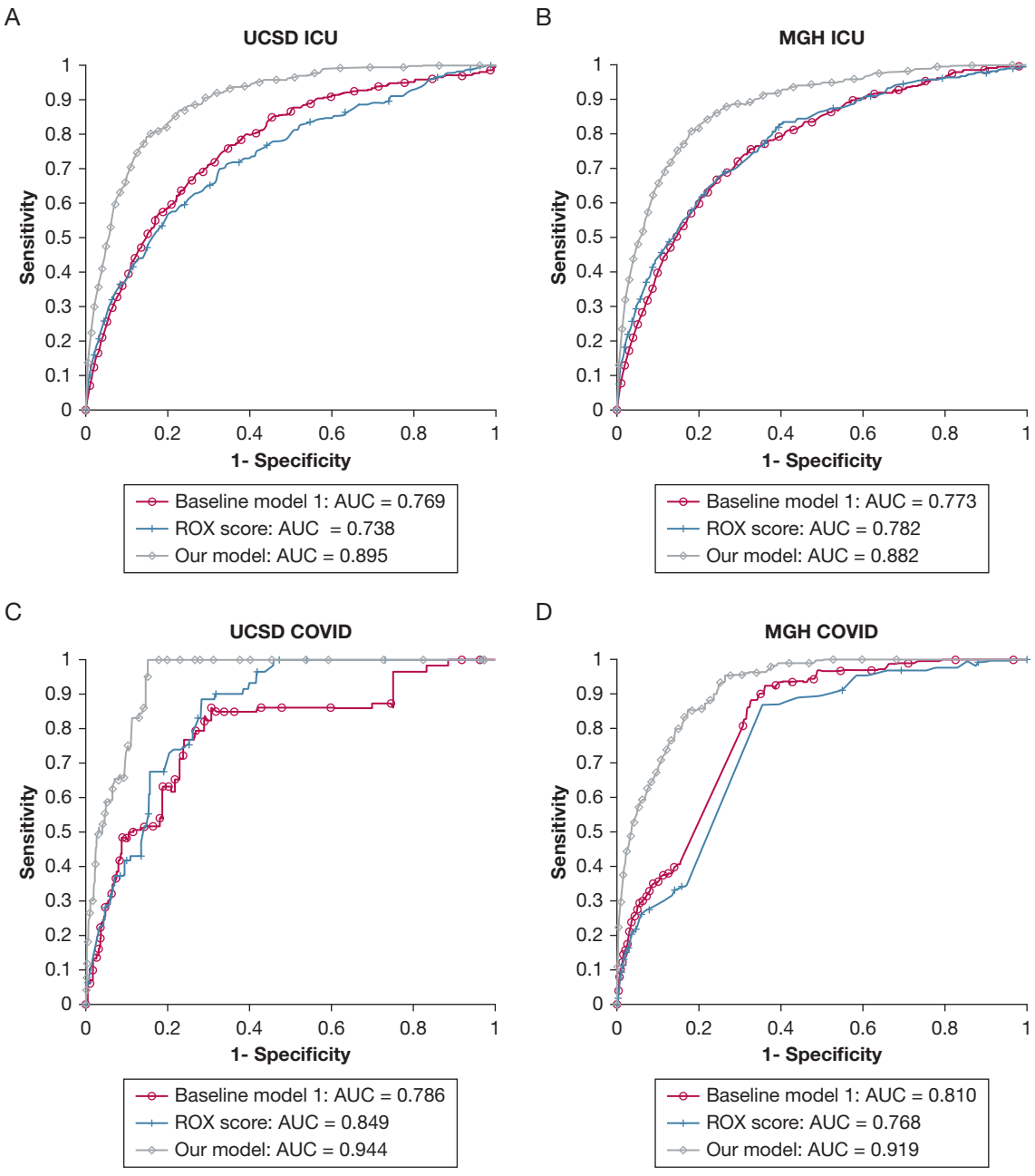

Figure 1 – *A-D, Line graphs showing the performance of the proposed and baseline models on the development and validation ICU cohorts and the two COVID-19 prospective validation cohorts. For a prediction horizon of 24 h, comparison of the proposed model vs two baseline models is shown on the development and validation ICU cohorts (A, B; P < .001) and prospective validation cohorts of patients with COVID-19 (C, D; P < .001). The baseline model 1 was a logistic regression model based on commonly used clinical variables (namely, heart rate, oxygen saturation, respiratory rate, and pH). AUC = area under the receiver operating characteristic curve; COVID-19 = coronavirus disease 2019; MGH = Massachusetts General Hospital; ROX = ratio of pulse oximetry/F_{IO_2} to respiratory rate; UCSD = University of California San Diego Health.*

calibration plots of VentNet on the development site testing set and the external validation cohort are shown in e-Figures 3, 4, 5A, 5B.

Figure 2A, 2B, shows heatmaps of the top 15 variables contributing to the increase in risk score up to 12 h before intubation for the development and the validation cohorts, respectively. Some of the most predictive features included respiratory rate, heart rate, temperature, chloride, oxygen saturation, platelet count, pH, and F_{IO_2}, among others. e-Figure 3 includes an illustrative example of clinical trajectory of a patient in the ICU, as well as the respective model predictions and the top contributing factors. Note that as shown in

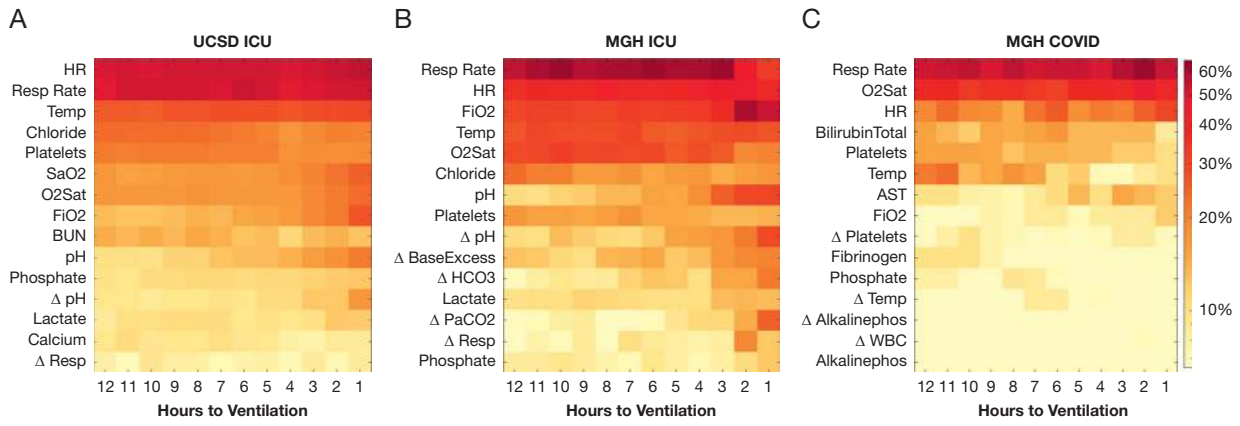

Figure 2 – *A-C, Heatmaps showing the population-level plot of top contributing factors to the increase in model risk score. The x-axis represents hours before onset time of mechanical ventilation. The y-axis represents the top factors (sorted by the magnitude of relevance score) across the patient populations at the development site (A), external validation site (B), and prospective COVID-19 cohort (C). Only dynamically changing variables are shown. Among the static factors, duration of time in hospital (to the current time) and sex (male) consistently were among the top factors. The heatmap shows the percentage of ventilated patients for whom a given variable was an important contributor to the risk score up to 12 h before intubation. See e-Appendix 1 and e-Figure 4 for more details. AST = aspartate transaminase; Δ = slope of change since last measurement; HR = heart rate; O2Sat = oxygen saturation; Resp = respiratory; SaO2 = saturation of arterial oxygen; Temp = temperature.*

e-Figure 4, a given risk factor can contribute to an increase in risk score by taking values either above or below the clinical reference range.

Model Performance on COVID-19 Populations

VentNet achieved superior performance when applied prospectively to the UCSD and MGH cohorts of patients with COVID-19 (AUC, 0.943 and 0.919, respectively). The corresponding specificities measured at 80% sensitivity level were 88.8% and 84.5%, respectively. See Figure 1C, 1D, and e-Figure 2C, 2D, for more information. Across both cohorts, performance of the VentNet was significantly better than the ROX score and the baseline model 1 ($P < .001$) (Fig 1, e-Fig 2). Additionally, the calibration plots of VentNet on both the UCSD and MGH COVID-19 cohorts are shown in e-Figure 5C, 5D.

Figure 2C shows a heatmap of the top 15 variables contributing to the increase in risk score up to 12 h before intubation for the COVID-19 cohort at the validation site. In addition to features listed above, other factors frequently contributing to the risk score in the COVID-19 population included total bilirubin, aspartate aminotransferase, fibrinogen, and phosphate, among others. Figure 3 includes an illustrative example of the clinical trajectory of a COVID-19 patient, as well as the respective model predictions and the top contributing factors.

Discussion

We demonstrated that a high-performing DL model (AUC > 0.88) can predict future need for MV 24 h in advance using commonly accessible EHR data. We externally validated all findings in patients from a separate academic center, as well as in two prospective cohorts of patients with COVID-19 (Fig 1). Because the proposed model can inform health care providers of the most relevant features contributing to the need for MV (Figs 2, 3), it provides an interpretable algorithm to aid clinicians with optimizing timing of tracheal intubation, better allocation of resources, and improving patient care. Importantly, the goal of algorithms such as this is not to replace clinical judgement, but rather to complement bedside care by providing predictions that can augment decision-making.

The COVID-19 pandemic has placed important strains on the health care system as the surge and long tail of critically ill patients continues to impact resource availability.[1] Despite having the highest number of ventilators and critical care beds per capita among developed countries, MV in the United States is still a finite resource.[8,9] Frontline providers in the pandemic noted that traditional risk stratification tools such as Modified Early Warning Score[35] and Quick Sequential Organ Failure Assessment score are inadequate to predict respiratory failure accurately in patients with COVID-19.[36] Recent data have shown that the ROX index has moderate usefulness for predicting tracheal intubation in patients with COVID-19.[37] However, VentNet showed a significantly higher AUC at all prediction windows compared with the ROX index. To our knowledge, this is the first study to demonstrate robust performance of a DL algorithm for early

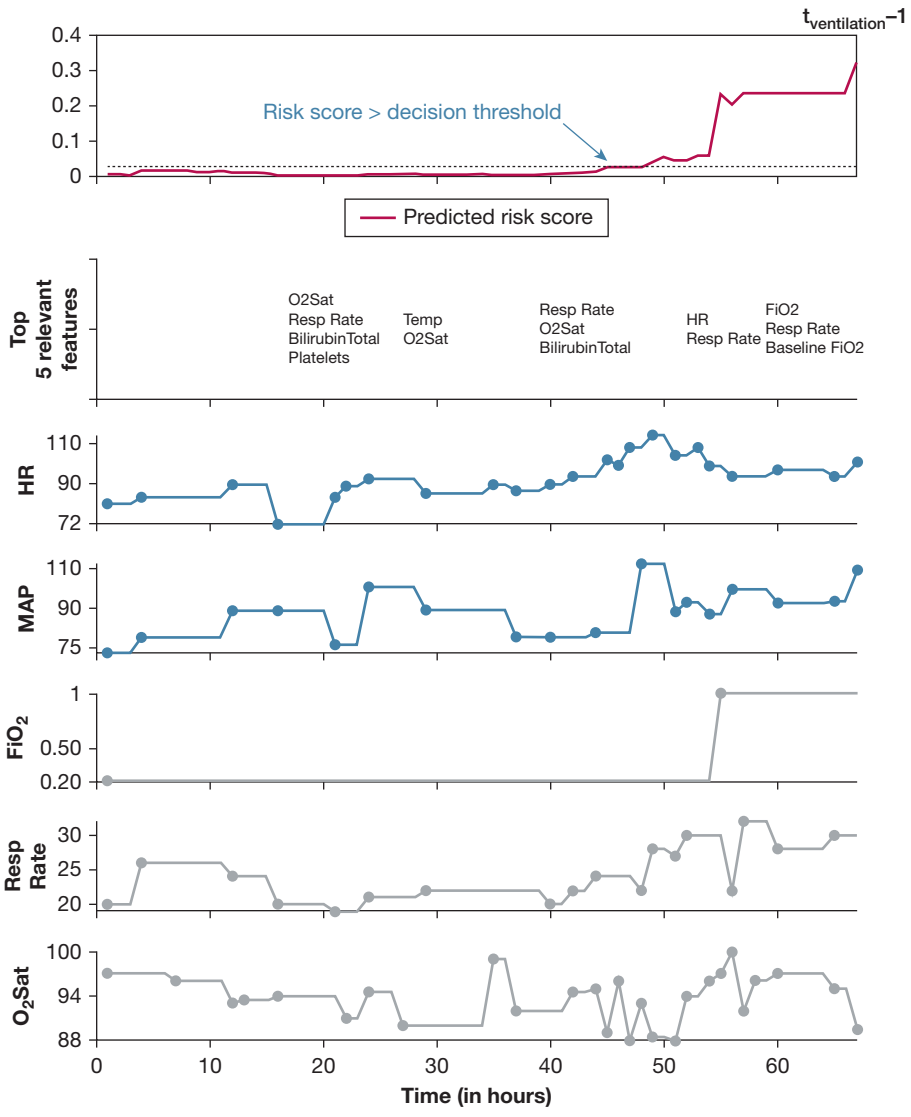

Figure 3 – *Illustrative example of a patient's trajectory over a 67-h window preceding intubation. The proposed algorithm crossed the prediction threshold at around hour 45 (highlighted by the red arrow), roughly 24 h before the onset time of mechanical ventilation. This 54-year-old woman with a history of hypothyroidism demonstrated fevers, chills, muscle aches, fever, sore throat, cough, and anosmia. She was admitted to the hospital for hypoxemia and a chest radiograph showing basilar patchy opacities present in the ED. She later showed positive results for COVID-19. Her oxygen requirements and work of breathing increased with a marked drop in oxygen saturation around hour 50. On the afternoon of the third day (hour 65) of hospitalization, she demonstrated rapidly progressive respiratory failure, was intubated, and was diagnosed with ARDS. For clarity, the top relevant features are shown every 5 h under the estimated risk scores. AST = aspartate transaminase; HR = heart rate; MAP = mean arterial pressure; O2Sat = oxygen saturation; Resp = respiratory; Temp = temperature.*

prediction of the need for MV in patients hospitalized with COVID-19.

We designed VentNet to be implemented in real time to augment clinician decision-making. All data input into VentNet were available to clinicians at the time of entry into the EHR. Such an algorithm can be implemented into the EHR, and we are actively pursuing this approach at our institutions. Previously, ML algorithms have been implemented into clinical workflow with improved clinical, statistical, or economic usefulness.[38] Additionally, we have included varying prediction windows to illustrate how VentNet performs at various time frames to illustrate potential uses (e-Fig 1). A shorter prediction horizon (eg, 6 h) may provide more clinically actionable information, whereas a longer prediction horizon (eg, 24-72 h) may inform population-level resource allocation. As anticipated, we observed a progressive drop in AUC when the prediction horizon increased from 6 to 48 h (from 0.950 to 0.845, respectively).

Our findings are important for a number of reasons. First, we developed and externally validated an interpretable DL algorithm that predicts the need for MV using commonly accessible clinical variables. Such findings could be used to facilitate optimal triage, more timely management, and resource use. Second, we showed with high predictive value the ability of our algorithm to function in different geographic settings in the United States and in varying cohorts. Third, our model used a sequential predictive approach such that ongoing clinical status was assessed to make important clinical predictions (see Fig 3 and e-Fig 3 for illustrative

examples). This strategy has advantages over a baseline assessment (eg, the Modified Early Warning Score and Quick Sequential Organ Failure Assessment) given the dynamic nature of critically ill patients. This approach paves the way for future implementation in real time at the point of care. Fourth, as shown in e-Tables 4 and 5 VentNet's predictions do not rely heavily on a single or a handful of clinical variables, and as such are more robust to data missingness. Thus, our model has both generalizability and portability and may have an impact not only on the current COVID-19 pandemic, but also on in the expected second wave and beyond.[39]

For a 24-h ahead prediction horizon, specificity of the model (on the MGH COVID-19 cohort) at 50% sensitivity was 96.5% (with a positive predictive value of 35.3%) vs 98.9% (with a positive predictive value of 39.2%) for 6 h. In terms of model optimization, one could argue the value in maximizing sensitivity, specificity, or both. In particular, during the COVID-19 pandemic, it has been argued that the avoidance of emergent procedures is a priority, because clearly a risk of viral transmission to providers exists and delays in intubation increase the risk of cardiovascular collapse.[40,41] Thus, a highly sensitive model may help to minimize the chance of a crash intubation,[42] which leads to poor clinical outcomes and may put providers at risk of unnecessary viral exposure. However, a highly specific model may be used to avoid unnecessary intubation,[15] and the associated risks of ventilator-induced lung injury, ventilator-associated pneumonia,[43] and sedation and associated delirium.[44,45] Additionally, a shorter prediction horizon (eg, 6 h) may provide more clinically actionable information, whereas a longer prediction horizon (eg, 24-72 h) may inform population-level resource allocation.

Despite its many strengths, this study includes a number of limitations. First, we defined the need for invasive MV in the EHR database based on the presence of PEEP and F_{IO_2} measurements. We believe that this definition is robust based on considerable experience, but acknowledge that some mislabeling (eg, noninvasive MV) could occur in any EHR-based criteria. Similarly, the delivery of noninvasive oxygen gives variable oxygen to the patient depending on inspiratory flow demand and breathing pattern; thus, our model likely could improve with more specificity from the EHR. Nonetheless, we view such misclassification as random and do not expect that any potential misclassifications would improve our model's performance artificially. Second, more generally, the proposed algorithm makes use of EHR data that was not designed originally for the analysis performed in our study. However, the superior performance of our algorithm, even in the presence of missing data, confirms its usefulness in a real-world clinical setting. Third, the COVID-19 pandemic has led to many changes in usual care, including potentially earlier intubation and avoidance of noninvasive ventilation, among others. Thus, one could argue that the need for intubation of these patients may be driven by factors unique to this epidemic. However, our model was trained and validated with historical data from major academic centers before the COVID-19 pandemic. Thus, the high observed AUCs speak to the robustness of the model, even in the face of rapid changes in practice patterns. Fourth, one could argue that the outcome of intubation and need for MV is somewhat subjective and could be a function of local practices or intrinsic bias inherent in such decisions. However, our ability to predict this clinically important outcome (need for MV) 6 to 24 h in advance suggests the value of this model. Moreover, traditional clinical parameters (heart rate, respiratory rate, pH, oxygen saturation) used to make intubation decisions performed relatively poorly compared with our DL algorithm (AUC, 0.769 vs 0.895 on the development site testing cohort). Despite these limitations, we view our new findings as robust and likely to lead to important advances in the care of COVID-19 patients. Furthermore, our approach may extend beyond the COVID-19 pandemic to guide optimal clinical care using advanced analytics as applied to the general ICU population, for example, to determine timing and selecting of appropriate pharmacologic therapies.

Interpretation

In this two-center observational study, we demonstrated that high-performance models can be constructed to predict the future need for MV in hospitalized patients, including those with COVID-19. By using open-source software, our validated algorithm is readily available for prospective studies aimed at determining the clinical usefulness of the proposed risk model for optimizing timing of tracheal intubation, better allocation of MV resources and staff, and improving patient care.

Acknowledgments

Author contributions: S. P. S, G. W., A. M. and S. N. were involved in the original conception and design of the work. S. P. S. and S. N. developed the network architectures, conducted the experiments and analyzed the data. P. P. and Y. S. assisted with acquisition of the data. S. P. S., G. W., M. C., A. M. and S. N. wrote the initial draft of the manuscript. All the authors assisted in editing and approval of the final version of the manuscript. S. N. and A. M. are guarantors for the paper.

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following: ResMed, Inc., provided a philanthropic donation to the University of California, San Diego, in support of a sleep center. A. M. received funding for medical education from Equillium, Corvus and Livanova. G. W. has received speaker's fees from Thermo-Fisher and consulting fees from General Electric. None declared (S. P. S, P. P., M. C., L. N. B., K. A. H., C. M. N., S. S. M., G. K. R., Y.-P. S., M. B. W., S. N.).

Role of sponsors: The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

Additional information: The e-Appendix, e-Figures, and e-Tables can be found in the Supplemental Materials section of the online article.

References

1. Emanuel EJ, Persad G, Upshur R, et al. Fair allocation of scarce medical resources in the time of Covid-19. *N Engl J Med*. 2020;382(21):2049-2055.
2. Feinstein MM, Niforatos JD, Hyun I, et al. Considerations for ventilator triage during the COVID-19 pandemic. *Lancet Respir Med*. 2020;8(6):e53.
3. Guan W, Ni Z, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720.
4. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506.
5. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061-1069.
6. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054-1062.
7. Cummings MJ, Baldwin MR, Abrams D, et al. Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet*. 2020;395(10239):1763-1770.
8. Halpern NA, Tan KS. United States resource availability for COVID-19. Society of Critical Care Medicine. https://www.sccm.org/Blog/March-2020/United-States-Resource-Availability-for-COVID-19. Accessed February 26 2021.
9. Truog RD, Mitchell C, Daley GQ. The toughest triage—allocating ventilators in a pandemic. *N Engl J Med*. 2020;382(21):1973-1975.
10. White DB, Katz MH, Luce JM, Lo B. Who should receive life support during a public health emergency? Using ethical principles to improve allocation decisions. *Ann Intern Med*. 2009;150(2):132-138.
11. World Health Organization. Infection prevention and control during health care when novel coronavirus (nCoV) infection is suspected Interim guidance. 19 March 2020. https://www.who.int/publications/i/item/10665-331495. Accessed February 26, 2021.
12. Hui DS. Severe acute respiratory syndrome (SARS): lessons learnt in Hong Kong. *J Thorac Dis*. 2013;5(Suppl 2):S122.
13. Respiratory care committee of Chinese Thoracic Society. Expert consensus on preventing nosocomial transmission during respiratory care for critically ill patients infected by 2019 novel coronavirus pneumonia [article in Chinese] [published online ahead of print February 20, 2020]. *Zhonghua Jie He He Hu Xi Za Zhi*. https://doi.org/10.3760/cma.j.issn.1001-0939.2020.0020.
14. Cheung JC-H, Ho LT, Cheng JV, Cham EYK, Lam KN. Staff safety during emergency airway management for COVID-19 in Hong Kong. *Lancet Respir Med*. 2020;8(4):e19.
15. Meng L, Qiu H, Wan L, et al. Intubation and ventilation amid the COVID-19 outbreak: Wuhan's experience. *Anesthesiology*. 2020;132(6):1317-1332.
16. Gattinoni L, Coppola S, Cressoni M, Busana M, Rossi S, Chiumello D. Covid-19 does not lead to a "typical" acute respiratory distress syndrome. *Am J Respir Crit Care Med*. 2020;201(10):1299-1300.
17. Gattinoni L, Chiumello D, Caironi P, et al. COVID-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive Care Med*. 2020;49(6):1099-1102.
18. Biddison LD, Berkowitz KA, Courtney B, et al. Ethical considerations: care of the critically ill and injured during pandemics and disasters: CHEST consensus statement. *Chest*. 2014;146(4):e145S-e155S.
19. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1(1):18.
20. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.
21. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-119.
22. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547-553.
23. Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382(18):1687-1695.
24. Ohno-Machado L. Data science and artificial intelligence to improve clinical practice and research. *J Am Med Inform Assoc*. 2018;25(10):1273.
25. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. 2014;16(1):441.
26. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29.
27. Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE - An interpretable and recurrent neural survival model for early prediction of sepsis. *Artif Intell Med*. 2021;113:102036.
28. Roca O, Caralt B, Messika J, et al. An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. *Am J Respir Crit Care Med*. 2019;199(11):1368-1376.
29. Goh KJ, Choong MC, Cheong EH, et al. Rapid progression to acute respiratory distress syndrome: review of current understanding of critical illness from COVID-19 infection. *Ann Acad Med Singapore*. 2020;49(1):1-9.
30. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;102(3):148-158.
31. Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med*. 2020;48(2):210-217.
32. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02). Association for Computing Machinery, New York, NY. July 2002:694-699. https://doi.org/10.1145/775047.775151.
33. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313(5786):504-507.
34. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
35. Stenhouse C, Coates S, Tivey M, Allsop P, Parker T. Prospective evaluation of a modified Early Warning Score to aid earlier detection of patients developing

critical illness on a general surgical ward. *Br J Anaesth*. 2000;84(5):663P.

36. Xie J, Tong Z, Guan X, Du B, Qiu H, Slutsky AS. Critical care crisis and some recommendations during the COVID-19 epidemic in China. *Intensive Care Med*. 2020;46(5):837-840.

37. Zucman N, Mullaert J, Roux D, Roca O, Ricard J-D. Prediction of outcome of nasal high flow use during COVID-19-related acute hypoxemic respiratory failure. *Intensive Care Med*. 2020;46(10):1924-1926.

38. Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*. 2020. https://doi.org/10.33590/emjinnov/19-00172.

39. Xu S, Li Y. Beware of the second wave of COVID-19. *Lancet*. 2020;395(10233):1321-1322.

40. Phua G-C, Govert J. Mechanical ventilation in an airborne epidemic. *Clin Chest Med*. 2008;29(2):323-328.

41. Wardi G, Villar J, Nguyen T, et al. Factors and outcomes associated with inpatient cardiac arrest following emergent endotracheal intubation. *Resuscitation*. 2017;121:76-80.

42. Flores MV, Cohen M. Preventing airborne disease transmission: implications for patients during mechanical ventilation. In: Esquinas A, ed. Noninvasive Ventilation in High-Risk Infections and Mass Casualty Events. Springer, Vienna; 2014. https://doi.org/10.1007/978-3-7091-1496-4_34.

43. Chastre J, Fagon J-Y. Ventilator-associated pneumonia. *Am J Respir Crit Care Med*. 2002;165(7):867-903.

44. Kotfis K, Williams Roberson S, Wilson JE, Dabrowski W, Pun BT, Ely EW. COVID-19: ICU delirium management during SARS-CoV-2 pandemic. *Crit Care*. 2020;24:1-9.

45. Dziadzko MA, Novotny PJ, Sloan J, et al. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. *Crit Care*. 2018;22(1):286.