Title

Development, deployment, and continuous monitoring of a machine learning model to predict respiratory failure in critically ill patients

Journal

JAMIA Open, 7(4)

ISSN

2574-2531

Authors

Lam, Jonathan Y

Lu, Xiaolei

Shashikumar, Supreeth P

et al.

Publication Date

2024-10-08

DOI

10.1093/jamiaopen/ooae141

Peer reviewed

JAMIA Open, 2024, **7(4)**, ooae141
https://doi.org/10.1093/jamiaopen/ooae141
Research and Applications

Research and Applications

Development, deployment, and continuous monitoring of a machine learning model to predict respiratory failure in critically ill patients

Jonathan Y. Lam, PhD[1], Xiaolei Lu, PhD[1], Supreeth P. Shashikumar, PhD[1], Ye Sel Lee, MS[1], Michael Miller, MD[2], Hayden Pour, MS[1], Aaron E. Boussina, PhD[1], Alex K. Pearce, MD[2], Atul Malhotra, MD[2], Shamim Nemati, PhD[1,*]

[1]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, United States, [2]Division of Pulmonary, Critical Care, and Sleep Medicine, University of California San Diego, La Jolla, CA 92093, United States

*Corresponding author: Shamim Nemati, PhD, Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive MC 0881, La Jolla, CA 92093, United States (snemati@health.ucsd.edu)

Abstract

Objectives: This study describes the development and deployment of a machine learning (ML) model called Vent.io to predict mechanical ventilation (MV).

Materials and Methods: We trained Vent.io using electronic health record data of adult patients admitted to the intensive care units (ICUs) of the University of California San Diego (UCSD) Health System. We prospectively deployed Vent.io using a real-time platform at UCSD and evaluated the performance of Vent.io for a 1-month period in silent mode and on the MIMIC-IV dataset. As part of deployment, we included a Predetermined Changed Control Plan (PCCP) for continuous model monitoring that triggers model fine-tuning if performance drops below a specified area under the receiver operating curve (AUC) threshold of 0.85.

Results: The Vent.io model had a median AUC of 0.897 (IQR: 0.892-0.904) with specificity of 0.81 (IQR: 0.812-0.841) and positive predictive value (PPV) of 0.174 (IQR: 0.148-0.176) at a fixed sensitivity of 0.6 during 10-fold cross validation and an AUC of 0.908, sensitivity of 0.632, specificity of 0.849, and PPV of 0.235 during prospective deployment. Vent.io had an AUC of 0.73 on the MIMIC-IV dataset, triggering model fine-tuning per the PCCP as the AUC was below the minimum of 0.85. The fine-tuned Vent.io model achieved an AUC of 0.873.

Discussion: Deterioration of model performance is a significant challenge when deploying ML models prospectively or at different sites. Implementation of a PCCP can help models adapt to new patterns in data and maintain generalizability.

Conclusion: Vent.io is a generalizable ML model that has the potential to improve patient care and resource allocation for ICU patients with need for MV.

Lay Summary

Earlier identification of patients at the highest risk of requiring mechanical ventilation (MV) offers an opportunity for timely medical interventions and efficient resource allocation. In this study, we developed a machine learning (ML) model called Vent.io for the prediction of MV up to 24 hours in advance using a combination of vital signs, laboratory measurements, comorbidities, medications, and demographic features. We trained Vent.io using intensive care unit (ICU) data from the University of California San Diego (UCSD) Health System and deployed it in our real-time predictive analytics platform with a Predetermined Changed Control Plan (PCCP) for continuous model monitoring that triggers model fine-tuning if performance drops below a specified area under the receiver operating curve (AUC) threshold of 0.85. Vent.io was prospectively validated after 1 month of deployment at UCSD and achieved an AUC of 0.908, meaning no model fine-tuning was required. We then simulated local deployment of Vent.io at an external site by evaluating Vent.io on the MIMIC-IV dataset. The resulting AUC of 0.73 was below the PCCP threshold, so Vent.io was fine-tuned using MIMIC-IV data, resulting in an AUC of 0.873. These results show Vent.io is generalizable and can aid clinicians in identifying high-risk patients for MV.

Key words: risk scoring system; machine learning; mechanical ventilation; electronic health records.

Introduction

Invasive MV is a vital intervention required for approximately 40% of the patients admitted to intensive care units (ICUs) due to severe respiratory failure, acute respiratory distress syndrome (ARDS), or other life-threatening conditions.[1] However, its use is complicated by the risk of ventilator-induced lung injury and complications resulting from prolonged MV.[2] Appropriate and timely triage of patients at the highest risk of requiring MV is critical. Earlier identification of this high-risk population offers an opportunity for timely medical interventions and allows hospital systems to allocate resources more efficiently.[3]

ML techniques are being increasingly integrated into medical practice[4] and may be especially valuable in the data-rich environment of the intensive care unit (ICU).[5] ML algorithms can help identify complex patterns, often before they are

Received: August 25, 2024; **Revised:** November 18, 2024; **Editorial Decision:** November 19, 2024; **Accepted:** November 25, 2024

obvious clinically, to enhance diagnostic and predictive capabilities to improve clinical care across different medical problems.[6] However, existing MV early prediction systems generally focus on model development and validation[7–11] and often fail to report on their deployment and ongoing monitoring, which is equally important. Deploying the MV early prediction system in clinical practice can provide timely reference for clinicians. However, a static model's predictive performance may deteriorate from development to deployment due to the shifts in patient populations, disease epidemiology, clinical care practices, and healthcare policies.[12] Therefore, dynamic monitoring is essential to be integrated into the model development process to build a reliable and durable MV early prediction system.

We previously developed a deep learning algorithm called VentNet to predict the need for MV up to 24 hours in advance using a combination of vital signs, laboratory measurements, and demographic features.[13] However, the model had limitations including lack of data on medications and comorbidities and lack of prospective validation and deployment. In the current study, we modified VentNet and developed an enhanced model named Vent.io. We incorporated an expanded labeling scheme to address various physiological states of respiratory failure and added 16 laboratory measurements, 11 SIRS and SOFA components, 12 types of medications, and 62 comorbidities to improve the generalizability of the model. The refined Vent.io model was then deployed in our predictive analytics platform, at the University of California San Diego (UCSD) Health System, for real-time early prediction of MV needs with a Predetermined Changed Control Plan (PCCP) for continuous model monitoring. The United States Food and Drug Administration (FDA) has published guidance for the ongoing maintenance and iterative improvement of clinical decision support software under the Software as a Medical Device (SaMD) designation.[14] As part of FDA guidance, a PCCP is recommended for SaMDs to ensure ongoing effectiveness in the face of any encountered changes, such as when a SaMD is deployed at a different setting or underlying data distributions are shifted. Formally, a PCCP is a plan outlining the planned modifications to a SaMD, the protocol for implementing and controlling those modifications, and the assessment of the impacts of the modifications.[15] For our use case, the PCCP systematically tracked the model's area under the receiver operating curve (AUC) over time. If performance fell below the specified AUC threshold of 0.85, the PCCP triggered model fine-tuning which involves retraining the model using prospective data. This fine-tuning process helps the model adapt to new patterns in the prospective cohort and maintain generalizability instead of being deployed with subpar performance. Overall, we aimed to develop a ML model for the prediction of MV using electronic health record (EHR) data, deploy it in the real-time setting, and evaluate its performance in a prospective cohort and a simulated external setting.

Methods

Study design and patient cohorts

We first conducted a retrospective cohort study using de-identified EHR data of all adult patients (≥18 years) who were admitted to the ICU between January 1, 2016 and December 31, 2023 at 2 hospitals within the UCSD Health System. Additionally, prospective data were collected from January 1, 2024 to January 31, 2024. Furthermore, we used data from ICU patients within the MIMIC-IV v2.2 database.[16] This study was completed in accordance with the ethical standards of the UCSD on human experimentation. IRB approved protocol #800258 with waiver of consent ("A Real-Time Multimodal Data Integration Model for Prediction of Respiratory Failure in Patients with COVID-19") was initially approved on August 30, 2021, with a latest approval date of February 8, 2024.

Patients were excluded if (1) their ICU length of stay was less than 5 hours, (2) they were mechanically ventilated before ICU admission, (3) there was no measurement of heart rate, blood pressure or labs prior to the prediction start time, or (4) had a Do Not Resuscitate (DNR) order in place. Timestamps up to 24 hours before and after surgery were also excluded to remove surgery-related ventilation events. For prediction purposes, patients were monitored throughout their ICU stay until either (1) the time of MV or (2) the time of transfer out of the ICU. To allow for adequate data collection, predictions commenced 4 hours after ICU admission and were updated on an hourly basis based on the newest clinical data.

The overall dataset was divided into a development cohort consisting of UCSD encounters and a local validation cohort consisting of MIMIC-IV encounters. The UCSD development cohort was further randomized into a training cohort (consisting of 80% of encounters) and testing cohort (consisting of the remaining 20% of encounters). The development cohort was used for model training and internal testing while the local validation cohort was solely used for model testing purposes with model parameters initialized from the final UCSD trained model. We utilized a custom 5-point labeling scale for MV (Section S1) for use during training to account for the various physiological states of respiratory failure. For model evaluation purposes, a Vent.io score of ≥3 was defined as the positive class, and a Vent.io score of <3 was defined as the control class.

Model features

Data from UCSD Health System were extracted from a clinical data repository (Epic Clarity; Epic Systems). Vent.io model features are similar to a previously published model called DETERIO[17] and consisted of 50 vital signs and laboratory measurements, 6 demographic features, 11 Systemic Inflammatory Response Syndrome (SIRS) and Sequential Organ Failure Assessment (SOFA) criteria, 12 medication categories, and 62 comorbidities (Section S2). The vital signs and laboratory measurements were grouped into 1-hour time series bins to account for varying data sampling frequencies. Variables sampled more than once per hour were resampled into hourly bins using the median. Updates were made hourly with new data and if no new data were present, existing values were carried forward for up to 24 hours. All remaining missing values were replaced using mean imputation. We reported missing data on an hourly basis for the 50 vital signs and laboratory measurements (Table S2). In addition to the 142 clinical variables, we calculated 150 features derived from the 50 vital signs and laboratory measurements. For each vital sign and laboratory measurement, we derived baseline values (mean value measured over the previous 72 hours), local trends (change since last measurement), and the time since the variable was last measured (TSLM). Predictions were made on an hourly basis using all 292 features.

Model development, evaluation, and statistical analyses

Vent.io is a 3-layer feedforward neural network of size 100, 80, and 64, similar to a previously published model for early prediction of sepsis called COMPOSER.[18] Vent.io was trained using a temporal difference learning approach to predict the future value of a patient state using the value iteration algorithm,[19] starting from fifth hour of ICU admission up to the first instance of MV or transfer out of the ICU. In this context, the patient state was represented as a 64-dimensional vector, which was mapped to a single state value through a fully connected neural network layer. All model development and training was done using Tensorflow 2.10.

Model performance was evaluated by thresholding on the predicted state value. For control patients, the model was trained to predict up to ICU discharge or 14 days, whichever occurred first. Vent.io included a conformal prediction module similar to COMPOSER designed to identify out-of-distribution samples, thereby establishing the model's "conditions for use."[20] The parameters of Vent.io were randomly initialized and optimized using the training dataset from the development cohort with L1-L2 regularization and dropout in the hidden layers to prevent overfitting. The decision threshold was chosen corresponding to 60% sensitivity at the encounter level based on clinical feedback to reduce the number of false positives. A predicted risk score beyond this threshold meant that Vent.io predicted that the patient would undergo MV within the prediction window (up to 24 hours before the time of MV T0). A predicted risk score less than the decision threshold meant that Vent.io did not predict MV within the prediction window. Additionally, Vent.io was made interpretable by calculating the relevance score of each input variable for every predicted risk score. To compute the relevance score, we took the derivative of the risk score with respect to all input features and multiplied it by the input features. The most relevant features contributing to the risk score have the largest magnitude of the relevance score with the direction of influence determined from the sign of the input gradients.

We have reported the median and interquartile range for all continuous variables and percentages for all binary variables. The area under receiver operating characteristic curve (AUC) has been reported at the hourly window level. Specificity (SPC), sensitivity (SEN), and positive predictive value (PPV) at a fixed decision threshold have been reported at the encounter level. The procedure to determine the number of true positives, false positives, true negatives, and false negatives required to compute SPC, SEN, and PPV at the encounter level has been described in Section S3. AUC was calculated under the same end-user clinical response policy as our prior paper[13] where alarms fired up to 72 hours in advance were suppressed and the model silenced for 6 hours after an alarm was fired.

Predetermined change protocol plan

Our PCCP for Vent.io is tailored to 2 scenarios: existing and new healthcare systems. In existing systems, model performance is monitored continuously over rolling 1-month windows, using real-time gold-standard labels derived from bulk FHIR data. If performance falls below the PCCP threshold, data from the preceding 4 months is extracted via bulk FHIR for model fine-tuning. In new healthcare systems, a one-time bulk FHIR data pull (over the preceding 5 months or more) establishes baseline performance. If performance is below the threshold, these data are temporarily split 80/20 for fine-tuning and testing, respectively. The model is deployed in the new healthcare system only if the testing performance exceeds the PCCP threshold. After deployment or fine-tuning, performance monitoring resumes on a rolling window basis. This approach ensures ongoing model quality and regulatory compliance, allowing for system-specific adjustments in both existing and new healthcare environments.

Prospective development at UCSD health

The Vent.io model was deployed in "silent mode" for real-time early prediction of MV within a 24-hour window in the ICU on a cloud-based platform. Our cloud-based analytics platform was developed to have real-time access to data elements in the EHR by leveraging the FHIR and HL7v2 standards.[21] The real-time platform extracted data at an hourly resolution of all the active patients across all of ICUs within UCSD Health System using FHIR APIs with OAuth 2.0 authentication. It processes an input feature set consisting of laboratory measurements, vitals measurements, comorbidities, medications, and demographics and passes these data to the Vent.io inference engine. The inference engine consisted of Vent.io microservice hosted within an EC2 instance. The Vent.io risk scores generated by the Vent.io pipeline were written to a flowsheet within the EHR using an HL7v2 outbound message. The schematic diagram of the real-time deployment pipeline is shown in Figure 1. The Vent.io pipeline was deployed in silent mode for the real-time prediction of requiring MV across all ICUs within the UCSD Health System starting from January 1, 2024 with the PCCP in place.

Results

Patient characteristics

From UCSD Health System, 26 045 and 276 ICU encounters were included in the development and real-time prospective validation cohorts, respectively, after applying the exclusion criteria. Patient characteristics from these 2 cohorts are listed in Table 1. To evaluate the external generalizability of the Vent.io model, 35 534 encounters from the MIMIC-IV database were used for local validation with patient characteristics listed in Table 2.

Model performance on the UCSD development and prospective cohorts

Vent.io achieved improved performance on the UCSD cohort based on a 10-fold cross-validation median AUC of 0.897 (IQR: 0.892-0.904) using the expanded feature set compared to a median AUC of 0.886 (IQR: 0.878-0.892) from VentNet. At a 60% sensitivity level, Vent.io had a median specificity of 0.825 (IQR: 0.812-0.841) and PPV of 0.162 (IQR: 0.148-0.176). The final model with the conformal prediction module using fixed parameters from the best-performing cross validation fold had an AUC of 0.908, sensitivity of 0.602, specificity of 0.81, and PPV of 0.174. Vent.io also demonstrated robust performance when applied to the prospective data from UCSD. The final model achieved a sensitivity of 0.632, specificity of 0.849, and PPV of 0.235. Since the AUC was above the minimum of 0.85 established by the PCCP, further model fine-tuning was not required. Top predictive features included respiratory rate, heart rate (HR),

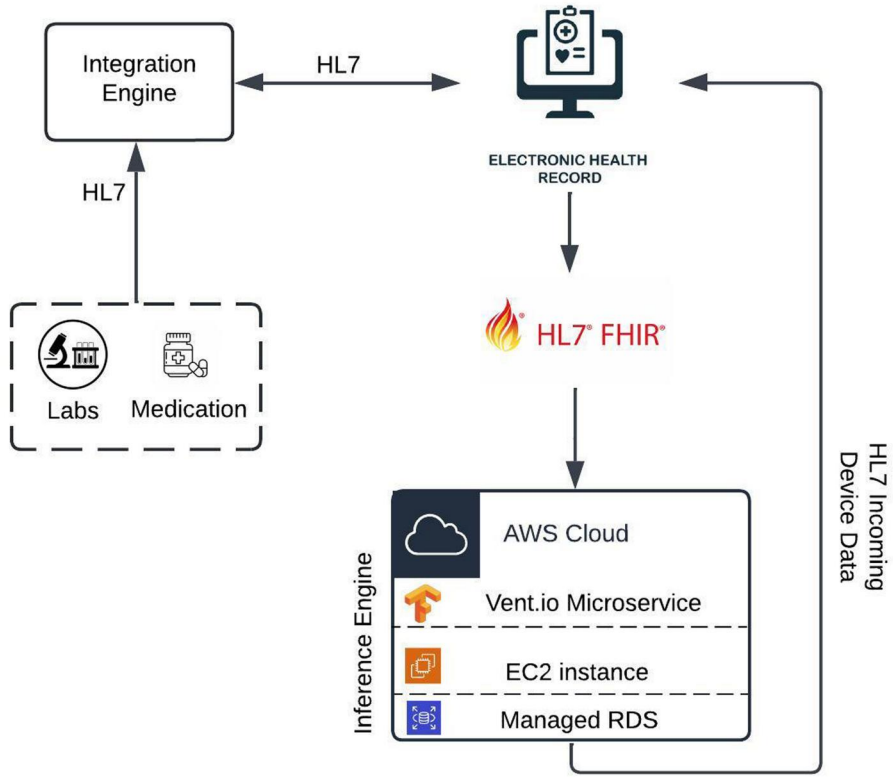

Figure 1. Schematic diagram of the Vent.io real-time deployment pipeline. The real-time platform extracts data at an hourly resolution of all active patients using FHIR APIs and passes the input feature set (consisting of laboratory measurements, vitals measurements, comorbidities, medications, and demographics) to the Vent.io inference engine. The Vent.io risk scores generated by the Vent.io pipeline are then written back to the EHR as a flow sheet item through an HL7 device data interface. The flowsheet then triggers a nurse facing Best Practice Advisory that alerts the caregiver that the patient is at risk of needing mechanical ventilation within a 24-hour window. AWS = Amazon Web Services; EC2 = Elastic Compute Cloud; FHIR = Fast Healthcare Interoperability Resources; HL7 = health level 7; RDS = relational database service.

Table 1. Patient characteristics for the UCSD cohort.

	Development		Prospective validation	
	Non-ventilated	Ventilated	Non-ventilated	Ventilated
No. of encounters (%)	24 715 (94.9%)	1330 (5.1%)	260 (94.2%)	16 (5.8%)
Age (in years), median [IQR]	61.7 [48.4-72.7]	61.6 [40.1-71.6]	65.0 [54.9-75.3]	62.0 [52.7-68.1]
Gender (male), %	58.4	63.4	56.2	68.8
Race				
White, %	48.7	49.1	45.4	37.5
African American, %	7.7	6.2	5.4	6.3
Asian, %	6.3	6.5	9.2	6.3
ICU length of stay (in hours), median [IQR]	51.8 [32.5-93.4]	222.3 [123.4-384.6]	73.2 [45.3-132.1]	337.4 [247.5-463.2]
CCI, median [IQR]	2 [0-3]	2 [1-3]	2 [1-5]	3 [2-6]
SOFA, median [IQR]	1 [2-4]	8 [10-12]	3 [1-4]	9 [7-11.5]
Mortality, %	3.5	33.6	2.7	18.8
Time from ICU admission to T0 (in hours), median [IQR]	N/A	28 [11-57]	N/A	15 [6.5-55.5]

CCI = Charlson comorbidity index; ICU = intensive care unit; SOFA = Sequential Organ Failure Assessment; T0 = time of mechanical ventilation initiation.

lymphocytes differential, blood urea nitrogen (BUN), albumin, oxygen saturation (O_2Sat), and white blood count (Figure 2A). In terms of directionality, an elevated respiratory rate, HR, and BUN and a lower lymphocyte differential count, albumin, and O_2Sat increased the risk for the need for MV (Figure 3A).

Model performance on the MIMIC-IV cohort

We used ICU encounters from the MIMIC-IV database to simulate the deployment of Vent.io at an external site. The final Vent.io model trained using UCSD data obtained an AUC of 0.73, lower than the minimum of 0.85 set by the PCCP, prompting a fine-tuning process. Starting with the parameters from the final Vent.io model, we fine-tuned Vent.io with different proportions of MIMIC-IV training data. Acceptable performance was achieved with 25% or more of the training data used to fine-tune the model (Figure 4). Respiratory rate, HR, BUN, O_2Sat, and WBC remained top features when evaluating Vent.io on MIMIC data with the same directionality while lymphocyte differential count and

Table 2. Patient characteristics for the MIMIC cohort.

	Non-ventilated	Ventilated
No. of encounters (%)	30 075 (84.6%)	5459 (15.4%)
Age (in years), median [IQR]	64 [51-76]	65 [55-75]
Gender (Male), %	60.8	52.0
Race		
White, %	69.2	69.4
African American, %	13.2	8.8
Asian, %	3.3	2.8
ICU length of stay (in hours), median [IQR]	141 [85-237]	257 [158-423]
CCI, median [IQR]	5 [2-7]	5 [3-7]
SOFA, median [IQR]	1 [1-3]	2 [1-3]
Mortality, %	5.8	19.7
Time from ICU admission to T0 (in hours), median [IQR]	N/A	10 [6-25]

CCI = Charlson comorbidity index; ICU = intensive care unit; SOFA = Sequential Organ Failure Assessment; T0 = time of mechanical ventilation initiation.

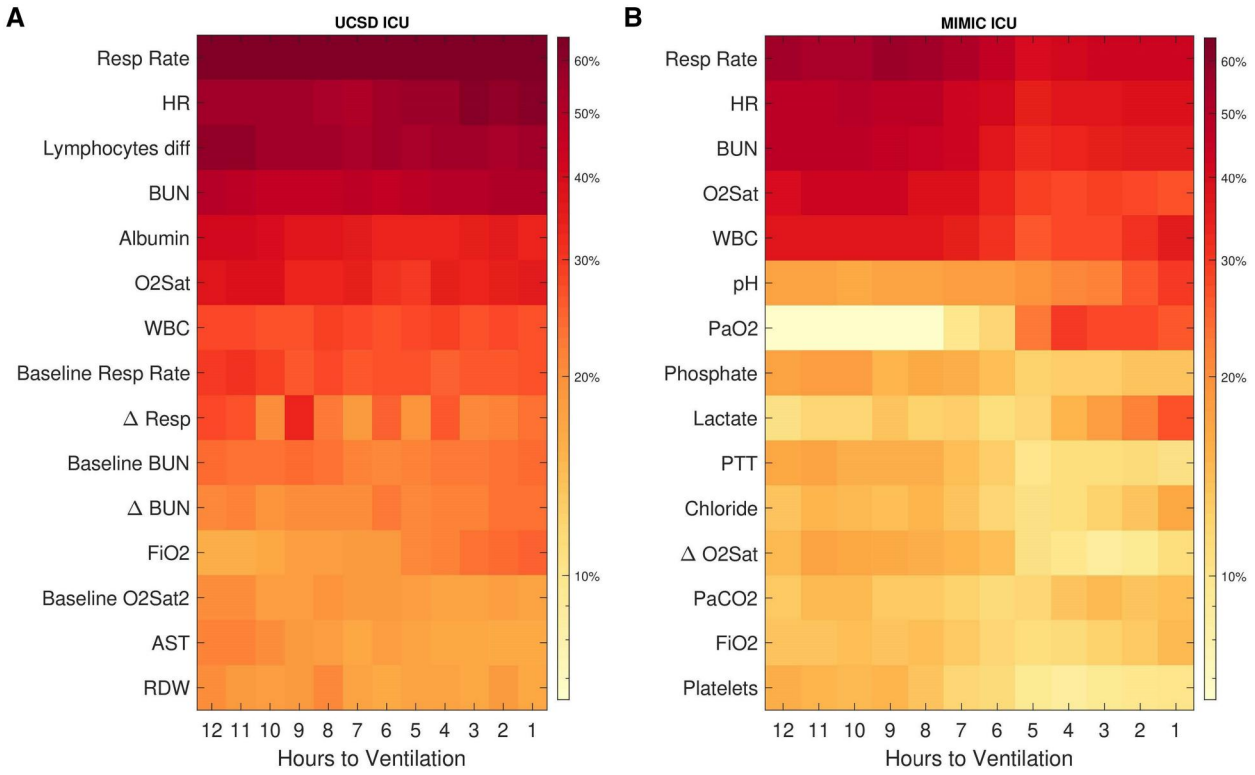

Figure 2. Population-level plot of top contributing factors to the increase in model risk score. The x-axis represents hours prior to onset time of mechanical ventilation. The y-axis represents the top factors (sorted by the magnitude of relevance score) across the patient populations at the (A) UCSD cohort and (B) MIMIC cohort. Resp rate = respiratory rate; HR = heart rate; Lymphocytes diff = lymphocyte differential count; BUN = blood urea nitrogen; O_2Sat = oxygen saturation; WBC = white blood count; AST = aspartate aminotransferase; RDW = red blood cell distribution width; PaO_2 = partial pressure of oxygen; PTT = partial thromboplastin time; $PaCO_2$ = partial pressure of carbon dioxide; FiO_2 = fraction of inspired oxygen.

albumin are replaced by pH and partial pressure of oxygen (Figures 2B and 3B).

Discussion

Our study adds to the existing literature in several important ways. First, we updated a deep learning model that allows the prediction of the need for MV for inpatients at risk of respiratory failure. Second, our model is externally generalizable with minor local retraining to a large external cohort of patients. Third, we have conducted a real-time implementation of our algorithms in the electronic health record which should form the basis for prospective randomized trials in the near future. Our updated model utilizes an expanded labeling scheme that incorporates intermediate phenotypes such as severe hypoxemia. The increased granularity of the labels enables our model to account for patients who are at increased risk for MV but have not been put on a ventilator. Furthermore, the addition of specific features such as medications and comorbid conditions leads to clinical actionability. Our algorithm also considers code status such as "Do Not Intubate/Resuscitate" and exclusion of patients intubated for surgical cases. Together, these changes enable Vent.io to make more accurate predictions.

Accurate and early prediction of patients at high risk for invasive MV is an important support tool for clinicians to

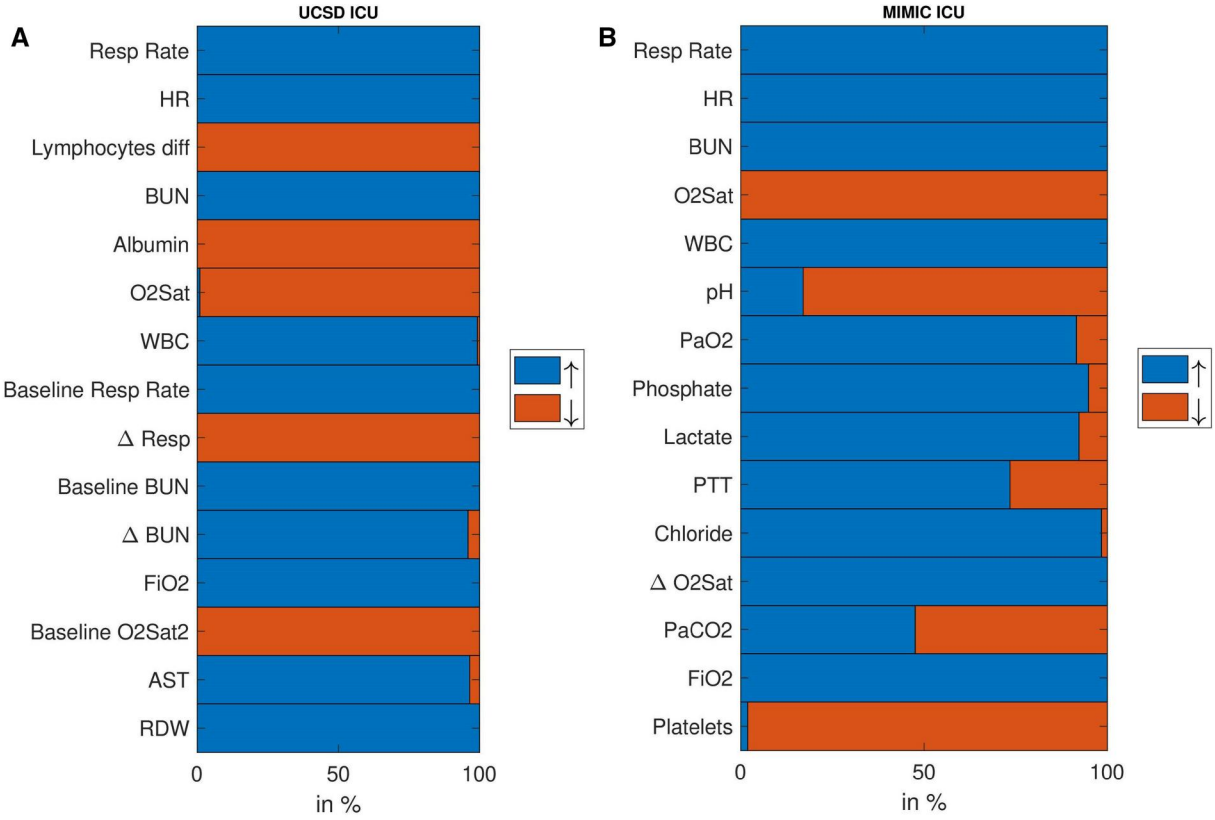

Figure 3. Directionality with respect to influence of top factors contributing to an increase in the risk score. The x-axis represents the percentage contribution of each feature to the risk score. The y-axis represents the top factors (sorted by the magnitude of relevance score) across the patient populations at the (A) UCSD cohort and (B) MIMIC cohort. Resp rate = respiratory rate; HR = heart rate; Lymphocytes diff = lymphocyte differential count; BUN = blood urea nitrogen; O_2Sat = oxygen saturation; WBC = white blood count; AST = aspartate aminotransferase; RDW = red blood cell distribution width; PaO_2 = partial pressure of oxygen; PTT = partial thromboplastin time; $PaCO_2$ = partial pressure of carbon dioxide; FiO_2 = fraction of inspired oxygen.

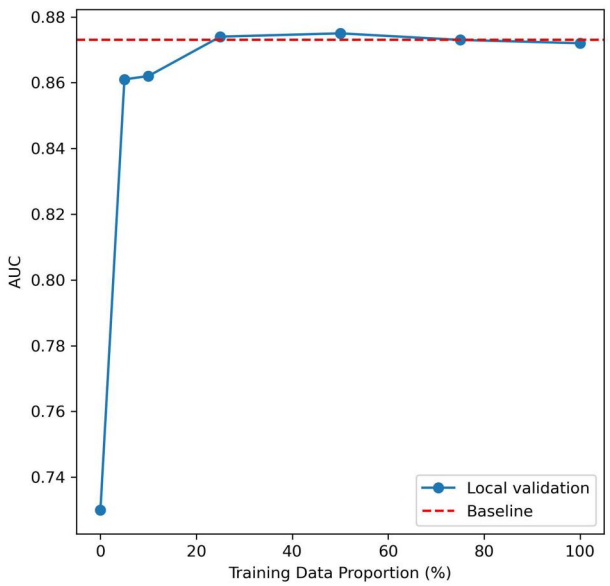

Figure 4. Local validation results using a fine-tuned Vent.io model on varying percentages of MIMIC training data.

guide clinical care and decisions. For VentNet, we previously showed that it outperformed the ROX index[22] in terms of AUC.[13] The retrospective performance of Vent.io exceeds an existing commercial algorithm for respiratory failure, CLEWICU,[23] in terms of sensitivity (60.2% vs 53.7%) and PPV (17.4% vs 4.7%) with the caveat that the models were evaluated on different datasets. Providing early identification allows additional time for clinicians to implement interventions such as diuretic or antimicrobial administration to avoid progression to MV as well as allocate resources appropriately if a patient does require intubation. For example, notifying the patient care team, via the EHR, that a patient is high risk for invasive MV in the next 12-24 hours can provide a trigger to promote patient re-assessment by the treatment team or even rapid response team evaluation. Hospital rapid response systems, including rapid response teams, emphasize the importance of early recognition of patient deterioration, resulting in improvements in patient outcomes such as hospital length or stay[24] and improved patient safety.[25]

Using ML models generated in one health system and applying them to another can lead to challenges such as degradation of model performance due to differences in variables such as EHR data constructs and clinical/administrative practices.[26] Transfer learning allows fine-tuning of a model on a small amount of local, site-specific data making the algorithm adaptable across different healthcare systems.[27–29] In this study, we provide an example of successful application of transfer learning techniques to generalize model performance at an external site. We found Vent.io performed similarly to the UCSD dataset once fine-tuned using only 25% of the MIMIC training data via transfer learning. This approach is practical and allows site-specific tailoring the model using a

small amount of data while maintaining model performance and ensures privacy of institutional patient data. We have previously applied this approach to the early detection of sepsis;[28,29] however, our novel use of this technique for a ML algorithm to predict the need for MV demonstrates the feasibility of using this approach across different disease processes.

Finally, we present one of the first examples of real-time live score generation for a MV prediction algorithm. Prior algorithms have largely used retrospective data or delayed data via data warehouses.[7–11] The integration of the model using FHIR allows real-time score generation, which is an essential step towards integration into clinical practice.[30] We also present our post-implementation monitoring protocol with integration of a PCCP to identify issues with model degradation and ensure consistent model performance. Although we did not need to fine-tune the model at the UCSD Health System, as the performance was acceptable per the PCCP, our real-time pipeline makes use of FHIR to assign gold-standard labels to patients that can be used to fine-tune the model if required. Implementing a real-time score generation process and ensuring continued model accuracy using a silent-mode prospective approach prioritizes patient safety. This successful real-time pipeline will lay the groundwork for future clinical integration and prospective randomized trials.

Despite our study's strengths, we acknowledge several limitations. First, our techniques rely on the electronic health record, which may be susceptible to errors in some cases. For example, we define the need for MV based on the documentation of PEEP and FiO_2 in the EHR. However, there may be delays or misclassification on this basis. We observed an increased prevalence of MV in the MIMIC cohort compared to the UCSD cohort likely due to lack of ventilation measurements for non-ED and non-ICU stays, meaning patients ventilated in non-ICU units would not be excluded if they were ventilated before ICU admission. Second, we have not conducted a randomized controlled trial using our approach. Thus, we cannot say with confidence that our new deep learning algorithm improves clinical outcomes. However, we are now in a strong position to design such studies based on our new findings.

Author contributions

Jonathan Y. Lam, Supreeth P. Shashikumar, Atul Malhotra, and Shamim Nemati were involved in the conception of the work. Supreeth P. Shashikumar, Hayden Pour, and Aaron E. Boussina contributed to acquisition of data and construction of the real-time pipeline. Jonathan Y. Lam, Xiaolei Lu, Supreeth P. Shashikumar, Ye Sel Lee, and Shamim Nemati refined the model, conducted the model experiments, and analyzed the data. Michael Miller, Alex K. Pearce, and Atul Malhotra provided clinical expertise and contributed to interpretation of the results. Jonathan Y. Lam, Alex K. Pearce, and Atul Malhotra wrote the manuscript. All authors contributed feedback and approved the final manuscript.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the National Heart, Lung, and Blood Institute (R01HL157985), the National Institute of Allergy and Infectious Diseases (R42AI177108), and the National Library of Medicine (R01LM013998 and T15LM011271).

Conflicts of interest

S.N., A.E.B., S.S., and A.M. are co-founders of a UCSD start-up, Healcisio Inc., which is focused on commercialization of advanced analytical decision support tools, and formed in compliance with UCSD conflict of interest policies. A.M. reports additional income from Eli Lilly, Livanova, Zoll, and Powell Mansfield. ResMed provided a philanthropic donation to UCSD. The remaining authors declare no competing interests.

Data availability

MIMIC-IV v2.2 data are publicly accessible from PhysioNet.[31] Access to the de-identified UCSD cohort may be made available by contacting the corresponding author and via approval from UCSD Institutional Review Board (IRB) and Health Data Oversight Committee (HDOC).

References

1. Bellani G, Laffey JG, Pham T, et al.; ESICM Trials Group. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*. 2016;315:788-800. https://doi.org/10.1001/jama.2016.0291
2. Beitler JR, Malhotra A, Thompson BT. Ventilator-induced lung injury. *Clin Chest Med*. 2016;37:633-646. https://doi.org/10.1016/j.ccm.2016.07.004
3. 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;150:132-138. https://doi.org/10.7326/0003-4819-150-2-200901200-00011
4. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25:24-29. https://doi.org/10.1038/s41591-018-0316-z
5. Wardi G, Owens R, Josef C, et al. Bringing the promise of artificial intelligence to critical care: what the experience with sepsis analytics can teach us. *Crit Care Med*. 2023;51:985-991. https://doi.org/10.1097/CCM.0000000000005894
6. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. https://doi.org/10.1038/s41746-018-0029-1
7. Yu L, Halalau A, Dalal B, et al. Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. *PLoS One*. 2021;16:e0249285. https://doi.org/10.1371/journal.pone.0249285
8. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27:1735-1743. https://doi.org/10.1038/s41591-021-01506-3
9. Kim Y, Kim H, Choi J, et al. Early prediction of need for invasive mechanical ventilation in the neonatal intensive care unit using artificial intelligence and electronic health records: a clinical study. *BMC Pediatr*. 2023;23:525. https://doi.org/10.1186/s12887-023-04350-1
10. Bendavid I, Statlender L, Shvartser L, et al. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. *Sci Rep*. 2022;12:10573. https://doi.org/10.1038/s41598-022-14758-x

11. Godoy MFD, Chatkin JM, Rodrigues RS, et al. Artificial intelligence to predict the need for mechanical ventilation in cases of severe COVID-19. *Radiol Bras*. 2023;56:81-85. https://doi.org/10.1590/0100-3984.2022.0049
12. Zhang A, Xing L, Zou J, et al. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat Biomed Eng*. 2022;6:1330-1345. https://doi.org/10.1038/s41551-022-00898-y
13. Shashikumar SP, Wardi G, Paul P, et al. Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation. *Chest*. 2021;159:2264-2273. https://doi.org/10.1016/j.chest.2020.12.009
14. FDA. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions. Accessed August 21, 2024. https://www.fda.gov/media/166704/download
15. FDA. Predetermined Change control plans for medical devices. Accessed August 21, 2024. https://www.fda.gov/media/180978/download
16. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:219. https://doi.org/10.1038/s41597-022-01899-x
17. Shashikumar SP, Le JP, Yung N, et al. Development and validation of a deep learning model for prediction of adult physiological deterioration. *Crit Care Explor*. 2024;6:e1151. https://doi.org/10.1097/CCE.0000000000001151
18. Shashikumar SP, Wardi G, Malhotra A, et al. Artificial intelligence sepsis prediction algorithm learns to say "I don't know." *NPJ Digit Med*. 2021;4:134. https://doi.org/10.1038/s41746-021-00504-6
19. Sutton RS. Learning to predict by the methods of temporal differences. *Mach Learn*. 1988;3:9-44. https://doi.org/10.1007/BF00115009
20. Le JP, Shashikumar SP, Malhotra A, et al. Making the improbable possible: generalizing models designed for a syndrome-based, heterogeneous patient landscape. *Crit Care Clin*. 2023;39:751-768. https://doi.org/10.1016/j.ccc.2023.02.003
21. Boussina A, Shashikumar S, Amrollahi F, et al. Development and deployment of a real-time healthcare predictive analytics platform. *Annu Int Conf IEEE Eng Med Biol Soc*. 2023;2023:1-4. https://doi.org/10.1109/EMBC40787.2023.10340351
22. 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;199:1368-1376. https://doi.org/10.1164/rccm.201803-0589OC
23. CLEW. CLEWICU—instructions for use. Accessed June 1, 2024. https://www.fda.gov/media/138372/download
24. Bellomo R, Goldsmith D, Uchino S, et al. Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates*. *Crit Care Med*. 2004;32:916-921. https://doi.org/10.1097/01.CCM.0000119428.02968.9E
25. Iyengar A, Baxter A, Forster AJ. Using medical emergency teams to detect preventable adverse events. *Crit Care*. 2009;13:R126. https://doi.org/10.1186/cc7983
26. Rockenschaub P, Hilbert A, Kossen T, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. *Crit Care Med*. 2024;52:1710-1721. https://doi.org/10.1097/CCM.0000000000006359
27. Youssef A, Pencina M, Thakur A, et al. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29:2686-2687. https://doi.org/10.1038/s41591-023-02540-z
28. Wardi G, Carlile M, Holder A, et al. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann Emerg Med*. 2021;77:395-406. https://doi.org/10.1016/j.annemergmed.2020.11.007
29. Holder AL, Shashikumar SP, Wardi G, et al. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the ICU. *Crit Care Med*. 2021;49:e1196-205-e1205. https://doi.org/10.1097/CCM.0000000000005175
30. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov*. 2020;6:45-47. https://doi.org/10.1136/bmjinnov-2019-000359
31. Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV. Accessed February 1, 2024. https://physionet.org/content/mimiciv/2.2/

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae141
Research and Applications