medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

A comprehensive ML-based Respiratory Monitoring System for Physiological Monitoring & Resource Planning in the ICU

Matthias Hüser[1,2,*], Xinrui Lyu[3,1,2,*], Martin Faltys[4,5,*], Alizée Pace[1,2,6,*], Marine Hoche[1], Stephanie Hyland[7], Hugo Yèche[1,2], Manuel Burger[1,2], Tobias M Merz[8,+], Gunnar Rätsch[1,2,5,9,10,+]

1 Department of Computer Science, ETH Zürich, Zürich, Switzerland;
2 Swiss Institute for Bioinformatics, Lausanne, Switzerland;
3 NEXUS Personalized Health Technologies, ETH Zürich, Zürich, Switzerland;
4 Department of Intensive Care Medicine, University Hospital, University of Bern, Bern, Switzerland;
5 Department of Intensive Care, Austin Hospital, Melbourne, Australia;
6 AI Center, ETH Zürich, Switzerland;
7 Microsoft Research, Cambridge, UK (current address);
8 Cardiovascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand;
9 Medical Informatics Unit, Zürich University Hospital, Zürich, Switzerland;
10 Department of Biology, ETH Zürich, Zürich, Switzerland;
* These authors contributed equally: Matthias Hüser, Xinrui Lyu, Martin Faltys, Alizée Pace;
+ These authors jointly supervised this work: Tobias M. Merz, Gunnar Rätsch; e-mail: tobiasm@adhb.govt.nz, gunnar.raetsch@inf.ethz.ch.

Abstract

Respiratory failure (RF) is a frequent occurrence in critically ill patients and is associated with significant morbidity and mortality as well as resource use. To improve the monitoring and management of RF in intensive care unit (ICU) patients, we used machine learning to develop a monitoring system covering the entire management cycle of RF, from early detection and monitoring, to assessment of readiness for extubation and prediction of extubation failure risk. For patients in the ICU in the study cohort, the system predicts 80% of RF events at a precision of 45% with 65% identified 10h before the onset of an RF event. This significantly improves upon a standard clinical baseline based on the SpO_2/FiO_2 ratio. After a careful analysis of ICU differences, the RF alarm system was externally validated showing similar performance for patients in the external validation cohort. Our system also provides a risk score for extubation failure for patients who are clinically ready to extubate, and we illustrate how such a risk score could be used to extubate patients earlier in certain scenarios. Moreover, we demonstrate that our system, which closely monitors respiratory failure, ventilation need, and extubation readiness for individual patients can also be used for ICU-level ventilator resource planning. In particular, we predict ventilator use 8-16h into the future, corresponding to the next ICU shift, with a mean absolute error of 0.4 ventilators per 10 patients effective ICU capacity.

Introduction

Respiratory failure (RF) is common among patients in intensive care units (ICUs) and is associated with high morbidity and mortality[1]. RF severity is defined by the P/F ratio (PaO_2/FiO_2 ratio) with values below 200 mmHg corresponding to moderate and below 100 mmHg to severe RF. Treating patients with RF involves a sequence of clinical evaluations. This includes identifying RF and the need for mechanical ventilation, tracking lung function improvements, determining the right time to stop mechanical ventilation, and assessing the risk of complications after extubation.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Optimizing clinical decision-making requires continuous monitoring of the patient state and prediction of the future clinical course. ICU physicians base their treatment decisions mostly on intermittent clinical assessments and evaluation of monitored vital signs stored in electronic patient-data management systems

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

(PDMS). In the increasingly complex ICU environment, clinicians are confronted with large amounts of data from a multitude of monitoring systems for numerous patients. The quantity of data increases the risk that clinicians do not readily recognize, interpret, and act upon relevant information, contributing to poorer patient outcomes as well as increased ICU resource expenditure[2]. These large data quantities are ideal for automatic processing by machine learning (ML) algorithms[3,4], which have been used to develop decision support systems for various conditions, such as acute respiratory distress syndrome (ARDS)[5–9], circulatory failure[10], sepsis[11–13], and renal failure[14].

We aim to develop a comprehensive, ML-based Respiratory Monitoring System (RMS) to simplify monitoring, expedite treatment of individual patients with RF, and optimize ICU resource planning. For individual patients, the system predicts the risk of RF and the need for mechanical ventilation, continuously monitors changes and improvements of the respiratory state, and predicts the probability of successful extubation. To facilitate total ICU resource management, we demonstrate how using respiratory state predictions from all individual patients admitted to the ICU enables estimating the future number of patients needing mechanical ventilation.

All models are developed on HiRID-II[15], a new open-source dataset containing more than 55,000 admissions to a tertiary care ICU in Switzerland, which forms an integral part of this work. The models for respiratory and extubation failure are externally validated in the Amsterdam University Medical Center database[16] (UMCdb).

We hypothesize that RMS can predict the relevant respiratory events throughout the treatment process of individual patients accurately and early; both in the development dataset and when validated in externally sourced data. In addition, we aim to show that ICU-level resource requirements for the respiratory treatment of patients can be accurately predicted by integrating the various RMS scores across patients in the ICU.

Results

Preparation of an extended HiRID dataset (HiRID-II)

We present the High time Resolution Intensive care unit Dataset II (HiRID-II), a substantial update to HiRID-I[15], that we aim to make available to the research community on physionet.org[17,18]. This new dataset contains 60% more ICU admissions than its predecessor (**Table 1, Extended Data Fig. 1a**). Additionally, the number of meta-variables increased from 209 to 310 by merging equivalent clinical concepts and including additional respiratory variables (**Extended Data Fig. 1b**). The dataset was k-anonymized with respect to the variables age, weight, height & gender, reducing the number of admissions from 60,503 to 55,858. To further reduce the risk of individual patient identification, admission dates were randomly shifted. To allow the assessment of model generalization to the future, the data set was divided into temporal splits while respecting k-anonymization (**Extended Data Fig. 1c**). To test generalization to other health systems, an external high-resolution evaluation data set was extracted from the Amsterdam UMCdb[16] and harmonized with the HIRID-II dataset (**Extended Data Fig. 1d**). Preliminary analysis of the HiRiD-II data set revealed strong correlations between occurrence of RF and extubation failure with ICU mortality, motivating our proposed respiratory monitoring system (**Extended Data Fig. 2**) and confirming prior results[1].

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

	HiRID-I	HiRID-II
Data set size		
Year of admission	2008-2016	2008-2019
No. of patients	33905	55858
No. of meta-variables	209	310
No. of variables	710	890
No. of patient years	356.73	552.77
Demographics		
Gender	Male: 63.5 %, Female: 36.5 %	Male: 63.1 %, Female: 36.9 %
Age [years]	66.0 [54.0,75.0]	66.0 [54.0,75.0]
Diagnostic group		
Neurologic / Neurologic surgery	17.4 % / 11.3 %	20.1 % / 13.8 %
Cardiovascular / Cardiovascular surgery	13.0 % / 23.7 %	10.9 % / 20.0 %
Gastrointestinal / Gastrointestinal surgery	5.0 % / 5.4 %	4.7 % / 5.6 %
Respiratory / Respiratory surgery	7.5 % / 1.9 %	6.9 % / 1.6 %
Trauma / Trauma surgery	4.5 % / 0.7 %	5.4 % / 1.0 %
Other	9.6 %	9.9 %
Outcomes		
ICU Mortality	6.1%	5.5 %

Table 1: Characteristics of the HiRID-I and HiRID-II datasets. Age is reported as median and interquartile range (IQR). The statistics are computed on the HiRID datasets after k-anonymization.

Development of a continuous monitoring system for respiratory management

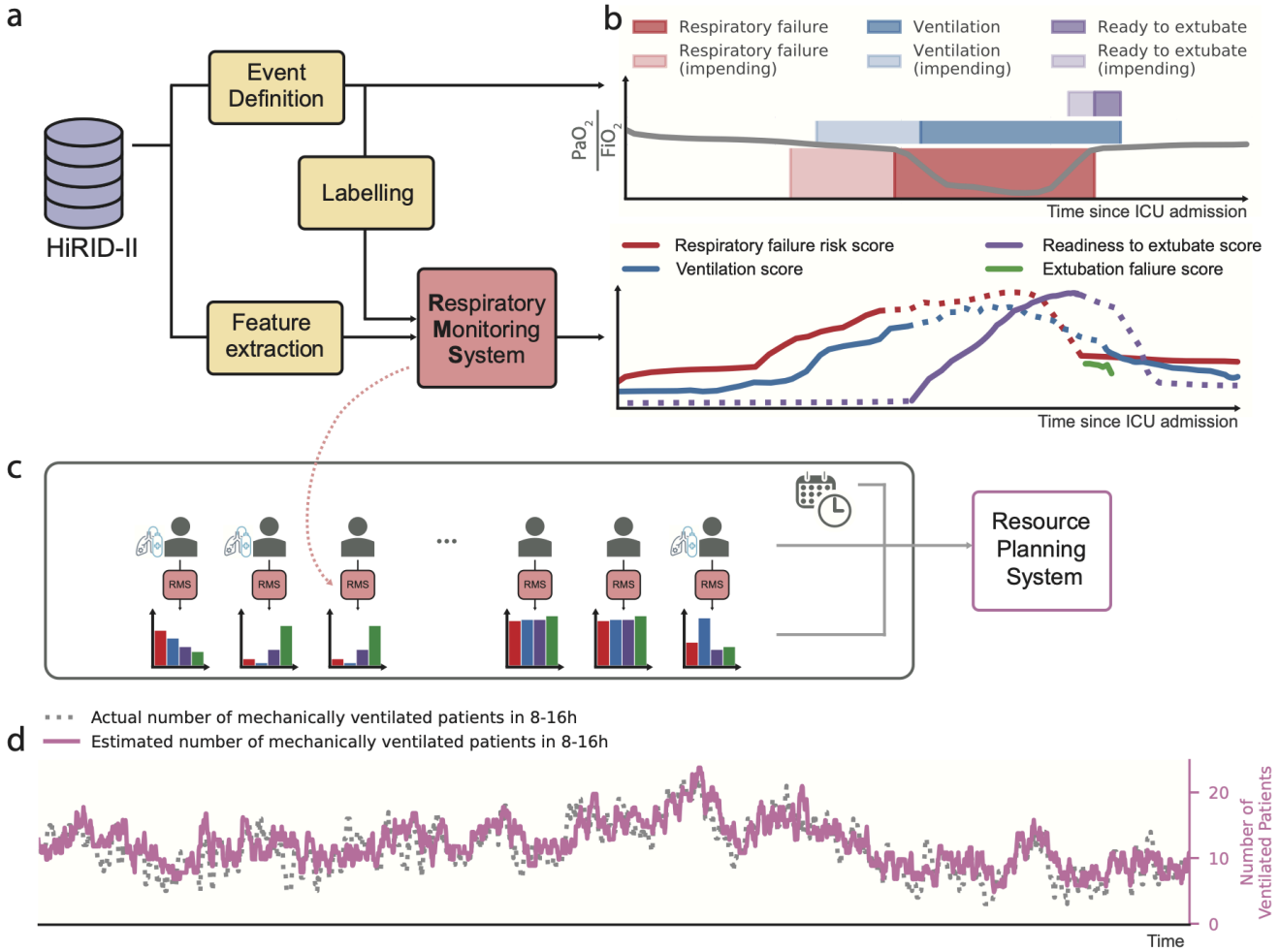
Continuous PaO_2 estimation The partial pressure of oxygen in arterial blood (PaO_2) is one of the main determinants of arterial oxygen content, a parameter that we aim to estimate continuously. The ratio of fraction of oxygen in the inspiratory gas (FiO_2) and PaO_2 (P/F ratio, PaO_2/FiO_2 in mmHg) is commonly used to determine the severity of RF[19]. To measure PaO_2, an arterial blood sample is necessary. Contrary to PaO_2, arterial oxygen saturation (SpO_2) can be continuously monitored in ICU patients using pulse oximetry. The underlying physiological principles governing the binding and release of oxygen to and from hemoglobin create a correlation between SpO_2 and PaO_2[20–22]. This relationship allows for the use of SpO_2 values to infer PaO_2 levels accurately. Firstly, we developed an algorithm to continuously estimate PaO_2 using SpO_2 and other relevant variables determining the hemoglobin-oxygen dissociation curve. This enables us to obtain PaO_2 estimates every five minutes. The algorithm outperforms the non-linear Severinghaus-Ellis baseline[23] for estimating PaO_2 values from non-invasive SpO_2 measurements (**Extended Data Fig. 3**).

Patient State Annotation and Labeling We aim to predict the risk of a patient developing RF within the next 24 hours throughout the ICU stay, with a risk score produced continuously every 5 minutes (**Fig. 1a**). For each time-point it was determined if a patient is currently in (moderate or severe) RF (P/F ratio < 200 mmHg), ventilated, or ready to be extubated. Readiness to extubate status at each time-point was defined using a clinical scoring system (REXT status score), and a score threshold was manually selected after inspection of the time series by an experienced ICU clinician (**Fig. 1b**). Current ventilation status was deduced from the presence of ventilator-specific requirements.

Positive labels for future RF are defined as time-points when the patient is not currently in failure, but RF occurs in the next 24 hours ("impending RF"), while a negative label is assigned if the patient remains stable in the next 24 hours. For every extubation event, we determine whether it failed (reintubation necessary within 48h after extubation) and use it as the label for extubation failure (EF). Labels for ventilation onset and readiness to extubate prediction are positive, if the patient is currently not

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

ventilated/ready-to-extubate, but will be in the next 24 hours (**Fig. 1b**). In HiRID-II, 43.7% and 46.2% of all patients had RF events and required mechanical ventilation, respectively. Moreover, the dataset contains 23,861 extubations of which 11.1% failed. As the original dates were removed during anonymization for HIRID-II, we used an additionally provided dataset with the admission dates of the ICU patients in order to reconstruct the number of patients within the ICU and the ventilator resource use.



e

Predictor	Prediction task	Horizon	Prevalence	# Input variables
RMS-RF	Newly P/F ratio < 200 mmHg	Next 24h	16%	15
RMS-EF	Reintubation	Next 48h	13%	20
RMS-VENT	Newly ventilated	Next 24h	8%	26
RMS-REXT	Newly ready-to-extubate	Next 24h	62%	26
Resource planning	Max # ventilators used in ICU	Next 4-24h	N/A	34

Fig. 1: Overview of the RMS decision support system for Respiratory State Management, and its extension for ICU-level resource planning. **a.** Flow diagram for the development of RMS predictors at the individual patient level. Time series were extracted from the HiRID-II database and gridded to a 5-minute resolution, and features were computed. Respiratory failure/ventilation/ready-to-extubate periods are annotated and machine learning labels created. **b.** The respiratory monitoring system consists of four scores which are active at different parts of the ICU stay, according to the respiratory and ventilation state of the patient **c.** Flow diagram for the development of a resource monitoring system at the ICU level. For all current patients in the ICU, the four scores are integrated to predict the probability that a patient will require mechanical ventilation within a future time horizon. The sum of the individual predictions and a number of ICU-level static features are used to obtain an estimate of the number of patients on mechanical ventilation within a future time window. **d.** Example of 3 months in an ICU, displaying the actual number of ventilated patients and the predicted number as estimated by RMS in the next 8 to 16 hours. **e.** Overview of prediction tasks solved by RMS for individual patients (RMS-RF/RMS-EF/RMS-VENT/RMS-REXT) as well as on the ICU-level. For RF, VENT and REXT we provide the event prevalences in the test set at times when the patient is stable, not ventilated, ventilated, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Development of RMS Predictors The developed RMS consists of four individual scores which are active at different stages of the RF management process. All four models are based on manual feature engineering and LightGBM[24] predictors, similar to what was previously described in Hyland et al.[10] Prior analyses on HIRID-I for circulatory and a related respiratory failure task have shown its superior performance compared to others, including deep learning models[10,25]. The predictor for RF (RMS-RF) uses 15 clinical variables (**Supplemental Table 3)**. As in Hyland et al.[10], the system raises an alarm, if the RF score raises above a certain threshold and is silenced for 4 hours afterwards; the alarm system is reset after the patient just recovered from an event and is able to raise an alarm again 30 minutes after the recovery. The extubation failure (RMS-EF) predictor uses 20 clinical variables (**Supplemental Table 3)**. The RMS-RF & RMS-EF variable sets were identified using greedy forward selection on the validation set of five data splits, separately for the two tasks. The models for the ventilator use (RMS-VENT) and extubation readiness (RMS-REXT) use the union of the parameters of the two main tasks, yielding a total of 26 variables (**Supplemental Table 3**).

We use the four risk scores to estimate mechanical ventilator resource requirements in the short-term future by training a meta-model (**Fig. 1c**). The resource planning problem is divided into two sub–problems; predicting the future ventilator use for already admitted ICU patients, and predicting near future ventilator requirement for newly admitted non-elective patients. We excluded elective patients as their resource use is typically known well in advance. The predictor uses date and time information as well as summary statistics regarding ventilator use and patient numbers from the ICU. A LightGBM[24] regressor is used to solve both sub-problems. For admitted ICU patients, it predicts the necessity for mechanical ventilation in the short-term future, as well as the total number of ventilators required for all admitted patients as an aggregate of the individual predictions (**Fig. 1d**).

Open Source Release All elements of the developed system, including data preprocessing, annotation, prediction task labeling (**Fig. 1e**), and both training and prediction pipelines are made available under an open source license facilitating the reproducibility and reuse of the methodology and results.

RMS-RF predicts RF early with high precision and reduces false alarms compared to clinical baselines

The early prediction of RF is crucial for timely intervention, potentially reducing the severity of patient outcomes and improving overall healthcare efficiency. By accurately forecasting these events, RMS-RF may not only improve clinical decision-making but also allow physicians to commence treatment early, thereby mitigating the risk of more severe respiratory complications. We observe that the developed early alarm system RMS-RF significantly outperforms a decision tree that uses the current value of the four most relevant respiratory parameters (SpO_2, FiO_2, PaO_2, and Positive End-Expiratory Pressure (PEEP)) as well as a clinical threshold-based system based on the SpO_2/FiO_2 ratio (**Fig. 2a**). It achieves an area under the alarm/event precision recall curve[10] (AUPRC) of 0.559 with an alarm precision of 45% at an event recall of 80%. Its underlying risk score has an area under the receiver operating characteristic curve (AUROC) of 0.839 (**Extended Data Fig. 4a**) and is well calibrated, in contrast to the two baselines (**Extended Data Fig. 4b**). The system detects 65% and 78% of events at least 10 hours before they occur when set to an event recall of 80% and 90%, respectively (**Fig. 2b**). Compared to the SpO_2/FiO_2 threshold-based system, our system generates two-thirds fewer false alarms per day on days where the patient experiences no respiratory failure (**Fig. 2c**). We find performance increases with more data up to 25% of the total dataset size (**Extended Data Fig. 4c**). Performance in patients from the cardiovascular and respiratory diagnostic groups is higher than average (alarm precisions 55% and 60% at 80% event recall, respectively). Lower performance is observed in neurologic and trauma patients (**Fig. 2d**). Performance varies in groups determined by age and gender[26] (**Extended Data Fig. 4d/e**). RMS-RF is inspectable to the clinician using SHapley Additive exPlanations (SHAP)[27] values and exhibits physiologically plausible relationships of risk and clinical variables (**Fig. 2e**, **Extended Data Fig. 5**).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

The proposed RMS-RF model only uses a small number of physiological parameters and ventilator settings. We excluded medication variables to reduce the effect of differences in medication policies in different hospitals. When externally validated in the Amsterdam UMCdb database[16], a somewhat reduced performance is observed when the HIRID-II-based model is used and no major performance gains are achieved by retraining using local data (**Fig. 2f;** 38% vs. 45% alarm precision at 80% event recall). A variant of RMS-RF including medication variables (RFS-RF-p) achieved only minor gains in internal HiRID performance (**Fig. 2g**) and exhibited poor transfer performance to UMCdb (**Extended Data Fig. 6a**).To understand these transfer issues, medication policy differences between HiRID-II and UMCdb were analyzed and could be attributed to the medications loop diuretics, heparin and propofol (**Fig. 2g**, **Extended Data Fig. 6b/c**).

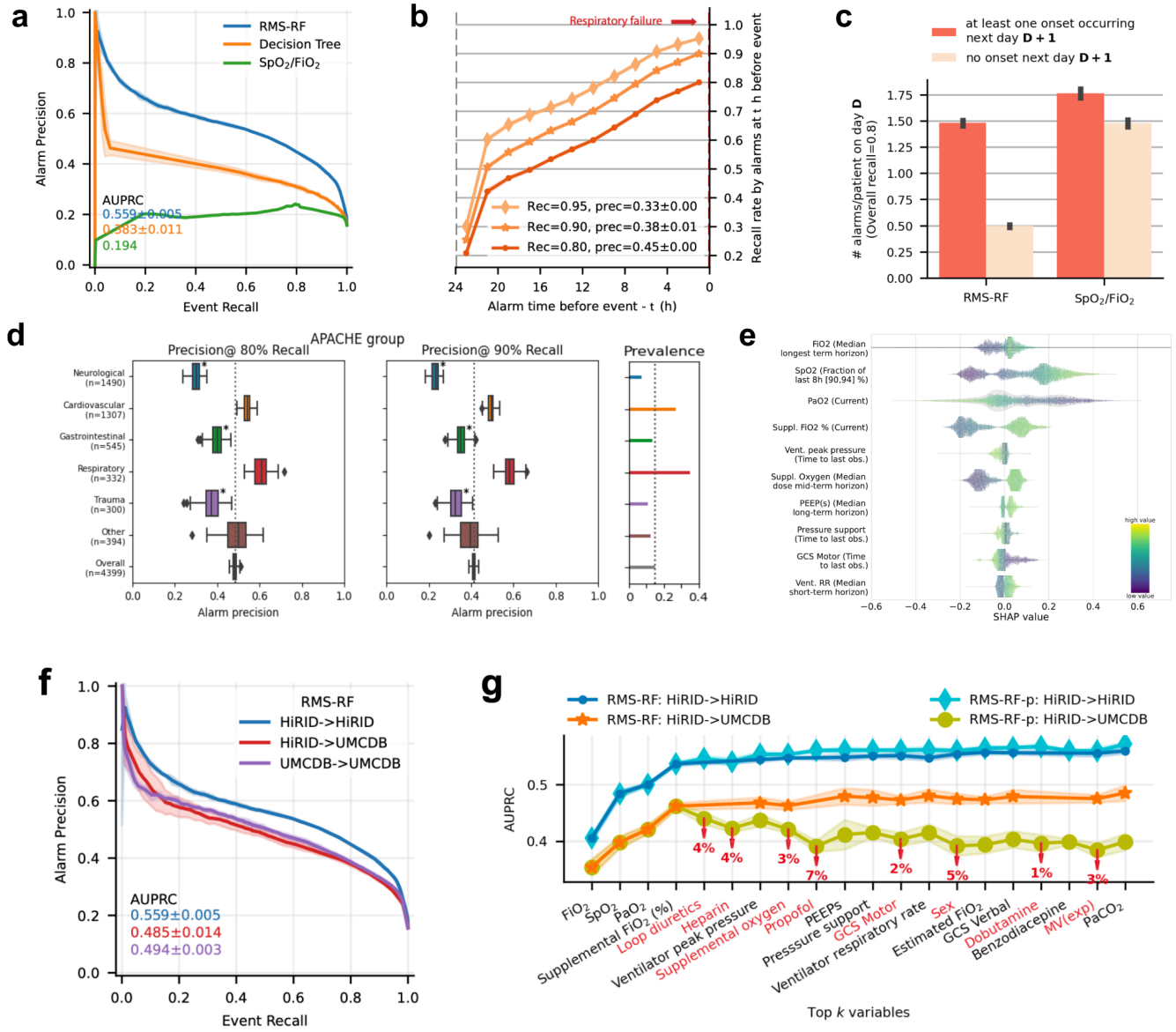
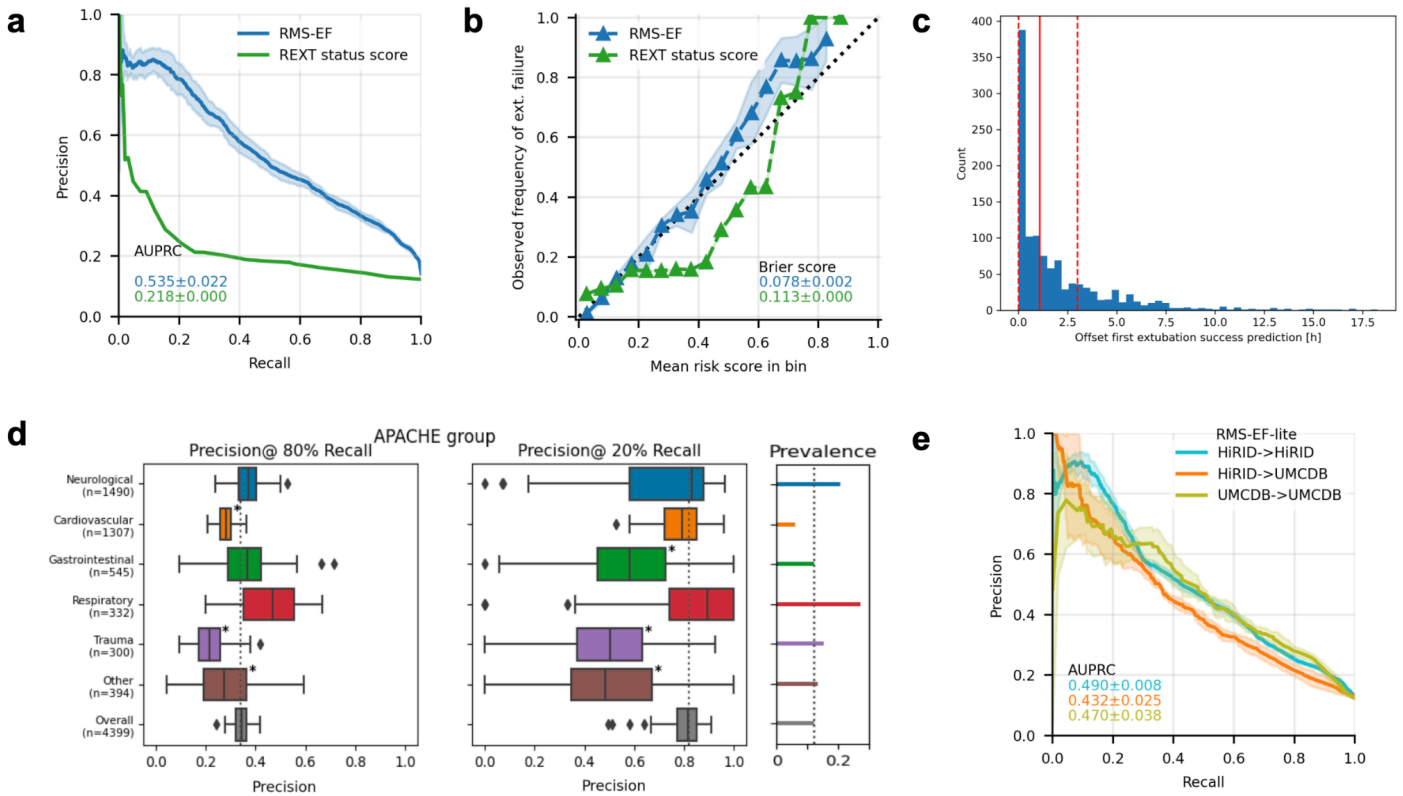


Fig. 2: RMS-RF: Model performance / feature inspection of the respiratory failure prediction model. a. Model performance of RMS-RF compared with a decision-tree based clinical baseline and a threshold-based alarm system based on the current SpO_2/FiO_2 ratio. **b.** The RMS-RF system's performance is evaluated by the earliness of its alarms. Specifically, this is measured as the proportion of events for which it provides early warnings at least a fixed time before a respiratory failure event, considering only those events that have a sufficient stability period beforehand. **c.** Comparison of generated false/true alarm counts of RMS-RF compared with a fixed threshold alarm system, both for patients with events, and patients without events on a given day. **d.** Performance of the RMS-RF model by admission group category, in terms of alarm precision at event recalls of 80% and 90%. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable event recall. **e.** Feature inspection using SHAP values for the most important features for predicting respiratory failure, depicting the relationship between feature values and SHAP values. **f.** External validation of RMS-RF in the Amsterdam UMCdb database[16]. Internal, transfer as well as retrain performance in UMCdb is displayed. **g.** RMS-RF performance changes as the most important variables are added incrementally to the model, for the internal HiRID setting, and the transfer setting to UMCdb. Model transfer issues between the two hospital centers exist if medication variables were included in RMS-RF, denoted as the RMS-RF-p model variant. Markers denote the variables included in the models, and red colors denote variables which decrease performance when added to the model in the transfer setting.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

RMS-EF predicts extubation failure with high precision and is well-calibrated

The accurate prediction of extubation failure is a critical aspect of patient management in intensive care, enabling clinicians to make informed decisions about the ideal timing of extubation. By utilizing RMS-EF to predict the risk of extubation failure, physicians could judiciously determine whether to proceed with or delay extubation based on a quantifiable risk threshold, potentially reducing the likelihood of complications associated with both, premature extubation or unnecessary prolongation of mechanical ventilation. We compare the developed RMS-EF predictor to a threshold-based scoring system, which counts the number of violations of clinically established criteria for readiness to extubate at the time point when the prediction is made (REXT status score). RMS-EF significantly outperforms the baseline (**Fig. 3a**) with an AUPRC of 0.535 and an AUROC of 0.865 (**Extended Data Fig. 7a**). We also analyzed calibration and observed high concordance between observed risk of extubation failure and RMS-EF with a Brier score of 0.078, in contrast to the baseline (**Fig. 3b**). The precision for predicting EF is 80% at a recall of 20% indicating that RMS-EF can confidently identify the highest risk patients. For 25% of correctly predicted successful extubations, RMS-EF would predict success at least 3h prior to the time point when extubation effectively takes place (**Fig. 3c**). As with RMS-RF, no major improvements are observed when using more than 25% of the training data (**Extended Data Fig. 7b**). Performance in sub-cohorts according to the diagnostic group is similar, with RMS-EF performing best in respiratory patients (**Fig. 3d**). We observe that the performance in female patients and older age groups is slightly inferior (**Extended Data Fig. 7c/d**). As RMS-EF is based almost exclusively on variables that are influenced by clinical policies which likely differ in different hospitals, it transfers poorly to the UMCdb database[16] (**External Data Fig. 7e**). However, a variant of RMS-EF can be constructed without medication variables, which transfers better to the UMCdb database with only slightly reduced internal performance (**Fig. 3e**; AUPRC 53.5% vs. 49% for HIRID). Accordingly, the analysis of medication policies revealed major differences for ready-to-extubate patients between HiRID-II and UMCdb (**Extended Data Fig. 7f/g**). SHAP value analysis[28] shows that the RMS-EF risk score is dependent on several parameters determined by treatment-policies, such as medications and ventilator settings (**Fig. 3f**, **Extended Data Fig. 8**). Severe loss of transfer performance resulted from the inclusion of sedatives and vasopressors in the model (**Fig. 3g**).



medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

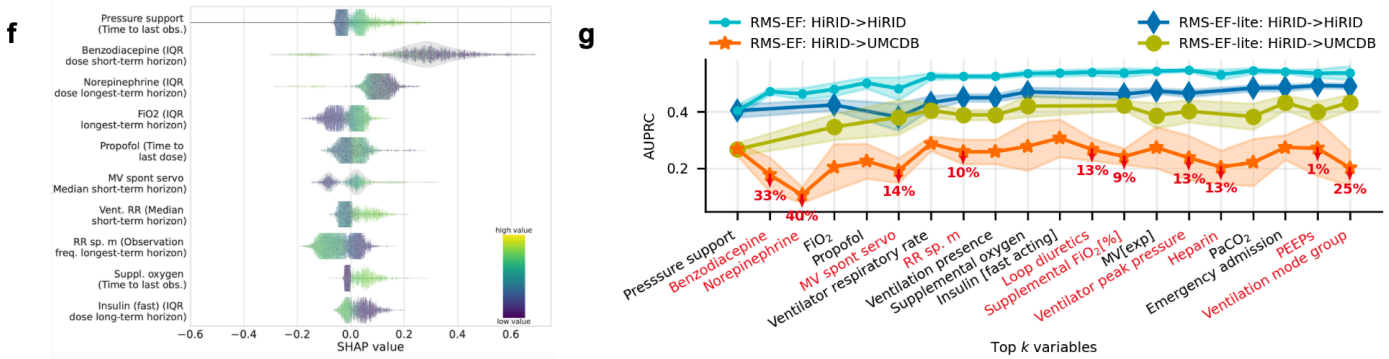


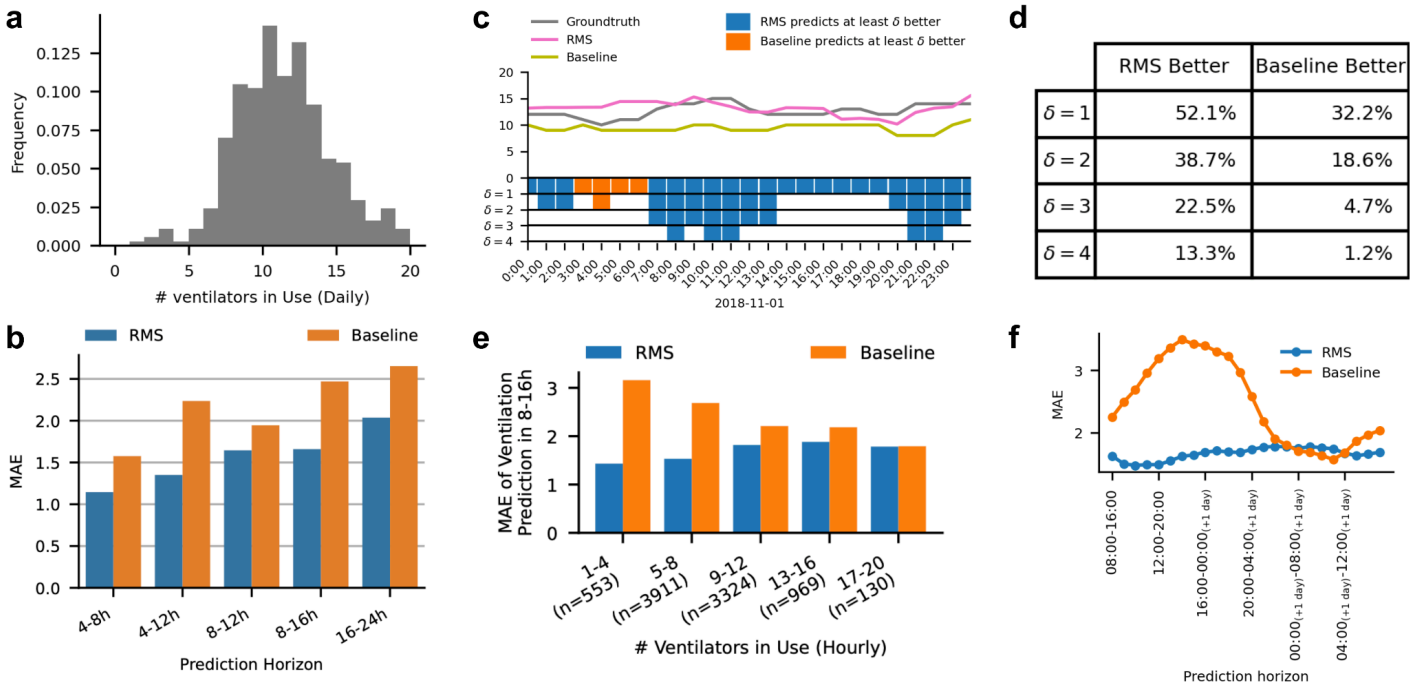
Fig. 3: RMS-EF: Model performance analysis and feature inspection. a. Model performance compared with a baseline based on clinically established criteria for readiness to extubate in terms of recall/precision. **b.** Risk calibration of the score for predicting extubation failure at the time of extubation, compared with the baseline. **c.** Distribution of time span between the earliest extubation success prediction of RMS-EF prior to the time point of successful extubation, for correctly predicted successful extubations. The earliest time is defined as the first time-point from which RMS-EF continuously predicts 'extubation success' while the patient is ready-to-extubate. Red dashed lines denote the 25, 75 percentiles, and the red solid line denotes the median, respectively. **d.** Performance in different sub-cohorts of the test set, according to diagnostic group, in terms of precision at 80% and 20% recall. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable recall. **e.** Performance of the RMS-EF-lite model, which is obtained by excluding medication variables from RMS-EF, when trained/tested on the HIRID-II database, transferred to the UMCdb data-base, and retrained in the UMCdb database. **f.** Summary of SHAP value vs. variable distribution for the most important feature of each of the top 10 important variables contained in the RMS-EF model. **g.** Performance of the RMS-EF, RMS-EF-lite models in the internal and transfer settings as variables are added incrementally to the model ordered by performance contribution (greedy forward selection performance on the validation set). Red marked percentages on the orange curve denote relative performance loss in the transfer, when adding the variable to the model. Variables are in red font if their inclusions leads to performance loss in the transfer setting.

Integrating RMS scores of individual patients for ICU-level resource planning

Using the predictions for the four models focusing on respiratory failure (RMS-RF), extubation failure (RMS-EF), ventilation onset (RMS-VENT), and readiness to extubate (RMS-REXT), we develop a combined model predicting the number of ventilators in use at a specific future horizon. Preliminary analysis of the HiRID-II dataset shows substantial variation in demand for ventilators each day, underscoring the need for a model to aid resource planning (**Fig. 4a**). In a first step, we evaluated ventilation onset (RMS-VENT) and readiness to extubate (RMS-REXT) prediction 24h prior to the event on a patient-level. We observe a high discriminative performance with AUROCs of 0.914 and 0.809 (**Extended Data Fig. 9a/b**), event-based AUPRCs of 0.528 and 0.910 (**Extended Data Fig. 9c/d**), respectively, and the models are well calibrated (**Extended Data Fig. 9e/f**).

We then train a meta-model using the four scores to predict ventilator usage in the ICU at future time horizons every hour (4-8h, 4-12h, 8-12h, 8-16h, 16-24h; **Fig. 4b**). We compare it with a baseline that predicts that the future ventilator resource remains unchanged. We observe that the proposed model clearly outperforms this baseline in terms of mean absolute error (MAE), with the largest relative gain in longer prediction horizons (**Fig. 4b**). We observe that in 39% of time-points the model's predictions are at least two ventilators closer to ground-truth, for predicting ventilator use in 8-16 hours into the future (**Fig. 4c/d**). RMS outperforms the baseline for the vast majority of ICU ventilator utilization scenarios (**Fig. 4e**) with the largest improvement over the baseline when the respirator use is below the maximum capacity (Fig. 4e) and for predictions of ventilator use during day hours (**Fig. 4f**).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



	RMS Better	Baseline Better
$\delta = 1$	52.1%	32.2%
$\delta = 2$	38.7%	18.6%
$\delta = 3$	22.5%	4.7%
$\delta = 4$	13.3%	1.2%

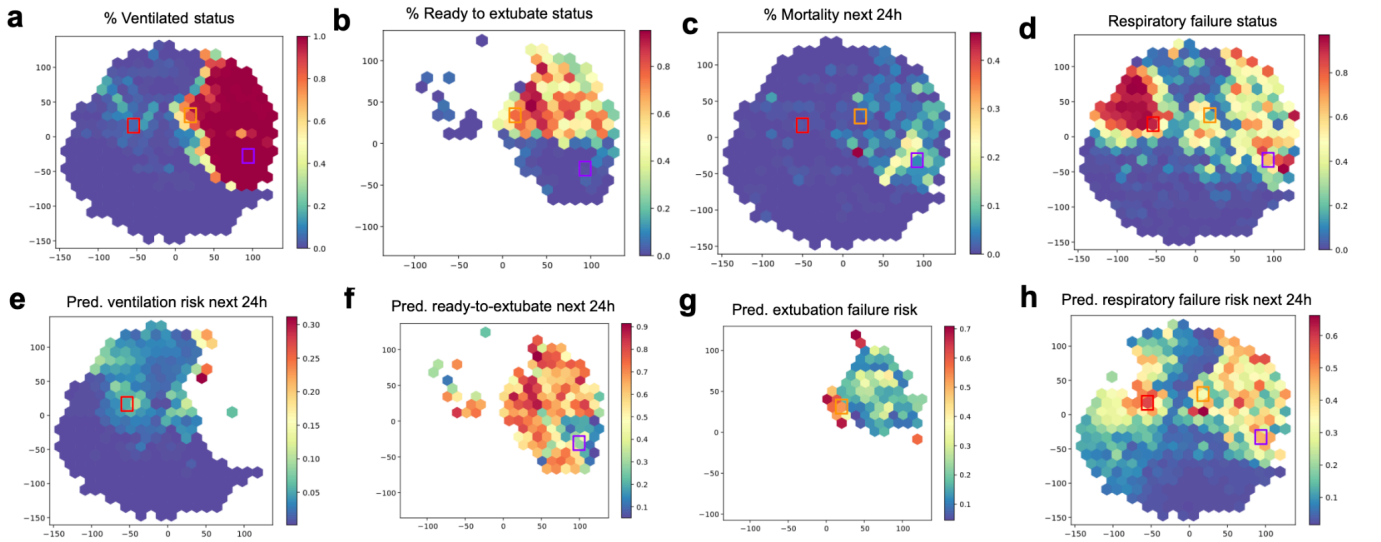
Fig. 4: Model performance of the integrated system for resource planning (RMS) in an ICU with an effective capacity of 42 beds. **a.** Observed ventilator usage pattern in the HiRID-II dataset, in terms of days on which a particular number of ventilators is used. **b.** Performance of RMS compared with the baseline, in terms of mean absolute error for predicting the maximum number of used ventilators at a fixed horizon in the future. **c**. Example of a typical ICU setting shown for a duration of one day, annotated with ground-truth, RMS and baseline predictions. In the rug plot, relevant better predictions are marked for the offsets 1-4. **d.** RMS compared with the baseline, for different absolute differences of predictions (1-4) in the rows, for a prediction horizon of 8-16 hours in the future. The table entries denote the proportion of time-points in which either RMS or the baseline is better by at least the difference of the row. **e.** Performance of RMS for different current ventilator ICU usage scenarios, ranging from low usage, to high usage, compared with the baseline, for a prediction horizon of 8-16 hours in the future. The number of time-points falling into each bin is denoted in parentheses. **f.** Performance of RMS, by hour of the day when the prediction is performed, compared with the baseline, for a prediction horizon of 8-16 hours in the future.

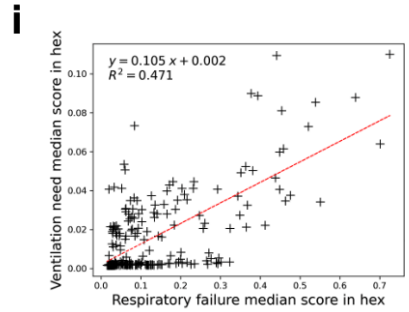
Explorative joint analysis of RMS scores throughout the ICU stay

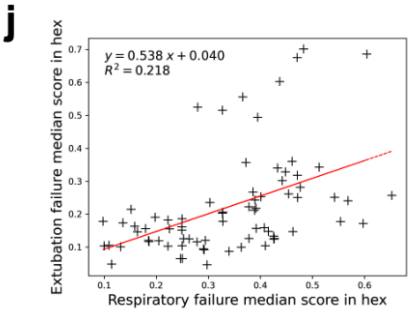
We analyzed the relationship of the four RMS scores produced at each time point of the ICU stay by embedding the most important parameters for respiratory failure and extubation failure prediction (union of the top 10 variables identified for each task, current value feature) using t-distributed Stochastic Neighbour Embedding (t-SNE[29]) with subsequent discretization into hexes. This approach produces a two-dimensional hex-map that defines subsets of comparable patient states that can be compared across different characteristics, i.e., between the panels for the hex. We observe that the space is divided into two distinct states, corresponding to time-points when the patient is ventilated or not ventilated (**Fig. 5a**). The region of ventilated patients is further subdivided, with patients in the upper part being more likely to be ready-to-extubate (**Fig. 5b**). As expected, the ventilated and not ready-to-extubate region has the highest observed 24h mortality (**Fig. 5c**). Patients currently experiencing respiratory failure are concentrated in a compact region in the non-ventilated space, as well as scattered throughout the ventilated space (**Fig. 5d**). States with high risk of future ventilation need according to RMS-VENT are close to the boundary of the ventilated region (**Fig. 5e**). Readiness to extubate scores show a less clear pattern, but scores tend to be higher in the upper part of the ventilated region, which is also enriched in states in which patients are ready-to-extubate (**Fig. 5f**). For RMS-EF, high scores are concentrated in two distinct regions at the edge of the ventilated region (**Fig. 5g**). Lastly, RMS-RF scores are high close to the boundary of patients already in respiratory failure (**Fig. 5h**). The median risk scores of hexes for respiratory failure/ventilation need are strongly positively correlated with an R^2 of 0.471 (**Fig. 5i**). Likewise, respiratory failure and extubation failure scores are moderately positively correlated (**Fig. 5j**). For RMS-EF/RMS-REXT scores, no correlation could be observed (**Extended Data Fig. 10**). For three exemplary hexes with predominantly (1) non-ventilated patients but high RMS-RF score, (2) ready-to-extubate patients but high RMS-EF score, and

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

(3) not-ready-to-extubate patients but high RMS-RF score, the distribution of clinical parameters was analyzed, showing plausible relationships with clinical parameters (**Fig. 5k**).







k

Variable	RF (high risk, non-vent)	EF (high risk, ready-exp)	RF (high risk, not-ready-exp)
RMS-RF score	0.59 [0.48,0.68]	0.34 [0.16,0.58]	0.45 [0.30,0.56]
RMS-EF score	Not active	0.51 [0.30,0.68]	Not active
RMS-VENT score	0.09 [0.04,0.17]	Not active	Not active
RMS-REXT score	Not active	Not active	0.56 [0.43,0.70]
SpO_2 [%]	96 [93,98]	97 [95,98]	96 [95,99]
PaO_2[mmHg]	81 [74,89]	91 [79,96]	87 [80,97]
FiO_2 [%]	40 [34,50]	34 [30,35]	35 [30,40]
GCS motor score	6 [6,6]	6 [5,6]	1 [1,1]
Pressure support	0 [0,0]	10 [10,12]	0 [0,0]
MV spont servo [L]	5 [5,5]	5 [5,5]	5 [2,5]
Vent. RR [1/min]	28 [24,31]	15 [13,20]	15 [13,19]
RRsp m. [1/min]	25 [21,27]	15 [13,20]	0 [0,4]
Supplemental oxygen [L/min]	4 [4,4]	0 [0,2]	0 [0,0]
Supplemental FiO_2 [%]	45 [45,45]	45 [34,45]	0 [0,0]
Vent. peak pressure [cmH_2O]	0 [0,0]	17 [15,18]	30 [27,34]
PEEPs [cmH_2O]	0 [0,0]	5 [5,5]	8 [5,10]
Benzodiacepine	0 [0,0]	0 [0,0]	0 [0,0]
Norepinephrine	0 [0,0]	0 [0,0]	0 [0,0]
Propofol	0 [0,0]	0 [0,0]	1.7 [1.1,3.3]
Insulin fast acting	0 [0,0]	0 [0,0.03]	0 [0,0]

Fig. 5: Joint analysis of the four scores produced by RMS overlaid on a t-SNE embedding based on important respiratory parameters. **a.** Hexes are colored by the proportion of time-points in the hex for which the patient is ventilated. **b.** Hexes are colored by the proportion of time-points in the hex for which the patient is ready-to-extubate given the patient is ventilated. **c.** Hexes are colored by observed 24h mortality risk. **d.** Hexes are colored by the proportion of time-points in the hex for which the patient is in respiratory failure. **e-h.** The color of the hex denotes the median RMS-VENT/RMS-REXT/RMS-EF/RMS-RF scores of the time points assigned to the hex, respectively. **i.** Relationship of median respiratory failure score (RMS-RF) and median ventilation need score (RMS-VENT) in hexes for time-points where both scores are active. The p-value of a Wald test for a non-zero regression line slope is $1.5 \cdot 10^{-36}$. **j.** Relationship of median respiratory failure score (RMS-RF) and median extubation failure score (RMS-EF) in hexes for time-points where both scores are active. The p-value of a Wald test for a non-zero regression line slope is $2.7 \cdot 10^{-5}$. **k.** Score and input value distribution of time points assigned to three selected hexes for the 16 variables used as input for the t-SNE. The median is reported, and numbers in square brackets refer to the interquartile range.

Discussion

We present a ML-based system for the *comprehensive monitoring of the respiratory state* of ICU patients. The respiratory monitoring system (RMS) consists of four highly accurate scoring models that predict the occurrence of respiratory failure, start of mechanical ventilation, readiness to extubate as well as extubation

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

failure. By combining the prediction scores of all admitted patients at any time point and by accounting for the likelihood of future admissions, RMS facilitates the accurate prediction of the near future cumulative number of patients requiring mechanical ventilation to help optimize resource allocation at the ICU level.

In conjunction with our study, we aim to release the *extensive HiRID-II dataset*, a rich resource for broad-scale analyses of ICU patient data. This dataset represents a significant advancement to HiRID-I, both in terms of number of included patients and clinical parameters. Our initial analysis of the HiRID-II dataset identified significant links: both the presence and duration of respiratory failure, as well as extubation failure, are associated with increased ICU mortality, highlighting distinct yet interconnected risk factors. These insights highlight the critical need for advanced alarm systems in clinical settings to reduce the risks associated with respiratory and extubation failure. The future availability of the HiRID-II dataset to the research community on Physionet[17,18] will open up numerous possibilities for further research, allowing for more in-depth investigations into various aspects of ICU patient care and outcomes.

RMS-RF predicts respiratory failure throughout the ICU stay, and alarms for impending failure are typically triggered at least 10 hours before the event. This early warning is sufficient to enable adjustments in the patient's medical management well in advance of the potential respiratory failure. It outperforms a baseline representing standard clinical decision-making based on SpO_2 and FiO_2, and reduces the number of false alarms by a factor of 3 at 80% event recall (Fig. 2c). RMS produces RF-specific alarms and silences them within a specified period of time after the model triggers an alarm, reducing alarm fatigue, which is a major issue for ventilator alarms[30]. Prior to respiratory failure, only 1.5 alarms per patient/day are raised, which is manageable for the clinical personnel, and unlikely to cause alarm fatigue. Reassuringly, only variables directly associated with respiratory physiology or ventilator settings were found consistently predictive of impending respiratory failure. RMS-RF demonstrates its highest precision in individuals admitted with cardiovascular or respiratory admission diagnoses, while its performance notably declines in neurologic patients. In these patients ventilatory management is often determined by the need to protect a compromised airway in patients with altered levels of consciousness and not by the presence of RF per se. A similar pattern was previously observed for circulatory failure[10] and suggests that patients in the neurologic category deserve additional attention and may need to be excluded in a clinical implementation of an early alarm system based on RMS-RF. To date, few externally validated ML models to continuously predict acute respiratory failure in the ICU have been reported. Recent works by Le et al.[8], Zeiberg et al.[31], and Singhal et al.[32] focus on mild respiratory failure (P/F index < 300 mmHg). Other models predict respiratory failure at the time of ICU admission or are only valid for specific cohorts[33–35].

RMS-EF predicts extubation failure and significantly outperforms a clinical baseline based on common clinical criteria for readiness to extubate status. The model is well calibrated, with almost ideal concordance of the prediction score and observed risk of extubation failure. A potential use case would be to assess the predicted EF risk when considering extubation for patients that are ready to extubate in order to decide whether to accelerate or delay the extubation of the patient. For instance, if the risk is very low, one may speed up extubation of patients that are ready to extubate. At about 80% recall, a quarter of correctly predicted extubation successes are recommended more than 3h before the actual extubation. This suggests that our model could help clinicians to extubate patients earlier. However, in our analysis we could not ascertain whether a patient was not extubated for another reason not apparent from the data, such as availability staff. For clinical use the model could also be operated at 20% recall with very high precision (80%), to identify patients with a high likelihood that extubation will be unsuccessful. This could guide attention towards critical patients, and may caution clinicians from prematurely extubating patients. For the prediction of extubation failure, various models have been proposed[36–41]. The largest cohorts to date were used in the works by Zhao et al.[41], who only validated the model in a cardiac ICU cohort, which limits the generalizability of the results, and Chen et al.[42], who restricted the evaluation to ROC-based metrics only, which makes clinical interpretation difficult.

Machine learning (ML) has previously been used to develop support systems for the management of RF patients in the ICU. These include models for recognition of acute respiratory distress syndrome (ARDS)[5–9] and COVID-19, pneumonitis patients[32,43], prediction of readiness-to-extubate[44–46], need for mechanical ventilation[47,48], and detection of patient-ventilator asynchrony[49]. Existing work focuses on single aspects of

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

RF management, often in specific patient cohorts only. Our approach aims to comprehensively monitor the respiratory state throughout the RF treatment process, by integrating relevant respiratory-system related tasks and allowing for joint analysis of risk scores and trajectories. We believe a single and universally applicable system is much more likely to be successfully implemented than multiple fragmented models pertaining to specific disease entities. A further distinguishing feature of RMS is the five-minute time resolution at which predictions are made, enabling longitudinal analysis of risk trajectories. This dynamic prediction paradigm is more flexible than traditional severity scores, which are evaluated at fixed time-points, such as at 24 h after ICU admission[50], mainly to predict ICU mortality[51].

For successful *external validation* of RMS-RF, it was key to exclude medication variables from the model, as their inclusion was detrimental to model transferability. We hypothesize that this difficulty is caused by the observed medication policy differences between the centers. Interestingly, ventilator settings, while also policy-dependent, do not appear to compromise transfer performance in the same way. Investigating and quantifying the underlying policy differences, which make transfer difficult, needs additional research. Model transferability is an emergent topic in robust machine learning for ICU settings, and recent works study it for sepsis[13,52,53] or mortality prediction[54]. Our results suggest that medication variables require special attention to enable transfer. In contrast to RMS-RF, we suggest that RMS-EF to be re-trained and fine-tuned using the data from the center where it should be applied. As extubation failure predictions are necessarily tied to policy, the policy differences between different centers proved more relevant than in the case of RMS-RF.

While clinical prediction models for individual patients have been extensively studied, *resource planning* in the ICU has received little attention in the ML literature, but came into renewed focus due to the COVID-19 crisis[55]. During the COVID-19 pandemic, the first ML-based models to predict ICU occupation were proposed, such as by Lorenzen et al.[55], who predict daily ventilator use up to 15 days into the future, as well as more generally hospitalization, using patient-specific features[56]. The proposed RMS clearly outperforms a baseline method for predicting future ventilator use at the ICU level. With a mean absolute error of 0.39 ventilators per 10 ICU patient beds used during the next shift (8 to 16 hours), it is sufficiently precise for practical purposes. Since resource allocation in the ICU depends on local policies and procedures, such a system likely needs to be retrained for every clinical facility for reliable predictions.

In this study, we developed comprehensive predictors of key aspects of respiratory failure management, including RF, EF, ventilation need, and readiness to extubate. These predictors collectively describe various aspects of a patient's respiratory health state in the ICU which can be used for exploratory analysis. The joint analysis and visualization of risk scores alongside other vital clinical variables yield discernible clusters that correspond to specific patient states, indicating the potential for risk stratification within the patient population. We observed a separation of patient states into two main clusters that align with ventilated and non-ventilated states, with substructures within these clusters, in particular the patients that are not ready to be extubated among the ventilated patients (Fig. 5b). A subset of patients with low predicted readiness to extubate within the next 24h have the highest mortality risk (Fig. 5f,c). These patients often have a low GCS, are more likely to require controlled ventilation modes, have higher ventilator peak pressure and require higher PEEP (Fig. 5k); all indicators of more severe underlying lung pathology. We can also identify a distinct cluster of patients that are clinically ready to be extubated, have a lower risk of RF but have a very high EF risk (Fig. 5g). These patients require relatively higher ventilation pressure support and have a low respiratory rate (Fig. 5k); again established risk factors for extubation failure. Among the patients that are not currently ventilated, we find a wide range of risks for RF. Those patients with the highest RF risks have low PaO_2, high (supplemental) FiO_2 and high respiratory rates (Fig. 5k). However, mortality risks are relatively low (Fig. 5c).

The hex-map visualization provides a snapshot of the ICU population at any given time and allows for the monitoring of patient states over time with updates, akin to those seen in methodologies like T-DPSOM[57,58]. This dynamic tracking is based on the automated integration of multiple respiratory state dimensions and uses nonlinear dimensionality reduction to provide the position of an individual patient on the map of respiratory health states. We hypothesize that this visualization could assist clinicians in identifying shifts in patient states, although the practical implications of this feature require further validation. This represents a different approach to previous works, that mainly tries to understand biological phenotypes of ARDS

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

patients[59–62] or longitudinal sub-phenotypes of a more specific patient set, like COVID-19 patients[63,64]. Overall, while the visualization provides an interesting perspective for an alternative modality for monitoring of respiratory state in the ICU and can serve as a tool for a more detailed exploration, the presented analysis is primarily exploratory. Further research is needed to substantiate the clinical relevance of the identified clusters and to explore how this system might integrate into the decision-making processes within the ICU.

Our study, while robust, has certain *limitations*. Unlike typical single-center studies, our research utilized data from two distinct centers, one for development and another for validation. This approach reduces the risk of overfitting models to a local patient cohort, although it is important to note that external applicability may still vary and retraining on local data will still be needed for parts of the proposed RMS. In our machine learning models, we have incorporated improvements based on our previous work. Unlike earlier systems heavily reliant on sporadic clinical measurements like lactate levels[10], our current model leverages continuous SpO_2 monitoring and ventilator data. This reduces the influence of clinician-driven decisions on our alarm systems, ensuring a more objective assessment of the patient's condition. However, a limitation remains in the retrospective nature of our data collection. Missing data was partially imputed for respiratory failure annotation, and while this aids in model development, it introduces potential biases. Additionally, our study could not evaluate the impact of system implementation in actual clinical practice, which might alter treatment or monitoring strategies (domain shift)[65]. Lastly, our assessment of the extubation failure (EF) risk score was limited to scenarios of actual extubation events. While we hypothesize that the accuracy of this score would be similar in patients nearing readiness for extubation, this cannot be definitively concluded from our retrospective data. Future prospective studies are needed to fully understand the implications of our model in a live clinical setting.

Overall, we have proposed a comprehensive monitoring system for the *entire respiratory failure management cycle*, including resource planning at the ICU level. We hypothesize that our system can facilitate early reassessment of deteriorating patients, enable rapid treatment and improve their outcomes. However, this has to be validated in prospective randomized controlled trials, assessing the impact of using RMS-RF/RMS-EF on patient outcomes. Using gradient-boosted decision trees for constructing RMS allows for the introspection on individual predictions using SHAP values, offering valuable insights to clinicians, and ultimately increasing trust in the predictions[66]. Resource planning at the ICU level, which has not only become an important topic in the context of the COVID-19 pandemic[55], is facilitated by a meta-model, built on top of RMS. Testing such an approach for resource planning and contrasting it with current clinical practice also lies in the scope of future clinical studies.

Acknowledgments

This project was supported by the Grant No. 205321_176005 of the Swiss National Science Foundation (to T.M.M. and G.R.), and grant #2022-278 of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institutes of Technology), and ETH core funding (to G.R.). We acknowledge discussions with and organizational, administrative or technical help by David Berger, Carmen Pfortmüller, Jörg Schefold, Daniel Vonder Mühll, Olga Mineeva, Quinten Johnson, Dinara Veshchezerova, David Meyer, Anastasia Escher, Nora Toussaint, Margarita Kuznetsova, Fedor Sergeev, Marc Zimmermann, Catherine Jutzeler, Karsten Borgwardt, Thomas Gumbsch, Bowen Fan, Jörg Goldhahn, Sonia Strangio, Ivo Schauwecker, Martina Baumann, Sergio Maffioletti, Bernd Rinn, Anna Wiegand, Diana Coman Schmidt, Matthew Levin, Robert Freeman, Thomas Fuchs, Emanuela Keller, Michael Krauthammer, Paul Elbers, and Patrick Thoral. Computational analyses were performed at the LeonhardMed Trusted Research Environment at ETH Zurich (https://sis.id.ethz.ch/services/sensitiveresearchdata/). The work by S.L.H. was done while she was working at ETH Zurich.

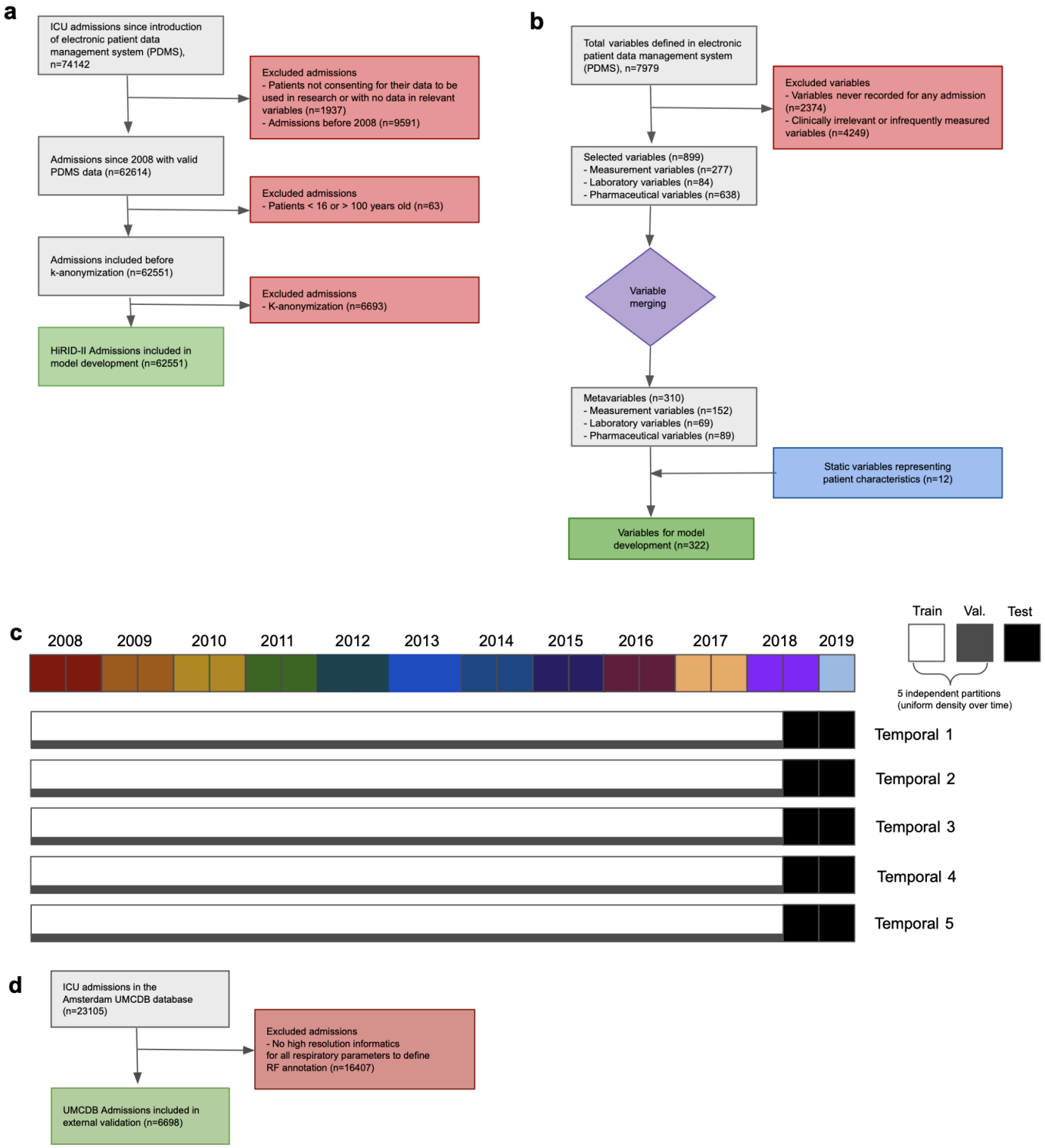
medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Author contributions

M.H., X.L., M.F., G.R., T.M.M. with input from S.L.H., M.Ho., A.P., H.Y., M.B. designed the experiments; M.F., T.M.M. selected and provided the clinical data and context; A.P. with input from M.F., G.R., T.M.M., X.L. k-anonymized the data set. X.L., M.F., A.P. with contributions from T.M.M., G.R. preprocessed and cleaned the HiRID-II data; X.L., M.F. harmonized the UMCdb data set with HiRID-II. M.H., M.F. with input from G.R., T.M.M, X.L., S.L.H. defined and developed the respiratory state annotations and labels; M.H., M.F. developed the continuous estimation algorithm for PaO2;. M.H., M.F. developed and extracted ML features; M.H. developed the pipeline for supervised learning including variable selection; M.H. with input from X.L., G.R., M.F., M.H. performed the fairness analysis in sub-cohorts. A.P., with input from M.H., M.F., T.M.M., G.R., performed analyses of treatment policy differences. X.L. with input from G.R., M.H., M.F., A.P. conceived and developed the model for resource planning, M.H. with input from G.R., M.F., X.L., T.M.M. implemented the joint analysis of RMS scores; T.M.M., G.R., M.F. conceived and directed the project; M.H., M.F., X.L., G.R., T.M.M, A.P., M.Ho. with input from H.Y., M.B. wrote the manuscript. X.L. with input from all authors created Fig. 1. All authors read the manuscript and provided critical feedback.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Extended data figures



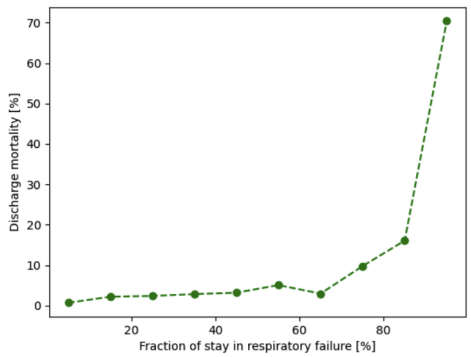
Extended Data Fig 1. Patient inclusion & Experimental design. a. Patient inclusion schema in the HiRID-II dataset. **b.** Inclusion of clinical parameters in the data extraction pipeline of the HIRID-II dataset. **c.** Split design schema for performance evaluation. A fixed test set consisting of admissions starting in Mid June 2018 to the end of 2019 was used, which is shared by all five temporal splits, and is marked by a black block in all five splits. The remaining patients were randomly partitioned five times into train and validation set, each defining a temporal split, which is indicated by the horizontal white and grey bars. **d.** Patient inclusion schema in the UMCdb dataset used for external validation.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

a

	Respiratory failure	No respiratory failure
Alive at discharge	22134 (90.7 %)	35029 (97.0 %)
Dead at discharge	2264 (9.3 %)	1077 (3.0 %)

b



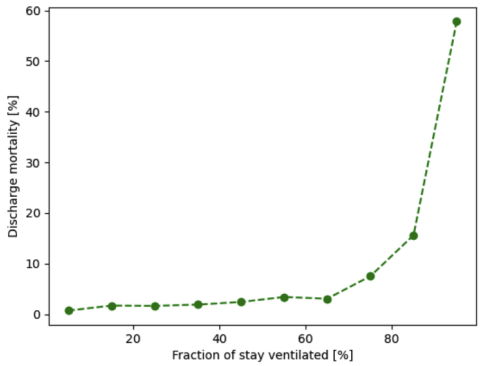
c

	Extubation failure	No extubation failure
Alive at discharge	2419 (92.1 %)	18716 (99.6 %)
Dead at discharge	208 (7.9 %)	67 (0.4 %)

d

	Mechanically ventilated	Not mechanically ventilated
Alive at discharge	22898 (88.8 %)	34265 (98.7 %)
Dead at discharge	2892 (11.2 %)	449 (1.3 %)

e

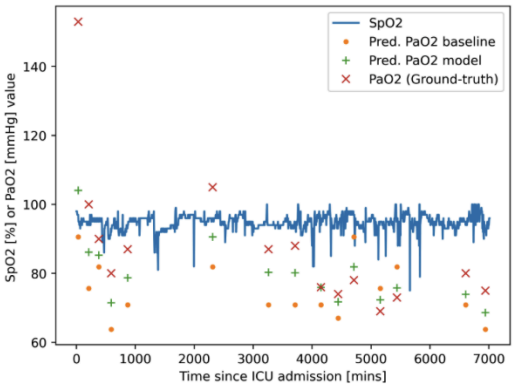


Extended Data Fig 2. Association of ICU Mortality with Respiratory Failure / Extubation Failure / Ventilation. a. Mortality statistics for patients with respiratory failure at some time during their ICU stay, and those without respiratory failure during their ICU stay. **b.** Relationship of ICU mortality rate with fraction of the stay in which patients experience respiratory failure. **c.** Mortality statistics for patients with extubation failure, and those without extubation failure but with at least one successful extubation. **d.** Mortality statistics for patients receiving mechanical ventilation during their ICU stay, and those not ventilated. **e.** Relationship of ICU mortality rate with fraction of their stay during which patients are mechanically ventilated.

a

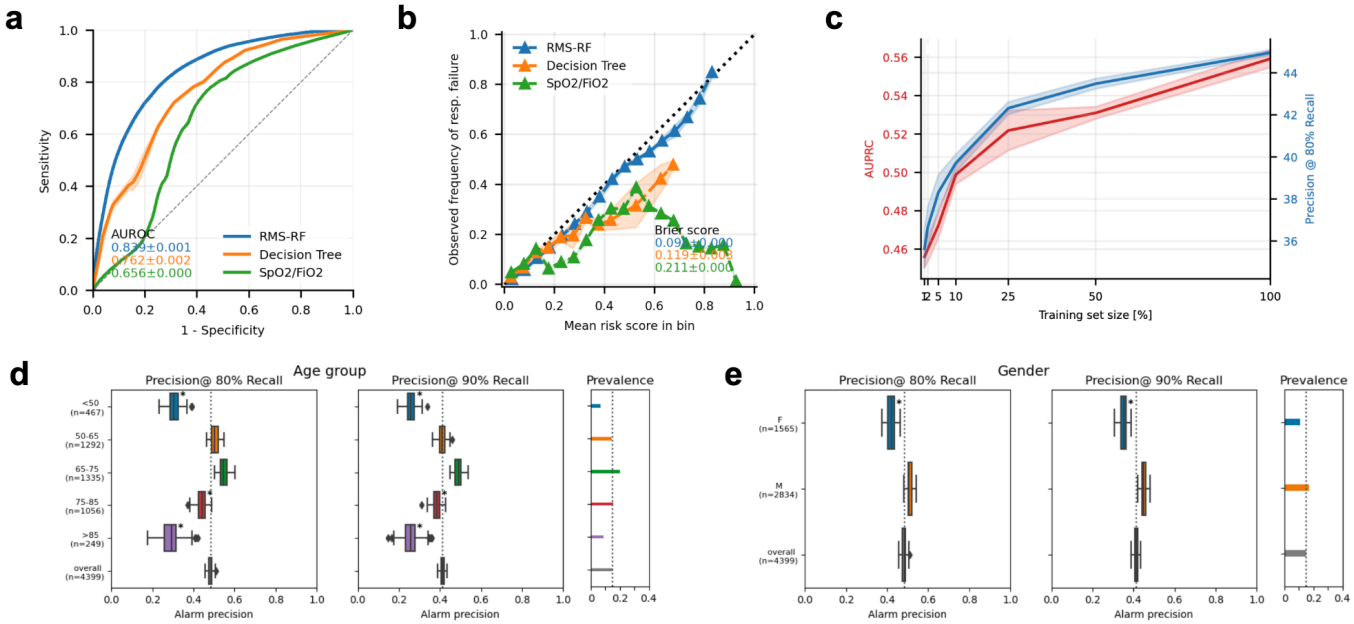
	Absolute error [mmHg]
PaO_2 estimation model	6.25 [2.74,13.80]
Severinghaus-Ellis Baseline	12.90 [5.45,32.94]

b

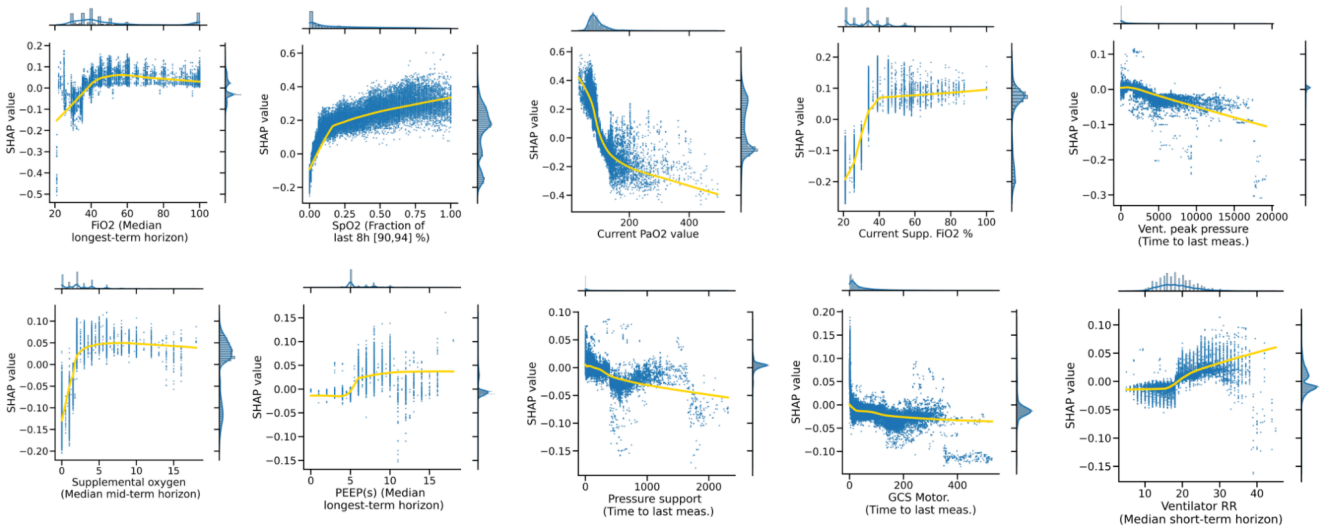


Extended Data Fig 3. Performance of PaO_2 estimation model. a. Performance evaluation of PaO_2-estimation model on HiRID-II test set in terms of MAE vs. ground-truth PaO_2 from invasive blood tests, compared with the non-linear Severinghaus-Ellis baseline. **b.** Example time series of predicted and ground-truth PaO_2 values, as well as SpO_2 values, and baseline predictions. A patient was selected for which the median absolute error of both model and baseline is close to their population median reported in panel a. The tendency of the PaO_2 model to predict closer to the ground-truth observations in case of outliers is clearly visible.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

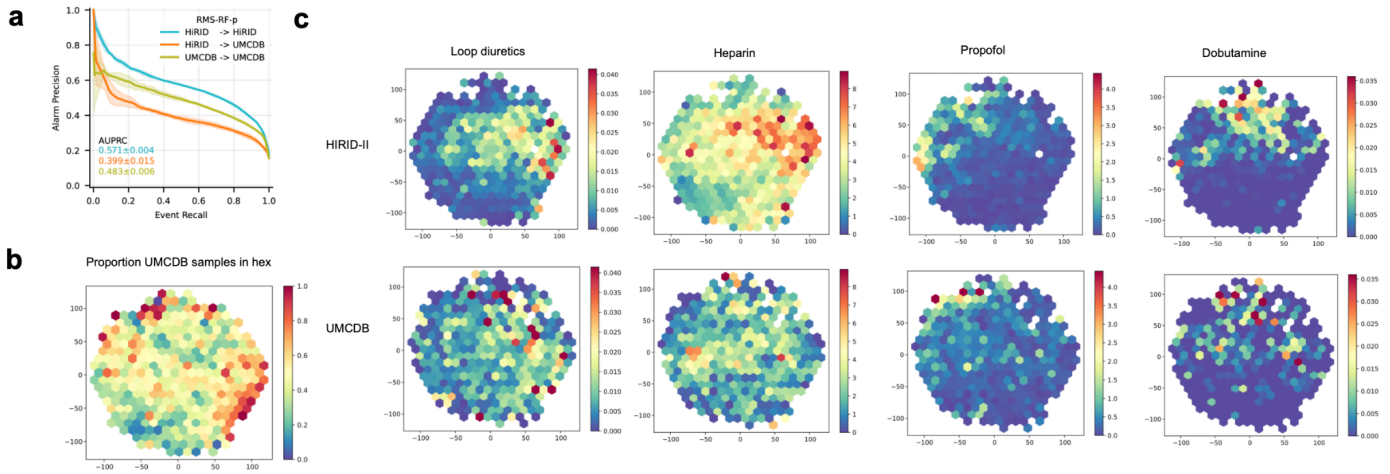


Extended Data Fig 4. Evaluation of RMS-RF. a. ROC-based performance of the RMS-RF score, compared with the two baselines. **b.** Calibration of the RMS-RF model compared with the two baselines. **c.** Performance of the RMS-RF model, as the training set size is varied, in terms of complete patients **d.** Performance of the RMS-RF model by age group, for event recalls of 80/90 %. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable event recall. **e.** Performance of the RMS-RF model by gender, for event recalls of 80/90 %. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable event recall.



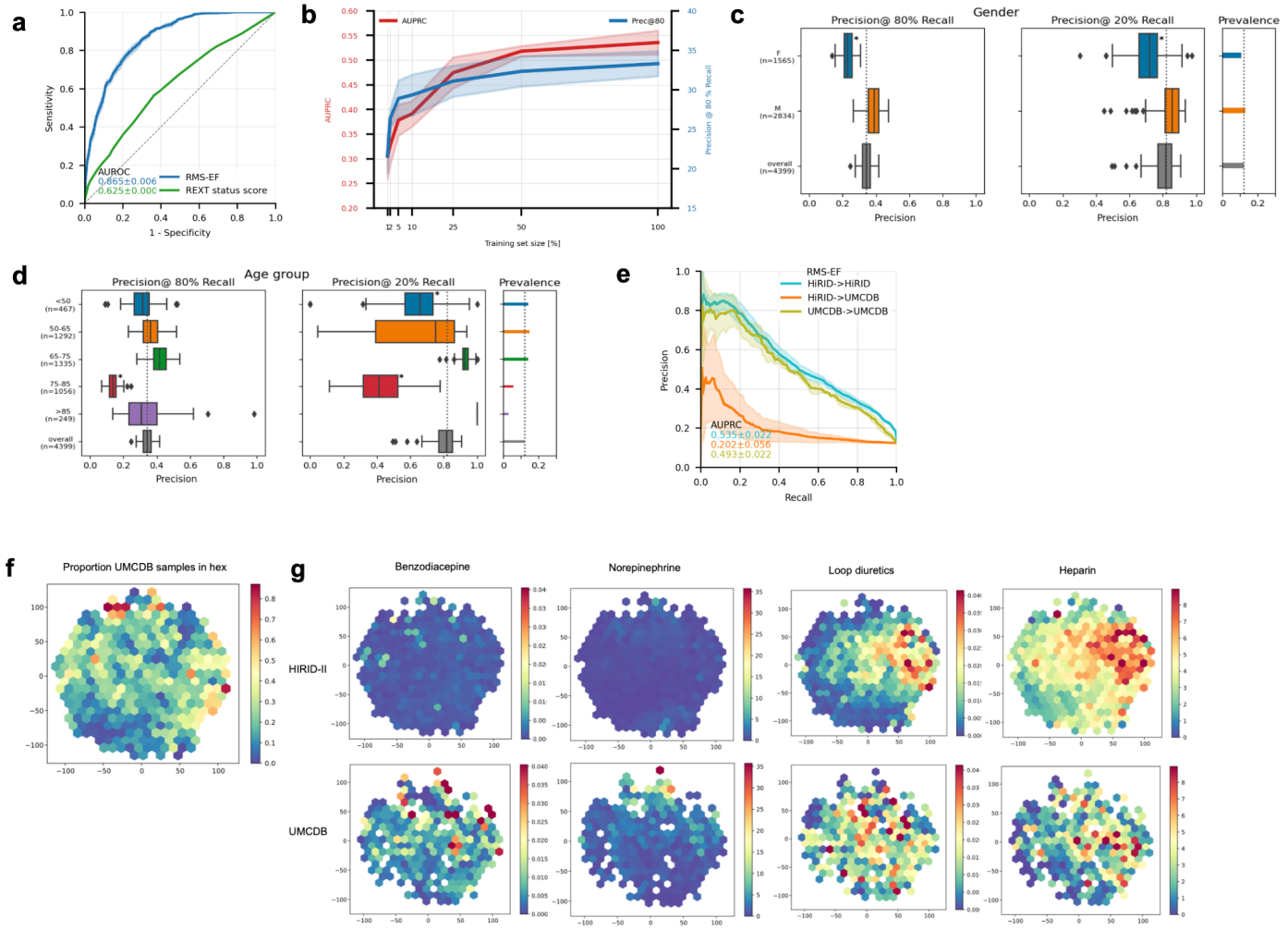
Extended Data Fig 5. Model introspection of RMS-RF. SHAP value - feature value interactions of the top feature of the top 10 most important variables contained in RMS-RF.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



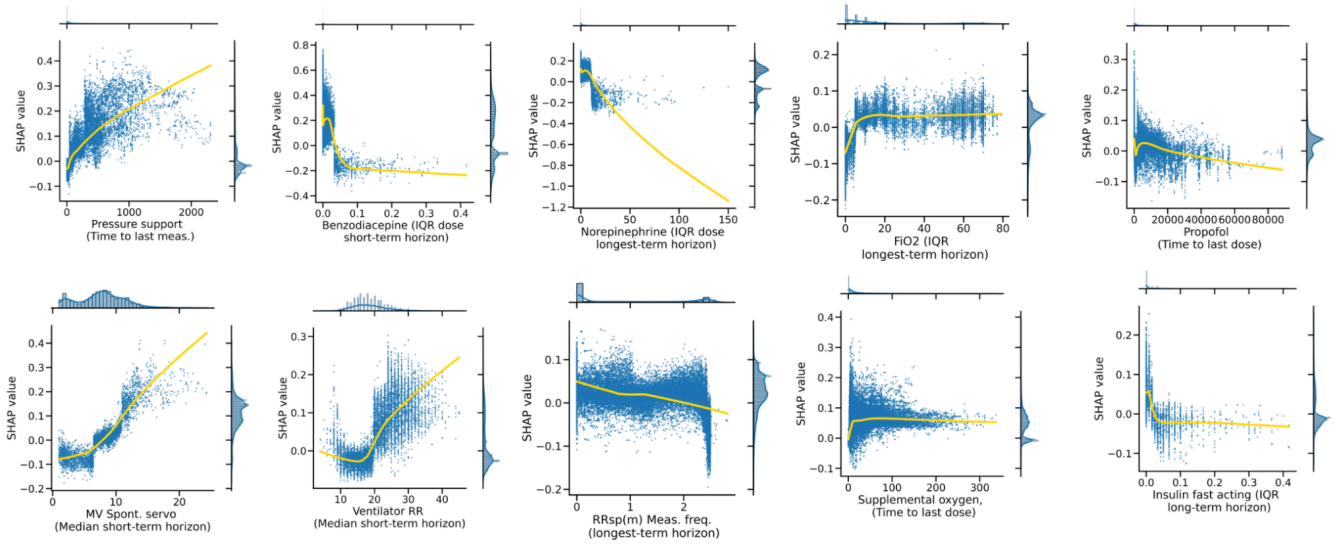
Extended Data Fig 6. External validation of RMS-RF-p / Medication policy comparison HiRID-II/UMCdb. a. Performance of the RMS-RF-p model, which additionally includes medication variables, when trained and tested on HIRID-II, transferred to UMCdb and retrained in the UMCdb database **b.** t-SNE embedding of time-points in the test set of a pooled dataset between samples from HiRID-II and UMCdb (1:1 ratio of two datasets), of physiological parameters. Only time-points when the patient is not in respiratory failure are taken into account, for which the RMS-RF-p model is active. The color indicates the proportion of time-points in the UMCdb dataset in a given hex. **c.** The same t-SNE embedding as in b is displayed separately for time-points from the HiRID-II dataset, and the UMCdb dataset, corresponding to the rows. The hexes in the t-SNE are colored by the mean drug dosage of all time-points assigned to the hex. The four medication variables, for which transfer issues of the RMS-RF-p model were detected, are analyzed in the columns. Medication policy differences are visible for all four variables, in particular for Heparin & Propofol.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

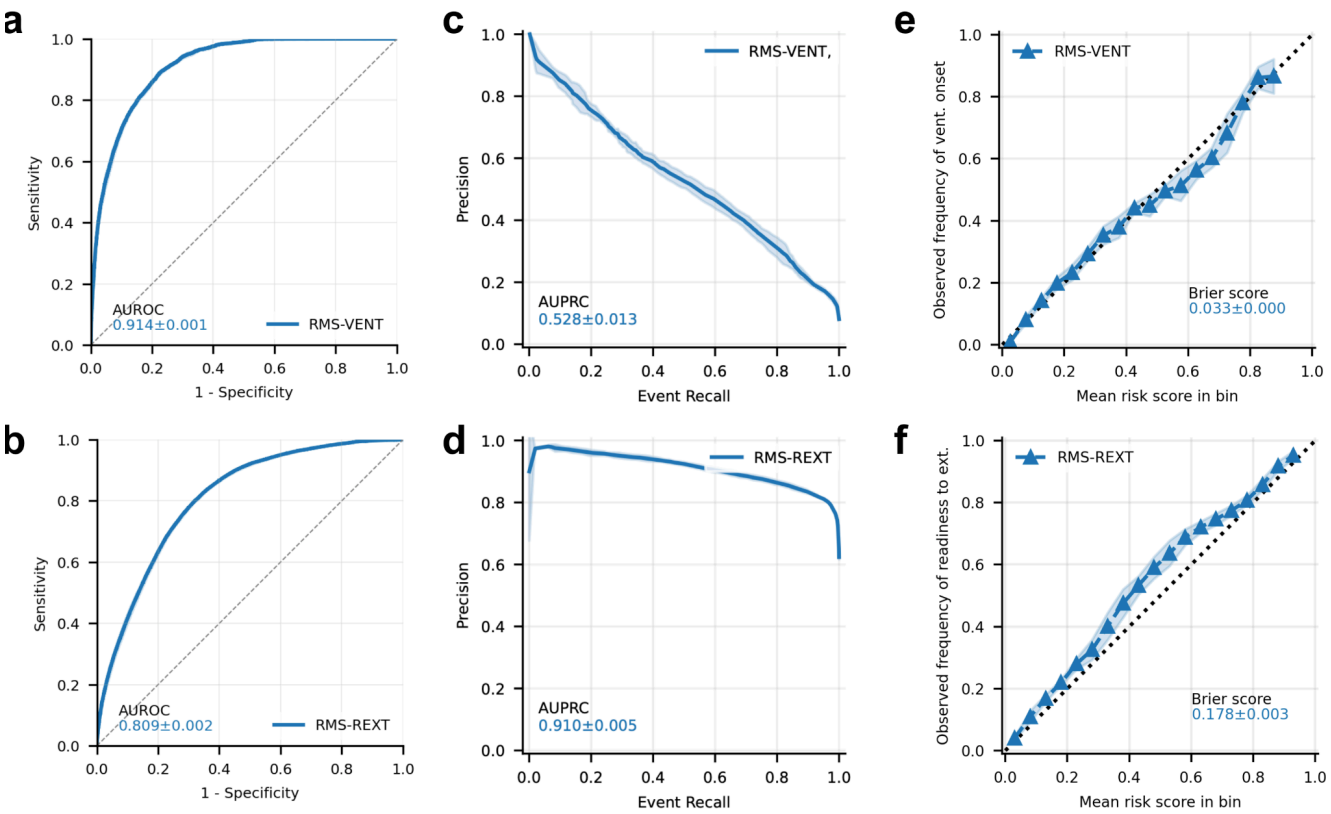


Extended Data Fig 7. Evaluation of RMS-EF / Medication policy comparison HiRID-II/UMCdb. a. ROC-based performance of RMS-EF, compared with the baseline. **b.** Performance of RMS-EF as the training set size is varied between 1 % and 100 % of the original dataset size, by subsampling complete patient records in the training set. **c.** Performance of RMS-EF for different genders, at recalls of 80/20 %. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable recall. **d.** Performance of RMS-EF for different age groups, at recalls of 80/20 %. The model was re-calibrated for each sub-group using information available at admission time, to achieve a comparable recall. **e.** Performance of the RMS-EF model, when trained/tested in the HIRID-II database, transferred to the UMCdb database, and retrained in the UMCdb database. **f.** t-SNE embedding of time-points in the test set of a pooled dataset between samples from HiRID-II and UMCdb (1:1 ratio of two datasets), of physiological input variables. Only time-points when the patient is ready-to-extubate are taken into account, for which the RMS-EF model is active. The color indicates the proportion of time-points in the UMCdb dataset in a given hex. **g.** The same t-SNE embedding as in g is displayed separately for time-points from the HiRID-II dataset, and the UMCdb dataset, corresponding to the rows. The hexes in the t-SNE are colored by the mean drug dosage of all time-points assigned to the hex. The four medication variables, for which transfer issues of the RMS-EF model were detected, are analyzed in the columns. Medication policy differences are visible for all four variables, in particular for Benzodiacepine & Norepinephrine.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

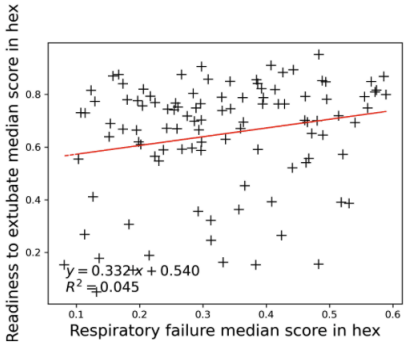


Extended Data Fig 8. Model introspection of RMS-EF. SHAP value - feature value interactions for the top feature of the top 10 most important variables contained in the RMS-EF model.



Extended Data Fig 9. Evaluation of RMS-VENT/RMS-REXT. a. ROC-based performance of RMS-VENT, predicting ventilation onset within the next 24h. **b.** ROC-based performance of RMS-REXT, predicting being newly ready to extubate within the next 24h. **c.** Event-based PRC of the RMS-VENT alarm system. **d.** Event-based PRC of the RMS-REXT alarm system. **e.** Calibration of the RMS-VENT score. **f.** Calibration of the RMS-REXT score.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



Extended Data Fig 10. Joint task analysis details. Scatter plot of median respiratory failure vs. median readiness to extubate score in the hexes analyzed in the explorative joint analysis of RMS scores (see Fig. 5). A light positive correlation between respiratory failure and readiness to extubate scores can be observed, which is barely significant at 5 % level. A Wald test which tests non-zero slope of the regression line (shown in red) was performed.

Supplemental Materials

Supplemental Table 1. Details on the clinical parameters extracted in the HiRID-II dataset (downloadable XLSX file).

Supplemental Table 2. Details on the imputation parameters, such as normal value, and imputation models, for the clinical parameters (downloadable XLSX file).

Supplemental Table 3. List of important variables used for computing complex features, as a basis for variable selection, and for building the final models RMS-RF/RMS-EF/RMS-VENT/RMS-REXT (downloadable XLSX file).

Supplemental Table 4. List of severity levels for computing 'instability history' features, for a subset of the important variables. (downloadable XLSX file).

Supplemental Table 5. Model training parameters and grid used for selection of hyperparameters for the LightGBM library (downloadable XLSX file).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

References

1. Vincent, J. L., Sakr, Y. & Ranieri, V. M. Epidemiology and outcome of acute respiratory failure in intensive care unit patients. *Crit. Care Med.* **31**, S296–9 (2003).
2. Donchin, Y. & Seagull, F. J. The hostile environment of the intensive care unit. *Curr. Opin. Crit. Care* **8**, 316–320 (2002).
3. Sanchez-Pinto, L. N., Luo, Y. & Churpek, M. M. Big Data and Data Science in Critical Care. *Chest* **154**, 1239–1248 (2018).
4. Wung, S.-F., Malone, D. C. & Szalacha, L. Sensory Overload and Technology in Critical Care. *Crit. Care Nurs. Clin. North Am.* **30**, 179–190 (2018).
5. Bai, Y., Xia, J., Huang, X., Chen, S. & Zhan, Q. Using machine learning for the early prediction of sepsis-associated ARDS in the ICU and identification of clinical phenotypes with differential responses to treatment. *Front. Physiol.* **13**, 1050849 (2022).
6. Lam, C., Thapa, R., Maharjan, J. & Rahmani, K. Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model …. *JMIR Medical* (2022).
7. Wu, J. *et al.* Early prediction of moderate-to-severe condition of inhalation-induced acute respiratory distress syndrome via interpretable machine learning. *BMC Pulm. Med.* **22**, 193 (2022).
8. Le, S. *et al.* Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). *J. Crit. Care* **60**, 96–102 (2020).
9. Ding, X.-F. *et al.* Predictive model for acute respiratory distress syndrome events in ICU patients in China using machine learning algorithms: a secondary analysis of a cohort study. *J. Transl. Med.* **17**, 326 (2019).
10. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).
11. Desautels, T. *et al.* Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform* **4**, e28 (2016).
12. Wang, D. *et al.* A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. *Front Public Health* **9**, 754348 (2021).
13. Moor, M. *et al.* Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine* **62**, 102124 (2023).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

14. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
15. Faltys, M. *et al.* HiRID, a high time-resolution ICU dataset. (2021) doi:10.13026/NKWC-JS72.
16. Thoral, P. J. *et al.* Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. *Crit. Care Med.* **49**, e563–e577 (2021).
17. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, (2000).
18. Moody, G. B., Mark, R. G. & Goldberger, A. L. PhysioNet: a web-based resource for the study of physiologic signals. *IEEE Eng. Med. Biol. Mag.* **20**, 70–75 (2001).
19. Rice, T. W. *et al.* Comparison of the SpO2/FIO2 ratio and the PaO2/FIO2 ratio in patients with acute lung injury or ARDS. *Chest* **132**, (2007).
20. Patel, S., Jose, A. & Mohiuddin, S. S. Physiology, Oxygen Transport And Carbon Dioxide Dissociation Curve. in *StatPearls* (StatPearls Publishing, 2023).
21. Wagner, P. D. Blood Gas Transport: Carriage of Oxygen and Carbon Dioxide in Blood. *Semin. Respir. Crit. Care Med.* **44**, 569–583 (2023).
22. Pandharipande, P. P. *et al.* Derivation and validation of Spo2/Fio2 ratio to impute for Pao2/Fio2 ratio in the respiratory component of the Sequential Organ Failure Assessment score. *Crit. Care Med.* **37**, (2009).
23. Brown, S. M. *et al.* Nonlinear Imputation of Pao2/Fio2 From Spo2/Fio2 Among Patients With Acute Respiratory Distress Syndrome. *Chest* **150**, 307–313 (2016).
24. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems 30* 3146–3154 (2017).
25. Yèche, H. *et al.* HiRID-ICU-Benchmark --- A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* **1**, (2021).
26. Hoche, M., Mineeva, O., Burger, M., Blasimme, A. & Rätsch, G. FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems. (2024).
27. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in*

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Neural Information Processing Systems 30 (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).

28. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv [cs.LG]* (2018).
29. Maaten, L. van der & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
30. Scott, J. B., De Vaux, L., Dills, C. & Strickland, S. L. Mechanical Ventilation Alarms and Alarm Fatigue. *Respir. Care* **64**, 1308–1313 (2019).
31. Zeiberg, D. *et al.* Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS One* **14**, e0214465 (2019).
32. Singhal, L. *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**, e0257056 (2021).
33. Bolourani, S. *et al.* A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation. *J. Med. Internet Res.* **23**, e24246 (2021).
34. Ferrari, D. *et al.* Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia-Challenges, strengths, and opportunities in a global health emergency. *PLoS One* **15**, e0239172 (2020).
35. Bendavid, I. *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.* **12**, 10573 (2022).
36. Igarashi, Y. *et al.* Machine learning for predicting successful extubation in patients receiving mechanical ventilation. *Front. Med.* **9**, 961252 (2022).
37. Huang, K.-Y. *et al.* A machine learning model for prediction of successful extubation in patients admitted to the intensive care unit. (2022).
38. Zeng, Z., Tang, X., Liu, Y., He, Z. & Gong, X. Interpretable recurrent neural network models for dynamic prediction of the extubation failure risk in patients with invasive mechanical ventilation in the intensive care unit. *BioData Min.* **15**, 21 (2022).
39. Wang, H. *et al.* Early prediction of noninvasive ventilation failure after extubation: development and validation of a machine-learning model. *BMC Pulm. Med.* **22**, 304 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

40. Otaguro, T. *et al.* Machine Learning for Prediction of Successful Extubation of Mechanical Ventilated Patients in an Intensive Care Unit: A Retrospective Observational Study. *J. Nippon Med. Sch.* **88**, 408–417 (2021).
41. Zhao, Q.-Y. *et al.* Development and Validation of a Machine-Learning Model for Prediction of Extubation Failure in Intensive Care Units. *Front. Med.* **8**, 676343 (2021).
42. Chen, T. *et al.* Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine. *IEEE Access* **7**, 150960–150968 (2019).
43. Lazzarini, N., Filippoupolitis, A., Manzione, P. & Eleftherohorinou, H. A machine learning model on Real World Data for predicting progression to Acute Respiratory Distress Syndrome (ARDS) among COVID-19 patients. *PLoS One* **17**, e0271227 (2022).
44. Sayed, M., Riaño, D. & Villar, J. Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning. *J. Clin. Med. Res.* **10**, (2021).
45. Jia, Y., Kaul, C., Lawton, T., Murray-Smith, R. & Habli, I. Prediction of weaning from mechanical ventilation using Convolutional Neural Networks. *Artif. Intell. Med.* **117**, 102087 (2021).
46. Zeng, L. *et al.* VentSR: A Self-Rectifying Deep Learning Method for Extubation Readiness Prediction. in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1369–1374 (2022).
47. Shashikumar, S. P. *et al.* Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation. *Chest* **159**, 2264–2273 (2021).
48. Siu, B. M. K., Kwak, G. H., Ling, L. & Hui, P. Predicting the need for intubation in the first 24 h after critical care admission using machine learning approaches. *Sci. Rep.* **10**, 20931 (2020).
49. Sottile, P. D., Albers, D., Higgins, C., Mckeehan, J. & Moss, M. M. The Association Between Ventilator Dyssynchrony, Delivered Tidal Volume, and Sedation Using a Novel Automated Ventilator Dyssynchrony Detection Algorithm. *Crit. Care Med.* **46**, e151–e157 (2018).
50. Lapp, L., Roper, M., Kavanagh, K., Bouamrane, M.-M. & Schraag, S. Dynamic Prediction of Patient Outcomes in the Intensive Care Unit: A Scoping Review of the State-of-the-Art. *J. Intensive Care Med.* **38**, 575–591 (2023).
51. Calvert, J. *et al.* Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg (Lond)* **11**, 52–57 (2016).
52. Roggeveen, L. *et al.* Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artif. Intell. Med.* **112**, 102003 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.23.24301516; this version posted January 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

53. Chen, Q. *et al.* Transferability and interpretability of the sepsis prediction models in the intensive care unit. *BMC Med. Inform. Decis. Mak.* **22**, 343 (2022).

54. Alves, T., Laender, A., Veloso, A. & Ziviani, N. Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation. in *2018 IEEE International Conference on Big Data (Big Data)* 1328–1336 (ieeexplore.ieee.org, 2018).

55. Lorenzen, S. S. *et al.* Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark. *Sci. Rep.* **11**, 18959 (2021).

56. Tariq, A. *et al.* Patient-specific COVID-19 resource utilization prediction using fusion AI model. *NPJ Digit Med* **4**, 94 (2021).

57. Manduchi, L., Hüser, M., Vogt, J., Rätsch, G. & Fortuin, V. DPSOM: Deep Probabilistic Clustering with Self-Organizing Maps. (2019).

58. Fortuin, V., Hüser, M., Locatello, F., Strathmann, H. & Rätsch, G. SOM-VAE: Interpretable Discrete Representation Learning on Time Series. (2018).

59. Calfee, C. S. *et al.* Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* **2**, 611–620 (2014).

60. Alipanah, N. & Calfee, C. S. Phenotyping in acute respiratory distress syndrome: state of the art and clinical implications. *Curr. Opin. Crit. Care* **28**, 1–8 (2022).

61. Wilson, J. G. & Calfee, C. S. ARDS Subphenotypes: Understanding a Heterogeneous Syndrome. *Crit. Care* **24**, 102 (2020).

62. Bos, L. D. *et al.* Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis. *Thorax* **72**, 876–883 (2017).

63. Bos, L. D. J. *et al.* Longitudinal respiratory subphenotypes in patients with COVID-19-related acute respiratory distress syndrome: results from three observational cohorts. *Lancet Respir Med* **9**, 1377–1386 (2021).

64. Joshi, R. & Szolovits, P. Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding. *AMIA Annu. Symp. Proc.* **2012**, 1276–1283 (2012).

65. Liley, J. *et al.* Model updating after interventions paradoxically introduces bias. (2020).

66. Stenwig, E., Salvi, G., Rossi, P. S. & Skjærvold, N. K. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med. Res. Methodol.* **22**, 53 (2022).