Online Methods

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.

Study design and setting

The study was designed as a retrospective cohort study to develop and validate a set of clinical prediction models that are combined to form an ML-based respiratory monitoring system. The study was performed using data from the Department of Intensive Care Medicine at the University Hospital Bern, an interdisciplinary unit admitting > 6,500 patients per year, which was used for model development and internal validation. For external validation, an open-source data-set from the Amsterdam University Medical Center, referred to as UMCdb[1], was used, and harmonized to match the same structure as the HiRID-II data-set.

Ethical approval and patient consent

The institutional review board (IRB) of the Canton of Bern approved the study (BASEC 2016 01463). The need for obtaining informed patient consent for patient data from our institution was waived owing to the retrospective and observational nature of the study. No IRB approval is required for the anonymous public external validation data-set from Amsterdam University Medical Center.

Participants and data sources

Details about participants and patient inclusion criteria in the two data sets are described in Table 1 / Extended Data Figure 1.

Data - HiRID II

For this work, we prepared the second version of the High time Resolution Intensive Care Unit Dataset (HiRID-II), consisting of high-temporal-resolution data from over 55,000 patient admissions to the intensive care units (ICUs) at the Bern University Hospital in Switzerland between January 2008 and June 2019.

HiRID-II is an increment over the first HiRID dataset released by Faltys et al. on Physionet[2], counting over 33,000 patients between January 2008 and August 2016. HIRID-II additionally includes patients without data for determining circulatory failure or receiving any form of full mechanical circulatory support (previously excluded from HiRID-I) and patient data between August 2016 and June 2019. The final dataset is obtained after applying exclusion criteria over 74142 initial admissions (see flow chart in Extended Data Figure 1).

HiRID-I includes 681 variables recorded in the PDMS (GE Centricity Critical Care, General Electrics) which are merged into 209 meta-variables based on their clinical concepts. HIRID-II records 218 more variables than the HiRID-I data-set, which yield 113 more meta-variables after variable merging. It is planned to release a version of HiRID-II on Physionet including the new admissions and variables. Details about meta-variables are listed in Supplemental Table 1.

Anonymization procedure

To ensure the anonymization of individuals in the data set, we followed the procedures successfully applied for the MIMIC-III and AmsterdamUMCdb datasets, which adopted the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor requirements and, in the case of AmsterdamUMCdb, also the European Union's General Data Protection Regulation (GDPR) standards[3,4].

This includes removal of all eighteen identifying data elements listed in HIPAA. Free text was removed from the database. Patient age, height and weight were grouped into bins of size 5, with patients aged 90 years and older binned together. K-anonymization was subsequently applied on patient age, weight, height and sex. This procedure was separately applied to the original HIRID I dataset (anonymized by Faltys et al.), to the additional set of training patients (2008-2018), and on the held-out test dataset (2018-2019).

Within these temporally distinct training and test sets, admission dates were shifted by a random offset to lie between 2100 and 2200, while preserving seasonality, time of day and day of week. Measurements and medications with changing units over time were standardized to the latest unit used, to ensure no inference on admission time could be carried out from units used.

Data splits

For designing the data splits in HiRID-II, the publicly-released temporal data split into development and test set was used as a basis, taken as output by the K-anonymization procedure described above. The test set of this split was held-out and not used prior to generating the final results for drafting the manuscript, to avoid subtle overfitting during the model design process. The development set was further divided randomly by complete patients, using 80 % of the patients to yield the final training set, and 20 % of the patients to yield the validation set. The final training set was used for model training. The validation set was used for selecting optimal hyperparameters, as well as early stopping of the training process. Performance in the validation set guided model design decisions in the prototyping phase as well as selection of clinical parameters using greedy forward selection. This splitting procedure into training and validation sets was repeated independently 5 times, to yield 5 splits.

The model development data set drawn from HiRID-II contains 51,457 patients and the test set 4,401 patients. A temporal splitting strategy analogous to the one presented by Hyland et al.[5] was used, but with only one fixed test set to minimize leakage of admission time information. Five independent partitions into training and validation set, containing 41,165 and 10,292 patients respectively, were extracted from the development dataset, to estimate variation of model performance.

For UMCDB a fixed test set consisting of 25 % of patients was drawn, which is shared by the 5 splits. The remaining patients formed the development data-set, which was 5 times partitioned in proportion 80:20 % by complete patients, to form 5 random splits.

Analysis platform

HiRID-II/UMCdb data was extracted to and all computational analyses were performed on a secure compute cluster environment located at ETH Zürich (https://scicomp.ethz.ch/wiki/Leonhard). Python3, with numpy, pandas, matplotlib and scikit-learn forming the backbone of the data-processing pipeline. For model training, the LightGBM[6] package was used. Processing was performed in a batched form across most steps of the processing pipeline, with the data of HiRID-II data-set being split into 100 batches, and the data of the UMCdb data-set being split into 50 batches.

Data preprocessing, Variable merging, Artifact rejection, Medication preprocessing

Similar to the preprocessing steps presented by Hyland et al.[5], we first remove different types of artifacts in the data, namely timestamp artifacts and variable-misnaming artifacts, out-of-range-value artifacts, and record duplication artifacts. For variables that encode cumulative values, we convert the cumulative values to a rate. The dose values of

medication variables are either converted to a rate or a binary indicator depending on their clinical relevance to respiratory failure defined by the clinicians. Drugs that are given in the form of discrete boluses were converted to continuous rates over a defined acting period. The acting period differs for different drugs and the details can be found in Supplemental Table 1.

After artifact removal and converting relevant variables to rates, we merge variables with the same or close clinical meaning/function into one single meta-variable. For example, the three temperature variables, i.e. core body temperature, rectal temperature, and auxiliary temperature, were merged into one meta-variable called temperature. Another example are drugs with the same compounds or compounds having the same pharmaceutical effect which will be merged as one medication group. For each meta-variable, we take the median value of the available measurements at each timepoint when any of the corresponding variables is measured.

Data imputation

The time grid for time gridding has step size of 5-minute and is defined between the first and last heart rate measurement of the intensive care unit stay. The time grid is cut off after 28 days, and for those patients exceeding this stay length only the first 28 days of observations are used, to avoid rare very long stays to bias the model development process. Patients without heart rate observations are excluded at this stage.

Besides the estimated values, at each grid point a binary measurement indicator column is introduced which is 1 if there was observation in the 5 minutes, and 0 otherwise. Further, a time to last measurement column is introduced which is -1 if there is no previous observation prior to the grid point, and equal to the number of minutes since the last observation otherwise.

“Dense imputation” where every value on the grid is finite is only used as a pre-processing stage for endpoint annotation and label definition, whereas data was gridded but not imputed at every grid point for feature extraction, to preserve some missingness patterns in the data. Here a missing indicator (NAN) is left at grid points where the value cannot be estimated. Subsequently we refer to these two modes as “feature imputation”, and “dense imputation” respectively.

For each meta-variable in the HiRID-II data schema, an imputation mode was defined, and then applied using the imputation algorithm. If there is no prior measurement before a grid point, the grid point is filled with a missing-value indicator (for “feature imputation”), or a clinically defined normal value (for “dense imputation”). If there was more than one observation with the same time-stamp, the mean of all such observations is used. The imputation modes are:

a) Indefinite forward filling: The last measurement is indefinitely forward filled to each point on the 5-minute grid later in time.
b) Limited forward filling using prior knowledge: Each measurement is forward filled up to a maximum time of k minutes, manually specified using medical knowledge
c) Limited data-adaptive forward filling: From the training set, the median and standard deviation of the observation intervals for a clinical parameter are estimated. Forward filling is then applied for up to 2*median(interval) + std(interval) minutes. This forms a fall-back if the forward filling horizon cannot be specified using prior knowledge.
d) Attribute to exact grid point: No forward filling is used if a measurement is only relevant for a very short time at the exact grid point where it was observed, so forward filling is limited to 5 minutes, i.e. only the grid point next to the measurement location contains the value.

For each variable in “dense imputation”, a normal value is specified which is used at grid points, where no value can be estimated. This is either specified by (a) prior clinical knowledge, (b) for weight, and the patient height is available, it is estimated from the patient’s BMI, which is defined using a look-up table.

The imputation modes used for each variable are listed in Supplemental Table 2, per clinical parameter..

For the variables ‘Cardiac output’, ‘Urine output’, ‘Fluid input’, ‘Fluid output’, imputation used special formulae to estimate the current value based on the patient’s observed/estimated height/weight and BMI values. More details are given in the Supplementary Table 2.

For static variables, analogously to time series variables, ‘dense imputation’ guarantees there are no missing values, and median/mode imputation based on statistics from the training set was used for continuous variables and categorical variables, respectively. ‘Feature imputation’ for static variables uses no imputation, missing values are left as NAN.

PaO_2 estimation

The annotation of respiratory failure depends on the availability of a current PaO_2 value. To measure PaO_2, an arterial blood sample (ABGA) of the patient has to be drawn and processed. Therefore, PaO_2 measurements are only available at intervals determined by its measurement frequency. For a continuous assessment of patient respiratory state using P/F ratio, estimates of PaO_2 values have to be used when its measurements are not available. In the clinical setting, continuously monitored pulse oximetry derived haemoglobin oxygen saturation (SpO_2) can be used to estimate the current PaO_2 value[7–9]. To reduce the effect of outliers, the SpO_2 time series was pre-processed with a percentile smoother (75% percentile kernel function, 30 min centralized kernel window). A literature review of existing models revealed that the non-linear parametric model by Ellis[8,9] performs best. We were able to further improve upon Ellis in PaO_2 estimation using 2 nested regularized L2 regression models by using 7 hand-crafted features, defined at each time-point of the stay. Polynomial features of degree 3 are then computed on these features to capture non-linear interactions explicitly.

- Last real SpO_2 measurement
- Last real PaO_2 measurement
- Last real SaO_2 measurement
- Last real pH measurement
- Time to last real SpO_2 measurement
- Time to last real PaO_2 measurement
- Closest SpO_2 measurement to the last PaO_2 measurement

This base model is nested into a meta-model which performs the final prediction. As input features the meta-model uses (1) the same polynomial features of the base model, (2) the prediction made by the base model as well as (3) the prior mistakes, i.e. signed offsets between base model prediction and ground-truth of the (up to) 10 prior PaO_2 real measurements. This allows the model to adapt using the context of previous wrong predictions to improve its predictions over the time of the ICU admission, and adapt to the patient physiology at hand.

The model is only trained using time-points where at last one prior measurement is available for each of SpO_2, PaO_2, SaO_2 and pH and a PaO_2 measurement was recorded at this time-point. The regression label is equal to the ground-truth PaO_2 measurement. During evaluation a prediction is only made if at least one PaO_2 measurement was previously observed. Before the first PaO_2 measurement, the normal imputation algorithm for PaO_2 (forward filling) is used instead of the prediction model.

For both base model and meta model a L2 regression loss function with a Huber regularizer were used, and using a separate validation set, the regularization weight alpha was optimized over the range [1.0,0.1,0.01,0.001,0.0001]. In the loss function the samples were weighted according to the formula $10+100/(1+\exp(0.025*(\text{Real}PaO_2-110)))$, to give true PaO_2 values close to the relevant decision boundaries for respiratory failure annotation higher weights. Model development of the PaO_2 estimation model did not use the held-out test set, which was not used prior to final preparation of figures.

FiO_2 estimation

For the calculation of the P/F ratio, estimates of FiO_2 values are necessary for every grid point. Three situations need to be distinguished: 1) the patient is breathing ambient air, i.e. FiO_2 = 21% (the ambient air oxygen fraction); 2) the patient is receiving supplemental oxygen and the corresponding FiO_2 is recorded in the data 3) for patients on mechanical ventilation FiO_2 is controlled by the ventilator and its value is recorded in the data. FiO_2 estimation at every grid point is implemented in the following way.

a) FiO_2 is forward filled from the last FiO_2 measurement, if (1) it was within the last 30 minutes, and (2) the patient was

estimated to be mechanically ventilated (using the ventilation detection algorithm described later) or the ventilation mode is NIV.
b) Otherwise, the two supplementary oxygen variables (Supplemental FiO_2 [%] and Highflow FiO_2 [%]) were considered, if a measurement was available in the last 12 hours. Hereby, Supplemental FiO_2 [%] takes precedence, if it was available in the last 12 hours.
b) If there was no measurement in the two supplementary oxygen FiO_2 variables in the last 12 hours, then an ambient air assumption was made, and FiO_2 was estimated at 21 %.

Estimation of the P/F index

The P/F index (or ratio) at each grid point is defined as PaO_2-estimate / FiO_2 estimate, where the PaO_2, FiO_2 estimates at the grid point were found by the two schemas explained above. As post-processing a Nadaraya Watson kernel smoother with a bandwidth of 20 was applied to the tentative P/F indices, to yield the final estimated P/F ratios per grid time point.

Respiratory failure annotation

Lung function is clinically evaluated using the ratio of blood oxygen partial pressure (PaO_2) and fraction of inspired oxygen (FiO_2), commonly referred to as P/F ratio[10]. A healthy person breathing room air is expected to have a P/F ratio of 475 mmHg (PaO_2: ~100 mmHg, room air FiO_2: 21 %). Current medical literature defines respiratory failure in three stages[11]:

Mild: 200 mmHg ≤ P/F ratio < 300 mmHg
Moderate: 100 mmHg ≤ P/F ratio < 200 mmHg
Severe: P/F ratio < 100 mmHg

A grid point is labeled with the 3 severity levels or ‘stable’ using a forward facing window of length 1 hour. If ⅔ of the grid points satisfied the severe criterion (<100 mmHg), it was labeled as ‘severe respiratory failure’, otherwise if ⅔ of the grid points satisfied the moderate criterion (<200 mmHg), it was labeled as ‘moderate respiratory failure’, otherwise if ⅔ of the grid points satisfied the mild criterion (<300 mmHg), it was labeled as ‘mild respiratory failure’. Otherwise the patient was labeled as ‘stable’ at the grid point. If for at least ⅔ grid points, the P/F ratio could not be estimated, the respiratory failure status of the grid point was set to ‘Unknown’. Additionally to satisfying the condition on the P/F ratio in the window, we also required that the patient was in a consistent ventilation state during the grid-points where the P/F ratio criterion is satisfied (patient is not ventilated, or patient is ventilated and PEEP is not densely available, or patient is ventilated and PEEP is densely available and satisfying PEEP>=4).

Because the labeling algorithm with a forward facing window can mis-label points on the right edges of events as ‘not in failure’, the right edges of events were manually corrected by scanning right-wards from the tentative right edge of the event and setting the grid point to the respective severity level if the current P/F ratio actually satisfied the criterion, but was mis-labeled as not satisfying the criterion due to the forward-facing 1 hour window.

As a last step, a postprocessing is performed where small events (length <= 4 hours) that are sandwiched between two other events, (1) at least one of which is longer than the sandwiched event, and (2) the 2 surrounding events have the same severity label, are relabeled to match the label of the surrounding events. In this way, spuriously labeled small respiratory failure events of length shorter than 4 hours will be deleted. Moreover small gaps between two respiratory failure events will be deleted and the two events merged together.

Ventilation status annotation

To derive ventilation status (0/1 binary) at each grid-point a voting algorithm was used, which
was informed by prior medical knowledge. Each criterion was evaluated per grid point, and depending on the outcome, positive or negative points were assigned. Positive points correspond to a higher likelihood of ventilation at a grid point, negative points to a lower likelihood of ventilation. Finally, a cut-off on the total sum of the points was specified, using prior medical knowledge, and by judging the correctness of endpoints visually using a time series visualization toolkit, developed for this project. Points assigned by the voting system are as follows:

a) +1 point, patient was admitted before 2009/12/06, to take into account different recording of ventilation information in the EHR before this data
b) +2 points: In a centered 30 minute window on the grid point, at least one EtCO2 measurement of >0.5 was observed
c) +1 point: The current estimated ventilation mode is 2 (controlled mode) or 3 (spontaneous mode)
d) -1 point: The current estimated ventilation mode is 1 (standby)
e) -2 points: The current estimated ventilation mode is 4 (NIV), 5 (High flow) or 6 (CPAP)
f) +1 point: Estimated TV is >0
g) +2 points: Tracheotomy indicator (vm313) or Intubation indicator (vm312) or Airway category (vm66) is 'Intubated' or 'Tracheotomy'
h) -1 point: No airway, Airway category (vm66) is 'Maske', 'Helm', 'Mundstueck', 'Nasenmaske'.

If the combined score is at least 4, at the grid point the ventilation status is 'True', 'False' otherwise.

Thereafter post-processing was applied

1) Remove gaps due to likely HR sensor disconnections, which could be caused by the patient leaving the ICU for imaging/operation and other procedures. If a gap in ventilation is observed and during the gap for fewer 50 % of the grid points, the following condition was true: At least one HR observation in the 10 minutes around the grid point exists, the gap is closed and the patient is assumed to be on mechanical ventilation, to avoid artificial gaps due to procedures.
2) Gaps of <15 minutes between successive ventilation events are closed, and the events merged together.
3) Gaps of <24 hours were closed, in case patient had a tracheotomy indicator just before and after the event.
4) Thereafter short ventilation events of length <45 minutes are deleted, if they do not occur at the beginning of the stay (i.e. no HR was recorded before the event), as these are likely to be spurious detections.

Readiness to extubate annotation

Readiness to extubate status was only annotated for time points where the patient was ventilated according to the criteria mentioned above. It was informed by medical prior knowledge and at each time-point the number of violations of commonly accepted extubation criteria were counted, to form a scoring system, as follows

1) If the ventilator mode is not 3 (spontaneous breathing), the patient cannot be extubated, before 2010 this criterion was not applied because the ventilator mode is sometimes incorrectly recorded in the data. A violation score of +9 is assigned.
2) If the current PEEP is >7, a violation score of +3 is assigned.
3) If the current pressure support is >10, a violation score of +3 is assigned
4) If the current FiO_2 is >0.4, a violation score of +3 is assigned
5) If the current TV is >0 and the rapid shallow breathing index (1000*RR/TV) is at least 105, a violation score of +3 is assigned
6) If the current RR is at least 35 breaths/minute, a violation score of +3 is assigned
7) If the current MV (Minute volume) is at least 10, a violation score of +3 is assigned
8) If the current P/F ratio (as estimated using the annotation algorithm for respiratory failure) is <= 150 mmHg, a violation score of +3 is assigned
9) If the current $PaCO_2$ is at least 50, a violation score of +3 is assigned
10) If GCS is <= 8, a violation score of +1 is assigned
11) If the current MAP is <= 60, a violation score of +1 is assigned
12) If the standardized dose of Norepinephrine is >0.05/kg or any dose of epinephrine/dobutamine/milrinone/levosimendan/theophyllin/vasopressin is given, a violation score of +1 is assigned
13) If current lactate is at least 2.5, a violation score of +1 is assigned

If the summed violation score from the 13 criteria at a time-point is <9, the patient is assumed (tentative) to be ready to be extubated, otherwise they are not ready to be extubated. To increase robustness of annotation, a backwards window of length 1 hour is used, and the patient is assumed to be ready to extubated only if for ⅔ of time-points in the last hour, they satisfied the violation criteria. The coefficients of the scoring system were obtained by fitting a model to

predict extubation failure from the input variables, and then rounding the coefficients to be integer. The threshold 9 was picked by visually inspecting the time series annotated with integer scores, against clinical plausibility.

Extubation failure task

For the purpose of extubation failure prediction, extubations which are from tracheotomy are not considered. An extubation is defined as a transition from ventilated to non-ventilated status, where the annotation algorithm for ventilation detection was used. The label for extubation failure was defined as positive, if the patient was re-intubated within the next 48 hours after the extubation event. The re-intubation was ignored if there is a HR measurement gap of ⅔ of the hour immediately prior to the re-intubation, which might indicate that the re-intubation was due to a procedure performed on the patient. If a valid reintubation occurred, the label for extubation failure was positive, otherwise negative. If the patient however died within the next 48 hours after extubation, and no re-intubation occurred, the label is treated as uncertain. Sample augmentation was used for training set/evaluation in the near vicinity of extubations, the prior 30 minutes before an extubation share the same label (i.e. extubation failure or no extubation failure) as the exact time-point of the extubation. In this way, the number of training samples is increased, and a clinically reasonable assumption is made that the physiological state reflecting likelihood of extubation failure does not change within a time span of 30 minutes.

Respiratory failure onset task

We are interested in predicting onset of oxygenation failure of only the moderate/severe level as previously defined. The machine learning label is only defined at time-points where the patient is not already in respiratory failure (P/F ratio <200 mmHg) and the annotation is not ‘unknown’. If they are currently stable or in mild respiratory failure, but will have at least moderate/severe respiratory failure at some point in the next 24 hours, the label is positive, otherwise negative. The label is undefined if the respiratory status is ‘unknown’ for the complete next 24 hours or at the current time-point.

Ventilation onset task

We are interested in predicting onset of mechanical ventilation, where ventilation presence was annotated using the score-based algorithm presented earlier. The machine learning label is only defined at time-points where the patient is not already ventilated and the annotation is not ‘unknown’. If they are currently not ventilated, but are mechanically ventilated at some point in the next 24 hours, the label is positive, except the offset is in the next 30 minutes from the current time-point. In the latter case, the time point is excluded from training/evaluation, to prevent any potential leakage of information from the future. The label is negative if the patient is not mechanically ventilated within the next 24 hours. The label is undefined if the ventilation status is ‘unknown’ for the complete next 24 hours, or at the current time-point.

Readiness to extubate onset task

We are interested in predicting a patient becoming newly ready to extubated, when they are currently not. The machine learning label is only defined at time-points where the patient is mechanically ventilated and is not currently ready to extubate, where readiness to extubate was annotated using the score-based algorithm presented earlier. If they are currently not ready to extubate, but are ready to extubate at some point within the next 24 hours, the label is positive, otherwise it is negative. The label is undefined if the readiness to extubate status is ‘unknown’ for the complete next 24 hours, or at the current time-point.

Feature extraction

To give our model a comprehensive view of the patient state the following feature classes were extracted on the clinical parameters available in the HiRID-II database.

- **Current value:** The current time grid value of the clinical parameters in the HiRID-II database was used directly as a feature.
- **Time since admission:** The time since admission was used as an individual feature.

- **Endpoint annotation variables:** The current estimated value of FiO_2 and the current ventilation status, as computed by the endpoint algorithm, were used as additional clinical variables. The current PaO_2 estimate was not used to avoid potential leakage of information from the future.
- **Multi-resolution summaries:** Various summary functions were computed over multiple horizons, including the last 10 hours, the last 26 hours, the last 63 hours, and the last 156 hours. These 4 horizon lengths correspond to the 20/40/60/80 percentiles of the available history across all time points in the training set. From the training set the expected number of measurements within the horizon was estimated, using the median observation interval of the parameter. If the expected number of measurements in the horizon is less than 5, the horizon is not used for feature computation. For ordinal variables, median/IQR/trend were used as the 3 summary functions. The trend is defined as the slope of a regression line fit over the values in the horizon. For binary variables, the mean was used as the only summary function. Note, for binary variables, the mean can be also interpreted as the proportion of the horizon in which a certain condition was true. For categorical variables, the mode was used as the only summary function. All 4 horizons were computed only for important variables which were determined using a preliminary variable importance selection step. The important variables are listed in Supplemental Table 3. For other variables, only the shortest horizon of the 4 horizons, for which the expected number of measurements exceeded 5, was used.
- **Measurement intensity:** The time to last real measurement was computed as a feature. If there was no such measurement, this feature was set to a large symbolic value. The measurement density was computed over the same multi-resolution horizons that were used for the last feature category. The measurement density is defined as the number of observations in the horizon divided by the horizon length.
- **Instability history:** If applicable for a variable, up to 3 severity levels were annotated using prior medical knowledge. The fraction of time spent in each severity state over the last 8 hours as well as over the entire stay up to now was extracted. This schema was used only for a subset of variables, which are among the important variables selected in a preliminary variable selection step. The severity levels and variables used for this feature class are listed in Supplemental Table 4.
- **Static variables:** Static variables are the same for all time-points of the patient time series and are finally concatenated to the feature vector. As static variables, the patient age, APACHE patient group, Patient group, gender, APACHE code, Emergency admission status, Surgical admission status, and height were used.

Variable selection

Variable selection was performed in a 2-step process, which used only the development set, and not the held-out test set.

a) Separately, for the 4 tasks, Respiratory failure/Extubation failure/Ventilation onset/Readiness to extubate, the 20 most important variables in terms of SHAP value magnitude were pre-selected on the validation set. The ‘SHAP importance’ of a feature was defined as the mean (over the 5 temporal splits) of the mean absolute SHAP value on predictions in the validation set, for that variable. The ‘SHAP importance’ of a variable was defined as the maximum of SHAP importance over the features derived from the clinical variable. In this way 20 variables are extracted per task. The union of the variables selected for the 4 tasks formed the initial set of ‘important variables’, which consisted of 31 variables, which are listed in Supplemental Table 3.
b) For the initial set of important variables, more complex features were computed, according to the description in the section on feature extraction above.
c) For the 2 main tasks presented here, Respiratory failure/Extubation failure, variables were greedily selected forward from the final set of complex features on 31 variables. In each step the variable was chosen, which yielded the highest time-point based AUPRC on the validation set, among the candidate variables to be added. The output of this procedure, which was run 5 times, per temporal split, is a forward trace of 31 variables, ranked by importance. The final importance of a variable was defined as the mean reciprocal rank over the 5 splits, yielding a ranked list of 31 variables. The extubation failure model used the top 20 variables, which included both medication and non-medication variables, and the respiratory failure model used the non-medication variables among the top 20 variables, which yielded a final set of 15 variables. For ventilation onset/readiness to extubate models, the union of the variables used for the respiratory failure and extubation failure models was used. The variables used in each model are listed in Supplemental Table 3.

Model training

The generated features for the variable sets of the RMS-RF (15 variables) and RMS-EF (20 variables), RMS-VENT/RMS-REXT (26 variables) predictors were passed to 4 gradient-boosted decision tree ensembles implemented in LightGBM (https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html), with one separate model per task. As LightGBM is robust to missing data, data was not imputed or standard scaled prior to training. Trees were added to the ensemble until performance did not improve for 50 epochs in the validation set, early stopping the training process. As a criterion guiding the early stopping the time-point based AUPRC on the validation set was used. Hyperparameters were optimized on the validation set, using the time-point based AUPRC as a criterion. Hyperparameters were fixed for RMS-RF/RMS-VENT/RMS-REXT, as experiments showed that early stopping was enough to find a configuration close to optimal. For RMS-EF, which has a small number of training set samples, a hyperparameter grid with 20 points was used, to select the optimal model. The parameters of the LightGBM model and the hyperparameter selection grid for RMS-EF are listed in Supplemental table 5. Prediction scores were generated for patients in the test set, using a separate LightGBM model for each of the 4 tasks. To allow more flexible evaluation, for resource planning and joint task analysis using t-SNE, predictions were generated at all time-points of test set patients, even when the ground-truth label is undefined, i.e. for RMS-RF, while the patients are already experiencing respiratory failure.

Resource planning

Mechanical ventilation planning helps ICU clinicians to better prepare for resource allocation and improve outcome. We develop a resource planning tool for mechanical ventilation that can predict how many mechanical ventilators will be needed in certain time-windows in the short-term future, which entails both new mechanical ventilation need for patients who are in the ICU and new patients who are admitted for emergency reason and in need of ventilation. To predict number of in-ICU patients who need ventilation in the near future, we use scores output from the machine learning models trained for predicting the four respiratory system related tasks, namely respiratory failure, mechanical ventilation need, readiness to extubate and the extubation failure as well as ICU-level information as features for a LightGBM model. The ICU-level information includes the hour of the day, weekday, number of patients who are already on mechanical ventilator in the past hour. To predict newly admitted emergency patients who need mechanical ventilators, we train a LightGBM model using only the ICU-level information.

Model calibration

For calibration evaluation, the prediction scores on the time-points in the test set where the label was defined were gathered. The scores were then binned between the minimum prediction score and maximum prediction score observed in the test set with a bin size of 0.05. The actual observed risk (proportion of true labels for time-points in the bin) was then computed per bin and plotted against the bin location. As evaluation metrics of calibration the Brier score was used. Models showed sufficient calibration using the raw scores, so re-calibration using isotonic regression was not needed.

Prevalence correction for external validation

Since the prevalence of positive events is different between the HiRID-II and the UMCdb test set, we correct the precision-recall curves for the performance on the test set of UMCdb such that the corrected prevalence matches with that in the HiRID-II test set by downscaling the false alarm number using the scaling factor s = (1/prev(HiRID-II)-1)/(1/prev(UMCDB)-1), as used by Hyland et al. [5]

Extubation-based evaluation (extubation failure)

Extubation failure was assessed using recall (percentage of extubation failures which were correctly predicted), and precision (percentage of extubation failure predictions which are correct, i.e. re-intubation occurs in the next 48h), yielding a PR curve, as well as recall/false positive rate, which defines the ROC curve.

Event-based evaluation (respiratory failure)

We used the same event-based evaluation scheme used by Hyland et al.[5], that measures the fraction of correctly predicted respiratory-failure events and the fraction of false-alarms.

External validation + prevalence correction

To allow external validation, the most important parameters for training predictive models in HiRID-II were matched to variables in the Amsterdam UMCdb database[1]. To enable endpoint annotation at a similar granularity as in HiRID-II, a subset of patients with high time resolution for respiratory parameters (n=6,698 patients) was included. A data set was then prepared by applying the same endpoint annotation and feature extraction pipeline as for HiRID-II. Because admission times are not exactly available in UMCdb, data were randomly divided by patient into a fixed test set containing 1674 patients, and a development dataset of 5024 patients. To retrieve variation estimates of performance, the development dataset was partitioned five times into training and validation sets. For external validation, we applied the models trained on HiRID-II dataset to the UMCdb dataset, and used prevalence correction to re-scale the false positive count, as described in the section on 'Prevalence correction'.

Sub-cohort / Fairness analyses

Patients grouping. We group patients by demographic characteristics (sex, age) but also clinical characteristics (APACHE admission group). For binary grouping, we compare one group versus the other, while for multi-categorical grouping we compare patients belonging to a group to all the other patients.

Patients bootstrapping. Due to the small number of patients composing certain cohorts of patients, we rely on bootstrapping. We create 100 bootstrap samples of the patients from the test set (i.e. we sample randomly with replacement the patients composing the test set). We can then compute for each bootstrap sample and for each of our patient cohorts the different performance metrics. Having several bootstrap samples to perform the analysis on, allows us to have an idea of the variability of the patients within each cohort and counterbalance for the small size of certain cohorts. At the end of the process, we have a distribution of each metric for each cohort. We give a graphical representation of this distribution for each group through box plots where the median, first quartile and third quartile over the 100 bootstrap samples are outlined.

Statistical testing. For each evaluated metric, we want to measure whether it is similar between groups of patients. We assume that the samples from one cohort of patients to the other are independent. However, we don't assume normality of the distribution. In order to compare the distribution of our metrics across different cohorts, we use the Mann-Whitney U test with a significance level of 0.1%. Since for each grouping we are performing multiple tests, we correct the p-value with Bonferroni correction. We test whether patients from a certain group are significantly worse off (according to the fairness metric) compared to patients not belonging to this group. On the graphical representation, we mark the groups that are significantly worse with a star.

Metrics used

For the respiratory failure task, we compute the precision at:

- 80% event-based recall for each cohort (the threshold will thus be different for each cohort)
- 90% event-based recall for each cohort (the threshold will thus be different for each cohort)

For the extubation failure task, we compute the precision at:

- 80% recall for each cohort (the threshold will thus be different for each cohort)
- 20% recall for each cohort (the threshold will thus be different for each cohort)

Finally, for both tasks, we also compute the corrected event-based AUPRC.

Model inspection using SHAP values

SHAP values for the positive class were extracted using a 'SHAP tree explainer' built for LightGBM ensembles (https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html), in the validation set and test set. In the validation set the mean absolute SHAP values of each feature were used to create an initial set of important variables for variable selection (refer to the section on 'Variable selection' above). In the test set the signed SHAP values (Interpretation: Large SHAP value means the feature value contributes to an increase of the prediction score)

were used to interpret the model's prediction, i.e. by plotting them against the feature value at the time-point the prediction is made.

Joint task analysis using t-SNE

For the joint task analysis the test set predictions for the 4 tasks (RMS-RF/RMS-EF/RMS-VENT/RMS-REXT). In principle, predictions were available at all time points in the patient stay for all tasks, irrespective of whether the label of the task is defined at the time-point. For computing the t-SNE embedding, only the current value features of 16 clinical parameters were used, which correspond to the union of the top 10 important variables for the RMS-RF/RMS-EF models respectively. As t-SNE requires dense input without missing values, the 'dense imputation' data (refer to the section on Imputation for its definition), was used as input to t-SNE. Prior to fitting of t-SNE the data was standard scaled such that each dimension has mean 0 and standard deviation 1, such that all variables have equal importance in the t-SNE input space. A random subsample of 150,000 time-points in the test set was drawn to ensure fast fitting of the embedding algorithm. The t-SNE was computed once, independent of task, as it only depends on the clinical parameters but not the prediction scores of the 4 models. For fitting t-SNE, the implementation available in the Python package scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html) was used with default parameters, and target dimension 2. In the t-SNE plots, only hexes with at least 30 assigned time-points are displayed, ignoring very rarely used parts of the embedding space.

Statistical methods

In result plots the solid curves refer to the mean of the performances obtained in the five experimental replicates, corresponding to the five temporal splits, as described in the section on data splits. Light shaded regions refer to the standard deviation of performances obtained in the five experiment replicates.

Data availability

More information on HiRID is available on hirid.intensivecare.ai. The newly curated data-set HiRID-II will be released on Physionet in the near future. The computer code used in this research is available at www.github.com/ratschlab/RMS under an open-source license.

References

1. 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).
2. Faltys, M., Zimmermann, M., Lyu, X., Hüser, M. & Hyland, S. HiRID, a high time-resolution ICU dataset (version 1.1. 1). *Physio. Net* (2021).
3. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
4. Amsterdam Medical Data Science. AmsterdamUMCdb website and documentation. https://amsterdammedicaldatascience.nl/#amsterdamumcdb.
5. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).
6. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information*

Processing Systems 30 3146–3154 (2017).

7. Brown, S. M. *et al.* Nonlinear Imputation of PaO2/FIO2 From SpO2/FIO2 Among Mechanically Ventilated Patients in the ICU: A Prospective, Observational Study. *Crit. Care Med.* **45**, 1317–1324 (2017).
8. Ellis, R. K. Determination of PO2 from saturation. *J. Appl. Physiol.* **67**, 902 (1989).
9. Severinghaus, J. W. Simple, accurate equations for human blood O2 dissociation computations. *J. Appl. Physiol.* **46**, 599–602 (1979).
10. Bernard, G. R. *et al.* The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am. J. Respir. Crit. Care Med.* **149**, 818–824 (1994).
11. ARDS Definition Task Force *et al.* Acute respiratory distress syndrome: the Berlin Definition. *JAMA* **307**, 2526–2533 (2012).