ARTICLES
https://doi.org/10.1038/s41591-020-0789-4

nature medicine

Check for updates

Early prediction of circulatory failure in the intensive care unit using machine learning

Stephanie L. Hyland[1,2,3,4,10], Martin Faltys[5,10], Matthias Hüser[1,4,10], Xinrui Lyu[1,4,10], Thomas Gumbsch[6,7,10], Cristóbal Esteban[1,4], Christian Bock[6,7], Max Horn[6,7], Michael Moor[6,7], Bastian Rieck[6,7], Marc Zimmermann[1], Dean Bodenham[6,7], Karsten Borgwardt[6,7,11] ✉, Gunnar Rätsch[1,2,3,4,7,8,11] ✉ and Tobias M. Merz[5,9,11] ✉

Intensive-care clinicians are presented with large quantities of measurements from multiple monitoring systems. The limited ability of humans to process complex information hinders early recognition of patient deterioration, and high numbers of monitoring alarms lead to alarm fatigue. We used machine learning to develop an early-warning system that integrates measurements from multiple organ systems using a high-resolution database with 240 patient-years of data. It predicts 90% of circulatory-failure events in the test set, with 82% identified more than 2 h in advance, resulting in an area under the receiver operating characteristic curve of 0.94 and an area under the precision-recall curve of 0.63. On average, the system raises 0.05 alarms per patient and hour. The model was externally validated in an independent patient cohort. Our model provides early identification of patients at risk for circulatory failure with a much lower false-alarm rate than conventional threshold-based systems.

Critical illness is characterized by the presence or risk of developing life-threatening organ dysfunction. Critically ill patients are typically cared for in intensive care units (ICUs), which specialize in providing continuous monitoring and advanced therapeutic and diagnostic technologies.

ICU physicians are presented with large quantities of data from many patients stored in electronic patient-data management systems (PDMS), and it is increasingly difficult to identify the most important information for care decisions. The limited ability of humans to process such quantities of information can lead to data overload, change blindness and task fixation[1]. This increases the risk that clinicians do not readily recognize, interpret and act upon relevant information[2,3]. ICU patients are not continuously supervised or assessed by ICU nurses or physicians. Low nurse-to-patient ratios and lack of intensivist presence have been associated with delays in establishing adequate treatment for deteriorating patients[4] and with higher ICU mortality[5–7].

Circulatory failure is common during critical illness, and monitoring of circulatory function is therefore an essential part of ICU patient management. The effects of circulatory failure are initially reversible in most patients[4,8,9], but repeated or prolonged episodes of hypotension and high-dose vasopressors may worsen the prognosis[10–12]. Care providers intermittently assess monitored vital signs and rely on alarms for individual physiological measurements to identify patients at risk of deterioration. These alarm systems do not utilize comprehensive patient information, and alarms are therefore often non-specific[13,14], leading to alarm fatigue[15], which was rated seventh on the list of top ten technology hazards from the ECRI Institute[16,17].

Machine-learning (ML) techniques excel in the analysis of complex signals in data-rich environments[18–20]. The abundance of data collected in the ICU and public availability of datasets such as MIMIC-III[21] and eICU[22] are key for developing the use of machine learning in this setting. Endpoints such as patient mortality[23] and length of stay (LOS)[24] are commonly tackled using predictive models. However, the accurate prediction of mortality or LOS is not of great importance for further treatment decisions after the initial decision to admit a patient to ICU. The prediction of events related to circulatory deterioration has been addressed by predicting ICU admission[25], the onset of treatment[26] or near-term mortality[27], and more specific aspects such as hypotension[28] and vasopressor use[29].

In this work, we develop a new approach to detect circulatory failure in ICU patients on the basis of medical knowledge, large-scale data analysis and state-of-the-art ML techniques. We construct two early-warning systems, named circEWS and circEWS-lite, that are of differing complexity and alert clinicians to patients at risk of circulatory failure within the next 8 h. We define a patient as being in circulatory failure if (1) arterial lactate is elevated ($\geq$2 mmol l^{-1}), and (2) either mean arterial pressure (MAP) $\leq$ 65 mmHg, or the patient is receiving vasopressors or inotropes (Extended Data Fig. 1a shows an example). We use the patient database from a large multidisciplinary ICU, containing routinely collected data from more than 54,000 ICU admissions, to train the early-warning systems. We have developed a comprehensive analysis framework including data preprocessing and cleaning, feature extraction and interpretation, and selection of large-scale supervised ML techniques to construct the early-warning systems.

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. [2]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA. [4]Medical Informatics Unit, Zürich University Hospital, Zürich, Switzerland. [5]Department of Intensive Care Medicine, University Hospital, University of Bern, Bern, Switzerland. [6]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland. [7]Swiss Institute for Bioinformatics, Lausanne, Switzerland. [8]Department of Biology, ETH Zürich, Zürich, Switzerland. [9]Cardiovascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand. [10]These authors contributed equally: Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch. [11]These authors jointly supervised this work: Karsten Borgwardt, Gunnar Rätsch, Tobias M. Merz. ✉e-mail: karsten.borgwardt@bsse.ethz.ch; gunnar.raetsch@inf.ethz.ch; tobiasm@adhb.govt.nz

To evaluate the performance of our systems, we established an alarm/event-based evaluation measure, which assesses the fraction of circulatory-failure events correctly predicted (that is, an alarm was raised for this event) and the false-alarm rate (that is, there was an alarm but no event). For external validation, we applied and tested our circEWS-lite system to the MIMIC-III database.

Results

Preparation of a high time resolution ICU dataset (HiRID). The full dataset contained a total of 7,333 routinely collected physiological variables, diagnostic test results and treatment parameters from 55,602 patient admissions to the ICU, resulting in more than 3 billion data observations. Continuous measurements are recorded every 2 min; therefore, the dataset contains more observations at a higher temporal resolution than the 2 publicly available ICU datasets (MIMIC-III[21], 312 million; eICU[22], 827 million observations)[30]. After applying exclusion criteria (Extended Data Fig. 2a,c), information on 710 variables from 36,098 patient admissions collected between January 2008 and June 2016 remained for further processing (Extended Data Fig. 3). There were 209 consistently measured variables used for model development after data merging by aggregating pharmaceutical variables and summarizing measurement modalities of physiological variables (Fig. 1a, Extended Data Fig. 2c and Supplementary Fig. 1). The data was resampled to a 5-minute resolution using adaptive imputation (Fig. 1b). The patient's circulatory state was annotated for each time point as 'circulatory failure' or 'no circulatory failure' (Fig. 1c). For 36.5% of time points, the annotation was not possible owing to lack of MAP or lactate measurements. These states were annotated as 'ambiguous'. Overall, we identified 45,886 circulatory-failure events in 11,046 patients, with mean event duration of 320 min. We found that ICU mortality correlated with longer duration and a higher number of events of circulatory failure (Extended Data Fig. 4).

Development of a continuous risk score for prediction of circulatory failure. Every 5 min, we aimed to determine the risk of a patient developing circulatory failure within the next 8 h using a continuous score. The features generated for the prediction task included static patient information, multi-scale summaries of time-series history, measurement intensity of variables and shapelet patterns (Fig. 1d–f and Supplementary Tables 1–3). We used SHAP (SHapley Additive exPlanations) values[31] to assess the influence of each feature on the classifier output: positive and negative SHAP values indicate an increase or decrease of the prediction score, respectively. The dataset contained 15 million time slices, of which 3.1% were labeled positive (circulatory failure within the next 8 h). An analysis framework for feature and model selection was developed. Among tested classifiers, gradient-boosted ensembles of decision trees (lightGBM[32]) offered the best performance (Extended Data Fig. 5c,d).

Two lightGBM classifiers with differing complexity were developed—the full and compact models. There were 5,278 features generated from the 209 variables in the HiRID data that were ranked according to mean absolute *SHAP* value[31], which indicates their importance for predictions. The full model uses the top 500 features, originating from 112 variables (Supplementary Table 4), and the compact model uses 176 features from 16 variables among the top 20 variables (Table 1) that are available in both the HiRID and MIMIC-III datasets. As a baseline, we developed a decision tree using only the last measurement of the variables included in the circulatory-failure definition (MAP, lactate and dose of vasopressors/inotropes), mimicking a threshold-based rule system. The areas under the receiver operator curves (AUROCs) of the full, compact and baseline models were 0.940, 0.939 and 0.883, respectively (Fig. 2a). For rare events, as in our case, predictions with high precision are more difficult to obtain than low false-positive rates. Hence, the areas under the precision-recall curves (AUPRCs) are more informative and were 0.467, 0.454, and 0.254 for the full, compact and baseline models, respectively (Extended Data Fig. 6a). The continuous scores are a good proxy for time to failure (Extended Data Fig. 7a,b). After re-calibrating the continuous scores post hoc using isotonic regression, we obtained almost ideal concordance between the score and the observed risk (Extended Data Fig. 7c), with an overall Brier score[33] of 0.02 and an area of 0.04 around the ideal calibration curve. We also tested calibration in various patient subgroups, and found that the model is well calibrated for most large patient subgroups, except patients with neurological conditions (see Supplementary Table 5).

The circulatory early-warning system. Our models generate a continuous prediction score every 5 min regarding the risk of circulatory failure within the next 8 h. A threshold-based warning system derived from this score could lead to alarms every 5 min, causing alarm fatigue. We therefore developed an alarm system that implements a silencing policy: once an alarm is triggered, subsequent alarms are suppressed for 30 min (Fig. 1g). The system is reset if the patient experiences circulatory failure and recovers. The effects of different factors in the circEWS alarm system are shown in Supplementary Tables 6–8. We applied this alarm algorithm to the predictions of our full and compact models, and named the resulting systems circEWS and circEWS-lite. The performance of the two systems is shown in Fig. 2b, using precision-recall curves[34] (PRC). Recall was defined as the fraction of events with any alarm in the preceding 8 h, and precision as the fraction of alarms which correctly predicted an event. Precision and recall measure performance on the raised alarms and occurring events and are clinically more meaningful than time point-based measures. We observed circEWS and circEWS-lite significantly outperforming the baseline (Extended Data Fig. 6b), also in a reclassification analysis (Supplementary Table 9, $P < 0.05$). We analyzed the number of

Fig. 1 | Model development overview. a–c, Data preparation. **a**, Data on patient admissions were exported from the ICU PDMS and filtered according to the inclusion and exclusion criteria. Clinically implausible values, variable-specific errors and other artifacts were automatically excluded using variable-specific algorithms. Variables coding the same active drug component with differing administration methods were merged to obtain the effective drug dose over time. Different monitoring modalities for the same parameter were merged. **b**, Adaptive imputation was performed by filling missing values using a patient and variable-specific imputation scheme to obtain a data point for every variable and time point in a 5-min time grid. **c**, The circulatory state was annotated according to the definition of circulatory failure. **d–f**, Machine learning. **d**, At each time point, four feature types (measurement intensity, multi-resolution summaries, instability history and shapelets) per variable were extracted. These feature types, together with static patient information representing patient characteristics, represented the data available at any given time point. **e**, To construct the binary prediction problem, each relevant time point was labeled as either 'positive' (circulatory failure (CF) will occur within the next 8 h) or 'negative' (circulatory failure will not occur within the next 8 h). **f**, A binary classifier to predict near-term circulatory failure was trained on the extracted labels and features. A gradient-boosted ensemble of decision trees was chosen as the classifier after comparison of different machine learning algorithms. **g**, Evaluation. The proposed early-warning system for circulatory deterioration (circEWS) consists of the trained binary classifier, a decision threshold and a policy of silencing for a short period after alarms. The system was evaluated on the basis of the fraction of alarms that are correct (precision), and the fraction of circulatory failure events that are correctly predicted (recall). CF, circulatory failure; HR, heart rate; IQR, interquartile range; MAP, mean arterial pressure; T, temperature.

alarms of a threshold-based alarm system with alarms triggered by abnormal values in key circEWS-lite variables (Supplementary Table 10) and found that 20–80 times more alarms were generated than by circEWS-lite at the same recall rate (Extended Data Fig. 6c).

The recall rate as a function of time before the occurrence of circulatory failure for fixed overall recall and precision is shown in Fig. 2c. We observe an increase in recall closer to the onset of circulatory failure, with 81.8% of the events identified more than 2 h in advance. The timeliness of circEWS alarms is illustrated in Fig. 2d, showing the temporal distribution of the first alarms and number of alarms in the 8-h window prior to deterioration. Considering the standard 8-h working shift common in the ICU, this would result in less than 1 alarm per patient per shift occurring, on average, 2 h and 32 min before circulatory failure. The effect of training set size on model performance was assessed by artificially subsampling patients at random and retraining the model. Model performance

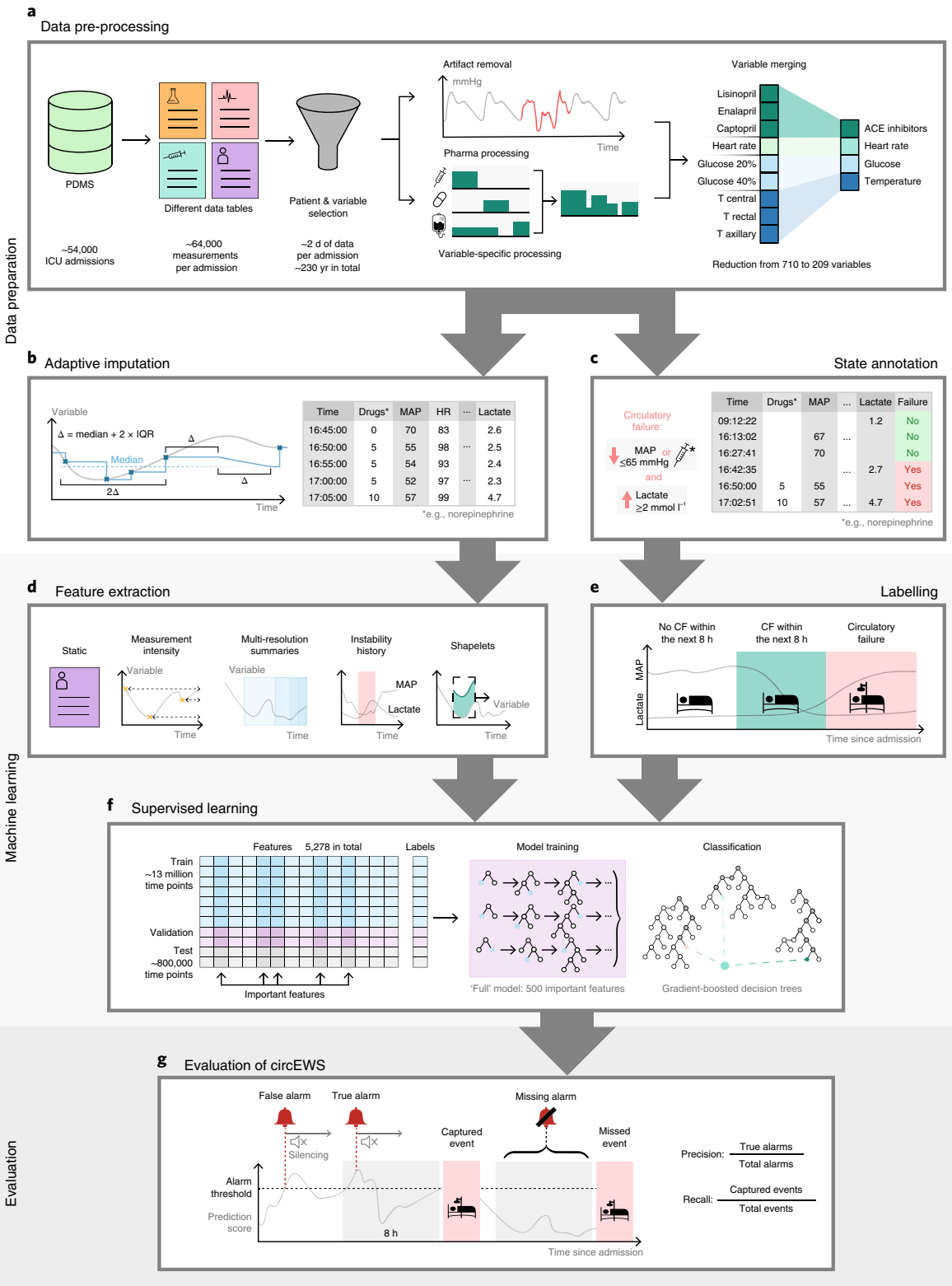


Table 1 | Top 20 ranked variables for the prediction of circulatory failure

Rank (std)	Variable	Important feature categories
1 (0.0)	Lactate	Current, Shapelet, Multi-resolution, Instability history, Measurement
2 (0.0)	MAP	Multi-resolution, Instability history, Current, Shapelet, Measurement
3 (5.3)	Time since ICU admission	N/A
4 (0.4)	Patient age	Static
5 (3.0)	Heart rate	Current, Multi-resolution, Measurement, Shapelet
6–9 (2.3)	Dobutamine, milrinone, levosimendan[a], theophylline[a]	Instability history, Multi-resolution
10 (5.3)	Cardiac output	Shapelet, Multi-resolution, Measurement
11 (3.5)	RASS	Current, Multi-resolution, Measurement
12 (34.6)	INR	Measurement, Multi-resolution, Current
13 (5.8)	Serum glucose	Multi-resolution, Current, Measurement
14 (4.4)	C-reactive protein	Multi-resolution, Current, Measurement
15 (7.9)	Diastolic BP	Multi-resolution, Shapelet, Measurement
16 (4.0)	Peak inspiratory pressure (ventilator)	Current, Measurement, Multi-resolution, Shapelet
17 (7.9)	Systolic BP	Current, Multi-resolution, Measurement, Shapelet
18 (10.6)	SpO_2	Multi-resolution, Shapelet, Measurement
19 (17.8)	Non-opioid analgesics[a]	Multi-resolution
20 (11.4)	Supplemental oxygen	Multi-resolution, Measurement, Current

The ranking was obtained by first ranking all 5,278 features according to their importance in explaining predictions of the development model and then greedily selecting clinical variables in a forward-selection procedure if they contribute to important features derived from these variables. The point estimate of the rank is obtained on the held-out data split, and the standard deviation of the rank was obtained on $n=5$ development splits of the data. The last column lists the important feature categories for a variable, that is, the feature categories that contribute to the top 50 features overall. The categories are sorted by decreasing importance in terms of rank in the list of top features. [a]Variables not contained in MIMIC-III; these were not used in the compact model (and hence not in the circEWS-lite system), as they appear to be less commonly available. BP, blood pressure; RASS, Richmond Agitation Sedation Scale; INR, international normalized ratio (prothrombin time); N/A, not applicable; SpO_2, peripheral oxygen saturation.

decreased with subsampling and did not show obvious saturation effects, even at the full size of the data (Extended Data Fig. 5a,b).

Model performance in different patient cohorts. In all subsequent analyses with fixed thresholds, we chose a threshold leading to a recall of 90.0%, resulting in a precision of 29.6% for circEWS.

We evaluated the performance of circEWS in different cohorts with varying age, gender, severity of illness, acute physiology and chronic health evaluation (APACHE) diagnostic groups, and compared medical and surgical as well as elective and emergency admissions (Fig. 3a–c and Extended Data Fig. 8a,b). We found similar performance across most diagnostic groups, with the exception of patients with neurological conditions, for whom the model performs worse, with an event recall of 76.6% (compared with 91.2% across other patients). For neurosurgical patients, the model exhibits a decreased precision of 8.1% compared with 30.0% in the rest of the patients. Patients with lower APACHE scores (0–15) have a lower precision of 19.7% compared with 30.5% in the rest of the cohort, which might be explained by their lower event prevalence. Emergency admissions have a lower recall of 88.2% compared with 93.6% for elective admissions, whereas surgical admissions have a higher recall of 92.3% compared with 87.7% in the rest of patients (Extended Data Fig. 8b). The recall rate increases with respect to the length and time since the last event (Extended Data Fig. 8c,d).

Model performance over time. Figure 3d shows how the performance of the model varies as time since admission increases. While the overall recall of the model is 90%, the performance is not uniformly distributed across a typical patient's stay, with recalls of over 95% attained within the first 8 h. After the first day, the overall recall of the model drops to around 83%. Using our dataset spanning 8 years, we analyzed how changes in medical practice and patient characteristics may impact model performance in the future. We simulated this by fixing a test set comprised of patients admitted in 2016, and 8 training sets for patients admitted in each of the years from 2008 to 2015, with each training set subsampled to have the same size (2,366 patients). In Fig. 3e, we report the performance of these eight models in terms of AUPRC and precision at fixed recalls (AUROC shown in Extended Data Fig. 8e). We observe an increase in performance the closer in time the test set is to the training set (Fig. 3e). When fitting an autoregressive model to the differences of the AUPRC values, we obtain a first-order term of size 0.14, which we interpret as the presence of a temporal drift[35]. This does not hold for the precision values (Fig. 3e), for which we can assume stationarity ($P = 5 \times 10^{-5}$, $n=8$ years, Dickley–Fuller test[35]).

Inspection of model features. In Fig. 4a, we list the top 15 features by mean absolute SHAP value and show the distribution across all predictions. Features from variables used to define circulatory failure rank highest. The relationship between feature value and SHAP value is illustrated in more detail for the features patient age and MAP in Fig. 4b,c, with further examples in Extended Data Fig. 9. Table 1 reports the 20 most relevant variables (a subset of them is used to define circEWS-lite). When removing each of these variables in turn, only the removal of lactate noticeably decreased performance (resulting AUPRC, 0.411 ± 0.037). Greedy forward selection of variables guided by performance on the validation set confirms lactate and MAP being the most important variables, as is also observed in the analysis based on SHAP values. The model performance begins saturating after adding around ten variables (Supplementary Table 11). Figure 4e shows the highest-ranking lactate shapelet as an example from the shapelet feature class, illustrating that the SHAP value of this feature increases 5.5 h before the onset of deterioration. While these analyses show the overall effect of the features, SHAP values can also be inspected for individual predictions to identify the influential features (see Extended Data Fig. 1b,c).

Training predictive models on observational data is associated with the risk that changing patient management patterns, such as additional monitoring modalities and/or frequency in anticipation of impending circulatory failure, are included as important model features. Such a model may perform well on the previously observed data, but will fail to generalize to scenarios with differing clinician behavior[36]. In Supplementary Table 12, we demonstrate that removing measurement-intensity-based features results in a drop in AUPRC of 0.024. To further assess the degree to which circEWS may suffer from this fragility, we analyzed the circEWS-lite model performance on an artificially re-sampled test set with regular measurement intervals per variable. The measurement intervals were

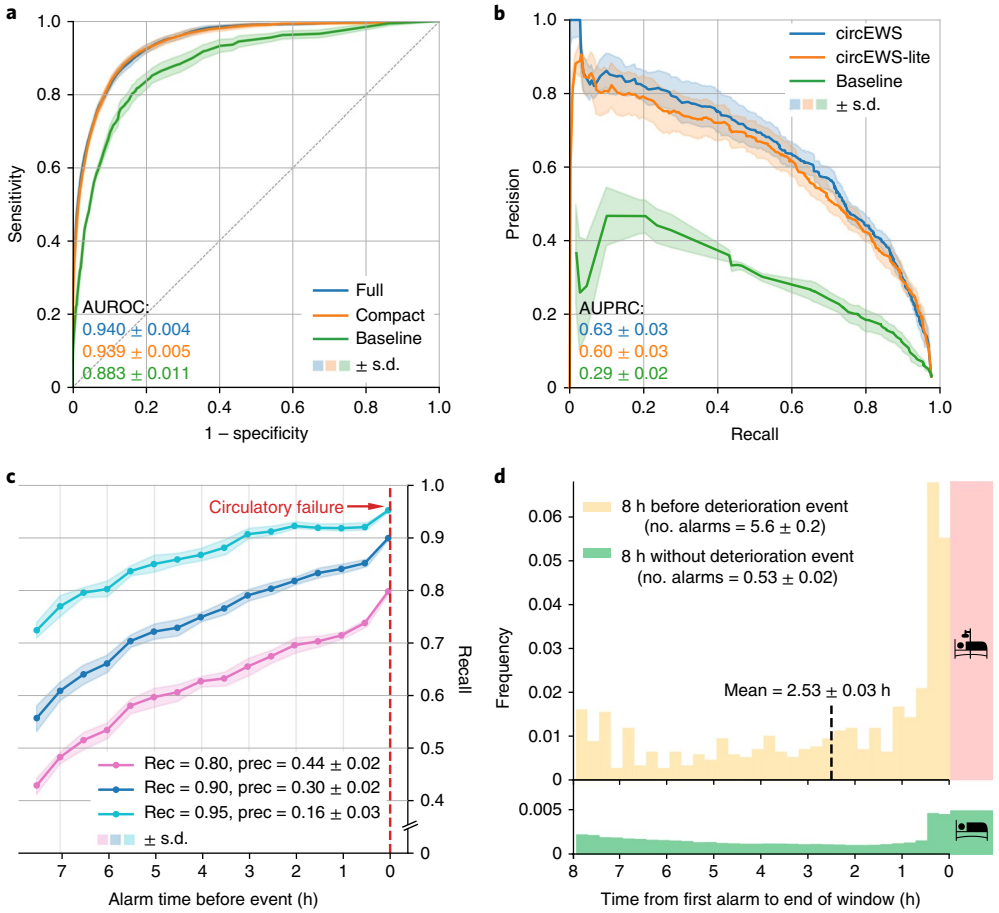


Fig. 2 | Model performance. a, Receiver-operating characteristic curve for the binary classification task of predicting circulatory failure, comparing the two proposed models with a baseline model. The full model contains 500 features (composed from 112 variables), and the compact model contains 176 features (composed from 16 variables). The baseline model used only variables included in the definition of circulatory failure and is based on a decision tree. **b**, Precision-recall curve for the circEWS and circEWS-lite alarm systems derived from the full and compact classification model from **a**. circEWS and circEWS-lite use a 30-min silencing period after every occurring alarm during which no new alarm is triggered. Recall was defined as the fraction of events for which the system correctly raised an alarm from 8 h to 5 min before the event. Precision was defined as the fraction of alarms that are in a window of 8 h prior to a circulatory failure event. **c**, The fraction of events that correctly trigger an alarm is reported for each 30-min interval during the time period 8 h before circulatory failure occurs. **d**, Top, the distribution of timing of the first alarm in the 8 h before an event. The mean time from the first alarm to deterioration was 2 h and 32 min. Bottom, the distribution of alarms in 8 h windows that were not immediately followed by an event. In **a**–**c**, solid curves were derived from the held-out split, variation estimates were derived from $n = 5$ independent experiments in the development splits. Prec, precision; rec, recall.

chosen to match the expected minimal measuring interval for each variable according to standard clinical practice (Supplementary Table 13). The model performance decreased from 0.60 to 0.55 AUPRC (Fig. 4d and Extended Data Fig. 10a,b). A model comparable to circEWS-lite was trained on a modified HiRID dataset that contained only the binarized feature information, that is, excluding the actual feature values and only providing information on time points of measurements. The performance of this model reduced drastically to 0.20 AUPRC on the original test set (Fig. 4d) and to 0.07 AUPRC on the re-sampled test set (Extended Data Fig. 10a,b; AUROC is 0.54, very close to the performance of the random classifier).

External validation. The publicly available ICU dataset MIMIC-III[21] was used for external validation. The 16 variables required for circEWS-lite were identified in MIMIC-III (Table 1 and Supplementary Table 4). We performed identical pre-processing of the MIMIC-III data with minor modifications to account for a lower time-resolution in MIMIC-III. We report the performance of circEWS-lite on the MIMIC-III test set as MIMIC (validation) in Fig. 5a,b. Additionally, a model constructed on HiRID was fine-tuned on the MIMIC-III dataset by linearly interpolating its scores with a model trained on the MIMIC-III data; results are reported as MIMIC (fine-tuned). In both cases, we corrected the label prevalence to be equal to the prevalence in HiRID, enabling comparison of precision in Fig. 5b (before correction MIMIC-III has 1.8% versus HiRID 3.1% positive labels). We observed a slight performance decrease when the circEWS models were applied on MIMIC-III. This can be explained by the higher temporal resolution of HiRID, as shown by the performance of circEWS-lite when trained on an artificially downsampled HiRID dataset to match the time resolution of MIMIC-III (Fig. 5c). As shown in Extended Data Fig. 7e,f, model calibration of both continuous score and alarm system remain suitable when circEWS-lite calibrated on HiRID is directly applied to the MIMIC-III data.

Discussion

We have demonstrated that two variants of a machine learning-based early-warning system (circEWS and circEWS-lite) can predict circulatory failure with very high recall—only a small fraction of events are missed. Since the prevalence of events is low, it is difficult to achieve a low false-alarm rate. Our system generates 2–3 false

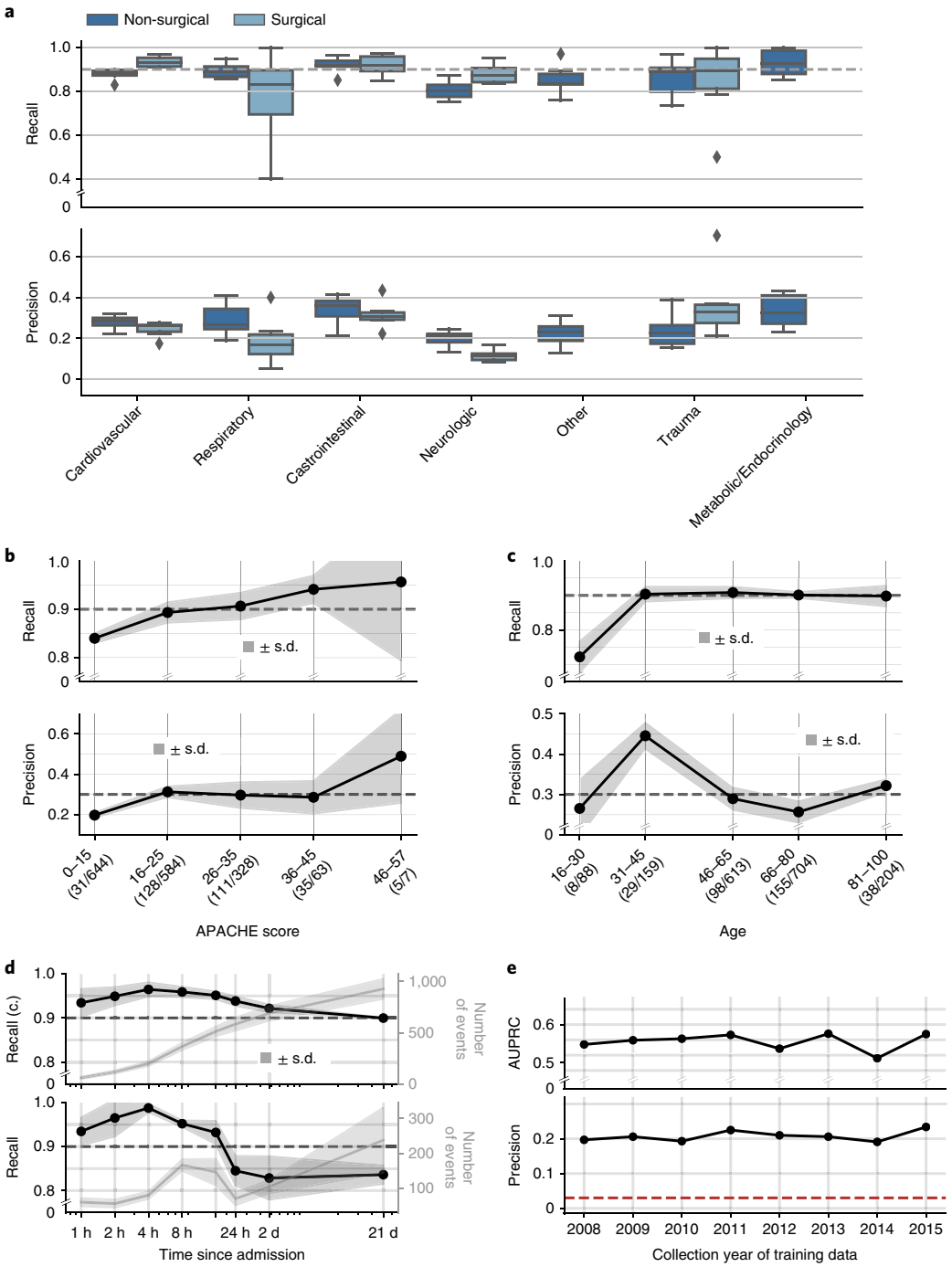


Fig. 3 | Model performance in different patient cohorts. a–e, Analyses use the circEWS model with a threshold corresponding to 90% recall (obtained on patients in the test set) corresponding to an overall precision of 30% and silencing of new alarms for 30 min. **a**, Recall and precision for patients in different APACHE diagnostic groups. Boxes in the box plot show IQR and the diamonds are outliers with values that lies outside the [minimum, maximum] range of the whiskers, where minimum = Q1 − 1.5 × IQR and maximum = Q3 + 1.5 × IQR (Q1, Q3 and IQR represent the first quartile, the third quartile and the interquartile range, respectively). **b**, Recall and precision for patients, as stratified by APACHE-III score. The notation (*a*/*d*) under each group name signifies that there were *a* numbers of patients with events among *d* numbers of patients in the group. **c**, Recall and precision as a function of patient age. **d**, Recall as a function of time since admission. Events (episodes of circulatory failure) are stratified on the basis of time lag after ICU admission. Top, the cumulative performance of the model; that is, at 8 h after admission the overall recall of the model is approximately 96%. Bottom, the recall for each indicated time interval. **e**, AUPRC (top) and precision at a fixed threshold (baseline prevalence shown in red) (bottom) as a function of the year for which the model was trained. Eight models were trained, each using one year of data between 2008–2015, and were tested on a dataset from 2016, for which we observe stationarity ($P = 5 \times 10^{-5}$, $n = 8$ years, Dickley–Fuller test). Box plots in **a** were derived from $n = 6$ independent experiments in the temporal splits; in panels **b–e**, solid curves were derived from the held-out split, and variation estimates were derived from $n = 5$ independent experiments in the development splits. *P* values for panels **a** and **b** (dependent 2-sample *t* test, Benjamini–Hochberg corrected): $P = 0.038$ for decreased event recall in patients with neurological conditions, $P = 0.0006$ for decreased precision in neurosurgical patients, $P = 0.0004$ for lower precision in patients with APACHE scores (0–15), $P = 0.039$ for lower recall in emergency admissions, $P = 0.039$ for higher recall in surgical admissions. $n = 6$ independent experiments in the temporal splits were used.

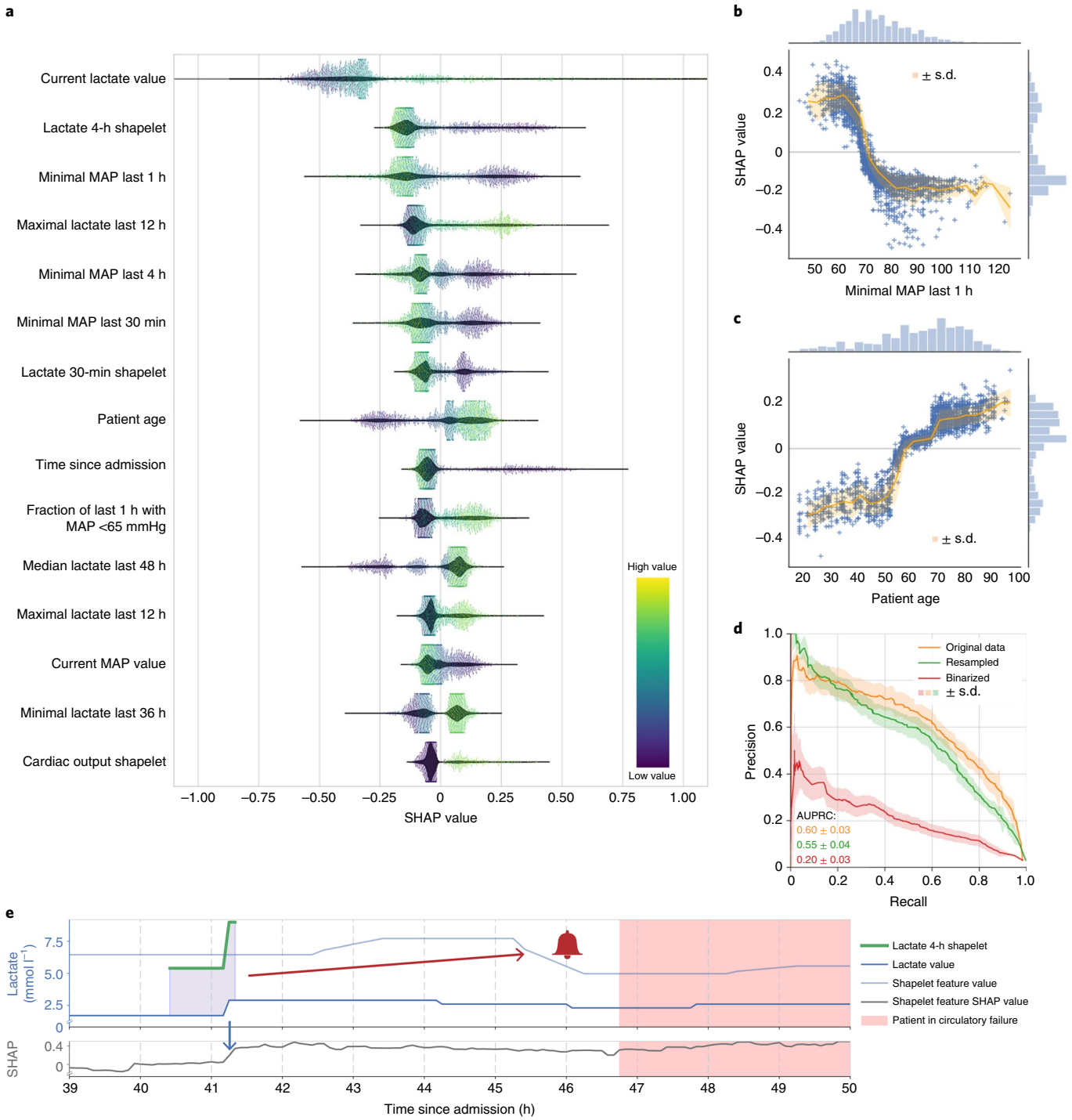


Fig. 4 | Feature inspection. a, The 15 features with highest mean absolute SHAP values. On the *y* axis, the black violin plot shows the full distribution of the SHAP values for each feature. The dot plot in the foreground shows a color coding of the actual value of the feature, resulting in the SHAP value as indicated on the *x* axis. The color coding is based on the percentile of the feature value with respect to the whole distribution. **b**,**c**, Scatter plots showing the relationship between feature value and SHAP value for the minimal MAP in the last hour (**b**) and for patient age (**c**). The orange line and shade represents the mean and s.d. of the regression line. The distributions of the SHAP and feature values are shown as histograms on the right and top of the scatter graph. The high variance in the SHAP value for a given feature value indicates a strong influence of other features. **d**, Precision-recall curves illustrating the performance of the circEWS-lite model on the original test set (Original) and on the resampled test set with regular sampling intervals (Resampled), as well as the performance of the circEWS-lite model trained on binarized data in the binarized test set (Binarized). The variation estimate was derived from $n = 5$ independent experiments in the temporal splits. **e**, Shapelet feature illustration. The lactate values are shown over the indicated time since admission. The lactate shapelet shows an increase in its SHAP value (gray line) 5.5 h before the patient suffers from a circulatory deterioration (blue arrow). The light blue line indicates the feature value that represents the L2 distance between the time series and the shapelet at a 4-h delay. The feature value drops right before the event, increasing the prediction score (red arrow).

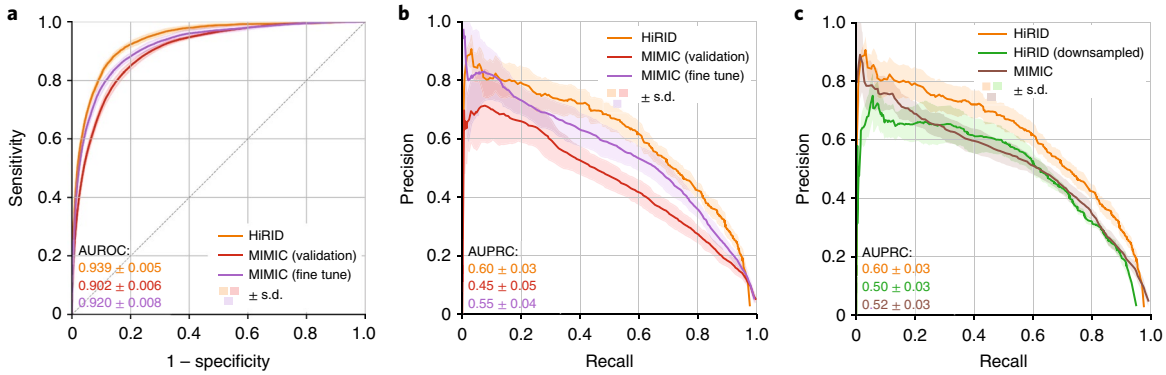


Fig. 5 | External validation of circEWS-lite, as tested on the MIMIC-III ICU dataset. a,**b**, AUROC (**a**) and AUPRC (**b**) for the models. HiRID performance of the circEWS-lite model that was trained and tested on the HiRID dataset. MIMIC (validation), performance of the circEWS-lite model that was trained on the HiRID dataset and tested on the MIMIC-III dataset. MIMIC (fine-tuned), performance of a model pre-trained on the HiRID dataset, then fine-tuned and tested on the MIMIC-III dataset. **c**, HiRID, AUPRC of the circEWS-lite model that was trained and tested on the HiRID dataset; HiRID downsampled, AUPRC of the circEWS-lite model trained on the HiRID dataset and tested on a modified HiRID dataset with artificially lowered data resolution equal to the resolution of the MIMIC-III dataset. MIMIC, AUPRC of the circEWS-lite model trained and tested solely on the MIMIC-III dataset. The reported precision on the MIMIC-III dataset was corrected to reflect the different prevalence of endpoints compared with HiRID dataset. The variation estimates were derived from experiments, considering six temporal splits in the HiRID cohort and five splits in the MIMIC-III cohort.

alarms per day and patient, which we consider very low compared with other warning systems in clinical practice and low enough to be of clinical utility. circEWS—based on 500 features from 112 clinical variables—performed best, but was only marginally better than circEWS-lite—which is based on 176 features derived from 16 of the 20 most important variables. The performance was similar irrespective of diagnosis, the severity of illness and age—with a few notable exceptions (patients with neurological conditions). As expected, the recall rate of the model was highest immediately prior to circulatory failure. Nevertheless, most events could be detected several hours in advance. The alarm system was tested in an independent patient group from a different hospital and demonstrated comparable performance as in the development data.

The main limitations of our study are related to the single-center design, which creates a risk of overfitting the model to the patient cohort and data at hand. However, the analyzed ICU admissions originate from a population covering the whole spectrum of ICU patients, and the external validation demonstrates the applicability of our model in other ICUs. Further, it was not possible to retrospectively identify patients in whom supra-normal blood pressure values were targeted. High blood-pressure targets are often set to maintain cerebral perfusion pressure in critically ill neurological or neurosurgical patients. These patients can have elevated lactate levels due to localized intracerebral ischemia and sympathetic activation[37,38], and therefore would fulfill our endpoint definition without being in circulatory shock. Their inclusion is likely to impair the model performance, and an inferior performance of circEWS was observed in this patient group.

circEWS was constructed with retrospective data collected in a clinical context. Our models rely on features whose presence depends on an active decision to measure by a healthcare provider. Excluding all parameters impacted by the decisions of healthcare providers from model development would eliminate most of our clinical information and is therefore not feasible. This opens up the possibility of bias by intensity of monitoring (related to bias by indication[36]), that is, changing patient monitoring patterns in anticipation of impending circulatory failure are included as important model features. Such a model would show a lower performance in a situation in which impending deterioration was not recognized already by a healthcare provider, defeating the purpose of an alarm system. This is especially relevant for prescribed measurements such as lactate, which may only be ordered if there is a concern about a patient. To understand this effect, we re-tested our model on an artificially regularly sampled test set (simulating the situation where the physician is unaware). We observed a small drop in performance compared with the original model, indicating that the model performance is only weakly dependent on bias by monitoring intensity. We also observed that measurement pattern-dependent features are not among the top 15 model features of circEWS-lite and the performance of a model trained only on intensity of monitoring patterns is poor.

Our data contains artifacts and errors, the removal (where possible) of which was automated. This ensures that similar performance can be expected once the model is applied on live data. The low prevalence of the endpoint was not artificially increased to improve apparent model performance, but left unchanged to mirror future applications of the model. Moreover, we use precision-recall measures to assess the model's performance, which better reflects the system's real-world performance than do ROC curves.

Conventional systems that help identify patients at risk of circulatory deterioration are based on variables known to determine circulatory function and tight alarm limits are set. The reported rates of monitoring alarms in ICUs vary from 6.5 to 53.1 per hour and patient in different studies[13,39–41], and the rate is estimated at about 10 alarms per hour in our data. Often lacking clinical relevance, such alarms can lead to alarm fatigue[15]. circEWS integrates individual patient information and a large variety of physiological measurements from multiple organ systems to achieve a manageable number of timely alarms. On average, the system raises an alarm every 16 h for a stable patient (0.048 alarms per patient hour in the held-out test set). Of all events, 82% are identified more than 2 h in advance. This compares favorably with the total number of alarms and the number of false alarms using conventional systems, and gives the physician enough time for early assessment. In our evaluation, we were not able to detect cases in which a treatment intervention prevented a deterioration event. These would be counted as false positives and decrease the estimated precision of circEWS. We therefore expect that the precision in practice is higher than our estimate.

The ability to establish which features contributed to a prediction ensures that this technology remains interpretable to its clinical users. Using SHAP values, we see that the model identified established predictors determined by circulatory state, but also time-series representations, information from other organ systems,

patient characteristics and treatment parameters. Using SHAP values as a generalized approach to identify the underlying cause of circulatory failure is not possible, but SHAP values may help generate clinical hypotheses for specific events. Only the exclusion of lactate from the top 20 variables identified by SHAP values resulted in markedly decreased performance. This indicates that there is redundancy across variables and features, which is also seen when we ablate entire feature categories. The high relevance of the lactate variable suggests that more frequent measurements of lactate might lead to an increase in model performance and allow for earlier detection of deterioration events[42,43]. The performance degradation of circEWS after the exclusion of lactate may be caused by the informativeness of the current lactate value estimate, the lactate monitoring intensity, or a combination of both. However, the dependence on monitoring intensity patterns of our model was shown to be minor.

To assess the external validity of circEWS, we applied it to the MIMIC-III dataset. We observed that if we apply circEWS-lite directly to MIMIC-III, its performance degrades markedly. However, a fine-tuned model trained on MIMIC-III and HiRID exhibits a performance that almost matches the performance in the HiRiD dataset. The remaining performance gap is likely related to the lower quantity of data available in MIMIC-III to train models (2.8 million versus 13 million time points in HiRID) and the lower temporal resolution (hourly sampling interval for physiological variables in MIMIC-III versus sampling frequency of 2 min in HiRID). A comparison of performance in MIMIC-III with an artificially downsampled HiRID dataset, in which variables were downsampled to the same frequency as MIMIC, confirmed the lack of temporal resolution as the main factor explaining the performance decrease. Even with the large training size available in HiRID, the model's performance has not yet saturated (Extended Data Fig. 5a,b). Therefore, the more limited data in MIMIC-III may also be a factor resulting in lower performance.

The practice of medicine changes with new research and improved technologies. ML methods trained on historical data are therefore susceptible to reduced performance associated with future deployment. Our results indicate a slight increase in model performance the closer the development is to the test set, providing evidence for this effect. Moreover, medical practice varies between providers as well as institutions. The importance of this locational dataset shift is illustrated in our results by the better performance of the locally re-calibrated MIMIC (fine-tuned) model versus the directly applied MIMIC (validation) model. Our model should therefore not be seen as an unalterable and universal scoring system similar to typical ICU scores. In fact, we suggest that other ICUs use our methodology and local data to develop their own models, rather than applying the unaltered circEWS models. In a clinical setting, it will be important to continually monitor the quality of predictions using new data to constantly develop and re-calibrate the models to account for temporal changes and developments in practice. Only after retraining, fine tuning and recalibrating the system for the specific setting can state-of-the-art performance be achieved.

ML techniques have been applied to tasks in radiology[44], pathology[45] and critical care[46,47] in retrospective clinical studies. Approaches spanning a spectrum of complexity have been developed to tackle clinical prediction problems, from linear models[48,49] to complex deep architectures[50]. In this work, we used gradient-boosted ensembles of decision trees owing to their observed superior performance in our application and ease of interrogation. This model class has been successfully applied in many different domains[51,52]. While we tested other models, including recurrent neural networks, we found these approaches inferior. This finding reflects recent observations[30] that careful feature design, combined with state-of-the-art ML approaches, can outperform deep learning. However, when more data is available for training the system, it is likely that more expressive deep architectures may ultimately prove superior.

Considering the demonstrated good performance of our models, we hypothesize that ML-based early-warning systems may help ICU staff to more rapidly identify patients at risk for development of circulatory failure with a much lower false-alarm rate than conventional threshold-based systems. Our data show that even short periods of circulatory failure over the length of stay are associated with an increase in ICU mortality, but do not prove a causal relationship. This finding is consistent with other clinical trials that indicate that repeated or prolonged episodes of hypotension and high dose vasopressors[11,53], as well as delays in shock treatment[9,54–58], are associated with higher mortality. We hypothesize that early identification and treatment of patients at risk of circulatory failure might lead to a reduction in mortality, but this hypothesis has to be tested in a future prospective study and cannot be concluded from our data. Prospective research on the impact of model implementation on patient outcomes has to be conducted before the clinical application of our models. Overall, we show that adaptive data-driven models have the potential to allow the shift from detection and treatment to automated prediction and, hopefully, prevention of organ system failure.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-0789-4.

Received: 9 April 2019; Accepted: 4 February 2020;
Published online: 9 March 2020

References

1. Ehrenfeld, J. M. & Cannesson, M. (eds) *Monitoring Technologies in Acute Care Environments: A Comprehensive Guide to Patient Monitoring Technology* (Springer Science & Business Media, 2013).
2. Fackler, J. C. et al. Critical care physician cognitive task analysis: an exploratory study. *Crit. Care* **13**, R33 (2009).
3. Wright, M. C. et al. Toward designing information display to support critical care. *Appl. Clin. Inform.* **07**, 912–929 (2017).
4. Duke, G., Green, J. & Briedis, J. Survival of critically ill medical patients is time-critical. *Crit. Care Resusc.* **6**, 261–267 (2004).
5. Numata, Y. et al. Nurse staffing levels and hospital mortality in critical care settings: literature review and meta-analysis. *J. Adv. Nurs.* **55**, 435–448 (2006).
6. Falk, A.-C. & Wallin, E.-M. Quality of patient care in the critical care unit in relation to nurse patient ratio: A descriptive study. *Intensive Crit. Care Nurs.* **35**, 74–79 (2016).
7. Wallace, D. J., Angus, D. C., Barnato, A. E., Kramer, A. A. & Kahn, J. M. Nighttime intensivist staffing and mortality among critically ill patients. *N. Engl. J. Med.* **366**, 2093–2101 (2012).
8. Rivers, E. et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N. Engl. J. Med.* **345**, 1368–1377 (2001).
9. De Luca, G., Suryapranata, H., Ottervanger, J. P. & Antman, E. M. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: every minute of delay counts. *Circulation* **109**, 1223–1225 (2004).
10. Lamontagne, F. et al. Pooled analysis of higher versus lower blood pressure targets for vasopressor therapy septic and vasodilatory shock. *Intensive Care Med.* **44**, 12–21 (2018).
11. Vincent, J.-L. et al. Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the MIMIC-III database. *Ann. Intensive Care* **8**, 107 (2018).
12. Martin, C. et al. Norepinephrine: not too much, too long. *Shock* **44**, 305–309 (2015).
13. Ruppel, H. et al. Testing physiologic monitor alarm customization software to reduce alarm rates and improve nurses' experience of alarms in a medical intensive care unit. *PLoS One* **13**, e0205901 (2018).
14. Simpson, K. R. & Lyndon, A. False alarms and overmonitoring: major factors in alarm fatigue among labor nurses. *J. Nurs. Care Qual.* **34**, 66–72 (2019).
15. Borowski, M. et al. Medical device alarms. *Biomed. Tech.* **56**, 73–83 (2011).
16. Top 10 health technology hazards for 2019. *ECRI Institute* (2019); https://www.ecri.org/top-ten-tech-hazards

17. Graham, K. C. & Cvach, M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *Am. J. Crit. Care* **19**, 28–34 (2010). quiz 35.
18. Dietterich, T. G. in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* vol. 2396 15–30 (Springer, 2002).
19. Krumholz, H. M. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff.* **33**, 1163–1170 (2014).
20. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
21. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
22. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
23. Ghassemi, M. et al. Unfolding physiological state: mortality modelling in intensive care units. *KDD* **2014**, 75–84 (2014).
24. Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O. & Sun, J. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2565–2573 (ACM, 2018).
25. Yoon, J., Alaa, A., Hu, S. & Schaar, M. in *Proceedings of The 33rd International Conference on Machine Learning* (eds. Balcan, M. F. & Weinberger, K. Q.) 1680–1689 (Proceedings of Machine Learning Research, 2016).
26. Ren, O. et al. in *2018 IEEE International Conference on Healthcare Informatics (ICHI)* 144–151 (IEEE, 2018).
27. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. multitask learning and benchmarking with clinical time series data. Preprint at https://arxiv.org/abs/1703.07771 (2017).
28. Ghosh, S., Feng, M., Nguyen, H. & Li, J. Hypotension risk prediction via sequential contrast patterns of ICU blood pressure. *IEEE J. Biomed. Health Inform.* **20**, 1416–1426 (2016).
29. Wu, M. et al. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *J. Am. Med. Inform. Assoc.* **24**, 488–495 (2017).
30. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
31. Lundberg, S. M. & Lee, S.-I. in *Advances in Neural Information Processing Systems* 30 (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
32. Ke, G. et al. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).
33. Brier, G. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1 (1950).
34. Davis, J. & Goadrich, M. in *Proceedings of the 23rd International Conference on Machine Learning* 233–240 (ACM, 2006).
35. Tsay, R. S. *Analysis of Financial Time Series* (John Wiley & Sons, 2005).
36. Fihn, S. D. et al. Insights from advanced analytics at the Veterans Health Administration. *Health Aff.* **33**, 1203–1211 (2014).
37. Brallier, J. W., Dalal, P. J., McCormick, P. J., Lin, H.-M. & Deiner, S. G. Elevated intraoperative serum lactate during craniotomy is associated with new neurological deficit and longer length of stay. *J. Neurosurg. Anesthesiol.* **29**, 388–392 (2017).
38. van Donkelaar, C. E. et al. Early circulating lactate and glucose levels after aneurysmal subarachnoid hemorrhage correlate with poor outcome and delayed cerebral ischemia: a two-center cohort study. *Crit. Care Med.* **44**, 966–972 (2016).
39. Cho, O. M., Kim, H., Lee, Y. W. & Cho, I. Clinical alarms in intensive care units: perceived obstacles of alarm management and alarm fatigue in nurses. *Healthc. Inform. Res.* **22**, 46–53 (2016).
40. Christensen, M., Dodds, A., Sauer, J. & Watts, N. Alarm setting for the critically ill patient: a descriptive pilot survey of nurses' perceptions of current practice in an Australian Regional Critical Care Unit. *Intensive Crit. Care Nurs.* **30**, 204–210 (2014).
41. Nuti, S. V. et al. The use of google trends in health care research: a systematic review. *PLoS One* **9**, e109583 (2014).
42. Wolf, A. et al. Evaluation of continuous lactate monitoring systems within a heparinized in vivo porcine model intravenously and subcutaneously. *Biosensors* **8**, E122 (2018).
43. Gouëzel, C. et al. Assessment of changes in lactate concentration with intravascular microdialysis during high-risk cardiac surgery using the trend interchangeability method. *Br. J. Anaesth.* **119**, 1110–1117 (2017).
44. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
45. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
46. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
47. Meyer, A. et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir. Med.* **6**, 905–914 (2018).
48. Verburg, I. W. M., de Keizer, N. F., de Jonge, E. & Peek, N. Comparison of regression methods for modeling intensive care length of stay. *PLoS One* **9**, e109684 (2014).
49. Vairavan, S., Eshelman, L., Haider, S., Flower, A. & Seiver, A. Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model. *Comput. Cardiol.* **39**, 393–396 (2012).
50. Choi, E. et al. in *Advances in Neural Information Processing Systems 29* (eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R.) 3504–3512 (Curran Associates, Inc., 2016).
51. Deng, L. et al. PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine. *BMC Bioinformatics* **19**, 522 (2018).
52. Sarica, A., Cerasa, A. & Quattrone, A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front. Aging Neurosci.* **9**, 329 (2017).
53. McIntyre, W. F. et al. Association of vasopressin plus catecholamine vasopressors vs catecholamines alone with atrial fibrillation in patients with distributive shock: a systematic review and meta-analysis. *JAMA* **319**, 1889–1900 (2018).
54. Seymour, C. W. et al. Time to treatment and mortality during mandated emergency care for sepsis. *N. Engl. J. Med.* **376**, 2235–2244 (2017).
55. Evans, J. et al. The impact of reducing intensive care unit length of stay on hospital costs: evidence from a tertiary care hospital in Canada. *Can. J. Anaesth.* **65**, 627–635 (2018).
56. Scholz, K. H. et al. Impact of treatment delay on mortality in ST-segment elevation myocardial infarction (STEMI) patients presenting with and without haemodynamic instability: results from the German prospective, multicentre FITT-STEMI trial. *Eur. Heart J.* **39**, 1065–1074 (2018).
57. Vincent, J.-L. & De Backer, D. Circulatory shock. *N. Engl. J. Med.* **369**, 1726–1734 (2013).
58. Ortolani, P. et al. Clinical impact of direct referral to primary percutaneous coronary intervention following pre-hospital diagnosis of ST-elevation myocardial infarction. *Eur. Heart J.* **27**, 1550–1557 (2006).

Methods

Study design and setting. The study was designed as a retrospective cohort study for the development and validation of a clinical prediction model. The study was performed at the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland (ICU), an interdisciplinary 60-bed unit admitting >6,500 patients per year. Data processing, model training and analyses were performed at the Departments of Computer Science as well as Biosystems Science and Engineering at ETH Zürich, Switzerland.

Ethical approval and patient consent. The institutional review board (IRB) of the Canton of Bern approved the study. The need for obtaining informed patient consent was waived owing to the retrospective and observational nature of the study.

Participants and data sources. The study included all patients admitted to the ICU in the period between the implementation of the ICU electronic PDMS (GE Centricity Critical Care, General Electrics) in April 2005 and August 2016. The PDMS was used to prospectively register patient health information, measurements of organ-function parameters, results of laboratory analyses and treatment parameters from ICU admission to discharge.

The study flow chart is presented in Extended Data Fig. 2a. Patient admissions prior to 2008 were excluded from the analysis owing to frequent changes in variable identifiers during the run-in phase of the PDMS implementation. Patients without data for determining circulatory failure and patients receiving any form of full mechanical circulatory support, younger than 16 years or older than 100 years, or actively declining the use of their data for research purposes were excluded.

Analysis platform. 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 and scikit-learn formed the backbone of the data-processing pipeline.

Artifact removal. Artifact removal and correction was performed using variable-specific algorithms to enable future live deployment and constituted a major effort. Four main types of artifacts were identified:

- Timestamp artifacts. Measurement time information was stored in two fields—the time the measurement was taken (SampleTime), and the time it entered the system (EnterTime). While the latter field was automatically filled, SampleTime can contain manual input errors, such as an incorrect month or year, disrupting the order of the time series, or falsely indicating unreasonably long ICU stays or gaps between measurements. Intervals longer than 1 d were identified and corrected as described in Supplementary Table 14. Timestamp artifacts existed in 3,530 (8%) of patient admissions.
- Variable-specific artifacts. Blood gas samples required a manual selection of the sample type as arterial or venous. As arterial is the default option, multiple venous samples were wrongly labeled as arterial. This was identified by comparing the oxygen saturation in the blood gas sample to the central venous saturation; if this difference was <10% of the s.d. of oxygen saturation (across the training data), the sample was re-labeled as venous. Patient height and weight were manually entered and sometimes accidentally interchanged. For weight and height measurements resulting in body mass indices (BMI) > 60 kg per m^2 or < 10 kg per m^2, height and weight were swapped if this resulted in a BMI in the range of 10–60 kg per m^2.
- Out of range artifacts. For each variable, a range of possible values (including pathologic values) was defined—these are reported in Supplementary Table 4 (column permitted range in variables tab). Values outside this range were deleted.
- Record duplication. The database contained records of the same variable for the same patient with the same timestamp. For non-pharmaceutical variables, one value was kept if the values of the duplicates were identical. Otherwise, we compared the standard deviation of the duplicates with the global standard deviation of the variable across all patients. If the former was <5% of the latter, we kept the mean of the duplicates. Otherwise, the duplicates were considered unreliable and deleted.
- For pharmaceutical variables, duplicates with an entry indicating a 'zero' dose were deleted. For duplicates of drugs applied as tablets or injections the sum of the recorded dose values was kept. For duplications with a status indicating none of the above, we took the mean of the dose.
- Processing pharmaceutical variables. We converted all pharmaceutical variables to either a rate or presence indicator. Drugs given as boluses such as injections and tablets were converted to an effective continuous rate over a time-period specified according to the estimated duration of action (Supplementary Table 4, column 'acting period (individual)' in drugs tab). In cases where a quantitative rate is not possible, we used a binary flag to indicate if the drug (or drug class; see next section) was present.

Variable merging. The PDMS contained many instances of the same parameter being recorded using different identifiers (for example, different dilutions of vasopressors, different probe locations for core temperature measurements). Moreover, specific variables were infrequently observed (for example, foscavir was observed <50 times), but belonged to a meta-category (such as 'Antiviral therapy'). To build a model that is less specific to the local patient cohort and setting, we used the following variable merging strategies, reducing our set of variables from 710 to 209. Identical medical core concepts recorded as different variable IDs were merged (for example, different probe locations for core temperature measurements). Identical pharmaceutical compounds were aggregated into one variable (Supplementary Fig. 1 and Supplementary Table 4).

Certain clinically less important compounds were aggregated to group variables regarding the targeted pharmaceutical effect (for example, non-opioid analgesics; Supplementary Table 4, columns 'drug' and 'constituent drugs (if relevant)' in drugs tab). This was performed for better temporal and inter-ICU generalizability by making the model features independent of the specific compound used. If this led to multiple measurements at the same time, the following strategies were used. For physiological parameters (such as temperature) or lab tests, we used the median of simultaneous measurements. For pharmaceutical variables, we used a weighted sum over simultaneous infusions, with weighting given by effective relative doses determined by analysis of the literature. Otherwise, we merged variables into a binary indicator denoting whether or not any drug from that class (for example, antibiotics) was present, or count how many drugs are present (Supplementary Table 4, column 'merging ratio' in drugs tab).

Circulatory state annotation. We annotated every 5-min interval of a patient's stay with their current circulatory state using 3 types of variables: lactate (arterial and venous), MAP and presence of vasoactive/inotropic drugs. The state was established using a window of 45-min duration centered on the current time point. To reduce spurious calls due to transient states, in each such window all conditions had to be independently true for 30 min (not necessarily consecutive).

We defined the following three states:

- Patient currently not in circulatory failure: if MAP is >65 mmHg, vasoactive/inotropic drugs are not present, and lactate is ≤2 mmol l^{-1}.
- Patient currently in circulatory failure: MAP is ≤65 mmHg, or (not exclusive) vasoactive/inotropic drugs are present and lactate is >2 mmol l^{-1}.
- Unknown/ambiguous: if any of the following conditions hold:
- No MAP or (interpolated) lactate is available in the 45-min windowMAP or vasoactive/inotropic drug criterion is met, but lactate is ≤2 mmol l^{-1}

To enable state annotation at all time points, we imputed lactate values between measurements. We linearly interpolated lactate values between measurements, unless the patient's lactate value had passed the threshold of 2 mmol l^{-1} in either direction. If a patient's state had changed, from either low to high lactate or vice versa, we linearly interpolated depending on the interval between the two measurements. If they were less than 6 h apart we interpolated for the full period. Otherwise, we forward/backward filled for a maximum of 3 h and the remaining time points were left missing.

To handle the starts/ends of the stay, we filled forward/backward. If the patient's first/last measurement was 'normal' (under the threshold), we backward/forward filled indefinitely. If the measurement was abnormal, we filled backward/forward for up to 3 h.

As this imputation scheme implicitly used information from the future, it was only used for annotating (and subsequently labeling) time points. Adaptive imputation and feature generation for model development were performed independently and as described below without using future information.

Labeling of future circulatory failure. All time points annotated as 'no circulatory failure' (that is, not currently in a circulatory event state) were labeled as 'positive' if circulatory failure occurs in the next 8 h, otherwise 'negative' (Fig. 1e). Ambiguously labeled time points were excluded from training and evaluation.

Patient-centered adaptive time series imputation. Our imputation strategy was based on the assumption that measurements are not missing at random, and that the level of missingness (that is, measurement rate) is informative for the rate of change of that variable. We used this measurement rate to define imputation parameters for each variable. These parameters were pre-computed on the training set. They consist of the median and interquartile range (IQR) of the sampling interval of each non-medication variable i, denoted as (m_i, iqr_i) below. The imputation process created a time grid with step size 5 min, starting and ending at the patient's first and last heart rate measurements, or ending at 28 d after admission (whichever was shorter). This provided a unified definition of 'beginning of stay' corresponding to the start of basic monitoring, and avoids biasing the data towards patients with very long stays. Values were imputed for all variables independently at each grid point using the following process. Prior to the first measurement of a given variable, or if the patient had no measurements, we filled it in using a normal value (Supplementary Table 4, column 'default value' in variables tab). If the last measurement, as seen from the grid point, was less than $m_i + iqr_i$ minutes away, we used forward filling from the last value. Otherwise, we linearly returned to the median of the last $2 \times (m_i + 2 \times iqr_i)$ min, as measured from the point where we entered the region where this imputation mode is applied,

for $2 \times (m_i + 2 \times iqr_i)$ minutes in total. After that, we assumed that the value stayed constant at this median value (indefinite forward-filling), until the next valid measurement, if any, at which we returned to step 2. Static variables were imputed according to either the mean or the mode value in the training data, for continuous and categorical values, respectively.

We found that a similar performance can be achieved using simpler imputation strategies like indefinite forward filling or leaving unobserved time points as missing values. While this is surprising, it should be seen in the perspective of our feature choices, which are robust to missing data, and the robustness to missing data inherent in decision-tree based methods. Adaptive imputation achieves a more regular data format and provides estimates of current values based on prior clinical knowledge, which could be useful in a real-time monitoring setting. The adaptive imputation method could provide advantages if certain components of our methodology are replaced with models less robust to missing data, such as deep neural networks (DNN) or logistic regression.

Feature generation. Feature generation took the imputed data as input and generated features sample-wise on the 5-min grid. The first 30 min of a stay were ignored for feature generation, because the history of vital signs and lab tests contained insufficient information to generate reliable features. Six types of features were generated for each time-grid point. They included the current estimated value of a variable, and five others (described in detail below). Besides these feature classes, we also added 'time since admission' as an individual feature.

- Static features. Six static features (age, indicator of surgical admission, indicator of emergency admission, APACHE diagnostic group, height and sex) were concatenated to each time-grid sample of a patient.
- Multi-resolution summaries. We constructed time windows of increasing size and extract summary statistics over each window to capture the temporal history of our data (Fig. 1d). The window sizes and statistics depended on the variable type and sampling rate. We classified each variable as either as high, low or medium frequency, according to its median sampling interval in the training set or estimated duration of action for drugs (Supplementary Table 4, in drugs tab). We defined four time windows for short, medium, long and very long time horizons for each frequency category, using prior clinical knowledge (Supplementary Table 1). We extracted the median, IQR and minimum and maximum values for continuous-valued variables, as well as a trend estimate for each of the time horizons. We reported a mean estimate rather than median for medications. For categorical and binary variables, we reported only the mode or mean, respectively. Additionally, we summarized the entire stay up to now with the summary functions mode, mean or median depending on the variable category. Our results suggest that considering several horizons (Supplementary Table 15) as well as different summary functions (Supplementary Table 16) increased performance.
- Instability history features. Assuming that patients who have already suffered circulatory instability are at increased risk of recurrence, we formed a set of features to capture the patient's history of instability. We encoded the current state, time to the last pathologic state as well as the density of pathological states in the past. All of these refer to logical subconditions of our circulatory failure definitions (Supplementary Table 2). We set a value of 30 d (larger than the maximum length of a stay) if no abnormal state was measured. The density was defined as the ratio of the duration that the state was active during the length of the stay so far.
- Measurement-intensity based features. Since imputation removed information about when and how often measurements were performed, we reintroduced some of this information with this feature class for vital sign measurements and lab tests. We computed the time since the last (non-imputed) measurement (30 d, if no measurement available), as well as the ratio of time points with measurements in the stay up to now.
- Shapelet-based features. A shapelet is a small time series subsequence that is discriminative for the class label and known to capture salient temporal dynamics of time series in a variety of application domains[59]. We used the computationally efficient S3M method[60]: for each variable, 300 subsequences were extracted with a padding of 5 min before any deterioration event, and these were labeled as cases. The same number of uniformly sampled time series from the remaining patients served as controls. The remaining parameters were adjusted according to the resolution of the variable (see Supplementary Table 3). As the resulting shapelet set might be very large (up to multiple 1,000 shapelets per variable), the subsequent shapelet selection step created a feasible number of shapelet features per variable (in our case 20 shapelets per variable and length) that was representative of the space of all shapelets with the min–max approach. First, the shapelet with the highest accuracy in differentiating cases from controls on the training dataset was selected. Afterward, shapelets were iteratively selected such that the minimal distance to the set of already selected shapelets was maximized until 20 shapelets are chosen. Supplementary Table 17 shows that the min–max sampling performed similarly as random sampling or selecting the top 20 shapelets. A shapelet was used to construct a set of features per time point by concatenating the L2 distances between the shapelet and the history of a patient's variable during the last 4 h. This history of distances (dist–hist) approach outperformed other feature computation approaches as shown in Supplementary Table 17: single distance (distance), the minimum over all distances in the last 4 h (min) and counting the number of shapelets that have occurred in the last 4 h (count).

Supervised learning of deterioration prediction. We defined a binary prediction task to be performed every 5 min while a patient is not in circulatory failure on the time-point-based labels. To avoid overestimating the prospective performance of our model by a random assignment of patient stays to training, validation and test sets, we used an experimental design in which the test set contained the most recently admitted patients in the cohort (Extended Data Fig. 2d); we call this a 'temporal split'. In a given temporal split, the full methodology was applied independently, including missing data imputation, feature extraction, model training and hyperparameter selection.

Five overlapping temporal splits were constructed, each containing admissions across 5 yr (Extended Data Fig. 2d). The admissions of the last year were split 1:1 to define the validation and test sets, respectively. The remaining earlier admissions were used as training set of that split. The start of each subsequent split was shifted by six months. Further, the most recent 10% of patient admissions (November 2015 to May 2016) were defined as the held-out evaluation set, which was not used for model development to avoid subtle overfitting to this dataset. This set was also split 1:1 into held-out validation and test sets. Using the rest of the available data as training data, this held-out set formed a special 'held-out' split, which was used to provide a point estimate of model performance. The five temporal splits were used to estimate the variability of model performance, containing disjoint test sets, and partially disjoint training sets. We report this variability as the s.d. over model performance in these splits. Lastly, an exploration split was defined, which assigns patients at random to training, validation and test set in proportions 8:1:1. This split was used to compare temporal generalization with the standard approach of randomly assigning admissions to training and test sets. Lastly, an exploration split was defined, which assigns patients at random to training, validation and test set in proportions 8:1:1. This split was used to compare temporal generalization to the standard approach of randomly assigning admissions to training and test sets and resulted in nearly identical model performance (AUROC in the temporal splits 0.934 versus AUROC in the random splits 0.937).

Statistical methods. If not otherwise indicated, solid lines and performance metrics displayed in figures and tables refer to the performance in the test set of the 'held-out' split as described in the previous section. Shaded areas and numbers in parentheses refer to the s.d. across the respective evaluation metric in the test sets of splits 'Temporal 1–5' ($n = 5$). The center line of box plots shows the median, and the lower and upper limit show Q1 and Q3 respectively, where Q1 is the first quartile and Q3 is the third quartile. The lower whisker shows the first datum greater than Q1 − 1.5 × (Q3 − Q1), and the upper whisker shows the last datum less than Q3 + 1.5 × (Q3 − Q1), and additional points are outliers. In violin plots, the bandwidth is set to 0.2 for computing the Gaussian kernel density estimate and the density is not extended past extreme data points. When analyzing the effect of temporal gap between the test data and training data in terms of year of training data collection, we used the Dickley–Fuller test ($P = 5 \times 10^{-5}$, $n = 8$ years) to test for stationarity (precision/AUPRC values, Fig. 3e), and the two-sided Wald test ($P = 0.051$, $n = 8$ years) for testing a non-zero slope in a linear regression line fit (AUROC, Extended Data Fig. 8e). Performance in sub-cohorts was tested using a paired two-sided two-sample t test corrected by the Benjamini–Hochberg procedure for multiple comparisons, hereby the samples were paired using the $n = 6$ temporal splits ('Temporal1–5'/'Held out' split) (Fig. 3).

Machine-learning approaches. We compared the following three state-of-the-art supervised ML techniques to learn to detect deterioration events:

- Gradient boosted ensemble of decision trees, and decision-tree baseline. The gradient boosting library lightGBM (version 2.2.1) was used for model fitting[32]. The hyperparameter settings maximizing the AUPRC on the validation set were used to generate the predictions on the test set, after refitting on the training set. The model training process was stopped if the AUPRC on the validation set did not improve over 50 consecutive fitting iterations, resetting the model state to the best iteration before early stopping. Since lightGBM can deal natively with categorical data, we did not one-hot-code such data before model fitting. As this model achieved highest performance during system development, it was used for further analyses. To obtain the decision tree baseline, we set the number of trees to 1.
- Logistic regression. The class SGDClassifier from the scikit-learn library (version 0.20.0) was used for model fitting. The strength of the regularization parameter was selected by maximizing the AUPRC on the validation set. Before fitting, continuous features were standardized (zero mean, one s.d.), and categorical features one-hot-encoded.
- LSTM-based recurrent neural network model. We constructed a long short-term memory (LSTM)[61] network comparison in TensorFlow 1.11.0. We used the same set of features as provided to lightGBM for fair comparison after intermediate results suggested worse performance when using only the raw variable values. Since a fraction of the features are static, a small-size single-layer-perceptron (SLP) was used alongside the LSTM to learn from the static

features. The LSTM and the SLP output the hidden states for the dynamic and the static features, respectively, and by linear combination these two hidden states are fed into the output layer. Before training, all non-categorical features were standardized and categorical features one-hot-encoded.

Hyperparameter settings and grids for all models are listed in Supplementary Tables 18–21, if not otherwise described.

Variable and feature selection. The importance of individual features was measured using mean absolute SHAP values of predictions made on the validation set for each temporal split. Before SHAP values were computed, the negative instances in the validation set were sub-sampled to achieve a balanced dataset. The variable ranking was obtained with a greedy forward selection approach by which the variable associated with the feature with the largest mean absolute SHAP value considered the most important variable. This variable and all its features were then removed from the ranking and the procedure was repeated. The final ranking of important clinical variables was determined using the held-out split. The standard deviation of the ranks is computed on the five temporal splits used for model development. Optimal model performance was obtained using 500 features, and removing more features degraded performance (Extended Data Fig. 5e,f). These features, comprising 112 variables, are provided to the full model. We further identified the top 20 clinical variables using the ranking procedure (see Table 1) and excluded four variables not identifiable in MIMIC ('non-opioid analgesics', supplemental oxygen and two inotropes). The resulting 176 features from the remaining 16 variables formed the compact model.

Model calibration. We have performed an analysis of calibration (observed risk versus raw prediction score) for our proposed full and compact models as well as a post-hoc calibration (Extended Data Fig. 7). The raw prediction scores produced by the machine learning model (LightGBM) were post-hoc calibrated on the validation set of each temporal split, and evaluated on the test-set. We used isotonic regression, which fits a rank-preserving transformation between the original scores and transformed scores that minimizes the deviation between the target label and the prediction score. We used the scikit-learn library (version 0.20.0) for fitting the isotonic regression model. Model calibration was evaluated using the area between the calibration curve and the ideal calibration curve, which represents perfect concordance between the prediction scores and the absolute risk. Twenty prediction score bins with regular spacing between the minimum/maximum prediction score produced by a model were used. Defining the observed risk as the average time to circulatory failure, we noticed that the raw scores are already well-calibrated. Temporal model calibration was evaluated using the area between temporal risk and its linear regression fit. Moreover, comparing the compact model on HiRID to the same model in the MIMIC (validation) setting, we observe only minor differences, which suggests that the calibration of the prediction score with respect to time-to-deterioration is not affected strongly by patients drawn from different populations in our scenario. We also evaluated model calibration in different subgroups of the HiRID cohort, the results are summarized in Supplementary Table 5. We observe the strongest deviations from calibration in patients with neurological conditions (for full and compact model) and for categories that have fewer than one hundred patients (which we attribute to statistical estimation errors).

Early-warning system and evaluation. A core contribution of this work is an early-warning system for circulatory failure within 8 h—circEWS. We built two variants, circEWS and circEWS-lite, based on the binary classifiers 'full' and 'compact' described above. The output of the classifier is a score between 0 and 1, which is converted to an alarm if it exceeds a fixed threshold. On top of this, we employed a silencing policy to reduce unnecessary repetitive alarms: for 30 min after an alarm is raised, any potential subsequent alarms were suppressed. If a patient experienced circulatory failure and recovered, the system was reset to allow new alarms after 25 min—this lag period ensures the patient is out of the circulatory failure event before the system is reset. The effects of different silencing periods and system reset times are shown in Supplementary Tables 6 and 7. Our objective was to evaluate circEWS in a clinically relevant context, focusing on the percentage of circulatory failure events the system is able to detect and the rate of false alarms. Model precision was defined as the fraction of alarms that correctly predict the onset of an event (a period of circulatory failure) within the next 8 h. Model recall was defined as the fraction of events that are captured by an alarm. This is analogous to exon prediction in gene finding[62]. Significance of performance differences in patient sub-cohorts was assessed using a $P < 0.05$ cutoff using dependent two-sample t tests corrected for multiple-comparison testing with the Benjamini–Hochberg procedure, matched on the 6 temporal splits in which the experiment was replicated.

Assessing bias by intensity of monitoring. To perform the analysis shown in Fig. 4d, we created an alternative version of the test set under a hypothetical setting in which clinicians follow a fixed measurement intensity for all patients, independent of any suspected circulatory deterioration. To achieve this, we defined 'normal' measurement intervals for all variables (Supplementary Table 13), and then resampled the data to conform to these intervals. If a patient has no measurement for a certain variable, we impute a normal value (Supplementary Table 4, column 'default value' in variables tab) with additive Gaussian noise to every time point on a semi-regular time grid. The mean interval time is equal to the corresponding baseline interval and the s.d. of the grid interval is equal to half of the baseline. For a patient who has measurements for a certain variable, we use the following procedure: if the first measurement happens within the first baseline interval time, then starting from the time of the first measurement, we identify the next time point by a time shift of 1 baseline interval. For every new time point thus reached, and if the closest measurement is within a window of size of the baseline interval centered at the current time point, we keep the closest measurement and remove all other measurements between the current and the last time point. Otherwise, we impute the current new time point with the last closest measurement value with additive Gaussian noise. If the first measurement occurs after the first baseline interval, we impute a random time point within the first interval with the normal value plus Gaussian noise, and repeat the same processes as described for the case that the first measurement happens within the first baseline interval. The standard deviation of the Gaussian noise model for each variable is the median of the s.d. of the corresponding measurement values during each patient stay for all patients.

We confirmed that the resampling behaved as expected by comparing the sampling interval distribution for patients with events and without events in both the original data and the resampled data (Extended Data Fig. 10c,f,g).

We further generated a 'binarized' test set, where all measurements were replaced by binary values indicating the presence or absence of a measurement at that time point. Specifically, a binarized value on the time grid was defined as 1 if there was a real measurement in the last 5 min prior to the grid point, and 0 otherwise (results shown in Fig. 4d and Extended Data Fig. 10a,b). This setting preserves only information from measurement patterns. Further, we defined a negative control that combines the binarization with the resampling procedure (which aims to remove measurement intensity information), which we assume retains no or very little measurement pattern information (Extended Data Fig. 10a,b). We observe that the AUROC in this setting is close to a random classifier (AUROC close to 0.5), confirming that the resampling strategy is effective, in that it removes the measurement pattern information almost completely.

External validation on MIMIC-III. MIMIC-III version 1.4 was used for external validation, including only patients admitted after 2008 and the introduction of MetaVision. Sixteen of the 20 most important variables (Table 1) were available and extracted. Non-opioid analgesics and supplementary oxygen could not be matched. The drugs levosimendan and theophylline were not used at Beth Israel Deaconess Center and were excluded. Since many of the artifact-removal steps described above are specific to HiRID, we applied only artifact removal using the same fixed variable-specific ranges on the MIMIC data. MIMIC data were converted into the correct format to be processed by the rest of the HiRID pipeline. Patient-state annotation, label generation and missing data imputation were performed as described above with minor modifications.

Imputation parameters. Part of our imputation pipeline required calculating the sampling interval for each variable. These intervals were not recomputed to provide similar data for validation to our existing model. Furthermore, we did not expect the ground truth of these values to vary much between ICUs, even if different down-sampling is used.

Time grid. The temporal resolution of MIMIC was different to that of HiRID. Nonetheless, to mirror the HiRID data structure as closely as possible, we resampled MIMIC to a 5-min grid, even if this introduces a large quantity of imputed data.

We consider two settings for evaluating circEWS-lite on MIMIC: MIMIC (validation), and MIMIC (fine-tuned). In the validation setting, we applied circEWS-lite on MIMIC, using the full dataset as a test set (in total 9,040 patients; Extended Data Fig. 2b). In the fine-tune setting, we applied the same processing as before, but re-trained the compact model to predict circulatory failure on MIMIC and linearly interpolated its prediction score with a base model trained on the HiRID data-set. The interpolation coefficient was optimized using event-based evaluation metrics on the validation set of each experimental split. We formed five replicates of MIMIC using Monte Carlo resampling, in each replicate assigning admissions at random to training/validation/test sets in the ratio 3:1:1. Since absolute admission times were not available in MIMIC, temporal splits could not be constructed. Our final performance estimate was the mean of the performance in each replicate's test set, and the error estimate was the standard deviation over replicates. In the fine-tune setting, each MIMIC training set contained approximately 4,600 patients, with ~1,600 patients in the test sets.

To evaluate whether the performance discrepancy between HiRID and MIMIC is due to the sampling frequency, we downsampled variables with higher frequencies in HiRID than in MIMIC so that the sampling frequencies of the corresponding variables will be the same in both datasets. The downsampling process of each of those variables keeps the last observations within the shifting time-windows of the same length as the average sampling interval of that variable in the MIMIC dataset. And the step-size of the shifting window is the same as the window size.

Prevalence correction for MIMIC. The test sets used for MIMIC (validation) and MIMIC (fine-tuned) have different prevalences of positive events compared to the test set of the HiRID dataset. Therefore, to enable a comparison of the performance of *circEWS-lite* in terms of precision and recall between HiRID and MIMIC, we

corrected the precision-recall curves for MIMIC in Fig. 5 and Extended Data Fig. 10d such that MIMIC would have the same positive event prevalence as HiRID. The uncorrected precision-recall curves are shown in Extended Data Fig. 10e. The correction was performed as follows. For the correction of the alarm/event-based precision-recall curves for MIMIC, we computed the AUPRC of an alarm system that was based on a random classifier and with the same silencing policy as circEWS-lite for both HiRID and MIMIC, which we denote by $prev_e$ (HiRID) and $prev_e$ (MIMIC), respectively. The event prevalence of the dataset is defined as the AUPRC of a random alarm system. We downscale the number of false alarms observed in the MIMIC test set with:

$$s = \frac{\left(\frac{1}{prev_e(\text{HiRID})} - 1\right)}{\left(\frac{1}{prev_e(\text{MIMIC})} - 1\right)}$$

to satisfy the assumption that the calibrated MIMIC dataset has the same event prevalence as HiRID.

The correction factor s is then multiplied with the false-alarm counts when computing precision on the MIMIC data in order to obtain corrected precision estimates.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

More information on HiRID is available on hirid.intensivecare.ai, and the full dataset can be downloaded from physionet.org. The computer code used in this research is available at www.github.com/ratschlab/circEWS under an open-source license.

References

59. Ye, L. & Keogh, E. in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 947–956 (ACM, 2009).
60. Bock, C. et al. Association mapping in biomedical time series via statistically significant shapelet mining. *Bioinformatics* **34**, i438–i446 (2018).
61. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
62. Engström, P. G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).

Acknowledgements

Funding for this work was provided by the Swiss National Science Foundation (grant no. 176005 to G.R. and T.M.M. and SNSF Starting Grant no. 155913 to K.B.). G.R., S.L.H. and K.B. received core funding from ETH Zürich. We gratefully acknowledge the helpful discussions with H. Strathmann and V. Gal. We thank V. Andreas, X. Bonilla, D. Sidebotham, N. Toussaint and I. Jarchum for proofreading the manuscript.

Author contributions

S.L.H., M. Hüser, X.L., M.F., G.R., T.M.M., K.B., T.G. designed the experiments; M.F., T.M.M. selected and provided the clinical data and context; X.L., M.F., S.L.H, M. Hüser with contributions from T.M.M., M.Z. and G.R. preprocessed and cleaned the data; S.L.H., M.F., T.M.M. with contributions from G.R., X.L., M.Hüser defined and developed the labeling of deterioration events; M. Hüser, M.F., with contributions from X.L., S.L.H., T.M.M., G.R. devised and implemented the adaptive imputation strategy; M. Hüser, M.F., X.L., S.L.H. developed and extracted non-shapelet features, T.G. and C.B. developed code for shapelet analysis; M.Hüser, X.L., S.L.H. developed the pipeline for supervised learning; X.L. implemented the LSTM model; T.G. implemented the decision tree baseline. C.E., C.B., M.F., M. Horn, M.M., B.R., D.B. contributed to various analyses of the data; T.M.M., G.R., S.L.H., M.F., K.B. conceived and directed the project; S.L.H., M.F., M.Hüser, X.L., T.G., T.M.M., G.R., K.B. wrote the manuscript with the assistance and feedback of all the other co-authors. S.L.H. and T.G. with input from all authors created Fig. 1.

Competing interests

The authors declare no competing interests.

Additional information

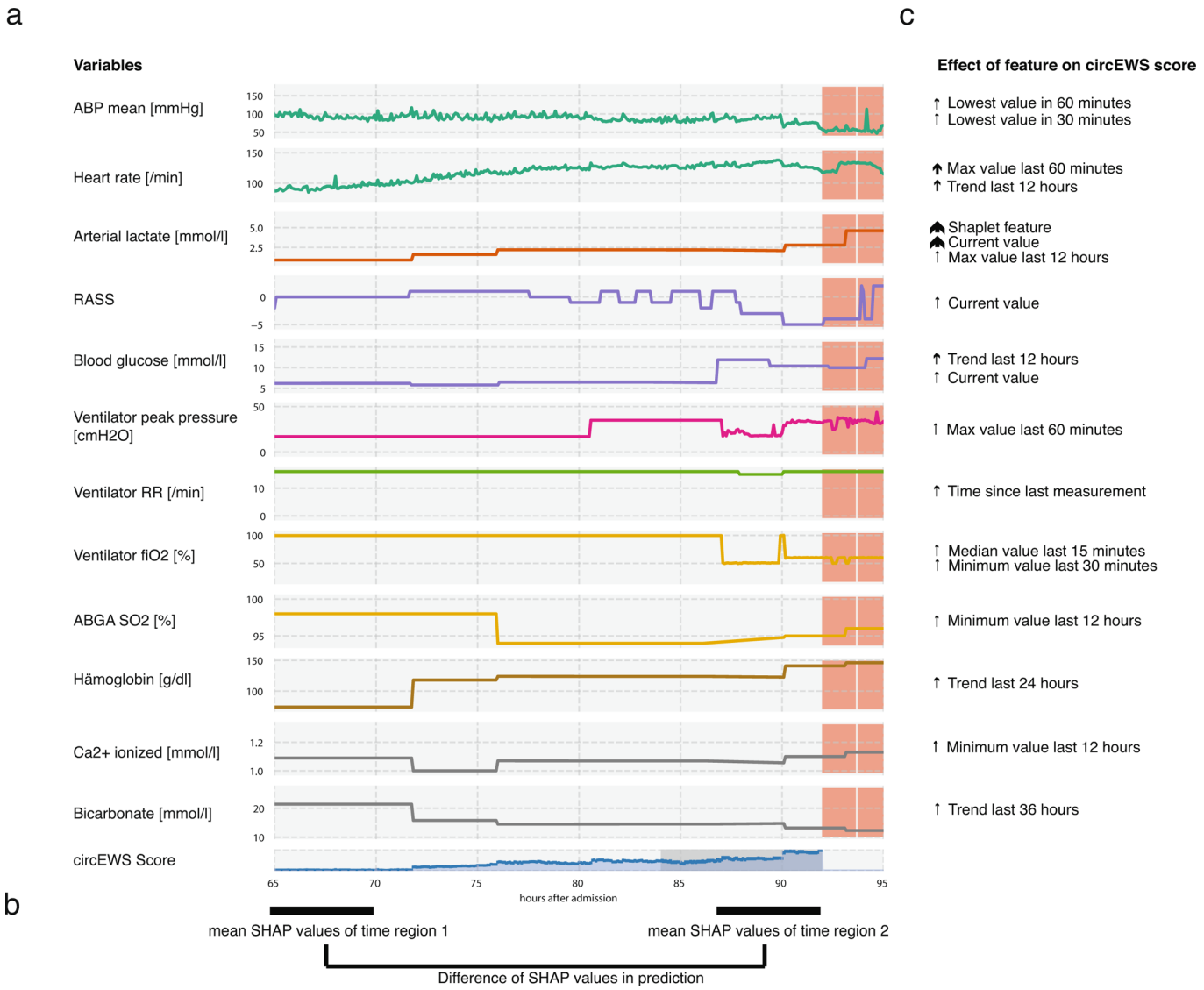
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0789-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0789-4.

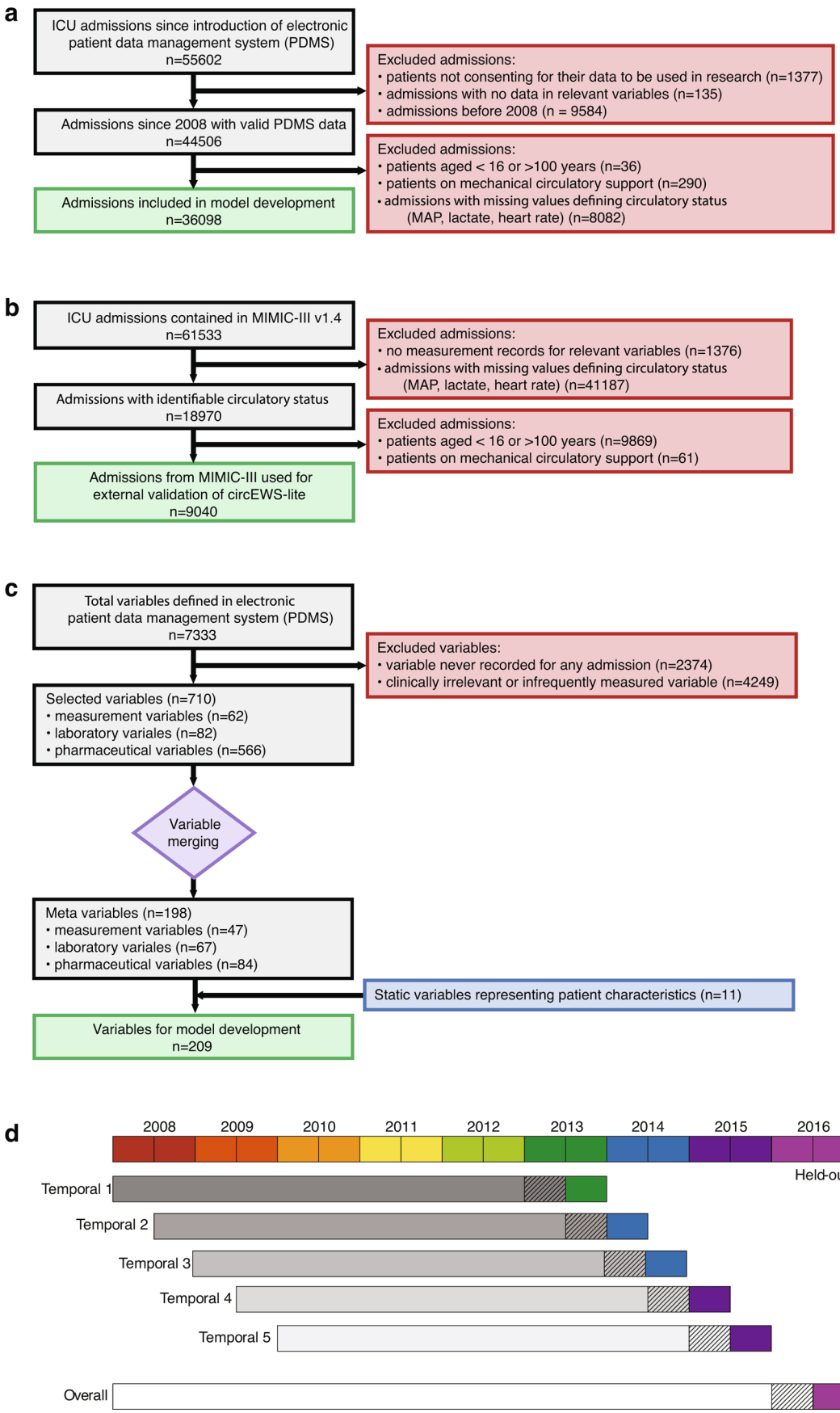
Correspondence and requests for materials should be addressed to K.B., G.R. or T.M.M.

Peer review information Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Example patient stay. a: Variables. Visualization of a partial example patient stay with time after admission on the x-axis. Shown are the recorded values for this patient for the top-ranking variables. The bottom time series shows the prediction score of circEWS. The red region denotes an area where the patient is in circulatory failure. The dark grey area overlaid on the score time series (bottom) denotes the region where the event should be predicted (8 h before the beginning of circulatory failure). For this patient, the alarm is triggered at 90.1 h after admission (1.9 h before the event). RASS: Richmond Agitation Sedation-Scale. **b**, **c**: Influential SHAP values. As a measure of the influence of each feature, we use the difference of the average SHAP values for this feature comparing the time region 1 [65, 70] and 2 [87, 92]. We plot the 20 features with the largest increase in SHAP values and indicate the magnitude of the change by the size of the arrow. (Patient information: 48-year-old female patient with urosepsis).

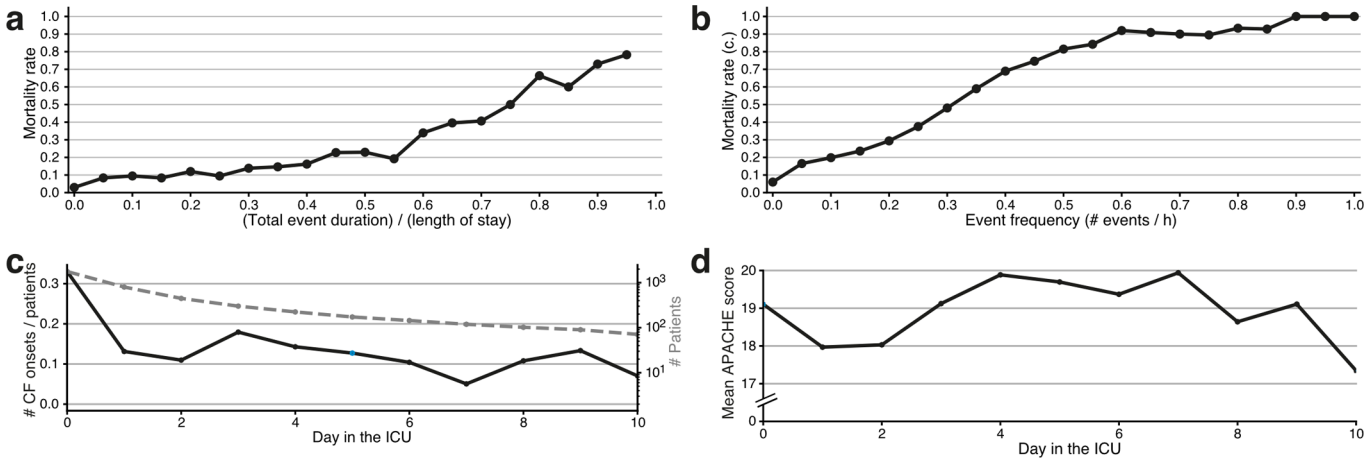


Extended Data Fig. 2 | See next page for caption.

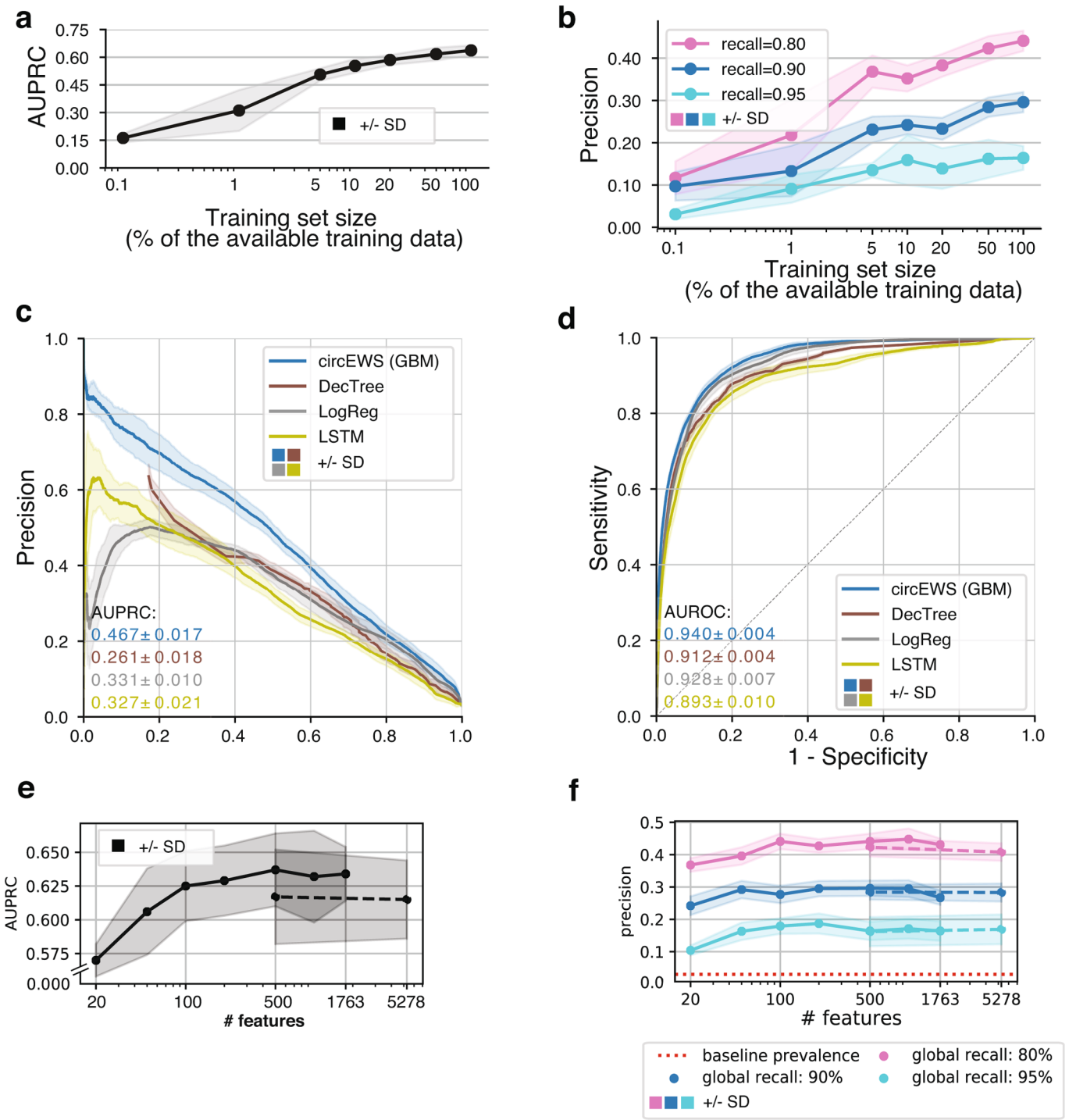
Extended Data Fig. 2 | Experimental design. a, Flow chart of the exclusion criteria applied to the HiRID patient cohort. **b**, Flow chart of the exclusion criteria applied to the MIMIC-III patient cohort. **c**, Flow chart of the exclusion, merging, and post-processing applied to the variables in the patient data management system of the HiRID cohort. **d**, Data split design. We formed five "replicate" splits with disjoint test sets (colored, dark), and partially overlapping training sets (grey). The validation and test sets consisted of the most recently admitted patients, whereas the validation patients (hatched) are from an earlier period. This enabled us to select a model more likely to be generalizable to the future. In addition, a held-out split was formed. The test set of the split was not used for developing the models described in the manuscript and only assessed for preparing the publication figures to avoid subtle overfitting to the dataset.

Sex			
Male	63.49 %	22919 admissions	
Female	36.51 %	13178 admissions	
Age (years)			
Median (Mean)	66 (62.98)		
Range	16-98		
Length of stay (days)			
Median (Mean)	0.93 (2.09)		
Range	0-84		
Admission type			
Emergency	55.49 %	20032 admissions	
Not emergency	42.49 %	15337 admissions	
Unknown	2.02 %	729 admissions	
Surgical status			
Yes	55.25 %	19944 admissions	
No	42.00 %	15070 admissions	
Unknown	3.00 %	1084 admissions	
APACHE diagnostic group			
Cardiovascular			
	Surgical	23.68 %	8547 admissions
	Nonsurgical	12.96 %	4677 admissions
Neurological			
	Surgical	11.32 %	4088 admissions
	Nonsurgical	17.36 %	6266 admissions
Respiratory			
	Surgical	1.87 %	675 admissions
	Nonsurgical	7.46 %	2694 admissions
Trauma			
	Surgical	0.71 %	258 admissions
	Nonsurgical	4.45 %	1608 admissions
Other		6.12 %	2209 admissions
Unknown		0.01 %	5 admissions
Circulatory dysfunction			
Patients with events	30.60 %	11046 admissions	
Mean #events per patient with events	4.16		
Mean event duration	320 minutes		
Mean time to first event	492 minutes		
Mortality	6.11 %	2205 admissions	
APACHE score			
Mean (std)	17.46 (7.86)		
Median (25%, 75 %)	16 (12,22)		
Range	[0,57]		

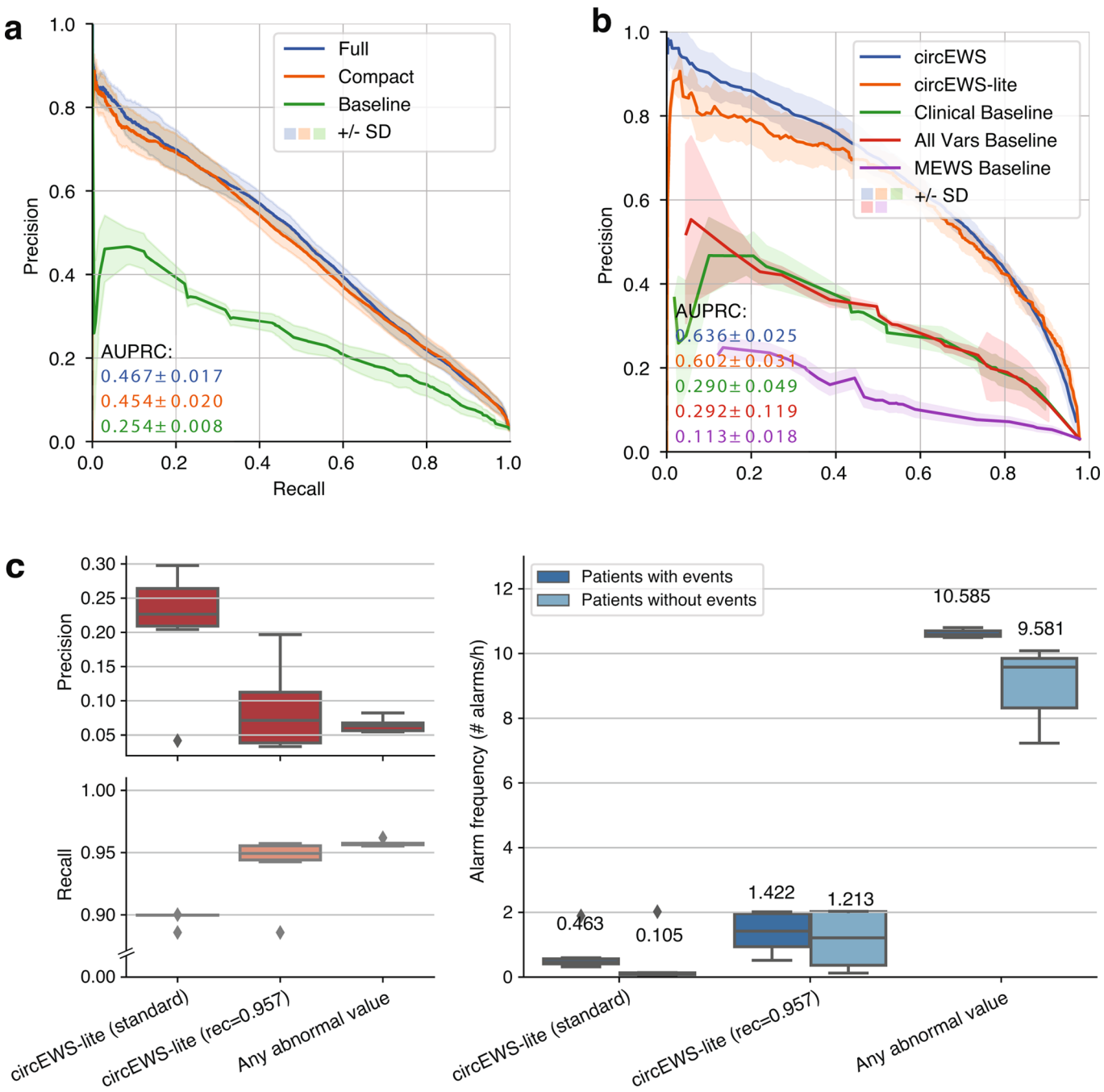
Extended Data Fig. 3 | Patient characteristics.



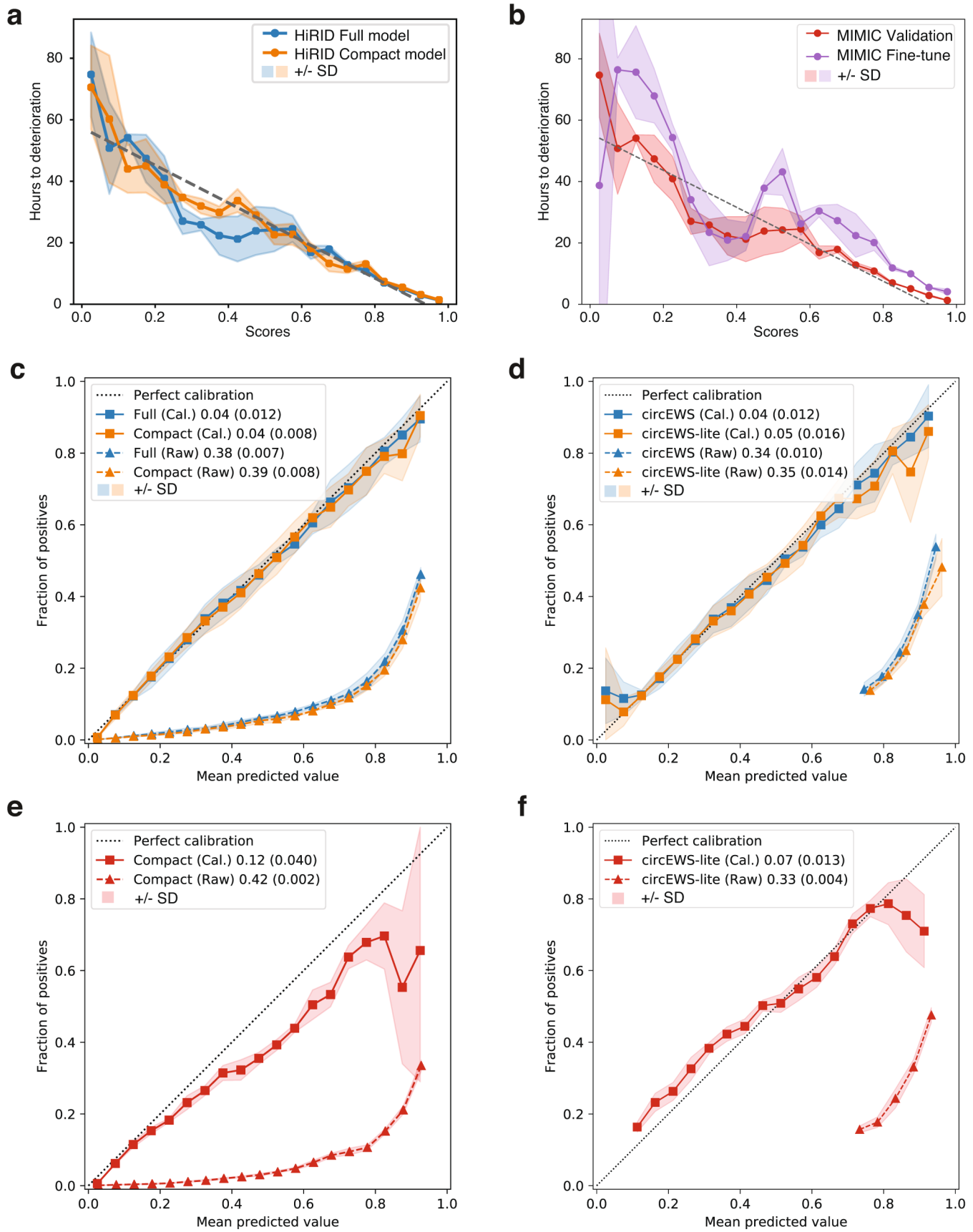
Extended Data Fig. 4 | Correlation of duration of circulatory failure and mortality. a, Mortality rate as a function of the duration of circulatory failure expressed as a fraction of length of stay in ICU. **b**, Cumulative mortality rate as a function of the frequency of occurrence of circulatory failure per hour of ICU admission. **c**: Decreasing trend of event onset occurrence by days of the patient stay in the ICU. **d**, First decreasing and then increasing trend of mean APACHE scores during the patient stay in the ICU.



Extended Data Fig. 5 | Model & Training. a, **b**: Effect of training set size. Analysis of the effect of training set size on model performance by artificially subsampling patients at random and retraining the model. This analysis was performed using the circEWS alarm system evaluation policy. We observed that model performance decreases drastically when subsampling to less than 5% of the original training set size, and that the model did not show obvious saturation effects as we move to the full size of the data. **c**, **d**: Comparison of machine learning models. Comparison of machine learning approaches using ROC/PR curves. A linear model baseline (logistic regression; "LogReg"), a tree-ensemble based method (based on lightGBM, "GBM"; used to construct circEWS), an individual decision tree (based on lightGBM, "DecTree"), and a recurrent neural network ("LSTM") were compared. The Tree models received identical input as given to GBM. The LogReg and LSTM received normalized feature values. We observed that gradient-boosting ensembles clearly outperform the other methods, followed by LogReg and LSTM/Tree models. **e**, **f**: Effect of number of features. We studied the effect of the number of features on model performance. Features were ranked by respective mean absolute SHAP values in a model trained on 50% of the data using all 5,278 features. The dashed line indicates results from the model trained on 50% of the data; using the full training data is computationally prohibitive. For 1,763 features the mean absolute SHAP value is non-zero. Inclusion of more features increases model performance until a saturation point occurs. SD: standard deviation.

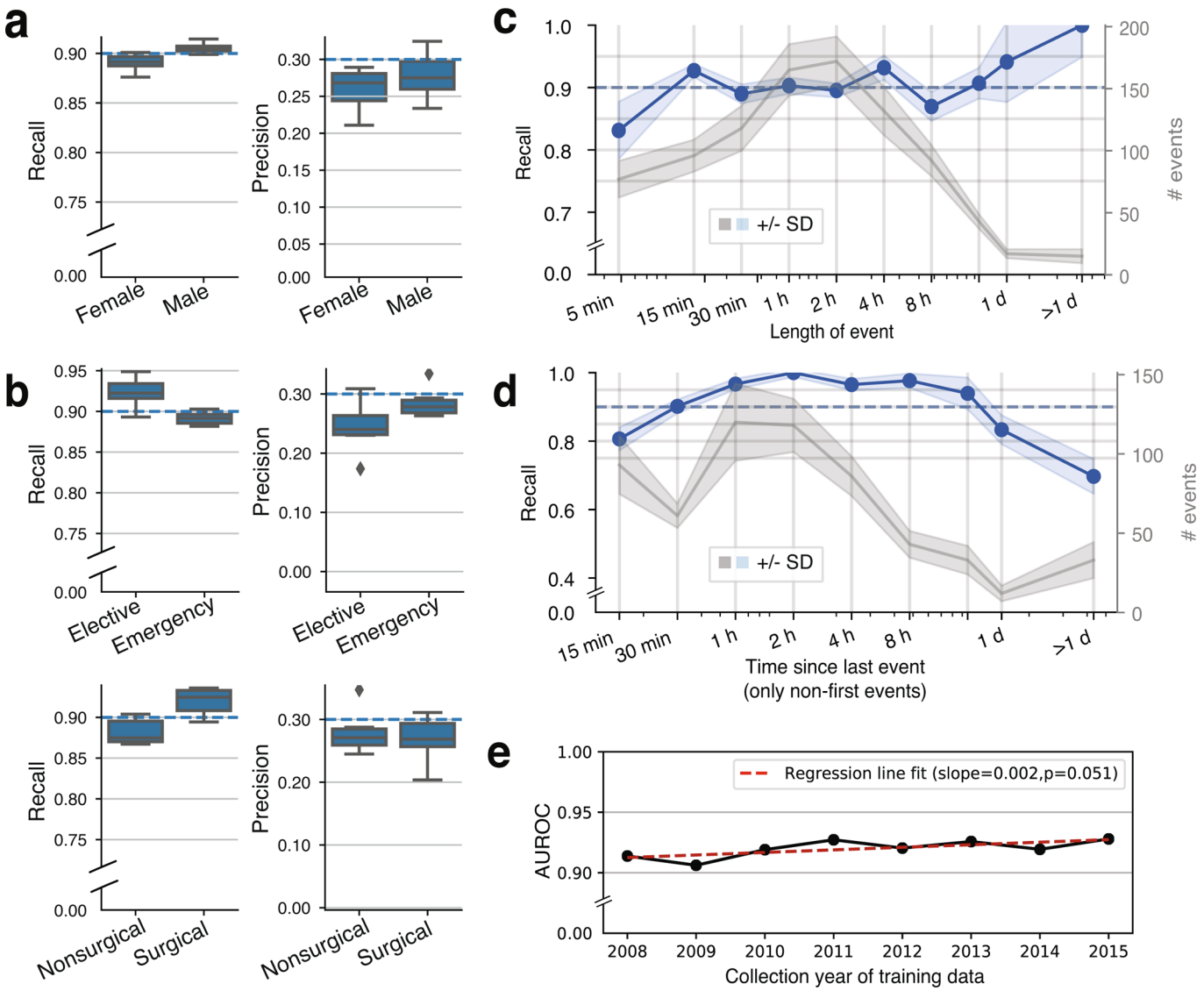


Extended Data Fig. 6 | Baseline Variations. a, Time-slice based labeling. Precision-recall curve comparing the full, compact, and baseline models on the task of predicting circulatory failure. **b**, Baseline variations. Additional decision-tree based baselines based on (**a**) all raw variables, (**b**) all variables included in the MEWS score (systolic blood pressure, heart rate, respiratory rate, temperature, AVPU score), (**c**) baseline including variables from endpoint definition (MAP, lactate). The MEWS baseline can be considered as a vital-sign-based baseline which mimics simple signal-processing algorithms based on constant thresholds. **c**, We consider a simple alarm system that raises an alarm whenever any abnormal value is observed for one of the circEWS-lite variables (excluding static and pharmaceutical variables). This system is comparable with current clinical practice in which individual monitoring modalities raise an alarm whenever observed values are abnormal. The event recall for this system is 0.957. The number of alarms for patients with and without events as well as the precision and recall rates of this system are compared to circEWS-lite with standard (recall = 0.90) and increased recall (recall = 0.957) setting. We observe that circEWS-lite generates 20 or 80 times fewer alarms than the abnormal-value based alarm system at the same recall rate for patients with or without events, respectively. The high number of alarms of the simple system can be partially explained by missing key features like alarm silencing and resetting, but also because a threshold-based system cannot combine information from multiple variables to make accurate predictions. SD: standard deviation.

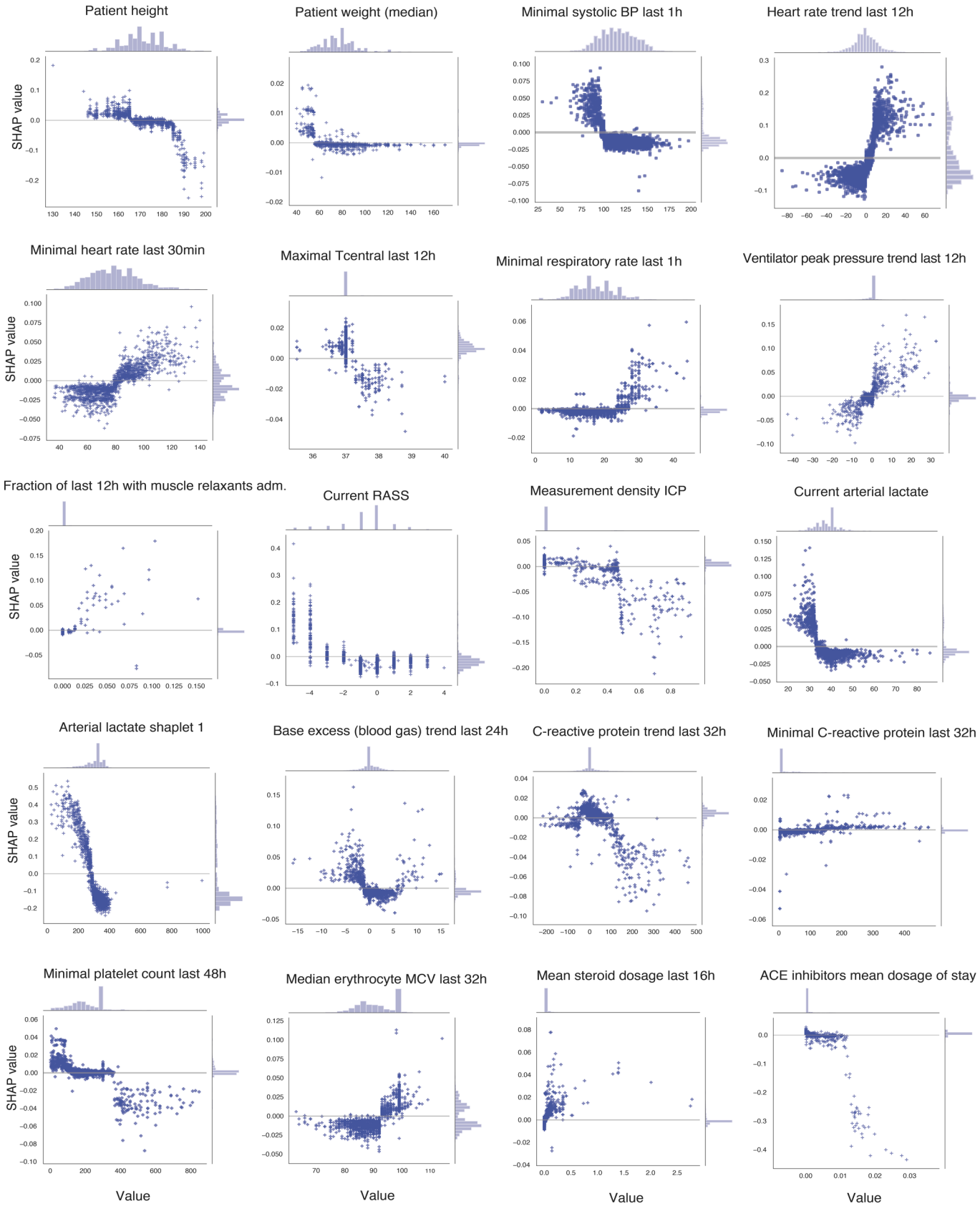


Extended Data Fig. 7 | See next page for caption.

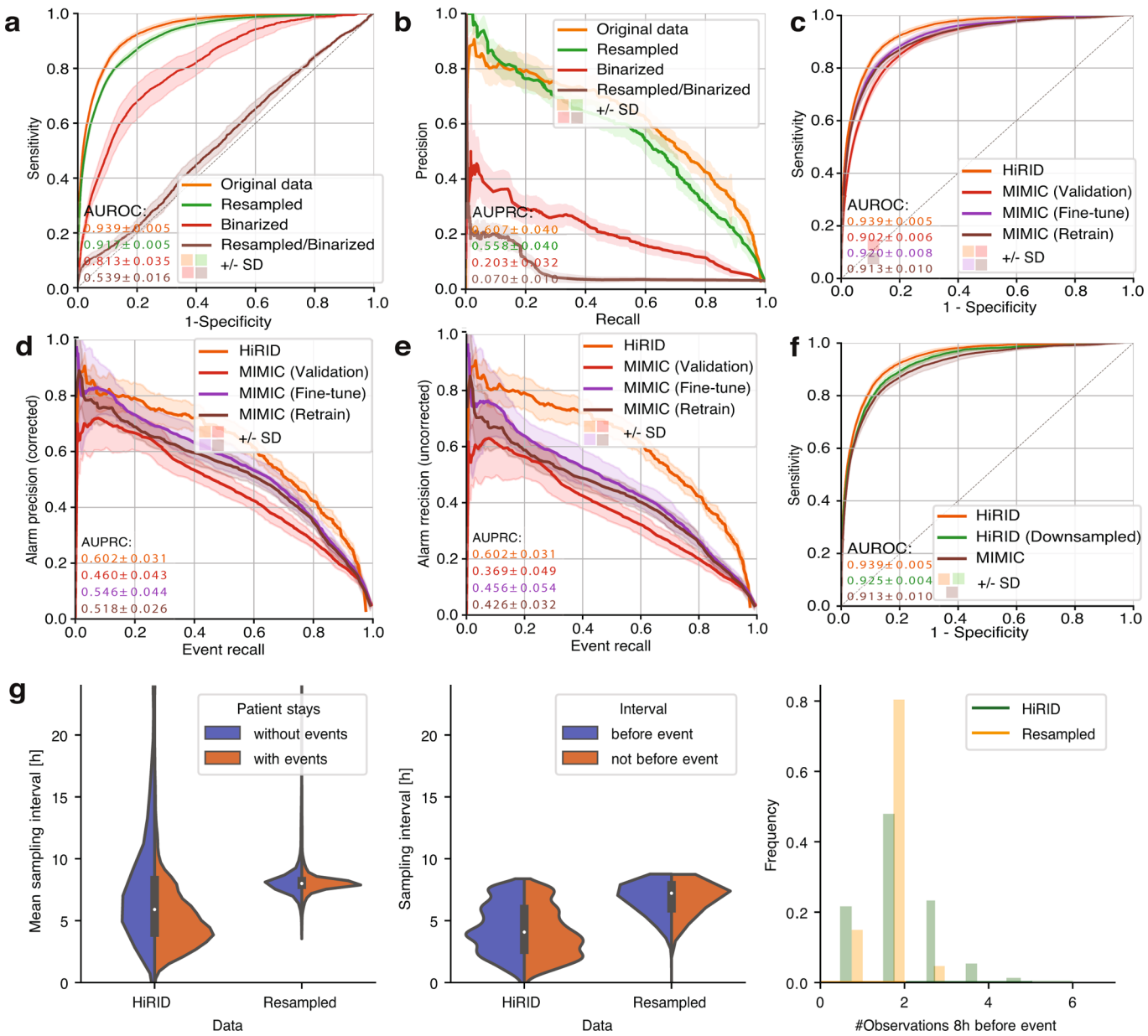
Extended Data Fig. 7 | Model calibration. a, **b**, Observed average time to circulatory failure vs. the raw prediction scores of the full and compact models on HiRID (**a**) and MIMIC (**b**). The dotted diagonal lines show a regression line fit to the observed average time to failure for the compact model (same model on HiRID and MIMIC, r-value −0.96/−0.93) **c**, **d**, Observed risk of circulatory failure within the next 8 h vs. the continuous risk scores (**c**) and the scores at time points when an alarm is produced (**d**) of the full and compact model (Raw) and their post-hoc calibrated counterparts (Cal.), (isotonic regression). The legend displays the mean/std of the absolute area between the calibration curve and the ideal calibration reference curve. Before computing areas, all curves were resampled to a regular grid covering [0,1]. The raw scores of the full and the compact model are not natural proxies for the probability of being in a 8 h window prior to circulatory failure due to the choice of the machine learning model (decision tree ensembles). However, they can be calibrated post-hoc using isotonic regression and then exhibit almost ideal calibration with a Brier score of 0.02 (not shown on plot) and an area around the diagonal reference curve of 0.04. **e**, **f**, Calibration of the (calibrated) compact model / circEWS-lite trained on HiRID and applied on MIMIC. The legend displays the mean/std of the absolute area between the calibration curve and the ideal calibration reference curve. Before computing areas, all curves were resampled to a regular grid covering [0,1]. We observe a slight overestimation of risk and a more unstable calibration across temporal splits. Calibration is still appropriate and does not seem to be strongly affected by patients drawn from a different population. The curves were corrected for the different label prevalence in the MIMIC cohort compared to the HiRID cohort. SD: standard deviation.



Extended Data Fig. 8 | Performance of circEWS in different patient sub-cohorts and over time. a, **b**, circEWS performance in sub-cohorts categorized by gender and admission types. **c**, Effect of event duration on model performance. Very short (=5 min duration) events had the lowest recall, indicating that these may be spuriously labeled events. The model appears to excel at identifying very long events, however, the sample size is low (15 events with a duration longer than a day). **d**, Effect of time since previous event on model performance. The model exhibits lower recall for very short (<30 min) time periods after a previous period of circulatory failure. **e**, Other temporal generalization results. AUROC values for the temporal generalization experiment where the AUROC obtained for training data from different years is shown in black, and a linear regression line fit is shown as a dotted red line. A non-zero slope is significant at 10 % level with two-sided p-value of 0.051 using the Wald test. SD: standard deviation.



Extended Data Fig. 9 | Further examples of relationship between SHAP value and feature value. Features were selected from the top 500 according to perceived clinical relevance and interpretability.



Extended Data Fig. 10 | External validation and bias by intensity of monitoring. Effect of resampling and/or binarizing measurements in the test set to study "bias by intensity of monitoring". Model performance does not strongly decrease if circEWS-lite is deployed to a test set with artificially created regular sampling for all variables ("Resampled" curve), compared with the original test set ("Original data" curve). The performance of a model based only on information about monitoring intensity using only binary measurement indicators has a much lower performance ("Binarized" curve). This effect is particularly strong when this model is applied to a data-set with both binarization/regular sampling ("Resampled & Binarized") and results in close-to random performance. **a**, ROC curve, **b**, PR curve showing event-based analysis of the alarm system. **c:** External validation of the compact model using ROC-based analysis. **d**, **e,** Event-based PR (un)corrected. Event-based based PR curve comparing the performance of the circEWS-lite alarm system on the HiRID data-set and the three external validation settings (MIMIC Validation, Retrain and Fine-tune). In d), the MIMIC curves are corrected for the substantially lower label prevalence in the MIMIC data-set, implying equal performance of random classifiers on both data-sets. In e), the curves are not prevalence corrected. **f**, Resolution comparison. Performance of the compact model in the original HiRID data, an artificially downsampled HiRiD data-set to approximate the time-resolution of the MIMIC data-set, as well as the MIMIC data-set, in terms of ROC-based metrics. **g**, The measurement sampling intervals for lactate comparing HiRID to the resampled version (targeting baseline frequencies) used for the "bias by intensity of monitoring" experiment. On the left, the average interval between measurements per patient is shown. The middle panel depicts the sampling interval in 8 h windows before events (orange) and in windows not before events (blue). The right panel shows the number of observations in the 8 h window before the event comparing the resampled to the original dataset. We observe that after resampling the distribution of sampling interval is unimodal and closely concentrated on the baseline frequency. SD: standard deviation.

nature research

Corresponding author(s): Tobias Michael Merz

Last updated by author(s): Dec 20, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PDMS; GE Centricity Critical Care, General Electrics, Helsinki, Finland
Data analysis	Python3, with numpy, pandas and scikit-learn

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

More information on HiRID is available on hirid.intensivecare.ai and the full dataset can be downloaded from physionet.org. The computer code used in this research is available at www.github.com/ratschlab/circEWS under an open-source license.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We report the use of a state-of-the-art machine learning techniques to construct two early warning systems for circulatory failure in ICU patients using a large ICU patient data base. The study included data that was collected for routine clinical management over a period of 10 years. All patients admissions recorded in the period between the implementation of the ICU electronic patient data management system (PDMS; GE 478 Centricity Critical Care, General Electrics, Helsinki, Finland) in April 2005 and August 2016 were analyzed and information from 36,098 patient admissions were used for model development. The PDMS was used to prospectively register patient health information, measurements of organ function parameters, results of laboratory analyses and treatment parameters from ICU admission to discharge. The sample size was determined by the size of the data base. There is no established method to calculated sample sizes for prognostic models using ML methods. For more standard model development methods using a multivariable regression framework a minimum sample size that allows for 10 to 20 events per predictor parameter (EPP) is generally recommended. Our data contains more than 45'000 circulatory failure events and we estimated to include less than 1000 predictor variables into the models.
Data exclusions	Patient admissions prior to 2008 were excluded from the analysis due to frequent changes in variable identifiers during the run-in phase of the PDMS implementation. Patients without data for determining circulatory failure and patients receiving any form of full mechanical circulatory support, younger than 16 years or older than 100 years, or actively declining the use of their data for research purposes were excluded.
Replication	The study reports the use of a state-of-the-art machine learning techniques to construct two early warning systems for circulatory failure in ICU patients. The model development requires many itinerations of data processing. External validation was performed using unrelated data from a different ICU cohort and hospital. We were able to successfully replicate our results in the external data set.
Randomization	Our study design did not include any allocation of subjects into different groups for comparison purposes. Therefore no randomization process was necessary.
Blinding	Our study design did not include any allocation of subjects into different groups for comparison purposes. Allocation concealment was therefore not necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging