International Journal of Medical Informatics 145 (2021) 104312

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation

Bernhard Wernly [a,b,*], Behrooz Mamandipoor [c], Philipp Baldia [d], Christian Jung [d], Venet Osmani [c]

[a] Department of Cardiology, Paracelsus Medical University of Salzburg, Austria
[b] Division of Cardiology, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden
[c] Fondazione Bruno Kessler Research Institute, Trento, Italy
[d] University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf, Medical Faculty, Division of Cardiology, Pulmonology and Vascular Medicine, Germany

ARTICLE INFO

Keywords:
critically ill
artificial intelligence
machine learning
deep learning
LSTM
ICU
risk stratification
intensive care unit
critical care
sepsis

ABSTRACT

Purpose: : To evaluate the application of machine learning methods, specifically Deep Neural Networks (DNN) models for intensive care (ICU) mortality prediction. The aim was to predict mortality within 96 hours after admission to mirror the clinical situation of patient evaluation after an ICU trial, which consists of 24-48 hours of ICU treatment and then "re-triage". The input variables were deliberately restricted to ABG values to maximise real-world practicability.

Methods: : We retrospectively evaluated septic patients in the multi-centre eICU dataset as well as single centre MIMIC-III dataset. Included were all patients alive after 48 hours with available data on ABG (n = 3979 and n = 9655 ICU stays for the multi-centre and single centre respectively). The primary endpoint was 96 -h-mortality.

Results: : The model was developed using long short-term memory (LSTM), a type of DNN designed to learn temporal dependencies between variables. Input variables were all ABG values within the first 48 hours. The SOFA score (AUC of 0.72) was moderately predictive. Logistic regression showed good performance (AUC of 0.82). The best performance was achieved by the LSTM-based model with AUC of 0.88 in the multi-centre study and AUC of 0.85 in the single centre study.

Conclusions: : An LSTM-based model could help physicians with the "re-triage" and the decision to restrict treatment in patients with a poor prognosis.

Take-home message: Machine learning based models could help physicians in the decision process evaluating therapy targets after an initial ICU-trial.

1. Introduction

Limited resources of intensive care units (ICUs), especially in terms of beds and personnel, require efficient management of resources and patients. This is particularly important when external stressors such as a pandemic with the outbreak of SARS-CoV-2 increase patient numbers and pressure on the triage system [1]. In order to optimize the allocation of intensive care resources, risk stratification and prediction of outcomes must be as precise and rapid as possible [2].

Machine learning is particularly suitable for predicting outcomes in intensive care settings because it integrates complex information from different data sources relatively easily. [3,4]. Several studies already proved the applicability of machine learning in the ICU setting; especially for sepsis detection machine learning based methods showed an earlier and precise detection of sepsis when compared to usual clinical scores like the SOFA score [5,6].

However, while machine learning can process complex and divergent information, the possibility of implementing an algorithm is reduced with the number of parameters taken into account (and thus necessary for the calculation). Therefore, with respect to the external validity of machine learning tools, it may be useful to voluntarily limit the number of parameters to common and common parameters. Arterial blood gas (ABG) tests, which are globally standardized on ICU and collected at a relatively high frequency, can be used for this purpose.

* Corresponding author at: Clinic of Internal Medicine II, Department of Cardiology Paracelsus Medical University of Salzburg, Müllner Hauptstraße 48 5020, Salzburg, Austria.
E-mail address: bernhard@wernly.at (B. Wernly).

https://doi.org/10.1016/j.ijmedinf.2020.104312
Received 10 June 2020; Received in revised form 26 September 2020; Accepted 20 October 2020
Available online 24 October 2020

Particularly in patients with sepsis - a multi-organ disease per se - multiple interlayers of information need to be interpreted [7]. The physicians currently integrate these data pieces and puzzles in clinical scores, individual judgment based on data, the physical and established guidelines, and at times, plain clinical gut feeling. However, work on ICU wards is often organised in shifts, and experience and judgment might differ significantly between physicians. Therefore, more objective measures of patients and disease trajectories could improve outcomes and decision process.

A particular challenge to clinicians is the decision when to restrict treatment [2,8]. Ideally, patients who do not benefit from intensive care should not be admitted to the ICU. However, this critical judgment is often impossible at admission when only limited data on the patient is available. Further, the *angst* of legal proceedings when limiting treatment is common among physicians. Therefore, in practice, many patients undergo an ICU-trial, even if not formally stated: patients in whom the benefit of intensive care is unclear are admitted to ICU and undergo unrestricted treatment for up to 48 hours [9]. During that time, further information on concomitant diseases, pre-admission frailty, and responsiveness to treatment are judged. Then, in theory, a decision about additional intensive care or a treatment restriction is made by the physicians together with other health-care providers, the patient, and his family. However, given the vast amount of information generated by a patient's ICU stay, this judgment is often tricky and outcome prediction remains challenging. Consequently, machine learning methods that intrinsically integrate a large amount of data could play an important role in supporting clinical decision making.

Typically, machine learning models integrate many data sources, such as biomarkers, ventilation settings, blood pressure and medications including catecholamines. However, availability of all this data differs significantly between ICUs, depending on how well they are equipped. Highly equipped ICUs might generate high quantities of data, however the application of machine learning models should not be limited to high-end ICUs only. As such, the aim of this study was to evaluate AI for mortality prediction based on a sweet-spot, combining the necessary amount of data while having broad applicability in ICUs in both secondary and tertiary settings. As ABG tests are widely available, this study aimed to evaluate the predictive capabilities of machine learning models based on the analysis of ABGs variables only. Specifically, we focused on sepsis, given the complex disease trajectory, which usually generates a broad array of data. Further, we aimed to mirror the clinical practice of an ICU-trial, where physicians need to judge the state of a patient after 48 hours. In this respect we develop a machine learning model based on Recurrent Neural Networks (RNN), specifically Long Short Term Memory (LSTM). We specifically chose LSTM, as this type of network is designed to capture temporal dependencies between variables (ABG values in our case) on longitudinal data. We evaluate our model based on the data of n = 3979 and n = 9655 ICU stays for the multi-centre and single centre respectively. Furthermore, we compare the performance of our model against baseline Logistic Regression, Baseline Lactate and SOFA Score.

While, there have been several works on mortality prediction [10], this is the first multi-centre study to address mortality prediction across diversely equipped ICUs, by relying solely on routinely available ABG values.

2. Methods

2.1. Data sources

We evaluate our method both on a multi centre dataset (eICU) as well as a single-centre dataset (MIMIC III). eICU dataset contains data associated with 200,859 admissions collected from 335 ICUs across 208 hospitals in the US admitted between 2014 and 2015 [11], while MIMIC-III dataset contains data associated with 61,532 distinct hospital admissions for adult patients (aged 16 years or above) admitted between 2001 and 2012 [15]. This study included all patients of both datasets diagnosed with sepsis based on the method established by Angus et al., in identifying patients using billing codes [12].

2.2. Ethics statement

The analysis using the eICU Collaborative Research Database (eICU-CRD) is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2). The data in the Medical Information Mart for Intensive Care (MIMIC) has been previously de-identified, and the institutional review boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the database for research. This study was an analysis of publicly available, anonymised databases with pre-existing institutional review board (IRB) approval; thus, no further approval was required.

2.3. Study subjects and lactate concentrations

We retrospectively evaluated a sub-group of patients of eICU and MIMIC-III databases. Patients datasets were analysed for the first 48 hours after the admission. All cases of mortality that occurred after the initial 48 hours and up to 96 hours were labelled as "dead" and were included in the model validation. Furthermore, we included all patients alive after 48 hours that had complete ABG data (n = 3853 patients, corresponding to 3979 ICU stays for the multi-centre study and n = 8061 patients, corresponding to 9655 ICU stays for the single centre study). The patient datasets were analysed separately, and the primary endpoint was ICU mortality between 48 to 96 hours after patient admission.

In addition to septic patients, we also evaluated the performance of the model for septic shock patients. For this evaluation, we selected patients that had initial lactate greater than 2.0 mmol/L (and > 18 mg/dL respectively) and documented the use of vasopressors, resulting in 1044 and 417 patients for the multi-centre and single-centre evaluation respectively. Based on this cohort, we selected a matching cohort (based on age and gender) of patients that were not under septic shock, resulting in 859 and 437 patients. Therefore, the overall dataset resulted in 1481 and 1276 patients with a mortality rate of 22.6% and 32.4% for both datasets respectively (for the single-centre and multi-centre analysis).

2.4. Statistical Analysis

Continuous variables are expressed as mean (± standard deviation) and compared using ANOVA. Categorical data are expressed as numbers (percentage). A chi-square test was applied to calculate differences between groups. Both univariable and multivariable logistic regression analysis to adjust for confounding factors for ICU mortality was done. As shown in Fig. 1a and Fig. 1b there were no strong linear correlations between each variable and the target outcome for both multi-centre and single-centre patient dataset.

To assess the predictive performance of our model, we calculated a range of common performance metrics, including positive predictive value (PPV), negative predictive value (NPV), sensitivity versus specificity, the area under the ROC curve (AUC) as well as the precision-recall plot and the area under its curve (AUPRC), also called average precision (AP). We also include Mathews correlation coefficient (MCC), which represents a correlation between the observed and predicted binary classifications. MCC is used in machine learning to measure the quality of binary classifications producing values that range from -1 (total disagreement between prediction and observation) to +1 (a perfect prediction).

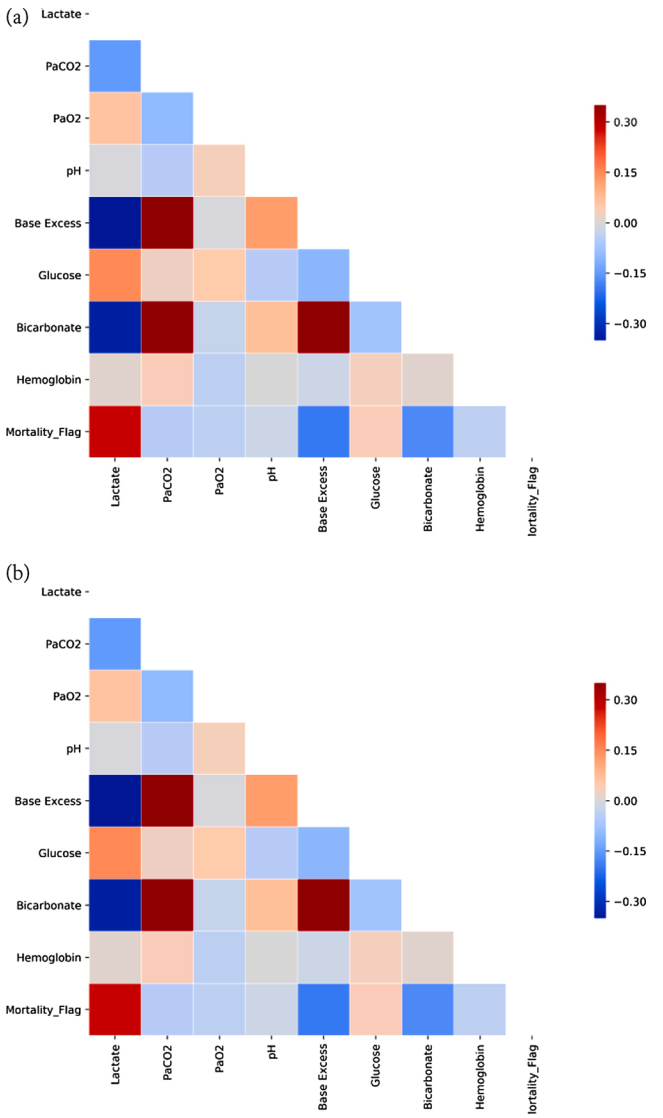

Fig. 1. a Pearson correlation between each variable and target outcome (denoted with Mortality Flag) for the multi-centre eICU dataset.
b Pearson correlation between each variable and target outcome (denoted with Mortality Flag) for the single-centre MIMIC III dataset. Note, the correlation interval is from -0.4 to 0.4 as defined by the scale on the right.

2.5. Machine learning model

The model predicts the likelihood that a patient will die in the next 48 hours based on the analysis of patient-specific variables, namely arterial blood gas (ABG) values. Even though additional variables were available in the dataset, we deliberately chose to include only ABG values as they are routinely measured, ensuring broad applicability of our method across diversely equipped ICUs. Only patients with features that were documented at least once were included, and the rest of the patients were excluded. The model was developed using Recurrent Neural Networks (RNN), a type of Deep Artificial Neural Network designed to learn temporal dependencies between variables. In particular, the long-short term memory (LSTM) network, a type of RNN designed to handle longitudinal time-series data was chosen [17]. The input layer of the LSTM network is composed of 8 neurons that receive time-ordered sequence of values of patients' ABG variables as shown in Fig. 2. At each time step, the network computes its internal state s_t, based on the input vector x_t of patient features. As time progresses, newly available data x_{t+1} are used to update the state of the network and generate the likelihood of mortality o_t given the current patient's state and their clinical history (past ABG values). As shown in Fig. 2 recurrent neural network parameters U, V, and W are adjusted automatically as the model learns to map mortality outcome to longitudinal ABG values (model training phase).

The neural network model is composed of a single hidden LSTM layer comprising 140 fully connected neurons, tanh activation function and Xavier normal weight initializer. The output layer is a sigmoid function used to classify each data sample in two outcomes. All the models are trained for 50 epochs with the batch size of 100 based on the binary cross-entropy loss function and Adam optimizer method with learning rate of 0.001. Furthermore, we used dropout of rate 0.4 to avoid model overfitting, such that our model can generalise its predictions also on data of future patients. Dropout is a regularisation method that discards a percentage of artificial neurons during the model derivation to force the network to learn a more robust internal representation.

Finally, the performance of the models was evaluated using stratified three-fold cross-validation with 10 times repetition. The data is randomly split into model derivation dataset (66.67%) and model validation dataset (33.33%) for each fold. To reduce possible bias and evaluate generalizability, we repeat three-fold cross-validation 10 times, where the prediction of each model is used to calculate the mean and standard deviation.

The model derivation dataset was further divided into training and tuning parts, with 75% of the data used to optimise the weights of the neural network model (training). Neural network hyperparameters (including learning rate, depth of the neural network, and size of the hidden layers) were also optimised using 25% of the training data (tuning). It is worth noting that other variants of RNNs, including GRU and Bidirectional LSTM [18,19], were also evaluated, however, there was no improvement in the prediction result.

3. Results

The multi-centre dataset contained 3979 ICU stays, while the single-centre dataset contained 9655 ICU stays, with complete ABG data (at least one measurement is available for each ABG variable during the first 48 hours of ICU stays) as shown in selection cohort in Fig. 5. Data preparation consisted of several steps in which possible outliers and noisy measurements were removed from the datasets using valid intervals for each ABG variable based on clinical knowledge. Secondly, handling slightly different frequencies of recording of the variables in the databases were done by aligning and ordering in time all the measurements. Finally, missing values were handled using forward filling strategy, which uses the last and nearest valid measurements for imputation by forward propagating each available measurement.

From the admission to 48 hour time point, 340 patients died in the multi-centre evaluation, while 507 patients died in the single centre evaluation, where their ABG values were not included in the model validation so as to mirror the clinical situation (the clinician only evaluates patients who are alive and still at the ICU).

Of the included ICU stays, 492 deaths were recorded until time point 96 hours for the multi-centre evaluation, while 858 for the single-centre. Compared to survivors, non-survivors evidenced higher age, lactate, and creatinine levels at admission for both single-centre and multi-centre dataset. They were clinically sicker, mirrored by higher initial SOFA scores as shown in Table 1a and Table 1b for both the single-centre and multi-centre patient dataset respectively (Table 2a and 2b).

3.1. Survival analysis results

Overall mortality was 19.1% (n = 759) and 19.3% (n = 1866) for the multi-centre and single-centre respectively, while mortality up to 96 hours was 12.4% and 8.9%. The SOFA score over 48 hours (using the worst clinical values) was only moderately predictive (AUC of 0.72 and 0.76) [13]. The initial lactate concentration was also moderately

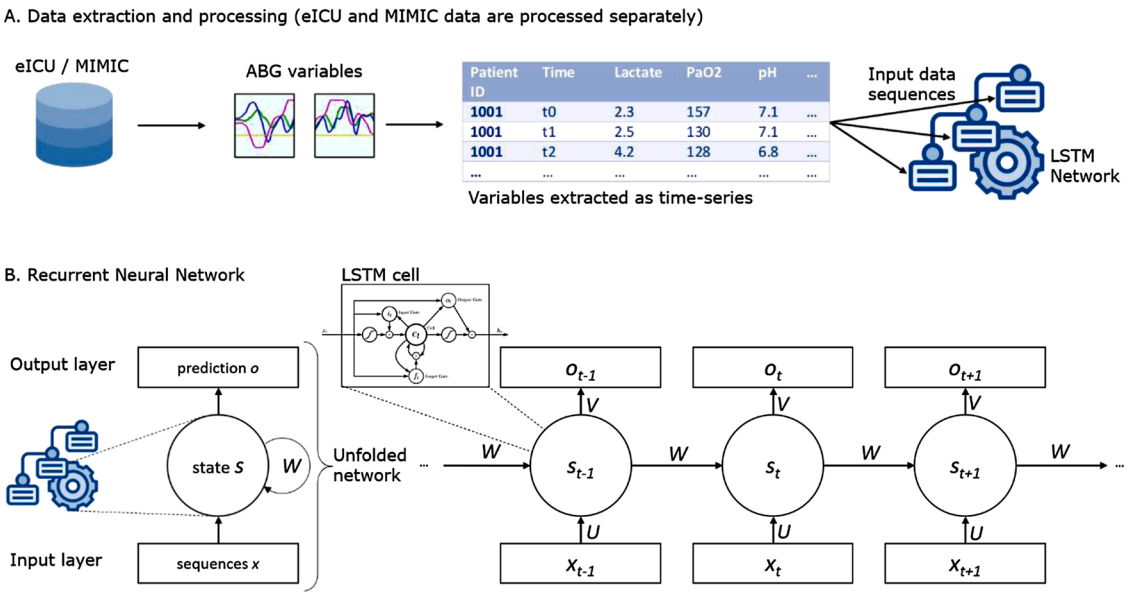

Fig. 2. Overview of data extraction, processing (A) and model development (B). Variables are extracted from the multi-centre eICU database and single-centre MIMIC III database and are processed separately. W, U, V represent neural network parameters (weight matrices for input, previous state and output respectively).

Table 1a
Baseline demographics at admission of survivors versus non-survivors (survival between 48 and 96 hours) for the single-centre MIMIC-III patient population.

	Survivors	Non-survivors	p-value
Female sex n (%)	3939 (45)	393 (46)	
Age (years)	65.6 ± 16.1	70.1 ± 14.7	<0.01
Weight (kg)	82.2 ± 26.3	80.4 ± 24.9	<0.01
O2 Saturation (%)	95.9 ± 6.9	93.5 ± 10.7	<0.01
PaCO2 (mmHg)	42.9 ± 13.8	42.8 ± 16.7	0.88
PaO2 (mmHg)	157 ± 115.6	133.2 ± 105.1	<0.01
pH	7.1 ± 0.6	7.0 ± 0.6	<0.01
MAP (mmHg)	10.7 ± 4.1	11.4 ± 4.4	<0.01
Heart Rate (bpm)	94.1 ± 21.4	100.0 ± 22.6	<0.01
Respiratory Rate (bpm)	20.3 ± 7.2	22.5 ± 7.1	<0.01
Body temperature (°C)	36.5 ± 2.7	36.1 ± 3.5	<0.01
CVP (mmHg)	13.6 ± 22.3	15.8 ± 25.2	0.015
Creatinine (g/dL)	1.7 ± 2.0	2.2 ± 1.6	<0.01
Hemoglobin (g/dL)	10.3 ± 2.0	10.3 ± 2.2	0.7
Lactate (mmol/L)	2.3 ± 1.8	4.4 ± 3.7	<0.01
Base Excess	−1.7 ± 5.9	−5.9 ± 7.8	<0.01
GCS Total (points)	9.7 ± 4.4	9.3 ± 4.5	<0.01
Glucose (mg/dL)	155.5 ± 76	155.7 ± 83.1	0.11
Bicarbonate (mmol/L)	22.6 ± 5.7	19.6 ± 6.7	<0.01
SOFA (points)	7.03 ± 3.6	11.3 ± 4.3	<0.01
SAPS (points)	21.2 ± 4.9	26.7 ± 5.8	<0.01
Ventilation n (%)	4161 (73)	425 (80)	0.05
Vasopressor use n (%)	364 (6)	182 (34)	0.02

Table 1b
Baseline demographics at admission of survivors versus non-survivors (survival between 48 and 96 hours) for the multi-center eICU patient population.

	Survivors	Non-survivors	p-value
Female sex n (%)	1699 (49)	248 (50)	
Age (years)	63.3 ± 15.4	67.8 ± 13.7	<0.01
Weight (kg)	86.3 ± 30.1	82.4 ± 31.8	<0.01
O2 Saturation (%)	95.7 ± 5.9	94.6 ± 7.3	<0.01
PaCO2 (mmHg)	40.7 ± 14.1	37.8 ± 15.2	<0.01
PaO2 (mmHg)	113 ± 76.7	125.7 ± 84.7	<0.01
pH	7.3 ± 0.1	7.2 ± 0.2	<0.01
MAP (mmHg)	12.5 ± 14.3	12.5 ± 4.6	<0.01
Heart Rate (bpm)	100.8 ± 22.4	102.4 ± 24.4	0.2
Respiratory Rate (bpm)	23.2 ± 8.4	24.5 ± 7.8	<0.01
Body temperature (°C)	35.5 ± 2.2	34.0 ± 2.2	<0.01
CVP (mmHg)	14.6 ± 28.4	17.7 ± 40	<0.01
Creatinine (g/dL)	2.01 ± 1.8	2.5 ± 1.7	<0.01
Hemoglobin (g/dL)	10.5 ± 2.2	10.2 ± 2.4	<0.01
Lactate (mmol/L)	2.5 ± 2.3	6.1 ± 4.4	<0.01
Base Excess	−3.4 ± 6.9	−10.5 ± 8.0	<0.01
GCS Total (points)	11.6 ± 3.6	10.8 ± 4.3	<0.01
Glucose (mg/dL)	157.1 ± 92	144.1 ± 89	<0.01
Bicarbonate (mmol/L)	21.8 ± 6.5	17 ± 6.4	<0.01
SOFA (points)	8.05 ± 3.9	11.2 ± 3.7	<0.01
APACHE (points)	78.6 ± 26.6	110 ± 30	<0.01
Ventilation n (%)	1530 (44)	292 (60)	<0.01
Vasopressor use n (%)	1558 (45)	349 (70)	<0.01

predictive (AUC of 0.80 and 0.70) for 96 -h-mortality. Logistic Regression showed good performance in the multi-centre evaluation with AUC of 0.82 (± 0.01), while its performance on the single-centre was an AUC of 0.81 (± 0.01). The LSTM-based model achieved the best performance in both studies with AUC of 0.88 (± 0.01) and 0.85 (± 0.01) with PPV of 0.60 (± 0.05) and 0.43 (± 0.07); NPV of 0.90 (± 0.03) and 0.96 (± 0.01); and MCC of 0.63 (± 0.05) and 0.47 (± 0.04) for the multi-centre and single-centre respectively. A tabular summary of the results can be found in Table 3 and Table 4. Predictive performance comparison of each model for both the multi-centre and single-centre evaluation is shown in Fig. 3a and Fig. 3b respectively with both AUC as well as area under the precision-recall curves (AUPRC). Furthermore, we also show the results of partial Area Under the Curve (AUC) along with partial c-statistic, developed in to illustrate the robustness of our model in addressing imbalanced datasets, as shown in Tables 5 and 6 [14].

3.2. *Prediction results on septic shock patients*

In addition to septic patients, we also evaluated the performance of the model for septic shock patients as described in the Methods section. The mean performance of the model on these patients resulted in AUC of 0.89 (± 0.03) for the multi-centre study and AUC of 0.93 (± 0.01) for the single-centre study and as shown in Fig. 4 a and 4b. Furthermore, other evaluation metrics also showed high performance with PPV of 0.72 (± 0.05), NPV of 0.90 (± 0.03) and MCC of 0.63 (± 0.05) for the multi-centre dataset; and PPV of 0.74 (± 0.05), NPV of 0.95 (± 0.02) and MCC of 0.70 (± 0.05) for the single-centre dataset.

4. Discussion

In this multi-centre retrospective analysis of a large cohort of

Table 2a
Baseline demographics at 48 hours of survivors versus non-survivors (survival between 48 and 96 hours) for the single-centre MIMIC-III patient population.

	Survivors	Non-survivors	p-value
Female sex n (%)	3939 (45)	393 (46)	
Age (years)	65.6 ± 16.1	70.1 ± 14.7	<0.01
Weight (kg)	82.2 ± 26.3	80.4 ± 24.9	<0.01
O2 Saturation (%)	96.7 ± 4.5	83.5 ± 23.0	<0.01
PaCO2 (mmHg)	40.4 ± 9.9	42.0 ± 17.0	0.018
PaO2 (mmHg)	109.9 ± 52.6	106.9 ± 66.8	<0.01
pH	7.2 ± 0.5	7.2 ± 0.3	<0.01
MAP (mmHg)	10.2 ± 4.7	14.2 ± 6.3	<0.01
Heart Rate (bpm)	87.6 ± 17.4	72.5 ± 44.0	<0.01
Respiratory Rate (bpm)	20.1 ± 6.2	17.6 ± 11.1	<0.01
Body temperature (°C)	36.9 ± 5.7	36.8 ± 17.0	<0.01
CVP (mmHg)	16.5 ± 35.5	26.1 ± 53.8	0.25
Creatinine (g/dL)	1.7 ± 2.3	2.3 ± 1.5	<0.01
Hemoglobin (g/dL)	10.0 ± 1.5	10.0 ± 2.0	0.42
Lactate (mmol/L)	1.7 ± 1.2	6.0 ± 5.3	<0.01
Base Excess	−0.4 ± 4.6	−7.3 ± 8.5	<0.01
GCS Total (points)	10.8 ± 3.7	6.6 ± 4.0	<0.01
Glucose (mg/dL)	138.1 ± 49.2	156.8 ± 88.2	<0.01
Bicarbonate (mmol/L)	23.5 ± 5.1	18.4 ± 6.8	<0.01
SOFA (points)	7.03 ± 3.6	11.3 ± 4.3	<0.01
SAPS (points)	21.2 ± 4.9	26.7 ± 5.8	<0.01
Ventilation n (%)	4161 (73)	425 (80)	0.05
Vasopressor use n (%)	364 (6)	182 (34)	0.02

Table 2b
Baseline demographics at 48 hours of survivors versus non-survivors (survival between 48 and 96 hours) for the multi-center eICU patient population.

	Survivors	Non-survivors	p-value
Female sex n (%)	1699 (49)	248 (50)	
Age (years)	63.3 ± 15.4	67.8 ± 13.7	<0.01
Weight (kg)	86.3 ± 30.1	82.4 ± 31.8	<0.01
O2 Saturation (%)	96.04 ± 5.0	85.3 ± 19.0	<0.01
PaCO2 (mmHg)	38.8 ± 10.3	38.2 ± 14.3	0.3
PaO2 (mmHg)	100 ± 50.5	108.5 ± 72.8	0.3
pH	7.4 ± 0.1	7.2 ± 0.1	<0.01
MAP (mmHg)	11.8 ± 14.0	14.6 ± 5.5	<0.01
Heart Rate (bpm)	90.5 ± 19	73.7 ± 43.7	<0.01
Respiratory Rate (bpm)	21 ± 7.1	19.6 ± 10.6	0.4
Body temperature (°C)	35.6 ± 7.3	33.4 ± 10.0	<0.01
CVP (mmHg)	19.5 ± 42.8	38.3 ± 78.6	<0.01
Creatinine (g/dL)	1.7 ± 1.5	2.5 ± 1.5	<0.01
Hemoglobin (g/dL)	10.0 ± 1.8	9.6 ± 2.1	<0.01
Lactate (mmol/L)	2.0 ± 1.8	7.9 ± 5.8	<0.01
Base Excess	−2.0 ± 5.8	−10.5 ± 8.2	<0.01
GCS Total (points)	12.0 ± 3.3	7.5 ± 4.1	<0.01
Glucose (mg/dL)	144.5 ± 59.3	146.6 ± 78.8	0.5
Bicarbonate (mmol/L)	23.5 ± 5.1	16.4 ± 6.6	<0.01
SOFA (points)	8.05 ± 3.9	11.2 ± 3.7	<0.01
APACHE (points)	78.6 ± 26.6	110 ± 30	<0.01
Ventilation n (%)	1530 (44)	292 (60)	<0.01
Vasopressor use n (%)	1558 (45)	349 (70)	<0.01

Table 3
Performance of the model in the multi-centre dataset (eICU) in both septic patients and patients in septic shock

Sepsis	AUC	PPV	NPV	MCC
Lactate	0.80	-	-	-
Sofa	0.72	0.23	0.92	0.21
LR	0.82 ± 0.01	0.48 ± 0.01	0.95 ± 0.01	0.48 ± 0.01
LSTM	0.88 ± 0.01	0.60 ± 0.05	0.96 ± 0.01	0.59 ± 0.06
Septic Shock	AUC	PPV	NPV	MCC
LSTM	0.89 ± 0.03	0.72 ± 0.05	0.90 ± 0.03	0.63 ± 0.05

patients with sepsis, the LSTM-based model derived on ABG values within 48 hours after admission was highly predictive for mortality within the next 48 hours. Of note, the model outperformed both lactate concentrations as well as the SOFA score. This comparison provides further evidence of the predictive power of our LSTM-based model bearing in mind that our model was derived using AGB values only, while SOFA considers the functioning of six major organ systems. The rationale of this study was to evaluate the accuracy of the LSTM model for ICU mortality prediction after 48 hours to mirror the clinical situation of patient evaluation after an ICU trial, which usually consists of 24-48 hours of ICU treatment and then "re-triage". We hypothesised that machine learning models could help to identify patients with a very high likelihood of not surviving the ICU stay, in whom potentially palliative treatment could be more appropriate compared to further intensive care. To enable potential wide-spread use of such an algorithm, the input variables were deliberately restricted to ABG values only, which are widely available in ICUs around the world.

Table 4
Performance of the model in the single centre dataset (MIMIC) in both septic patients and patients in septic shock

	AUC	PPV	NPV	MCC
Lactate	0.70	-	-	-
Sofa	0.76	0.24	0.95	0.26
LR	0.81 ± 0.01	0.35 ± 0.01	0.96 ± 0.01	0.40 ± 0.01
LSTM	0.85 ± 0.01	0.43 ± 0.07	0.96 ± 0.01	0.47 ± 0.04
Septic Shock	AUC	PPV	NPV	MCC
LSTM	0.93 ± 0.01	0.74 ± 0.05	0.95 ± 0.02	0.70 ± 0.05

On the one hand, the voluntary restriction to ABG is also associated with some disadvantages. For example, the integration of additional laboratory values such as renal, cardiac and inflammatory biomarkers could refine the granularity of the data and improve the predictivity of the model. Furthermore, this is a retrospective analysis of ABG values, which were primarily initially collected according to clinical criteria. Thus, it is very likely that ABG was determined at higher frequencies in sick patients or patients with volatile lactate concentrations. These are a priori limitations of our analysis. On the other hand, the frequency, granularity and distribution of ABG values reflect clinical reality. Since medicine is an art and AI can only be another new (and in our opinion promising) arrow in the clinicians' quiver, this unbiased collection of ABG values can also be interpreted as a strength, since ABG values should continue to be collected according to clinical viewpoints and not primarily with AI in mind.

Based on our results, LSTM-based models could help ICU physicians in several ways. First, an algorithm predicting mortality at high accuracy, with low false-positive rates, could help physicians in the decision process for treatment limitation after an initial "ICU-trial" consisting of full-blown ICU treatment for 24 to 48 hours. Considering this possible implantation of machine learning models in the ICU, the ethical challenges accompanied by these models for prognostication become evident [15]. Although physicians will not - and must not – rely solely on machine learning models to predict outcomes in the critically ill, such a model may become part of a decision in a matter of life and death [15].

It is necessary and important - for all the enthusiasm for the great potential that the implementation of AI holds for the improvement of clinical decision processes and outcomes - to emphasize that a clinical decision - especially "about life and death" - can never be made on the basis of an AI-derived algorithm alone.

Therefore, we consider our model primarily thesis-generating and a proof-of-concept that LSTM-based model can accurately predict mortality in critically ill patients even when evaluated across hundreds of diverse ICUs and hospitals. In this study, for example, age or gender were not included as a variable in the AI model, to avoid bias against older patients or a gender. However, after the prospective evaluation of a machine learning algorithm, the careful integration of further information could even increase its predictive capabilities.

These considerations underscore the importance of another principle challenge with AI and risk prediction: depending on the specific AI

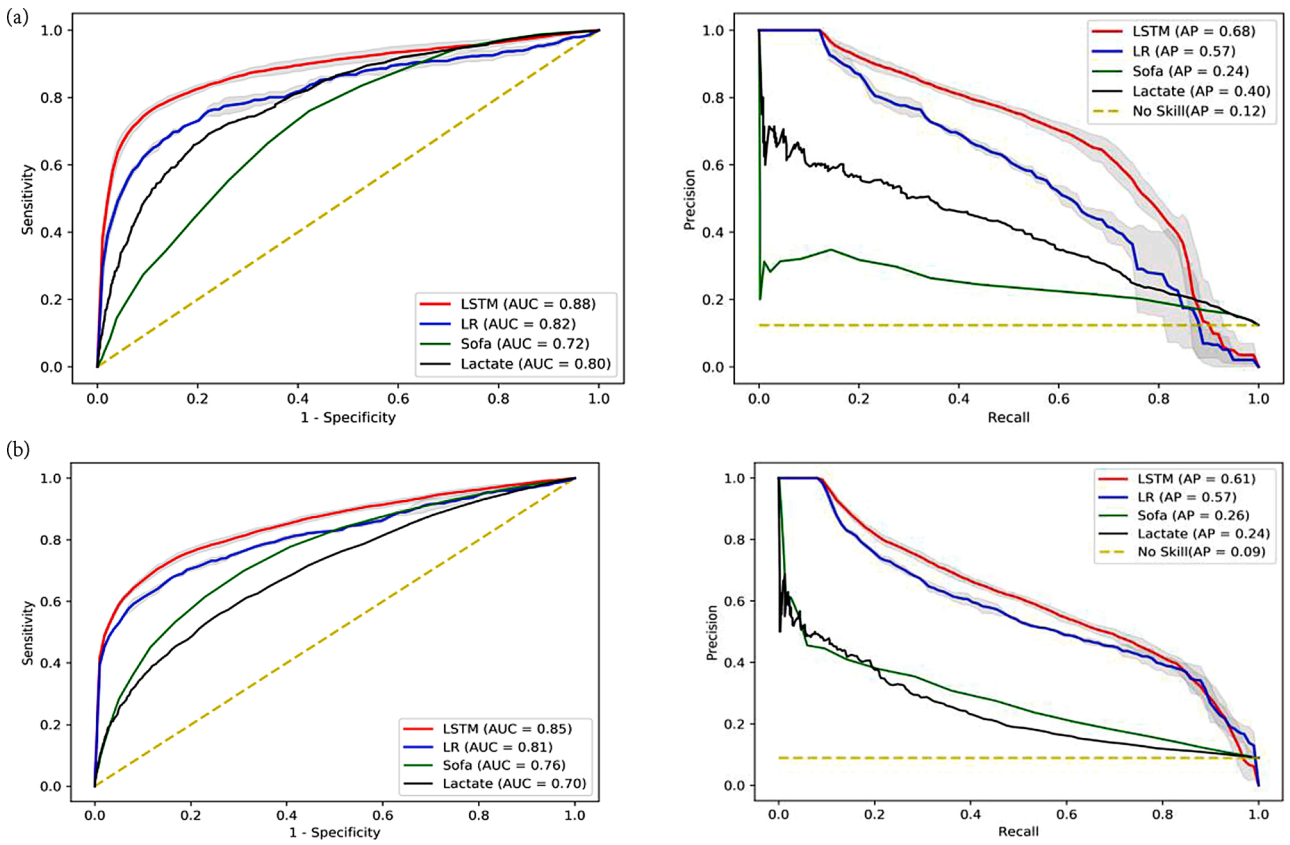

Fig. 3. a Multi-centre Receiver Operating Characteristic Curves (left) and Precision-Recall Curves (right) of machine learning model (LSTM) compared to Logistic Regression (LR), SOFA scores and Lactate (at admission) in Predicting Patient Mortality. Grey bands indicate standard deviation. No Skill refers to baseline performance.
b Single-centre Receiver Operating Characteristic Curves (left) and Precision-Recall Curves (right) of machine learning model (LSTM) compared to Logistic Regression (LR), SOFA scores and Lactate (at admission) in Predicting Patient Mortality. Grey bands indicate standard deviation. No Skill refers to baseline performance

Table 5
Partial AUC as well as Concordant Partial AUC for the multi-centre dataset

FPR		Lactate	Sofa	LR	LSTM	LSTM Septic Shock
[0.00 - 0.33]	pAUC	0.19	0.13	0.22	0.25	0.25
	pAUCc	0.44	0.33	0.48	0.54	0.55
[0.33 - 0.66]	pAUC	0.29	0.27	0.28	0.30	0.31
	pAUCc	0.20	0.21	0.17	0.17	0.17
[0.66 - 1.00]	pAUC	0.32	0.32	0.32	0.33	0.33
	pAUCc	0.16	0.18	0.17	0.17	0.17
Whole [0.00 - 1.00]	sum (AUC)	0.80	0.72	0.82	0.88	0.89

Table 6
Partial AUC as well as Concordant Partial AUC for the single-centre dataset

FPR		Lactate	Sofa	LR	LSTM	LSTM Septic Shock
[0.00 - 0.33]	pAUC	0.15	0.21	0.21	0.23	0.28
	pAUCc	0.36	0.43	0.48	0.51	0.59
[0.33 - 0.66]	pAUC	0.25	0.24	0.28	0.29	0.32
	pAUCc	0.18	0.16	0.16	0.17	0.17
[0.66 - 1.00]	pAUC	0.30	0.31	0.32	0.33	0.33
	pAUCc	0.16	0.17	0.17	0.17	0.17
Whole [0.00 - 1.00]	sum (AUC)	0.70	0.76	0.81	0.85	0.93

methodology applied, humans know little to nothing about the algorithm's composition and how it weighs the individual variables, even though this is an active area of research [16–19]. This is a fundamental difference compared to models derived from multivariable logistic regression – even in complex models, the regression coefficients of each variable are readily available and understandable for humans [16]. In that regard, the medical research community must carefully consider and evaluate the distinct methods available in machine learning: both supervised and unsupervised methods do have potential advantages but also disadvantages [17,20].

Sepsis is a multi-organ disease with many clinical faces [7]. While the mortality in this dataset was relatively low, the high predictiveness of our model persisted in a sub-group analysis on patients suffering from septic shock. Mortality prediction using machine learning models based on ABG values might, therefore, be applicable in both patients with sepsis and full-blown septic shock. Further, this approach might apply to other ICU admission reasons, but these speculations are beyond the scope of this paper. As in other disease entities such as cardiogenic shock, further factors such as coronary status, reperfusion therapy, or mechanical support might influence outcomes we focused on septic patients for this study [21].

Currently, physicians integrate clinical judgment, intensive care scores, biomarker concentrations such as lactate, as well as gut feeling to predict outcomes in complex diseases [22–25]. In this analysis, our model outperformed established tools to predict risk in ICU patients, both SOFA score as well as lactate concentration at baseline. Whereas the outperformance of lactate comes at a little surprise as the model evaluated in this study integrates lactate concentrations, the superior

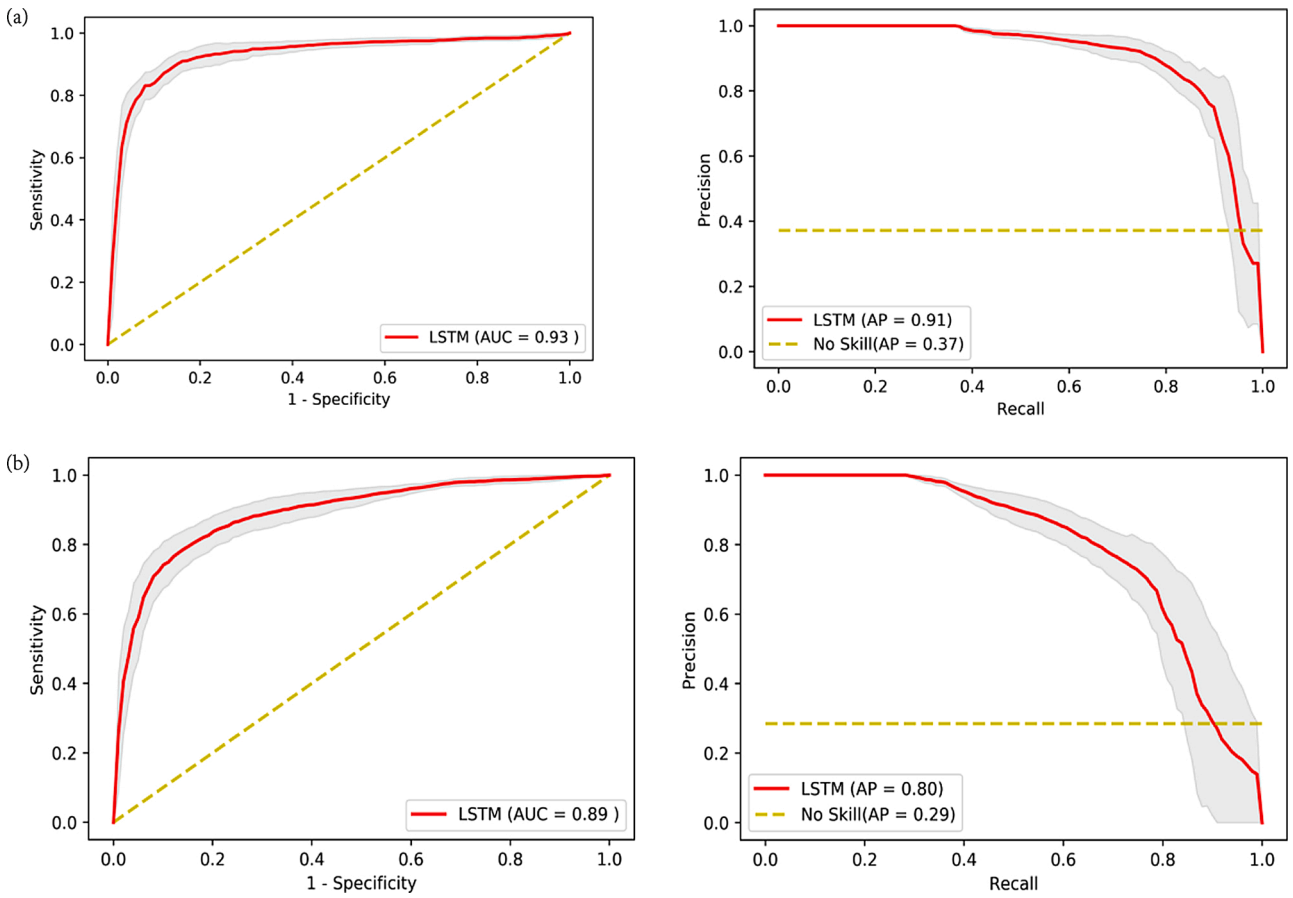

Fig. 4. a Multi-centre Receiver Operating Characteristic Curves (left) and Precision-Recall Curves (right) of machine learning model (LSTM) to predict mortality for patients in septic shock. Grey band indicates standard deviation. No Skill refers to baseline performance.
b Single-centre Receiver Operating Characteristic Curves (left) and Precision-Recall Curves (right) of machine learning model (LSTM) to predict mortality for patients in septic shock. Grey band indicates standard deviation. No Skill refers to baseline performance

predictability compared to SOFA, which consists of information about multiple organ systems as well as other clinical information, is noteworthy. Further, our model outperformed a multivariable logistic regression model based on the very same variables. This finding is in accordance with previous studies reporting superior outcomes of machine learning versus logistic regression [26–29]. McWilliams et al. showed that a machine learning algorithm outperformed logistic regression in assessing the readiness for the discharge of ICU patients [28]. In the study of Meiring et al., logistic regression outperformed an ICU scoring tool, but deep learning derived algorithm yielding even higher predictiveness for ICU mortality (33). In another study in patients suffering from previous cardiac arrest, again, machine learning enhanced the predictiveness in comparison to established scoring tools [26]. Pirracchio et al. integrated the variables necessary to compute the SAPS II score in a machine learning algorithm, which had a very high accuracy with an AUC of 0.94 [27].

Still, compared to these previous studies, the algorithm in this study is based only on widely available ABG values, as compared to other variables unavailable in non-tertiary settings or cost-sensitive environments. This fact also deliberates our study from past studies, which showed an earlier detection of sepsis in machine learning based models. Those studies mainly considered vital parameters such as heart ratio, peripheral oxygen saturation an body temperature, whereas our algorithm only relied on arterial blood gas analysis values [29,30]. Future studies may show the potential benefit of combining both algorithms and enriching them with more given information on ICUs. Depsite the aBGA based algorithm evaluated in this study could, therefore, easily be trained in individual ICUs to mirror the particular local situation and help ICU physicians in distinct settings.

5. Limitations

First, this is a retrospective study lacking a randomisation process, prospective screening, and inclusion of patients and a control group, therefore this study can only be thesis-generating. Second, no specific protocol for the collection of ABG values (e.g., specific timespan and/or clinical situations when to document ABG) was applied, which could further dispose of the study to selection bias. On the other hand, this mirrors a real-world scenario, where – likely – ABG values are taken based on clinical needs and after careful consideration of the ICU staff. Third, we focused on patients admitted to ICU for sepsis: The diagnosis of sepsis has, therefore, to be established before this algorithm is applied. Furthermore, the retrospective definition and identification of septic patients is not an easy task. We have decided to use the method of Angus et al., which is however not uncontroversial. On the other hand, Johnson et al. compared several methods for identification of septic patients without detecting relevant differences. We therefore consider the method of Angus et al. a pragmatic and established approach [12, 31]. Further, the results in patients not admitted to ICU might differ. Fourth, patients with missing values were excluded from this analysis, which could, again, lead to selection bias, however in clinical practice ABG values are readily available in patients with sepsis.

6. Conclusion

In this study, we evaluated an LSTM-based model on sepsis patients

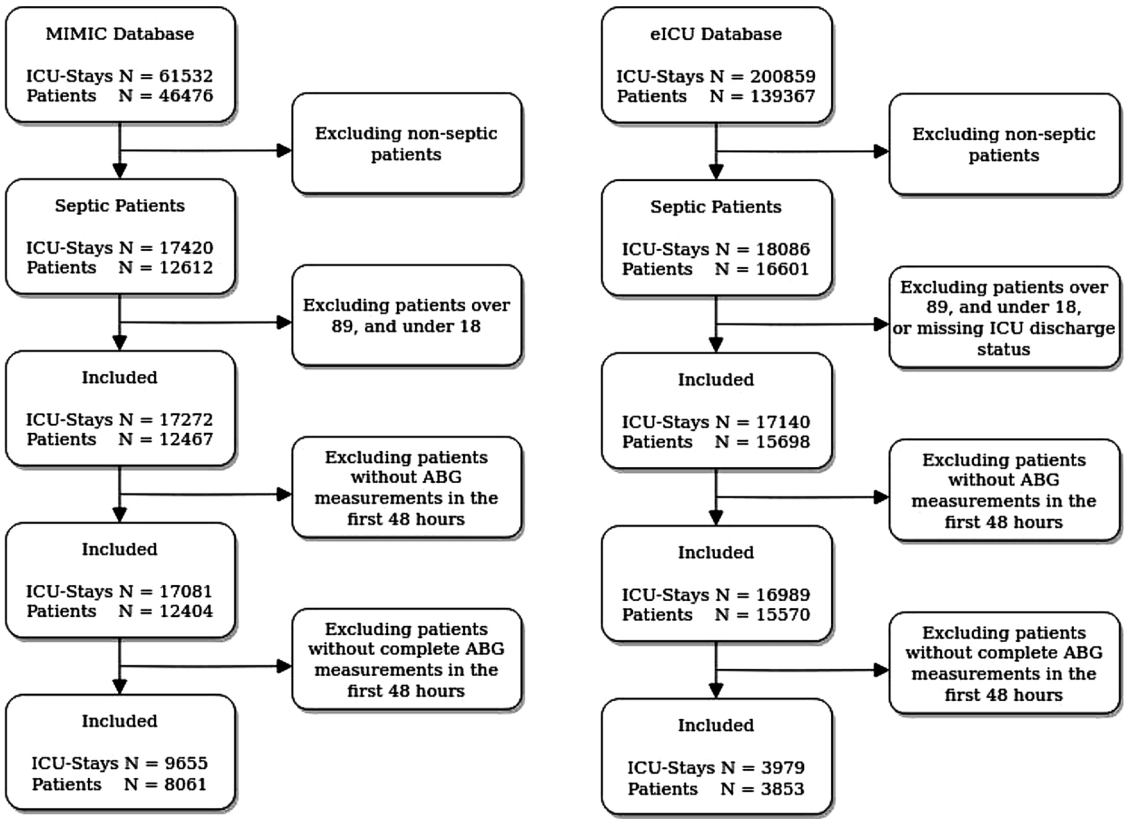

Fig. 5. Cohort selection diagram for both MIMIC dataset and eICU dataset.

admitted to ICU, where the model enhanced the capabilities of ICU mortality prediction after an ICU trial. The model was evaluated and validated in a multi-centre approach. Ethical considerations arise, but such models could help physicians with the "re-triage" and ultimately, the decision to restrict treatment in patients with a dismal prognosis.

Ethical Approval and Consent to participate

This study was an analysis of publicly available, anonymised databases with pre-existing institutional review board (IRB) approval; thus, no further approval was required.

Consent for publication

This study was an analysis of publicly available, anonymised databases with pre-existing institutional review board (IRB) approval; thus, no further approval was required.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Funding

No (industry) sponsorship has been received for this investigator-initiated study.

Availability of Data and Materials

All data relevant for this study will be provided by the authors upon specific request without restriction.

Acknowledgments

We acknowledge the support of all investigators of the eICU and MIMIC-III group.

References

[1] X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, et al., Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study, Lancet Respir Med. (2020).
[2] B. Guidet, H. Flaatten, A. Boumendil, A. Morandi, F.H. Andersen, A. Artigas, et al., Withholding or withdrawing of life-sustaining therapy in older adults (&/= 80 years) admitted to the intensive care unit, Intensive Care Med. 44 (7) (2018) 1027–1038.
[3] G. Gutierrez, Artificial Intelligence in the Intensive Care Unit, Crit Care. 24 (1) (2020) 101.
[4] C.A. Lovejoy, V. Buch, M. Maruthappu, Artificial intelligence in the intensive care unit, Crit Care. 23 (1) (2019) 7.
[5] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, et al., Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach, JMIR Med Inform. 4 (3) (2016) e28.
[6] G.L. Ambrosini, W.H. Oddy, M. Robinson, T.A. O'Sullivan, B.P. Hands, N.H. de Klerk, et al., Adolescent dietary patterns are associated with lifestyle and family psycho-social factors, Public Health Nutr. 12 (10) (2009) 1807–1815.
[7] I. Rubio, M.F. Osuchowski, M. Shankar-Hari, T. Skirecki, M.S. Winkler, G. Lachmann, et al., Current gaps in sepsis immunology: new opportunities for translational research, Lancet Infect Dis. 19 (12) (2019), 422-36.
[8] J. Siewiera, D. Tomaszewski, J. Piechocki, A. Kubler, Withholding and withdrawing life-sustaining treatment: Experiences in limiting futile therapy from three Polish intensive care departments, Adv Clin Exp Med. 28 (4) (2019) 541–546.
[9] G. Leblanc, A. Boumendil, B. Guidet, Ten things to know about critically ill elderly patients, Intensive Care Med. 43 (2) (2017) 217–219.
[10] B.E. Keuning, T. Kaufmann, R. Wiersema, A. Granholm, V. Pettila, M.H. Moller, et al., Mortality prediction models in the adult critically ill: A scoping review, Acta Anaesthesiol Scand. 64 (4) (2020) 424–442.
[11] T.J. Pollard, A.E.W. Johnson, J.D. Raffa, L.A. Celi, R.G. Mark, O. Badawi, The eICU Collaborative Research Database, a freely available multi-center database for critical care research, Sci Data. 5 (2018) 180178.
[12] D.C. Angus, W.T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, M.R. Pinsky, Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care, Crit Care Med. 29 (7) (2001) 1303–1310.

[13] F.L. Ferreira, D.P. Bota, A. Bross, C. Melot, J.L. Vincent, Serial evaluation of the SOFA score to predict outcome in critically ill patients, JAMA. 286 (14) (2001) 1754–1758.
[14] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, et al., A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Med Inform Decis Mak. 20 (1) (2020) 4.
[15] M. Beil, I. Proft, D. van Heerden, S. Sviri, P.V. van Heerden, Ethical considerations about artificial intelligence for prognostication in intensive care, Intensive Care Med Exp. 7 (1) (2019) 70.
[16] A.J. London, Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability, Hastings Cent Rep. 49 (1) (2019) 15–21.
[17] P.I. Dorado-Diaz, J. Sampedro-Gomez, V. Vicente-Palacios, P.L. Sanchez, Applications of Artificial Intelligence in Cardiology, The Future is Already Here. Rev Esp Cardiol (Engl Ed). 72 (12) (2019) 1065–1075.
[18] Y. Nohara, K. Iihara, N. Nakashima, Interpretable Machine Learning Techniques for Causal Inference Using Balancing Scores as Meta-features, Conf Proc IEEE Eng Med Biol Soc. 2018 (2018) 4042–4045.
[19] A. Holzinger, P. Kieseberg, E. Weippl, A.M. Tjoa (Eds.), Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI2018, Springer International Publishing, Cham, 2020.
[20] D. Dey, P.J. Slomka, P. Leeson, D. Comaniciu, S. Shrestha, P.P. Sengupta, et al., Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review, J Am Coll Cardiol. 73 (11) (2019) 1317–1335.
[21] B. Wernly, C. Seelmaier, D. Leistner, B.E. Stahli, I. Pretsch, M. Lichtenauer, et al., Mechanical circulatory support with Impella versus intra-aortic balloon pump or medical treatment in cardiogenic shock-a critical appraisal of current data, Clin Res Cardiol. 108 (11) (2019) 1249–1257.
[22] C. Jung, S. Bueter, B. Wernly, M. Masyuk, D. Saeed, A. Albert, et al., Lactate Clearance Predicts Good Neurological Outcomes in Cardiac Arrest Patients Treated with Extracorporeal Cardiopulmonary Resuscitation, J Clin Med. 8 (3) (2019).
[23] M. Masyuk, B. Wernly, C. Jung, Prognostic relevance of serum lactate kinetics: a powerful predictor but not Chuck Norris in Intensive Care Medicine, Intensive Care Med. 45 (8) (2019) 1174–1175.
[24] J.L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonca, H. Bruining, et al., The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine, Intensive Care Med. 22 (7) (1996) 707–710.
[25] J. Castela Forte, A. Perner, I.C.C. van der Horst, The use of clustering algorithms in critical care research to unravel patient heterogeneity, Intensive Care Med. 45 (7) (2019) 1025–1028.
[26] S. Nanayakkara, S. Fogarty, M. Tremeer, K. Ross, B. Richards, C. Bergmeir, et al., Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study, PLoS Med. 15 (11) (2018) e1002709.
[27] R. Pirracchio, M.L. Petersen, M. Carone, M.R. Rigon, S. Chevret, M.J. van der Laan, Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study, Lancet Respir Med. 3 (1) (2015) 42–52.
[28] C.J. McWilliams, D.J. Lawson, R. Santos-Rodriguez, I.D. Gilchrist, A. Champneys, T.H. Gould, et al., Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. BMJ Open. 9 (3) (2019) e025925.
[29] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R.L. Davis, R. Kamaleswaran, A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier, Int J Med Inform. 122 (2019) 55–62.
[30] R. Kamaleswaran, O. Akbilgic, M.A. Hallman, A.N. West, R.L. Davis, S.H. Shah, Applying Artificial Intelligence to Identify Physiomarkers Predicting Severe Sepsis in the PICU, Pediatr Crit Care Med. 19 (10) (2018) e495–e503.
[31] A.E.W. Johnson, J. Aboab, J.D. Raffa, T.J. Pollard, R.O. Deliberato, L.A. Celi, et al., A Comparative Analysis of Sepsis Identification Methods in an Electronic Database, Crit Care Med. 46 (4) (2018) 494–499.