Journal of Clinical Monitoring and Computing (2022) 36:1087–1097
https://doi.org/10.1007/s10877-021-00739-4

ORIGINAL RESEARCH

Prediction of blood lactate values in critically ill patients: a retrospective multi-center cohort study

Behrooz Mamandipoor[1] · Wesley Yeung[2,3] · Louis Agha-Mir-Salim[2,4] · David J. Stone[5] · Venet Osmani[1] · Leo Anthony Celi[2,6,7]

Received: 9 March 2021 / Accepted: 1 July 2021 / Published online: 5 July 2021

Abstract

Elevations in initially obtained serum lactate levels are strong predictors of mortality in critically ill patients. Identifying patients whose serum lactate levels are more likely to increase can alert physicians to intensify care and guide them in the frequency of tending the blood test. We investigate whether machine learning models can predict subsequent serum lactate changes. We investigated serum lactate change prediction using the MIMIC-III and eICU-CRD datasets in internal as well as external validation of the eICU cohort on the MIMIC-III cohort. Three subgroups were defined based on the initial lactate levels: (i) normal group (< 2 mmol/L), (ii) mild group (2–4 mmol/L), and (iii) severe group (> 4 mmol/L). Outcomes were defined based on increase or decrease of serum lactate levels between the groups. We also performed sensitivity analysis by defining the outcome as lactate change of > 10% and furthermore investigated the influence of the time interval between subsequent lactate measurements on predictive performance. The LSTM models were able to predict deterioration of serum lactate values of MIMIC-III patients with an AUC of 0.77 (95% CI 0.762–0.771) for the normal group, 0.77 (95% CI 0.768–0.772) for the mild group, and 0.85 (95% CI 0.840–0.851) for the severe group, with only a slightly lower performance in the external validation. The LSTM demonstrated good discrimination of patients who had deterioration in serum lactate levels. Clinical studies are needed to evaluate whether utilization of a clinical decision support tool based on these results could positively impact decision-making and patient outcomes.

Keywords Resuscitation · Lactate · Critical illness · Deep learning · Time series

✉ Venet Osmani
vosmani@fbk.eu

1 Fondazione Bruno Kessler Research Institute, Trento, Italy

2 Laboratory for Computational Physiology, Harvard-MIT Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

3 University Medicine Cluster, National University Hospital, Kent Ridge, Singapore

4 Faculty of Medicine, University of Southampton, 12 University Rd, Southampton SO17 1BJ, UK

5 Departments of Anesthesiology and Neurosurgery, and the Center for Advanced Medical Analytics, University of Virginia School of Medicine, Charlottesville, VA 22908, USA

6 Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

7 Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA

Abbreviations

ABG	Arterial blood gas
AI	Artificial intelligence
AP	Average precision
AUC	Area under the receiver operating characteristic curve
AUPRC	Area under the precision recall curve
FDA	Food and drug administration
ICU	Intensive care unit
LR	Logistic regression
LSTM	Long short-term memory
MCC	Mathews correlation coefficient
MIMIC	Medical information mart for intensive care
NPV	Negative predictive value
PPV	Positive predictive value
PRC	Precision recall curve
RF	Random forest
ROC	Receiver operating characteristic
SHAP	Shapley additive explanations

SOFA Sequential organ failure assessment
STROBE Strengthening the reporting of observational studies in epidemiology

1 Introduction

Hyperlactatemia is a medical condition caused by accumulation of lactate and hydrogen ions in the bloodstream and tissues, usually as a result of tissue hypoxia and systemic hypoperfusion. It is commonly observed and treated in critical care conditions such as severe heart failure, sepsis, or other forms of shock. Both the magnitude and rate of change of serum lactate elevation are strong predictors of mortality [1].

When a patient is admitted to an ICU for certain conditions such as shock or trauma, a serum lactate level may be obtained in addition to a number of other laboratory tests. Certain tests including metabolic and hematologic (i.e. complete blood count) panels are routinely obtained on a daily basis during the acute phase of critical illness. However, it is rarely appropriate or indicated to test serum lactate in this kind of routine, periodic basis. Rather, serum lactate values are obtained in a targeted fashion based on the clinical context, or more explicitly, on the perceived stability of the patient. Reliance on the provider to order the test introduces variation that can impact the outcome of the patient. In this work, we address the issue of whether the clinical determination of timing for subsequent serum lactate samples can be improved by application of artificial intelligence techniques to available data.

Blood (serum, or plasma when the lactate is measured in an anticoagulated sample with an arterial blood gas) lactate reduction during the initial hours of intensive care unit (ICU) admission has been shown to be associated with improved survival [2–5], while persistently high and increasing levels are associated with poor outcomes [6]. Resuscitation guided by serum lactate levels has also been shown to be associated with reduced hospital mortality [7]. At present, continuous lactate monitoring is not yet available [8]. While frequent, periodic serum lactate measurements might seem the next best choice, these approaches involve downsides, including risk of anemia from repeated blood draws [9, 10], need for frequent venipunctures, or use of a central venous catheter to draw blood that comes with an infection risk [11, 12], and cost. Many unnecessary samples are also likely to be drawn when periodic studies are ordered. Most importantly, in the absence of the availability of continuous serum lactate measurements, the optimal approach to periodic or repeated determination of serum lactate level simply remains uncertain, and in current practice is likely to rest on the variable and individualized experiences of practitioners as well as the inevitable exigencies and vicissitudes of clinical workflow in a demanding environment. The fundamental clinical question is whether the serum lactate levels are likely to be increasing, or in the case of already elevated levels, showing no improvement (which is considered negatively from a clinical standpoint). A data driven trigger for determining the need and timing for repeat serum lactate testing would be a significant advance in standardizing and potentially improving care processes as well as clinical laboratory utilization in this setting.

Given the strong prognostic utility of serum lactate [1], continuously predicting the trajectory of serum lactate values would be clinically useful as a tool that would optimize the number of serum lactate tests by reducing unnecessary testing while providing a reminder for necessary, and presumably useful, subsequent testing. A prediction of increasing serum lactate levels could alert clinicians to potential deterioration and prompt confirmatory testing with a blood draw. On the other hand, a prediction of stable (in the case of previously normal levels) or improving serum lactate levels would prevent unnecessary blood draws. Machine learning algorithms may be useful in both prompting repeat blood draws likely to yield actionable information, and in reducing the number of unnecessary repeat testing [13].

We hypothesize that: (1) clinical variables during the first 48 h of ICU admission can predict the trajectory of serum lactate values during that time, and that (2) patients classified into normal, mild and severe groups, based on their initial serum lactate measurements, manifest different factors affecting this trajectory. In this work, we describe an approach to detecting worsening hyperlactatemia in ICU patients on the basis of input of expert clinical knowledge, state-of-the-art analytical techniques, and large, high-resolution, multi-center datasets to construct three models to identify patients at risk of worsening hyperlactatemia within the first 48 h of ICU admission.

2 Methods

2.1 Data sources

The Medical Information Mart for Intensive Care (MIMIC-III, v1.4) is a longitudinal, single-center database maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) which contains data associated with 53,423 distinct ICU admissions for adult patients (aged 16 years and older) admitted to critical care units between 2001 and 2012 [14] at the Beth Israel Deaconess Medical Center. It is a teaching hospital of Harvard Medical School with 673 licensed beds, including 493 medical/surgical beds, 77 critical care beds, and 62 OB/GYN beds.

The eICU Collaborative Research Database (eICU-CRD) contains data associated with 200,859 admissions collected from 335 ICUs across 208 hospitals in the US admitted between 2014 and 2015 [15].

2.2 Study design

We retrospectively evaluated a subgroup of adult patients (age ≥ 18 years) from the MIMIC-III and eICU-CRD datasets that had at least 2 serum lactate measurements recorded within the first 48 h of ICU admission as well as an ICU length of stay greater than or equal to 24 h. The selected patients were further divided into three subgroups based on their initial serum lactate levels: (i) normal group (< 2 mmol/L), (ii) mild group (2 to 4 mmol/L) and (iii) severe group (> 4 mmol/L). The present study is reported in accordance with the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) statement.

2.3 Definition of outcomes

The outcome was the trajectory of the second serum lactate measurement which was categorized into a positive or negative outcome based on the initial subgroup to which the patient belonged. For the normal group, a negative outcome was defined as a serum lactate increase to mild or severe levels, while a positive outcome was defined as a value that remained within the normal level. A similar approach was taken for the other two subgroups, as shown in Table 1, where the increase in lactate levels between groups corresponds to a negative outcome.

2.4 Sensitivity analyses

We conducted a sensitivity analysis to investigate whether a 10% change in serum lactate levels (rather than between groups) influences predictive performance of the model. The 10% change was chosen because serum lactate non-clearance, defined as a serum lactate decrease of less than 10%, is associated with an increased risk of mortality [2–6]. Details and results of this analysis are presented in Online Appendix 1.

We also conducted an additional sensitivity analysis to investigate whether the difference in time between the two serum lactate measurements has any effect on the prediction performance of lactate deterioration. For this analysis, we restricted the cohort to only those patients that had the subsequent serum lactate measured within 8 h of the preceding lactate measurement. The 8-h interval was chosen based on Surviving Sepsis Campaign guidelines [16] that recommend serum lactate be measured every 6 h. We allowed for a 2-h delay to account for situations where the serum lactate might be measured but not immediately recorded in the patient's health record. Details and results of these analyses are presented in Online Appendix 2.

2.5 Variable selection

We selected 54 variables identified by ICU clinicians and the related literature as relevant to serum lactate deterioration and available in both MIMIC-III and eICU-CRD within 48 h of admission. These include selected laboratory values, vital signs, patient demographics, and nursing care data obtained during the admission assessment as shown in Online Appendix 3. Variables also included values obtained through an arterial blood gas (ABG) of the previous measurement, which had to be sampled at least two hours (discretization interval) before the timestamp of the predicted serum lactate level. Laboratory variables were discretized into two-hour intervals as experiments revealed better model performance compared to models developed on hourly time windows. Outliers were addressed by defining a clinically valid interval. The variables were normalized using zero mean and scaling to unit variance. Linear correlation (Pearson) between the top 10 highest correlated variables with serum lactate is shown in Online Appendix 3.

2.6 Missing values imputation strategy

We evaluated several imputation strategies using both data-driven approaches and in combination with clinical heuristics. In our previous work, we evaluated twelve different imputation strategies, including strategies based on mean, multiple imputation (chained equations), random

Table 1 Definition of outcomes for each patient subgroup

Initial lactate value	Outcome	
	Negative	Positive
Normal (< 2 mmol/L)	Serum lactate **increases** to mild or severe group levels	Serum lactate **remains** within the normal group
Mild (2–4 mmol/L)	Serum lactate **increases** to severe group or **remains** within the mild group levels	Serum lactate **decreases** to normal group levels
Severe (> 4 mmol/L)	Serum lactate **increases** or **remains** within the severe group levels	Serum lactate **decreases** to mild or normal group levels

forest, and autoencoders for prediction of serum lactate levels [17, 18]. This previous work has shown that an imputation strategy based on mean and indicator variables, where a Boolean variable is added to indicate whether a value is missing or not, provides the best performance. We therefore set out to evaluate this strategy first, and subsequently compare it with the most common strategy based on the mean value. The results showed that the indicator imputation strategy provided better performance (in terms of AUC) than using the mean value alone. However, using an indicator variable degrades model interpretability and variable ranking, due to the increase in the number of Boolean variables (an indicator is added for each variable with missing values).

As such, we opted for a fill-forward imputation strategy applied to each ICU stay by forward propagation of all the valid measurements. This approach provided an optimal trade-off between model performance and interpretability. Furthermore, we investigated whether a strategy based on clinical heuristics would further improve performance. Using this strategy, we defined an imputation method for each individual variable, as shown in Online Appendix 3. This strategy provided 2% (±0.85) AUC better performance on average than using the mean value. As a result, after splitting datasets into train and test set, we used clinical heuristics combined with the fill-forward method for the imputation of missing values. It should be noted that no lactate values were imputed.

2.7 Experimental evaluation methodology

2.7.1 Model development and experimentation

We evaluated the performance of three machine learning algorithms—logistic regression (LR), random forest (RF), and long short-term memory (LSTM). LR is an algorithm capable of predicting class probabilities using predictor variables, by adjusting the coefficients of the logit function, RF is an ensemble learning method constructing multiple decision trees and then producing class probabilities as outputs [19], while LSTM is a type of Deep Artificial Neural Network designed to learn temporal dependencies between variables and process longitudinal time-series data [20].

We spilt the data randomly into a derivation cohort (80%) and validation cohort (20%), where hyperparameters of all the models were optimized using random search on the validation set, detailed in Online Appendix 4. The final models were internally validated using stratified five-fold cross validation with 5 repetitions for both MIMIC-III and eICU-CRD datasets. For the external validation, we derived models on the eICU-CRD patient cohort and validated them on the MIMIC-III cohort.

2.7.2 Performance evaluation

We assessed each model by computing the area under the receiver operator characteristic curve (AUC-ROCs) and the area under the precision-recall curve (AUPRCs), also called Average Precision (AP). We also provide additional performance metrics, including calibration, Positive Predictive Value (PPV), Negative Predictive Value (NPV), F-1 score (showing the balance between PPV and sensitivity) and Matthews correlation coefficient MCC (used to measure the quality of classification between our algorithms) for each model. These performance metrics are detailed in Online Appendix 5, whereas calibration performance is detailed in Online Appendix 6.

2.7.3 Model interpretability

We conducted a model interpretability analysis to understand how the model ranked the importance of variables when predicting serum lactate trajectory. We used the SHAP (Shapley Additive exPlanations) method whose objective is to explain a prediction output of a machine learning model by computing the contribution of each variable to the prediction [21]. The SHAP method computes Shapley values, where those variables with the largest absolute values are the most important. Based on Shapley values, we ranked each variable based on importance, including a ranking of the top ten variables.

3 Results

3.1 MIMIC-III cohort

From the 61,532 overall admissions in MIMIC-III, 12,502 admissions matched our selection criteria (11,083 patients). The cohort selection diagram is shown in Online Appendix 3.

The MIMIC-III cohort had 29,337 serum lactate values recorded within the first 48 h of ICU admission with close to half in the normal group (46.9%) with the remaining values in the mild (37.7%) and severe groups (15.4%). The average patient age was 64.4 (±16.6) with 42% female patients. The most common admission diagnosis was sepsis, followed by pneumonia, with an overall median length of stay of 4.2 (IQR, [2.4–8.9]) days and mortality rate of 20.7% as shown in Table 2. Detailed patient characteristics and subgroup differences for both MIMIC-III and eICU-CRD are shown in Online Appendix 3.

For the MIMIC-III cohort, positive outcomes (see Table 1) with respect to serum lactate trajectory were observed in 87.1% (n = 11,977) of subsequent lactate measurements in the normal group, 40.5% (n = 4,485) in

Table 2 MIMIC-III cohort characteristics based on the initial lactate measurement

	Overall	Normal	Mild	Severe
Age* (SD)	64.4 (± 16.6)	64.6 (± 16.5)	64.4 (± 16.7)	62.8 (± 17.0)
Gender (% male)	7211 (58%)	4502 (57%)	3777 (58%)	1380 (58%)
Length of Stay** (IQR)	4.2 ([2.4–8.9])	4.7 ([2.7–9.2])	4.5 ([2.5–9.4])	5.1 ([2.6–11.5])
Mortality	20.7%	16.6%	21.5%	37.9%
Admission diagnosis	Sepsis (5.0%) Pneumonia (3.8%)	Sepsis (4.6%) Pneumonia (4.3%)	Sepsis (5.3%) Pneumonia (3.0%)	Sepsis (5.9%) Abdominal pain (2.2%)

*Represents mean

**Represents median

the mild group, and 39.5% (n = 1785) in the severe group. Figure 1 summarizes the performance of each model for each subgroup. We have also calculated precision-recall for different thresholds to devise curves that compare the performance of the models in the presence of imbalanced datasets, as is the case with the normal group where positive outcomes (87.1%) significantly exceed negative outcomes (12.9%).

RF and LR performed similarly for the normal group with AUCs for both of 0.74 (95% CI 0.738–0.748 vs 0.732–740), while RF outperformed LR for the other two subgroups. The LSTM model performed best across all the three subgroups, achieving an AUC of 0.77 (95% CI 0.762–0.771) for the normal group, 0.77 (95% CI 0.768–0.772) for the mild group, and 0.85 (95% CI 0.840–0.851) for the severe group. The models were well calibrated for the mild and severe groups as shown in Online Appendix 6, while the normal group demonstrated less accurate calibration.

We also investigated the performance of the models when removing the patients with elective admission type. However, the results showed no statistically significant difference between the overall cohort and the cohort without elective admissions in terms of AUC performance.

3.2 eICU-CRD cohort

From the 200,859 admissions in the eICU-CRD, 17,452 admissions (16,283 patients) matched our selection criteria, as detailed in the cohort selection diagram in Online Appendix 3. The eICU-CRD study cohort had 39,389 serum lactate values recorded within the first 48 h of ICU admission with 39.7% in the normal group, 35.4% in the mild group,

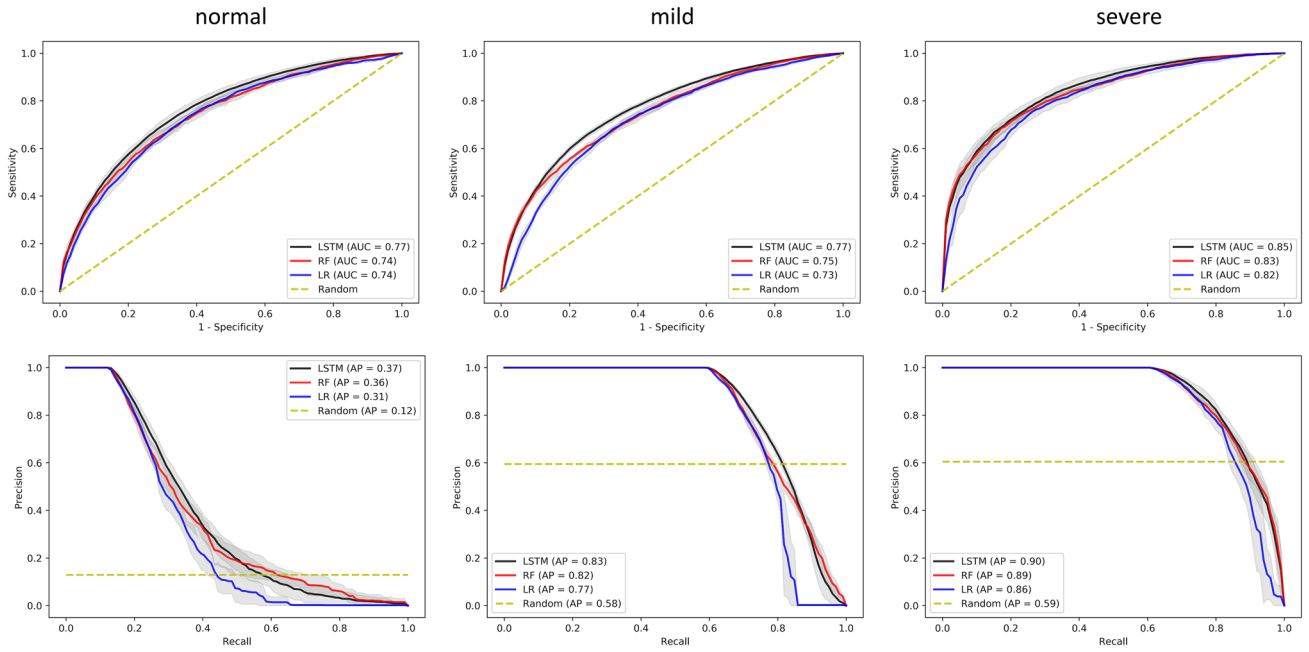

Fig. 1 Performance of each model in the MIMIC-III cohort across the three patient subgroups. Top row represents AUC-ROC, while the bottom AU-PRC. Confidence intervals are shown in grey

and 24.9% in the severe group. The average patient age was 62.2 (± 16.1) with 45% female patients. The most common admission diagnosis was sepsis, followed by cardiac arrest, with a median length of stay of 3.5 (IQR, [2.0–6.6]) days and mortality rate of 15.4% as shown in Table 3.

For the eICU-CRD cohort, positive outcomes with respect to serum lactate trajectory were observed in 87.2% (n = 13,640) in the normal group, 40.4% (n = 5,638) in the mild group, and 36.0% (n = 3,528) in the severe group. Figure 2 summarizes the performance of each model for each subgroup. RF performed slightly better than LR for the mild group with an AUC of 0.73 (95% CI 0.723–0.731) versus an AUC of 0.69 (95% CI 0.683–0.693), while both models had similar performances for the other two subgroups. The LSTM model performed best across all the three subgroups, achieving an AUC of 0.72 (95% CI 0.707–0.724) for the normal group, 0.74 (95% CI 0.735–0.745) for the mild group, and 0.84 (95% CI 0.837–0.848) for the severe group.

3.3 External validation of the eICU-CRD model on the MIMIC-III cohort

In addition to evaluating serum lactate deterioration prediction within MIMIC-III and eICU-CRD individually, we conducted an external validation where a model derived from the eICU-CRD was validated on the MIMIC-III patient cohort. This was done to investigate the generalizability of our method on independent patient data and its potential utility as a clinical decision support tool. We followed the same cohort selection criteria and derived the model using

Table 3 eICU-CRD cohort characteristics

	Overall	Normal	Mild	Severe
Age* (SD)	62.2 (± 16.1)	61.9 (± 16.0)	62.3 (± 16.3)	61.8 (± 16.0)
Gender (% male)	9520 (55%)	5142 (54%)	4772 (56%)	2649 (55%)
Length of Stay** (IQR)	3.5 ([2.0–6.6])	3.7 ([2.1–6.7])	3.7 ([2.1–6.9])	3.8 ([2.1–7.2])
Mortality	15.4%	10.0%	15.3%	29.4%
Admission diagnosis	Sepsis, pulmonary (12.7%) Cardiac arrest (8.1%)	Sepsis, pulmonary (12.8%) Cardiac arrest (7.5%)	Sepsis, pulmonary (13.3%) Cardiac arrest (8.7%)	Cardiac arrest (15.0%) Sepsis, pulmonary (11.3%)

*Represents mean

**Represents median

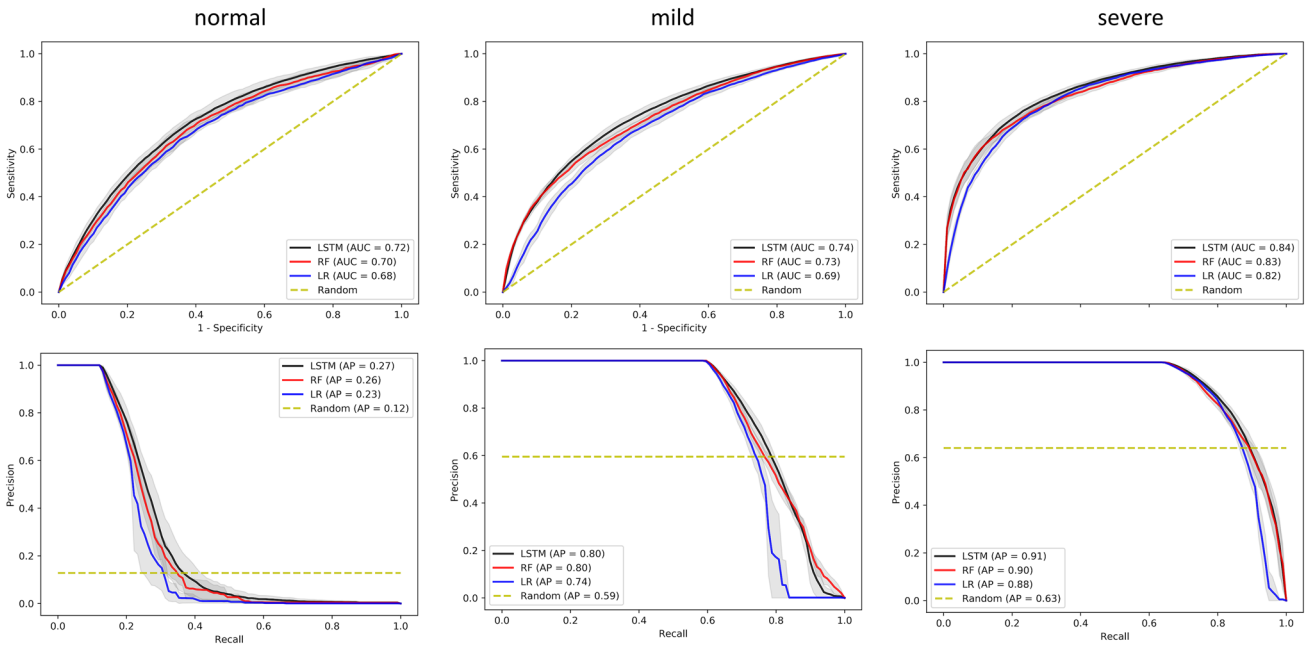

Fig. 2 Performance of each model in the eICU-CRD cohort across the three patient subgroups. Top row represents AUC-ROC, while the bottom AU-PRC. Confidence intervals are shown in grey

the eICU-CRD patient cohort, while the MIMIC-III cohort was used as a test set. The results are detailed in Fig. 3.

It is typically much more challenging to achieve similar results with external validation in comparison to internal validation. However, the performance of our method remained at similar levels to the internal studies. External validation results demonstrated deepened differences in performance between baseline algorithms (LR) and machine learning approaches (LSTM and RF): the performance of LSTM was much closer to that of RF, outperforming it only in the normal group. While, in contrast to RF, the LSTM is equipped to capture temporal dependencies between serum lactate measurements, the majority of temporal sequences are quite short, especially in the severe group where serum lactate measurements are more frequent (see distribution of timing between subsequent lactate measurements in Online Appendix 3). LSTM and RF achieved AUCs of 0.74 and 0.83 for the mild and severe groups respectively, with a lower performance for the normal group with AUCs of 0.70 and 0.69.

3.4 Sensitivity analyses

We also conducted sensitivity analyses using an alternative definition of serum lactate prediction outcome where we defined a change of at least 10% of serum lactate levels as an increase or decrease. The results of this analysis are detailed in Online Appendix 1, where the AUC of the LSTM model decreased to 0.76 (95% CI 0.755–0.766) from 0.77 (95% CI 0.762–0.771) for the normal group; to 0.67 (95% CI 0.661–0.672) from 0.77 (95% CI 0.768–0.772) for the mild group; and to 0.75 (95% CI 0.746–0.758) from 0.85 (95% CI 0.840–0.851) for the severe group. The changes in model performance are likely due to the fact that percent changes in serum lactate may be more sensitive than changes from one category to the next. This is not unlikely as these percentage changes in most instances require a smaller value change to register than do categorical changes.

The second sensitivity analysis focused on investigating whether the time difference between subsequent serum lactate measurements has any effect on serum lactate deterioration prediction performance (detailed in Online Appendix 2). The results of this analysis showed a decrease in AUC performance of the LSTM model for the mild and severe group to an AUC of 0.75 (95% CI 0.744–0.751) from 0.77 (95% CI 0.768–0.772); and to 0.83 (95% CI 0.831–0.840) from 0.85 (95% CI 0.840–0.851), respectively. For the normal group, the performance increased slightly to an AUC of 0.78 (95% CI 0.772–0.783) from 0.77 (95% CI 0.762–0.771). These changes are not statistically significant.

3.5 Variable importance

As we derived three different models for each of the subgroups, we also calculated variable rankings separately for each model. Therefore, the top three ranked variables for the model of the normal group were the prior serum lactate, serum glucose, and anion gap; for the mild group

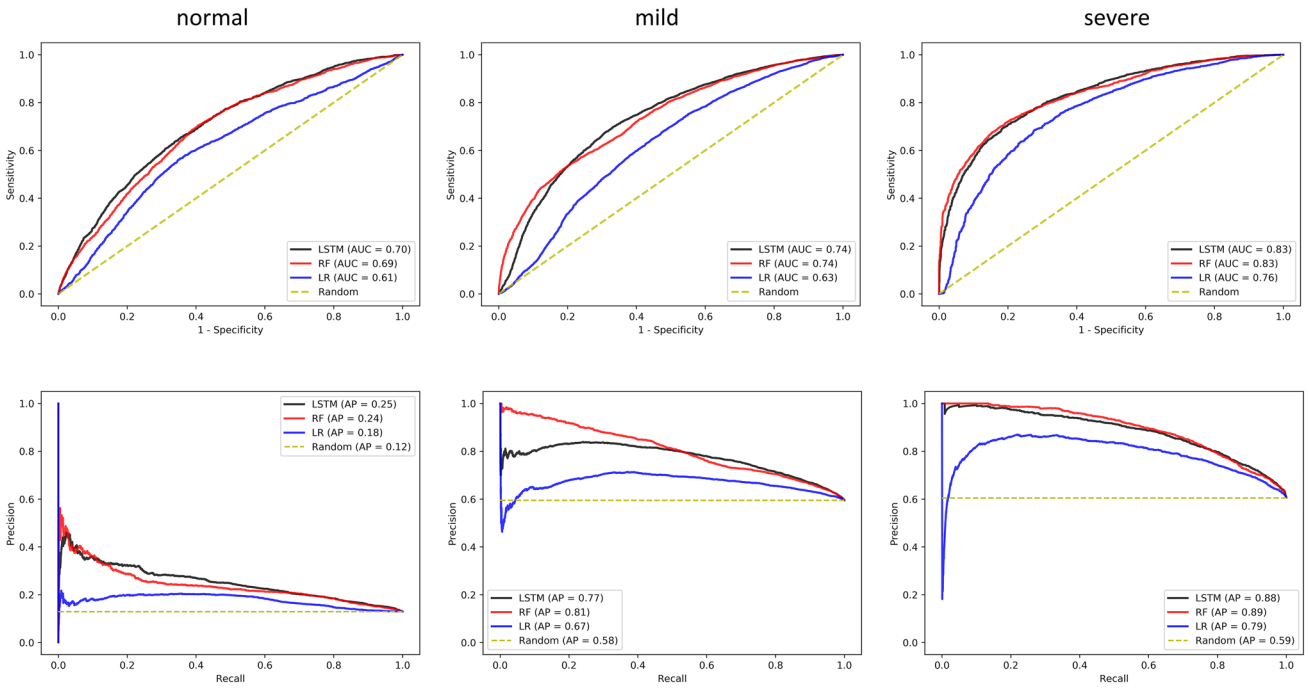

Fig. 3 Performance of each model derived in the eICU-CRD cohort and externally validated on the MIMIC-III cohort across the three patient subgroups. Top row represents AUC-ROC, while the bottom AU-PRC

model- prior serum lactate, respiratory rate, and serum glucose; while, for the model of the severe group, the most important variables were prior serum lactate, prior arterial base excess, and Glasgow Coma Score value. A graphical representation of the top 10 variables for each model and their ranking is provided in Fig. 4.

4 Discussion

The LSTM models were the most accurate in predicting deterioration of serum lactate values in all three serum lactate level subgroups in the MIMIC-III cohort, with an AUC of 0.77 (95% CI 0.762–0.771) for the normal group, 0.77 (95% CI 0.768–0.772) for the mild group, and 0.85 (95% CI 0.840–0.851) for the severe group.

We observed different patterns of importance of the variables among the patient subgroups. For example, in the subgroup with normal baseline serum lactate levels, the prior serum lactate measurement was an important predictor of deterioration of serum lactate values, followed by serum glucose, anion gap, temperature and heart rate. The mild and severe groups additionally showed respiratory rate, GCS, and base excess as important variables in predicting serum lactate levels. Arterial pH, base excess, serum bicarbonate, and serum anion gap values all reflect the acid–base balance [22], while decreased partial pressure of carbon dioxide and increased respiratory rate are the results of physiologic responses to metabolic acidosis [23]. Urine output is a function of volume status, and renal function, and is affected by vascular perfusion to that organ, which in turn is intrinsically linked with tissue acidosis and lactate metabolism [24, 25]. Heart rate can be affected by many diverse factors and has been found to be independently associated with in-hospital mortality [26]. Elevated serum bilirubin could be a marker of hepatic metabolic dysfunction, commonly referred to as "shock liver", which has also been found to be an important predictor of survival [27]. Moreover, lactate is also metabolized by the liver so that hepatic dysfunction can independently contribute to worsening hyperlactatemia [28, 29]. Clearly, it makes sense that these values would be and are strong determinants of the ongoing state of lactate levels. In addition to forming the basis of our predictive models, knowing the relative weights of these values in contributing to observed lactate levels is also useful to know for clinicians making these decisions. While severe changes in blood pressure or pH are rather obvious indicators that another lactate value should be obtained, there is also a more subtle constellation of changes in laboratory and vital sign values that should drive clinicians to consider rechecking lactate levels.

Clinical decision support (CDS) modalities must be accurate, useful, and usable, and fit as seamlessly as possible into clinicians' workflows. Lindsell et al. [30] recently stated that "Designing a useful AI tool in health care should begin with asking what system change the AI tool is expected to precipitate." In this case, the change would consist of implementing a tool that would optimize the number of serum lactate tests by reducing unnecessary testing while providing a reminder for necessary, and presumably useful, subsequent testing.

The net change in test frequency would not be an adequate metric for evaluating the impact of this process change because any decrease in unnecessary testing could be offset by an increase in indicated testing. One metric that could be employed would be the relative (compared to baseline) percentage of repeat serum lactate values that demonstrated values that were clinically actionable (e.g., crossing the threshold from normal to mild, or from mild to high). But the critical metric would be whether the more focused identification of serum lactate anomalies contributes to improved outcomes in those patients who, at some time in their clinical course, have elevated serum lactate levels of some degree. The ultimate analysis of the value of such a CDS tool really requires a systems level approach that incorporates the classic ICU metrics of mortality and LOS, but also considers costs, fluid balance and renal function, impacts on workflows, and even the detection of adverse event outlier cases where the CDS leads the clinicians astray.

We would envision that the preliminary version of a CDS model would be updated every hour in order to make

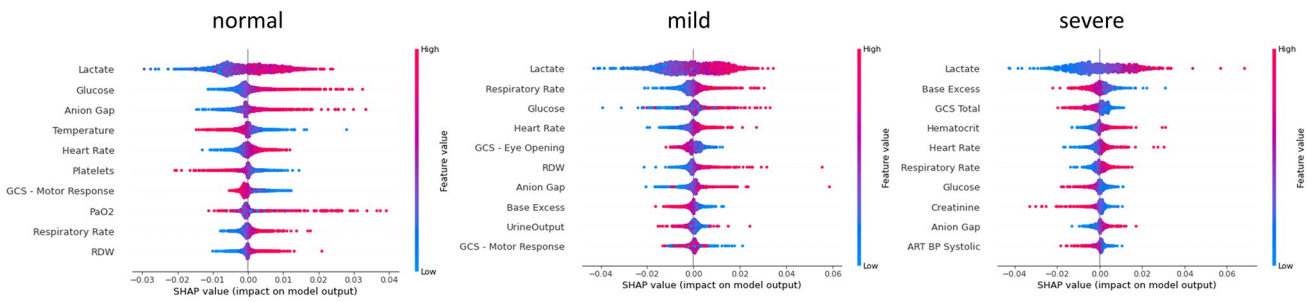

Fig. 4 Variable importance ranking for each LSTM-based model of the patient subgroups derived from the MIMIC-III cohort

predictions employing events and values, newly captured over that interval. The algorithm would incorporate all pertinent variables in the interval and determine which contribute to the possible need for an additional serum lactate sample, e.g., new evidence of sepsis (e.g. increasing SOFA score), increasing anion gap, or increasing respiratory rate. The specific element of 'figuring out' the new determination only applies to the LSTM based model in which newly available data can actually update the state of the neural network. Unlike the static algorithms of LR and RF, the LSTM algorithm would change dynamically during use so that any CDS tool based on this approach would require special Food and Drug Administration (FDA) approval. The FDA is currently studying what will be best practices in terms of approving such innovative yet evolving instruments in clinical care. Further advances in the quality of the CDS would involve such dynamically evolving tools that learn continuously and provide automatic feedback to improve the model; learning how to best incorporate continuously tracked values such as HR and RR; identifying patient, disease, and unit level characteristics that could make the CDS a more precision tool; and the addition of new input variables as clinical medicine evolves.

Since all available data up to the time of prediction would be employed, even the fixed algorithms (LR, RF) would become potentially more accurate over time. For example, if there are two (or more) prior serum lactate values entered in the laboratory information system, then subsequent predictions should benefit from the creation of a potentially more robust trajectory than if only a single prior lactate is available (as is the case for this paper). Potentially, the AUCs generated in this instance would be at least as high, and likely higher than those we have reported, making the tool a progressively more accurate and useful one.

The relatively suboptimal model performance (AUC 0.72–0.83) compared to other machine learning in healthcare publications (including ours), highlights the challenge of predicting the trajectory of serum lactate during critical illness. Features pertaining to the immunologic response that are specific to a patient, likely are only partially captured in the clinical data currently collected in the process of care. However, we assert that having a model with decent discrimination to inform clinicians when to check serum lactate is still an improvement compared to the variation across clinicians with regard to when serum lactate is ordered. We suspect the model performance can be improved by training on a larger dataset. An acceptable precision should be set if the intent is to reduce unnecessary testing and by how much. An acceptable recall should be set if the intent is to detect deterioration early and increase patient monitoring or move the patient to a higher level of care. Furthermore, inclusion of treatments administered would further improve the performance of the model and address some of the challenges with predicting the trajectory of serum lactate.

While our results are specifically calculated for serum lactate, the method could, upon appropriate contextual recalculation of the algorithm, be applied to other often repeated laboratory tests such as serum glucose or hemoglobin, similarly alerting clinicians of the need to recheck a value on the basis of a predicted high probability of a clinically important and potentially actionable change in that value.

Our results suggest that it is feasible to predict future deterioration of serum lactate levels using routine clinical observations. However, there are several limitations. Validation studies on a target population are necessary before these models can safely be deployed in a clinical setting. The models also require regular recalibration to capture shifts in clinical practice and patient profiles over time. With these safeguards in place, our models could serve as a point-of-care tool that assists in the prediction of serum lactate values. This could provide an early warning for potential deterioration, prompting confirmatory testing of serum lactate levels (potentially automatically in the future) while at the same time reducing the number of unnecessary serum lactate blood tests for stable patients, and enable clinicians to have a more personalized approach to the care of critically ill patients. The approach allows for clinicians to independently determine the need for serum lactate values, so that it represents a supplement to, rather than a replacement of, clinical judgment.

To the best of our knowledge, this is the first study that attempts to use physiological and routinely measured markers to predict serum lactate deterioration. Strengths of this study include robust machine learning methodology able to capture temporal relationships between time-varying variables, while at the same time providing interpretability of the results. However, as this was a retrospective study, there were missing data for some of the variables. While our external validation results are encouraging, the models trained on the MIMIC-III data were obtained from a single center in the United States, and its performance might not generalize to external populations; the eICU-CRD data were restricted to ICUs in the United States and might not generalize to global populations. Furthermore, a bias in the model might have been introduced by selective measurement of serum lactate. The ideal dataset to train the model would have serum lactate drawn from every patient every hour. However, as such a dataset does not exist a prospective validation would be required to determine if such a bias exists and to what extent.

Our intention with this study was to provide insight into appropriate methodology and present a sound approach in predicting a biochemical marker for clinical application as opposed to a highly accurate ungeneralizable model.

Clinical prediction models demonstrate their true utility only if they can positively impact clinical practice and

patient outcomes without unacceptably negative impacts on workflow and/or costs. Apart from validation studies as described above, prospective evaluation in a clinical setting is required to measure the impact such models have on clinical decision making by nurses and physicians, and whether the subsequent changes in practice translate to desirable outcomes related to number of serum lactate tests ordered, rates of organ failure, length of ICU stay, and hospital mortality. We provide the parameters used to develop the models in the supplementary information to enable replication and extension of this work in other patient populations.

5 Conclusion

Compared to other clinical outcome prediction algorithms, our model performance seems suboptimal. Serum lactate is challenging to predict as lactate metabolism results from a complex interplay of factors pertaining to the patient, factors pertaining to the disease or injury, treatments, and patient response to treatment, and perhaps a signature genetically encoded host response. Omics data may be a proxy of the last element but is currently not captured in routine patient care. The difficulty with serum lactate prediction is in effect a missing data issue. Despite this, the LSTM model provided the highest performance in predicting lactate value deterioration in critically ill patients, followed by RF and LR. This suggests that the use of machine learning might be a useful adjunct in helping to predict serum lactate deterioration in a manner that can inform clinician decision-making. Further studies are needed to evaluate its utility in clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10877-021-00739-4.

Author contributions LAC, VO and BM conceived the presented idea. VO and BM developed the theory and performed the computations. BM, LAMS, WY and DJS led the drafting of the manuscript. LAC and VO supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Funding LAC is funded by the National Institute of Health through NIBIB R01 EB017205.

Data availability The datasets analyzed in the current study are publicly available in the MIMIC-III repository (https://mimic.physionet.org/) and eICU-CRD repository (https://eicu-crd.mit.edu/).

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval The data in MIMIC-III was 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. The analysis using the 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). All experiments were performed in accordance with relevant guidelines and regulations.

References

1. Liu Z, et al. Prognostic accuracy of the serum lactate level, the SOFA score and the qSOFA score for mortality among adults with Sepsis. Scand J Trauma Resusc Emerg Med. 2019;27(1):51. https://doi.org/10.1186/s13049-019-0609-3.
2. Arnold RC, et al. Multicenter study of early lactate clearance as a determinant of survival in patients with presumed sepsis. Shock. 2009. https://doi.org/10.1097/SHK.0b013e3181971d47.
3. Nguyen HB, et al. Early lactate clearance is associated with improved outcome in severe sepsis and septic shock. Crit Care Med. 2004. https://doi.org/10.1097/01.CCM.0000132904.35713.A7.
4. Suistomaa M, Ruokonen E, Kari A, Takala J. Time-pattern of lactate and lactate to pyruvate ratio in the first 24 h of intensive care emergency admissions. Shock. 2000. https://doi.org/10.1097/00024382-200014010-00002.
5. Bruno RR, et al. Failure of lactate clearance predicts the outcome of critically Ill septic patients. Diagnostics. 2020. https://doi.org/10.3390/diagnostics10121105.
6. Claridge JA, Crabtree TD, Pelletier SJ, Butler K, Sawyer RG, Young JS. Persistent occult hypoperfusion is associated with a significant increase in infection rate and mortality in major trauma patients. J Trauma. 2000. https://doi.org/10.1097/00005373-200001000-00003.
7. Jansen TC, et al. Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. Am J Respir Crit Care Med. 2010. https://doi.org/10.1164/rccm.200912-1918OC.
8. Daurat A, Dick M, Louart B, Lefrant JY, Muller L, Roger C. Continuous lactate monitoring in critically ill patients using microdialysis. Anaesth Crit Care Pain Med. 2020. https://doi.org/10.1016/j.accpm.2020.05.018.
9. Horwitz SMC, et al. Anemia and blood transfusion in critically ill patients. J Am Med Assoc. 2002. https://doi.org/10.1001/jama.288.12.1499.
10. Tosir P, Kanitsa N, Kanitsa A. Approximate iatrogenic blood loss in medical intensive care patients and the causes of anemia. J Med Assoc Thail. 2010;93(Suppl 7):S271–6.
11. Ong EL, Lim NL, Koay CK. Towards a pain-free venepuncture. Anaesthesia. 2000. https://doi.org/10.1046/j.1365-2044.2000.01124.x.
12. McCormick RD, Maki DG. Epidemiology of needle-stick injuries in hospital personnel. Am J Med. 1981. https://doi.org/10.1016/0002-9343(81)90558-1.
13. Cismondi F, et al. Reducing unnecessary lab testing in the ICU with artificial intelligence. Int J Med Inform. 2013. https://doi.org/10.1016/j.ijmedinf.2012.11.017.
14. Johnson AEW, et al. MIMIC-III, a freely accessible critical care database. Sci data. 2016;3: 160035. https://doi.org/10.1038/sdata.2016.35.
15. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely

available multi-center database for critical care research. Sci Data. 2018;5(1): 180178. https://doi.org/10.1038/sdata.2018.178.
16. Casserly B, et al. Lactate measurements in sepsis-induced tissue hypoperfusion: results from the surviving sepsis campaign database. Crit Care Med. 2015. https://doi.org/10.1097/CCM.0000000000000742.
17. Mamandipoor B, Majd M, Moz M, Osmani V. Blood lactate concentration prediction in critical care. Stud Health Technol Inform. 2020;270:73–7. https://doi.org/10.3233/SHTI200125.
18. Mamandipoor B, Majd M, Moz M, Osmani V. Blood lactate concentration prediction in critical care patients: handling missing values. Stud Health Technol Inform. 2020;270:73–7.
19. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, 1995. pp. 278–282
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997. https://doi.org/10.1162/neco.1997.9.8.1735.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions, 2017
22. Kraut JA, Madias NE. Metabolic acidosis: pathophysiology, diagnosis and management. Nat Rev Nephrol. 2010. https://doi.org/10.1038/nrneph.2010.33.
23. Eichenholz A. Respiratory Alkalosis. Arch Intern Med. 1965. https://doi.org/10.1001/archinte.1965.03870050053009.
24. MacEdo E, Malhotra R, Bouchard J, Wynn SK, Mehta RL. Oliguria is an early predictor of higher mortality in critically ill patients. Kidney Int. 2011. https://doi.org/10.1038/ki.2011.150.
25. Vincent JL, et al. The clinical relevance of oliguria in the critically ill patient: analysis of a large observational database. Crit Care. 2020. https://doi.org/10.1186/s13054-020-02858-x.
26. Grander W, Muellauer KM, Duenser MD. Heart rate before ICU discharge: a simple and readily available predictor of short- and long-term mortality from critical illness. Eur Heart J. 2013. https://doi.org/10.1093/eurheartj/eht308.p1434.
27. Kramer L, Jordan B, Druml W, Bauer P, Metnitz PGH. Incidence and prognosis of early hepatic dysfunction in critically ill patients—a prospective multicenter study. Crit Care Med. 2007. https://doi.org/10.1097/01.CCM.0000259462.97164.A0.
28. Gladden LB. Lactate metabolism: a new paradigm for the third millennium. J Physiol. 2004;558(1):5–30. https://doi.org/10.1113/jphysiol.2003.058701.
29. Phypers B, Pierce JT. Lactate physiology in health and disease. Contin Educ Anaesth Crit Care Pain. 2006. https://doi.org/10.1093/bjaceaccp/mkl018.
30. Lindsell CJ, Stead WW, Johnson KB. Action-informed artificial intelligence—matching the algorithm to the problem. JAMA. 2020. https://doi.org/10.1001/jama.2020.5035.