Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records

Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Kirstine Belling, Søren Brunak, Anders Perner

Summary

Background Many mortality prediction models have been developed for patients in intensive care units (ICUs); most are based on data available at ICU admission. We investigated whether machine learning methods using analyses of time-series data improved mortality prognostication for patients in the ICU by providing real-time predictions of 90-day mortality. In addition, we examined to what extent such a dynamic model could be made interpretable by quantifying and visualising the features that drive the predictions at different timepoints.

Methods Based on the Simplified Acute Physiology Score (SAPS) III variables, we trained a machine learning model on longitudinal data from patients admitted to four ICUs in the Capital Region, Denmark, between 2011 and 2016. We included all patients older than 16 years of age, with an ICU stay lasting more than 1 h, and who had a Danish civil registration number to enable 90-day follow-up. We leveraged static data and physiological time-series data from electronic health records and the Danish National Patient Registry. A recurrent neural network was trained with a temporal resolution of 1 h. The model was internally validated using the holdout method with 20% of the training dataset and externally validated using previously unseen data from a fifth hospital in Denmark. Its performance was assessed with the Matthews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUROC) as metrics, using bootstrapping with 1000 samples with replacement to construct 95% CIs. A Shapley additive explanations algorithm was applied to the prediction model to obtain explanations of the features that drive patient-specific predictions, and the contributions of each of the 44 features in the model were analysed and compared with the variables in the original SAPS III model.

Findings From a dataset containing 15 615 ICU admissions of 12 616 patients, we included 14 190 admissions of 11 492 patients in our analysis. Overall, 90-day mortality was 33·1% (3802 patients). The deep learning model showed a predictive performance on the holdout testing dataset that improved over the timecourse of an ICU stay: MCC 0·29 (95% CI 0·25–0·33) and AUROC 0·73 (0·71–0·74) at admission, 0·43 (0·40–0·47) and 0·82 (0·80–0·84) after 24 h, 0·50 (0·46–0·53) and 0·85 (0·84–0·87) after 72 h, and 0·57 (0·54–0·60) and 0·88 (0·87–0·89) at the time of discharge. The model exhibited good calibration properties. These results were validated in an external validation cohort of 5827 patients with 6748 admissions: MCC 0·29 (95% CI 0·27–0·32) and AUROC 0·75 (0·73–0·76) at admission, 0·41 (0·39–0·44) and 0·80 (0·79–0·81) after 24 h, 0·46 (0·43–0·48) and 0·82 (0·81–0·83) after 72 h, and 0·47 (0·44–0·49) and 0·83 (0·82–0·84) at the time of discharge.

Interpretation The prediction of 90-day mortality improved with 1-h sampling intervals during the ICU stay. The dynamic risk prediction can also be explained for an individual patient, visualising the features contributing to the prediction at any point in time. This explanation allows the clinician to determine whether there are elements in the current patient state and care that are potentially actionable, thus making the model suitable for further validation as a clinical tool.

Funding Novo Nordisk Foundation and the Innovation Fund Denmark.

Introduction

Improved prediction of a patient's risk of dying would be of value in decision making in the intensive care unit (ICU) setting. Several prognostic scores such as the Simplified Acute Physiology Score (SAPS), the Acute Physiologic Assessment and Chronic Health Evaluation, and the Mortality Prediction Model have been developed for ICU populations,[1–4] but low precision impedes patient-level use.[1,5,6] Furthermore, most existing scores are static as they are calculated from

Lancet Digital Health 2020; 2: e179–91

Published Online March 12, 2020 https://doi.org/10.1016/S2589-7500(20)30018-2

See Comment page e152

Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (H-C Thorsen-Meyer MD, A B Nielsen PhD, A P Nielsen MD, B S Kaas-Hansen MD, P J Chmura MSc, K Belling PhD, Prof S Brunak PhD); **Department of Intensive Care, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark** (H-C Thorsen-Meyer, Prof A Perner PhD); **Clinical Pharmacology Unit, Zealand University Hospital, Roskilde, Denmark** (B S Kaas-Hansen); **Department of Anesthesiology and Intensive Care, Odense University Hospital, Odense, Denmark** (Prof P Toft DMSc, J Schierbeck MD, T Strøm PhD); **Department of Clinical Research, University of Southern Denmark, Odense, Denmark** (Prof P Toft, J Schierbeck, T Strøm); **Centre for IT, Medical Technology and Telephony Services, Capital Region of Denmark, Copenhagen, Denmark** (M Heimann BSc); **and Daintel, Lyngby, Denmark** (L Dybdahl MSc, L Spangsege MSc, P Hulsen BSc)

Correspondence to: Prof Søren Brunak, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark **soren.brunak@cpr.ku.dk**

Research in context

Evidence before this study

Much effort has been put into the development of scoring systems to predict the risk of mortality for critically ill patients. Traditionally, linear models such as logistic regression analysis have been used to construct such prognostication tools. However, in recent years, scores based on non-linear machine learning have emerged. We searched MEDLINE for studies published in any language from inception to May 27, 2019, using the terms ("mortality" OR "survival") AND ("machine learning" OR "artificial intelligence" OR "neural network") AND ("intensive care" OR "critical care"). We found 158 papers, of which 34 were original studies relevant for our study. Six of the studies used machine learning to develop dynamic or real-time mortality prediction models and all reported superior performance of their models compared with traditional logistic regression models. Hence, there is a growing amount of evidence that machine learning models can provide a more accurate outcome prediction to support decision making when dealing with critically ill patients. Still, none of the six studies included a methodology to make the model directly interpretable. Importantly, machine learning predictions will need to be transparent to ensure medical professionals embrace this new technology.

Added value of this study

We developed a deep learning model based on routinely collected data capable of providing real-time predictions of mortality risk for critically ill patients in intensive care units. The model is updated every hour and integrates new observations as they arrive, thus mimicking the reasoning of a medical professional. This approach allows the model not only to learn from actual values of the variables at a given time, but also from the temporal trend in the data obtained hourly. Finally, the model provides real-time explanations of the features that drive a prediction. Such a comprehensible deconvolution allows medical professionals to combine the temporal predictions with their existing beliefs to facilitate decision making.

To our knowledge, this is the first time a machine learning model produces explanations in a longitudinal fashion as the patient's condition develops. Such temporal rankings of features might assist medical professionals in deciding on timing of interventions during admissions.

Implications of all the available evidence

The study shows that it is possible to make dynamic and easily interpretable models that predict mortality in critically ill patients. Such models can deliver new insight into complex interactions, non-linearities, and the importance of trends in the explanatory variables. The model is based on few variables and, as such, is only meant as a proof-of-concept study. We are currently working on a more extensive model to investigate to what extent the predictive power can be further improved by introducing more information about the patients. Before this kind of model can be used as a decision support tool, the results need to be confirmed in a prospective clinical trial.

data obtained during the first day of ICU admission. Lacking adequate tools for patient-level prognostication, clinicians could resort to subjective judgment, which is prone to bias.[7] In addition, the events that markedly change the prognosis for an individual patient can be missed.

Modern ICUs generate vast amounts of data in a continuous stream containing valuable information on subsequent patient outcomes. The data are generally heterogeneous and comprise both structured and unstructured information with irregular sampling, artifacts, and varying degree of completeness, which challenges traditional statistical models. Machine learning can extract information from incomplete, complex data and provide insights to support clinical decision making. Recent in-silico research has yielded machine learning-based mortality prediction models using ICU data that were superior to traditional methods.[8–10] While it is not surprising that a complex model can outperform a simpler one, improved performance comes at a price: these models might be perceived as so-called black boxes, which could limit their acceptance among clinicians and raise legal and ethical concerns. Recently, algorithms that explain patient-specific predictions have emerged that might increase the understanding of and trust in machine learning prediction models.[11,12] This could, in turn, facilitate the translation of machine learning models into clinical decision-support tools.

In this Article, we report on the development of a machine learning model that produces hourly patient-level predictions of 90-day mortality in patients admitted to the ICU.

Methods

Overview

The workhorse of our mortality prediction model is an artificial neural network with a long short-term memory (LSTM) architecture that integrates static baseline data and accruing data with a setup approximating the SAPS III model to link it to current clinical practice (figure 1). To substantiate the clinical use of the model and increase its explainability, we visualise the features that drive patient-specific predictions with 1-h intervals.

This study was approved by the Danish Patient Safety Authority (3–3013–1723), the Danish Data Protection Agency (DT SUND 2016–48, 2016–50 and 2017–57) and the Danish Health Data Authority (FSEID 00003724). This paper adheres to relevant items in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement.[13]

Data sources

For model development, we retrospectively collated data covering the period Sept 6, 2011, to May 22, 2016, from electronic patient records (EPRs), the Central Person Registry (CPR), and the Danish National Patient Registry (DNPR) for patients admitted to four mixed medical and surgical ICUs in the Capital Region of Denmark. We forewent prior power calculations and used all accessible data. The EPR system used in the ICUs in the study period was the Critical Information System (CIS; developed by Daintel, Copenhagen, Denmark), customised for ICUs to store demographic and high-frequency data collected from equipment such as monitors, ventilators, and infusion pumps. The DNPR is a nationwide registry containing data on all procedures and diagnoses encoded in a Danish adaptation of the International Classification of Diseases.[14] Additionally, laboratory values were extracted from Labka II (DXC Technology; Tysons, VA, USA) and BCC (CGI; Montreal, QC, Canada) clinical laboratory information systems.

For model validation, we obtained an external dataset also stored in the CIS format from a fifth hospital located in the Region of Southern Denmark, including all patients admitted there between June 7, 2012, and Jan 27, 2017.

To adhere to the SAPS III model, we constrained both the development and validation cohorts to patients older than 16 years with an ICU stay lasting more than 1 h. To have available follow-up, we further limited the cohorts to patients with a Danish civil registration number (all Danish residents) who were retained (not emigrated or reported missing) in the DNPR at least 90 days after their ICU admission. The outcome—all-cause 90-day mortality after the date of ICU admission—was obtained from the CPR.

Variables and features

Data included in SAPS III are patient characteristics before ICU admission, type of admission, and markers of physiological derangement during the first hour in the ICU.[15] These variables are a mix of static information (eg, demographics and diagnoses), daily obtained information (eg, laboratory values), and data obtained with high sampling rate (eg, from monitoring equipment). To remove questionable records from our development and validation datasets, we applied the plausible ranges used during development of the SAPS III model (appendix p 6).[15] Minimum value for leucocytes and maximum value for systolic blood pressure (SBP) were considered too strict, and were set to 0 billion per L instead of 1 billion per L for leucocytes and 300 mm Hg instead of 200 mm Hg for SBP. Measurements outside these ranges were set as missing. We were unable to extract information for three SAPS III variables in an automated way: use of major therapeutic options before ICU admission, reasons for ICU admission, and acute infection at ICU admission. Consequently, these variables were excluded from our model.

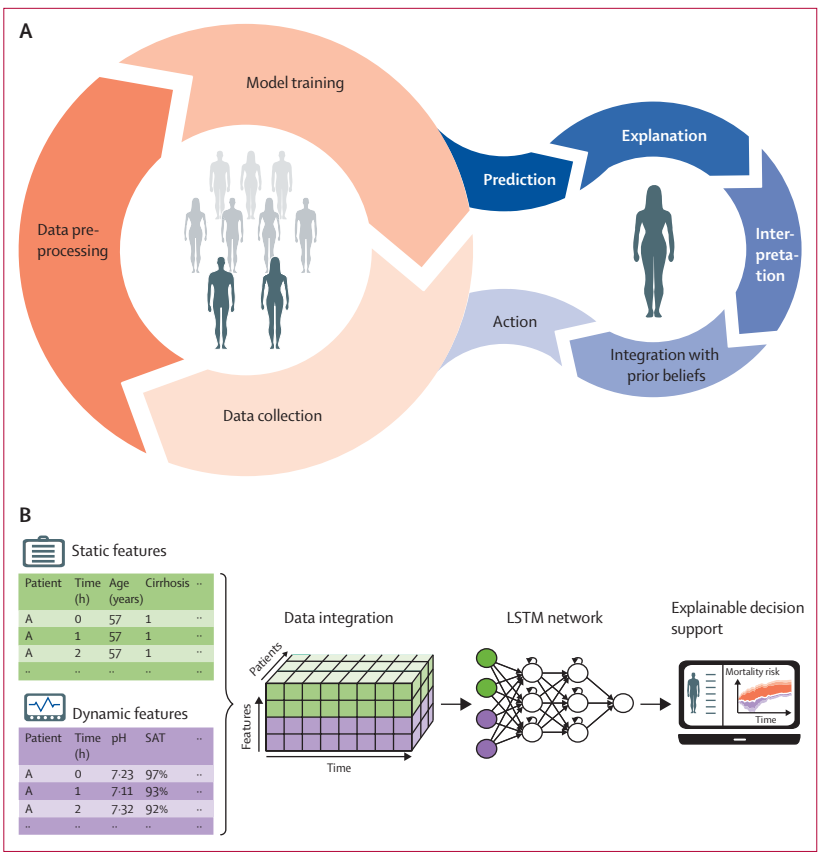


Figure 1: **Study overview: from raw data to explainable decision support**
(A) Raw data were obtained from a large ICU cohort, cleaned, and pre-processed. A mortality risk prediction model was trained using an LSTM neural network, which was updated hourly. The model can provide individual and consecutive mortality predictions for patients in the ICU. (B) Static features and dynamic features (ie, physiological time-series data), shown with different colours, were integrated as they became available during the ICU stay. The LSTM network learned relevant inferences over time to predict mortality. The predictions were rendered explainable and potentially actionable by visualising the features that drove the prediction at every timepoint. ICU=intensive care unit. LSTM=long short-term memory.

In the following, we distinguish variables from features: the former represent the raw input variables from SAPS III, whereas the latter are computed from the input variables during pre-processing.[16] To optimise the data for computation, continuous variables were aggregated to 1-hourly values. Variables potentially measured more than once per hour were aggregated (down-sampling), yielding three new features: minimum, maximum, and median. Variables typically measured less frequently than every hour were aggregated to hourly medians (up-sampling). Laboratory values obtained up to 24 h before ICU admission were included in the baseline aggregation upon admission. Missing values, including absent datapoints due to up-sampling, were imputed by last observation carried forward (LOCF). Population means in the datasets were used for missing values occurring before the first actual measurement. Non-binary categorical data were dummy coded and static variables were repeated at each timepoint (figure 1).

See Online for appendix

Model development

We applied a deep learning method to predict 90-day mortality in patients admitted to the ICU. Specifically, we used a recurrent neural network consisting of LSTM units capable of updating its prediction hourly by integrating new data as they accrue and learning from the temporal development of the features.[17] We chose hourly predictions to strike a balance between continuously obtaining new predictions and keeping the complexity of the model manageable. Essentially, an LSTM model takes a (temporal) sequence of input data, then learns and retains the patterns in these longitudinal data useful for the prediction task at hand. Technically, an LSTM neuron has three main gates: the input gate determines whether to let new inputs in, the forget gate determines whether to discard currently used information because it is not or no longer important, and the output gate determines whether to let the input affect the output at the time step in question. As such, LSTM networks incorporate data from the past to make predictions about the future, making them suitable for time-series prediction. LSTM networks, however, have many parameters to estimate, which requires large datasets and a powerful computing infrastructure.

We used the holdout method and split ICU admissions into a training dataset (80%) and a test dataset (20%) for internal validation. All performance metrics were derived from the test dataset and the external validation dataset. To deal with overfitting in model selection, hyperparameters and model architecture were chosen using a five-fold cross-validation on the training data.[18] When a patient had more than one ICU admission, all ICU stays were assigned to the same cross-validation fold to avoid information leaking between the folds;[19] allowing admissions from the same patient across different cross-validation folds would entail that some information, such as comorbidities, used for training could also be present in the validation set, thus compromising its independence. We remedied the imbalanced nature of the dataset, which had more survivors than non-survivors, by keeping all admissions from the minority class (ie, non-survivors) and a new randomly selected subset of equal size from the majority class (ie, survivors) in each epoch in the training process. In each epoch (ie, one cycle through the training dataset), we leave out some data: however, over many epochs the LSTM will see the full dataset.[20,21] To find the optimal hyperparameters, we tested all 336 combinations of the following settings: number of hidden LSTM neurons (1, 4, 8, 16, 32, 64, or 128), number of hidden layers (1, 2, or 3), dropout (0, 0·2, 0·4, 0·6), optimisation function (RMSprop or Adagrad), and balancing of classes (on or off). The activation function (sigmoid) and number of batches (n=1) were fixed during training. The number of hidden neurons and layers determine the complexity of the model, while dropout randomly drops input neurons from the network during training to prevent overfitting. During the training process, the optimisation function determines how to gradually modify the model parameters to lower the prediction error quantified by the loss function.

Explainable predictions

We applied a Shapley additive explanations (SHAP) algorithm to our prediction model to obtain explanations of the features that drive patient-specific predictions to mitigate the issue of black-box predictions.[11,12] SHAP is a model-agnostic representation of feature importance where the impact of each feature on a particular prediction is represented using Shapley values inspired by cooperative game theory.[11,22,23] A Shapley value states, given the current set of feature values, how much a single feature in the context of its interaction with other features contributes to the difference between the actual prediction and the mean prediction. That is, the sum of the Shapley values for all features plus the mean prediction equals the actual prediction.[11,22,23] Importantly, this is not the same as direct feature effects known from (generalised) linear models. The SHAP value for a feature should not be seen as its direct, and isolated effect, but as its compound effect when interacting also with the other features.

For comparison, we also show how the variables in the original SAPS III model contribute to the predictions. We calculate these contributions using a background distribution of SAPS III admission scores similar to the distribution in the original SAPS III paper.[24] The calibration for northern Europe was used to calculate the probability of non-survival.

Model performance

To evaluate the ability to discriminate survivors from non-survivors, we used Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), positive and negative predictive values, and positive and negative likelihood ratios. The MCC is defined as

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

where TP and TN indicate true positives and negatives, and FP and FN indicate false positives and negatives. MCC is a quality measure for classification models and can be seen as a discretisation of the Pearson correlation for binary variables,[25] taking values between 1 (all examples correctly predicted) and –1 (all examples incorrectly predicted). MCC is considered a balanced measure, rendering it useful for imbalanced datasets.[26]

Positive and negative likelihood ratios are used for assessing the value of performing a diagnostic test. The sensitivity and specificity of the model are used to determine whether a positive or negative prediction usefully changes the probability that a patient will die

within 90 days from ICU admission.[27] The positive likelihood ratio is calculated as

$$\frac{\text{sensitivity}}{1-\text{specificity}} = \frac{P(\text{positive test} \mid \text{disease})}{P(\text{positive test} \mid \text{no disease})}$$

and the negative likelihood ratio as

$$\frac{1-\text{sensitivity}}{\text{specificity}} = \frac{P(\text{negative test} \mid \text{disease})}{P(\text{negative test} \mid \text{no disease})}$$

Bootstrapping was used to construct 95% CIs around the estimates using 1000 bootstrap samples of mortality prediction probabilities with replacement.

Model calibration

We gauged the calibration visually by inspecting how hour-specific calibration curves aligned with the diagonal line that represented perfect calibration.[28,29] We adopted the usual approach for binary outcomes of plotting means of decile-binned predictions on the *x*-axis and means of the observed outcomes in the patients in each bin on the *y*-axis.[30,31] To quantify the calibration, we computed calibration slopes and intercepts independently for each hourly prediction. To aggregate these to compound metrics, we fitted intercept-only meta-regressions on the slopes and intercepts, weighted by the number of patients still in the cohort at the respective timepoints. We assessed the sensitivity of the calibration results by dividing the predictions into 15 and 20 bins. Because the predictions resulting from the initial training were globally pessimistic (ie, predicting higher risk than observed; appendix p 3), we used isotonic regression for prediction calibration to obtain more reliable isotonic predictions.[32,33] The calibration analyses were carried out for both initial and isotonic predictions.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The first and corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

For model development, we obtained data on 12 616 patients with 15 615 ICU admissions. 11 492 patients with 14 190 ICU admissions were eligible for inclusion in the model development dataset, of which we allocated approximately 20% of patients (2299 patients with 2825 admissions) to the holdout test dataset (figure 2). The table shows baseline characteristics of the patients in the training dataset and holdout test dataset using the data from their first ICU admission. In the development dataset, the median age was 65 years (IQR 52–75) and 4816 (41·9%) were female. 1815 (15·7%) patients died in the ICU, 3389 (29·5%) in hospital, and 3802 (33·1%) by 90 days after ICU admission.

When predicting 90-day mortality after ICU admission, the predictive performance of our model increased over time (figure 3). The AUROC upon ICU admission in the holdout test dataset was 0·73 (95% CI 0·71–0·74) and increased to 0·85 (0·84–0·87) at 72 h. The corresponding MCCs were 0·29 (0·25–0·33) and 0·50 (0·46–0·53), respectively. When evaluating the performance relative to time elapsed since admission, for some patients, the time of prediction will approach the time of death. To deal with this issue, we also evaluated the predictive performance relative to the time of discharge (figure 3). At time of discharge from the ICU, the model achieved an AUROC of 0·88 (0·87–0·89) in the holdout test dataset, whereas 24 h before discharge the AUROC was 0·82 (0·80–0·84); the corresponding MCCs were 0·57 (0·54–0·60) and 0·46 (0·42–0·50), respectively.

We assessed the external validity of the model on our external validation dataset, comprising 5827 unique ICU

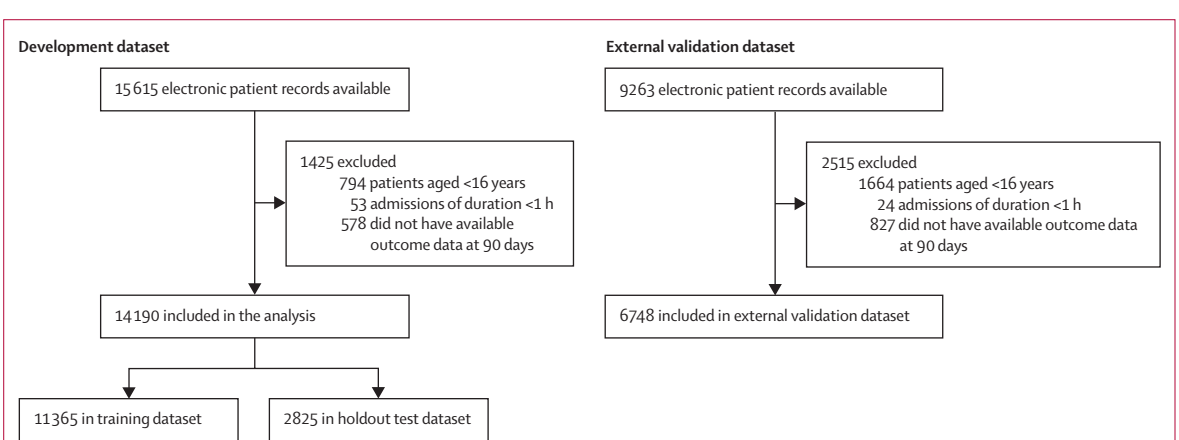


Figure 2: **Study profile**

patients with a total of 6748 admissions (figure 2). Overall, the predictive ability of the model was less accurate in the external validation dataset according to AUROC or MCC, with an MCC of 0·29 (95% CI 0·27–0·32) and AUROC of 0·75 (0·73–0·76) at admission, 0·41 (0·39–0·44) and 0·80 (0·79–0·81) after 24 h, 0·46 (0·43–0·48) and 0·82 (0·81–0·83) after 72 h, and 0·47 (0·44–0·49) and 0·83 (0·82–0·84) at the time of discharge (figure 3). However, in the initial part of the ICU stay, the model performed slightly better in the external validation dataset compared with the holdout test dataset. Given differences in mortality rates and premorbid status of the patients in the two populations, such as a substantially lower proportion of patients with heart failure in the training data (table), a decrease in model performance was to be expected.

	Development dataset (n=11 492)		External validation dataset (n=5827)
	Training dataset (n=9193)	Holdout test dataset (n=2299)	
Age, years	65 (52–75)	65 (52–74)	66 (55–75)
Sex			
Female	3842 (41·8%)	974 (42·4%)	2419 (41·5%)
Male	5351 (58·2%)	1325 (57·6%)	3408 (58·5%)
Comorbidities			
AIDS	27 (0·3%)	5 (0·2%)	6 (0·1%)
Cancer therapy	390 (4·2%)	103 (4·5%)	288 (4·9%)
Chronic heart failure	29 (0·3%)	11 (0·5%)	800 (13·7%)
Cirrhosis	516 (5·6%)	120 (5·2%)	184 (3·2%)
Haematological cancer	436 (4·7%)	103 (4·5%)	231 (4·0%)
Metastatic cancer	372 (4·0%)	96 (4·2%)	431 (7·4%)
Admission category			
Medical	5323 (57·9%)	1324 (57·6%)	2713 (46·6%)
Scheduled surgery	612 (6·7%)	175 (7·6%)	1401 (24·0%)
Unscheduled surgery	3258 (35·4%)	800 (34·8%)	1713 (29·4%)
Type of surgery			
Transplantation			
Liver, kidney, or pancreas	175 (1·9%)	43 (1·9%)	9 (0·2%)
Combined kidney and pancreas, or other transplantation	7 (0·1%)	0	0
Cardiac surgery	265 (2·9%)	58 (2·5%)	119 (2·0%)
Trauma	474 (5·2%)	116 (5·0%)	207 (3·6%)
Neurosurgery	247 (2·7%)	57 (2·5%)	233 (4·0%)
Intra-hospital location before ICU admission			
Emergency room	2909 (31·6%)	691 (30·1%)	1549 (26·6%)
Other ICU	913 (9·9%)	236 (10·3%)	617 (10·6%)
Hospital ward, recovery unit, or operating room	5371 (58·4%)	1372 (59·7%)	3661 (62·8%)
Length of hospital stay before ICU, days	1·0 (0·0–4·0)	1·0 (0·0–4·0)	1·0 (0·0–2·0)
Length of ICU stay, days	1·9 (0·8–5·1)	2·0 (0·9–5·0)	1·3 (0·8–3·8)
Number of ICU admissions			
1	7642 (83·1%)	1916 (83·3%)	5135 (88·1%)
2	1146 (12·5%)	286 (12·4%)	540 (9·3%)
3	277 (3·0%)	66 (2·9%)	112 (1·9%)
≥4	128 (1·4%)	31 (1·3%)	40 (0·7%)
Mortality			
ICU mortality	1444 (15·7%)	371 (16·1%)	690 (11·8%)
In-hospital mortality	2709 (29·5%)	680 (29·6%)	1203 (20·6%)
90-day mortality	3054 (33·2%)	748 (32·5%)	1512 (25·9%)

Data are n (%) or median (IQR). For patients with multiple admissions, the data provided are from the first admission. ICU=intensive care unit.

Table: **Baseline characteristics of the ICU patients in the training, test, and external validation datasets**

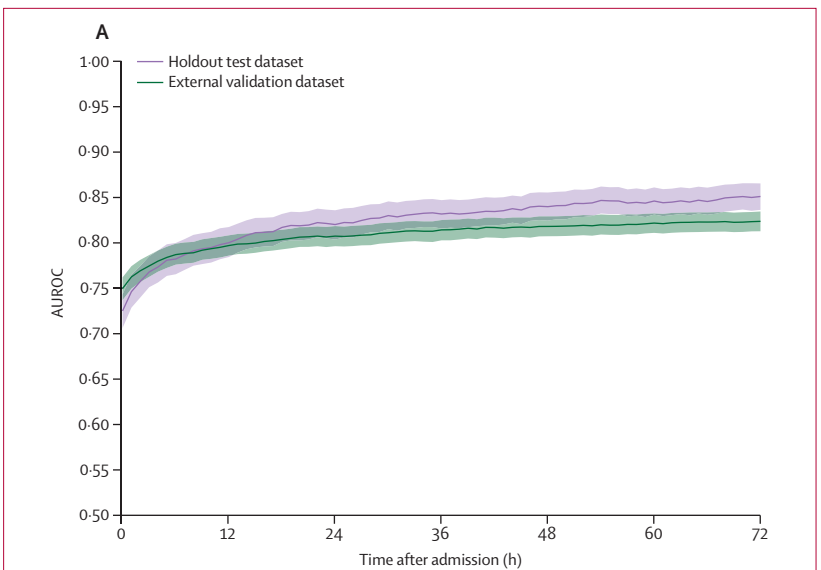
The calibration plots show that the hourly predictions, taken at face value, consistently overestimate the risk, whereas the isotonic predictions lie snugly around the diagonal (appendix p 3); early and late predictions deviate more, probably due to less available information (early predictions) and fewer patients (late predictions). Very early predictions are generally inferior (ie, less well calibrated), but the estimates converge after a few hours (appendix p 4). The compound isotonic calibration slope of 1·00 (95% CI 0·99 to 1·01) and intercept of –0·01 (–0·01 to –0·01) are close to ideal and robust to changes in binning of predictions (appendix p 7).

When considering the contribution of each of the 44 features in the model (figure 4), it is not surprising that age at admission has the greatest impact on the predictions, with older age driving the predictions towards non-survival and younger age driving the predictions towards survival. This is in keeping with the fact that age is the variable potentially yielding the second-most points in the SAPS III score. Most binary features predominantly influence mortality prediction when present in a uni-directional manner towards either survival or non-survival (eg, admission type of scheduled surgery pulls the prediction towards survival). For non-binary features in general, low values will drive mortality prediction towards either survival or non-survival and high values will drive the prediction in the opposite direction, although exceptions exist. An example is low median heart frequency (second top feature; figure 4), which generally drives mortality prediction towards survival but can be seen to drive predictions towards non-survival for some patients. For comparison, figure 4B illustrates the contributions for the features in the original SAPS III score. There are two marked differences: for all but one feature, the effect on mortality is unidirectional in the SAPS III score, whereas in our model, features can drive the prediction in either direction; and in the SAPS III score, individual features generally can have a greater effect on the prediction than in our model.

The dynamic risk prediction can also be explained at any given time for a particular patient; we illustrate a representative case from the holdout test cohort (figure 5). The patient was an 83-year-old female with a history of hypertension and paroxysmal atrial fibrillation. She was transferred from the medical ward to the ICU with hypoxic respiratory failure due to a community-acquired pneumonia, with a SAPS III score of 75 points at admission. She was initially treated with intermittent

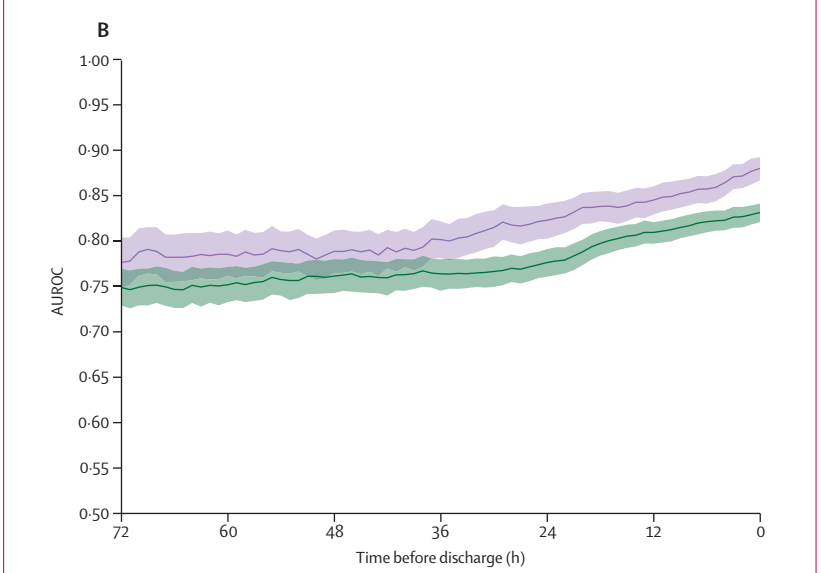
non-invasive ventilation but intubated 26 h after admission due to insufficient treatment response. Due to sedation and atrial fibrillation with a rapid ventricular response, the patient developed hypotension and vasopressor treatment was initiated 36 h after admission. Her condition gradually deteriorated from 40 h onwards, with an increasing oxygen demand and development of delirium. The patient died in the ICU 98 h after admission. In this case, age at admission drives mortality prediction towards non-survival throughout (dark orange ribbon), whereas median heart frequency is the most important feature pulling the prediction down towards survival for the bulk of the stay, along with median SBP and leucocytes. Some features can drive the prediction towards survival at one timepoint and towards non-survival at other timepoints (eg, minimum Glasgow Coma Scale [GCS]) or vice versa (eg, leucocytes), and others can oscillate between the two (eg, minimum SBP; figure 5A). In the same patient, the three most important contributions to the SAPS III model prediction at each hour all drive towards non-survival, with age at admission the most influential, followed by intra-hospital location before ICU and oxygenation, which are occasionally overtaken in importance by maximum heart frequency and minimum SBP (figure 5B). When considering the relative importance of all included features on the predictions for the full holdout test dataset over time, as illustrated by the mean rank, we note that mechanical ventilation gains importance (relative to the other features) over time, whereas creatinine, leucocytes, and platelets seem to lose importance, with their respective ranks diminishing (figure 6).

We further detail the importance of selected features and provide a visual example of how they interact (appendix p 9). When considering the GCS, it is clear that lower GCS is associated with a higher relative risk of non-survival (appendix p 9). Yet, due to feature interactions, the range of relative risks within each GCS level is quite wide; the imputed GCS values (using the population mean, which lay between 12 and 13) have essentially no impact on the prediction. The same pattern is observed for minimum and maximum hourly SBP, with low SBP associated with increased relative risk of non-survival and vice versa (appendix p 9). The final graph shows how the contributions from minimum and maximum SBP interact: the



AUROC	0·73 (0·71–0·74)	0·80 (0·78–0·82)	0·82 (0·80–0·84)	0·83 (0·82–0·85)	0·84 (0·82–0·86)	0·85 (0·83–0·86)	0·85 (0·84–0·87)
MCC	0·29 (0·25–0·33)	0·40 (0·36–0·44)	0·43 (0·40–0·47)	0·46 (0·43–0·50)	0·47 (0·44–0·51)	0·49 (0·45–0·52)	0·50 (0·46–0·53)
PPV	0·61 (0·57–0·65)	0·70 (0·66–0·74)	0·67 (0·64–0·71)	0·67 (0·64–0·70)	0·69 (0·65–0·71)	0·73 (0·70–0·77)	0·73 (0·70–0·76)
NPV	0·73 (0·71–0·75)	0·76 (0·74–0·78)	0·79 (0·77–0·80)	0·81 (0·79–0·82)	0·81 (0·79–0·82)	0·79 (0·78–0·81)	0·80 (0·78–0·82)
LRP	2·99 (2·60–3·49)	4·51 (3·90–5·28)	3·97 (3·51–4·56)	3·89 (3·47–4·37)	4·19 (3·74–4·74)	5·28 (4·61–6·11)	5·19 (4·56–5·96)
LRN	0·72 (0·68–0·75)	0·61 (0·57–0·65)	0·53 (0·49–0·56)	0·47 (0·42–0·50)	0·46 (0·42–0·50)	0·50 (0·46–0·54)	0·48 (0·44–0·51)
Prop	100·0% (n=2825)	89·3% (n=2522)	70·1% (n=1981)	59·7% (n=2825)	51·3% (n=1448)	45·2% (n=1276)	39·6% (n=1119)

AUROC	0·75 (0·73–0·76)	0·79 (0·78–0·81)	0·80 (0·79–0·81)	0·81 (0·80–0·82)	0·81 (0·80–0·83)	0·82 (0·81–0·83)	0·82 (0·81–0·83)
MCC	0·29 (0·27–0·32)	0·39 (0·36–0·41)	0·41 (0·39–0·44)	0·43 (0·41–0·46)	0·43 (0·40–0·45)	0·45 (0·43–0·48)	0·46 (0·43–0·48)
PPV	0·54 (0·51–0·56)	0·58 (0·55–0·60)	0·54 (0·52–0·56)	0·55 (0·53–0·57)	0·61 (0·59–0·64)	0·59 (0·57–0·61)	0·59 (0·56–0·61)
NPV	0·79 (0·78–0·80)	0·82 (0·81–0·83)	0·86 (0·85–0·87)	0·86 (0·85–0·87)	0·83 (0·82–0·84)	0·86 (0·85–0·86)	0·86 (0·85–0·87)
LRP	3·08 (2·81–3·38)	3·65 (3·37–3·96)	3·11 (2·91–3·32)	3·29 (3·08–3·51)	4·19 (3·87–4·58)	3·84 (3·59–4·14)	3·76 (3·52–4·03)
LRN	0·70 (0·67–0·73)	0·57 (0·54–0·59)	0·45 (0·42–0·48)	0·43 (0·40–0·46)	0·53 (0·50–0·56)	0·45 (0·42–0·48)	0·43 (0·41–0·46)
Prop	100·0% (n=6748)	89·2% (n=6020)	58·7% (n=3962)	50·5% (n=3407)	41·2% (n=2781)	36·9% (n=2488)	31·6% (n=2132)



AUROC	0·78 (0·75–0·80)	0·79 (0·76–0·81)	0·79 (0·76–0·81)	0·80 (0·78–0·82)	0·82 (0·80–0·84)	0·84 (0·83–0·86)	0·88 (0·87–0·89)
MCC	0·37 (0·31–0·42)	0·41 (0·36–0·46)	0·39 (0·34–0·44)	0·41 (0·37–0·46)	0·46 (0·42–0·50)	0·52 (0·48–0·56)	0·57 (0·54–0·60)
PPV	0·66 (0·61–0·71)	0·70 (0·65–0·74)	0·68 (0·64–0·72)	0·70 (0·66–0·74)	0·74 (0·71–0·78)	0·77 (0·74–0·80)	0·80 (0·77–0·83)
NPV	0·72 (0·68–0·75)	0·73 (0·70–0·76)	0·73 (0·70–0·76)	0·74 (0·72–0·77)	0·76 (0·74–0·78)	0·80 (0·78–0·82)	0·82 (0·80–0·83)
LRP	2·83 (2·39–3·44)	3·40 (2·86–4·14)	3·22 (2·76–3·82)	3·73 (3·19–4·40)	4·70 (3·97–5·53)	6·35 (5·43–7·60)	7·83 (6·70–9·35)
LRN	0·57 (0·51–0·63)	0·54 (0·48–0·59)	0·56 (0·51–0·61)	0·55 (0·51–0·60)	0·52 (0·48–0·56)	0·48 (0·44–0·51)	0·43 (0·40–0·47)
Prop	39·6% (n=1118)	45·1% (n=1275)	51·2% (n=1446)	59·6% (n=1684)	69·9% (n=1974)	89·2% (n=2519)	100·0% (n=2825)

AUROC	0·75 (0·73–0·77)	0·75 (0·73–0·77)	0·76 (0·74–0·78)	0·76 (0·75–0·78)	0·78 (0·76–0·79)	0·81 (0·80–0·82)	0·83 (0·82–0·84)
MCC	0·38 (0·34–0·42)	0·35 (0·31–0·39)	0·38 (0·35–0·41)	0·38 (0·34–0·41)	0·38 (0·35–0·41)	0·44 (0·41–0·46)	0·47 (0·44–0·49)
PPV	0·59 (0·56–0·62)	0·57 (0·54–0·60)	0·58 (0·55–0·61)	0·57 (0·54–0·60)	0·57 (0·54–0·59)	0·58 (0·55–0·60)	0·61 (0·59–0·63)
NPV	0·78 (0·75–0·80)	0·77 (0·75–0·79)	0·79 (0·77–0·81)	0·80 (0·78–0·82)	0·81 (0·79–0·82)	0·85 (0·84–0·86)	0·86 (0·85–0·87)
LRP	2·25 (2·04–2·47)	2·24 (2·06–2·46)	2·37 (2·20–2·59)	2·52 (2·30–2·74)	2·62 (2·42–2·84)	3·75 (3·48–4·06)	4·16 (3·86–4·49)
LRN	0·44 (0·40–0·49)	0·50 (0·46–0·54)	0·45 (0·41–0·49)	0·48 (0·45–0·53)	0·48 (0·44–0·52)	0·47 (0·44–0·50)	0·45 (0·42–0·47)
Prop	31·5% (n=2128)	36·8% (n=2486)	41·1% (n=2776)	50·4% (n=3402)	58·6% (n=3957)	89·0% (n=6008)	100·0% (n=6748)

Figure 3: **Model performance in the holdout test dataset and external validation dataset as a function of observation period**
(A) AUROC as a function of time after ICU admission and (B) AUROC as a function of time before ICU discharge. The metrics for each timepoint in the graphs are displayed in the tables below with 95% CIs in parentheses. AUROC=area under the receiver operating characteristic curve. MCC=Matthews correlation coefficient. PPV=positive predictive value. NPV=negative predictive value. LRP=likelihood ratio positive. LRN=likelihood ratio negative. Prop=proportion of the total number of test patients admitted at a given timepoint. ICU=intensive care unit.

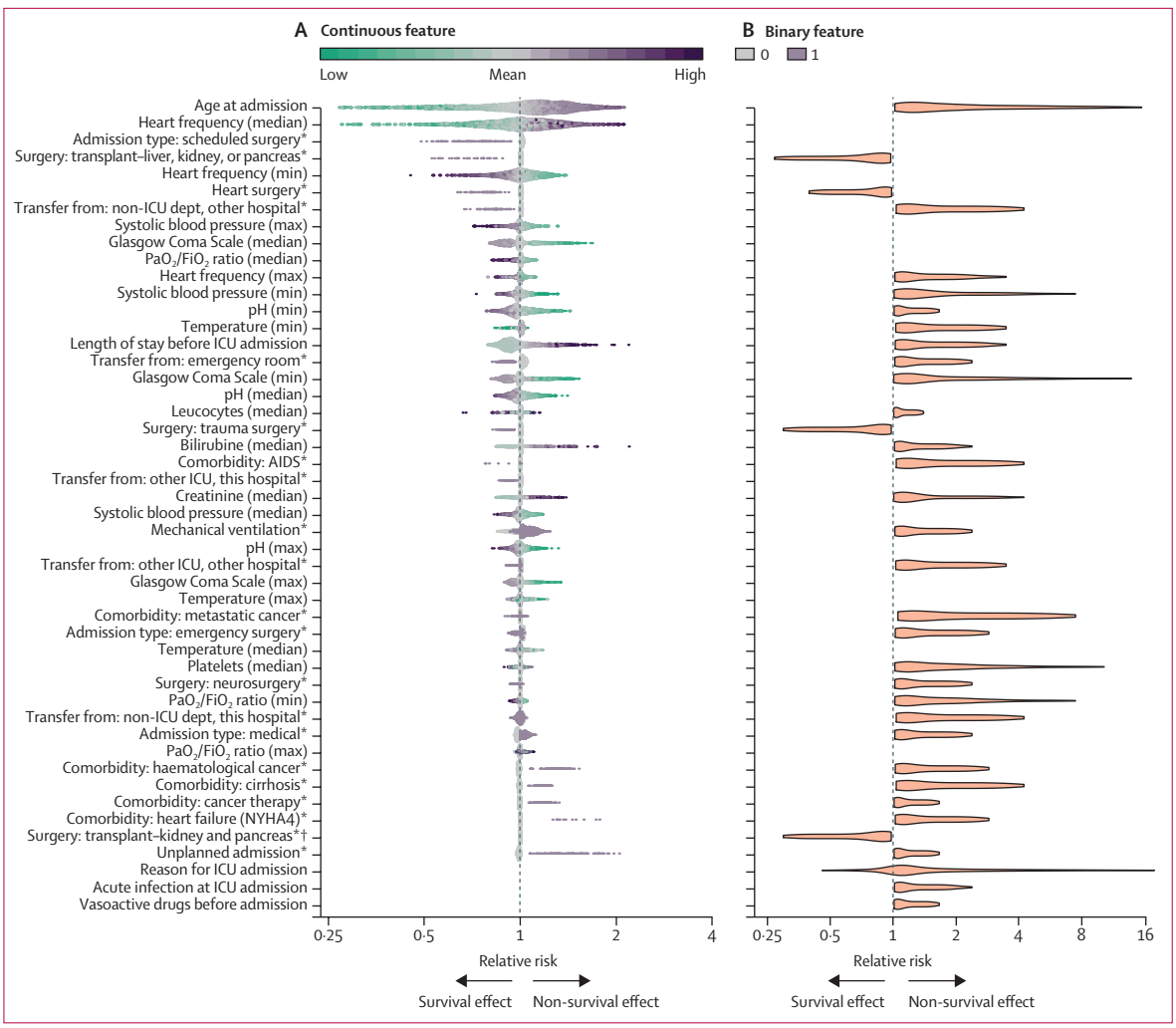


Figure 4: **The impact of the input features on predictions**

(A) The model includes both continuous and binary input features. Continuous features vary from low to high values, whereas binary features are either present or absent. Each dot represents the impact of a feature on the mortality prediction for one patient at a given point in time. As new mortality estimates are hourly, any given patient can be represented multiple times, depending on the duration of the ICU admission. Dots to the left represent patients with feature values that pull the prediction towards survival and dots to the right represent patients with feature values that drag the prediction towards non-survival. (B) The theoretical impact of the input features in the original SAPS III model on predictions. The calculations are based on a distribution of SAPS III admission scores between 25 and 110. The calibration for northern Europe is used. ICU=intensive care unit. dept=department. NYHA=New York Heart Association. SAPS=Simplified Acute Physiology Score. *Binary feature. †Combined kidney and pancreas, or other transplantation.

negative effect of having low minimum SBP can be countered by high maximum SBP within the same hour.

We additionally made a decision curve analysis to quantify the potential benefit of guiding treatment based on predictions from our model (appendix p 5).[34]

Discussion

In this study, we developed a risk prediction model providing dynamic, individual predictions of 90-day mortality of ICU patients. The model was trained on 44 binary and continuous SAPS III variables for more than 9000 patients hospitalised in four ICUs in the Capital Region of Denmark between 2011 and 2016. The model was updated at 1-h intervals and calibrated for more reliable predictions. Model performance increased over time and achieved a performance of AUROC of 0·88 (95% CI 0·87–0·89) and MCC of 0·57 (0·54–0·60) at time of discharge. The model was made explainable and the top features driving mortality prediction were identified both for an individual and the full holdout test dataset. Importantly, in the analysis of individual mortality predictions over time, we found that one feature can drive the prediction towards survival at one timepoint and towards non-survival at another. The predicted outcome varies: at time of admission, it encompasses in-ICU, in-hospital, and post-discharge mortality. Patients who die while at the ICU are probably quite different from patients who are discharged and die at home before the 90-day

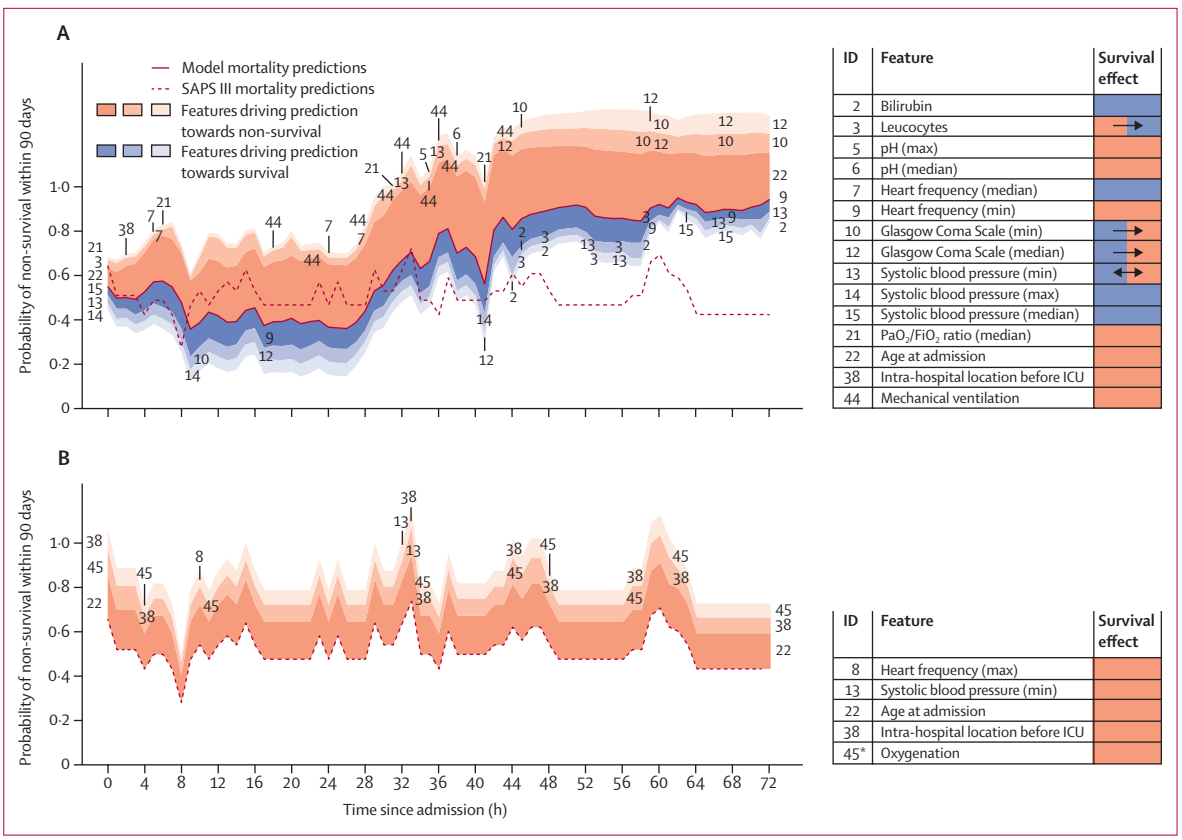


Figure 5: **Impact of input features on the dynamic mortality prediction for a single patient using our model (A) and the original SAPS III score (B) for the first 3 days of ICU admission**

(A) Lines show the mortality predictions from our model and the SAPS III score as they evolve over time (ie, higher values indicating higher mortality risk while lower values indicate higher chance of survival). The shaded areas show the three most important input features driving the 90-day mortality prediction towards either non-survival (orange) or survival (blue) during the first 72 h of the ICU admission. High opacity reflects high relative feature importance. Numbers are used to identify the features; labels are added whenever a feature is outranked by another. (B) The mortality prediction using the SAPS III model with the three most important features driving the prediction towards non-survival. For the depicted patient, there are no features pulling in the direction of survival. ICU=intensive care unit. SAPS=Simplified Acute Physiology Score. *In the machine learning model, the SAPS III oxygenation variable is split into its sub-components of mechanical ventilation and PaO_2/FiO_2 ratio.

mark. Thus, the model adapts to account for the changing nature of the predicted outcome, making it more useful than one-off scores such as SAPS that are computed only once with data obtained during the first day of ICU admission. This finding underpins the importance of continuously updating decision support tools, which adapt to the evolving clinical picture and provide real-time guidance to clinicians. Such dynamic tools are likely to be more useful than the static scores that are currently implemented. Clinical decision making in the early stages of admission—eg, whether to commence treatment and how aggressively to treat a patient—is very different from decisions made later on, such as a decision whether to withdraw life-sustaining treatment.

We also found that certain features could compensate for one another. An example was the negative effect of having a low minimum SBP could be countered by a high maximum SBP within the same hour. Thus, the occurrence of low SBP values has less impact if the condition is correctable, which makes sense from a clinical point of view. Overall, we see that features have complex interactions over time and the ambiguity of features further emphasises the need for real-time machine learning-based decision support. We note that the features interact in a complex non-linear manner, unlike two-way or three-way interactions often used in generalised linear models. The LSTM architecture allows us to model complex, multidimensional interactions but this also makes clinical interpretation of the results more difficult and thus requires caution.

Some input features did not have the anticipated impact on the predictions. For instance, the comorbidities of metastatic cancer and AIDS did not alter predictions much; however, when they did, they often pulled the prediction towards survival. The reason for this counter-intuitive association might be that the model was unable to learn the true importance of the conditions due to the small prevalence in the training dataset. Another possible explanation is that patients with these conditions die early in their ICU stay due to physiological derangements that

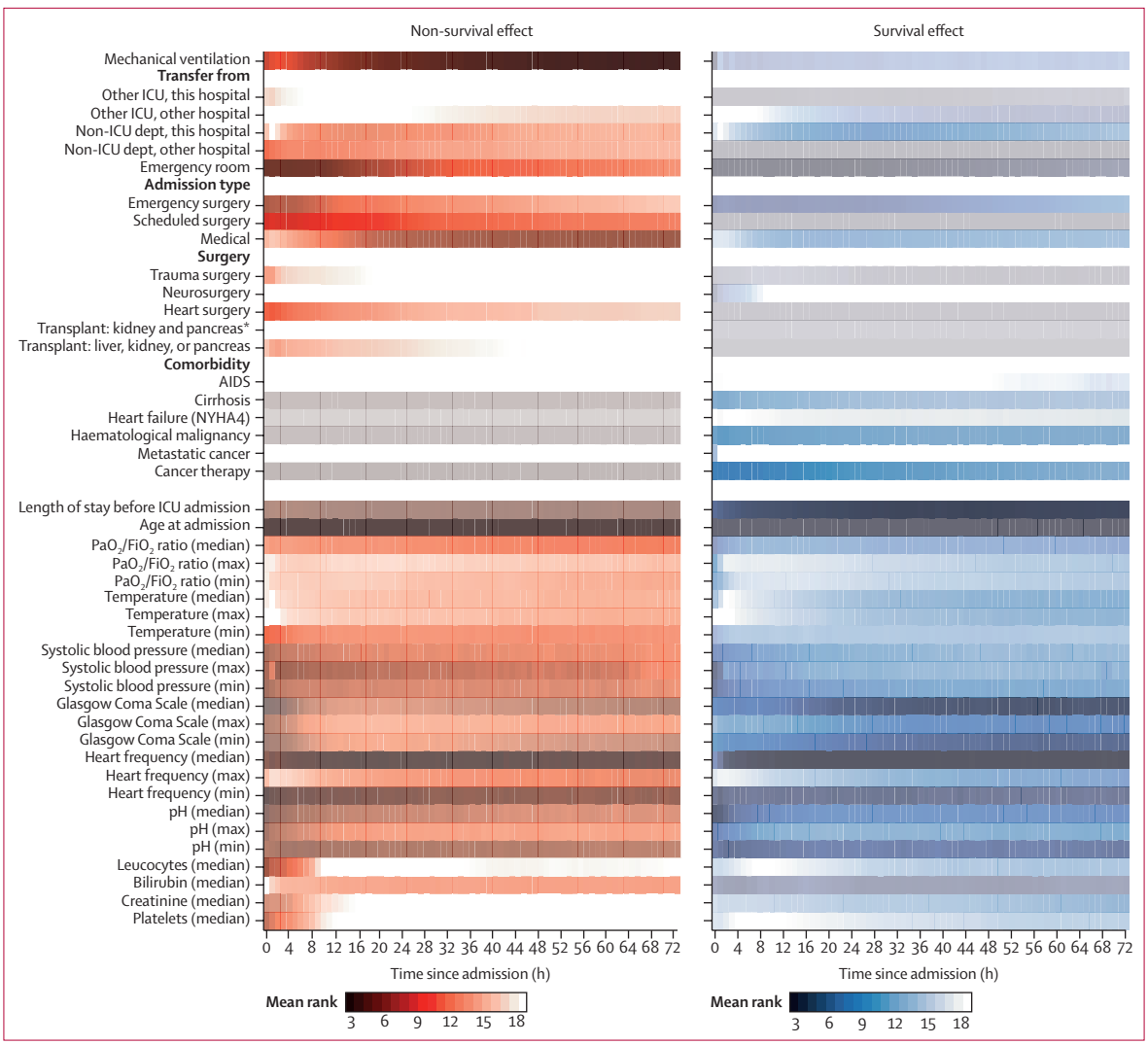


Figure 6: **Impact of input features on the dynamic mortality prediction for the full holdout test dataset population (2825 admissions) for the first 3 days of ICU admission**

The relative importance of all 44 features on the mortality predictions for the full holdout test dataset during the first 72 h of ICU admission. The left column shows contributions driving predictions towards non-survival whereas the right column shows those driving towards survival. ICU=intensive care unit. dept=department. NYHA=New York Heart Association. SAPS=Simplified Acute Physiology Score. *Combined kidney and pancreas, or other transplantation.

cannot be mitigated and thus the comorbidities become less discriminating. Furthermore, patients with metastatic cancer or AIDS are, to some extent, selectively triaged to the ICU—ie, only the younger and non-terminally ill patients will be admitted. Another example of features not having the anticipated impact is seen in figure 6. An almost undetectable SBP (near 0 mm Hg) is only modestly associated with increased mortality and an SBP as high as 300 mm Hg appears to have a beneficial effect. Again, a plausible explanation might be that the model is unable to learn the importance of the extreme values due to a low prevalence in the training dataset. Additionally, the extreme SBP measurements are to some extent artifactual in the clinical setting and likely to have occurred due to flushing or occlusion of the arterial line. The model thus learns to moderate the impact of the extreme values.

The choice of method reflects the nature of a dynamic patient-level prediction problem from the perspective of the clinician: a patient's mortality risk is constantly evolving and depends on the past as well as the current condition. An LSTM network is a special kind of recurrent neural network composed of LSTM units capturing long-range dependencies from sequential data via gated cells, determining whether or not to maintain information based on the importance it assigns to the information. This way, an LSTM-based machine learning prediction model—unlike, for example, logistic regression models—both learns from information about the temporal development and the interaction between the features.

Because we are missing three of the SAPS III variables, and because we chose 90-day overall mortality as the outcome measure, we are not able to do a direct comparison with SAPS III. However, in the original SAPS III study, Moreno and colleagues found that in a cohort from northern Europe, the SAPS III model had an AUROC of 0·814.[15] External validation studies have revealed AUROCs of 0·69 (95% CI 0·63–0·75) and 0·81 (0·79–0·93) in Denmark and Norway, respectively.[35,36] Our model had a higher predictive performance compared with these results, which we confirmed using an external dataset obtained after the study was completed.

Previous studies have applied similar methods to accomplish real-time predictions in an ICU setting. In a recent study, Meyer and colleagues described a recurrent neural network-based model for real-time prediction of bleeding, renal failure, and mortality in a cohort of cardiac surgery patients.[37] As in our study, they base their model on routinely collected data. However, they only report the discriminative performance of the model, not the calibration. This is an important issue if the model is intended for making predictions for single patients. Furthermore, the matter of model explainability is not addressed.

Meiring and colleagues showed that ICU prognostication can be improved by applying a dynamic approach accounting for changes in physiological parameters over the course of several days.[38] In contrast to our study, they report that the best performance is accomplished around 2 days into the ICU admission. The reason for this discrepancy might be that they only use daily measurements for each feature and their neural network architecture is not well suited for dealing with time-series data. Hence, the full information hidden in temporal trends in the data was not exploited.

The European General Data Protection Regulation of 2018 communicates concerns with black-box predictions. It states that individuals have the right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing" when automated decision making is used.[39,40] There are several examples of dubious conclusions drawn from automated decision models based on machine learning.[41] An advantage with our model is that SHAP made our model explainable both in terms of the importance of individual features for ICU patient survival in general and those at patient level at any given timepoint. Thus, using such a model for decision support gives the clinician real-time information of the patient's risk of dying and the specific features currently pulling towards non-survival. The importance of a continuously updated mortality prediction is shown by our finding that one feature can drive predictions towards either survival or non-survival depending on the timepoint of prediction during the ICU stay.

As in all secondary uses of health-care data, we had some missing data. We used LOCF to impute missing data, knowing that this approach is usually not a reputable imputation method.[42] LOCF can be problematic in at least two ways: it can distort temporal covariate tendencies causing misclassification of exposures and bias in unpredictable ways, and it can introduce statistically dependent replicates. We would argue that it makes sense in our case, because the absence of a datapoint is not necessarily void of information. Indeed, much can be inferred when it comes to measurements, especially in a setting as controlled as that of the ICU: the very absence of, for example, a pH value might simply mean that the physician actively chose not to run the analysis again because there was no need. This is arguably often the reality, so carrying the most recent values forward as proxies for missing values might be clinically meaningful. Along this line, artificial missingness was introduced because of the up-sampling of variables not measured every hour. In this case, the use of LOCF is just an imitation of the reasoning of a medical professional, who would derive a clinical assessment on the basis of the available knowledge. The use of LOCF yields replicates that are not statistically independent and could be considered a form of pseudo-replication. This could affect model performance and generalisability, but through cross-validation and regularisation during training, we expect this to have little real effect.

Our study is retrospective and based on ICU data from a densely populated, but rather small geographical area in Denmark. Hence, the model might be biased and reflect the clinical guidelines and treatment decisions made in this area. However, using data from fairly homogeneous settings can render the model useful for exactly the kinds of patients that physicians encounter in their daily work. Besides, the physicians who recorded the data, ordered the tests, and intervened when they saw fit did this with the aim of providing the best possible patient care and did not consider that the data might subsequently be used for prediction purposes. In this way, although retrospective, the data are unlikely to reflect information bias that is otherwise known to haunt prospective studies.

During training, the model was optimised to predict the chance of survival 90 days from ICU admission given data series with a fixed length. The trained model is able to make predictions on new data of varying lengths between admission and discharge, but it is not very accurate at the point of ICU admission. At this point, there are few datapoints available for model training and LSTM models are explicitly optimised for time-series prediction. We acknowledge that an ideal model should have a better performance at early stages as well, but it would not be possible to extract information about the temporal trends, and the model would then serve another purpose than that of this model and would require a different design or a combination of methods.

The present model is based on relatively few variables taken from SAPS III and, as such, the study is intended

as a proof of concept. However, replication in a large validation dataset obtained from another geographical region could verify its robustness and confirms that it is ready to be turned into a clinical decision support tool to be tested in a randomised controlled trial. We intend to do future studies using this model, and are working on implementing the model into our new electronic medical record system (Epic; Verona, WI, USA).

A recent study successfully combined 10 years of disease history before ICU admission with measures from the first 24 h of ICU stay to predict mortality.[10] Our model provides a more accurate prediction of mortality, probably due to the high granularity of the ICU data included in our LSTM model. Thus, adding more detailed disease history to the present model might increase the performance even further and increase the performance upon ICU admission. Additionally, much information is hidden in the clinical notes in which physicians and other health-care professionals collect detailed phenotypic data. Adding such data might also improve the predictive ability.

Many machine learning methods are still opaque, and we have made progress using SHAP to open up and gauge what drives predictions. SHAP values, however, cannot resolve algorithmic bias should such prevail. Algorithmic bias is a genuine concern in the context of machine learning prediction models and comes about because these models have no underlying causal structure: they make predictions entirely on the basis of what humans have done before. This lack of a causal structure also means that prediction models can perform suboptimally when applied to minority populations because the algorithm has only seen few such patients. Thus, albeit explainable, our model is not necessarily fully actionable: age, for example, was the most important feature, but cannot be manipulated by the clinician. Furthermore, we cannot know if clinicians will act and if this action—eg, further correction of low blood pressure—will change the outcome even though low blood pressure strongly influences predictions. During training, the model learned a lot about correlations but nothing about causality. However, these new insights into complex feature interactions might guide our search for causal relations. To achieve actionable models, we would need to build the statistical model on a causal model of how physiological factors interact and react to interventions. Causal models reflect our best guess for the data-generating process, and allow for counterfactual reasoning;[43] this notion is not new but is yet to converge with powerful machine learning methods. Because the ICU is a fairly controlled environment with many objective measurements available, it could be an interesting setting for combining these two disciplines. We gauge model performance by several different measures—eg, AUROC and MCC—but none of these measures encapsulate if a model prediction will result in a favourable change in patient care and outcome.[44] The next step in the process of establishing clinically applicable machine learning models is randomised clinical trials.

In conclusion, we developed an explainable LSTM model for ICU 90-day mortality prediction from a total dataset of more than 14 000 admissions of 11 000 patients from four mixed ICUs in Copenhagen, Denmark, with external validation. The predictive performance improved over the timecourse of an ICU stay. Model interpretation showed that input features can interact and compensate for one another and can pull towards survival at one timepoint and towards non-survival at another. None of these observations can be obtained from current static prognostic scores. Yet, before this kind of model can be used as a bedside tool, the results need to be confirmed in a randomised clinical trial.

Contributors

SB and AP conceived the study, which was designed in detail by H-CT-M. H-CT-M did the data analysis, which was interpreted by H-CT-M, ABN, APN, BSK-H, PT, JS, TS, KB, SB, and AP. PJC, MH, LD, LS, TS, and PH extracted and handled the data. All authors contributed to the preparation of the Article and approved the final version.

Declaration of interests

LD, LS, and PH are employed by Daintel (as of Jan 1, 2020, Cambio has purchased Daintel), which is taking part in the BigTempHealth project funded by the Danish Innovation Fund, grant 5184–00102B. AP reports grants from Ferring and the Novo Nordisk Foundation. SB reports personal fees from Intomics and Proscion, as well as grants from the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001). All other authors declare no competing interests.

Data sharing

Data are available for use in secure, dedicated environments via application to the Danish Patient Safety Authority and the Danish Health Data Authority. Python code used for training of neural networks is available in the appendix (pp 10–11).

Acknowledgments

This study was funded by the Novo Nordisk Foundation (grant agreement NNF14CC0001) and the Innovation Fund Denmark (grant agreement 5153–00002B). We thank the staff at Rigshospitalet, Bispebjerg Hospital, Hvidovre Hospital, Herlev Hospital, and Odense University Hospital for their role in data acquisition.

References

1 Glance LG, Osler TM, Dick AW. Identifying quality outliers in a large, multiple-institution database by using customized versions of the Simplified Acute Physiology Score II and the Mortality Probability Model II0. *Crit Care Med* 2002; **30:** 1995–2002.

2 Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; **100:** 1619–36.

3 Vincent JL, de Mendonça A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on 'sepsis-related problems' of the European Society of Intensive Care Medicine. *Crit Care Med* 1998; **26:** 1793–800.

4 Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; **270:** 2478–86.

5 Zimmerman JE, Draper EA, Wagner DP. Comparing ICU populations: background and current methods. In: Sibbald WJ, Bion JF, eds. Evaluating critical care. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 121–39.

6 Salluh JIF, Soares M. ICU severity of illness scores. *Curr Opin Crit Care* 2014; **20:** 557–65.

7 Kahneman D, Lovallo D, Sibony O. Before you make that big decision. *Harv Bus Rev* 2011; **89:** 50–60,137.

8 Johnson AEWW, Ghassemi MM, Nemati S, Niehaus KE, Clifton D, Clifford GD. Machine learning and decision support in critical care. *Proc IEEE* 2016; **104:** 444–66.

9 Aczon M, Ledbetter D, Ho L, et al. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. *arXive* 2017; published online Jan 23. arXiv:1701.06675 (preprint).

10 Nielsen AB, Thorsen-Meyer H-C, Belling K, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *Lancet Digital Health* 2019; **1:** e78–89.

11 Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *Adv Neur In* 2017; **1:** 4765–74.

12 Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?' In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM 2016: 1135–44.

13 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162:** W1–73.

14 Sandegaard JL, Schmidt SAJ, Sørensen HT, Pedersen L, Ehrenstein V, Schmidt M. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015; **7:** 449.

15 Moreno RP, Metnitz PGH, Almeida E, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; **31:** 1345–55.

16 Guyon I. An introduction to variable and feature selection. *J Mach Learn Res* 2003; **3:** 1157–82.

17 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; **9:** 1735–80.

18 Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; **11:** 2079–107.

19 Chollet F. Deep learning with Python. Shelter Island, NY: Mannning Publications, 2017.

20 Haibo He, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009; **21:** 1263–84.

21 Baldi P, Brunak S. Bioinformatics—the machine learning approach. Cambridge, MA: MIT Press, 2001.

22 Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2:** 749–60.

23 Shapley LS. A value for n-person games. In: Roth AE, ed. The Shapley value. Cambridge: Cambridge University Press, 1988: 31–41.

24 Metnitz PGH, Moreno RP, Almeida E, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intensive Care Med* 2005; **31:** 1336–44.

25 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; **405:** 442–51.

26 Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* 2017; **12**: e0177678.

27 McGee S. Simplifying likelihood ratios. *J Gen Intern Med* 2002; **17:** 647–50.

28 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014; **33:** 517–35.

29 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models. *Epidemiology* 2009; **21:** 128–38.

30 Steyerberg EW. Clinical prediction models. New York, NY: Springer New York, 2009.

31 Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Informatics Assoc* 2018; **25:** 969–75.

32 Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, eds. Advances in large margin classifiers. Cambridge, MA: MIT Press, 1999: 61–74.

33 Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *arXiv* 2017; published online June 14. arXiv:1706.04599 (preprint).

34 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26:** 565–74.

35 Christensen S, Johansen MB, Christiansen CF, Jensen R, Lemeshow S. Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care. *Clin Epidemiol* 2011; **3:** 203–11.

36 Strand K, Søreide E, Aardal S, Flaatten H. A comparison of SAPS II and SAPS 3 in a Norwegian intensive care unit population. *Acta Anaesthesiol Scand* 2009; **53:** 595–600.

37 Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018; **6:** 905–14.

38 Meiring C, Dixit A, Harris S, et al. Optimal intensive care outcome prediction over time using machine learning. *PLoS One* 2018; **13:** e0206862.

39 EU. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off J Eur Communities* 2016; **59:** 1–88.

40 Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Heal Aff Anal Heal Care* 2014; **33:** 1139–47.

41 O'Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy. New York, NY, USA: Crown Publishing Group, 2016.

42 Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials* 2016; **13:** 161–68.

43 Pearl J. Causality. Cambridge: Cambridge University Press, 2009.

44 Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. *JAMA* 2019; published online Aug 8. DOI:10.1001/jama.2019.10306.