THELANCETRM-D-18-00076R3
S2213-2600(18)30300-X
Embargo: September 28, 2018—23:30 (BST)
Doctopic: Primary Research

Machine learning for real-time prediction of complications in critical care: a retrospective study

Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, Carsten Eickhoff

Summary

Background The large amount of clinical signals in intensive care units can easily overwhelm health-care personnel and can lead to treatment delays, suboptimal care, or clinical errors. The aim of this study was to apply deep machine learning methods to predict severe complications during critical care in real time after cardiothoracic surgery.

Methods We used deep learning methods (recurrent neural networks) to predict several severe complications (mortality, renal failure with a need for renal replacement therapy, and postoperative bleeding leading to operative revision) in post cardiosurgical care in real time. Adult patients who underwent major open heart surgery from Jan 1, 2000, to Dec 31, 2016, in a German tertiary care centre for cardiovascular diseases formed the main derivation dataset. We measured the accuracy and timeliness of the deep learning model's forecasts and compared predictive quality to that of established standard-of-care clinical reference tools (clinical rule for postoperative bleeding, Simplified Acute Physiology Score II for mortality, and the Kidney Disease: Improving Global Outcomes staging criteria for acute renal failure) using positive predictive value (PPV), negative predictive value, sensitivity, specificity, area under the curve (AUC), and the F_1 measure (which computes a harmonic mean of sensitivity and PPV). Results were externally retrospectively validated with 5898 cases from the published MIMIC-III dataset.

Findings Of 47 559 intensive care admissions (corresponding to 42 007 patients), we included 11 492 (corresponding to 9269 patients). The deep learning models yielded accurate predictions with the following PPV and sensitivity scores: PPV 0·90 and sensitivity 0·85 for mortality, 0·87 and 0·94 for renal failure, and 0·84 and 0·74 for bleeding. The predictions significantly outperformed the standard clinical reference tools, improving the absolute complication prediction AUC by 0·29 (95% CI 0·23–0·35) for bleeding, by 0·24 (0·19–0·29) for mortality, and by 0·24 (0·13–0·35) for renal failure (p<0·0001 for all three analyses). The deep learning methods showed accurate predictions immediately after patient admission to the intensive care unit. We also observed an increase in performance in our validation cohort when the machine learning approach was tested against clinical reference tools, with absolute improvements in AUC of 0·09 (95% CI 0·03–0·15; p=0·0026) for bleeding, of 0·18 (0·07–0·29; p=0·0013) for mortality, and of 0·25 (0·18–0·32; p<0·0001) for renal failure.

Interpretation The observed improvements in prediction for all three investigated clinical outcomes have the potential to improve critical care. These findings are noteworthy in that they use routinely collected clinical data exclusively, without the need for any manual processing. The deep machine learning method showed AUC scores that significantly surpass those of clinical reference tools, especially soon after admission. Taken together, these properties are encouraging for prospective deployment in critical care settings to direct the staff's attention towards patients who are most at risk.

Funding No specific funding.

Lancet Respir Med 2018

Published Online
September 28, 2018
http://dx.doi.org/10.1016/
S2213-2600(18)30300-X

Department of Cardiothoracic and Vascular Surgery, Deutsches Herzzentrum Berlin, Berlin, Germany (A Meyer MD, D Zverinski MSc, B Pfahringer MSc, J Kempfert MD, S H Sündermann MD, Prof C Stamm MD, Prof V Falk MD)**; Institute of Imaging Science and Computational Modelling** (Prof T Kuehne MD), **Department of Cardiovascular Surgery** (S H Sündermann, Prof V Falk)**, and Berlin Center for Regenerative Therapies** (Prof C Stamm)**, Charité – Universitätsmedizin Berlin, Berlin, Germany; DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany** (A Meyer, Prof C Stamm, Prof V Falk)**; Department of Computer Science, ETH Zurich, Zurich, Switzerland** (D Zverinski, Prof T Hofmann PhD, C Eickhoff PhD)**; Berlin Institute of Health, Berlin, Germany** (A Meyer, B Pfahringer)**; and Center for Biomedical Informatics, Brown University, Providence, RI, USA** (C Eickhoff)

Correspondence to:
Dr Alexander Meyer, Department of Cardiothoracic and Vascular Surgery, German Heart Center Berlin, 13353 Berlin, Germany
meyera@dhzb.de

Introduction

Machine learning is the study and development of systems that can learn from and make predictions on data without the need to be explicitly programmed, and is particularly useful in settings where signals and data are produced at a faster rate than the human brain can interpret. Intensive care treatment is highly challenging for care teams and generates massive amounts of data, and is therefore an optimal target for applying machine learning techniques with the goal of supporting clinical decision making.

Despite frequent reviews and editorials concerning the potential revolutionary impact of machine learning in medicine,[1–11] translation to practical solutions for critical care patients' benefit is non-existent. Translating machine learning approaches to clinical practice is challenging for several reasons. First, some machine learning methods, such as reinforcement learning,[12] require prospective interaction with patients. In the early learning stages, this could mean a dramatically increased risk of adverse events. Second, data recording in electronic health record (EHR) systems is designed and optimised for reporting, liability, and billing purposes rather than informing clinical intelligence systems.[3] Third, data are often organised and stored across a variety of systems, requiring integration and harmonisation before being used in

Research in context

Evidence before this study

Artificial intelligence-augmented care is an emerging field. Consequently, the existing literature is relatively sparse. We searched MEDLINE and arXiv for the term ((“real-time prediction”) OR (“deep learning”) OR (“real-time scoring”) OR (“machine learning”) OR (“artificial intelligence”)) AND ((intensive OR critical) care) with no language restrictions or date limitations. We retrieved 510 MEDLINE results and 252 arXiv results, 72 of which were relevant original studies. The relevant prior evidence included 18 articles investigating real-time prediction approaches. None of these articles used a deep learning methodology. Most of the articles described the prediction of sepsis and mortality, using often curated or open datasets such as the MIMIC-III dataset. All studies described a specific approach predicting a single outcome. At the time of writing, prediction of sepsis in real time is the topic with most available evidence.

Added value of this study

We developed deep learning models to predict severe complications following cardiothoracic surgery. These models used uncurated clinical datasets to predict three endpoints. By contrast with standard clinical risk scores, our approach was not based on the average patient but used cohort data to inform predictions. This approach yields higher accuracy for each individual patient. The selected clinical variables reflect the range of routinely collected information at intensive care units for all postoperative patients, removing the need for any additional manual data collection or annotation. The deep learning methods we implemented achieved superior predictive power and timeliness compared with three standard-of-care baselines.

Implications of all the available evidence

A real-time complication prediction system based on deep learning outperforms the selected standard-of-care baselines in timeliness and accuracy, even when acting on a real, uncurated data stream. We are currently deploying our system in our intensive care unit and will do a trial to confirm the results prospectively to enable its use in the clinical routine.

automated reasoning. Finally, patient data recorded in clinical information systems (such as vitals monitoring, laboratory values, and medications) are prone to having missing values, heterogeneity, errors, and artifacts, potentially introducing significant levels of noise to the decision process.

Existing clinical applications remain largely academic in nature and model patient outcomes such as mortality on the basis of synthetic, manually curated, or heavily distorted datasets[13–15] that often do not reflect the whole dimension of signals and complexity faced in modern critical care environments.

In this work, we investigate the use of deep learning techniques in postoperative cardiac surgery care in a real-world setting. In a retrospective study including uncurated intensive care cases, we assess the merit of a predictive machine learning approach to increase quality of care and patient safety. We support our findings with a retrospective validation study on intensive care cases from another intensive care unit.

Methods

Datasets

See Online for appendix

We analysed electronic health record data from a German tertiary care centre for cardiovascular diseases (German Heart Center Berlin) of adult patients (≥18 years of age at the time of surgery) who underwent major open heart surgery from Jan 1, 2000, until Dec 31, 2016. We included coronary artery bypass grafting, valve surgery, aortic surgery, assist device surgery, pericardial surgery, and heart and lung transplantations. All patients who had catheter-based interventions were excluded except for transcatheter aortic valve implantations. Surgical re-exploration due to postoperative bleeding, postoperative renal failure requiring renal replacement therapy, and postoperative in-hospital mortality were defined as relevant outcomes; we therefore labelled patients as “complication occurred” or “complication did not occur”. To obtain a balanced dataset, for each of these groups an equally sized control group of patients who did not experience the relevant complication was generated. These controls were sampled uniformly at random from all cases of adult patients with the same initial surgery selection criteria.

Patient datasets were analysed for the first 24 h after the initial surgery. All cases of bleeding, mortality, and renal failure that occurred after the initial 24 h were labelled as “complication occurred”, but were analysed as though the outcome occurred 24 h after the initial surgery.

This study was approved by the institutional data protection officer and ethics committee (EA2/180/17).

Model development

A recurrent deep neural network (RNN) was used that models the likelihood that a patient might experience future complications on the basis of patient-specific attributes (so-called features; for details see appendix pp 4–5). We chose objectively collected markers such as patient demographics, vital parameters, coagulation tests, and bleeding rate to be the features for our model. Some of these features are static and not subject to change during hospitalisation (eg, age or sex) whereas others might be susceptible to substantial changes over time (eg, blood pressure). We included all such features that were documented for at least 50% of patients in our model. Missing values were imputed with the patient’s last measured value for that feature (see appendix p 3).

Measurements of dynamic clinical markers were collected at 30-min intervals and used as input to a gated recurrent unit network. Static patient variables were replicated at each timepoint.[16] The network computes an internal state s_t at each timepoint t, based on input vector $\mathbf{x}_t$ of patient features; as time progresses, these states are updated as new inputs $\mathbf{x}_t$ are read (figure 1; see appendix pp 1–3 for further details). At each timepoint t, the likelihood o_t of a specific complication (eg, renal failure) occurring given the patient's current situation was computed and evaluated against the true class label of whether or not the patient would eventually develop the relevant complication (ie, the observed outcome). Network parameters U, V, and W (figure 1) were initialised randomly and were successively refined as the model learned to recognise the various clinical complications from historic episodes, the so-called training data.

The performance of the trained network was assessed by a two-fold approach. First, the network was used to predict the observed outcomes in a separate dataset from the training data, the test set. This provided an unbiased estimate of model performance. As comparators, we measured the performance of three clinical reference tools (see appendix pp 5–6): the standard method for recognising postoperative bleeding requiring operative revision is

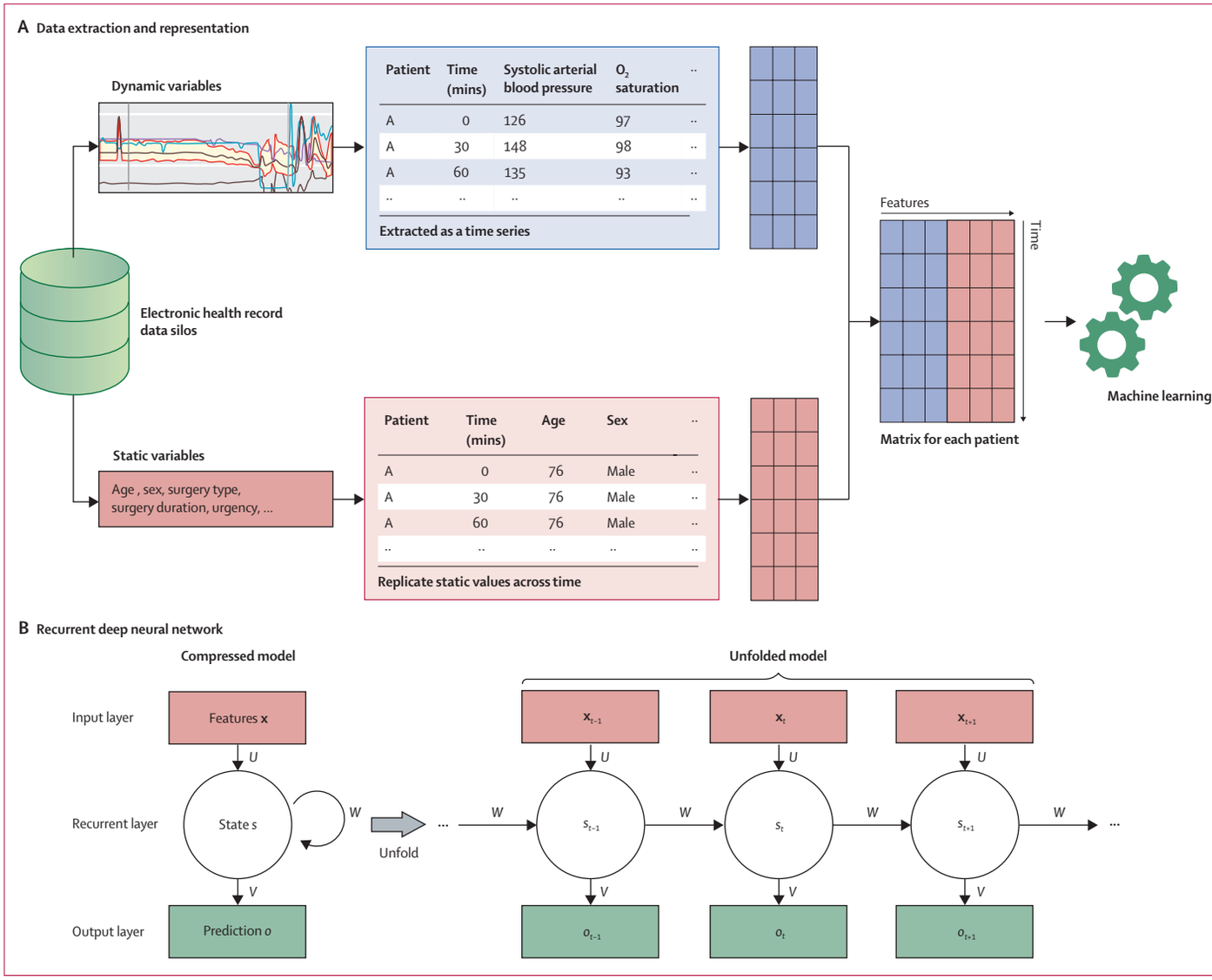


Figure 1: **Architectural overview of data extraction and representation (A) and a schematic illustration of a recurrent deep neural network (B)**
(A) Dynamic variables are extracted as a time series, whereas static variables are replicated across time. Using these values, a matrix consisting of all features and labels is generated for each patient, representing the individual clinical episode over time. (B) The recurrent deep neural network is comprised of an input layer of features **x**, a simple recurrent layer of states *s*, and an output layer of predictions *o*. *U*, *V*, and *W* are trainable parameters of the neural network represented as matrices of real-valued numbers. *U* maps raw patient data to an internal model state. *W* further updates and modifies this state by information from the previous time steps. Finally, *V* translates the internal state into the most likely outcome to predict.

given by an algorithm specific to post-cardiothoracic care formalised by Bojar;[17] renal failure requiring renal replacement therapy was identified via the retrospective Kidney Disease: Improving Global Outcomes stage III calculation;[18] and patient mortality was predicted using the Simplified Acute Physiology Score II (SAPS II).[19] None of these clinical reference tools had been developed for this exact task of predicting these outcomes; we accounted for this limitation by modifying the tools to give an optimal estimate of the comparative metric (appendix pp 5–6). Finally, aside from the global performance evaluation (ie, irrespective of the time elapsed since the original surgery), the model and baselines were evaluated as time passed since surgery. This experiment captures ability of these various approaches to correctly indicate clinical outcomes as early as possible.

The model was developed under a ten-fold cross-validation scheme, using 90% of the data for training and validation. The remaining 10% was reserved for performance measurement and the reported scores were based on the results of this test set.

	Features
Patient information (four features)	Age, sex, height, weight
Information relating to initial surgery (nine features)	Anaesthesia type, American Society of Anesthesiologists Score, cardioplegic solution, aortic cross-clamp time, cardiopulmonary bypass time, anaesthetic monitoring time, surgery duration, surgery type, urgency
Vital signs (11 features)	Systolic, mean, and diastolic arterial pressure; systolic, mean, and diastolic pulmonary artery pressure; central venous pressure; ventilator FiO_2 setting; heart and respiratory frequency; body temperature
Arterial blood gas (nine features)	Bicarbonate, glucose, haemoglobin, oxygen saturation, partial pressure of carbon dioxide and oxygen, pH level, potassium, sodium
Laboratory results (17 features)	Albumin, bilirubin, urea, C-reactive protein, creatine kinase, γ-glutamyltransferase, glutamic oxaloacetic transaminase, haemoglobin , haematocrit, international normalised ratio, creatinine, white blood cell count, lactate dehydrogenase, magnesium, partial thromboplastin time, platelets, prothrombin time
Balance output (two features)	Bleeding rate, urine flow rate

General patient and surgery information remains static whereas vital signs, blood gas analysis, laboratory results, and output are tracked at regular intervals. FiO_2=fraction of inspired oxygen.

Table 1: **Feature overview**

	Cases (n=47 559)
Patient demographics	
Number of patients	42 007
Age, years	67·6 (59·1–74·7)
Sex	
Female	12 803 (30·5%)
Male	29 204 (69·5%)
Outcome distribution	
Bleeding	2322 (4·9%)
Mortality	2972 (6·2%)
Renal failure	452 (1·0%)
Surgery type distribution	
Coronary artery bypass grafting plus valve surgery	5071 (10·7%)
Aortic surgery	3866 (8·1%)
Ventricular assist device surgery	1671 (3·5%)
Pericardial surgery	1528 (3·2%)
Valve surgery	11 756 (24·7%)
Coronary artery bypass grafting	19 569 (41·1%)
Transplantation	936 (2·0%)
Other	3162 (6·6%)
Surgery characteristics	
Number of emergency surgeries	7650 (16·1%)
Number of surgeries involving cardiopulmonary bypass	38 128 (80·2%)
Surgery time, h	3·8 (2·8–4·8)

Data are n (%) or median (IQR).

Table 2: **Baseline characteristics of the complete dataset**

External validation

We did external retrospective validation using patients from the MIMIC-III dataset, which contains data associated with 53 423 distinct hospital admissions for adult patients admitted to critical care units during 2001–12.[20] We selected a cohort of 5898 patients that had undergone major open heart surgery at one of the centres included in the MIMC-III dataset. Most of the features we used in our development dataset were available in MIMIC-III, with the important exception of detailed information about the type of original surgery performed. Very little was changed in the modelling, but we did add regularisation to the original model to help to avoid overfitting. More details of the procedures and modifications for external validation are listed in the appendix (p 11).

Statistical analysis

We assessed the predictive performance of our model using a range of common performance metrics: positive predictive value (PPV), negative predictive value, sensitivity, specificity, and area under the curve, all of which range from 0 (low performance) to 1 (high performance). In the balanced two-class setting, random guessing would correspond to a score of 0·5. To reflect the frequently observed trade-off between precision (ie, PPV) and sensitivity, we also include the F_1 measure[21] that computes a harmonic mean of both scores, penalising methods for disproportionately favouring either of the two measures over the other. Accuracy is a commonly applied metric from the machine learning domain that is defined as:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives, TN the true negatives, FP the false positives, and FN the false negatives. For each complication, a clinical reference tool was included as a reference point for the quality of the RNN-based predictions (see appendix pp 3–6 for additional

	Bleeding		Mortality		Renal failure	
	Training (n=4180)	Test (n=464)	Training (n=5350)	Test (n=594)	Training (n=813)	Test (n=91)
Cases with event	2090 (50%)	232 (50%)	2675 (50%)	297 (50%)	407 (50%)	45 (49%)
Patient demographics						
Age, years	67 (57–74)	66 (57–74)	66 (57–74)	66 (58–74)	68 (59–74)	68 (60–73)
Sex						
Female	1226 (29%)	133 (29%)	1602 (30%)	167 (28%)	241 (30%)	33 (36%)
Male	2954 (71%)	331 (71%)	3748 (70%)	427 (72%)	572 (70%)	58 (64%)
Surgery type distribution						
Coronary artery bypass grafting plus valve surgery	602 (14%)	67 (14%)	675 (12%)	92 (15%)	115 (14%)	15 (16%)
Aortic surgery	336 (8%)	41 (9%)	502 (9%)	60 (10%)	130 (16%)	18 (20%)
Ventricular assist device surgery	400 (10%)	47 (10%)	600 (11%)	58 (10%)	6 (1%)	3 (3%)
Pericardial surgery	116 (3%)	11 (2%)	236 (4%)	22 (4%)	26 (3%)	1 (1%)
Valve surgery	1044 (25%)	109 (23%)	1160 (22%)	110 (19%)	240 (30%)	27 (30%)
Coronary artery bypass grafting	1501 (36%)	164 (35%)	1745 (33%)	200 (34%)	265 (33%)	22 (24%)
Transplantation	173 (4%)	25 (5%)	262 (5%)	31 (5%)	30 (4%)	5 (5%)
Other	8 (<1%)	0	188 (4%)	21 (4%)	1 (<1%)	0
Surgery characteristics						
Emergency surgeries	1099 (26%)	133 (29%)	1715 (32%)	180 (30%)	219 (27%)	32 (35%)
Surgeries involving cardiopulmonary bypass	3747 (90%)	412 (89%)	4375 (82%)	494 (83%)	709 (87%)	80 (88%)
Surgery time, h	4·2 (3·3–5·7)	4·1 (3·1–5·6)	4·3 (3·2–6·3)	4·3 (3·2–6·6)	4·4 (3·4–5·9)	4·6 (3·4–6·5)

Data are n (%) or median (IQR).

Table 3: **Baseline characteristics across training and test sets for each modelling task**

information on the evaluation process). We used the Wilcoxon signed-rank test to compare the accuracy levels of the clinical reference tool against the RNN-based predictions. A two-tailed p value of less than 0·05 was considered significant. Statistical analyses were done using R version 3.3.0.

Role of the funding source

The funding sources had no involvement in the study design; collection, analysis, and interpretation of data; writing of the report; or the decision to submit the paper for publication. The corresponding author had full access to all the data and final responsibility to submit for publication.

Results

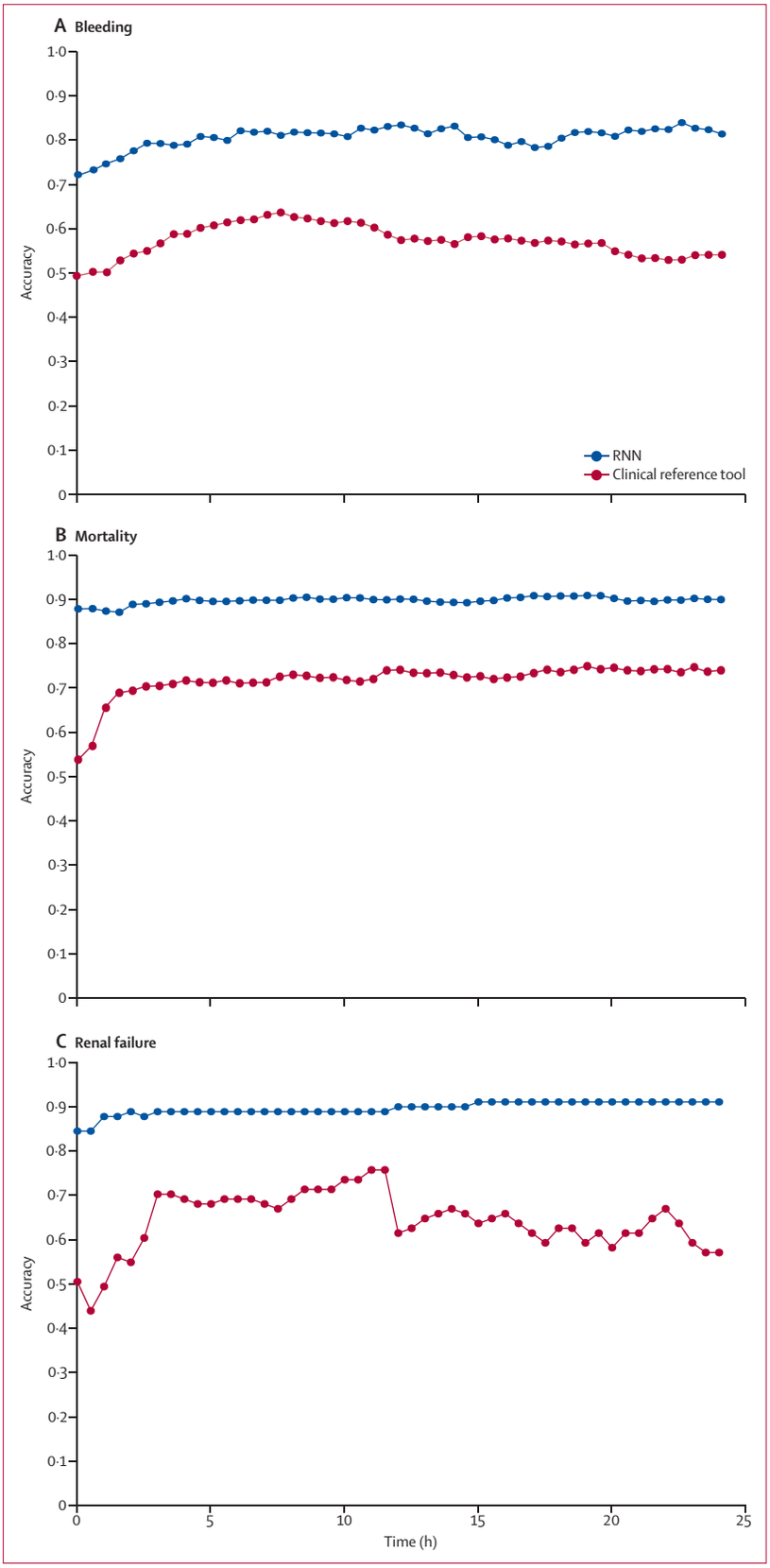
The complete dataset comprised 47 559 intensive care cases, corresponding to 42 007 patients, with information available on 52 patient features (tables 1, 2). In total, we included 11 492 admissions, which corresponded to 9269 patients (table 3).

Overall, when considering the performance scores in the balanced test dataset (table 4), postoperative bleeding seems to be more difficult to predict by either method than mortality or renal failure, for which all compared methods—including RNN—obtained better results overall. For all tasks, the RNN approach provided significantly better accuracy levels than the respective clinical reference tool (absolute accuracy improvement in bleeding 0·22, p=0·0007; mortality 0·17, p=0·0060; renal failure 0·17, p=0·0008), determined by a Wilcoxon signed-rank test. The exception was a lower than reference specificity in bleeding prediction that was accompanied by a more than tripled sensitivity score as compared with the commonly applied Bojar textbook procedure.[17] The RNN model was able to improve absolute complication prediction area under the curve (AUC) by 0·29 (95% CI 0·23–0·35) for bleeding, by 0·24 (0·19–0·29) for mortality, and by 0·24 (0·13–0·35)

	Accuracy	PPV	NPV	Sensitivity	Specificity	F_1	AUC
Bleeding							
RNN	0·80	0·84	0·77	0·74	0·86	0·79	0·87
Bojar	0·58	0·81	0·54	0·21	0·95	0·33	0·58
Mortality							
RNN	0·88	0·90	0·86	0·85	0·91	0·88	0·95
SAPS II	0·71	0·68	0·74	0·78	0·63	0·73	0·71
Renal failure							
RNN	0·90	0·87	0·94	0·94	0·86	0·90	0·96
KDIGO	0·73	0·87	0·67	0·53	0·92	0·66	0·72

The F_1 measure computes a harmonic mean of sensitivity and PPV. PPV=positive predictive value. NPV=negative predictive value. AUC=area under the curve. RNN=recurrent deep neural network. SAPS=Simplified Acute Physiology Score. KDIGO=Kidney Disease: Improving Global Outcomes.

Table 4: **Predictive performance of fully converged models in the balanced test dataset by method used**



for renal failure (p<0·0001 for all three analyses). For further side-by-side comparison, the results of the RNN-based predictions with the sensitivity set to the corresponding clinical reference tool sensitivity as well as the performance metrics on an unbalanced dataset are shown in the appendix (pp 6–8).

We also evaluated both the RNN method and the relevant clinical reference tools in terms of their performance over time, to account for the fact that additional information often helps both human and machine reasoners in the decision-making effort. At time 0 h, immediately after the original surgery had been concluded and the patient had arrived in the intensive care unit, all methods were less accurate due to rapidly changing and unstable initial patient state and missing information (eg, lab results arriving with delay; figure 2). Over time, as additional information arrived, thus stabilising the recognition of each patient's state, both RNN and clinical reference tool performance increased until performance scores were eventually saturated. Depending on the outcome being predicted, this initial prediction time lag differs. Whereas bleeding prediction required several hours of observing the patient to unfold its full potential, mortality and renal failure could be predicted with high accuracy almost immediately after admission to the intensive care unit (figure 2). Notably, in all three settings at all times, even immediately after conclusion of the original surgery, the RNN method was considerably more accurate than the clinical reference tools would ever become during their individual periods of peak performance.

External validation on the MIMIC-III heart surgery cohort confirmed the previous observations (table 5). Predictions of bleeding events with both Bojar's algorithm and RNN were generally less accurate than those targeting mortality or renal failure. For all three predictive endpoints, the RNN predictions resulted in significantly higher accuracies than those originating from the respective clinical reference tools (absolute accuracy improvement in bleeding 0·04, p=0·011; mortality 0·11, p=0·0018; renal failure 0·16, p<0·0001), according to a Wilcoxon signed-rank test. As in the study cohort, bleeding and renal failure prediction of the respective clinical reference tools was higher at the cost of drastically reduced sensitivity (table 5). The RNN method overall improved predictive performance in terms of absolute AUC scores by 0·09 (95% CI 0·03–0·15; p=0·0026) for bleeding, by 0·18 (0·07–0·29; p=0·0013)

Figure 2: **Detailed performance analysis in the test set**
Clinical reference tools were the Bojar algorithm[17] for postoperative bleeding, the Simplified Acute Physiology Score[19] for patient mortality, and the Kidney Disease: Improving Global Outcomes stage III calculation[18] for renal failure. Previous overall performance numbers are broken down according to the time elapsed since the original surgery using all time slices of each included patient in the test set. RNN=recurrent deep neural network.

for mortality, and 0·25 (0·18–0·32; p<0·0001) for renal failure. Unsurprisingly, these improvements, while consistent, were milder than those observed in the study cohort, given the reduced size of the eligible cohort (5898 patients *vs* 47 559 patients) and a reduced set of available features.

In our predictive analyses of the MIMIC-III validation cohort, we observed a similar pattern as in the study cohort (figure 3). Consistently, at each timepoint, the RNN models showed higher accuracy than the respective clinical reference tools, leading to a clear separation of graphs. Again, bleeding prediction with RNN showed the strongest improvement over time whereas mortality and renal failure prediction almost immediately performed at their respective maximum levels of accuracy.

	Accuracy	PPV	NPV	Sensitivity	Specificity	F_1	AUC
Bleeding							
RNN	0·70	0·70	0·70	0·70	0·69	0·70	0·75
Bojar	0·66	0·94	0·59	0·34	0·98	0·50	0·66
Mortality							
RNN	0·74	0·75	0·73	0·73	0·76	0·74	0·81
SAPS II	0·63	0·59	0·74	0·86	0·41	0·70	0·63
Renal failure							
RNN	0·82	0·85	0·80	0·79	0·86	0·82	0·91
KDIGO	0·66	0·88	0·60	0·37	0·95	0·53	0·66

The F_1 measure computes a harmonic mean of sensitivity and PPV .PPV=positive predictive value. NPV=negative predictive value. AUC=area under the curve. RNN=recurrent deep neural network. SAPS=Simplified Acute Physiology Score. KDIGO=Kidney Disease: Improving Global Outcomes.

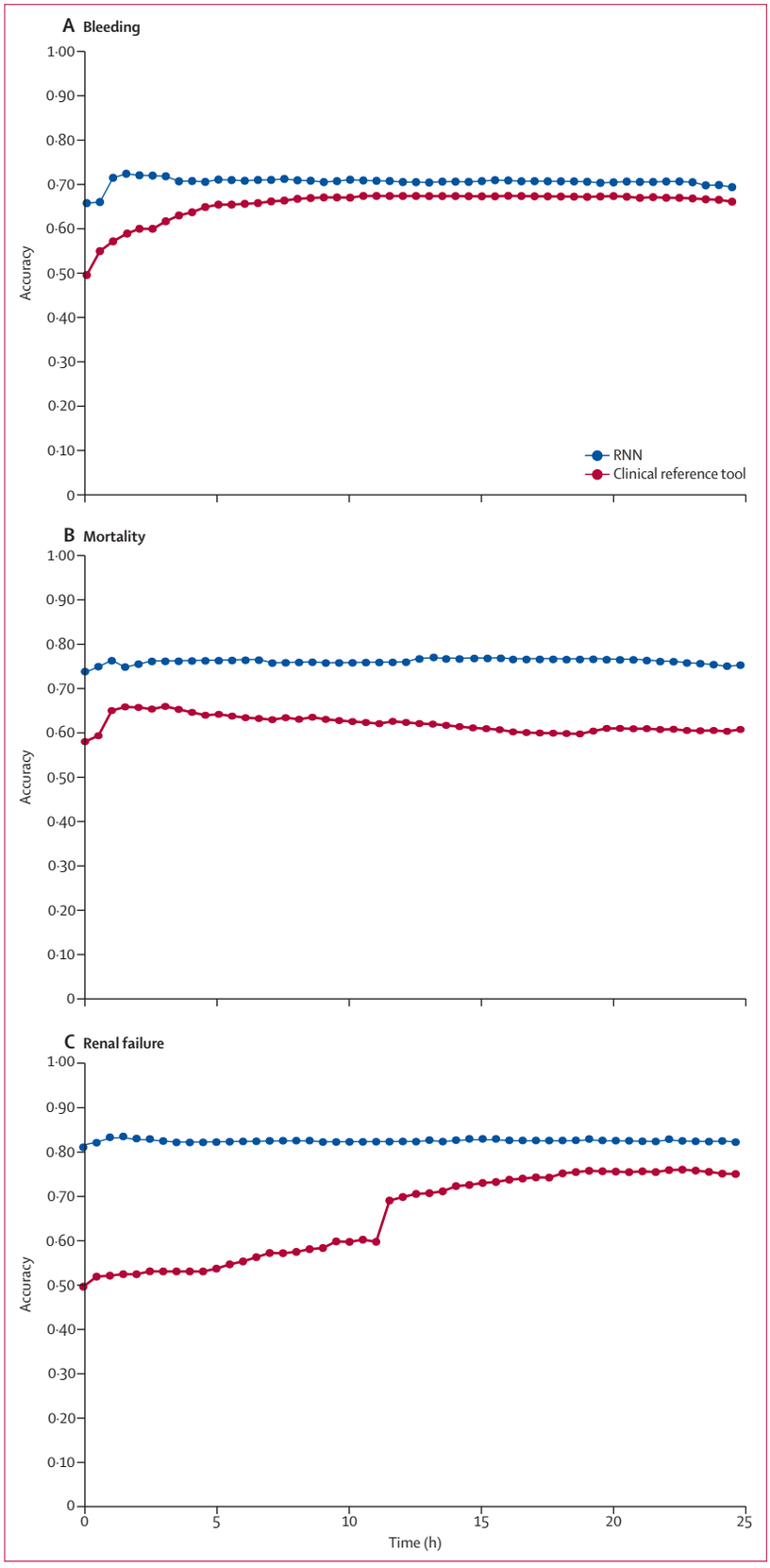
Table 5: **Predictive performance of fully converged models on MIMIC-III validation cohort by method used**

Discussion

A real-time diagnostic and prognostic prediction model based on a machine learning algorithm and routinely collected clinical data during critical care was established and validated. The deep learning models incorporated static and dynamic variables and scrutinised their changes over time. We noted in each modelled outcome a high predictive performance (AUC ≥ 87 for all models) that is not commonly observed in current clinical prognostic models. The proposed approach has several advantages over conventional clinical risk models: instead of a generic cohort, we used a highly specific one—ie, critical care patients post cardiothoracic surgery; we included static and dynamic variables and considered their changes over time; clinical staff did not have to collect extra variables for a prognostic model (such as the well known APACHE II[22] or SAPS II[19] models for prediction of mortality during critical care, which require manual addition of additional data); and we used a recurrent deep learning method that allowed predictions to change over time, on the basis of a patient's development during the time of observation, thus accounting for the evolving and often rapidly changing clinical state and treatment decisions. The developed prediction models are not limited to predicting and hence preventing complications but might also serve the purpose of monitoring the success of interventions used; this, however, will have to be evaluated in another study.

Several authors have claimed that the availability of so-called big data is a prerequisite for successful application of machine learning.[1,4,9] Although there is no exact definition for this term, our study shows that a relatively small cohort of 47 559 historic patient cases, not including any imaging, so-called omics, or textual information, can substantially outperform the current prognostic standard of care. Even during external validation on a substantially smaller dataset, using a heart-surgery-specific subset of the published MIMIC-III dataset,[20] our proposed method continued to outperform the clinical reference tools. In addition, the range of available features in MIMIC-III was smaller than that available in our dedicated study cohort. These two challenges—fewer cases and less information per case—make the external validation results even more notable and stress the robustness of our approach. These observations motivate a new clinical paradigm: each hospital and each clinical unit could leverage their own data to generate knowledge and predictive power for the specific patient cohort treated, resulting in a tailored approach within the specific context of each unit (eg, a stroke unit has different patient cohorts than does a cardiothoracic unit or an emergency room). Even within a specific intensive care unit, striking differences in care practice and patient cohort composition might occur over time. This variation partly explains the difficulty of achieving reproducible results in the critical care environment[23] and stresses the relevance for cohort-specific models. Incorporation of additional data categories such as omics or imaging data is possible and will probably improve the applicability and versatility of deep learning analytics in the clinical setting even further.

The results of this study confirm the view of Obermeyer and colleagues about the importance of applying state-of-the-art algorithms that "learn rules from data" to find patterns of covariates over time.[4,5] An important issue is the handling of potentially biased data. For our model, this was addressed by incorporating only data that were routinely collected and objectively measured during postoperative critical care and excluding any direct traces of human intelligence, such as the ordering of blood transfusions or any pharmacological prescriptions. This ensures that the models are exclusively trained on systemic patient properties rather than people's reactions to them. Verghese and colleagues[11] further stress the challenges when dealing with EHR data and Obermeyer and colleagues[4] emphasise the importance of the curation of EHR data. Conversely, we believe that comprehensive manual, a-priori curation of EHR data would prohibit any real-time use of EHR data-driven systems—ie, to ensure scalable translation into day-to-day clinical care, our proposed method directly processed uncurated data.



Finally, transparency and traceability of the decision-making process of artificial intelligence systems must be made available to physicians. Opening up the often opaque machine-learning process is an active field of ongoing research,[24,25] the importance of which is being increasingly recognised—eg, by the European Union's General Data Protection Regulation that postulates a "right to explanation" for automated individual decision-making systems.[26]

Our study had several limitations. It was a retrospective study: although result quality was considerable on its own, the model's predictive power was further compared with a range of established clinical reference tools; however, none of these tools were the perfect fit for this purpose because they had not been developed for this exact task. To account for this, each standard score was modified to give an optimal estimate of the comparative metric, resulting in a highly conservative view of the performance differences between baselines and deep learning models that is likely to underestimate the comparative benefits of the RNN method (see appendix pp 5–6). Compared with synthetic datasets, the real-world clinical data used here were very noisy—our dataset contained mismatches, contradictions, errors, and omissions. The signal-to-noise ratio for clinical questions (eg, diagnosis or treatment choice) to be answered by patient data can be very low. However, this is not an exclusive characteristic of this particular dataset, but rather a common property of real clinical data and hence unavoidable unless data collection techniques improve; it is therefore important that deep learning models be shown to work successfully with noisy data because this is the reality in which the models will be applied. In the retrospective validation cohort, the same limitations on data quality and baseline performance also hold. Furthermore, the annotation quality of clinical outcomes for bleeding and renal failure as we defined it is limited by the lack of detailed information on the original surgery (eg, the exact procedure done and by whom, and exact finding reports from the surgery). The small size of the validation cohort coupled with low incidence rates of the investigated outcomes—in particular mortality—limit the reliability of the validated results. External validation using prospectively collected data, which remains to be done, could address this limitation.

Intensive care units provide a highly challenging environment that confronts physicians with a demanding case load and require rapid decision making.[27] The handling of a continuous stream of massive amounts of noisy data, such as laboratory results, clinical and physiological measurements, imaging, and—increasingly—

Figure 3: **Detailed performance analysis on MIMIC-III validation cohort**
Clinical reference tools were the Bojar algorithm[17] for postoperative bleeding, the Simplified Acute Physiology Score[19] for patient mortality, and the Kidney Disease: Improving Global Outcomes stage III calculation[18] for renal failure. RNN=recurrent deep neural network.

omics information can easily go beyond the information processing capacity of the human operator (ie, intensive care physicians). Our proposed deep-learning-based and cohort-specific modelling can detect physiological signatures and complex relationships in the data stream before they become apparent to the critical care team. The rational of predicting a future event several hours before its full manifestation is to avoid a delay in treatment; for example, a delayed operative revision of substantial postoperative bleeding leads ultimately to excessive transfusion and haemorrhagic shock. Risks such as these can potentially be avoided by pointing the care team's attention to patients at risk as early as possible.

The resulting models could be directly integrated into EHR systems, thus ensuring all necessary data are accessible in real time for the models and no additional data input is necessary. Predictions could then be calculated automatically by the EHR system at a chosen interval—eg, a new prediction every minute. These predictions could then be stored back in the patient's EHR or made available to the clinician in various ways, such as in a dedicated bedside monitoring device, a computer application, or a tablet or smartphone app.

Given the human limits in parallel data processing capacity,[28] the ever increasing size and diversity of datasets,[29,30] and growing economic constraints,[31,32] machine-learning-based clinical decision support could serve as an enhancement and safety tool. Diagnostic models, in particular, can substantially increase patient safety by reducing time to diagnosis, thereby granting physicians additional time to react and intervene. This time advantage can be crucial because early changes in organ function are strongly related to patient outcomes.[33–36] The expectations towards machine learning systems are high but appear to be justified in light of the results presented here.

In conclusion, real-time, highly accurate scoring based on specific patient cohorts and individual clinical settings with routinely collected data during critical care is feasible and has the potential to augment the physician's decision-making process. Recurrent deep learning models significantly outperform standard clinical baselines. Real-time event prediction holds the potential to reduce the variability in clinical care by overcoming issues of intensivist workload and experience, ultimately resulting in increased patient safety.

Contributors

AM and CE conceived the study and drafted the manuscript. DZ, AM, and CE designed the methods. AM and DZ obtained the data and contributed to study design. JK, TK, CS, SHS, TH, and VF contributed to study design. DZ and BP did the analyses under the supervision of CE and AM. All authors contributed in result interpretation and critically revised the manuscript.

Declaration of interests

AM reports grants from Berlin Institute of Health, during the conduct of the study; and grants from Edwards Lifesciences, Abbott, and Maquet; and grants and other support from Medtronic, outside of the submitted work. AM, DZ, and CE have a patent on using machine learning to predict health conditions (US Provisional Patent Application 62/677,890) pending to AIPHIA Capital Partners. VF reports grants and other support from Abbott, Medtronic, Boston Scientific, and Edwards Lifesciences; and other support from Biotronik, Berlin Heart, and Novartis Pharma, outside of the submitted work. VF is also on the advisory board for Medtronic, Berlin Heart, Novartis Pharma, and Boston Scientific. All other authors declare no competing interests.

Acknowledgments

AM participates in the BIH Charité Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. CE is funded by the Swiss National Science Foundation (SNSF) Ambizione Program (grant agreement number 174025).

References

1 Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017; **69:** 2657–64.

2 Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 2017; **376:** 2507–09.

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

4 Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016; **375:** 1216–19.

5 Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. *N Engl J Med* 2017; **377:** 1209–11.

6 McKenna J. Big data: big promise. *Eur Heart J* 2017; **38:** 470–71.

7 Celi LA, Mark RG, Stone DJ, Montgomery RA. 'Big data' in the intensive care unit. Closing the data loop. *Am J Respir Crit Care Med* 2013; **187:** 1157–60.

8 Maslove DM, Lamontagne F, Marshall JC, Heyland DK. A path to precision in the ICU. *Crit Care* 2017; **21:** 79.

9 Lancet. Artificial intelligence in health care: within touching distance. *Lancet* 2017; **390:** 2739.

10 Topol E. Digital medicine: empowering both patients and clinicians. *Lancet* 2016; **388:** 740–41.

11 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2017; published online Dec 20. DOI:10.1001/jama.2017.19198.

12 Sutton RS, Barto AG. Reinforcement learning: an introduction. *Trends Cogn Sci* 1998; **3:** 360.

13 Grnarova P, Schmidt F, Hyland SL, Eickhoff C. Neural document embeddings for intensive care patient mortality prediction. *arXiv* 2016; published online Dec 1. http://arxiv.org/abs/1612.00467 (preprint).

14 Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. *arXiv* 2016; published online June 6. http://arxiv.org/abs/1606.01865 (preprint).

15 Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with RNNs. *arXiv* 2016; published online June 13. http://arxiv.org/abs/1606.04130 (preprint).

16 Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* 2014; published online Dec 11. http://arxiv.org/abs/1412.3555 (preprint).

17 Bojar RM. Manual of perioperative care in adult cardiac surgery. Hoboken, NJ: Wiley-Blackwell, 2011.

18 Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin Pract* 2012; **120:** c179–84.

19 Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; **270:** 2957–63.

20 Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3:** 160035.

21 Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. In: Losada DE, Fernández-Luna JM, eds. Advances in information retrieval. Berlin: Springer, 2005: 345–59.

22 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; **13:** 818–29.

23 Ospina-Tascón GA, Büchele GL, Vincent J-L. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? *Crit Care Med* 2008; **36:** 1311–22.

24 Che Z, Purushotham S, Khemani RG, Liu Y. Interpretable deep models for ICU outcome prediction. *AMIA Annu Symp Proc* 2017; **2016:** 371–80.

25 Suresh H. Clinical intervention prediction and understanding using deep networks. *arXiv* 2017; published online May 23. https://arxiv.org/pdf/1705.08498.pdf (preprint).

26 Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a 'right to explanation'. *arXiv* 2016; published online June 28. http://arxiv.org/abs/1606.08813 (preprint).

27 Donchin Y, Seagull FJ. The hostile environment of the intensive care unit. *Curr Opin Crit Care* 2002; **8:** 316–20.

28 Halford GS, Baker R, McCredden JE, Bain JD. How many variables can humans process? *Psychol Sci* 2005; **16:** 70–76.

29 Schneeweiss S. Learning from big health care data. *N Engl J Med* 2014; **370:** 2161–63.

30 Bell G, Hey T, Szalay A. Computer science. Beyond the data deluge. *Science* 2009; **323:** 1297–98.

31 Pastores SM, Dakwar J, Halpern NA. Costs of critical care medicine. *Crit Care Clin* 2012; **28:** 1–10.

32 Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* 2010; **38:** 65–71.

33 Sakr Y, Lobo SM, Moreno RP, et al. Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome. *Crit Care* 2012; **16:** R222.

34 Gajic O, Urrutia LE, Sewani H, Schroeder DR, Cullinane DC, Peters SG. Acute abdomen in the medical intensive care unit. *Crit Care Med* 2002; **30:** 1187–90.

35 Gaieski DF, Mikkelsen ME, Band RA, et al. Impact of time to antibiotics on survival in patients with severe sepsis or septic shock in whom early goal-directed therapy was initiated in the emergency department. *Crit Care Med* 2010; **38:** 1045–53.

36 Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med* 2006; **34:** 1589–96.