BRAIN
ORIGINAL ARTICLE

A recurrent machine learning model predicts intracranial hypertension in neurointensive care patients

Nils Schweingruber,[1] Marius Marc-Daniel Mader,[2,3] Anton Wiehe,[1,4] Frank Röder,[1,4] Jennifer Göttsche,[2] Stefan Kluge,[5] Manfred Westphal,[2] Patrick Czorlich[2] and Christian Gerloff[1]

The evolution of intracranial pressure (ICP) of critically ill patients admitted to a neurointensive care unit (ICU) is difficult to predict. Besides the underlying disease and compromised intracranial space, ICP is affected by a multitude of factors, many of which are monitored on the ICU, but the complexity of the resulting patterns limits their clinical use. This paves the way for new machine learning techniques to assist clinical management of patients undergoing invasive ICP monitoring independent of the underlying disease.

An institutional cohort (ICP-ICU) of patients with invasive ICP monitoring ($n = 1346$) was used to train recurrent machine learning models to predict the occurrence of ICP increases of ≥22 mmHg over a long (>2 h) time period in the upcoming hours. External validation was performed on patients undergoing invasive ICP measurement in two publicly available datasets [Medical Information Mart for Intensive Care (MIMIC, $n = 998$) and eICU Collaborative Research Database ($n = 1634$)].

Different distances (1–24 h) between prediction time point and upcoming critical phase were evaluated, demonstrating a decrease in performance but still robust AUC-ROC with larger distances (24 h AUC-ROC: ICP-ICU 0.826 ± 0.0071, MIMIC 0.836 ± 0.0063, eICU 0.779 ± 0.0046, 1 h AUC-ROC: ICP-ICU 0.982 ± 0.0008, MIMIC 0.965 ± 0.0010, eICU 0.941 ± 0.0025). The model operates on sparse hourly data and is stable in handling variable input lengths and missingness through its nature of recurrence and internal memory. Calculation of gradient-based feature importance revealed individual underlying decisions for our long short time memory-based model and thereby provided improved clinical interpretability.

Recurrent machine learning models have the potential to be an effective tool for the prediction of ICP increases with high translational potential.

1 Department of Neurology, University Medical Centre Hamburg-Eppendorf, Hamburg 20246, Germany
2 Department of Neurosurgery, University Medical Centre Hamburg-Eppendorf, Hamburg 20246, Germany
3 Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
4 Department of Informatics, University of Hamburg, Hamburg, 22527, Germany
5 Department of Intensive Care Medicine, University Medical Centre Hamburg-Eppendorf, Hamburg 20246, Germany

Correspondence to: Nils Schweingruber
Martinistraße 52, Department of Neurology
University Medical Centre Hamburg-Eppendorf
20246 Hamburg, Germany
E-mail: n.schweingruber@uke.de

Received July 30, 2021. Revised October 24, 2021. Accepted November 19, 2021. Advance access publication February 9, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of the Guarantors of Brain.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Keywords: machine learning; deep learning; traumatic brain injury; intracranial pressure; intensive care unit

Abbreviations: AUC = area under the curve; CPP = cerebral perfusion pressure; eICU = collaborative research database; ICP = intracranial pressure; ICU = intensive care unit; MIMIC = Medical Information Mart for Intensive Care; PR = precision recall; ROC = receiver operating characteristics; TBI = traumatic brain injury

Introduction

Data acquisition and storage has exponentially increased and developed in the medical field over the last 10 years. Medical data include vital parameters, laboratory values, imaging data, and genetic information. The vast amount of available data brings new complexity to the medical field. The ability to fully process and interpret such big data is limited even for experienced specialists. If patterns become too multidimensional and complex, reproducible conclusions and optimal decisions cannot be made anymore. This is exactly where machine learning-based decision support can help. Easy accessibility of medical data due to digital storage in combination with recent developments in the field of machine learning hold the potential for automatized data processing to make predictions, look for certain patterns, and classify these data.[1]

Severely ill patients admitted to an intensive care unit (ICU) are at risk to deteriorate within a short period of time and therefore must be more closely monitored than any other patient in the hospital. The medical regimen is adapted according to the alarms of the monitoring system if parameters outside a pre-set range are detected. Laboratory values, imaging data, and other information are interpreted by physicians and allow for further treatment decisions. Due to a great amount of constantly changing information, which is acquired in real time, not all discrete changes, particularly their combinations in all values, might be fully recognized by the medical staff. Setting individual alarm thresholds often is not standardized, leading to alarm fatigue which jeopardizes alarm safety.[2–4] Furthermore, patterns which are believed to be unassociated with the current medical problem can go unnoticed. For this reason, new technologies hold the potential to improve treatment of severely ill patients, as shown in sepsis[5] therapy or the prediction of circulatory[6] or renal failure.[7]

Neurointensive care units (Neuro-ICU) house a plethora of neurological conditions, including intracerebral haemorrhage (ICH), subarachnoid haemorrhage (SAH), ischaemic stroke, and traumatic brain injury (TBI). Moreover, patient demographics, general health statuses and comorbidities can vary.[8,9] Many of diseases treated in the Neuro-ICU can lead to an elevated intracranial pressure (ICP). Clinical management of intracranial hypertension often relies on invasive ICP measurement via external ventricular drainage or intraparenchymal probe.[10–12] Long-sustained pressure phases should be avoided to protect affected and non-affected surrounding brain tissue from secondary deterioration or entrapment.[13–16] Clinical guideline recommendations for ICP thresholds can vary between groups of patients, but treatment of an elevated ICP should be initiated immediately in order to keep the periods of an elevated ICP as short as possible.[17–19] One of the most common reasons for an elevated ICP is cerebral oedema, which can develop focally or in a diffuse way.[20] Oedema, whether vasogenic or cytotoxic, is greatest between 24–72 h after ictus.[21] Unless a neurosurgical evacuation of a mass effect or treatment of an underlying acute hydrocephalus are indicated, Neuro-ICU regimens mainly depend on the usage of pharmacological agents for either the treatment of a brain oedema or induction of a deep sedation for reduction of the brain metabolism. If these actions do not control the ICP, a decompressive craniectomy can be considered.[20,22,23]

Machine learning has been used mainly to process the waveform signal obtained from invasive monitoring.[24–26] Besides the plain ICP signal, other waveform data like ECG or haemodynamics have been used to support predictions of intracranial hypertensive phases.[26–28] Other studies have included clinical measurements like the Glasgow Coma Scale (GCS),[29] the time-series summary statistics of the first 24 h[30] or multiscale waveform metrics.[31] A 30-min time window to an critical ICP increment was chosen to be sufficient for clinical decision-making.[32] Recent approaches try to integrate medical imaging information to predict ICP[33,34] or intracranial pathologies.[35] On the other hand the ICP signal itself can be used to predict ventriculitis with machine learning.[36] Recurrent machine learning in contrast to other machine learning methods, is better suited to learn time-dependent information since the output is fed back into the input of the model during training.[37]

This study aims to explore the potential of recurrent machine learning to predict an elevated ICP in Neuro-ICU patients. External validation is confirmed in publicly available Medical Information Mart for Intensive Care (MIMIC) and eICU Collaborative Research Database (eICU). Besides the goal of robust models for ICP prediction, methods to calculate individual feature importance are used to reflect underlying decisions made by the algorithm to improve clinical acceptance and translational value.

Materials and methods

Study design and setting

Model development and internal validation were performed in a retrospective single centre cohort. Our institutional and interdisciplinary department of intensive care medicine operates 140 high-care ICU beds and treats ~8200 patients per year.

Ethical approval and patient consent

The study protocol was reported to the local ethics committee (reference number WF-059/20) and was conducted according to the Declaration of Helsinki. Written informed consent was waived for this kind of study since all datasets have been de-identified prior to processing and evaluation for the purposes of the study.

Participants and data sources

The ICU is equipped with Dräger Monitoring systems. Patient information, laboratory values, blood-gas samples, and vital parameters are stored centrally in the Dräger® supported Software Integrated Care Manager (ICM). Dräger® supports a tool to search for certain treatments and situations applied; the tool is called ICMiq. It searches in a passively saved reporting database. All patients with a parenchymal ICP probe (Codman Microsensor, Integra

LifeSciences) during the study period (01/2008–01/2020) were included.

Preprocessing

Figure 1A demonstrates the workflow of the inclusion criteria, the preprocessing, training, and extraction of feature importance of the study. The obtained de-identified data were preprocessed with R (Tidyverse package[38]). The detailed preprocessing procedure can be obtained through the publicly available GitHub repository: https://github.com/agschweingruber/icp. Blood gas analysis (BGA) and laboratory values are stored directly and automatically in the system. Certain laboratory values are obtained once a day (e.g. CRP, white blood cells) and blood gas analysis at least every 4 h in patients under invasive ventilation. Sampling was performed more often when patients e.g. suffer under critical invasive breathing situations or present an increase in the ICP. Compared to the external cohorts, the institutional dataset (ICP-ICU) cohort had a much higher frequency of laboratory measurements. All physicians and nursing staff are trained in the documentation system, and vital signs are digitally documented at least once every hour or in case of special events by the nursing staff. Medication is assigned manually in the system and does not change automatically. Scores are applied by drop-down menus in the software interface. Only continuous medication was considered for model training. Groups of drugs were defined through their active ingredients (e.g. narcotics with propofol or ketamine). A dictionary of defined groups based on their strings is also supported in the study repository. The preprocessing was evaluated through a dimensionality reduction to find the fewest differences between datasets (Supplementary Fig. 1). All values were averaged when available over 1 h and this average was used as Input. Vital parameters like the ICP were available one to two times in the ICP-ICU dataset and on a 5-min resolution in the MIMIC and eICU dataset. More information about the general preprocessing and alignment to the external datasets (MIMIC-III and eICU) procedure can be obtained from the Supplementary material and the public GitHub repository.

Defining intracranial hypertensive phases: targets

Targets, which are the variable to be predicted by the model, reflect critical ICP phases and were defined based on hours with a critical ICP event. An hour was defined as a critical phase if at least one ICP measurement was ≥22 mmHg. This threshold was chosen based on the distribution of all ICP measurements (both internal and external) in surviving patients, which indicated 21 mmHg as the 95th percentile. A maximum of two consecutive critical hours were defined as a short critical phase. More than two consecutive critical hours were defined as a long critical phase (Fig. 4A). The targets were defined according to the temporal proximity of the critical phase (1–10 and 24 h). Targets were only defined when ICP measurements were available. Nevertheless, the complete ICU trajectory of a patient was used for training purposes.

Supervised recurrent learning

A common approach to deal with sequential data in the medical domain is to break sequences into fixed-size blocks. A model such as gradient boosted trees or even simple linear regressions can then operate over all time steps at once. Sequences that are shorter than the block size can be padded to match, but very long sequences require a prediction for every part of it. This hinders the model from learning long-term dependencies unless heavy feature engineering is applied, such as adding the long-term variance of features from the time-steps before the block.

The study aimed at creating a model that is robust to missing features and to raw datasets from various clinical sources. A long short-term memory (LSTM)[31] cell is a recurrent neural network unit that can operate on arbitrary sequence lengths and decide what information to remember or forget.

The training and tuning took place on 80% of the data from the ICP-ICU dataset (training set). The validation took place on 20% of the ICP-ICU dataset (test set) and the whole external cohorts. For more information about the training, tuning, hardware, software and used packages refer to the Supplementary material.

Statistics

To obtain the receiver operating characteristic (ROC) and precision recall (PR) curve, the according sensitivity (recall), specificity, and precision were calculated. Predictions of deep learning models range continuously between zero and one and a threshold must be set to classify a prediction into true or false. To visualize the performance of machine learning models independently from a set threshold, ROC and PR are used. ROC curves simulate the trade-off between specificity and sensitivity (a perfect classifier would have the AUC-ROC of one). PR curves demonstrate the trade-off of precision (positive predictive value) and recall (sensitivity), the higher the sensitivity the lower the positive predictive value will be [a perfect classifier would also have the area under the curve (AUC)-PR of one].

To calculate a possible accuracy of the model predicting long critical phases, the optimal threshold was chosen for the highest value, when the false positive rate (1 − specificity) was subtracted from the true positive rate (sensitivity).

The AUC was calculated (yardstick R-Package). The mean and the standard deviation (SD) were calculated based on the prediction of five independent models. Visualization was done using ggplot2 and patchwork. Tables were created using the package gt table. Post-finishing (layout and alignment of text) was done using Adobe Illustrator.

Feature importance

Neural networks can have complex architectures (multiple different layers) and certain randomness in classification (bias). In essence and practical terms, feature importance of recurrent machine learning models indicates a potential role of an input feature for a certain prediction. In contrast to normal statistical models, neural networks can have a certain randomness of prediction and that is also true for their consideration of input features. A feature importance method normally repeats its calculation several times to build an average importance of each feature. The main goal of this study was to be able to calculate a feature importance for every individual input timestep and over the whole individual past of a patient. These individual feature importances are an important factor as to why the present study was based on a sequence-to-sequence LSTM. To calculate feature importance the recently introduced method SmoothGrad (SG)[38] and integrated gradients (IG)[39] was used. See the Supplementary material for detailed information. Feature importance for the prediction of long and short phases by five independent models was calculated on the ICP-ICU test set, the MIMIC dataset, and the eICU dataset.

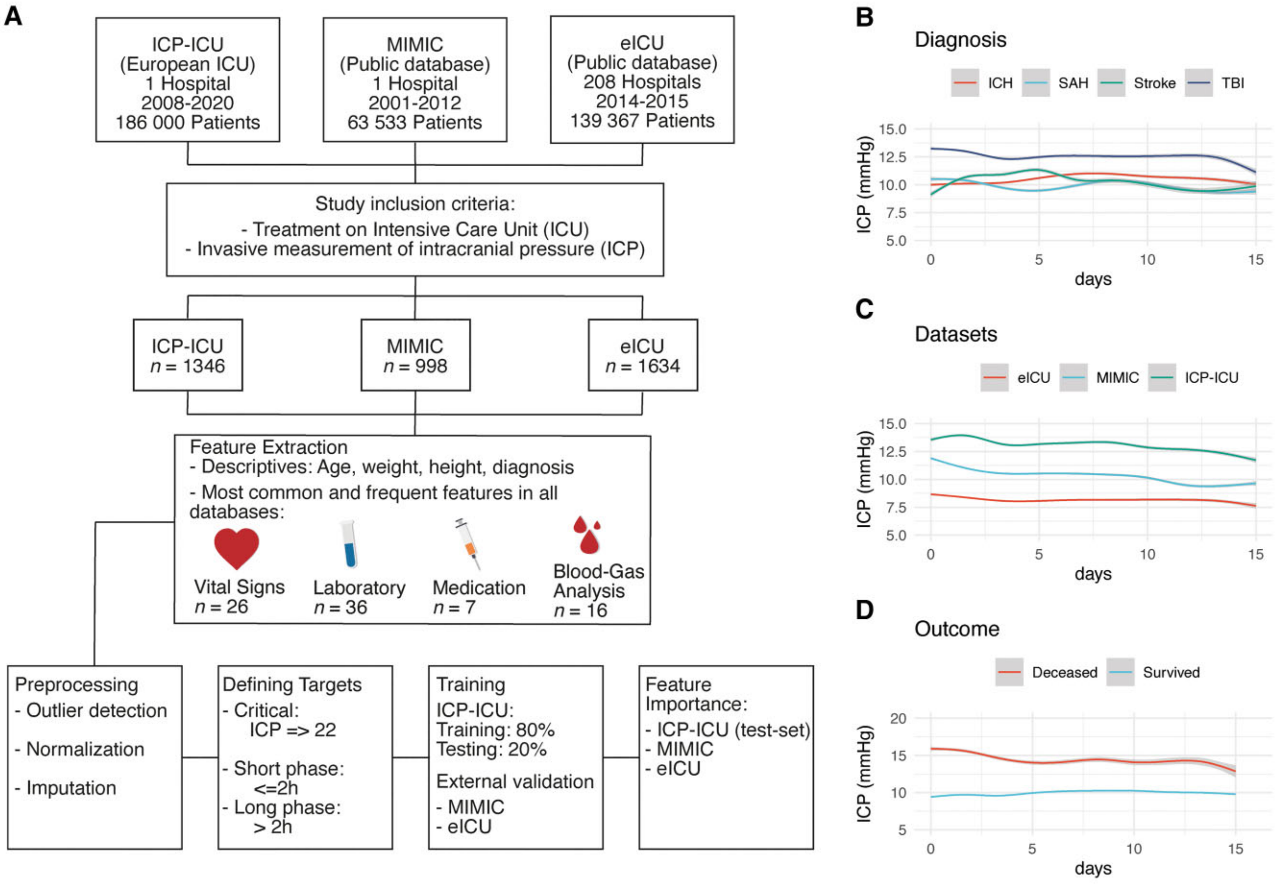

Figure 1 Overview of study design and ICP dynamics. (**A**) Workflow of data acquisition, preprocessing, and training. The institutional dataset is labelled as ICP-ICU which was processed the same as the external datasets: MIMIC-III (Medical Information Mart for Intensive Care) and the eICU (eICU Collaborative Research Database). Invasive measurement of intracranial pressure (ICP) and treatment in ICU were the main inclusion criteria. (**B**–**D**) ICP values over the first 15 days in ICU (mmHg) are depicted according to the diagnosis (**B**), to the dataset (**C**), and to the outcome. Values are shown as a generative additive model with standard deviation in grey.

Data and material availability

Publicly available data (MIMIC-III and eICU) can be found: PhysioNet.org (for credentialed users only).[40] The local dataset is not publicly available because no informed patient consent or ethical approval exist for this purpose. If other investigators are interested in performing additional analyses, requests can be made to the corresponding author and analyses will be performed in collaboration with the University Medical Centre Hamburg-Eppendorf.

Code availability: https://github.com/agschweingruber/icp.

Results

Study population

In total, 1346 patients with an invasive ICP monitoring were admitted to our institutional ICU between January 2008 and January 2020. The cumulative time of these patients accounted for the treatment of ∼63.5 years (Table 1). The most common diagnoses among the study cohort were TBI (33.9%), ICH (20%), stroke (16.1%), SAH (11.1%), and tumours (4.4%) (Table 1). The ICP course over time differed among diagnoses, datasets, and outcomes reflecting the heterogeneity of the study cohort (Fig. 1B and C and Supplementary Tables 1 and 2). Patients with TBI (12.5 ± 9.5 mmHg) and deceased outcome (14.7 ± 15.8 mmHg) had the highest mean ICP during the entire ICU stay. The local dataset ICP-ICU demonstrated the highest average of ICP measurements compared to the publicly accessible ICU datasets MIMIC and eICU. Outcome was comparable between the ICP-ICU and the MIMIC dataset (both with 27.2% deceased patients, Table 1). The ICP-ICU cohort comprised the most severely ill patient population reflected by a lower GCS (median 3 IQR: 4) throughout the entire stay (Supplementary Table 2). Approximately half of the cumulative ICU stay took place without invasive ICP monitoring (ICP-ICU 57.5% MIMIC 65.8% and eICU 55%) (Table 1). Short (≤2 h) critical phases were similarly distributed across outcome groups (deceased 1.4–2.7%, survived 0.8–2.3%). Long phases (>2 h) were more prevalent in the deceased outcome group (deceased 2.3–8.5%, survived 0.8–2%) (Table 1).

Different distances between prediction time point and upcoming critical phase

To evaluate the influence of the temporal distance between prediction time point and target, different models were trained up to 24 h in advance (Fig. 2). An extension of the temporal proximity to 24 h resulted in a robust ROC performance (AUC-ROC ICP-ICU test set = 0.826 ± 0.0071, MIMIC = 0.836 ± 0.0063, eICU = 0.779 ± 0.0046) indicating a sufficient prediction. However, extension of the temporal distance led to a reduced PR (AUC PR ICP-ICU test set = 0.29 ± 0.0069, MIMIC = 0.18 ± 0.0049, eICU = 0.30 ± 0.0089). Reducing the time to

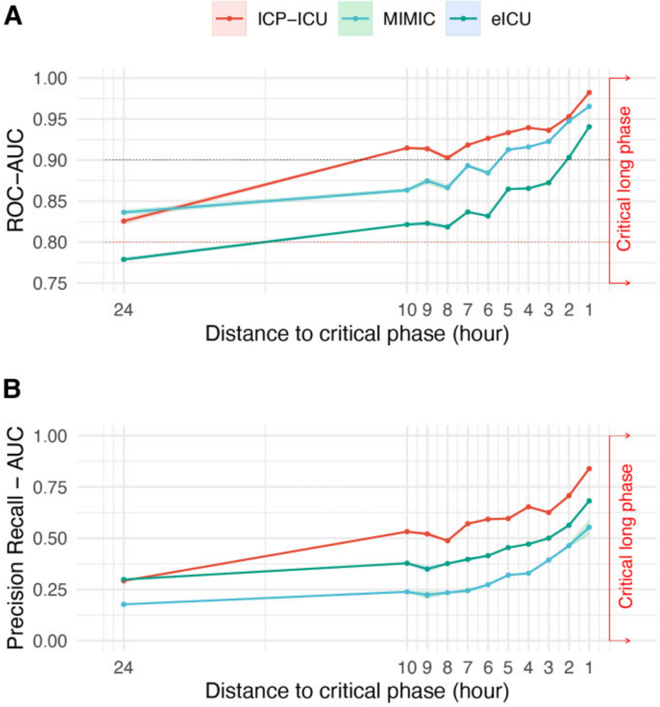

Figure 2 Performance of models with different distances between prediction time point and upcoming critical phase. (**A**) Five independent models were trained on different splits of training data from our institutional dataset to predict critical phases up to 24 h in advance. The AUC of ROC curves and (**B**) PR curves with the corresponding standard deviation of five independent models (ribbon) are depicted for each trained hour (1–10 h and 24 h). Performance on the underlying test set of the institutional dataset (ICP-ICU in red) and external datasets MIMIC (green) and eICU (blue) are depicted separately.

target to 1 h outperformed the other proximities according to AUC-ROC (ICP-ICU test set = 0.98 ± 0.0008, MIMIC = 0.965 ± 0.0010, eICU = 0.941 ± 0.0025) (Supplementary Table 3). In line with the AUC-ROC, the accuracy of the model improved for the ICP-ICU test set from the 24 h distance [75.1% (confidence interval, CI 74.4–75.8%)] to the 1-h distance [93.1% (CI 92.9–93.3%)]. This was also true for the validation on MIMIC [24 h: 76.2% (CI 75.6–76.7%), 1 h: 90.7% (CI 90.1–91.2%)] and eICU [24 h: 71.2% (CI 70.6–71.9%), 1 h: 87.1% (CI 86.6–87.6%), Supplementary Table 4]. The specificity between the 1 h and the 2 h model showed only little differences [1 h: ICP-ICU 91.6% (CI 90.7–92.5%), 2 h ICP-ICU 89.0% (CI 87.3–90.5%)] compared to larger differences in sensitivity [1 h: ICP-ICU 94.6% (CI 93.9–95.3%), 2 h ICP-ICU 89.0% (CI 87.5–90.5%), Supplementary Table 4]. For further in-depth analysis of model performance the 2 h distance to target was chosen, since the clinically more relevant longer distance of 2 h to the critical event outweighs the decrease in sensitivity and specificity.

Model performance on the prediction of long critical phases 2 h in advance

The model with 2 h proximity to predict long (>2 h) and short (≤2 h) critical phases had an AUC-ROC of 0.95 (±0.0009) in the test set [with the best PR for long phases (AUC-PR) 0.71 (±0.0067)] (Fig. 3 and Supplementary Table 5). Additionally, the transition of results from ICP-ICU cohort to the publicly available, external validation datasets was robust (MIMIC, *n* = 998, 50 treatment years, with an AUC-ROC of 0.948 ± 0.0025 and eICU, *n* = 1440, 60 treatment years with an AUC-ROC of 0.903 ± 0.0033). Besides the validation in external datasets, predefined subgroups were then analysed concerning model performance. The model performance was stable regarding outcome and weeks of treatment. PR was higher in the deceased group (death on ICU, AUC-PR 0.73 ± 0.0054, survived 0.456 ± 0.0121) and in the first week of treatment (first week AUC-PR 0.593 ± 0.008, third week 0.371 ± 0.0228). The days of all patients' ICU trajectories were partitioned into two groups to estimate model performance with missing data. A day was defined as 'Less Missing' when it had fewer than 700 data-points missing from a total of ∼2016 possible data-points per day, splitting the days into 49.8% (Less Missing) and 50.2% (More Missing). The model performance was better when less data were missing (Less Missing: AUC-ROC 0.956 ±0.0012 and More Missing: 0.899 ± 0.0045) (Fig. 3E and Supplementary Table 5) and especially in the first week of ICU surveillance [AUC-ROC 0.922 (±0.0024) Fig. 3F]. Concerning different diagnoses, the performance was best in patients suffering from TBI (AUC-ROC 0.918 ± 0.0018, Fig. 3D). For the institutional cohort (ICP-ICU) the accuracy was 89.2% (CI: 88.9–89.5%) with a sensitivity of 88.6% (CI: 88.2–89.1%) and a specificity of 89.8% (CI: 88.9–90.7%). MIMIC had an accuracy of 80.7% (CI: 80.0–81.4%) and eICU of 78.3% (CI: 77.2–79.3%) (Supplementary Table 4).

Feature importance

To demonstrate the advantage of the applied method to calculate feature importance, an individual example of an ICP trajectory was depicted (Fig. 4A). This patient demonstrated a long critical ICP phase in the beginning and a few short phases at the end of the invasive ICP measurement period (Fig. 4A). The according individual feature importance is shown as a heat map to provide an

Table 1 Patient characteristics

Dataset	ICP-ICU[a]	MIMIC[b]	eICU[b]
Descriptive			
Years of age	54.8 (±17.8)	55.8 (±18.9)	54.8 (±17.7)
Weight, kg	79.8 (±15.0)	80.2 (±18.7)	83.6 (±23.7)
Height, cm	173.4 (±9.3)	170.8 (±8.3)	170.3 (±10.9)
Outcome			
Deceased	381 (27.2%)	266 (27.2%)	284 (17.4%)
Survived	1021 (72.8%)	713 (72.8%)	1350 (82.6%)
Gender			
Female	562 (40.1%)	447 (45.7%)	733 (44.9%)
Male	840 (59.9%)	532 (54.3%)	901 (55.1%)
Diagnosis			
TBI	475 (33.9%)	240 (24.5%)	270 (16.5%)
ICH	294 (21.0%)	160 (16.0%)	375 (22.9%)
Stroke	226 (16.1%)	64 (6.5%)	135 (8.3%)
Miscellaneous	190 (13.5%)	186 (19.0%)	567 (34.7%)
SAH	156 (11.1%)	252 (25.7%)	179 (11.0%)
Tumour	61 (4.4%)	77 (7.9%)	108 (6.6%)
ICP values			
Deceased, mmHg	19.5 (±18.5)	10.8 (±6.4)	10.3 (±14.6)
Survived, mmHg	11.6 (±5.9)	10.2 (±5.7)	7.9 (±7.4)
Critical phases: deceased, h			
Overall time	95 084 (100.0%)	49 509 (100.0%)	50 001 (100.0%)
No ICP measurement	44 612 (46.9%)	24 912 (50.3%)	20 391 (40.8%)
Not critical	39 651 (41.7%)	22 720 (45.9%)	24 796 (49.6%)
Long phase	8221 (8.6%)	1151 (2.3%)	3483 (7.0%)
Short phase	2600 (2.7%)	726 (1.5%)	1331 (2.7%)
Critical phases: survived, h			
Overall time	480 783 (100.0%)	311 786 (100.0%)	448 625 (100.0%)
No ICP measurement	274 435 (57.1%)	204 169 (65.5%)	243 971 (54.4%)
Not critical	192 982 (40.1%)	102 428 (32.9%)	184 847 (41.2%)
Long phase	4980 (1.0%)	2527 (0.8%)	9342 (2.1%)
Short phase	8386 (1.7%)	2662 (0.9%)	10 465 (2.3%)

ICH = intracerebral haemorrhage; SAH = subarachnoid haemorrhage.
[a] ICP-ICU is an institutional database that was used for training.
[b] MIMIC, eICU are publicly accessible databases not used for training.

overview of all important features that accounted for a given prediction (Fig. 4B). Higher ICP and cerebral perfusion pressure (CPP) values were positively, and higher haemodynamic parameters were negatively correlated with the occurrence of long phases. The total GCS was positively correlated with the occurrence of long phases.

Aggregating all the feature importances (ICP-ICU test set, MIMIC, eICU) revealed that the prediction of long critical phases relied mainly on ICP and for short critical phases the model took more input features into account (Fig. 4C and Supplementary Fig. 4). The ICP, the mean arterial pressure (MAP) and the CPP were the most important dynamic predictors for critical long phases. Therefore, a higher ICP, MAP or CPP correlates with a critical phase 2 h for both long and short phases. For BGA, sodium and bicarbonate correlate with the occurrence of long critical phases, and higher glucose and chloride was negatively correlated with the prediction of long phases. Continuous medication also demonstrated an influence on the model prediction. A higher continuous dosage of opioids correlated with upcoming critical phases, while narcotics like propofol showed the opposite effect. Laboratory values are less frequently acquired but can influence the model prediction. The most important values were thrombocytes, erythrocytes, mean corpuscular haemoglobin (MCH) and blood urea nitrogen (BUN) (Fig. 4C). To visualize the overall feature importance over time, a heat map was drawn to show possible time dependent changes in feature importance and all datasets (Supplementary Fig. 4B).

Discussion

The present study demonstrates that recurrent machine learning techniques can effectively be used to predict critical phases of intracranial hypertension in patients with invasive ICP measurement. This was achieved by the usage of stored real-world data from the Neuro-ICU over the past decade and the validation on publicly available open-source ICU data. The prediction for critical long phases was still sufficient when it was done 24 h in advance. A closer distance to the critical phase resulted in a more precise prediction.

A forerun of 2 h to react upon a prediction appeared reasonable upon presented results. Two hours is enough time to prepare an anticipated reaction in clinical settings, like adapting sedative medication, addressing invasive breathing conditions, or considering other more invasive procedures. Comparable proximities for the predictions of sepsis were chosen and have already been evaluated in prospective studies to significantly reduce mortality.[41] When implemented in Neuro-ICU, models with different prediction

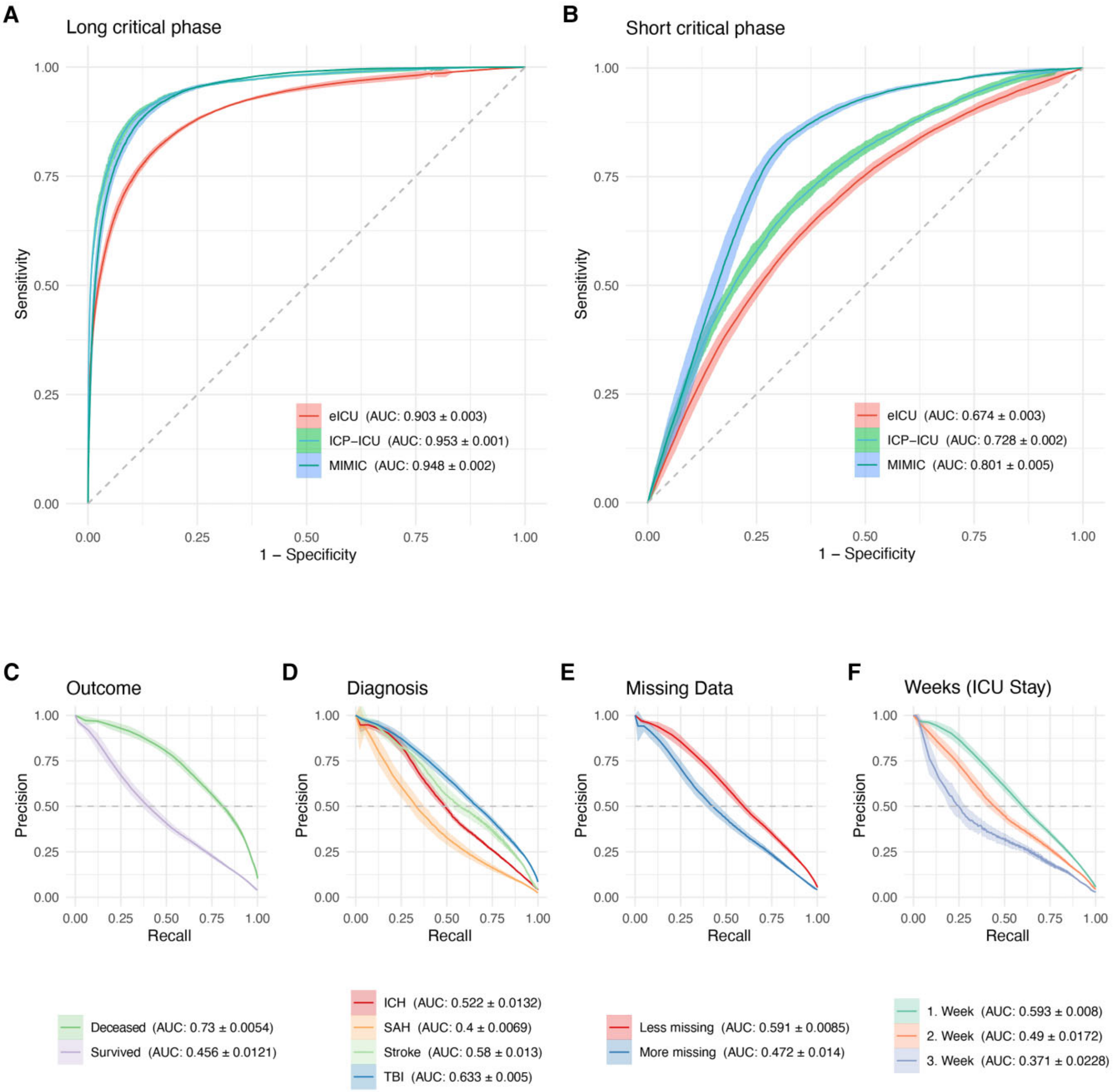

Figure 3 Predicting critical phases 2 h in advance. (**A**) ROC curves are shown of the model predicting critical phases of ICP values of ≥22 mmHg for more than two consecutive hours (>2 h) and are referred to as long critical phases. (**B**) Critical phases under 2 h (≤2 h) are referred to as short critical phases. A whole hour was defined as critical (target) when one single ICP value measurement that hour was ≥22 mmHg. Ribbons represent the standard deviation of five independent models. Model performance on external datasets is also shown MIMIC (blue) and eICU (red) (**A** and **B**). Model performance according to certain subgroups was drawn as a PR curve. Outcome (**C**) is defined as deceased on ICU stay. (**D**) Diagnosis is defined by main diagnosis of ICD-10. (**E**) Missing data dichotomy was done by defining two groups of days per patient. One group had fewer than 77% missing data-points (of a total of 2016 possible data-points per day), splitting the days into two groups (49.8% Less Missing and 50.2% More Missing). (**F**) To show a possible decline in model performance over the time course of ICU stay, all days are grouped according to their week. For each group or subgroup drawn in a different colour, the corresponding AUC and the standard deviation are shown in the legend.

proximities could be combined to reach a higher certainty when the critical target approaches.

Missingness of data in clinical routines is an issue in real-world settings.[42] Whole data streams can suddenly be interrupted due to varying and unexpected reasons. Recurrent machine learning models can make predictions on very long and very short input sequences, making it more robust towards sudden missing data at certain points of individual time courses. Implementing as many input features as possible supports possible predictions when certain inputs drop.

Another advantage of the presented approach, besides using the underlying real-world data, is the broad spectra of pathologies leading to intracranial hypertension. Previous studies training models to predict ICP increment are mainly trained on data from patients suffering from TBI.[25,27,30,32,43–46] The presented approach does not discriminate between different diagnoses. Generally, the relevant

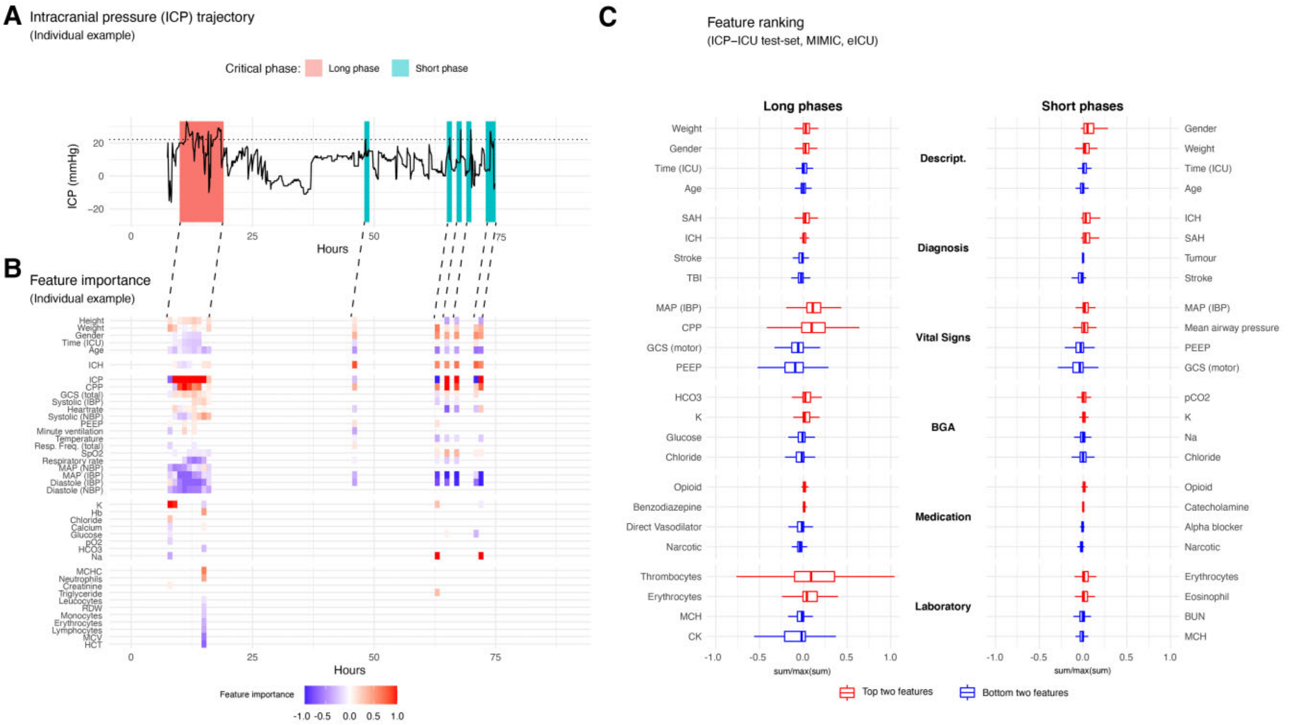

Figure 4 Feature importance of the prediction of long and short critical phases of intracranial pressure. (**A**) A representative ICU trajectory of an individual patient with invasive ICP monitoring is presented, having a long critical phase in the beginning and several shorter critical phases at the end. The individual ICP course is depicted over time (h); the horizontal dashed line represents our threshold for the definition of critical phases (ICP 22 mmHg). (**B**) Gradient based saliencies were calculated from five independent models based on the prediction 2 h in advance of the critical phases. All other features which had a low influence are not shown for that trajectory. The lines connecting the saliency and ICP plot demonstrate the predictive horizon. The prediction takes place 2 h in advance and the important features for that prediction at that time are demonstrated. The colour scale is continuous between −1 (blue) and 1 (red). Values being positive are red (to be considered as bad) because their higher values are positively correlated with the prediction of critical phases. Negative values (blue) represent negatively correlated values with the positive prediction. (**C**) To have a broader view on the top features over all validation datasets (ICP-ICU test set, MIMIC and eICU), the sum of all saliencies per timestep were calculated. The *top* (red) and *bottom* (blue) two features are shown for each group. Descript. = patient characteristic, diagnosis, vital signs, BGA, medication, laboratory value; and for each target long (*left*) and short (*right*) critical phase. The lower and upper hinges of box plots correspond to the first and third quartiles (the 25th and 75th percentiles) the middle line of the median. The upper and the lower whisker extends from the hinge to the largest and smallest value no further than 1.5 × IQR from the hinge.

diagnosis is already available upon Neuro-ICU admittance and this information played an important for the prediction of critical phases (Fig. 4C).

Aiming at clinical implementation of machine learning-based decision support, a real time visualization of features influencing the individual prediction of the model will be important. The proposed visualization of feature importance can serve as an example of how a clinical software interface for decision support could be designed (Fig. 4A and B). This presented approach for explainable artificial intelligence will play a crucial role in implementing these tools in direct patient care. The Neuro-ICU team must be informed about underlying decisions made by the models to predict critical ICP phases. A clinician should see individual feature trajectories of patients suffering ICP at one glance.

Implementing the model in clinical routine could lead to a much more preventive handling of patients with invasive ICP monitoring. An ICP-driven management of TBI patients is suggested to be associated with a lower 6-month mortality.[12] Besides the sole ICP measurement a common clinical approach is a multimodal neuromonitoring in specialized Neuro-ICUs.[47] Indices of the ICP waveform and blood pressure monitoring can be used to calculate an optimal CPP.[48] The distance of actual and optimal CPP is associated with a higher mortality.[49] Other modalities like near-infrared spectroscopy (NIRS),[50] transcranial Doppler ultrasound (TCD)[51] and cerebral microdialysis[52] can be used to get an information about the cerebral blood flow, the underlying autoregulation or tissue oxygenation of patients on Neuro-ICU. All modalities can lead to a much more precise and individual management of ICP on Neuro-ICU. Besides being a further block in this chain of modalities to ICP guidance, the presented recurrent machine learning approach could be augmented by the integration of multimodal neuromonitoring to achieve an even better performance for the prediction of critical ICP phases in the future. This though presupposes larger clinical real-world databases of multimodal neuromonitoring data to train and validate new machine learning architectures. Nevertheless, its advantage compared to common clinical practice needs to be evaluated in prospective randomized controlled trials and cannot be stated at this point.

The feature importance shed light into limitations of the presented study. Data-driven bias that has been learned by the models can be unveiled. Examples could be the pupil size or the gender of the patients. On both longer and shorter phases, the left pupil size seems to have a negative influence on the prediction of long phases (Supplementary Fig. 4). The pupil size in the underlying trainings cohort was wider in the group of patients that died (left pupil deceased 2.84 ± 1.32 mm and survived 2.49 ± 0.89 mm). The external cohorts show a wider pupil in the surviving cohort (eICU left 3.04 ± 0.90 mm MIMIC left 3.07 ± 0.77 mm). Though the training process was sufficient to achieve generalizable results, the model focusing on the left pupil can be misleading based on this learned bias. Gender and race discrimination plays a big role in recent critical discussions about biased decisions made by machine learning models. This bias can also be found in the predictions made by the proposed model.[53] Male gender was predominant in the training cohort (59.9%) but also in external validation (MIMIC 54.3% and eICU 55.1%). According to the outcome, female patients died less often (deceased female ICP-ICU: 39.3%, MIMIC: 44.7%, eICU: 35.6%). This points to a less precise performance of the presented models in female patients (AUC-PR female 0.473 ± 0.0136 and male 0.596 ± 0.0101). One answer to this could be a more diverse dataset and a local retraining of models if they were pre-trained on biased data. Until then, a constant evaluation of predictions is necessary also considering potentially changing composition of cohorts in an uncertain future.

Shorter phases were much harder to predict since a quite strict definition of critical targets was taken. An hour was determined to be critical when only one ICP measurement was ≥22 mmHg. ICP can rise easily to ≥22 mmHg in the context of physiological body functions. Nevertheless, it seems quite intuitive to have a time dependency in the prediction made by the model. Long-sustained critical phases need different clinical management than a few short critical phases. A stricter definition of short phases concerning ICP increment in future studies based on larger cohorts could lead to a better performance to predict shorter critical phases.

Based on the presented results, the hypothesis can be generated that recurrent machine learning-based prediction of critical ICP phases can lead to a much more precise and anticipated treatment. Prospective studies will be conducted next to evaluate presented models for their performance in clinical settings. No statement can be made about influence on patients' outcomes when these models are used to make predictions in the Neuro-ICU. There is a need for prospective evaluation on the underlying predictions concerning patient safety and outcomes, but also on acceptance by Neuro-ICU treating staff.

Conclusion

Predicting critical ICP phases with recurrent machine learning has several advantages such as dealing with variable input length, bias-free imputation and individual per time step calculable gradient-based feature importance. Recurrent machine learning models are feasible and could become an effective tool for the prediction of ICP increases with high translational potential for the prospective use in clinical studies.

Funding

Young research funding of the Werner Otto Foundation; NVIDIA Academic Hardware Grant.

Competing interests

N.S., M.M.M., A.W., F.R., J.G., S.K., M.W., P.C. and C.G. report no conflicts of interests in relation to the submitted work. Outside the submitted work, C.G. reports personal fees from Amgen, personal fees from Boehringer Ingelheim, personal fees from Daiichi Sankyo, personal fees from Abbott, personal fees from Prediction Biosciences, personal fees from Novartis, and personal fees from Bayer.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
2. Bach TA, Berglund L-M, Turk E. Managing alarm systems for quality and safety in the hospital setting. *BMJ Open Qual*. 2018; 7(3):e000202.
3. Bridi AC, da Silva RCL, de Farias CCP, Franco AS, dos Santos VdeLQ. Reaction time of a health care team to monitoring alarms in the intensive care unit: implications for the safety of seriously ill patients. *Rev Bras Ter Intensiva*. 2014;26(1): 28–35.
4. Poole S, Shah N. Addressing vital sign alarm fatigue using personalized alarm thresholds. *Pac Symp Biocomput*. 2018;23:472–483.
5. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11): 1716–1720.
6. Hyland SL, Faltys M, Hüser M, *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*. 2020;26(3):364–373.
7. Zimmerman LP, Reyfman PA, Smith ADR, *et al.* Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. *BMC Med Inform Decis Mak*. 2019;19(1):16.
8. Howard RS, Kullmann DM, Hirsch NP. Admission to neurological intensive care: who, when, and why?. *J Neurol Neurosurg Psychiatry*. 2003;74 (Suppl 3):iii2–iii9.
9. Ridley S, Burchett K, Gunning K, *et al.* Heterogeneity in intensive care units: Fact or fiction? *Anaesthesia*. 1997;52:531–537.
10. Carney N, Totten AM, O'Reilly C, *et al.* Guidelines for the management of severe traumatic brain injury, fourth edition. *Neurosurgery*. 2017;80(1):6–15.
11. Chesnut RM, Temkin N, Carney N, *et al.* A trial of intracranial-pressure monitoring in traumatic brain injury. *N Engl J Med*. 2012;367(26):2471–2481.
12. Robba C, Graziano F, Rebora P, *et al.* Intracranial pressure monitoring in patients with acute brain injury in the intensive care unit (SYNAPSE-ICU): an international, prospective observational cohort study. *Lancet Neurol*. 2021;20(7):548–558.
13. Lane PL, Skoretz TG, Doig G, Girotti MJ. Intracranial pressure monitoring and outcomes after traumatic brain injury. *Can J Surg*. 2000;43(6):442–448.
14. Bulger EM, Nathens AB, Rivara FP, Moore M, MacKenzie EJ, Jurkovich GJ. Management of severe head injury: institutional variations in care and effect on outcome. *Crit Care Med*. 2002; 30(8):1870–1876.
15. Valentin A, Lang T, Karnik R, Ammerer HP, Ploder J, Slany J. Intracranial pressure monitoring and case mix-adjusted mortality in intracranial hemorrhage. *Crit Care Med*. 2003;31(5): 1539–1542.
16. Farahvar A, Gerber LM, Chiu Y-L, Carney N, Härtl R, Ghajar J. Increased mortality in patients with severe traumatic brain injury treated without intracranial pressure monitoring. *J Neurosurg*. 2012;117(4):729–734.

17. Sorrentino E, Diedler J, Kasprowicz M, *et al.* Critical thresholds for cerebrovascular reactivity after traumatic brain injury. *Neurocrit Care*. 2012;16:258–266.
18. Bratton SL, Chestnut RM, Ghajar J, *et al.* Guidelines for the management of severe traumatic brain injury. VIII. Intracranial pressure thresholds. *J Neurotrauma*. 2007;24(Suppl. 1):S55–S58.
19. Sauvigny T, Göttsche J, Czorlich P, Vettorazzi E, Westphal M, Regelsberger J. Intracranial pressure in patients undergoing decompressive craniectomy: New perspective on thresholds. *J Neurosurg*. 2018;128:819–827.
20. Cook AM, Morgan Jones G, Hawryluk GWJ, *et al.* Guidelines for the acute treatment of cerebral edema in neurocritical care patients. *Neurocrit Care*. 2020;32:647–666.
21. Marmarou A. A review of progress in understanding the pathophysiology and treatment of brain edema. *Neurosurg Focus*. 2007;22(5):E1.
22. Burgess S, Abu-Laban RB, Slavik RS, Vu EN, Zed PJ. A systematic review of randomized controlled trials comparing hypertonic sodium solutions and mannitol for traumatic brain injury: implications for emergency department management. *Ann Pharmacother*. 2016;50(4):291–300.
23. Czosnyka M, Pickard JD, Steiner LA. Principles of intracranial pressure monitoring and treatment. *Handb Clin Neurol*. 2017; 140:67–89.
24. Quachtran B, Hamilton R, Scalzo F. Detection of intracranial hypertension using deep learning. *Proc IAPR Int Conf Pattern Recogn*. 2016;2016:2491–2496.
25. Zhang F, Feng M, Pan SJ, *et al.* Artificial neural network based intracranial pressure mean forecast algorithm for medical decision support. *Annu Int Conf IEEE Eng Med Biol Soc*. 2011;2011:7111–7114.
26. Naraei P, Nouri M, Sadeghian A. Toward learning intracranial hypertension through physiological features: A statistical and machine learning approach. In: *2017 Intelligent Systems Conference, IntelliSys 2017*. Vol 2018-Janua. IEEE; 2018:395–399.
27. Hamilton R, Xu P, Asgari S, *et al.* Forecasting intracranial pressure elevation using pulse waveform morphology. In: Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009. IEEE; 2009:4331–4334.
28. Lee HJ, Kim H, Kim YT, Won K, Czosnyka M, Kim DJ. Prediction of life-threatening intracranial hypertension during the acute phase of traumatic brain injury using machine learning. *IEEE J Biomed Health Inform*. 2021;25(10):3967–3976.
29. Raj R, Luostarinen T, Pursiainen E, *et al.* Machine learning-based dynamic mortality prediction after traumatic brain injury. *Sci Rep*. 2019;9(1):17672.
30. Güiza F, Depreitere B, Piper I, Van Den Berghe G, Meyfroidt G. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: Development and validation in a multicenter dataset. *Crit Care Med*. 2013;41(2):554–564.
31. Hüser M, Kündig A, Karlen W, De Luca V, Jaggi M. Forecasting intracranial hypertension using multi-scale waveform metrics. *Physiol Meas*. 2020;41(1):014001.
32. Myers RB, Lazaridis C, Jermaine CM, Robertson CS, Rusin CG. Predicting intracranial pressure and brain tissue oxygen crises in patients with severe traumatic brain injury. *Crit Care Med*. 2016;44(9):1754–1761.
33. Vasseneix C, Najjar RP, Xu X, *et al.* Accuracy of a deep learning system for classification of papilledema severity on ocular fundus photographs. *Neurology*. 2021;97(4):e369–e377.
34. Miyagawa T, Sasaki M, Yamaura A. Intracranial pressure based decision making: Prediction of suspected increased intracranial pressure with machine learning. *PLoS One*. 2020;15:e0240845.
35. Wang X, Shen T, Yang S, *et al.* A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *NeuroImage Clin*. 2021;32:102785.
36. Megjhani M, Terilli K, Kalasapudi L, *et al.* Dynamic intracranial pressure waveform morphology predicts ventriculitis. *Neurocrit Care*. 2022;36(2):404–411.
37. Schmidhuber J. Deep learning. *Scholarpedia*. 2015;10(11):32832.
38. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: Removing noise by adding noise. arXiv, arXiv:1706.03825
39. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. Vol 7. 2017:5109–5118.
40. Goldberger AL, Amaral LAN, Glass L, *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101-(23):e215–e220.
41. Fleuren LM, Klausch TLT, Zwager CL, *et al.* Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. 2020;46(3):383–400.
42. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, *et al.* Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157–166.
43. Scalzo F, Hamilton R, Asgari S, Kim S, Hu X. Intracranial hypertension prediction using extremely randomized decision trees. *Med Eng Phys*. 2012;34(8):1058–1065.
44. Hu X, Xu P, Asgari S, Vespa P, Bergsneider M. Forecasting ICP elevation based on prescient changes of intracranial pressure waveform morphology. *IEEE Trans Biomed Eng*. 2010;57(5):1070–1078.
45. Güiza F, Depreitere B, Piper I, *et al.* Early detection of increased intracranial pressure episodes in traumatic brain injury: External validation in an adult and in a pediatric cohort. *Crit Care Med*. 2017;45(3):e316–e320.
46. Naraei P, Sadeghian A. A PCA based feature reduction in intracranial hypertension analysis. In: Canadian Conference on Electrical and Computer Engineering. 2017:1–6.
47. Maas AIR, Menon DK, David Adelson PD, *et al.* Traumatic brain injury: Integrated approaches to improve prevention, clinical care, and research. *Lancet Neurol*. 2017;16(12):987–1048.
48. Czosnyka M, Pickard JD. Monitoring and interpretation of intracranial pressure. *J Neurol Neurosurg Psychiatry*. 2004;75(6):813–821.
49. Depreitere B, Güiza F, Van den Berghe G, *et al.* Pressure autoregulation monitoring and cerebral perfusion pressure target recommendation in patients with severe traumatic brain injury based on minute-by-minute monitoring data. *J Neurosurg*. 2014;120(6):1451–1457.
50. Rivera-Lara L, Geocadin R, Zorrilla-Vaca A, *et al.* Validation of near-infrared spectroscopy for monitoring cerebral autoregulation in comatose patients. *Neurocrit Care*. 2017;27(3):362–369.
51. Zeiler FA, Cardim D, Donnelly J, Menon DK, Czosnyka M, Smielewski P. Transcranial Doppler systolic flow index and ICP-derived cerebrovascular reactivity indices in traumatic brain injury. *J Neurotrauma*. 2018;35(2):314–322.
52. Martini RP, Deem S, Yanez ND, *et al.* Management guided by brain tissue oxygen monitoring and outcome following severe traumatic brain injury. *J Neurosurg*. 2009;111(4):644–649.
53. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1–35.