ARTICLE

Machine Learning to Predict Delayed Cerebral Ischemia and Outcomes in Subarachnoid Hemorrhage

Jude P.J. Savarraj, PhD, Georgene W. Hergenroeder, PhD, Liang Zhu, PhD, Tiffany Chang, MD, Soojin Park, MD, Murad Megjhani, PhD, Farhaan S. Vahidy, PhD, Zhongming Zhao, PhD, Ryan S. Kitagawa, MD, and H. Alex Choi, MD

Neurology® 2021;96:e553-e562. doi:10.1212/WNL.0000000000011211

Correspondence
Dr. Choi
Huimahn.A.Choi@uth.tmc.edu
or Dr. Savarraj
jude.p.savarraj@uth.tmc.edu

Abstract

Objective

To determine whether machine learning (ML) algorithms can improve the prediction of delayed cerebral ischemia (DCI) and functional outcomes after subarachnoid hemorrhage (SAH).

Methods

ML models and standard models (SMs) were trained to predict DCI and functional outcomes with data collected within 3 days of admission. Functional outcomes at discharge and at 3 months were quantified using the modified Rankin Scale (mRS) for neurologic disability (dichotomized as good [mRS ≤ 3] vs poor [mRS ≥ 4] outcomes). Concurrently, clinicians prospectively prognosticated 3-month outcomes of patients. The performance of ML, SMs, and clinicians were retrospectively compared.

Results

DCI status, discharge, and 3-month outcomes were available for 399, 393, and 240 participants, respectively. Prospective clinician (an attending, a fellow, and a nurse) prognostication of 3-month outcomes was available for 90 participants. ML models yielded predictions with the following area under the receiver operating characteristic curve (AUC) scores: 0.75 ± 0.07 (95% confidence interval [CI] 0.64–0.84) for DCI, 0.85 ± 0.05 (95% CI 0.75–0.92) for discharge outcome, and 0.89 ± 0.03 (95% CI 0.81–0.94) for 3-month outcome. ML outperformed SMs, improving AUC by 0.20 (95% CI −0.02 to 0.4) for DCI, by 0.07 ± 0.03 (95% CI −0.0018 to 0.14) for discharge outcomes, and by 0.14 (95% CI 0.03–0.24) for 3-month outcomes and matched physician's performance in predicting 3-month outcomes.

Conclusion

ML models significantly outperform SMs in predicting DCI and functional outcomes and has the potential to improve SAH management.

From the Departments of Neurosurgery (J.P.J.S., G.W.H., T.C., R.S.K., A.C.), Internal Medicine (L.Z.), and Neurology (F.S.V.), McGovern Medical School, Center for Precision Health, School of Biomedical Informatics (Z.Z.), and Human Genetics Center, School of Public Health (Z.Z.), The University of Texas Health Science Center at Houston; and Department of Neurology (S.P., M.M.), Columbia University, NY.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

Glossary

ANN = artificial neural network; **AUC** = area under the receiver operating characteristic curve; **CI** = confidence interval; **CV** = cross-validation; **DCI** = delayed cerebral ischemia; **EMR** = electronic medical record; **GB** = gradient boost; **HH** = Hunt-Hess scale; **IQR** = interquartile range; **IRB** = institutional review board; **IVH** = intraventricular hemorrhage; **LR** = logistic regression; **mFS** = modified Fisher Scale; **ML** = machine learning; **mRS** = modified Rankin Scale; **RF** = random forest; **ROC** = receiver operating characteristic; **SAH** = subarachnoid hemorrhage; **TCD** = transcranial Doppler; **WBC** = white blood cell.

After subarachnoid hemorrhage (SAH), delayed cerebral ischemia (DCI) is the largest contributor to poor functional outcomes. DCI is characterized by neurologic worsening occurring between 4 and 21 days after the initial hemorrhage, affecting 20%–30% of patients with SAH. Early identification of DCI and prognostication of functional outcomes are critical aspects in SAH management. Currently, the hematoma volume in the subarachnoid space (quantified by the modified Fisher Scale [mFS][1]) and the clinical severity at admission (quantified by the Hunt-Hess scale [HH][2]) are the validated predictors of DCI and functional outcomes, respectively.[3] HH, typically ascertained by a neurologic examination at admission, is the most widely used predictor of short-term and long-term functional outcomes. Clinicians can subjectively prognosticate patient outcomes based on all available clinical information.

Machine learning (ML) can learn from complex data to find hidden features that can improve predictions. It often requires large data samples ("big data"), which has recently become more accessible,[4] resulting in successful clinical applications.[5–7] ML[8] can objectively learn from hundreds of variables and samples to provide inferences. The electronic medical record (EMR) is a rich repository of archived data (including laboratory data and vital signs) primarily collected for the day-to-day management of patients with SAH. Previous studies show that several EMR measures including white blood count panel (white blood cells [WBCs],[9–11] neutrophil count,[12] platelets,[13] erythrocytes[14]), measures of coagulation and fibrinolysis,[15] serum glucose,[16] and sodium,[17] and vital signs (including ECG[18] and blood pressure[18]) are either marginally or strongly associated with DCI and functional outcomes.[19] We hypothesized that ML models can learn these associations to accurately predict DCI and functional outcomes and outperform standard models. Our objective was to develop ML models that predict DCI and functional outcomes using standard clinical and laboratory measures captured in the EMR. We compared the performance of ML models with that of clinician-based prognostication in predicting 3-month outcomes.

Methods

Standard Protocol Approvals, Registrations, and Patient Consents

We received written informed consent from all patients (or guardians of patients) participating in the study. We received approval from the institutional institutional review board (IRB) and this study was conducted under an institutional IRB-approved protocol.

Study Population, Enrollment Criteria, and Clinical Endpoints

We conducted a retrospective analysis of a prospectively collected cohort of consented patients with SAH admitted between July 2009 and August 2016 to the neuroscience intensive care unit of our tertiary medical center. Patients with SAH with traumatic etiology and those who died or were discharged within 3 days of admission were excluded for developing the ML model for DCI prediction. Patients who died within 5 days of admission were excluded in the development of ML models to predict functional outcomes. Patients for whom EMR data were not available post-SAH were also excluded. The clinical severity at admission was quantified using HH[20] (supplementary data: HH, doi.org/10.5061/dryad.2rbnzs7kk). The hematoma volume on admission CT was quantified using the modified Fisher scale (supplementary data: modified Fisher scale, doi.org/10.5061/dryad.2rbnzs7kk). The functional outcome at discharge and at 3 months was quantified using the modified Rankin Scale (mRS) (supplementary data: functional outcome assessment, doi.org/10.5061/dryad.2rbnzs7kk). DCI status was ascertained using an established definition[21] (supplementary data: DCI assessment, doi.org/10.5061/dryad.2rbnzs7kk). Members of the clinician prognostication team (a board-certified neurocritical care attending physician, a neurocritical care nurse, and a neurocritical care fellow) were asked between days 1 and 3 postadmission to prospectively predict the 3-month mRS of patients with SAH over a 16-month period (supplementary data: Clinician prognostication, doi.org/10.5061/dryad.2rbnzs7kk).

Data Extraction and Imputation

Values of several laboratory and vital measures (supplementary data: List of EMR measures initially extracted, doi.org/10.5061/dryad.2rbnzs7kk), for each 24-hour period, from the day of admission to the 3rd day postadmission were obtained. Only measures that were routinely available for most participants and at regular intervals were included. If multiple recordings of the measure for each day were available (as in the case of vital signs), then the maximum, minimum, average, and SD of each measure was calculated as a separate variable for each day. The SD of the laboratory values for each day were excluded as most laboratories are typically done once or twice a day. Missing data were imputed procedurally (supplementary data: Data preparation and imputation, doi.org/10.5061/dryad.2rbnzs7kk). Raw EMR data were stored in a

MySQL database (MySQL 8.0) and queried by SQL language. The data preparation and imputation was preformed using the Python programming language (v3.6).

Machine Learning Methodology

We first established a baseline clinical model using variables validated in literature (standard model) for predicting DCI and functional outcomes. Next, we compared the ML models with the standard models. The standard model for predicting DCI is a logistic regression (LR) model that included age and the modified Fisher scale.[1] The mRS was dichotomized into a binary response—good (mRS ≤ 3) or poor (mRS ≥ 4)—to train standard and ML models that predict functional outcomes. The standard model for predicting functional outcomes is an LR model that included age and HH. Since several candidate ML models exist, we heuristically investigated models governed by tradeoffs between performance and complexity. Models investigated included support vector machines, random forest, gradient boost, and artificial neural networks (ANNs). The dataset was split into a training and test set (supplementary data: Data split for training/test split protocol, doi.org/10.5061/dryad.2rbnzs7kk). A stratified 10-fold cross-validation (CV) approach was used to train the ML models on the training set and tune the model measures. The model with the best average AUC on the training set (using the 10-fold CV) was chosen and its performance on the test set was evaluated and reported. The ML models were developed using the scikitlearn and tensorflow libraries available in the Python programming language (v3.6).

Statistical Methodology

The performance metrics in this study are sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) (table S1, doi.org/10.5061/dryad.2rbnzs7kk). The metrics of the ML model, the standard model, and the clinician's predictions on the test sets were compared. The training and the test set were randomly stratified and the characteristics are shown (table 1). Receiver operating characteristic (ROC) curves were statistically compared using the DeLong test.[22] To dichotomize the predictions from the ROC curves, the optimal cutoff threshold that corresponded to the maximum Youden Index[23] was selected. McNemar test was used to compare the binary predictions of standard models, ML model, and clinician evaluation.[24] ML development and statistical analysis were performed using python (v3.6) and MedCalc (v18.11.3)[25] software. A complete overview of the methodology is presented in figure 1, A–E.

Data Availability

All anonymized data used in this study are not available in a public domain, but request for data may be considered by the principal investigator.

Results

Demographics and Data Split Approaches

Overall, 451 patients were consented to participate during the study period. Among them, 290 (64%) were female. Baseline characteristics of these patients are shown in table S2 (doi.org/10.5061/dryad.2rbnzs7kk). The average age was 54 years (interquartile range [IQR] 45–63). A total of 223 participants (60%) had a history of hypertension, 60 (17%) had a history of hyperlipidemia, and 44 (12%) had diabetes. The median HH was 3 (IQR 2–3), the median mFS was 3 (IQR 3–3), and an intraventricular hemorrhage (IVH) was observed in 240 (66%) participants on admission CT. A total of 88 (21%) of the participants developed DCI. The median discharge mRS was 3 (IQR 1–4) and the median 3-month mRS was 1 (IQR 0–4).

Data from 399 participants were available for the development of the DCI prediction model and they were randomly stratified into a training set (80%, ~319 participants) and a test set (20%, ~80 participants). Data from 393 were available for the development of the discharge mRS prediction model and they were randomly stratified into a training set (80%, ~314 participants) and a test set (20%, ~79 participants) (table 1). Data from 240 participants were available for the development of the 3-month outcome prediction model of which the clinician prediction was available for 90 participants. For 3-month outcome, these 90 participants with clinician prediction data were used as the test set. The remainder of the data (150 participants) were used as the training set to develop the ML models and standard models (figure 1E). The rationale for the number of participants available for each model is discussed (see supplementary data: Subject and variable availability, doi.org/10.5061/dryad.2rbnzs7kk).

Performance of ML and Standard Model in DCI and Discharge Outcomes

Among the ML models tested, ANN performed well. The ANN model had a 10-fold CV AUC of 0.78 ± 0.16 on the training set (supplementary data: DCI model, doi.org/10.5061/dryad.2rbnzs7kk). The ML model and the standard model were evaluated on the test set and the AUC of the ML model was higher than the standard model (0.75 ± 0.07, 95% confidence interval [CI] 0.64–0.84 vs 0.56 ± 0.07, 95% CI 0.44–0.66, $p = 0.08$, figure 2A). The performances of the ML model and standard model at the optimal cutoff threshold (sensitivity: 0.82 vs 0.79, and, specificity: 0.72 vs 0.25) were significantly different ($p < 0.01$, McNemar test, table 2).

For discharge outcomes, the AUC of the ML model (see supplementary data: Model measures functional outcomes, doi.org/10.5061/dryad.2rbnzs7kk) was significantly higher than the standard model (0.85 ± 0.05, 95% CI [0.75–0.92] vs 0.78 ± 0.06, 95% CI [0.67–0.86], which was a 0.07 [95% CI −0.0018 to 0.14] improvement). The assessment of the ML model (at the optimal cutoff threshold) was significantly different than the standard model (sensitivity: 0.75 vs 0.58 and specificity: 0.87 vs 0.90, respectively, $p < 0.05$, McNemar test). This best performing ML model used a combination of EMR variables and 1 measure that was derived by human intuition—the HH score. The ML model that only used EMR

Table 1 Demographics of Training and Test Set in Delayed Cerebral Ischemia (DCI): Discharge and 3-Month Outcome Cohort

	DCI		Discharge outcome		3-month outcome	
Patient demographics	**Train (n = 319)**	**Test (n = 80)**	**Train (n = 314)**	**Test (n = 79)**	**Train (n = 150)**	**Test (n = 90)**
Age, y	52 (44–62)	56 (45–65)	54 (44–63)	54 (46–62)	54 (44–62)	56 (45–64)
Female	211 (66)	53 (66)	209 (66)	51 (63)	102 (67)	55 (61)
Hypertension	170 (60)	41 (57)	159 (57)	47 (69)	74 (59)	52 (64)
Hyperlipidemia	46 (17)	8 (12)	45 (17)	10 (15)	21 (17)	20 (24)
Diabetes	33 (12)	7 (10)	34 (13)	5 (7)	15 (12)	8 (9.8)
Clipping	96 (30)	29 (36)	103 (32)	26 (32)	42 (27)	25 (27)
mRS	3 (1–4)	3 (2–4)	3 (1–4)	3 (1–4)	1 (0–2)	1 (1–4.7)
Mortality	29 (9)	7 (8)	21 (6)	6 (7)	12 (8)	17 (18)
DCI	70 (21)	18 (22)	67 (22)	20 (25)	33 (22)	25 (28)
HH	3 (2–3)	3 (2–3)	3 (2–3)	3 (2–3)	2 (2–3)	3 (2–3)
mFS score 3	181 (69)	43 (65)	174 (69)	44 (65)	78 (65)	54 (69)
IVH	187 (67)	42 (60)	179 (65)	46 (67)	81 (66)	56 (69)

Abbreviations: HH = Hunt-Hess scale; IVH = intraventricular hemorrhage; mFS = modified Fisher Scale; mRS = modified Rankin Scale.
Values are median (IQR) or n (%).

variables (excluding the HH score) had an AUC 0.81 ± 0.05 (95% CI 0.71–0.89) (figure 2C).

Analysis of DCI and Discharge Outcome Models

After developing the ML models, we attempted to interpret the ML models. The ML model that had the highest AUC for predicting DCI used 31 derived variables from the EMR. The variables included age and laboratory test results such as hemoglobin, sodium, WBC, platelets, and creatinine (figure 2B). Since ANN models are difficult to interpret, we employed analysis using the gradient boost (GB) and random forest (RF) model,[26] both tree-based models that are relatively less challenging to interpret. We trained the GB model to predict DCI using only these 31 variables and ranked these variables based on relative importance. The top 20% of the ranked variables included sodium, WBC, and neutrophils. The ML model for discharge mRS prediction included 8 derived variables including glucose, segmented neutrophil levels, variations in systolic blood pressure, WBC, hematocrit levels, and lymphocytes (figure 2D). This model also included age and HH score. The best performing ML model was obtained by using a combination of EMR variables and the clinician-derived HH score.

Performance of ML, Standard Model, and Clinician Prognostication in Predicting 3-Month Mutcomes

Both ML models previously developed to predict discharge outcome (one model that included only EMR and another that included both EMR and HH score) were retrained to predict 3-month outcomes. For 3-month outcome prediction, the AUC of the ML model that included only EMR variables (0.89 ± 0.03, 95% CI 0.8–0.94) and the AUC of the ML model that included both EMR variables and HH score (0.893 ± 0.03, 95% CI 0.81–0.94) were significantly ($p < 0.05$) higher than the AUC of the standard model (0.75 ± 0.06, 95% CI 0.65–0.83), The ML model that included the EMR and HH variables resulted in a 0.13 ± 0.05 (95% CI 0.038–0.24) improvement in the AUC over the standard model (figure 3).

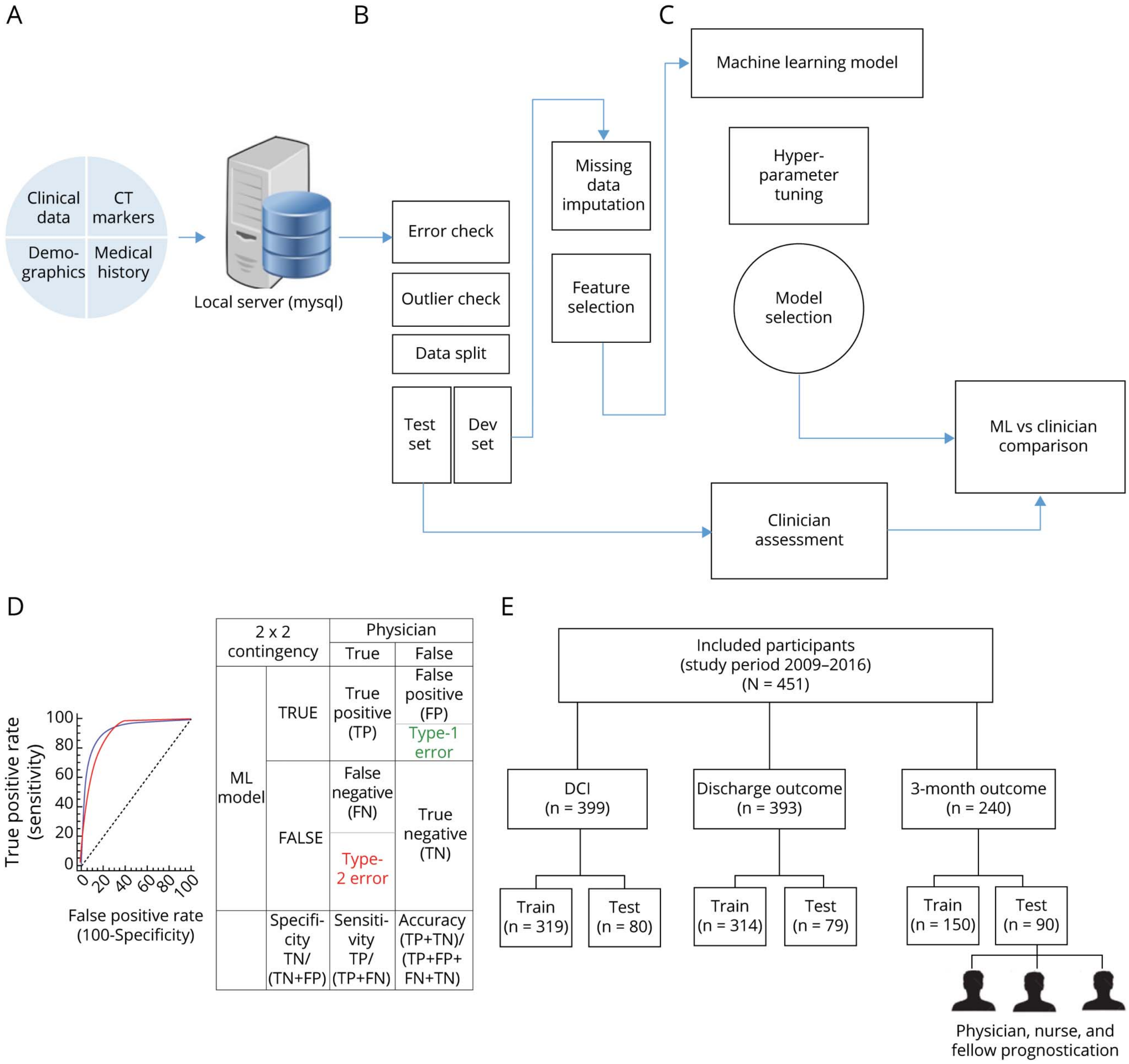
Among clinical team members, the attending physician (sensitivity 0.88, specificity 0.95) outperformed the nurse (0.86 and 0.85) and fellow (0.81 and 0.75). The sensitivity and specificity of the ML model (at the optimal threshold) was 0.91 and 0.64, respectively. The attending physician's prognostication was numerically higher than the ML model, but the difference in performance was not statistically significant ($p > 0.05$, McNemar test).

Discussion

There are 3 main findings in this study. ML models predict DCI and functional outcomes better than standard models. They match physician's performance in predicting 3-month outcomes. ML models that include variables derived from clinician insight are marginally better than ML models that exclude them.

ML results in a 36% improvement in AUC over the SM in predicting DCI. Except for the mFS, the variables used by the ML model are routinely available, potentially allowing for

Figure 1 Architectural Overview of Data Extraction, Preprocessing, Machine Learning (ML) Model Development, and Performance Comparison

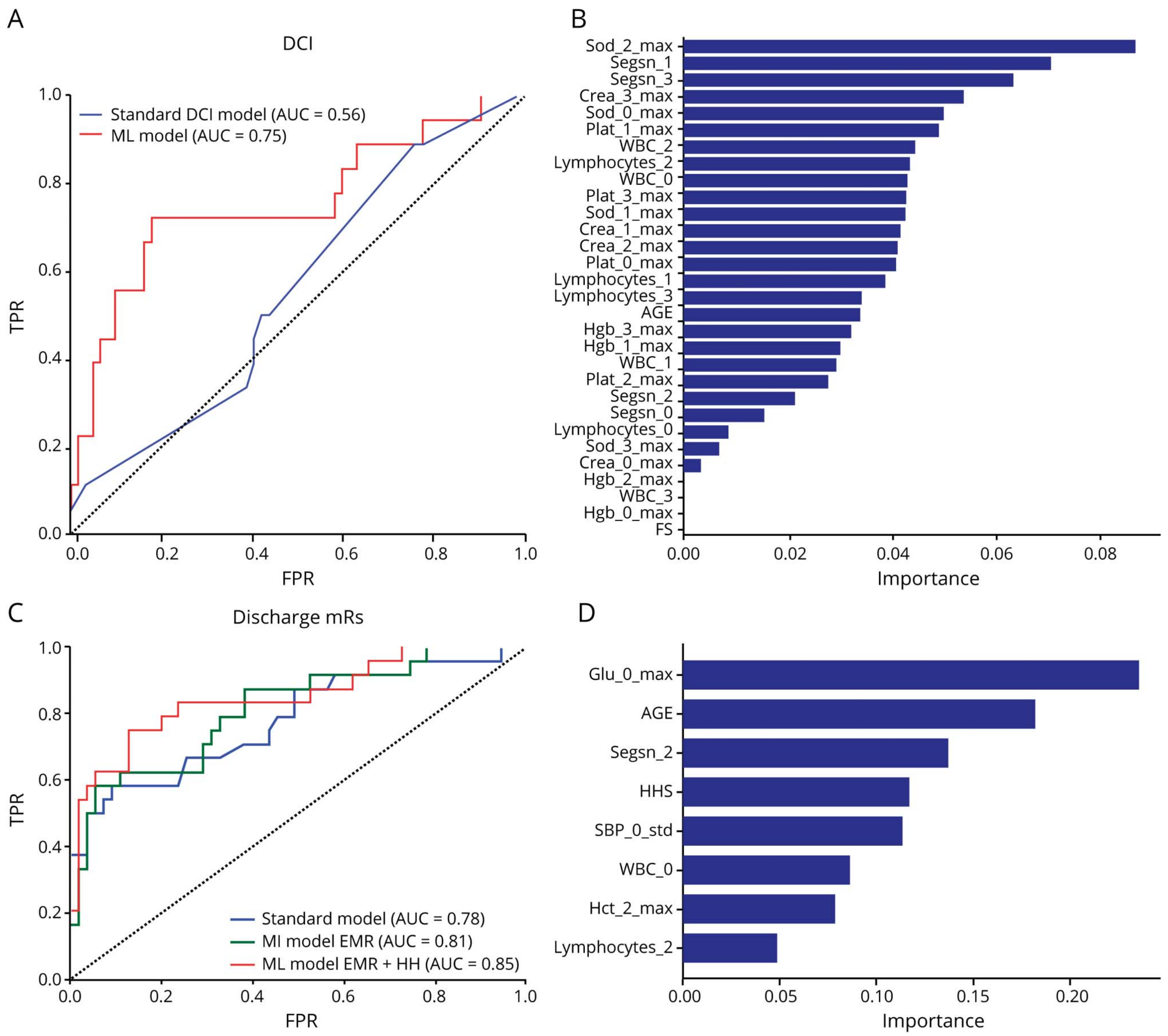


2 x 2 contingency		Physician	
		True	False
ML model	TRUE	True positive (TP)	False positive (FP) Type-1 error
	FALSE	False negative (FN) Type-2 error	True negative (TN)
	Specificity TN/(TN+FP)	Sensitivity TP/(TP+FN)	Accuracy (TP+TN)/(TP+FP+FN+TN)

(A) Data (static and dynamic) extraction from the electronic medical record (EMR). Data from the EMR is stored in a local server and queried. (B) Data quality is checked for errors, outliers, and split into a training set and test set followed by missing data imputation. (C) The development set was used to train and develop a ML model. (D) The test set is used to evaluate the performance of the ML model, standard clinician, and physician. The performances are statistically compared. (E) Overview of participant allocation for training of the ML models and testing. DCI = delayed cerebral ischemia.

easier integration into existing EMR systems. The variables identified in the DCI prediction ML model including sodium levels, hemoglobin, WBC, and segmented neutrophils are relevant in the pathophysiologic mechanisms of DCI. For instance, sodium imbalances can be attributed to SAH-induced release of natriuretic peptides[17] linked to DCI.[27,28] The WBCs ML model can be attributed to inflammatory response after SAH,[29–31] which has been shown to be associated with poor outcomes and DCI.[11,32–34] Lymphocytic infiltrators in aneurysm walls are associated with cerebral vasospasm, a precursor to DCI, suggesting a direct mechanism between cerebral vasospasm/DCI and WBC level.[35] Serum neutrophils[12] and erythroid abnormalities (including hemoglobin and hematocrit levels) are linked to DCI.[14] The superior AUC of the ANN/ML is likely due to its ability to leverage the combined predictive value of these variables. Previous implication of the identified variables in DCI literature adds credence to the physiologic plausibility of the model. It is noteworthy that some ML models, using continuous physiologic data, predict DCI with an AUC of 0.77.[36]

Figure 2 Evaluation of Machine Learning (ML) Model and Standard Model Performance on the Same Test/Evaluation Set



(A) Comparison of receiver operating characteristic (ROC) curves of the standard delayed cerebral ischemia (DCI) prediction model with the ML model. The area under the ROC curve (AUC) of the ML model (0.75 ± 0.07, 95% confidence interval [CI] 0.65 to 0.84) was higher (0.19 ± 0.11, 95% CI −0.02 to 0.42, $p = 0.08$, DeLong test) than the AUC of the standard model (0.56 ± 0.07, 95% CI 0.44 to 0.66). The standard model is a logistic regression model that included age and the modified Fisher score. The ML model included the modified Fisher score and other variables including white blood cells (WBC), neutrophils, lymphocytes, platelets, creatinine, sodium, and hemoglobin. (B) Variables used in the ML model are ranked by the gradient boost model based on their importance. Among the 31 variables included in the ML model, variables pertaining to serum sodium, neutrophil, creatinine, and WBC count levels during the pre-DCI phase were ranked most important in predicting impending DCI. (C) Comparison of the standard model and 2 ML models (one that included only electronic medical record [EMR] variables and another that included both EMR variables and the Hunt-Hess scale [HH] score). The standard model is a logistic regression model that included age and HH score. The AUC of the ML model that included only EMR variables (0.81 ± 0.05, 95% CI 0.71 to 0.89) was higher than the AUC of the standard model (0.78 ± 0.06, 95% CI 0.67 to 0.86), but the differences were not statistically significant ($p = 0.6$). The AUC of the ML model that included both the EMR measures and the HH score (0.85 ± 0.05, 95% CI 0.75 to 0.92) was significantly higher than the AUC of the standard model ($p = 0.05$, DeLong test) (D) Variables in the best ML model were ranked using the random forest model to identify relative importance of variables in prognosticating discharge outcomes. The variables are named with the prefix of the EMR measure (e.g., Sod = sodium), followed by the day in which it was recorded and then the aggregation (max, min, SD, or average). For instance, Sod_3_Max denotes the maximum value of sodium levels at day 3 after admission. Besides age and clinical severity, glucose levels at admission, segmented neutrophil count, and variations in systolic blood pressure were ranked among the top variables used by the model. mRS = modified Rankin Scale; SBP = systolic blood pressure.

ML models can make systemic inferences, as opposed to reductionist approaches of traditional approaches. Among several reported risk factors for DCI, with the exception of the mFS, none has translated into a practical tool. For instance, a proposed WBC threshold of 12.1 × 10^9 cells/L has an AUC of 0.63 (only slightly better than the AUC of the mFS score ~0.57) and hence nongeneralizable.[37] Other approaches to predict DCI using different modalities have been proposed. A cEEG-based approach reports DCI prediction 24 hours prior to onset with high sensitivity and specificity. Transcranial Doppler (TCD) and CT angiography/CT perfusion have been used for DCI prognostication but with limitations. TCDs only achieve

Table 2 Summary of Model Performances on the Test Set

	Accuracy	Sensitivity	Specificity	AUC
DCI				
Standard model	0.55	0.79	0.25	0.55
ML model	0.8	0.82	0.72	0.75
Discharge mRS				
Standard model	0.81	0.58	0.90	0.78
ML EMR	0.81	0.58	0.94	0.81
ML EMR + HH	0.84	0.75	0.87	0.85
3-month mRS				
Standard model	0.75	0.90	0.56	0.75
ML EMR	0.84	0.92	0.69	0.89
ML EMR + HH	0.81	0.91	0.64	0.89
Attending	0.89	0.88	0.95	NA
Nurse	0.86	0.86	0.85	NA
Fellow	0.80	0.81	0.75	NA

Abbreviations: AUC = area under the receiver operating characteristic curve; DCI = delayed cerebral ischemia; EMR = electronic medical record; HH = Hunt-Hess scale; ML = machine learning; mRS = modified Rankin Scale score; NA = not available.

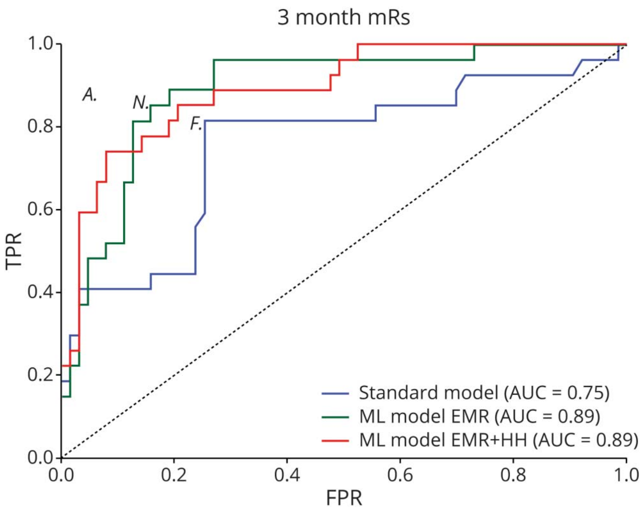
reasonable sensitivity and specificity on day 8 of SAH, which is too late for early risk stratification and intervention.[39] CT angiography or CT perfusion require iodinated contrast injections, additional exposure to radiation for patients, and they lack the sensitivity and specificity to be used in routine clinical practice.[40] importantly, these methods are surrogate markers of cerebral vessel narrowing (cerebral vasospasm), which is only one of the many processes important to the development of DCI. Unlike EEG and TCD, the ML approach uses only routine clinical variables (which are already available as part of standard care); it avoids the need for expensive instrumentation (as in the case of EEG and TCD). The output of the ML model (which is simply a probabilistic risk score between 0 and 1) is easily translatable in most settings because, unlike EEG and TCD, expertise of trained technician is not required. ML utilizes hundreds of variables in tandem. ML augments the precision medicine paradigm, wherein patient-specific risk can be evaluated. To verify whether the ML model's high AUC is attributed to model complexity (rather than the choice of variables alone), a logistic regression model was trained using all 31 variables and its performance evaluated on the test set. The AUC of this model was 0.65, which was 13% less than the AUC of the ML model, implying that high AUC of the ML can be attributed to its complex model architecture.

The AUCs of the ML models were 9% and 18% higher than the standard models in predicting discharge and 3-month functional outcomes, and both models included only 8 variables (of which only one—the HH score—was derived by physician examination). Even ML models that only relied on EMR variables (blinded to the HH score) showed 4% and 18% increase in AUC over standard models. Variables in the ML model include glucose, WBC, and variations in systolic blood pressure (markers of systemic inflammation), which were previously linked to poor outcomes in SAH. Impaired glucose metabolism is frequent after SAH,[41] likely a stress response due to disruption in the metabolic processes and inhibition of insulin release via the sympathetic noradrenergic nerves[42]; it exacerbates neurologic injury[43] and contributes to poor outcomes.[44] Early systemic inflammatory processes have been shown to be associated with poor outcomes and ML models likely leverage the combined predictive value of variables linked to these processes to predict outcomes with high AUCs.

Among clinicians, compared to the nurse and fellow, the attending physician had the highest sensitivity and specificity scores (0.88 and 0.95) for predicting outcome (table 2). The sensitivity and specificity of the ML model (0.91 and 0.64) did not significantly differ from that of the physician's assessment. The demonstrated sensitivity and high specificity of the physician in prognosticating 3-month outcome highlights the importance of factors that influence human decision-making processes, which are difficult to be objectively ascertained. These factors are likely to have a bigger effect on long-term outcomes than EMR variables collected <3 days of admission.

It is important to contrast the differences in clinician and ML predictions. Physicians consider a more holistic set of information in their decision-making as compared to the ML model, which has information limited to clinical and

Figure 3 Evaluation of 3-Month Outcome Prediction by Machine Learning (ML) Models



(A) Comparison of the standard model, a ML model that included only electronic medical record (EMR) variables, and another ML model that included both EMR variables and the Hunt-Hess scale (HH) score. The standard model is a logistic regression model that included age and HH score. The area under the receiver operating characteristic curve (AUC) of the ML model that included only EMR variables (0.89 ± 0.03, 95% confidence interval [CI] 0.8 to 0.94) and the AUC of the ML model that included EMR variables and HH score (0.89 ± 0.03, 95% CI 0.81 to 0.94) was significantly ($p < 0.05$) higher than the AUC of the standard model (0.75 ± 0.06, 95% CI 0.65 to 0.83). The AUCs of both the ML models were not statistically different ($p = 0.93$) (B) The FPR vs TPR of attendings, nurses, and fellows are shown (indicated by A, N, and F in the receiver operating characteristic plot at the locations [0.05, 0.88], [0.15, 0.86], and [0.25, 0.81], respectively). Among the clinicians, the assessment of the attending was most accurate (sensitivity of 0.88, specificity of 0.95). The best performing ML model (ML EMR + HH) had a sensitivity of 0.91 and specificity of 0.64. The 2 assessments (clinician and ML model) are compared (difference = −9.09%, 95% CI −18.35 to 0.17, $p = 0.09$, McNemar test). mRS = modified Rankin Scale.

laboratory measures during the first 3 days of admission. The developed ML models are blinded to any information about preexisting functional status of the patient. However, physician assessments are aided by past medical history and examination of the patient. They possess an intuition from years of treating patients via which they have developed their aptitude (albeit subjective) to prognosticate outcomes. Even experienced physicians are limited by inability to process several variables at the same time.[45] However, the actual outcome/impending complication (like DCI) can depend on several hundred variables. ML can learn objectively and can handle numerous variables. Variability in physician-to-physician assessment—due to experience, education, and hospital protocols—is also absent or minimized in ML models. Other issues, like time pressure, cognitive biases, fatigue, information load, and behavioral confirmation effects,[46] are less of a concern in ML. ML can offer unique perspective on the patient's condition and can serve as a decision support tool in the management of SAH. However, clinical judgement is necessary to interpret the ML results and implement a corresponding plan of action.

First, reproducibility in ML studies is a concern,[47] wherein performance measures observed in one cohort may not be generalizable to others. This may be due to overfitting and variations in treatment protocols at different centers. To minimize overfitting, we used a 10-fold CV training approach and report performance measures on a separate test set that was not used in the training process. The difference in the 10-fold AUC of the DCI prediction ML model on the training and testing data is small, indicating that the model was not overfitting. To minimize effects due to varied treatment protocols, only data from within a few days of admission were included in the ML model, as this acute period is more likely to reflect a patient's physiologic status and is less influenced by variations in the interventions following admission due to the short time span. Second, even though our models were trained with hundreds of patients, this is still a relatively small dataset in comparison with recent ML studies, which include thousands of patients. All other factors unperturbed, ML models trained with larger samples have better performance.[48] Though we included data from a 7-year period, the incidence of SAH is low and larger sample sizes are challenging. Third, the clinicians adjudicating DCI were unblinded to the clinical case. While DCI was adjudicated prospectively, the EMR data were queried retrospectively and the informatician developing the ML model was blinded to the clinical case. Fourth, due to the large number of participants included and variables included, missing data were prevalent. We removed all variables that were missing for >30% of the participants and used diligent imputation strategies and the model is not sensitive to missing values. Finally, it is challenging to interpret the ML models. This challenge—termed a "black box" problem[49]—is a problem in other domains as well. Though we employed the RF analysis to rank the variables, a thorough understanding of the decision-making process of ML could further our understanding of the disease process. Nevertheless, successful identification of impending DCI has great potential to improve SAH management.

ML improves prediction of DCI and functional outcomes compared to standard models. It matches attending physicians' performance in predicting 3-month outcomes. Its performance must be evaluated in patient cohorts from other centers. In the future, the model can be expanded to include other variables including imaging and specimen biomarkers to improve performance.

Acknowledgment

The investigators thank the Neuroscience Research Repository for assistance with obtaining the specimens and the CCTS pilot program and Dr. Roxanne George for her contributions and ideas in drafting the manuscript.

Study Funding

No targeted funding reported.

Disclosure

The authors report no disclosures relevant to the manuscript. Go to Neurology.org/N for full disclosures.

Publication History

Received by *Neurology* August 26, 2019. Accepted in final form September 21, 2020.

Appendix Authors

Name	Location	Contribution
Jude P.J. Savarraj, PhD	Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Georgene W. Hergenroeder, PhD	Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Liang Zhu, PhD	Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Tiffany Chang, MD	Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Soojin Park, MD	Department of Neurology, Columbia University, New York, NY	Study concept and design, critical revision of manuscript for intellectual content
Murad Megjhani, PhD	Department of Neurology, Columbia University, New York, NY	Study concept and design, critical revision of manuscript for intellectual content
Farhaan S. Vahidy, PhD	Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Zhongming Zhao, PhD	Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston	Study concept and design, critical revision of manuscript for intellectual content
Ryan S. Kitagawa, MD	Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center	Study concept and design, critical revision of manuscript for intellectual content
H. Alex Choi, MD	Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center	Study concept and design, critical revision of manuscript for intellectual content

References

1. Frontera JA, Claassen J, Schmidt JM, et al. Prediction of symptomatic vasospasm after subarachnoid hemorrhage: the modified Fisher scale. Neurosurgery 2006;59:21–27.
2. Adams HP. 21: Clinical scales to assess patients with stroke. In: Grotta JC, Albers GW, Broderick JP, et al, eds. Stroke (Sixth Edition) [online]. London: Elsevier; 2016: 308–325.e6. Available at: sciencedirect.com/science/article/pii/B9780323295444000219. Accessed March 6, 2019.
3. Naval NS, Kowalski RG, Chang TR, Caserta F, Carhuapoma JR, Tamargo RJ. The SAH Score: a comprehensive communication tool. J Stroke Cerebrovasc Dis 2014;23: 902–909.
4. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interf 2018;15:20170387.
5. Deo RC. Machine learning in medicine. Circulation 2015;132:1920–1930.
6. Machine Learning in Healthcare [online]. Available at: nature.com/collections/zbkpvddmhm. Accessed December 21, 2018.
7. Artificial Intelligence in Healthcare: Past, Present and Future: Stroke and Vascular Neurology [online]. Available at: svn.bmj.com/content/2/4/230. Accessed December 21, 2018.
8. Machine Learning and its Applications: A Review: IEEE Conference Publication [online]. Available at: ieeexplore.ieee.org/document/8070809. Accessed December 21, 2018.
9. Rothoerl RD, Axmann C, Pina A-L, Woertgen C, Brawanski A. Possible role of the C-reactive protein and white blood cell count in the pathogenesis of cerebral vasospasm following aneurysmal subarachnoid hemorrhage. J Neurosurg Anesthesiol 2006;18:68–72.
10. Kasius KM, Frijns CJM, Algra A, Rinkel GJE. Association of platelet and leukocyte counts with delayed cerebral ischemia in aneurysmal subarachnoid hemorrhage. Cerebrovasc Dis 2010;29:576–583.
11. McGirt MJ, Mavropoulos JC, McGirt LY, et al. Leukocytosis as an independent risk factor for cerebral vasospasm following aneurysmal subarachnoid hemorrhage. J Neurosurg 2003;98:1222–1226.
12. Tao C, Wang J, Hu X, Ma J, Li H, You C. Clinical value of neutrophil to lymphocyte and platelet to lymphocyte ratio after aneurysmal subarachnoid hemorrhage. Neurocrit Care 2017;26:393–401.
13. Hirashima Y, Hamada H, Kurimoto M, Origasa H, Endo S. Decrease in platelet count as an independent risk factor for symptomatic vasospasm following aneurysmal subarachnoid hemorrhage. J Neurosurg 2005;102:882–887.
14. Da Silva IRF, Gomes JA, Wachsman A, de Freitas GR, Provencio JJ. Hematologic counts as predictors of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage. J Crit Care 2017;37:126–129.
15. Nina P, Schisano G, Chiappetta F, et al. A study of blood coagulation and fibrinolytic system in spontaneous subarachnoid hemorrhage: correlation with Hunt-Hess grade and outcome. Surg Neurol 2001;55:197–203.
16. Feng W, Tauhid S, Goel S, Sidorov EV, Selim M. Hyperglycemia and outcome in intracerebral hemorrhage: from bedside to bench-more study is needed. Transl Stroke Res 2012;3:113–118.
17. Hannon MJ, Finucane FM, Sherlock M, Agha A, Thompson CJ. Disorders of water homeostasis in neurosurgical patients. J Clin Endocrinol Metab 2012;97:1423–1433.
18. Naredi S, Lambert G, Edén E, et al. Increased sympathetic nervous activity in patients with nontraumatic subarachnoid hemorrhage. Stroke 2000;31:901–906.
19. Roederer A, Holmes JH, Smith MJ, Lee I, Park S. Prediction of significant vasospasm in aneurysmal subarachnoid hemorrhage using automated data. Neurocrit Care 2014; 21:444–450.
20. Hunt WE, Hess RM. Surgical risk as related to time of intervention in the repair of intracranial aneurysms. J Neurosurg 1968;28:14–20.
21. Vergouwen Mervyn DI, Marinus V, Jan VG, et al. Definition of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage as an outcome event in clinical trials and observational studies. Stroke 2010;41:2391–2395.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.
23. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J 2008;50:419–430.
24. Trajman A, Luiz RR. McNemar chi2 test revisited: comparing sensitivity and specificity of diagnostic examinations. Scand J Clin Lab Invest 2008;68:77–80.
25. MedCalc [online]. Available at: medcalc.org.
26. Breiman L. Random forests. Mach Learn 2001;45:5–32.
27. Sviri GE, Feinsod M, Soustiel JF. Brain natriuretic peptide and cerebral vasospasm in subarachnoid hemorrhage. Clinical and TCD correlations. Stroke 2000;31:118–122.
28. Qureshi AI, Suri MFK, Sung GY, et al. Prognostic significance of hypernatremia and hyponatremia among patients with aneurysmal subarachnoid hemorrhage. Neurosurgery 2002;50:749–755; discussion 755–756.
29. Savarraj J, Parsha K, Hergenroeder G, et al. Early brain injury associated with systemic inflammation after subarachnoid hemorrhage. Neurocrit Care 2018;28:203–211.
30. Savarraj JP, McGuire MF, Parsha K, et al. Disruption of thrombo-inflammatory response and activation of a distinct cytokine cluster after subarachnoid hemorrhage. Cytokine 2018;111:334–341.
31. Ahn SH, Savarraj JP, Pervez M, et al. The subarachnoid hemorrhage early brain edema score predicts delayed cerebral ischemia and clinical outcomes. Neurosurgery 2018; 83:137–145.
32. Spallone A, Acqui M, Pastore FS, Guidetti B. Relationship between leukocytosis and ischemic complications following aneurysmal subarachnoid hemorrhage. Surg Neurol 1987;27:253–258.
33. Neil-Dwyer G, Cruickshank J. The blood leucocyte count and its prognostic significance in subarachnoid haemorrhage. Brain 1974;97:79–86.
34. Choi HA, Bajgur SS, Jones WH, et al. Quantification of cerebral edema after subarachnoid hemorrhage. Neurocrit Care 2016;25:64–70.
35. Holling M, Jeibmann A, Gerss J, et al. Prognostic value of histopathological findings in aneurysmal subarachnoid hemorrhage. J Neurosurg 2009;110:487–491.
36. Park S, Megjhani M, Frey HP, et al. Predicting delayed cerebral ischemia after subarachnoid hemorrhage using physiological time series data. J Clin Monit Comput 2019;33:95–105.
37. Al-Mufti F, Misiolek KA, Roh D, et al. White blood cell count improves prediction of delayed cerebral ischemia following aneurysmal subarachnoid hemorrhage. Neurosurgery 2019;84:397–403.
38. Rosenthal ES, Biswal S, Zafar SF, et al. Continuous electroencephalography predicts delayed cerebral ischemia after subarachnoid hemorrhage: a prospective study of diagnostic accuracy. Ann Neurol 2018;83:958–969.

39. Carrera E, Schmidt JM, Oddo M, et al. Transcranial Doppler for predicting delayed cerebral ischemia after subarachnoid hemorrhage. Neurosurgery 2009;65:316–323; discussion 323–324.

40. Cremers CHP, van der Schaaf IC, Wensink E, et al. CT perfusion and delayed cerebral ischemia in aneurysmal subarachnoid hemorrhage: a systematic review and meta-analysis. J Cereb Blood Flow Metab 2014;34:200–207.

41. Dorhout Mees SM, van Dijk GW, Algra A, Kempink DRJ, Rinkel GJE. Glucose levels and outcome after subarachnoid hemorrhage. Neurology 2003;61:1132–1133.

42. Järhult J, Falck B, Ingemansson S, Nobin A. The functional importance of sympathetic nerves to the liver and endocrine pancreas. Ann Surg 1979;189:96–100.

43. Kumari S, Anderson L, Farmer S, Mehta SL, Li PA. Hyperglycemia alters mitochondrial fission and fusion proteins in mice subjected to cerebral ischemia and reperfusion. Transl Stroke Res 2012;3:296–304.

44. Badjatia N, Topcuoglu MA, Buonanno FS, et al. Relationship between hyperglycemia and symptomatic vasospasm after subarachnoid hemorrhage. Crit Care Med 2005;33:1603.

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

46. Burgess DJ, Fu SS, van Ryn M. Why do providers contribute to disparities and what can be done about it? J Gen Intern Med 2004;19:1154–1159.

47. Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018;359:725–726.

48. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009;24:8–12.

49. Medicine TLR. Opening the black box of machine learning. Lancet Respir Med 2018;6:801.