Stroke

ORIGINAL ARTICLE

Deep Learning Versus Neurologists: Functional Outcome Prediction in LVO Stroke Patients Undergoing Mechanical Thrombectomy

Lisa Herzog, Dr sc. nat.; Lucas Kook, Dr. sc. nat.; Janne Hamann, Dr med; Christoph Globas, PD Dr med; Mirjam R. Heldner, Dr med; David Seiffge, Dr med; Kateryna Antonenko, PD Dr med; Tomas Dobrocky, Dr med; Leonidas Panos, Dr med; Johannes Kaesmacher, PD Dr med; Urs Fischer, Prof Dr med; Jan Gralla, Dr med; Marcel Arnold, Dr med; Roland Wiest, Dr med; Andreas R. Luft, Dr med; Beate Sick, Dr; Susanne Wegener, Dr med

BACKGROUND: Despite evolving treatments, functional recovery in patients with large vessel occlusion stroke remains variable and outcome prediction challenging. Can we improve estimation of functional outcome with interpretable deep learning models using clinical and magnetic resonance imaging data?

METHODS: In this observational study, we collected data of 222 patients with middle cerebral artery M1 segment occlusion who received mechanical thrombectomy. In a 5-fold cross validation, we evaluated interpretable deep learning models for predicting functional outcome in terms of modified Rankin scale at 3 months using clinical variables, diffusion weighted imaging and perfusion weighted imaging, and a combination thereof. Based on 50 test patients, we compared model performances to those of 5 experienced stroke neurologists. Prediction performance for ordinal (modified Rankin scale score, 0–6) and binary (modified Rankin scale score, 0–2 versus 3–6) functional outcome was assessed using discrimination and calibration measures like area under the receiver operating characteristic curve and accuracy (percentage of correctly classified patients).

RESULTS: In the cross validation, the model based on clinical variables and diffusion weighted imaging achieved the highest binary prediction performance (area under the receiver operating characteristic curve, 0.766 [0.727–0.803]). Performance of models using clinical variables or diffusion weighted imaging only was lower. Adding perfusion weighted imaging did not improve outcome prediction. On the test set of 50 patients, binary prediction performance between model (accuracy, 60% [55.4%–64.4%]) and neurologists (accuracy, 60% [55.8%–64.21%]) was similar when using clinical data. However, models significantly outperformed neurologists when imaging data were provided, alone or in combination with clinical variables (accuracy, 72% [67.8%–76%] versus 64% [59.8%–68.4%] with clinical and imaging data). Prediction performance of neurologists with comparable experience varied strongly.

CONCLUSIONS: We hypothesize that early prediction of functional outcome in large vessel occlusion stroke patients may be significantly improved if neurologists are supported by interpretable deep learning models.

GRAPHIC ABSTRACT: A graphic abstract is available for this article.

Key Words: machine learning ■ outcome prediction ■ stroke

Large vessel occlusion (LVO) stroke treatment has rapidly evolved. Mechanical thrombectomy (MT) for patients with stroke due to LVO has significantly increased the number of patients with favorable outcome.[1,2] However, about 50% of patients treated with MT still experience major functional disability or death after

Correspondence to: Susanne Wegener, Prof Dr med., Department of Neurology, University Hospital Zurich, Frauenklinikstrasse 26, 8091 Zurich, Switzerland. Email susanne.wegener@usz.ch

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.123.042496.
For Sources of Funding and Disclosures, see page xxx.

Stroke is available at www.ahajournals.org/journal/str

ORIGINAL ARTICLE

Nonstandard Abbreviations and Acronyms

CV	cross validation
DL	deep learning
DWI	diffusion weighted imaging
LVO	large vessel occlusion
MCA	middle cerebral artery
MRI	magnetic resonance imaging
mRS	modified Rankin scale
MT	mechanical thrombectomy
NIHSS	National Institutes of Health Stroke Scale
PWI	perfusion weighted imaging
TMAX	time to maximum

stroke.[3,4] In their daily work, neurologists estimate functional outcome, particularly when therapeutic decisions are challenging. In addition, they disclose the expected functional outcome early to patients and their families, who want to know the patient's chance for functional recovery versus risk for lifelong disability or death.

In routine clinical practice, neurologists use a combination of clinical variables and brain scans from computed tomography or magnetic resonance imaging (MRI) for functional outcome prediction. Currently, treatment decisions and early prognoses in patients with LVO stroke are heavily based on diffusion weighted imaging (DWI) and perfusion weighted imaging (PWI).[4] A small infarct core and large tissue at risk were previously shown to be associated with favorable functional outcome, which is why these imaging features are used to select patients for MT.[5,6] Yet, more recent trials indicated that patients with large infarct core or small tissue at risk may benefit from MT as well.[7,8] Consequently, the benefit of considering acute stroke MRI in addition to clinical variables for functional outcome prediction is not clear.

Generally, functional outcome prediction in acute stroke patients is difficult. Practicing physicians perform poorly in estimating functional disability at discharge[9] and after 6 months[10] when being provided with established clinical variables. Different scores like the DRAGON[11] or ASTRAL (Acute Stroke Registry and Analysis of Lausanne)[12] were developed to support functional outcome prediction and have shown to outperform physicians.[13] However, they are rarely used in clinical practice, mainly because they ignore imaging data.[14,15] Machine learning models for predicting functional outcome using imaging data and combinations of imaging and clinical data were implemented and shown to achieve similar or higher prediction performances than the scores.[16] Particularly deep learning (DL) models, which learn relevant imaging features, have achieved outstanding results on image analysis tasks.[17] However, DL models for functional outcome based on DWI and PWI besides clinical variables are lacking. Moreover, although imaging data and clinical variables can easily be integrated in one DL model,[15] such models usually lack interpretability, preventing neurologists from trusting the results. To summarize, a clinically suitable functional outcome prediction model is expected to provide an excellent prediction performance, comparable to or better than expert neurologists, integrating clinical variables, and imaging data but at the same time, being transparent for neurologists to relate results to their expert knowledge.[1]

The goal of this study was to develop a reliable prediction model for functional outcome at the acute pretreatment stage and to evaluate clinical variables and MRI data regarding their value for indicating functional outcome in patients with LVO treated with MT. In a previous study, we found that outcome prediction in those patients was not improved when using extracted core and mismatch imaging features in addition to clinical variables.[16] However, we assumed that DWI and PWI are highly relevant for functional outcome prediction, which is why we here used a novel class of DL models, learning relevant imaging features, and yielding interpretable parameter estimates for clinical variables, in addition to high prediction performances. Furthermore, we assessed outcome predictions of neurologists based on clinical variables, MRI, and a combination thereof and compared them to the model predictions using a subset of 50 patients.

METHODS

Data Availability

The corresponding author had full access to all the data in the study and takes responsibility for its integrity and data analysis. Investigators may request access to anonymized individual patient data. Before use of the data, proposals need to be approved by an independent review panel at Inselspital Bern. A signed data sharing agreement will then be approved.

Cohort Description

We retrospectively collected data from LVO stroke patients hospitalized between January 2012 and August 2017 at Inselspital Bern, Switzerland. All patients had middle cerebral artery (MCA) M1-segment infarction based on MR angiography and underwent MT. All patients received acute stroke MRI including DWI/PWI. Decision for MT required detecting an LVO and a relevant clinical deficit. Mismatch or core size was usually not considered for treatment decisions within a time window of 6 hours from symptom onset, according to international guidelines.[18] Patients were treated with either Solitaire AB, Solitaire stent retriever, or, rarely, with aspiration catheters only. Generally, the first-line approach during the study period was stent retriever–based thrombectomy using the Solitaire Flow Restoration device and a balloon guide catheter whenever possible.[19] Recanalization success was scored by experienced interventional neuroradiologists from post-MT angiogram according to the TICI scale, with TICI score 2b-3 considered successful recanalization. All patients had the same access to state-of-the-art rehabilitation after stroke.

Patients with missing outcome (modified Rankin scale [mRS] score at 3 months), missing imaging data, or imaging data of insufficient quality were excluded. Furthermore, we excluded patients with previous territorial infarction evident on MRI or with additional vessel occlusions other than clot extension to the internal carotid artery or distal MCA branches revealed by angiography. Data from this cohort were previously analyzed in the study by Hamann et al.[16] For functional outcome prediction, we only considered clinical variables and imaging data available before treatment. Clinical variables included patient characteristics, risk factors, prior medication, clinical scores, and other stroke-related information (Table 1). Imaging data included DWI and PWI with calculated maps of cerebral blood flow, cerebral blood volume, and time to maximum (TMAX) >6 s. Detailed information about the imaging protocol can be found in the study by Hamann et al.[16]

Standard Protocol Approvals and Patient Consent

The methods were performed in accordance with relevant guidelines and regulations and approved by the cantonal Ethics Commission of the Canton of Bern (No. 231/14). Written informed consent was available for all participants.

Descriptive Statistics

Descriptive statistics were used to summarize the patient cohort. Frequencies and percentages and medians and interquartile ranges were calculated to describe categorical and continuous clinical variables, respectively. Measures are reported for all patients and for patients with favorable (mRS score 0–2) and unfavorable (mRS score 3–6) outcomes.

Functional Outcome Prediction Models

We developed interpretable DL models based on (1) clinical variables, (2) MRI, and (3) a combination thereof to predict mRS at 3 months. The models belong to a novel class of interpretable (deep) neural networks, enabling analyzing combinations of unstructured data like images and structured tabular data such as clinical variables.[20] The models provide probabilities for the 7 mRS classes and parameter estimates for clinical variables like odds ratios quantifying the effect of those variables on the outcome. As input data, we used clinical variables and imaging data considered as most important by expert stroke neurologists. Clinical variables were age, systolic blood pressure, diabetes, hypertension, smoking, prior stroke, functional independence before stroke (mRS score 0–2), admission National Institutes of Health Stroke Scale (NIHSS), treatment with intravenous thrombolysis, and time to groin puncture. Imaging data included DWI and TMAX perfusion maps.

DL models were evaluated in a 5-fold cross validation (CV) to account for the rather small sample size of 222 patients (Figure S1). Therefore, we split the data into 5 test sets of equal size. Data not contained in the respective test set were used for training and validation (80:20 ratio). In each fold, we trained 5 randomly initialized versions of the model to consider uncertainty in model parameters and potentially improve prediction performance.[21] The predictions of the 5 folds were then averaged to 1 final prediction. We imputed missing values of clinical variables (Table 1) using missForest.[22] Missing variables in the test data were imputed based on the imputed training data sets.

Table 1. Descriptive Statistics for Baseline Clinical Variables

	All (n=222)	Favorable (n=119)	Unfavorable (n=103)
Demographics			
Age, y	73.54 (20)	68.80 (18.2)	79.20 (17.7)
Sex (female)	134 (60.4%)	72 (60.5%)	62 (60.2%)
Risk factors			
Diabetes	39 (17.6%)	15 (12.6%)	24 (23.3%)
Atrial fibrillation (n=1)	92 (41.6%)	39 (33.1%)	53 (51.4%)
Hypercholesterolemia (n=2)	133 (60.5%)	75 (63%)	58 (57.4%)
Hypertension	148 (66.7%)	69 (58%)	79 (76.7%)
Smoker (n=20)	49 (24.3%)	35 (31%)	14 (15.7%)
CHD	35 (15.8%)	19 (16%)	16 (15.5%)
Peripheral artery disease (n=19)	8 (3.9%)	3 (2.7%)	5 (5.4%)
Prior stroke	30 (13.5%)	15 (12.6%)	15 (14.6%)
Medication			
Oral anticoagulation	21 (9.5%)	10 (8.4%)	11 (10.7%)
Statin therapy (n=1)	49 (22.2%)	25 (21.2%)	24 (23.3%)
Antihypertensive therapy	130 (58.6%)	61 (51.3%)	69 (67%)
On admission			
Independent before stroke (n=23)	184 (92.5%)	104 (96.3%)	80 (87.9%)
NIHSS	13 (9)	12 (7)	15 (9.5)
Systolic blood pressure, mm Hg (n=8)	153 (36)	147 (29)	160 (35)
Diastolic blood pressure, mm Hg (n=9)	82 (25)	81 (22)	84.5 (27.5)
Glucose, mmol/L (n=14)	6.5 (2.02)	6.2 (1.7)	6.75 (2.5)
HbA1c, % (n=41)	5.8 (0.6)	5.8 (0.5)	5.8 (0.9)
LDL, mmol/L (n=27)	2.4 (1.4)	2.5 (1.3)	2.4 (1.4)
HDL, mmol/L (n=34)	1.4 (0.6)	1.3 (0.6)	1.4 (0.6)
TG, mmol/L (n=35)	1.3 (0.8)	1.3 (0.8)	1.3 (0.9)
CRP, mg/L (n=21)	3 (5)	3 (3)	4 (8)
INR (n=21)	1.0 (0.1)	1.0 (0.1)	1.0 (0.1)
Infarct side (left)	99 (44.6%)	54 (45.4%)	45 (43.7%)
Additional occlusions	38 (17.1%)	17 (14.3%)	21 (20.4%)
IVT	103 (46.4%)	58 (48.7%)	45 (43.7%)
Onset to imaging, min (n=4)	132 (210.5)	128 (187)	149.5 (243)
Onset to groin puncture, min (n=7)	216 (231)	210.5 (190.75)	230 (271)
Collateralization status (n=4)			
Good	115 (52.8%)	69 (58.5%)	46 (46%)
Moderate	78 (35.8%)	39 (33.1%)	39 (39%)
Poor	25 (11.5%)	10 (8.5%)	15 (15%)

Median (interquartile range) and frequency (percentage) for continuous and categorical clinical variables for all patients and for patients with favorable (mRS score, 0–2) and unfavorable (mRS score, 3–6) outcome. The number of missing values is indicated in brackets. CRP indicates c-reactive protein; CHD, coronary heart disease; HbA1c, hemoglobin A1c; HDL, high-density lipoprotein; INR, international normalized ratio; IVT, intravenous thrombolysis; LDL, low-density lipoprotein; mRS, modified Rankin Scale; NIHSS, National Institutes of Health Stroke Scale; and TG, triglycerides.

Clinical variables were normalized by subtracting the mean and dividing through the SD, both calculated based on the training data of the respective CV fold. This ensures that corresponding parameter estimates are directly comparable while larger values in either direction (negative or positive) indicate higher importance. DWI and TMAX perfusion maps were preprocessed such that 3-dimensional image volumes were of dimension 128×128×28. Pixel values were normalized for improved model training. Detailed methods, data preparation information, model architecture, implementation, and training procedures are summarized in the Supplemental Material (Supplemental Methods; Figures S2 and S3). Prediction model development and validation were performed according to TRIPOD guidelines.[23] Code is available on GitHub (https://github.com/liherz/functional_outcome_prediction_dl_vs_neurologists).

Functional Outcome Prediction of Stroke Neurologists

The most important criterion for a model to be considered in clinical practice is its prediction performance in comparison to current stroke experts. Therefore, we asked 2 male and 3 female neurologists with a median (interquartile range) age of 38 (9) and 10 (1) years of working experience in stroke, currently employed at tertiary stroke centers in Switzerland, to predict mRS at 3 months for randomly selected patients based on (1) clinical variables, (2) MRI, and (3) a combination thereof. We provided several routinely collected clinical variables, summarized in Table 1, as well as DWI and PWI (cerebral blood flow, cerebral blood volume, TMAX maps). Clinical variables only and imaging data only were from 50 different patients (100 patients in total). The combination of clinical variables and imaging data was of the same 50 patients as provided for predictions based on clinical variables only. This allowed us to evaluate improvements in mRS prediction when adding imaging to clinical data. The experts were instructed not to switch between the 3 modes and process cases sequentially.

Evaluation Procedure

We assessed prediction performance of DL models and stroke neurologists for estimating ordinal (mRS score, 0–6) and binary (mRS score, 0–2 versus 3–6) functional outcome. All reported measures are test performances on patient data that were not used during any stage of the model building process (Figure S1). To evaluate DL models for binary functional outcome prediction, predicted probabilities for mRS score 0 to 2 and 3 to 6 were summed up. For stroke neurologists, a prediction of mRS score ≤2 was assigned to class mRS score 0 to 2, a prediction of mRS score >2 to class mRS score 3 to 6. To achieve a common prediction across all raters, we averaged ordinal, respectively, binary rater predictions for each patient. For the DL models evaluated in the CV, we report the negative log likelihood to take into account the whole predictive distribution of the mRS. Furthermore, we calculated accuracy and quadratic weighted Cohen κ to measure the agreement between true and estimated ordinal predictions. Accuracy shows the percentage of correctly classified patients, calculated as the proportion of true positive and true negative designations combined divided by the total. The quadratic weighted Cohen κ ranges between 0 and 1 while penalizing predictions further away from the true one more harshly. For binary predictions, we considered accuracy and area under the receiver operating characteristic curve values. Calibration was assessed with brier scores and

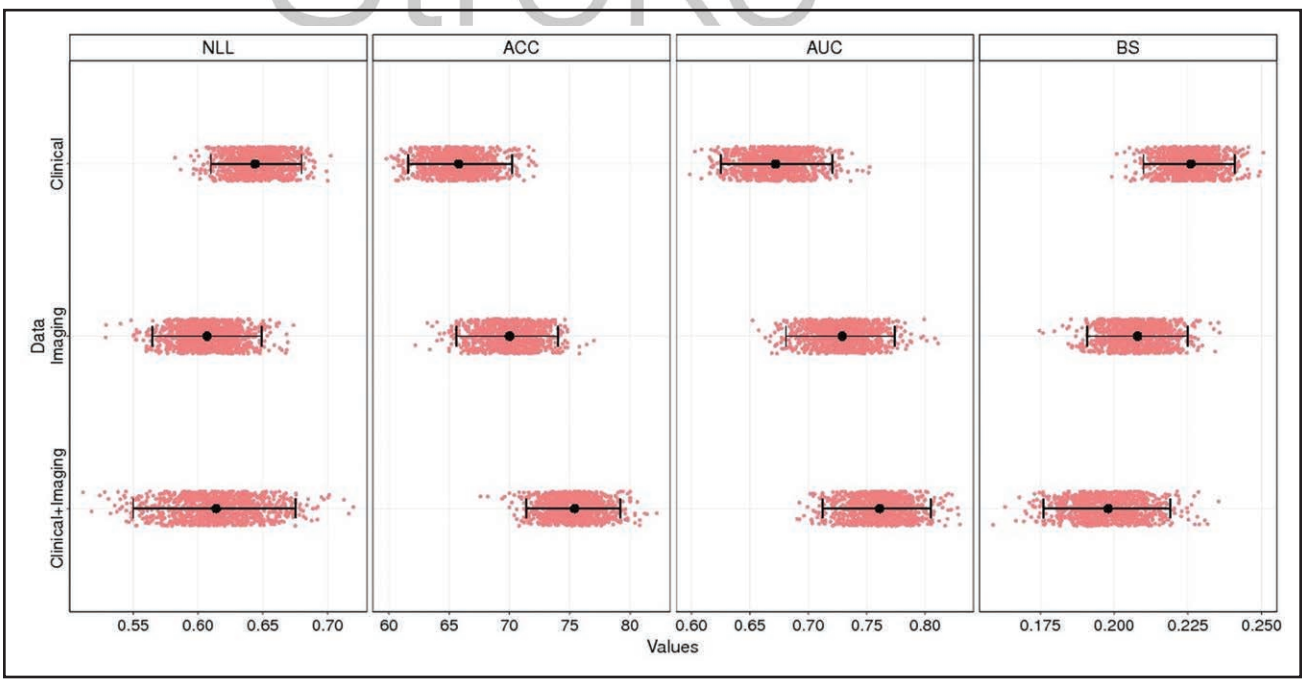

Figure 1. Performance of deep learning models for binary functional outcome prediction.
Test negative log likelihood (NLL), accuracy (ACC), area under the receiver operating characteristics curve (AUC), and Brier score (BS) for models based on clinical data, diffusion weighted imaging, and a combination thereof when evaluated in a 5-fold cross validation. Lower values of NLL and BS and higher values of ACC and AUC indicate better performance. The single points show the 1000 bootstrap metrics used to calculate medians and 95% CIs.

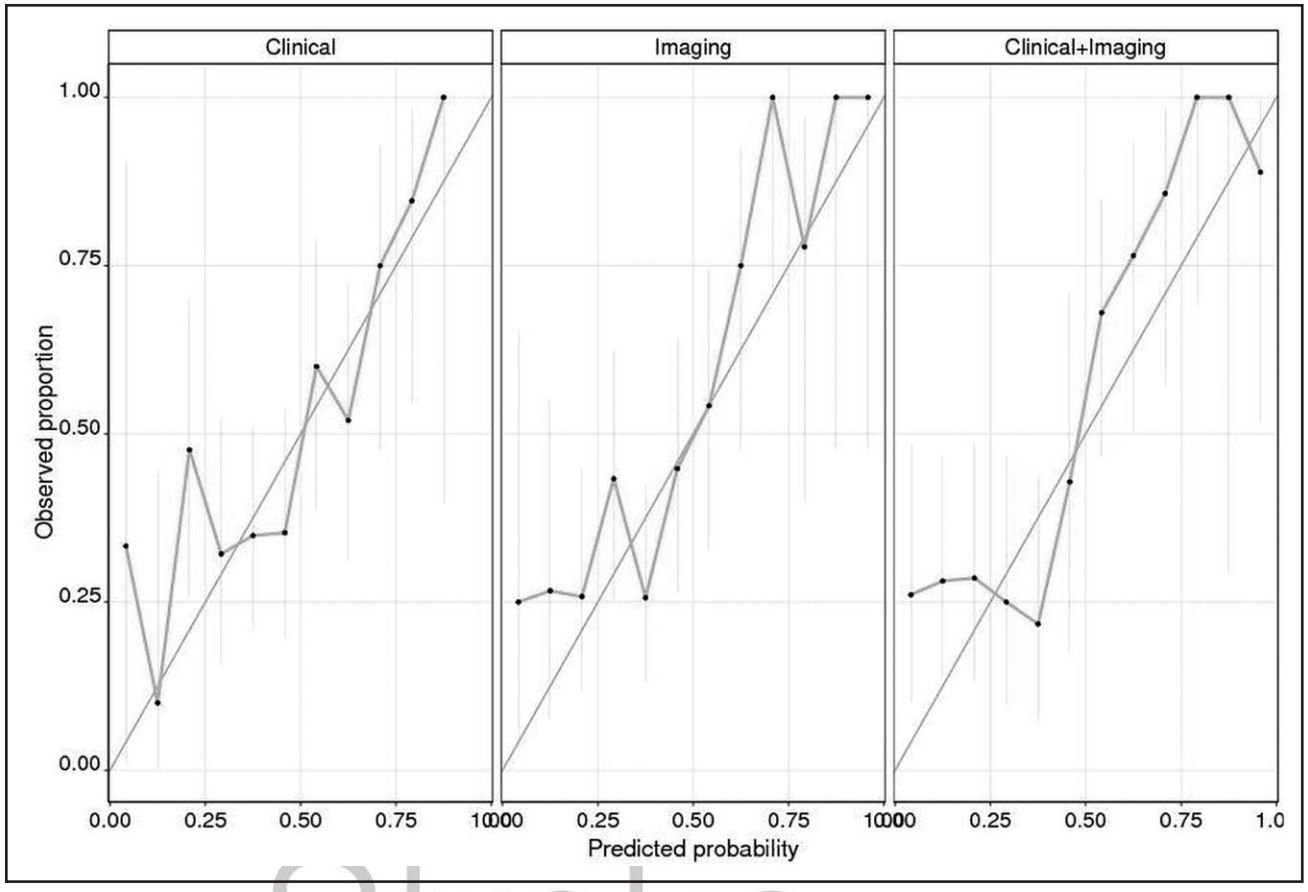

Figure 2. Calibration of binary deep learning model predictions.
The figure shows the calibration of the predicted probabilities for binary functional outcome. To obtain calibration curves, we split the predicted probabilities into 11 intervals of equal size and calculated the observed proportion and the average predicted probabilities of an unfavorable outcome. In case of perfect calibration, the points fit the straight line.

calibration plots. We compared predictions of DL models and stroke neurologists from those patients who were available to the neurologists. Discriminatory measures included accuracy and quadratic weighted Cohen κ, respectively, sensitivity and specificity in case of binary functional outcome. We constructed 95% CIs for all metrics by taking 1000 bootstrap samples from the test data of size 500 and computing the 2.5th, 50th, and 97.5th percentiles of the resulting 1000 bootstrap metrics. For evaluation of interrater reliability, we calculated Fleiss Kappa.[24]

RESULTS

Cohort Description

We screened 578 patients with MCA-M1 occlusion. We then excluded patients due to missing angiography (n=4) or initial MR perfusion (n=267), imaging data of insufficient quality (n=49), previous infarct signs (n=3), additional cerebral vessel occlusion (n=6), and missing outcome (n=27; Figure S4). While history of stroke was no exclusion criterion, we excluded those 3 patients due to large areas of signal loss within the newly ischemic region. Reasons for patients to not undergo MRI were MR contraindications (eg, metal implants), as well as stroke-induced or preexisting comorbidities. These included inability to lie flat (risk of aspiration or vomiting), inability to lie still (agitation, disorientation, and aphasia), or patients requiring intubation or other forms of continuous assistance due to clinical instability. In critically ill or agitated patients, and in cases of insufficient renal function, no PWI was performed. The final data set for analysis consisted of 222 patients who were treated with (n=217) or attempted to be treated with (n=5) MT. The distribution of the mRS at 3 months was as follows: n_0=32 (14.41%), n_1=48 (21.62%), n_2=39 (17.58%), n_3=32 (14.41%), n_4=32 (14.42%), n_5=5 (2.25%), and n_6=34 (15.32%). Clinical variables are summarized in Table 1. Overall, a favorable outcome was achieved in 54% of the patients. Of the patients, 78% were successfully recanalized. Five patients experienced symptomatic intracerebral hemorrhage. The median (interquartile range) age was 73.54 (20) years; the admission NIHSS score was 13 (9). Patients with favorable outcome were younger, suffered less often from atrial fibrillation or hypertension, were more often smokers, and less often exposed to antihypertensive therapy. On admission, patients with favorable outcome were more often

independent before stroke (mRS score, ≤2) and showed lower NIHSS, systolic blood pressure, glucose, and CRP (c-reactive protein) values.

Functional Outcome Prediction Models

Performance of DL models for binary functional outcome prediction when evaluated via CV based on the 222 patients is summarized in Figure 1.

Figure 2 shows calibration plots. For the sake of simplicity, we only report results of models using imaging data in terms of DWI. Using TMAX perfusion maps alone or in combination with DWI yielded similar results but increased model variability (Figures S5 and S6). The DL model based on a combination of clinical variables and DWI achieved the best results (area under the receiver operating characteristic curve, 0.761 [0.712–0.805]). Performance of the model based on DWI only was similar in terms of negative log likelihood but slightly worse when considering accuracy and area under the receiver operating characteristic curve. Providing DWI only yielded slightly better results than using clinical variables only (area under the receiver operating characteristic curve, 0.729 [0.681–0.774] versus 0.672 [0.625–0.721]). The negative log likelihood showed a higher variability of the DL model based on clinical variables and DWI compared with the models based on 1 modality. Brier scores and calibration plots indicated good calibration in all 3 models (Brier score, 0.226 [0.21–0.241] for clinical variables, 0.208 [0.191–0.225] for DWI, and 0.198 [0.176–0.219] for clinical variables and DWI). Ordinal functional outcome prediction was worse compared with binary functional outcome prediction but yielded similar results when we compared the 3 models (Figures S7 and S8).

Functional Outcome Prediction Models Versus Stroke Neurologists

Binary prediction performance of DL models and stroke neurologists, based on the same 50 randomly selected patients, is shown in Figure 3. The DL model integrating clinical variables and DWI achieved the best results (accuracy, 72% [67.8%–76%]). Performance of models using clinical variables (accuracy, 60% [55.4%–64.4%]) and DWI only (accuracy, 61.4% [56.6%–65.6%]) was comparable but lower than those of the DL model using both modalities. Stroke neurologists (accuracy, 60% [55.8%–64.21%]) and model achieved similar performances

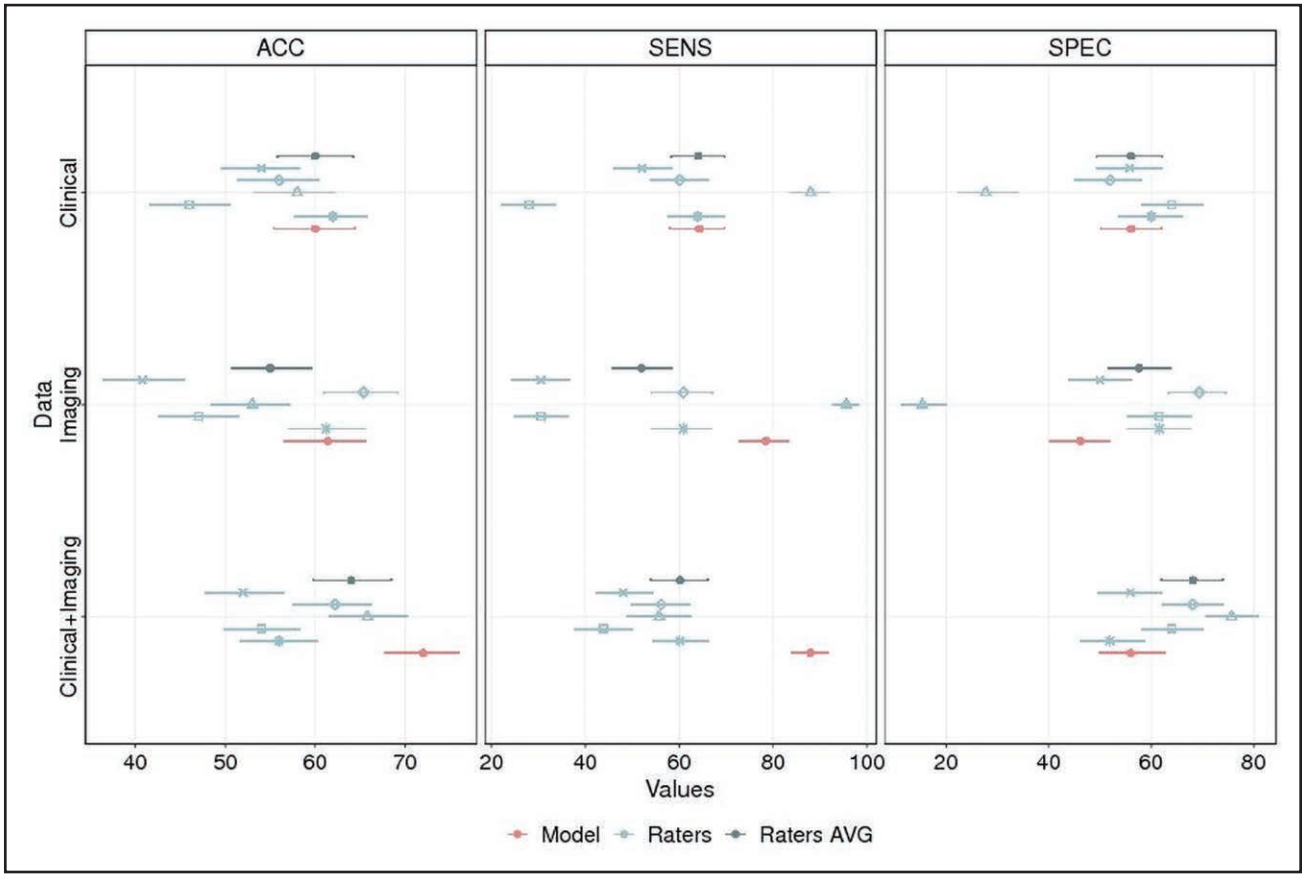

Figure 3. Models vs raters: binary functional outcome prediction.
The figure shows the results in terms of accuracy (ACC), sensitivity (SENS), and specificity (SPEC) for models (red), individual raters (light blue), and the raters average (AVG; dark blue) when predicting binary functional outcome based on different input data. The raters AVG is based on the binarized raters' predictions averaged for each patient. For all metrics, higher values indicate better prediction performances.

based on clinical variables only. However, when imaging data were provided, alone or in combination with clinical variables, DL models outperformed stroke neurologists significantly (accuracy, 61.4% [56.6%–65.6%] versus 55% [50.8%–59.6%] for imaging data and 72% [67.8%–76%] versus 64% [59.8%–68.4%] for clinical variables and imaging data). DL models outperformed neurologists in terms of sensitivity, that is, they more often correctly classified patients with favorable outcome. On the contrary, neurologists achieved similar or slightly higher specificities when being provided with imaging data. Apart from 1 rater, all raters achieved similar or slightly better results when adding imaging data to clinical variables. Again, performance of models and stroke neurologists for ordinal functional outcome prediction was worse but yielded similar results (Figure S9).

Overall, we observed large differences in functional outcome prediction between raters. According to Fleiss classification, there was only slight agreement when predicting binary functional outcome (κ, 0.205 for clinical variables, 0.134 for imaging data, and 0.216 for clinical variables and imaging data). Interrater agreement for ordinal functional outcome prediction was poor (κ, 0.07 for clinical variables, 0.06 for imaging data, and 0.072 for clinical variables and imaging data).

Clinical Predictors for Functional Outcome

The most important predictors for the 2 models integrating clinical variables were admission NIHSS and age (Figure 4). Note, odds ratios for clinical variables indicate the risk for a worse outcome when the respective variable is increased by 1 SD while all other variables are held constant. Like the models, the majority of stroke neurologists considered admission NIHSS and age as highly important for estimating functional outcome (Table 2).

DISCUSSION

Functional outcome prediction in patients with LVO stroke is challenging. We presented DL models, which do not only integrate imaging but also clinical data by simultaneously providing interpretable parameter estimates for the variables. We showed that our model based on DWI and clinical variables reliably predicts functional outcome and outperforms current stroke experts. Like the DL model, most neurologists achieved the best performance when being provided with clinical and imaging data. However, although all neurologists had comparable experiences, there was only slight agreement in outcome prediction. This is in line with previous studies, which showed that practicing physicians predict functional outcome relatively inaccurate based on clinical variables.[9,10] Our models could, therefore, be a valuable, objective tool for supporting neurologists in prognostic decision-making.

To apply functional outcome prediction models in clinical practice, a high prediction performance is indispensable. Although different scores for functional outcome prediction exist and have shown to outperform physicians, those scores ignore imaging data.[13] In our previous

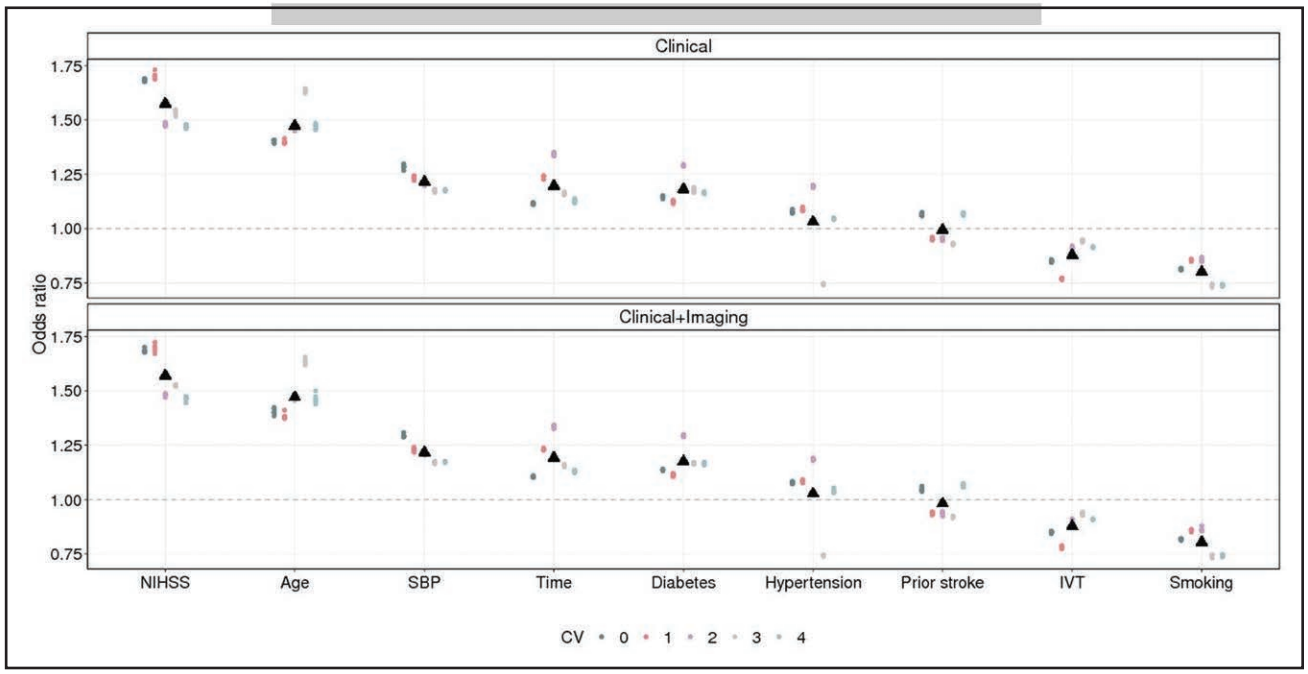

Figure 4. Odds ratios for clinical variables.
The figure visualizes the estimated odds ratios for clinical variables obtained with the model based on clinical variables only and the model based on clinical variables and diffusion weighted imaging. Variables are normalized and sorted with respect to decreasing effect size. CV indicates cross validation; and IVT, intravenous thrombolysis.

ORIGINAL ARTICLE

Table 2. Importance of Variables According to Stroke Neurologists

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Clinical parameters	Age	Independent before stroke	Admission NIHSS	Age	Admission NIHSS
	Admission NIHSS	Age	Independent before stroke	Admission NIHSS	Infarct side
	SBP	Additional occlusions	Age	Collateral score	Onset to imaging, min
Imaging parameters	DWI	DWI	DWI	DWI	DWI
	CBF/CBV	CBV	Mismatch	CBV	Mismatch
	TMAX	CBF	CBV/CBF		CBV

The table summarizes clinical and imaging parameters that were mentioned by stroke neurologists as the 3 most important predictors indicating functional outcome. CBF indicates cerebral blood flow; CBV, cerebral blood volume; DWI, diffusion weighted imaging; NIHSS, National Institutes of Health Stroke Scale; SBP, systolic blood pressure; and TMAX, time to maximum.

work, analyzing the same patient cohort, we have shown similar or slightly higher performances compared with the THRIVE score[25] when predicting favorable functional outcome using clinical variables and extracted brain imaging features.[16] However, only the DL models developed in this study further improved outcome prediction when adding imaging to clinical data. This highlights the importance for analyzing raw imaging data with DL models and indicates that DWI and PWI contains more relevant information than the traditionally considered imaging features. To further investigate those imaging modalities, we plan to highlight image regions that were most important for predicting functional outcome. This might give new insights into stroke pathophysiology and further enhance our understanding of stroke. Adding PWI to DWI did not improve outcome prediction, which is in line with previous observations.[16] Besides modeling raw imaging data, integrating clinical variables appears important. Using clinical variables or DWI only resulted in lower prediction performances. However, those performances were comparable to other studies predicting functional outcome with machine learning approaches using clinical variables[14] or CT Angio.[17] Although DL models for functional outcome prediction based on clinical and imaging data exist,[15] those models are not interpretable. In contrast, our models provide odds ratios for clinical variables, enabling a reliable quantification of model predictions.

Our study has limitations. One is the rather small sample size. This is because we only included patients with MCA-M1 occlusion and acute stroke MRI for increased quality and decreased variability. To account for this, we applied a CV setting, used pretrained neural networks, and performed data augmentation when modeling imaging data.[26,27] Nonetheless, we expect our models to perform better when more data are available. However, prediction performances were comparable or even higher than in other studies with similar[28] or larger sample sizes.[14,17] Based on these findings and the limited width of computed CIs, we are confident that our results allow valid conclusions about predictive performance of our models, despite the rather small sample size. As an additional limitation, our data set is monocentric, and since we only included patients with MCA-M1 occlusion, results are not directly generalizable to other patients with LVO strokes. Moreover, we cannot completely rule out bias toward patients with a better outcome due to exclusion of patients for whom no 3-month mRS score was available.

In summary, our interpretable DL models reliably estimate functional outcome and chances for recovery early after LVO stroke in patients treated with MT. They first analyze a combination of clinical variables and imaging data and simultaneously provide interpretable parameter estimates. They build the basis for future research for treatment decision-making and show the potential of interpretable machine learning models to transform health care. Moreover, in a direct comparison with stroke neurologists, they outperformed current experts who showed a high variability in prediction performance. We, therefore, hypothesize that functional outcome prediction of stroke neurologists may be significantly improved if they are supported by such models. Currently, we plan a prospective, controlled clinical trial to investigate how stroke physicians arrive at their predictions and whether prediction performance increases when neurologists are supported by our models.

ARTICLE INFORMATION

Received January 5, 2023; final revision received April 20, 2023; accepted April 26, 2023.

Affiliations

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland (L.H., L.K., B.S.). Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Switzerland (L.H., L.K., B.S.). Department of Neurology (L.H., J.H., C.G., A.R.L., S.W.) and Clinical Neuroscience Center (L.H., J.H., C.G., A.R.L., S.W.), University Hospital and University of Zurich, Switzerland. Department of Neurology (M.R.H., D.S., K.A., L.P., J.K., U.F., M.A.), Department of Diagnostic and Interventional Neuroradiology (T.D., J.G., R.W.), and Department of Diagnostic, Interventional and Pediatric Radiology (J.K.), Inselspital, University Hospital and University of Bern, Switzerland. Department of Neurology, University Hospital Basel, Switzerland (U.F.).

Sources of Funding

This project was funded by the Swiss National Science Foundation (PP00P3_202663), the University of Zurich Clinical Research Priority Program (CRPP) Stroke, and a Novartis FreeNovation grant.

Disclosures

Dr Heldner reports grants from the Bangerter Foundation, the Swiss National Science Foundation, SITEM (Center for Translational Medicine and Biomedical Entrepreneurship) Support Funds, and the Swiss Heart Foundation, all outside the submitted work. Dr Antonenko reports a grant from the Swiss National Science Foundation. Dr Dobrocky reports consultant fees for MicroVention, Inc. Dr Fischer reports grants from Alexion, CSL Behring, Medtronic, Penumbra, Inc, Phenox, Inc, Rapid Medical, Ltd, and Stryker. Dr Gralla reports interests with Medtronic USA, Inc. Dr Arnold reports consultancy fees from Amgen, AstraZeneca, Bayer, Bristol Myers Squibb, Covidien, Daiichi Sankyo, Novartis, Pfizer, Boehringer Ingelheim, Covidien, Medtronic, and Novo Nordisk. Dr Luft reports consultancy fees from Amgen and Moleac, Ltd. Susanne Wegener received research funds by the Swiss National Science Foundation, the UZH Clinical research priority program (CRPP) stroke, the Swiss Heart foundation, the Zurich Neuroscience Center (ZNZ), speaker honoraria from Springer, Teva Pharma, and consultancy fees from Bayer and Novartis. The other authors report no conflicts.

Supplemental Material

Supplemental Methods
Figures S1–S9
References 29–31

REFERENCES

1. Goyal M, Ospel JM, Kappelhof M, Ganesh A. Challenges of outcome prediction for acute stroke treatment decisions. *Stroke*. 2021;52:1921–1928. doi: 10.1161/STROKEAHA.120.033785
2. Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, Schonewille WJ, Vos JA, Nederkoorn PJ, Wermer MJ, et al; MR CLEAN Investigators. A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med*. 2015;372:11–20. doi: 10.1056/NEJMoa1411587
3. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, McTaggart RA, Torbey MT, Kim-Tenser M, Leslie-Mazwi T, et al; DEFUSE 3 Investigators. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med*. 2018;378:708–718. doi: 10.1056/NEJMoa1713973
4. Khandelwal P, Yavagal DR, Sacco RL. Acute ischemic stroke intervention. *J Am Coll Cardiol*. 2016;67:2631–2644. doi: 10.1016/j.jacc.2016.03.555
5. Campbell BC, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, Yan B, Dowling RJ, Parsons MW, Oxley TJ, et al; EXTEND-IA Investigators. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med*. 2015;372:1009–1018. doi: 10.1056/NEJMoa1414792
6. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, Roy D, Jovin TG, Willinsky RA, Sapkota BL, et al; ESCAPE Trial Investigators. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. 2015;372:1019–1030. doi: 10.1056/NEJMoa1414905
7. Campbell BCV, Majoie CBLM, Albers GW, Menon BK, Yassi N, Sharma G, van Zwam WH, van Oostenbrugge RJ, Demchuk AM, Guillemin F, et al; HERMES Collaborators. Penumbral imaging and functional outcome in patients with anterior circulation ischaemic stroke treated with endovascular thrombectomy versus medical therapy: a meta-analysis of individual patient-level data. *Lancet Neurol*. 2019;18:46–55. doi: 10.1016/S1474-4422(18)30314-4
8. Yoshimura S, Sakai N, Yamagami H, Uchida K, Beppu M, Toyoda K, Matsumaru Y, Matsumoto Y, Kimura K, Takeuchi M, et al. Endovascular therapy for acute stroke with a large ischemic region. *N Engl J Med*. 2022;386:1303–1313. doi: 10.1056/NEJMoa2118191
9. Saposnik G, Cote R, Mamdani M, Raptis S, Thorpe KE, Fang J, Redelmeier DA, Goldstein LB. JURaSSiC: accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology*. 2013;81:448–455. doi: 10.1212/WNL.0b013e31829d874e
10. Geurts M, de Kort FAS, de Kort PLM, van Tuijl JH, Kappelle LJ, van der Worp HB. Predictive accuracy of physicians' estimates of outcome after severe stroke. *PLoS One*. 2017;12:e0184894. doi: 10.1371/journal.pone.0184894
11. Strbian D, Meretoja A, Ahlhelm FJ, Pitkäniemi J, Lyrer P, Kaste M, Engelter S, Tatlisumak T. Predicting outcome of IV thrombolysis-treated ischemic stroke patients: the DRAGON score. *Neurology*. 2012;78:427–432. doi: 10.1212/WNL.0b013e318245d2a9
12. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology*. 2012;78:1916–1922. doi: 10.1212/WNL.0b013e318259e221
13. Ntaios G, Gioulekas F, Papavasileiou V, Strbian D, Michel P. ASTRAL, DRAGON and SEDAN scores predict stroke outcome more accurately than physicians. *Eur J Neurol*. 2016;23:1651–1657. doi: 10.1111/ene.13100
14. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, Hill MD, Sajobi TT. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models. *Front Neurol*. 2020;11:889. doi: 10.3389/fneur.2020.00889
15. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study. *Acad Radiol*. 2020;27:e19–e23. doi: 10.1016/j.acra.2019.03.015
16. Hamann J, Herzog L, Wehrli C, Dobrocky T, Bink A, Piccirelli M, Panos L, Kaesmacher J, Fischer U, Stippich C, et al. Machine-learning-based outcome prediction in stroke patients with middle cerebral artery-M1 occlusions and early thrombectomy. *Eur J Neurol*. 2021;28:1234–1243. doi: 10.1111/ene.14651
17. Hilbert A, Ramos LA, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, Barros RS, van der Schaaf I, Dippel D, Roos Y, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput Biol Med*. 2019;115:103516. doi: 10.1016/j.compbiomed.2019.103516
18. Powers WJ, Derdeyn CP, Biller J, Coffey CS, Hoh BL, Jauch EC, Johnston KC, Johnston SC, Khalessi AA, Kidwell CS, et al; American Heart Association Stroke Council. 2015 American Heart Association/American Stroke Association focused update of the 2013 guidelines for the early management of patients with acute ischemic stroke regarding endovascular treatment: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2015;46:3020–3035. doi: 10.1161/STR.0000000000000074
19. Pereira VM, Gralla J, Davalos A, Bonafe A, Castano C, Chapot R, Liebeskind DS, Nogueira RG, Arnold M, Sztajzel R, et al; the STAR Investigators. Prospective, multicenter, single-arm study of mechanical thrombectomy using solitaire flow restoration in acute ischemic stroke. *Stroke*. 2013;44:2802–2807. doi: 10.1161/STROKEAHA.113.001232
20. Kook L, Herzog L, Hothorn T, Dürr O, Sick B. Deep and interpretable regression models for ordinal outcomes. *Pattern Recogn*. 2022;122:108263.
21. Kook L, Gotschi A, Baumann P, Hothorn T, Sick B. Deep interpretable ensembles. 2022. https://doi.org/10.48550/arXiv.2205.12729
22. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112–118. doi: 10.1093/bioinformatics/btr597
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55–63. doi: 10.7326/M14-0697
24. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378. doi: 10.1037/h0031619
25. Flint AC, Cullen SP, Faigeles BS, Rao VA. Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score. *AJNR Am J Neuroradiol*. 2010;31:1192–1196. doi: 10.3174/ajnr.A2050
26. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. *BMC Med Imaging*. 2022;22:69. doi: 10.1186/s12880-022-00793-7
27. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
28. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. *Stroke*. 2018;49:1394–1401. doi: 10.1161/STROKEAHA.117.019740
29. He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *Ieee I Conf Comp Vis*. 2015;1026–1034. doi: 10.1109/Iccv.2015.123
30. Solovyev R, Kalinin AA, Gabruseva T. 3D convolutional neural networks for stalled brain capillary detection. *Comput Biol Med*. 2022;141:105089. doi: 10.1016/j.compbiomed.2021.105089
31. Herzog L, Kook L, Gotschi A, Petermann K, Hansel M, Hamann J, Durr O, Wegener S, Sick B. Deep transformation models for functional outcome prediction after acute ischemic stroke. *Biom J*. 2022. doi: 10.1002/bimj.202100379