communications medicine

Article

https://doi.org/10.1038/s43856-024-00666-w

Machine learning for early dynamic prediction of functional outcome after stroke

Check for updates

Julian Klug [1], Guillaume Leclerc[2], Elisabeth Dirren[1] & Emmanuel Carrera [1]

Abstract

Background Prediction of outcome after stroke is critical for treatment planning and resource allocation but is complicated by fluctuations during the first days after onset. We propose a machine learning model that can provide hourly predictions based on the integration of continuous variables acquired within 72 h of hospital admission.

Methods We analyzed 2492 admissions for ischemic stroke in the Geneva University Hospital from 01.01.2018 to 31.12.2021, amounting to 2'131'752 unique data points. We developed a transformer model that continuously included clinical, physiological, imaging, and biological data recorded within 72 h of admission. This model was trained to generate hourly predictions of mortality and morbidity. Shapley additive explanations were used to identify the most relevant predictors to explain outcomes for each patient. The MIMIC-III database was used for external validation.

Results Our transformer model predicts mortality, with an area under the receiver operating characteristic curve of 0.830 (95% CI 0.763–0.885) on admission, reaching 0.893 (95% CI 0.839–0.933) 72 h later for a 3-month outcome. Validated in an independent cohort, it outperforms all static models. Based on their mean explanatory weights, the top predictors included continuous clinical evaluation, baseline patient characteristics, timing from admission to acute treatment, and markers of inflammation and organ dysfunction.

Conclusions The performance of our transformer model demonstrates the potential of machine learning models integrating clinical, physiological, imaging, and biological variables over time after stroke. The clinical applicability of our model is further strengthened by access to hourly updated predictions along with accompanying explanations.

Plain language summary

Stroke is the most frequent cause of disability in industrialized countries. To determine the best treatment and allocate resources, an early and accurate prediction of outcome is essential. Although modern stroke units gather a continuous stream of data, existing tools for outcome prediction are rarely used as they are static and fail to adapt to the evolving condition of the patient.We developed a machine learning model, a computer system learning from existing data, to provide real-time predictions of in-hospital mortality and 3-month outcomes. Our model was able to provide accurate hourly prediction of outcome based on regularly updated clinical data obtained from the patient.This study demonstrates the potential of integrating the continuous data stream recorded in the electronic health record after stroke. Similar predictive models could help personalize treatment planning, empower patients and their families through counseling, and facilitate resource allocation.

Stroke remains the first cause of handicap in industrialized countries[1]. Patient trajectories after the initial event are highly heterogenous and outcomes range from full recovery to severe, debilitating neurological deficits with severe functional impairment or death. Predicting outcomes therefore enables personalized treatment planning through risk stratification, empowers patients and their families through counseling, facilitates resource allocation, and enhances the overall stroke recovery process[2,3]. Predicting stroke outcome also contributes to research and development of new treatments and helps designing clinical trials[4–6].

Acute stroke patients are routinely admitted to information-rich environments, usually stroke or intensive care units (ICU) for the first days after the index event. Despite the amount of data available, traditional[7–10] and machine learning[11–14] models are not commonly used to predict long-term functional outcome. Two main reasons may explain the limited clinical relevance of the previously reported models. First, existing scores are static, based on data acquired on admission or shortly afterwards, and do not include variables that are continuously or repeatedly recorded during the first days of hospitalization. Available models are thus largely blind to the early course of the patients' condition, which can be marked by dramatic clinical and physiological fluctuations during the first days after onset[15]. In other words, these models do not investigate, for instance, how the variability of parameters like arterial blood pressure or heart rate over time or excursion in blood values beyond normal ranges, may negatively affect long-term functional outcome. Technical constraints have long limited the use of

[1]Stroke Research Group, Department of Clinical Neurosciences, University Hospital and Faculty of Medicine, Geneva, Switzerland. [2]Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. e-mail: emmanuel.carrera@hcuge.ch

the continuous streams of information, with traditional statistical models struggling with heterogeneous multi-dimensional data, irregular sampling, and artefacts[16,17]. More complex machine learning models using gradient-boosted trees and recurrent neural networks have recently demonstrated their usefulness in outcome prediction in patients admitted to ICU, with the integration of a wide range of continuous physiological variables[18–22]. In acute and critically ill patients, newly developed transformer models might perform even better, as they are particularly well suited for sequential data, such as arterial blood pressure or heart rate[23–25]. Classically used for language processing, they have been shown to perform well on medical text. In the setting of acute ischemic stroke, the transformer architecture has been used for the automatic retrieval of contraindications to thrombolysis[26], automated information extraction from radiology reports[27], as well as improving the prediction of outcome when combined with admission data in stroke patients[28]. The vector resulting from the decomposition of an image can be analyzed analogously to a sequence of words by so called Vision transformers[29], yielding promising results in the automated detection of stroke in computed tomography (CT), magnetic resonance (MRI), and Doppler images[30–32]. Interpreting the EHR data as a sequence of events, transformers have been used as a warning system for stroke-associated pneumonia and the estimation of overall cardiovascular risk[33,34]. However, these models have for now been limited to the detection of specific complications or to single timepoints, but have not yet been applied to the real-time outcome prediction after stroke. The second main reason that limits the clinical use of machine learning models relates to their lack of transparency. Improvement in model performance has come at the cost of greater complexity, creating so-called black box models[35]. In the management of patients, predictions that are not associated with mechanistic explanations show limited acceptance and come with legal and ethical challenges[36]. On the other hand, algorithms that probe and explain individual predictions offer an opportunity to deploy predictive models to complex environments in which understanding and trust are key[37–39].

Here, we first develop a transformer machine learning model to predict in-hospital and 3-month clinical outcomes in a large population of ischemic stroke patients based on all available continuous and non-continuous variables. We investigate its performance to output hourly updated predictions. We then extract, from this model, the explanatory weights of the most important inputs to identify the variables that more strongly predict the predefined outcomes. In this study, we use data from 2492 acute ischemic stroke admissions to the Geneva Stroke Center to train the predictive model. This model is externally validated in a population of stroke patients included in the MIMIC-III database[40]. The transformer model achieves a good discriminative performance of high clinical relevance for both in-hospital and 3-month clinical outcomes and outperforms existing static models. It accurately tracks dynamic trends over time and offers insights into the variables most strongly associated with acute events during the hospital stay. Based on their mean explanatory weights, the top variables driving predictions include continuous clinical evaluation, baseline patient characteristics, timing from admission to acute treatment, and markers of inflammation and organ dysfunction.

Methods

Study design and setting

The study was designed as a retrospective cohort study for the development and validation of a clinical prediction model. The study was performed at the Stroke Center of the Geneva University Hospitals, Switzerland. The study was conducted in accordance with the Helsinki Declaration and approved by the local institutional review board (Commission cantonale d'éthique de la recherche de Genève 2016-01445). Consent was waived in accordance with Article 34 of the Swiss Federal Act on Human Research due to the retrospective nature of the study. This manuscript adheres to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement (TRIPOD)[41].

Participants

We extracted all admissions for acute ischemic stroke to the Stroke Center of the Geneva University Hospital between January 1st 2018 and December 31st 2021. 2492 admissions for ischemic stroke (2359 patients) were identified during the study period among the 5154 admissions with the diagnoses of ischemic, hemorrhagic strokes, transient ischemic attacks, and stroke mimics (Supplementary Fig. 1). The diagnosis of acute ischemic stroke was prospectively recorded by the treating physician and verified by MRI as part of standard of care. Exclusion criteria were patients younger than 18 years old, admitted more than 7 days after onset, with a hospital stay shorter than 12 h, as well as in-hospital strokes, secondary transfer, lack of follow-up data, and active opposition to data use for research purposes.

Data sources

Data used in this study were extracted from the hospital electronic health record (EHR) used in the daily management of patients to collect and store patient health information, including physiological parameters, laboratory values, clinical evaluations, and imaging data. In addition, data were extracted from the Swiss Stroke Registry (SSR) in which all patients characteristics, in-hospital management information, and outcomes are prospectively collected from the EHR, for audit and research purposes[42]. As a standard of care, the clinical outcome at 3 months was prospectively assessed during an in person visit by a senior consultant and recorded in the SSR, using the modified Rankin scale (mRS). This scale, ranging from 0 (no symptoms at all) to 6 (death)[43], is routinely used in stroke populations to evaluate outcome at 3 months[44].

Outcome measures

We defined in-hospital mortality, as well as two dichotomous outcome measures at 3 months: (1) mortality and (2) good functional outcome as endpoints for all analyzes. A "good" functional outcome was defined as mRS 0-2 (functionally independent) and "poor" functional outcome as mRS 3-6 (functionally dependent or death) (Supplementary Table 7)[44–46].

Feature timeseries generation

Variables were selected by consensus between the investigators based on existing models[11–13,20] among static information obtained at admission (e.g., demographics, comorbidities), daily sampled information (e.g., laboratory measurements), and values sampled at higher sampling rates (e.g., vital signs and clinical evaluation) (Supplementary Data 1). Imaging data was collected for a random subset of patients to evaluate its impact on model performance (Supplementary Table 1). All variables were pre-processed into features optimized for computation and were aggregated to hourly values based on existing strategies (Supplementary Methods 1)[19,20]. Data from the first 72 h after admission were used based upon prior work in ICU populations[20] and coinciding with the phase of ischemic stroke with the highest risk of deterioration[47] as well as with the length of stay in the acute care setting (intermediate or intensive care unit)[48,49].

Supervised learning of outcome prediction

All models were trained for three binary predictions tasks defined according to the three predefined outcome measures. The models were developed, fine-tuned, and selected on a random subset of the Geneva Stroke Dataset (Supplementary Fig. 2). To avoid overfitting during model development, 5-fold cross-validation was used for the selection of hyperparameters and model architecture. To segregate development and test datasets, we split all patients randomly, with 20% used for the final evaluation in the holdout test dataset (internal validation). The holdout test dataset was used only once for the testing of the models selected through cross-validation. When a patient had more than one admission, all admissions were assigned to the same split to avoid information leaking between datasets. There were no readmissions in the holdout test dataset. We remedied the imbalanced nature of the dataset by stratifying for outcome at each split. The prevalence of the endpoint was not artificially inflated to enhance model performance, but rather maintained in its original state to accurately reflect future model usage.

Machine learning model development

We compared three state-of-the-art supervised machine-learning techniques for the three tasks. All models had to update their prediction by integrating new data from a continuous stream and learning from the temporal development of the features. We have opted for hourly predictions to allow for clinically relevant continuous prediction with manageable model complexity. Two of the evaluated models (transformer, long short-term memory model) take a sequence of input data and learn longitudinal patterns as well as interactions between the input features to predict their final output, making them well-suited for timeseries. Decision tree ensembles were elected as third model type, as they have traditionally performed well on tabular data on similar clinical tasks[19,50]. As they are not inherently built for timeseries, they were evaluated on an aggregated view at each timepoint (Supplementary Methods 4).

Transformer encoder model: Transformers have been traditionally used in language processing and are composed of layers made of two main components: self-attention and feedforward neural networks[23]. The self-attention mechanism analyzes the relationships between each item in a sequence. This is used to weigh each item, driving the focus of the model to the features most relevant to the task. The feedforward network takes the output of the self-attention layer and applies two linear transformations to reach a new representation. This is passed on to the next layer of the model, where the process is repeated. By stacking multiple layers of self-attention and feedforward networks, the transformer encoder can model complex relationships between sequential features. The final representation is passed through a linear projection to obtain a binary classification. The traditionally associated decoder is not needed for classification tasks.

Long short-term memory model (LSTM): A neuronal LSTM unit can store values over arbitrary time intervals and has three gates that regulate the flow of information into and out of the neuron. The input gate determines whether to let new inputs in, the forget gate determines whether to discard currently used information when it is no longer deemed important, and the output gate determines whether to let the input affect the output at the current time step[51].

Gradient boosted decision tree ensemble (XGB): The XGBoost library (version 1.6.1) was used for model fitting[52]. XGBoost iteratively adds decision trees to an ensemble, where each subsequent tree corrects the errors of the previous tree. The final classification of this model is determined by taking the weighted average of the predictions made by all the decision trees in the ensemble.

All models underwent hyperoptimization of training parameters (Supplementary Methods 3). The hyperparameter settings maximizing the median area under the receiver operating characteristic curve (ROC AUC) across folds on the optimization (validation) set were used to generate the predictions on the test set and external validation set. We further evaluated the performance of the final model on selected subgroups based on age, sex, severity of initial presentation, initial treatment, and SARS-CoV-2 infection during the same admission. As the transformer model achieved the highest performance during system development, it was used for further analyzes.

Baseline clinical models

We compared our three machine learning models against four scores that are used in clinical practice for the prediction of functional outcome. We selected three point-based stroke prognostic scores, namely Totaled Health Risks in Vascular Events Score (THRIVE)[7], Houston Intra-Arterial Therapy score[10] and the Stroke Prognostication using Age and NIH Stroke Scale index (Span100)[9], as well as THRIVE-C, a validated logistic equation model[8]. These scores were selected as comparators because they were validated in similar patient populations. We further selected the three most salient clinical variables: admission NIHSS, pre-stroke disability, and IVT timing. These were used as input to train a simple custom logistic regression model on the development set, to be evaluated in the internal and external set test sets. As THRIVE-C performed best among clinical models in our dataset, it was used as reference for further comparison.

Explainable predictions

The Shapley additive explanations (SHAP)[37] algorithm was used to obtain explanations of how individual features drive patient-specific predictions from the previously obtained model. SHAP applies a game theoretic approach to identify the contribution of each feature to the prediction by considering all possible coalitions of features. It then assigns a value to each feature that represents its contribution to the prediction, considering the impact of the other features: positive and negative SHAP values indicate an increment or decrement of the prediction score, respectively. On a patient level, SHAP-values were used to identify the most relevant features at each timepoint. Features driving major changes in predicted outcome probability were identified by maximum change in SHAP-value at the same timepoint (Fig. 1).

Impact of access to imaging data

To investigate the added benefit of including imaging features, we selected a random a subset of patients (10%) for which perfusion imaging features obtained from the admission CT were extracted when available (Supplementary Table 1). A new model was then trained having access to imaging features for the selected subgroup for the prediction of functional outcome at 3 months. It was then evaluated on the subset of the holdout test set for which imaging features had been collected, once with access to imaging features, once without access to imaging features.

External validation

MIMIC-III version 1.4[40] was used for external validation using in-hospital mortality and death at 3 months as outcomes. MIMIC-III is a large, single-center database of patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, USA between 2008 and 2014. MIMIC-III was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care, and all protected health information was deidentified. After appropriate training in the protection of human research participant data, access can be requested over an online portal[53].

All patients over 18 years old admitted for the first time for acute ischemic stroke were identified. The identification of patient records was based on a two-step process with first a screening based on ICD-9 code and admission diagnosis, followed by the manual verification of inclusion and exclusion criteria by two of the investigators (ED, JK). Exclusion criteria included death during the 72-h surveillance period, admissions under 12 h, absence of a detailed neurological admission exam in the clinical notes, in-hospital stroke, and initial management in another hospital (secondary transfer). Most of the previously described variables were readily available. Neurological and functional status at admission, stroke onset, and intravenous and intra-arterial therapy timings, as well as comorbidities and admission medication were manually extracted from the admission and discharge notes by two of the investigators (ED, JK). A previously published database[54] of all individual items of the National Institutes of Health Stroke Scale (NIHSS)[55] for patients in MIMIC-III was used to obtain the details of the neurological exam at admission. Neurological evaluation scores (NIHSS, Glasgow Coma Scale (GCS)) were estimated from recurrent clinical examinations throughout the ICU stay. Mortality dates were available for all patients. Follow-up mRS was not recorded in the MIMIC-III population.

Data cleaning and pre-processing were performed as described above with the following minor modifications. For three variables with a large number of missing values (>2/3 of subjects), the first measurement was imputed with the median of the development population if necessary (Supplementary Data 1). The parameters used for winzorising and normalization came from the development population.

No retraining or finetuning of the model was undertaken before external validation on the MIMIC-III dataset. The ability to compensate for missing values further served as a sensitivity analysis.

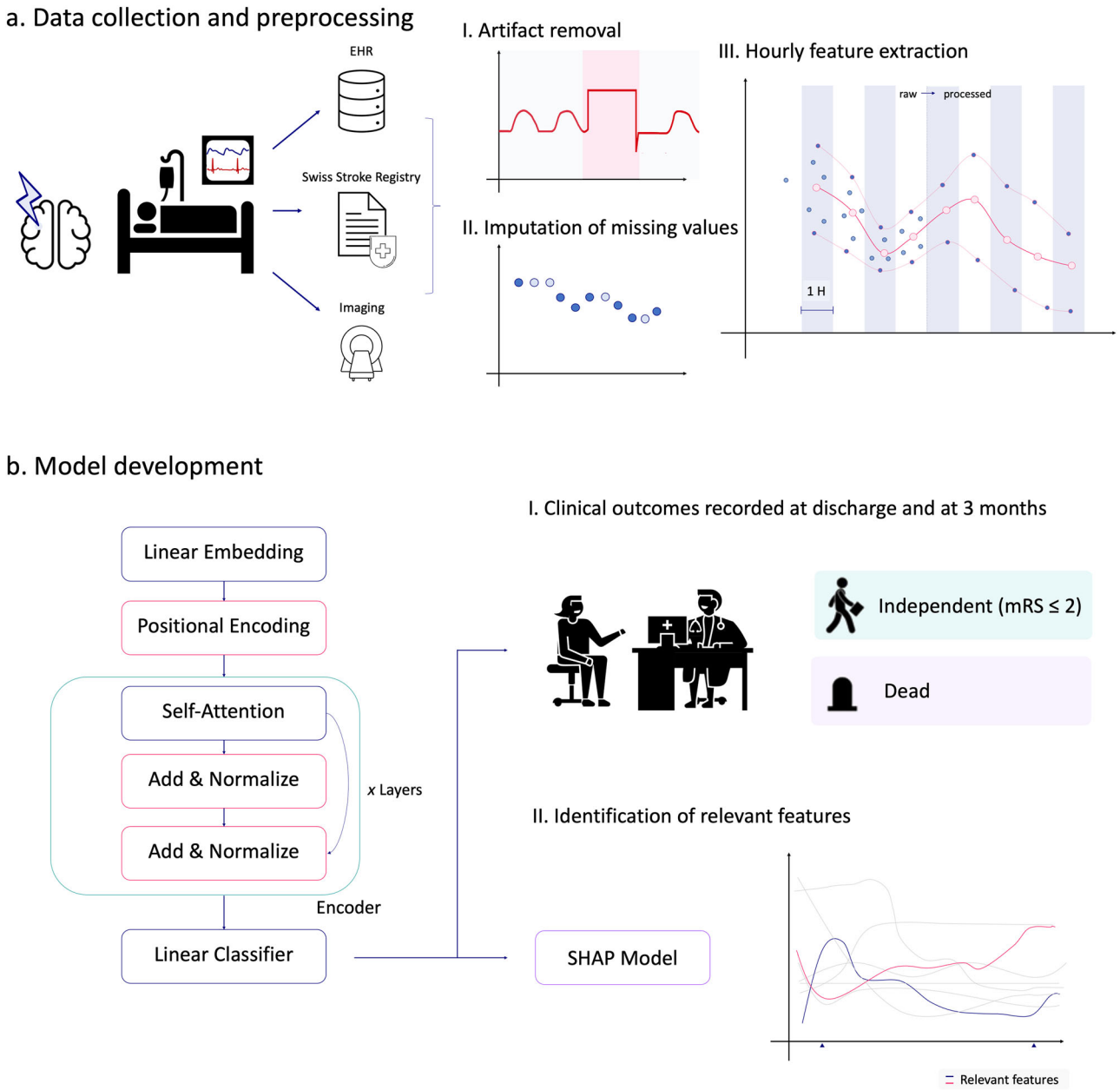

Fig. 1 | Data collection, pre-processing, and development process. a All patients admitted for acute ischemic stroke between 01.01.2018 and 31.12.2021 were identified. Relevant clinical variables were recorded prospectively in the electronic health record (EHR) and the Swiss stroke registry. **a**.**I** Artifacts were removed by identifying values exceeding physiologic plausible ranges. **a**.**II** Missing observations were imputed via last observation carried forward. **a**.**III** All variables were sampled hourly, or downsampled to hourly medians, maximums, and minimums. **b**.**I** A transformer encoder model was trained to predict clinical outcomes at discharge (mortality) and at 3 months (mortality and functional independence assessed using the modified Rankin scale (mRS)). **b**.**II** SHapley Additive exPlanations (SHAP) were used to identify relevant features driving the predictions.

Statistics and reproducibility

If not otherwise indicated, reported performance metrics in text and tables, as well as solid lines in figures, refer to the performance in the holdout test dataset. To evaluate the models discriminative ability, we used Matthew's correlation coefficient (MCC)[56] and ROC AUC[57]. The models were evaluated on test set data and ranked based on the primary discriminatory metric of ROC AUC. We further report the accuracy, positive predictive value, sensitivity, and specificity. Confidence intervals (95%) were obtained through boot-strapping of 1000 samples with replacement. Shaded areas refer to the standard deviation and were derived from $n = 5$ independent experiments from the cross-validation splits. Interquartile ranges (IQR) are used to represent variation across data.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Geneva stroke (development) dataset

Selection of variables and preprocessing of high-frequency data (Fig. 1). The Geneva Stroke Dataset contained data from the 2492 admissions to the Geneva Stroke Center between 01.01.2018 and 31.12.2021 with the diagnosis of acute ischemic stroke (Supplementary Fig. 1). Of the 2359 unique patients included in the study 2244 patients had one admission and 115 patients had multiple admissions for ischemic stroke during the study period. The variables of interest,

Table 1 | Description of training, holdout test, and external validation datasets

	Geneva stroke dataset		MIMIC-III
	Training dataset (n = 1812)	Holdout test dataset (n = 441)	External validation dataset (n = 247)
Age, years	76 (64–84)	77 (66–85)	73 (58–82)
Sex, male	1006 (55%)	242 (54%)	122 (49%)
Pre-stroke disability, mRS	0 (0–1)	0 (0–1)	2 (1–2)
BMI, kg/m^2	24 (21–27)	24 (21–27)	30 (25–48)
Hypertension	1238 (68%)	287 (65%)	168 (68%)
Diabetes	393 (21%)	85 (19%)	54 (21%)
Atrial fibrillation	377 (20%)	96 (21%)	77 (31%)
NIHSS on admission	3 (1–7)	3 (1–9)	13 (7–19)
IVT	414 (22%)	115 (26%)	159 (64%)
IAT	220 (12%)	54 (12%)	50 (20%)
Disability at 3 months, mRS	2 (0–4)	2 (0–4)	/
Mortality at 3 months	258 (14%)	62 (14%)	51 (20%)

Data from the Geneva Stroke Dataset as used for the mortality analysis. Values are presented as mean (IQR) and N (%).
BMI body-mass index, *NIHSS* National Institute of Health stroke scale, *IVT* intravenous thrombolysis, *IAT* intra-arterial treatment, *mRS* modified Rankin scale.

including patients' characteristics, physiological and acute treatment data, as well as laboratory tests, were recorded as a standard of care during the first 72 h of admission. 84 features were used for model development (Supplementary Data 1). Among them, 50 variables were obtained once or less than once per hour (for instance age, sex, or laboratory values). These data were upsampled or aggregated to hourly median values. Seven variables (NIHSS, systolic, diastolic, and mean blood pressure, oxygen saturation level, respiratory and heart rates) were recorded more often than once per hour and were downsampled to hourly medians, maximums, and minimums, resulting in 21 additional features. One-hot encoding of categorical variables resulted in 13 further features[58]. Patient outcomes were in-hospital mortality, as well as status at 3 months determined using mortality and the mRS, a 7-point disability scale ranging from 0, no symptom to 6, death. After preprocessing, the dataset contained 2,131,752 unique datapoints, with a mean of 857 observations per admission.

The development datasets for the prediction of in-hospital mortality, 3-month mortality, and good functional outcome (mRS 0-2) contained respectively 2492, 2253, and 2245 admissions (90–100% of all ischemic stroke admissions) after exclusion of patients without follow-up (Supplementary Fig. 1). In the development dataset used for mortality prediction, the median age was 76 years (IQR 64–84) and 1006 admissions (55%) concerned male patients. Upon admission, the median NIHSS was 3 (IQR 1–7) and the median pre-stroke mRS was 0 (IQR 0–1). 258 (14%) patients had died at 3 months and the median follow-up mRS was 2 (IQR 0–4). Table 1 summarizes baseline patients' characteristics in the training, holdout test, and external validation (MIMIC-III) datasets (see below).

Development of a continuous risk prediction model of outcome

We aimed to continuously predict three dichotomous clinical outcomes: (1) in-hospital mortality, (2) 3-month mortality, and (3) good functional outcome (mRS 0-2)[44] at 3-months. We developed a transformer model and compared its performance with other traditional and machine learning algorithms (LSTM and XGB)[23,51,52]. Predictions were computed every hour using data collected during the first 72 h after admission. The models were evaluated and compared at 72 h after admission on holdout test data and ranked based on the primary discriminatory metric of ROC AUC for all outcomes[57]. Model parameters obtained after optimization are reported in Supplementary Methods 3. A model inherently capable of dealing with timeseries data was preferred as this reduces the need for manual feature extraction and can give insight to the handling of time within the model. Overall, the high ROC AUC of the transformer model to predict in-hospital mortality, 0.911 (95% CI 0.843–0.951), 3-month mortality, 0.893 (95% CI 0.839–0.933) and good functional outcome, 0.894 (95% CI 0.863–0.922) can be considered as highly clinically relevant (Fig. 2 and Supplementary Data 2). The corresponding MCCs were 0.486 (95% CI 0.318–0.629), 0.565 (95% CI 0.437–0.676) and 0.638 (95% CI 0.567–0.711)[56]. Among all models, the transformer performed slightly better to predict mortality compared to LSTM and XGB and was similar to predict good functional outcome (mRS 0-2) (Supplementary Table 2). The transformer model (as well as XGB and LSTM) performed largely better than all traditional models, of which the best performing model, THRIVE-C[8], achieved a ROC AUC in the prediction of in-hospital mortality and 3-month outcomes of 0.840 (95% CI 0.815–0.866), 0.787 (95% CI 0.765–0.807) and 0.811 (95% CI 0.797–0.824) respectively (Fig. 2 and Supplementary Data 2). The performance of the transformer model was preserved across predefined subgroups; being slightly better in more severely affected patients (NIHSS > 5) (Supplementary Tables 3–5).

Model performance over time

The final transformer model had access to all features during the first 72 h after admission. Intermediate predictions can be obtained at any time point. In other words, the model produced an hourly prediction for all outcomes (in-hospital mortality, survival, and functional status at 3 months), continuously updated based on incoming data. Overall, the performance of the model improved gradually during the first 72 h after admission as information load increased (Fig. 3). The highest daily gain in performance was observed within the first 24 h of admission. The ROC AUC of the model to predict good functional outcome at three months increased from 0.835 (95% CI 0.795–0.869) on admission to 0.878 (95% CI 0.846–0.908) at 24 h and 0.894 (95% CI 0.863–0.922) at 72 h. The increase in ROC AUC for the prediction of in-hospital and 3-month mortality was more gradual, from respectively 0.765 (95% CI 0.656–0.854) and 0.830 (95% CI 0.763–0.885) at admission to 0.838 (95% CI 0.756–0.906) and 0.850 (95% CI 0.791–0.902) at 24 h and 0.911 (95% CI 0.843–0.951) and 0.893 (95% CI 0.839–0.933) at 72 h. This observation underlines the dynamic characteristics of our model with a prediction that can be adjusted as new information is continuously integrated. Clinically, it demonstrates the importance of acute stroke monitoring as made available in dedicated stroke units[59]. For instance, application of our model based on updated information could help determine, without delay, when a patient, initially with a favorable prognosis becomes, a few hours later, at risk of unfavorable outcome (Fig. 4 and Supplementary Fig. 7).

Feature inspection and explanations of model predictions

Features included static baseline data in combination with a stream of continuously measured variables, as well as a representation of their variability. We used the SHAP algorithm[37] to obtain explainable predictions at any given timepoint. SHAP values weight the impact of each feature on the model output: positive and negative SHAP values indicate an increment or decrement of the prediction score, respectively. We determined the top 10 predictors of good functional outcome and mortality by their mean absolute SHAP value (Fig. 5). These top 10 features included continuous clinical evaluation scores (NIHSS and GCS), baseline patient characteristics (age, prestroke functional status), timing from admission to acute treatment (intravenous thrombolysis), as well as blood markers of inflammation and organ dysfunction (leukocyte count, C-reactive protein, urea, and sodium). On a patient level, the main factors driving the individual prediction can be identified at any point in time, serving as an explanation and opportunity to back-check the model for the end-user. The model can further explicit the evolution of the predicted risk in the last hours. Inflection points in patient trajectories can be used to point out variables most relevant to the change in

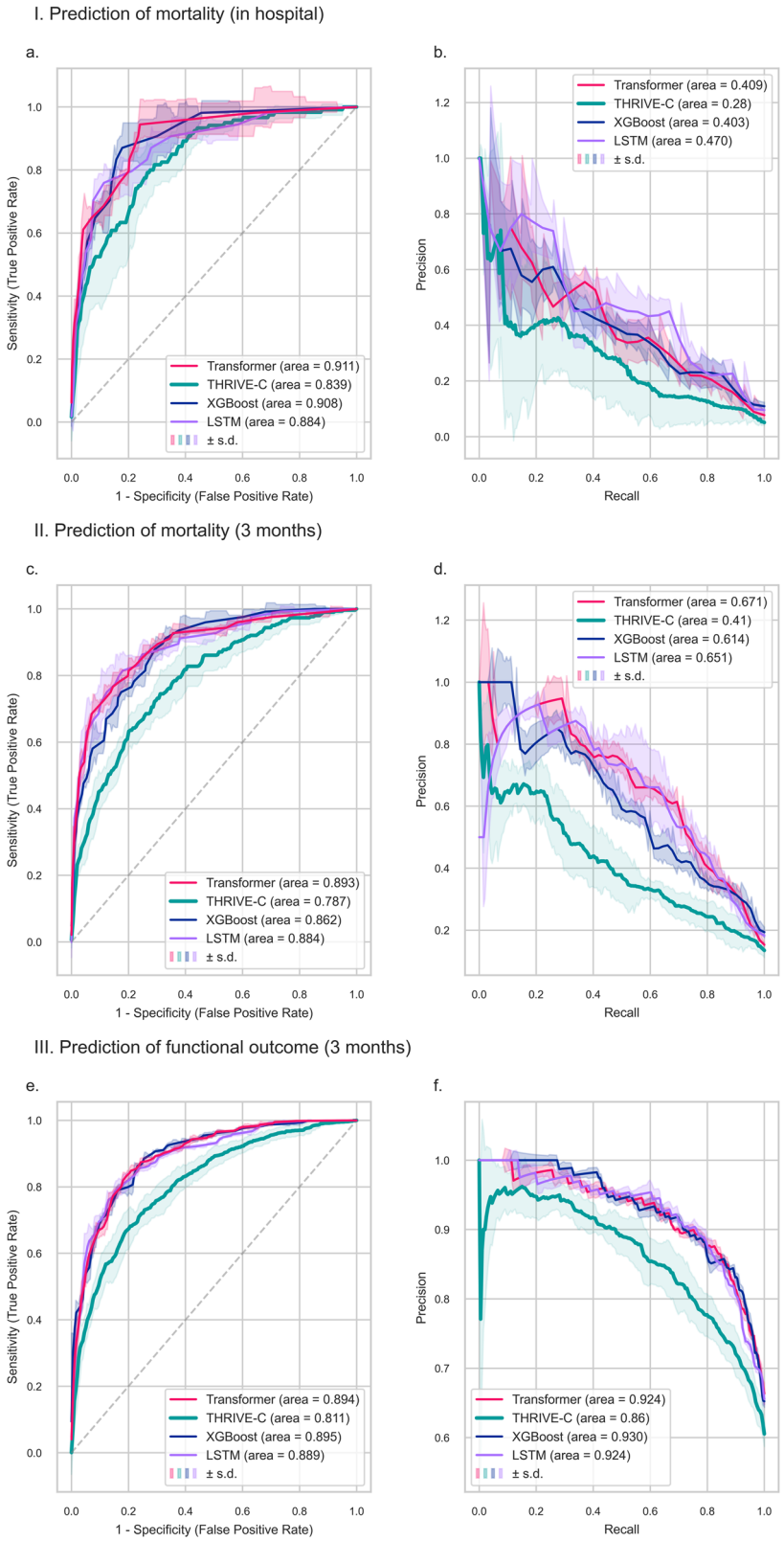

Fig. 2 | Receiver operating and precision-recall curves for the prediction of outcomes at discharge (death) and at 3 months (good functional outcome and death) in the Geneva Stroke Dataset. Comparison of the developed transformer model (red), gradient boosted tree ensemble (XGBoost, blue), long short-term memory model (LSTM, purple) and best performing clinical model (THRIVE-C, green) in the Geneva Stroke hold-out test dataset (internal validation) for the prediction of in hospital death (I), as well as death (II) and good functional outcome (III) at 3 months. (**a**, **c**, **e**) show ROC curves. (**b**, **d**, **f**) show precision-recall curves. Solid curves were derived from the held-out test split, variation estimates (shaded areas, standard deviation) were derived from $n = 5$ independent experiments from the cross-validation splits. s.d standard deviation, ROC receiver operating characteristic.

model prediction and thus likely associated with the change in patient condition (Fig. 4 and Supplementary Fig. 7).

Impact of access to imaging data

Imaging data was available for 150 (66.6%) patients of the randomly selected subset. The performance in the prediction of good functional at 3 months of the transformer model increased slightly when given access to imaging features achieving a ROC AUC of 0.974 (95% CI 0.913–1.000) and MCC of 0.577 (95% CI 0.395–0.756). This compares to a ROC AUC of 0.969 (95% CI 0.895–1.000) and MCC of 0.569 (95% CI 0.395–0.778) in the same cohort when imaging data was not available (Supplementary Table 6).

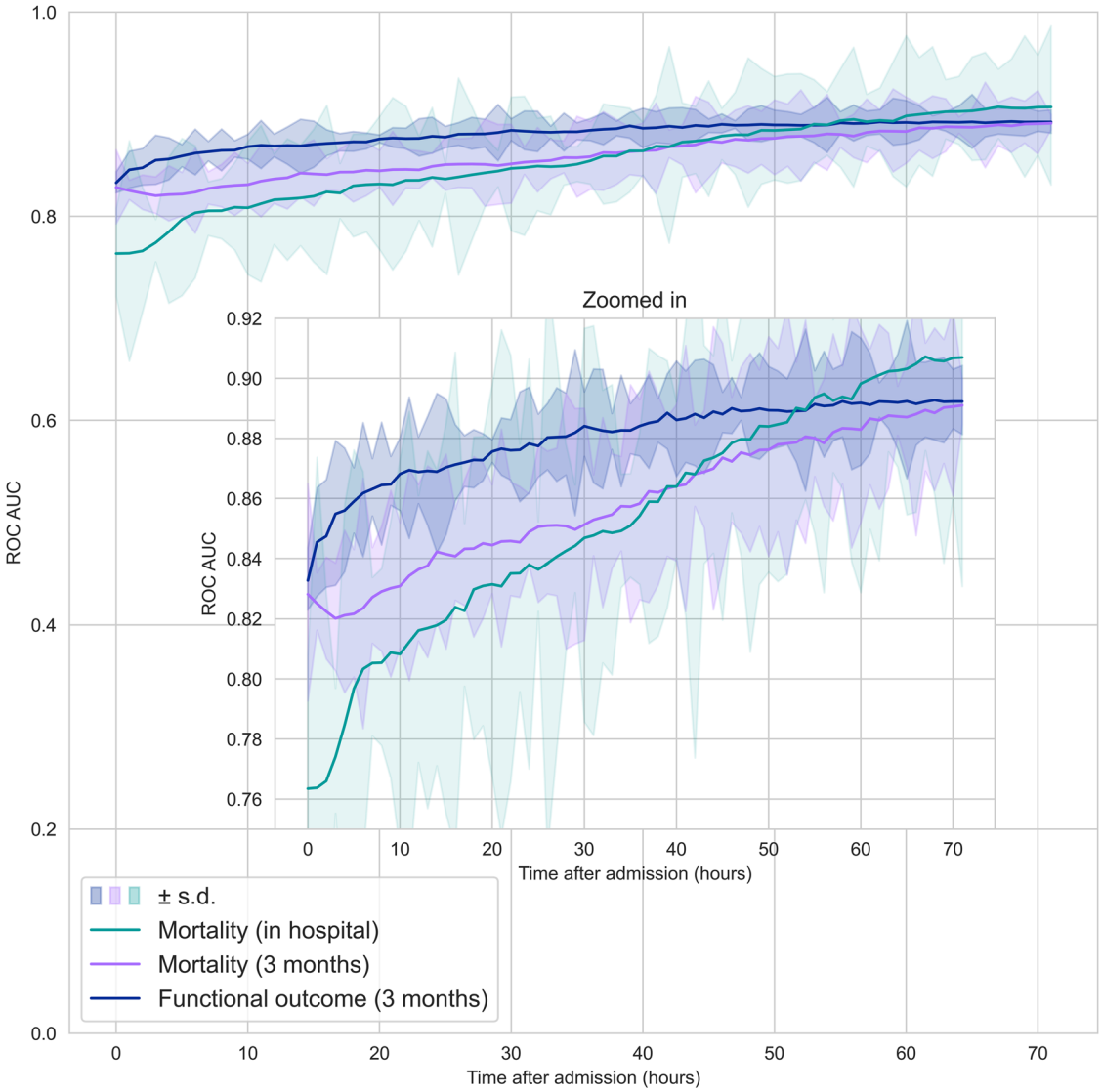

Fig. 3 | Performance across the first 72 h after admission of the final model (expressed as ROC AUC) for all main outcomes in the Geneva Stroke Dataset. Solid curves were derived from the held-out test split, variation estimates (shaded areas, standard deviation) were derived from $n = 5$ independent experiments from the cross-validation splits. ROC AUC area under the receiver operating characteristic curve, s.d standard deviation.

External validation of mortality prediction on MIMIC-III

For external validation, we used the publicly available ICU dataset MIMIC-III as it was the only dataset reporting stroke patients with continuous physiological variables[40]. A total of 744 admissions for acute ischemic stroke were identified. After manual chart review, 247 admissions (247 patients) were retained. 497 patients had to be excluded because of missing detailed admission or discharge data (Supplementary Fig. 1). In the external validation dataset (Table 1), the median age was 73 years (IQR 58–82) and 122 patients (49%) were male. Compared to the development dataset, patients in the external validation dataset were more likely to present with more severe neurological deficits, with a median NIHSS of 13 (IQR 7–19), and received acute treatment more frequently (intravenous thrombolysis (64%) or intra-arterial thrombectomy (20%)). MIMIC-III patients had a worse prestroke functional status (mRS of 2 (IQR 1–2)) and higher 3-month mortality ($n = 51$ (20%)). Because blood C-reactive protein, N-terminal prohormone of brain natriuretic peptide (BNP), and fibrinogen were only rarely recorded, imputation was done with the population medians of the development dataset for missing values. The ability to compensate for these missing values serves as an internal sensitivity analysis. For the prediction of in-hospital mortality and mortality at 3 months, the transformer model achieved a ROC AUC of 0.876 (95% CI 0.791–0.932) and 0.875 (95% CI 0.820–0.919) and a MCC of 0.293 (95% CI 0.168–0.382) and 0.448 (95% CI 0.338–0.553) respectively. Detailed follow-up mRS and imaging data were not available in the external validation cohort.

Discussion

We have developed a transformer model integrating static baseline data and continuous physiological variables to predict in-hospital mortality and clinical outcomes at 3 months. During the first 72 h after admission, the model provided an accurate, real-time, prediction of outcome, which was continuously adjusted according to the newly acquired clinical and physiological data. This model was additionally able to identify patient-specific features that drive outcome prediction.

Our transformer model achieved a discriminative performance of high clinical relevance, with static information available on admission enriched by the dynamic measures recorded during the following 72 h. To predict in-hospital mortality and clinical outcomes at 3 months after stroke, the ROC AUCs of the model ~0.8 on admission, reaching nearly 0.9 at 72 h. On admission, the initial performance of our model, using, by default, only the first available datapoint, was comparable to previous models predicting

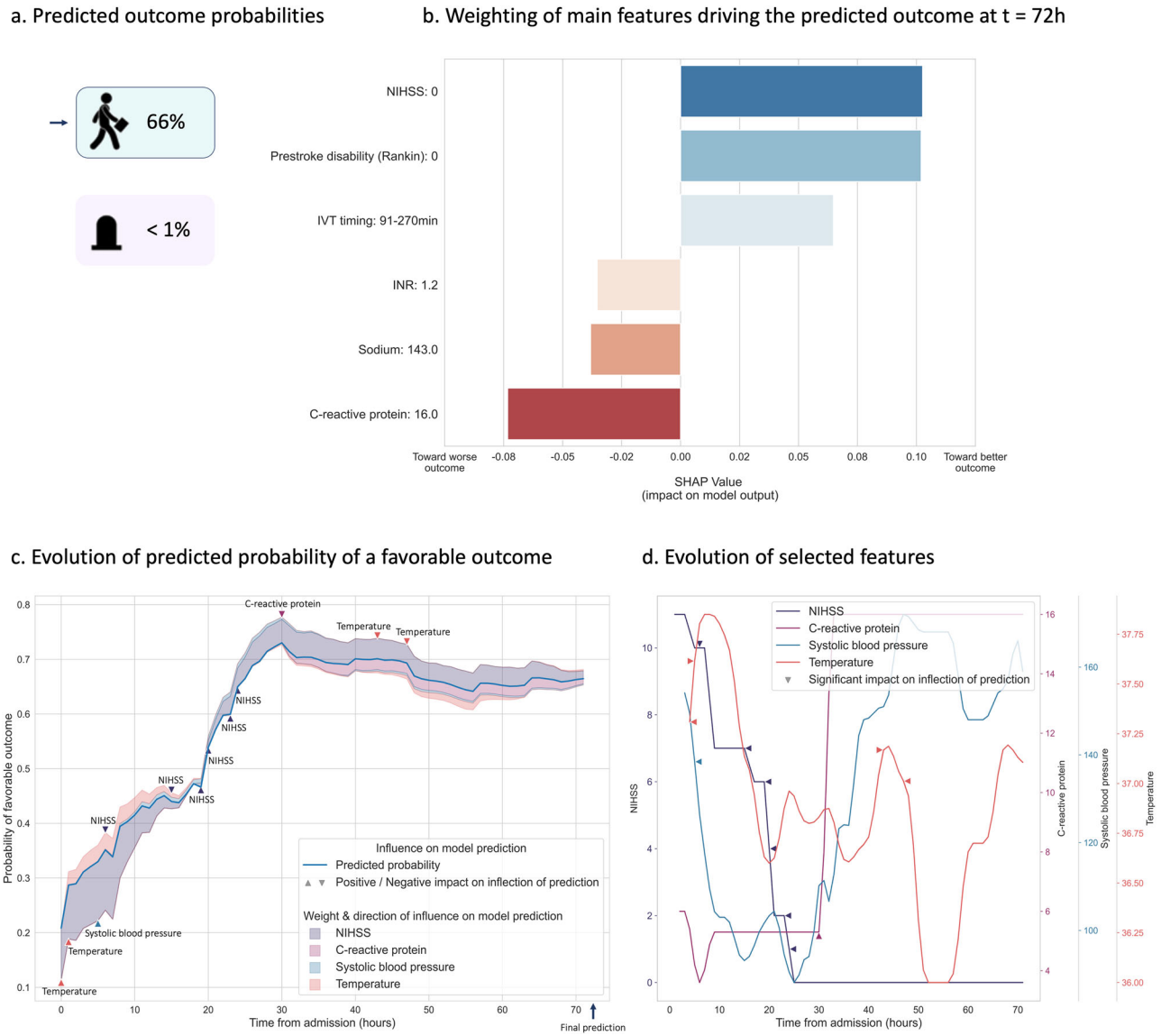

Fig. 4 | Example of the prediction and identification of relevant features for a given patient. a Prediction, at 72 h, of good functional outcome (mRS 0-2) and death at 3 months. **b** Top 3 positive and negative driving features for the prediction of functional outcome by SHAP value. **c** Evolution of the prediction of functional outcome at 3 months (dark blue line) during the first 72 h of admission. The x-axis represents time from admission in hours, and the y-axis represents the predicted probability of good functional outcome at 3 months. Inflection points in predicted probability are pointed out with colored triangles, associated with the feature driving the changes in prediction. The shaded areas represent the influence of selected features at any given point in time. Features above the bold blue line represent features influencing the model towards predicting a higher probability of good functional outcome. Features below the bold blue line represent features pulling the model towards predicting a lower probability of good functional outcome. The magnitude of the effect is represented by the distance to the bold blue line. Further help for interpretation of this figure is provided in Supplemental Fig. 2. **d** Evolution of predictive features across time. The features driving changes in the predicted probability of good outcome are plotted in colored lines. mRS modified Rankin scale, NIHSS National Institute of Health stroke scale, IVT intravenous thrombolysis, INR international normalized ratio.

functional outcome after stroke using baseline information only[12,14,60,61]. However, as it integrates all data available during the 72 first hours of hospitalization, the performance of our model increased considerably over time, performing far better at 72 h than on admission and also better than existing static models[12,14,60,61]. One major additional benefit of our approach is the ability of the model to provide an accurate prediction of long-term outcome at any time within 72 h of admission based on the integration of continuous and non-continuous variables, as previously demonstrated in ICU patients to predict mortality[20]. For the physician in charge, this continuously updated online prediction is highly relevant, as changes in clinical status and physiological variables occur frequently at any time point in the first hours/days after stroke. Furthermore, for the clinical team, it is currently almost impossible to integrate the increasing flow of data becoming available, resulting in lost opportunities to intervene in individual patient trajectories. The computational help that is proposed here, enables healthcare professionals to make more comprehensive decisions based on all available data.

Whereas prior literature has focused on gradient-boosted tree models[19,61] and recurrent neural networks with long short-term memory units[20,62], we provide evidence that transformer models can be highly relevant in a clinical setting. Transformer models have recently gained widespread attention, as they perform well on a variety of data modalities, including sequential data, with the development of large language models based on generative pretrained transformers (GPT)[63], as well as automated image analysis[29]. On such tasks, prior machine learning models struggle with long-range dependencies[64] which are handled by the transformers

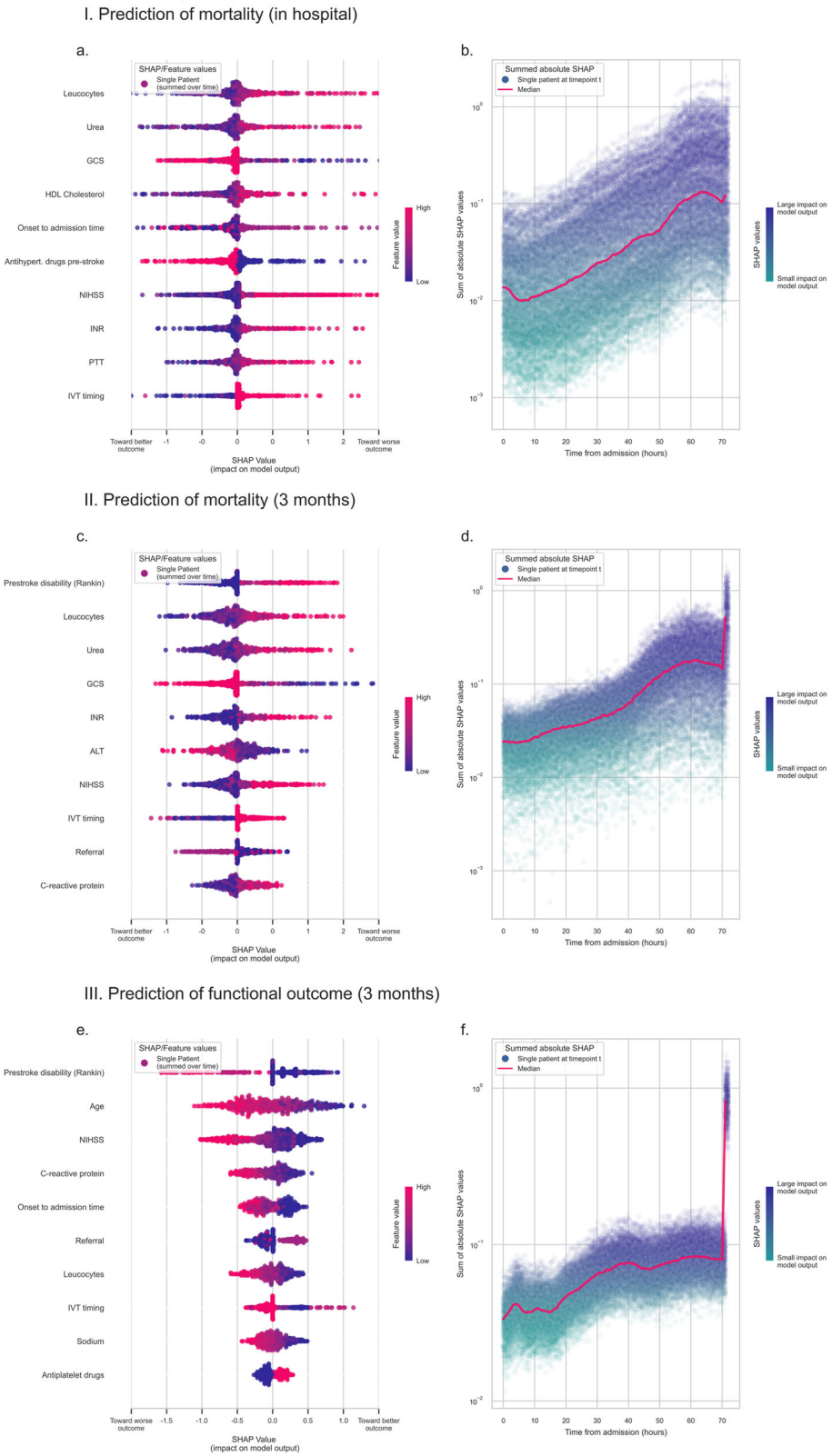

Fig. 5 | Most important features and timesteps as identified by SHAP values for the prediction of outcomes at discharge (death) and at 3 months (good functional outcome and death) in the model evaluated at 72 h after admission. a, **c**, **e** SHAP values summed over all timesteps for each feature are represented on the y-axis for all subjects in the test set ($n = 441$). The dot plot shows a color coding of the actual value of the feature, resulting in the SHAP value as indicated on the x-axis. The color coding is based on the percentile of the feature value with respect to the whole distribution. Please note that since predicted outcomes represent opposite patient trajectories, the moral quality of the x-axis is inversed between [**a**, **c**] and **e**. **b**, **d**, **f** Sum of absolute SHAP values of all features for all subjects of the test set over time ($n = 441$). The dot plot shows a single color-coded SHAP value for every timestep (log-scale). The hourly median for all subjects is plotted as a red curve. NIHSS National Institute of Health stroke scale, GCS glasgow coma scale, IVT intravenous thrombolysis, INR international normalized ratio, PTT partial thromboplastin time, HDL high-density lipoprotein.

multi-head attention mechanism and positional embeddings. For each element in a sequence, the transformer model, on the contrary to LSTM or tree models combined with summarization methods, can consider the interactions with every other element in the sequence simultaneously without relying on intermediary states. Consequently, transformers can scale easily to longer sequences and thus handle longer time spans and data of higher resolution[65]. Today, transformer models are considered the state-of-the-art architecture for scaling up to large datasets, as they reduce training and inference times by allowing for parallel computation, processing whole sequences at once, and avoiding the sequential dependency of prior architectures[23,66]. Recent developments in both hardware and software make transformers fast and lightweight in terms of resources[67–69]. The results

obtained here, with a transformer model performing similarly or better than the XGB and LSTM models even on a medium-sized dataset without overfitting, are encouraging and are a strong suggestion that, with larger quantities of high-quality data, performance would drastically improve. With the increasing deployment of monitoring technology, clinicians are expected to interpret a wide range of data along with their evolution over time. Our results underline the ability of transformer models to process multimodal inputs and to handle timeseries, with promising healthcare applications. As an example, the system, as developed here, could easily be adapted to integrate further data modalities, notably by adding information contained in clinical notes and raw imaging data[70]. Nonetheless, the overall developed framework is agnostic to the architectural backbone and can be equally used with the slightly less performant, but simpler XGB model.

The extraction of SHAP values facilitates the clinical applicability of the model. We were able to identify the variables most strongly associated with outcome, which included specific patient pre-stroke characteristics, clinical severity, acute management data as well as markers of organ dysfunction. Our analysis highlighted the importance of blood markers of infection and inflammation that are frequent complications after stroke and important predictors of unfavorable outcome or death[71]. In acute and critically ill patients admitted to the emergency department and ICU, SHAP values are increasingly used to identify variables associated with mortality, shock, and organ failure[19,20,50,72]. Compared to previous use of SHAP values, we went one step further and demonstrated that this technique can be used to identify, over time, individualized features driving prediction. Clinically, this technique may help determine which features impact on the change in prediction in an individual patient, at a given timepoint. Although SHAP values alone may not provide a comprehensive explanation for the underlying cause of adverse outcomes in the acute phase of stroke, they may facilitate the generation of clinical hypotheses on a personalized basis.

We used the MIMIC-III dataset to test the external validity of the developed model[40]. We acknowledge differences between the Geneva developmental and MIMIC validation dataset characteristics. The differences concern, first, the patient inclusion period (2008–2014 in the MIMIC-III and 2018-2021 in the Geneva Databases). Furthermore, the Geneva Database included all stroke patients admitted to the Geneva University Hospital, whereas the MIMIC-III data was restricted to those admitted to the ICU, resulting in a group of more severely affected patients in the MIMIC-III database. We used this dataset because it is, to our knowledge, the only available dataset containing stroke patients with continuous physiological monitoring variables. Nevertheless, despite the differences in datasets characteristics, we were able to show that the model's ability to predict death in hospital and at 3 months was preserved. To foster further research in the field, we will make our dataset available to the research community to facilitate the external validation of future models.

In addition to the differences in database characteristics, our study has several limitations. Deploying machine learning models is further associated with inherent risks. The single-center design of our development cohort creates a risk of overfitting to a specific patient population. As such models easily overfit their training data, they may perpetuate biases contained in the initial dataset resulting in systematic errors in their ability to classify subgroups of patients. However, the preserved performance recorded when evaluating our model on the external MIMIC-III dataset makes overfitting highly unlikely in the present study, with a model performing well across a range of subgroups. To date, there is no standard for the validation of post-hoc explainability methods and their deployment may lead to over-confidence in model predictions by confirmation bias[73]. Clinicians must be aware that explainable models will not be more performant than a black-box model and that the provided explanations are only approximations of the inner workings of the model. Nonetheless, we are convinced of the utility of SHAP values as used in the current study, for model back-checking and hypothesis generation[74]. Finally, it is important to mention that most predictive models purely rely on statistic association, do not account for future treatments and interventions, and do not imply causation. This is of importance when making decisions regarding patient treatment based on model output. For instance, a predicted favorable outcome should only cautiously be interpreted as an indication to adapt the usual clinical treatment, as the patient will only progress to the predicted outcome if receiving the same subsequent care as it was in the development cohort. We report encouraging results when adding imaging features as input to our model, in-line with recently published literature[12,75,76]. Although larger than most previously published cohorts, the analysis of the imaging subset, however, remains limited by a small sample size, as the extraction of imaging features is challenging[77] and is not available for all patients. As perfusion imaging data was not available in the MIMIC-III cohort, external validation of the imaging model was not performed. Future predictive models should further investigate the added benefit of imaging features or even include raw imaging data as input to the model. We have limited our analysis to the first 72 h after admission. Although the overall hospital stay of stroke patients is generally longer[78,79], the model is not informed by any event occurring outside this timeframe. This design choice targets the period with the highest risk of deterioration after acute ischemic stroke[47], representing a critical timepoint for early interventions targeting the peri-infarct and the risk of early reoccurrence[80,81]. This is further reflected in the length of stay in an acute care ward with continuous monitoring, which is on average 72 h in our center and in international cohorts[48,49,82]. A similar timeframe has successfully been used to predict 90-day mortality in a cohort of patients admitted to the ICU[20]. We choose to focus on medium-term instead of short-term endpoints as outcomes for the proposed model. Although this introduces a latency between the timing of the prediction and the occurrence of the outcome, this reflects current recommendations[44,83] and trial designs[84,85] as the mRS assessed at 3 months correlates well with long-term death and disability[45]. This design choice targets the pragmatic patient-centered questions that arise at the bedside from healthcare personnel and family alike: Will the patient return to his prior functional status? Will the patient suffer from handicapping deficits? Or will the patient die from the insult?[86]

We have demonstrated that a machine learning model can provide accurate and dynamic prediction of outcome in the acute phase of ischemic stroke. These results may have an immediate clinical impact at the bedside for the clinical team, the patient, and his family, by providing accurate information early after stroke onset. On a broader scale, the development of dynamic machine learning models opens new perspectives for other critically-ill patients admitted to information-rich environments, such as ICUs and emergency departments[20,50]. To ensure convenient clinical usability of our model, it is critical to ensure proper integration of the various and complex variables which are increasingly recorded in the EHR. In this context, the development of collaborative teams including IT specialists and clinicians will be essential. Finally, for a further and complete validation of our model, future randomized clinical trials are needed to investigate whether interventions based on machine learning models improve outcome in stroke patients. It is only when this last step will be successful that predictive machine models will be used confidently at the bedside by the clinical teams.

Data availability

The Geneva Stroke dataset cannot be made publicly available to protect patient confidentiality but is available from the corresponding author upon reasonable request. MIMIC-III can be freely accessed over the physionet.org platform. Researchers need to formally request access after completing a recognized course in protecting human research participants that includes Health Insurance Portability and Accountability Act requirements and signing a data use agreement, outlining appropriate data usage and security standards[53]. Supplementary Data 1 and 2 respectively contain the list of all variables used along with their accepted range and missingness and performance metrics for all evaluated models to predict predict in hospital mortality, as well as good functional outcome (mRS 0-2) and mortality at 3 months in the Geneva Stroke (internal validation) and MIMIC-III datasets (external validation). Data used for the creation of Figs. 2–4 is available at https://doi.org/10.5281/zenodo.13694271[87]. For Fig. 2, ground truth labels and final predictions for all models and target outcomes are available. For

Fig. 3, ground truth labels and predictions at every timestep are available for every fold and every target. For Fig. 5, SHAP values derived from the final transformer model are available for every target outcome and every timestep. For Fig. 4, predictions for an anonymized example patient are available.

Code availability

The computer code, as well as model weights used in this study are open-source and available at https://github.com/JulianKlug/OPSUM (or https://doi.org/10.5281/zenodo.11473804[88]) and https://doi.org/10.5281/zenodo.10848070[89] respectively. An interactive demo can be accessed at https://opsum.julianklug.com/.

Received: 23 October 2023; Accepted: 5 November 2024;
Published online: 13 November 2024

References

1. Tsao, C. W. et al. Heart disease and stroke statistics—2022 update: a report from the American Heart Association. *Circulation* **145**, e153–e639 (2022).
2. Kwakkel, G. & Kollen, B. J. Predicting activities after stroke: what is clinically relevant? *Int. J. Stroke* **8**, 25–32 (2013).
3. Cho, J. S. et al. Hospital discharge disposition of stroke patients in Tennessee. *South Med. J.* **110**, 594–600 (2017).
4. Holliday, E. et al. Developing a multivariable prediction model for functional outcome after reperfusion therapy for acute ischaemic stroke: study protocol for the Targeting Optimal Thrombolysis Outcomes (TOTO) multicentre cohort study. *BMJ Open* **10**, e038180 (2020).
5. Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial intelligence for clinical trial design. *Trends Pharmacol. Sci.* **40**, 577–591 (2019).
6. Kavalci, E. & Hartshorn, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci. Rep.* **13**, 121 (2023).
7. Flint, A. C. et al. THRIVE score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in VISTA. *Stroke* **44**, 3365–3369 (2013).
8. Flint, A. C. et al. Improved ischemic stroke outcome prediction using model estimation of outcome probability: the THRIVE-c calculation. *Int. J. Stroke* **10**, 815–821 (2015).
9. Saposnik, G., Guzik, A. K., Reeves, M., Ovbiagele, B. & Johnston, S. C. Stroke prognostication using age and NIH stroke scale. *Neurology* **80**, 21–28 (2013).
10. Hallevi, H. et al. Identifying patients at high risk for poor outcome after intra arterial therapy for acute ischemic stroke. *Stroke* **40**, 1780–1785 (2009).
11. Heo, J. et al. Abstract 194: machine learning-based model can predict stroke outcome. *Stroke* **49**, A194–A194 (2018).
12. Bacchi, S. et al. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study. *Acad. Radiol.* **27**, e19–e23 (2020).
13. Hamann, J. et al. Machine-learning-based outcome prediction in stroke patients with middle cerebral artery-M1 occlusions and early thrombectomy. *Eur. J. Neurol.* **28**, 1234–1243 (2021).
14. Akay, E. M. Z. et al. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. *Stroke* **54**, 1505–1516 (2023).
15. Seners, P., Turc, G., Oppenheim, C. & Baron, J.-C. Incidence, causes and predictors of neurological deterioration occurring within 24 h following acute ischaemic stroke: a systematic review with pathophysiological implications. *J. Neurol. Neurosurg. Psychiatry* **86**, 87–94 (2015).
16. Johnson, A. E. W. et al. Machine learning and decision support in critical care. *Proc. IEEE Inst. Electr. Electron Eng.* **104**, 444–466 (2016).
17. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
18. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir. Res.* **4**, e000234 (2017).
19. Hyland, S. L. et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).
20. Thorsen-Meyer, H.-C. et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Health* **2**, e179–e191 (2020).
21. Nielsen, A. B. et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *Lancet Digit. Health* **1**, e78–e89 (2019).
22. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
23. Vaswani, A. et al. Attention is All you Need. In *Advances in Neural Information Processing Systems* Vol. 30 (Curran Associates, Inc., 2017).
24. Islam, S. et al. A Comprehensive survey on applications of transformers for deep learning tasks. Preprint at http://arxiv.org/abs/2306.07303 (2023).
25. Li, Y. et al. Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inf.* **27**, 1106–1117 (2023).
26. Cutforth, M. et al. Acute stroke CDS: automatic retrieval of thrombolysis contraindications from unstructured clinical letters. *Front. Digit. Health* **5**, 1186516 (2023).
27. Rosario, H. D. et al. Applications of natural language processing for the management of stroke disorders: scoping review. *JMIR Med. Informatics* **11**, e48693 (2023).
28. Huang, R. et al. Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. *Comput. Biol. Med.* **155**, 106176 (2023).
29. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
30. Chen, S. et al. MSA-YOLOv5: multi-scale attention-based YOLOv5 for automatic detection of acute ischemic stroke from multi-modality MRI images. *Comput. Biol. Med* **165**, 107471 (2023).
31. Ayoub, M. et al. End to end vision transformer architecture for brain stroke assessment based on multi-slice classification and localization using computed tomography. *Comput. Med. Imaging Graph.* **109**, 102294 (2023).
32. Lo, C.-M. & Hung, P.-H. Predictive stroke risk model with vision transformer-based Doppler features. *Med. Phys.* **51**, 126–138 (2024).
33. Dai, L. et al. A clinically actionable and explainable real-time risk assessment framework for stroke-associated pneumonia. *Artif. Intell. Med.* **149**, 102772 (2024).
34. Antikainen, E. et al. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. *Sci. Rep.* **13**, 3517 (2023).
35. Breiman, L. Statistical modeling: the two cultures. *Stat. Sci.* **16**, 199–231 (2001).
36. Bonkhoff, A. K. & Grefkes, C. Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. *Brain* **145**, 457–475 (2022).
37. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inform. Process. Syst.* Vol. 30 (Curran Associates, Inc., 2017).
38. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining 1135–1144 (Association for Computing Machinery, 2016). https://doi.org/10.1145/2939672.2939778.
39. Zihni, E. et al. Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome. *PLoS ONE* **15**, e0231166 (2020).
40. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
41. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
42. Swiss Stroke Registry. https://www.neurovasc.ch/portrait/komitees/swiss-stroke-registry/.
43. van Swieten, J. C., Koudstaal, P. J., Visser, M. C., Schouten, H. J. & van Gijn, J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* **19**, 604–607 (1988).
44. Outcome measures in stroke. European Stroke Organisation. https://eso-stroke.org/outcome-measures-stroke-modified-rankin-scale-ordinal-logistic-regression/ (2017).
45. Ganesh, A., Luengo-Fernandez, R., Wharton, R. M. & Rothwell, P. M. Ordinal vs dichotomous analyses of modified Rankin Scale, 5-year outcome, and cost of stroke. *Neurology* **91**, e1951–e1960 (2018).
46. Bath, P. M. W. et al. Statistical analysis of the primary outcome in acute stroke trials. *Stroke* **43**, 1171–1178 (2012).
47. Birschel, P., Ellul, J. & Barer, D. Progressing stroke: towards an internationally agreed definition. *Cerebrovasc. Dis.* **17**, 242–252 (2003).
48. Liu, S. D., Rudd, A. & Davie, C. Hyper acute stroke unit services. *Clin. Med.* **11**, 213–214 (2011).
49. WHO. WHO Guidelines for Management of Stroke. https://extranet.who.int/ncdccs/Data/MNG_D1_1.%20Clinical%20guideline%20of%20Acute%20Stroke%20.pdf (2012).
50. Sundrani, S. et al. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. *Npj Digit. Med.* **6**, 1–10 (2023).
51. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
52. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785–794 (2016) https://doi.org/10.1145/2939672.2939785.
53. Johnson, A., Pollard, T. & Mark, R. MIMIC-III clinical database. *PhysioNet*, https://doi.org/10.13026/C2XW26 (2015).
54. Wang, J., Huang, X., Yang, L. & Li, J. National institutes of health stroke scale (NIHSS) annotations for the MIMIC-III database. PhysioNet https://doi.org/10.13026/GYJG-0T90 (2021).
55. Brott, T. et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* **20**, 864–870 (1989).
56. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
57. Lusted, L. B. Signal detectability and medical decision-making. *Science* **171**, 1217–1219 (1971).
58. Suits, D. B. Use of dummy variables in regression equations. *J. Am. Stat. Assoc.* **52**, 548–551 (1957).
59. Indredavik, B., Slørdahl, S. A., Bakke, F., Rokseth, R. & Håheim, L. L. *Stroke Unit Treat. Stroke* **28**, 1861–1866 (1997).
60. van Os, H. J. A. et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front. Neurol.* **9**, 784 (2018).
61. Xie, Y. et al. Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *Am. J. Roentgenol.* **212**, 44–51 (2019).
62. Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M. J. & Campbell, R. H. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS ONE* **14**, e0218942 (2019).
63. Brown, T. et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems* Vol. 33 1877–1901 (Curran Associates, Inc., 2020).
64. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. In *Proc. 30th International Conference on Machine Learning* III-1310-III-131 (ICML, 2013).
65. Zaheer, M. et al. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems* Vol. 33 17283–17297 (Curran Associates, Inc., 2020).
66. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling vision transformers. Preprint at http://arxiv.org/abs/2106.04560 (2022).
67. Mehta, S. & Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. Preprint at https://doi.org/10.48550/arXiv.2110.02178 (2022).
68. Yu, C., Chen, T., Gan, Z. & Fan, J. Boost vision transformer with GPU-friendly sparsity and quantization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 22658–22668 (IEEE Computer Society, 2023).
69. Rasley, J., Rajbhandari, S., Ruwase, O. & He, Y. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 3505–3506 (Association for Computing Machinery, 2020). https://doi.org/10.1145/3394486.3406703.
70. Liu, Y. et al. Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model. *Stroke* https://doi.org/10.1161/STROKEAHA.123.044072 (2023).
71. Elkind, M. S. V., Boehme, A. K., Smith, C. J., Meisel, A. & Buckwalter, M. S. Infection as a stroke risk factor and determinant of outcome after stroke. *Stroke* **51**, 3156–3168 (2020).
72. Lauritsen, S. M. et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif. Intell. Med.* **104**, 101820 (2020).
73. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
74. Cutillo, C. M. et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *Npj Digit. Med.* **3**, 1–5 (2020).
75. Ozkara, B. B. et al. Utilizing imaging parameters for functional outcome prediction in acute ischemic stroke: a machine learning study. *J. Neuroimaging*. https://doi.org/10.1111/jon.13194 (2024).
76. Herzog, L. et al. Deep learning versus neurologists: functional outcome prediction in LVO stroke patients undergoing mechanical thrombectomy. *Stroke* **54**, 1761–1769 (2023).
77. Klug, J. et al. Integrating regional perfusion CT information to improve prediction of infarction after stroke. *J. Cereb. Blood Flow Metab.* https://doi.org/10.1177/0271678X20924549 (2020).
78. Huang, Y.-C. et al. The impact factors on the cost and length of stay among acute ischemic stroke. *J. Stroke Cerebrovasc. Dis.* **22**, e152–e158 (2013).
79. Chang, K.-C. et al. Prediction of length of stay of first-ever ischemic stroke. *Stroke* **33**, 2670–2674 (2002).
80. Johnston, K. C. et al. Intensive vs Standard treatment of hyperglycemia and functional outcome in patients with acute ischemic stroke: the shine randomized clinical trial. *JAMA* **322**, 326–335 (2019).
81. Liu, L. et al. Early versus delayed antihypertensive treatment in patients with acute ischaemic stroke: multicentre, open label, randomised, controlled trial. *BMJ* **383**, e076448 (2023).

82. Olma, M. C. et al. Extent of routine diagnostic cardiac work-up at certified German stroke units participating in the prospective MonDAFIS study. *Neurol. Res. Pract.* **5**, 21 (2023).
83. Lees, K. R., Broderick, J. P., Selim, M. H. & Molina, C. A. Early vs. late assessment of stroke outcome. *Stroke* **47**, 1416–1419 (2016).
84. Lees, K. R. et al. Contemporary outcome measures in acute stroke research. *Stroke* **43**, 1163–1170 (2012).
85. Xiong, Y., Wakhloo, A. K. & Fisher, M. Advances in acute ischemic stroke therapy. *Circ. Res.* **130**, 1230–1251 (2022).
86. Filipovic, M. Patient decision making in anesthesia and intensive care medicine [Patientenwille in Anästhesie und Intensivmedizin]. in *Challenges in Anesthesia [Herausforderungen in der Anästhesie]* (Barbara Meyer-Zehnder, Thierry Girard, 2024) (In press).
87. Klug, J. Figure data for 'machine learning for early dynamic prediction of functional outcome after stroke'. zenodo https://doi.org/10.5281/zenodo.13694272 (2024).
88. Klug, J. & Leclerc, G. JulianKlug/OPSUM: OPSUM: 3-month outcome transformer. https://doi.org/10.5281/zenodo.11473805 (2024).
89. Klug, J. & Leclerc, G. Model weights for 'machine learning for early dynamic prediction of functional outcome after stroke'. https://doi.org/10.5281/zenodo.8195709 (2024).

Acknowledgements

This work was supported by the Swiss National Science Foundation (32003B_215285) and the Swiss Heart Foundation (FF 18071).

Author contributions

J.K. conceived the study, which was designed in detail by J.K., E.D., and E.C. Data was extracted by J.K., E.D. and E.C. Data pre-processing was done by J.K., and verification was done by J.K. and G.L. Algorithm design and testing were realized by J.K. and G.L. Data analysis was done by J.K., and interpretation was done by J.K., G.L., E.D., and E.C.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00666-w.

Correspondence and requests for materials should be addressed to Emmanuel Carrera.

Peer review information *Communications Medicine* thanks Joo Heung Yoon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. [Peer reviewer reports are available.].