Benchmarking Stroke Outcome Prediction through Comprehensive Data Analysis – NeuralCup 2023

Anna Matsulevits (Ph.D)[1,2], Pedro Alvez (Ph.D)[3], Manfredo Atzori (Ph.D)[4], Ahmad Beyh (Ph.D)[5], Maurizio Corbetta (MD, Ph.D)[3], Federico Del Pup (Ph.D)[3], Lilit Dulyan (Ph.D)[6], Chris Foulon (Ph.D)[1,2], Thomas Hope (Ph.D)[7], Stefano Ioannucci (Ph.D)[8], Gael Jobard (Ph.D)[1], Hervé Lemaitre (Ph.D)[1], Douglas Neville (Ph.D)[9], Victor Nozais (Ph.D)[1], Christopher Rorden (Ph.D)[10], Orionas-Vasilis Saprikis (Ph.D)[11], Igor Sibon (MD, Ph.D)[12], Christoph Sperber (Ph.D)[13], Alex Teghipco (Ph.D)[14], Bertrand Thirion (Ph.D)[15], Louis Fabrice Tshimanga (Ph.D)[4], Roza Umarova (Ph.D)[11], Ema Birute Vaidelyte (Ph.D)[16], Emiel van den Hoven (Ph.D)[11], Esteban Villar Rodriguez (Ph.D)[17], Andrea Zanola (Ph.D)[4], Thomas Tourdias (MD, Ph.D)*[18,19], Michel Thiebaut de Schotten (Ph.D)*[1,2]

Short title: **Benchmarking Stroke Outcome Prediction**

[1] Groupe d'Imagerie Neurofonctionnelle, Institut des Maladies Neurodégénératives-UMR 5293, Centre national de la recherche scientifique (CNRS), CEA, University of Bordeaux, Bordeaux, France.

[2] Brain Connectivity and Behaviour Laboratory, Sorbonne Universities, Paris, France.

[3] Department of Neurosciences and Mental Health, Neurology, Hospital de Santa Maria, CHLN, Lisbon, Portugal; Language Research Laboratory, Faculty of Medicine, Universidade de Lisboa, Lisbon, Portugal.

[4] Padova Neuroscience Center (PNC), University of Padova, Padova, Italy.

[5] Natbrainlab, Sackler Institute for Translational Neurodevelopment, Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, Psychology & Neuroscience, Denmark Hill, London SE5 8AF, UK; Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, UK.

[6] Donders Centre for Brain Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands.

[7]Wellcome Trust Centre for Neuroimaging, University College London, UK.

[8] Université de Fribourg, Visual and Cognitive Neuroscience Lab, Fribourg, Switzerland.

[9] Wellcome Centre for Human Neuroimaging, Department of Imaging Neuroscience, Institute of Neurology, University College London, WC1N 3AR, UK.

[10] McCausland Center for Brain Imaging, University of South Carolina, Columbia, SC, USA.

[11] University Medical Center Freiburg, Department of Neurology and Neurophysiology, Freiburg, Germany.

[12] Neurology Department, CHU Bordeaux, UMR-5287-CNRS, EPHE PSL Research University, Bordeaux, France.

[13] Inselspital, Universitätsspital Bern, Switzerland.

[14] Department of Psychology, University of South Carolina, Columbia, SC 29208, USA.

[15] Inria-Saclay, Palaiseau; Universite Paris Saclay; CEA Saclay, Paris, France.

[16] Universitäre Psychiatrische Dienste Bern (UPD) AG, Switzerland.

[17] Universitat Jaume, Department of Basic Psychology, Clinical Psychology and Psychobiology, Castelló de la Plana, Castelló, Spain.

[18] Centre Hospitalier Universitaire (CHU) de Bordeaux, Neuroimagerie diagnostique et thérapeutique, Bordeaux, France.

[19] University Bordeaux, INSERM, Neurocentre Magendie, U1215, Bordeaux, France.

* Thomas Tourdias and Michel Thiebaut de Schotten contributed equally to this work.

Correspondence to **Anna Matsulevits**

Neurodegeneratives Diseases Institute, CNRS UMR 5293, Université de Bordeaux – Case 28, Centre Broca Nouvelle Aquitaine, 146 rue Léo Saignat – CS 61292, F-33076 Bordeaux

E-mail: anna.matsulevits@hotmail.com

Abstract

Stroke is a significant cause of mortality and long-term disability worldwide, with variable recovery trajectories posing substantial challenges in anticipating post-event care and rehabilitation planning. The NeuralCup 2023 consortium was established to address these challenges by comparing the predictability of stroke outcome models through a collaborative, data-driven approach. This study presents the consortium's findings, which involved 15 participating teams worldwide. Using a comprehensive dataset, which included clinical and imaging data, we conducted an open competition to identify and compare predictors of motor, cognitive, and neuropsychological (emotional) outcomes one-year post-stroke. Analyses incorporated both traditional and novel methods, including machine learning algorithms. These efforts culminated in the search for 'optimal recipes' for predicting each domain through an exhaustive exploration of the features of all the approaches. Key predictors included lesion characteristics, T1-weighted MRI sequences, and demographic factors. Notably, integrating FLAIR imaging and white matter tract analysis emerged as crucial to improving the accuracy of cognitive and motor outcome predictions, respectively. These findings advocate for a tailored, multifaceted approach to stroke outcome prediction, underscoring the potential of collaborative data science in addressing complex neurological prognostication challenges. This study also sets a new benchmark methodology in stroke research, offering a foundational step toward personalized care strategies that could significantly impact recovery planning and quality of life for stroke survivors.

1. Introduction

Knowledge about a future event empowers individuals to adequately prepare and take measures to influence outcomes in their favor. In healthcare, mathematical and statistical modeling has enabled the use of past observations to anticipate outcomes following various life-changing events such as brain damage. Stroke, as the second leading cause of death, has seen a 70% increase in incidence in the past two decades[1], resulting in 6.55 million deaths and 101 million survivors in 2019[1]. Predictive frameworks in stroke are still invaluable for providing realistic prognostic expectations to the patients and their families, accurately planning post-acute care and enhancing clinical research trials by identifying patients with an expected homogenous outcome. This will allow to counter-balance individuals with similar outcome trajectories in clinical trials, increasing the statistical power of such studies.

Forecasting behavioral outcomes post-stroke is a longstanding challenge[2], but the optimal predictive biomarkers, neuroimaging methodology, or algorithms are difficult to pinpoint due to significant variations in study resources, methods, and purpose. Neuroimaging has been promising in revealing brain alterations due to infarct, particularly helping individual prediction. Previous work has demonstrated the importance of stroke volumes[3], stroke location[4], disconnection[5], functional pattern[6], co-existence of small vessel disease[7], and pre-existing brain atrophy[8]. These insights influence prediction parameters (for a review, see[9,10], though methodologies vary from conventional regression models[11] to advanced machine learning algorithms[5]. However, few frameworks consider cognitive and neuropsychological outcomes on top of the more visible motor outcomes[12], and most focus on 3-month outcomes, limiting long-term outcome investigations. In addition, many studies mislabel statistical associations as 'predictions'[13,14] which do not forecast accurate outcomes in new data. Advancing precision medicine requires data-driven approaches that assess model performance on unseen data to predict outcomes in new cohorts[15]. Yet, limited efforts place their existing predictive methods into a comparative context using the same dataset. Recently, systematic reviews have compared outcome prediction[16,17] but have often excluded machine learning algorithms and failed to provide comprehensive comparisons.

To address these gaps, we organized an open competition involving 15 teams of experts in stroke outcome prediction from across the globe at the NEURAL conference in Bordeaux in May 2023. Participants used a comprehensive dataset, including motor, cognitive, and neuropsychological scores one year after stroke, as well as neuroimaging, lesion, and demographic data to predict outcomes on an unseen out-of-sample dataset. Following this, we conducted a statistical analysis of the input and method combinations to derive the best recipe for predicting a wide range of stroke outcomes.

2. Methods

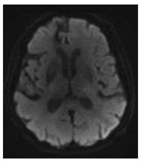
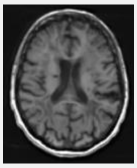
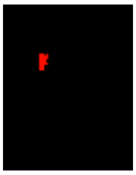
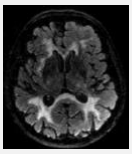
Dataset

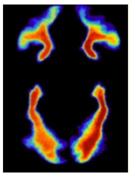
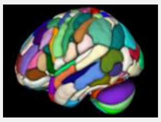
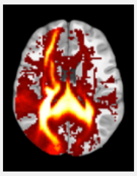
The dataset used for the stroke outcome prediction is based on a prospective observational study, "Brain Before Stroke (BBS)" cohort, conducted between June 2012 and February 2015. The main results and secondary objectives have already been published[11,18–20]. The study was approved by the local research ethics committee, and written consent was obtained from the patients before inclusion. The BBS cohort has not been previously used to compare different prognosis models side-by-side, which is the objective of the present study.

Briefly, the BBS cohort recruited consecutive stroke patients who underwent clinical and MRI evaluations at initial (24 to 72 hours) and chronic (1 year) states. Primary inclusion criteria were men and women over 18 years old with a clinical diagnosis of minor-to-severe supratentorial cerebral infarct (NIHSS between 1 and 25) at the stroke onset. Exclusion criteria were history of symptomatic cerebral infarct with a functional deficit (pre-stroke modified Rankin Scale [mRS] score ≥1), infratentorial stroke, history of severe cognitive impairment (dementia), or psychiatric disorders according to axis 1 of the DSM-IV[21] criteria except for major depression, coma, pregnant or breast-feeding women, and contraindications to MRI.

We mainly aimed to compute and compare imaging modalities' predicting abilities but also included demographic data such as age and sex, knowing their predictive value[22]. The baseline imaging used for prediction consisted of MRI performed between 24 and 72 hours after stroke onset on a 3T scanner (Discovery MR750w, GE Healthcare). We provided the raw Diffusion-Weighted Images (DWI) and the infarct masks that had already been manually delineated[20] regarding the critical role of stroke volume and location in the outcome[12,23]. Additionally, we included 3D FLAIR and 3D T1 sequences, which, while capable of detecting stroke-related changes at this stage, offer valuable insights into the overall integrity of the brain beyond the lesion site[24]. Furthermore, the teams were free to use any openly available brain atlases[25,26] or additional tools to compute supplementary inputs such as the white matter tracts and/or their disconnection[27–30] by the infarct. In summary, each team was free to include any input combination in their predictive models, which resulted in the following 9 different inputs: “age”, “gender”, “DWI”, “T1”, “segmented lesions”, “FLAIR”, “tracts”, “parcellation atlases”, and “disconnectome”. The parameters of acquisition and additionally used inputs are summarized in Table 1.

Table 1. Inputs provided for the outcome prediction, as well as additional inputs (tracts, parcellation atlases, disconnectome) used by the teams.

Inputs	Data dimensionality	Data example
age	continuous	64.2849315 years
gender	discrete	female (1)
DWI	3D nifti image, continuous intensity values	
T1	3D nifti image, continuous intensity values	
segmented lesions	3D nifti image, binary values	
FLAIR	3D nifti image, continuous intensity values	

tracts	3D nifti image	
parcellation atlases	3D nifti image	
disconnectome	3D nifti image, continuous probability values	

For 1-year post-stroke, teams had to predict motor, cognitive, and psychological outcomes. Motor outcome was based on an expanded version of the Fugl-Meyer (FM) score (range: 0–242) which provides a detailed evaluation of upper and lower limb functions with an excellent inter and intra-rater reliability[31]. Cognitive outcome was assessed with the Montreal Cognitive Assessment (MoCA) which is a validated tool for screening cognitive impairment after stroke[32] (range: 0–30). To assess more details on the affected domains, we also considered the sub-scores of the MoCA (Attention, Visuospatial, Denomination, Language, Abstraction, Orientation, Recall) and the Isaacs test set (IST) which evaluates the executive functions through categorical verbal fluency[33] (range: 0–max. number of orally produced words for different categories within 1 minute). Psychological outcome was assessed using the Hospital Anxiety and Depression Scale (HAD-Anxiety and HAD-Depression)[34] (range: 0–21 for each test).

From the original BBS dataset, we excluded individuals with inadequate baseline neuroimaging or missing 1-year values resulting in 237 patients (see Supplementary Figure 1 for a detailed flowchart). The remaining data was split into a training dataset (n=187) and an out-of-sample validation dataset (n=50) through pseudorandomization to ensure comparable distribution of the outcome scores (Supplementary Figure 2). Table 2 provides patients' demographics and neuropsychological scores of the whole dataset.

Table 2. Patient demographics and average scores (with standard deviations) for the training and validation cohort.

	training		**validation**	
Score	**Mean ($\bar{x}$)**	**SD**	**Mean ($\bar{x}$)**	**SD**

Age	65,44	13,48	66,29	12,61
FM	224,93	30,43	224,66	31,16
MoCA	24,13	4,93	24,38	5,78
IST	30,48	6,41	30,66	9,23
HAD-A	6,07	3,80	5,78	4,27
HAD-D	4,54	3,98	4,10	3,03

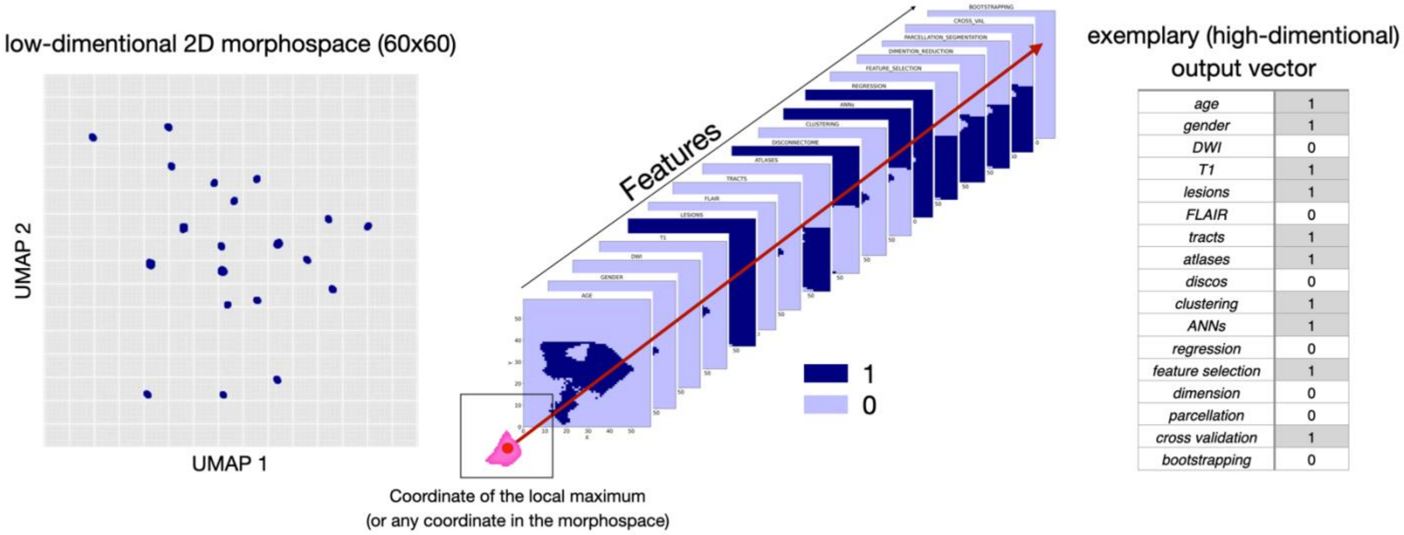
Evaluation of the prediction outcomes

The participating teams were provided with data and instructed to produce clinical score predictions for 50 out-of-sample patients. There were no restrictions placed on the use of inputs or methods. The teams' performances were primarily ranked based on the overall Pearson mean R^2 outcome[35], but we also examined additional statistical outcomes such as the Mean Squared Error (MSE) and Mean Absolute Error (MAE) losses (for formulas see Supplementary Formula 1 and Formula 2). To assess clinical relevance, we also used the area under the receiver operating characteristic (ROC) curve for motor, cognitive, and psychological tests (FM, MoCA, IST, HAD-A, and HAD-D) with clinically validated thresholds to define poor outcomes as follows: FM ≤ 100, MoCA ≤ 25, IST ≤ 28, HAD-A ≥ 8, HAD-D= ≥ 8.

Exhaustive evaluation of the used approaches

After evaluating teams' performances, we aimed to quantify which combinations of inputs and methods led to the most accurate stroke outcome prediction. The teams utilized different methods which were categorized into eight main classes that represent different strategies to obtain the prediction *(Artificial Neural Networks, Regression)*, to extract and represent the data (*Clustering, Feature Selection, Dimension Reduction, Parcellation/Segmentation)*, and to validate the prediction *(Cross Validation, Bootstrapping)*. More details on each class, alongside examples of specific techniques used by the participating teams, can be found in Supplementary Table 1.

To explore the similarities and differences between the teams we used Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)[36] on a matrix summarizing the observed feature combinations and obtained R^2 for each score. This created a 2D discrete morphospace – 60x60 grid to allow statistics on the UMAP latent space – where teams were localized based on the combinations of inputs (among the nine listed in Table 1) and methods (among the eight listed in Supplementary Table 1) they used. Similar approaches were located close to each other while differing methods were further apart. The low dimensionality of the space allowed us to utilize the matrix for the next analysis step. Following previous work[5], using the FSL tool *randomise*[37] we identified areas in the morphospace that were associated with a high R^2 score for each test. The *randomise* analysis yielded t-statistic maps distinguishing significant from non-significant regions, marking the presumably 'best' prediction points that we defined based on the local maxima. With the UMAP inverse transformation (*inverse transforms*) it is possible to generate a high dimensional data sample given a location in the low dimensional embedding space, meaning we can obtain an input data vector from coordinates in the morphospace even for coordinates that are not corresponding to the initial input data points. Therefore, we then used the UMAP inverse for the combination of variables that were assigned to the points where the t-value for the *randomise* test was the highest, associated with the best-predicting performance. After applying the inverse option of the analysis, we obtained a separate matrix for each feature, each containing 3600 values (60x60 space). Every value in the matrix represented a point in the UMAP space, corresponding to one out of seventeen (total number of features: 8 inputs + 9 methods) components of a potential feature vector. We binarized each feature matrix using the threshold 0.5 and concatenated the seventeen feature matrices that summarize the inputs and methods we were investigating. Subsequently, we were able to obtain a combination of used and discarded features for the previously identified local maxima (as well as any other coordinate in the space), revealing an estimation of the theoretically best combination for accurately predicting the one-year outcome for each clinical domain (see Figure 1).


low-dimentional 2D morphospace (60x60)
UMAP 2
UMAP 1
Features
AGE
GENDER
DWI
T1
LESIONS
FLAIR
TRACTS
ATLASES
DISCONNECTOME
CLUSTERING
ANNs
REGRESSION
FEATURE_SELECTION
DIMENTION_REDUCTION
PARCELLATION_SEGMENTATION
CROSS_VAL
BOOTSTRAPPING
1
0
Coordinate of the local maximum
(or any coordinate in the morphospace)
exemplary (high-dimentional) output vector

age	1
gender	1
DWI	0
T1	1
lesions	1
FLAIR	0
tracts	1
atlases	1
discos	0
clustering	1
ANNs	1
regression	0
feature selection	1
dimension	0
parcellation	0
cross validation	1
bootstrapping	0

Figure 1: Visualization of the process for obtaining the theoretically optimal feature combination for predictions: inversing the *randomise* analysis yields heatmaps for each analyzed feature. After binarizing these maps, we investigated the overlap of the local maxima of the clinical tests with the binarized feature maps (1 representing the presence of the feature, and 0 representing the absence of a feature in the final combination 'optimal recipe').

3. Results

Patient characteristics, participating teams and their prediction approaches

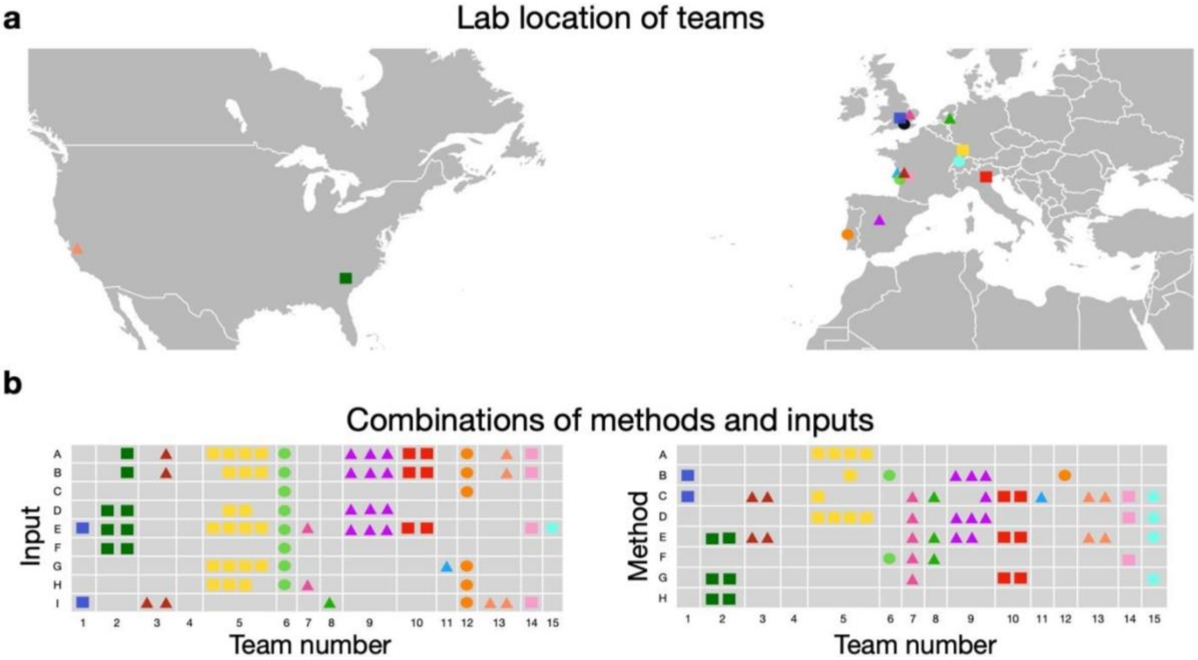


Figure 2: Summary of participating teams and the approaches taken for all predictions. **a** Locations of the teams' affiliated labs. **b** Summary of different inputs (A: age, B: gender, C: DWI, D: T1, E: lesion, F: FLAIR, G: tracts, H: atlases, I: disconnectome) and methods (A: clustering, B: artificial neural networks, C: regression, D: feature selection, E: dimensionality reduction, F: parcellation, G: cross-validation, H: bootstrap) used for each prediction. (Figure modified from[38]).

Comparison of the evaluation metric R^2

The performance of each team's model was evaluated by comparing the predicted scores for the 50 out-of-sample patients with their actual recorded scores. The mean proportion of variance explained (R^2) was calculated for each of the 24 prediction models (Figure 3a) from the 6 main scores: FM motor, FM total, MoCa, IST, HAD-A, and HAD-D. Team 5, in their third prediction obtained the highest overall R^2= 0.311, followed by the same team's first prediction with R^2 =0.238, and the second prediction of team 2 (R^2 =0.182). The mean R^2 values displayed a strong disparity of the predictive performances according to the type of outcome. The prediction of the motor outcome at one year could reach R^2 as high as 0.611 for team 5. On the other hand, the prediction of the neuropsychological outcome was the worst with R^2 not higher than 0.034 to anticipate post-stroke depression (HAD-D) and 0.138 for post-stroke anxiety (HAD-A). The predictive performances of cognitive status were intermediate between motor and mood prediction. Exhaustive information describing the results of all teams can be found in the supplementary material (Supplementary Table 2 and Supplementary Table 3).

As secondary analyses, we also looked at how each team's model could classify patients' outcomes as poor or good based on clinically relevant thresholds. The ROCs of the five best models are shown in

Supplementary Figure 3 and confirmed the highest performances for motor prediction followed by cognitive and mood predictions. As an additional metric for evaluation, we calculated the mean average errors (MAEs) and mean squared errors (MSEs) of the teams' predictions for the main test scores, which can be found in Supplementary Table 6.

Evaluation with UMAP and randomise approach

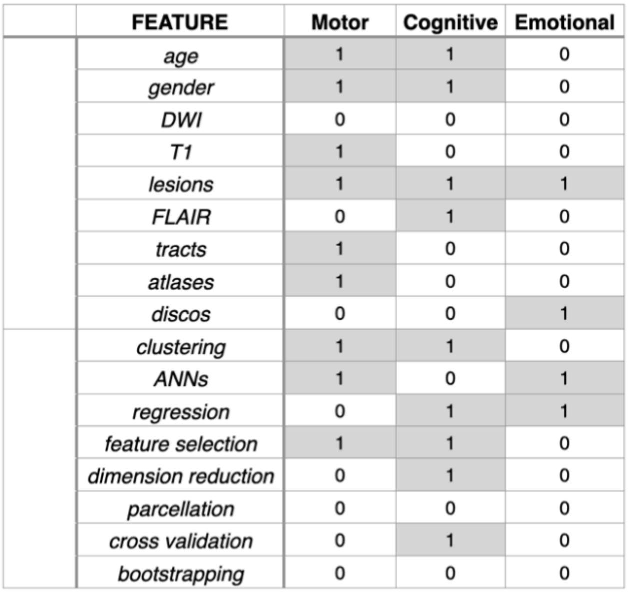
To investigate the nuances of different combinations of features (input variables and methodological approaches) used by the teams, we reduced their interactions into a UMAP morphospace for each outcome. These representations, combined with t-statistic maps (that indicate locations of the space associated with higher R^2) highlight how close or far each of the 24 models is from the best prediction. The observed pattern indicated that overall, the proximity of the models' placement in the morphospace was closer to the local maxima for motor outcomes and cognitive outcomes compared to psychological outcomes (Figure 3c and d). In other words, the prediction models show generally closer distances to the local maxima (indicated through the highest t-value) of the test scores FM, MoCA, and IST compared to the scores HAD-A and HAD-D, which is reflected in the achieved prediction accuracies.

Moving back from the local maximum in the morphospace to the combination of features associated with this location, we identified the optimal combinations of features (defined as the best approaches within the boundaries of the given context) for each of the motor, cognitive, and neuropsychological domains (see Table 3 and Figure 1). The overview reveals that the best predictions always required the lesion mask as an input. Adding information about the overall status of the brain by including the FLAIR modality improved predictions on the cognitive domain whereas utilizing white matter tracts and atlases seem to be beneficial for anticipating motor outcomes. Conversely, predicting higher cognitive outcomes in the neuropsychological domain benefited more from a global disconnectivity analysis.

Investigating the methodological approaches, the most prominent and successful methods were clustering, ANNs (Artificial Neural Networks), regression, and feature selection. However, the utilization of the different methods varies (e.g., different clustering), making comparison less straightforward than with inputs. For an exhaustive listing of combinations for each single score of FM, MOCA, IST, HAD-A, and HAD-D, please refer to Supplementary Table 4.

Examining the obtained t-statistics maps from the first step, we can identify separate clusters of significant t-values (Supplementary Figure 4). Repeating the above-described procedure for the local maxima of each cluster separately led to slightly different feature combinations for different clusters of the same test. For instance, the analysis of the local maximum of one cluster (coordinates: 25, 42) that was highly predictive for MoCA resulted in the combination of FLAIR, clustering, feature selection, and dimension reduction while the analysis of the second cluster (local maximum coordinates: 44, 28) resulted in a different combination of features, that included age, sex, T1, lesions, ANNs, regression, and feature selection (Supplementary Figure 4). This example, along with similarly varying results for different significant clusters of the HAD-A and HAD-D tests, demonstrates that there is potentially more than one recipe that leads to a decent prediction of a score. The differences found in the feature combinations for the tests also show that there is not one single recipe for an overall good prediction that covers all stroke outcomes. Instead, different combinations of features work differently well depending on the type of outcome being predicted.

Table 3. Summary of the optimal combination of features to use for long-term (1 year) predictions of motor, cognitive, and emotional stroke outcomes.



	FEATURE	Motor	Cognitive	Emotional
	age	1	1	0
	gender	1	1	0
	DWI	0	0	0
	T1	1	0	0
	lesions	1	1	1
	FLAIR	0	1	0
	tracts	1	0	0
	atlases	1	0	0
	discos	0	0	1
	clustering	1	1	0
	ANNs	1	0	1
	regression	0	1	1
	feature selection	1	1	0
	dimension reduction	0	1	0
	parcellation	0	0	0
	cross validation	0	1	0
	bootstrapping	0	0	0

Note: Resulting combinations of features to use for the motor, cognitive, and neuropsychological (emotional) domains. The first column of the table lists the features (inputs and methods) that were investigated, followed by the columns ‘Motor’, ‘Cognitive’, and ‘Emotional’ that represent the domains of the neuropsychological scores the prediction performance of the teams was evaluated on. Gray cells containing ‘1’ indicate which features belong to the optimal, ‘theoretically best’ combination that leads to the best prediction of a given domain. White cells containing ‘0’ represent the features that do not contribute to the optimal feature combination.

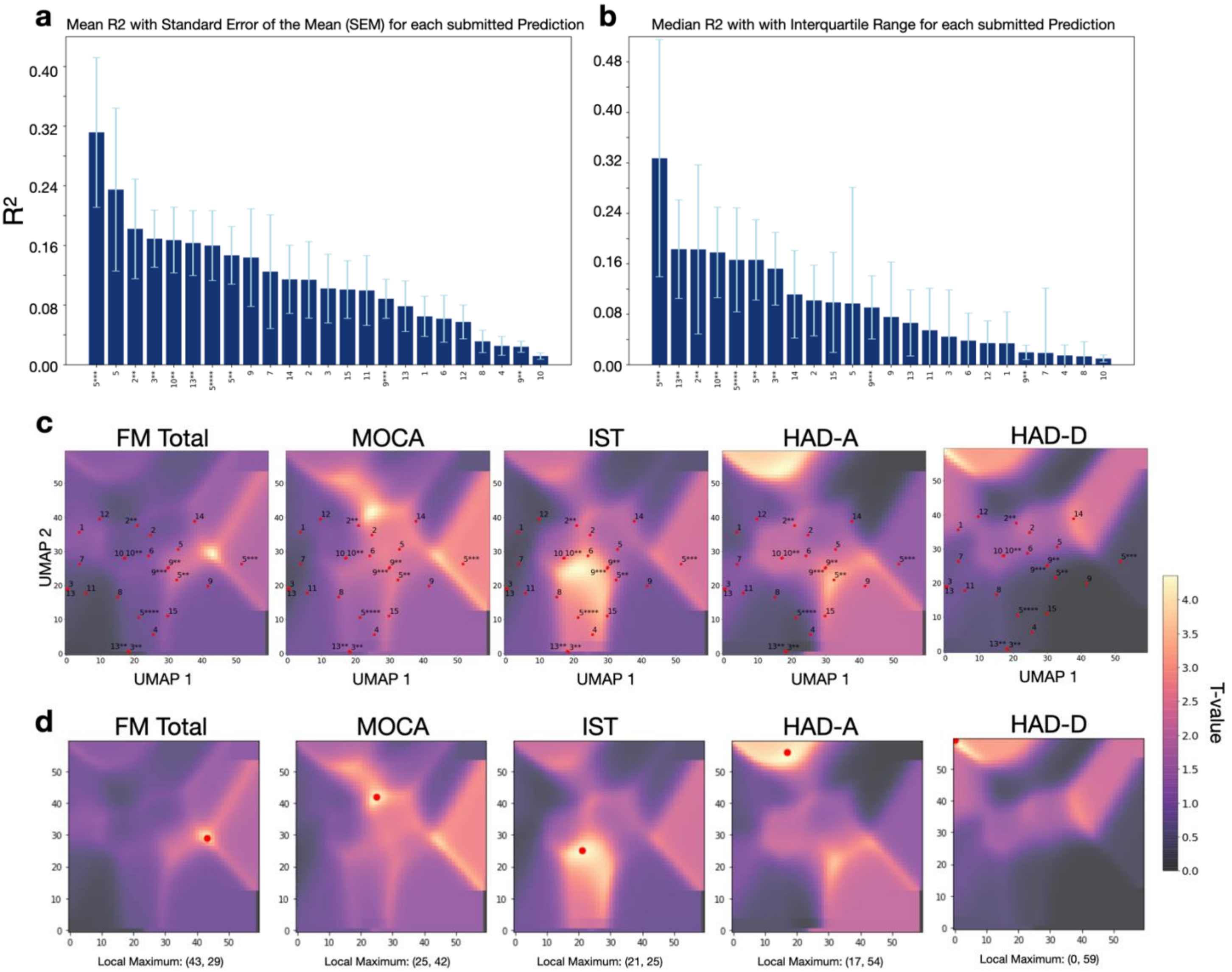


Figure 3: **a** Mean R^2 comparison for all submitted predictions (motor, cognitive, and psychological outcomes) of five neuropsychological scores (FM total, MoCA, IST, HAD-A, HAD-D) sorted ascendingly from the highest score to the lowest score across all teams whose number is indicated on the x-axis. The stars indicate the prediction number, the whiskers indicate the standard error of the mean (SEM). **b** Median R^2 comparison for all submitted predictions of the same scores, sorted ascendingly with whiskers indicating the interquartile range (IRQ). **c** T-statistic maps for each clinical outcome test, obtained from the FSL *randomise* analysis. The yellow regions indicate a significant t-value, the purple regions indicate a non-significant t-value. The UMAP distribution of all teams is plotted on the t-statistic maps. **d** Local maximum of the t-statistic map for each analyzed outcome score.

4. Discussion

Our study demonstrates the feasibility of making long-term predictions for various stroke outcomes. By analyzing different feature combinations, we have discovered that there is no one-size-fits-all solution for predicting stroke outcomes. Instead, the effectiveness of various approaches varies depending on the specific outcome domain being predicted, such as motor, cognitive, or neuropsychological scores. Our systematic comparison of observed and non-observed feature utilization (through UMAP projection and its inverse) is a novel approach to benchmark different models. It has initiated an exploration into the interaction of inputs and methods, contributing to the ongoing effort to establish a consensus in this field.

The principal conclusion of this manuscript derives from an in-depth analysis of the methodologies employed by the participating teams in constructing their predictive models and the identification of the features instrumental in facilitating optimal predictions. This investigation has elucidated distinct ‘optimal recipes’ for various outcome scores, which reflect the different brain mechanisms in response to different tasks. These findings underscore that not all inputs and methodologies exhibit equivalent efficacy in the precise prognostication of specific stroke outcomes. Lesion-symptom mapping, with its rich history in studying the brain structure-function relationships, remains a widely used approach for analyzing and explaining stroke outcomes[39]. In our case, for predictions of all domains (motor, cognitive, emotional), the location and size of the lesion were consistently identified as pivotal factors.

Some other factors were specific to a domain. For instance, the prediction of cognitive scores, such as the MoCA, was found to derive greater benefit from FLAIR imaging. This modality not only delineates the infarct but also captures underrepresented characteristics of cerebral damage beyond the lesion site, particularly white matter hyperintensities indicative of small vessel disease[40]. Such proxies of brain health[41], if applied with the appropriate combination of methods (feature selection, dimension reduction, and clustering), could provide sufficient information to predict outputs on the cognitive domain. The importance of FLAIR imaging in predicting MoCA scores is consistent with the previous literature correlating cerebral integrity and potential frailty with cognitive decline and suboptimal recovery post-stroke[42] , attributed to diminished neural plasticity beyond the lesion site[43]. Recently, a research group linked microstructural and macrostructural biomarkers from the normal-appearing brain matter in FLAIR images (texture, intensity, volume) to cognitive function[43] supporting this hypothesis. In exploring the most predictive motor domain, it is noteworthy that motor outcome predictions uniquely benefit from incorporating additional data derived from tracts and atlases. This observation aligns well with foundational research conducted over a century ago, highlighting the critical role of structures now known as the corticospinal tract in motor impairment[1].

In contrast to post-stroke motor impairments that are visible and have been studied exhaustively over the last decades[44–46], higher-order symptoms such as depression or anxiety have only recently been recognized and included in the clinical assessment for stroke patients[47,48]. Thus, prediction approaches used for motor impairment studies cannot simply be translated to other functions yielding the same results. In our analysis, while out-of-sample motor score predictions exceeded the predictive accuracy of several comparable attempts reported in the literature[49–51], depression, and HAD-Anxiety marked the least predictive outcomes with the evaluation scores of $R^2_{\text{HAD-D}}$ =0.034 and $R^2_{\text{HAD-A}}$=0.138. Being able to similarly predict emotional outcomes is crucial, given that one in three stroke survivors is affected by post-stroke depression[52]. Unfortunately, cognitive alterations and mood disorders are frequently overlooked, while they are associated with suboptimal recovery, increased risk of a further stroke, decreased quality of life, and mortality[53,54]. The importance of depression as an independent predictor for functional long-term stroke outcome was already established over two decades ago[55]. Hence, psychological outcomes following a stroke represent a crucial component for further exploration and prediction that are likely to benefit from more complex clinic-radiological models that include additional variables (e.g., genetic vulnerability, biological markers) to capture and explain more variance.

Additionally, more than one recipe could predict cognitive and psychological scores with comparable accuracy (see Supplementary Figure. 4, Supplementary Table 5). These distinct patterns of feature combinations can be attributed to different factors associated with the same impairment. This demonstrates that the 'optimal recipes', defined as the most efficient or effective solution under a given set of conditions, we have identified are not static but amenable to refinement with integrating a broader array of inputs and methods.

With the novel analytical framework selected for this study, we were able to shed light on the observed and estimate the non-observed feature combinations, suggesting avenues for future investigations. Nonetheless, these insights are constrained by the scope of the applied methodologies and clinical data's utilization (and availability). Therefore, it remains crucial to advocate for embracing open-science principles in the community, and we have made the training dataset available to encourage others to evaluate their predictive models against our cohort. By starting a crowd-sourcing initiative, we are committed to improving stroke outcome predictions and encouraging participation from other teams (to participate or test your algorithm, download the dataset on our website [http://neuralcup.bcblab.com]). The interplay of features helpful in predicting stroke outcomes of diverse domains warrants deeper investigation, and collaboration, methodological exchanges, and data-sharing in the field will greatly advance our understanding.

While investigating predictive inputs and methods is a landmark in stroke outcome research, this study acknowledges its constraints. First and foremost, the quality of the prediction is dependent on the specificity of the behavioral and cognitive assessment. In the present study, while the behavioral assessment might be considered standard neurological practice, the granularity of the cognitive measures might have hampered the predictions. Future sharing initiatives should provide a larger dataset with a more in-depth cognitive examination. Embedding the teams' observed approaches into the UMAP framework is an innovative step in visualizing potential feature combinations that were directly observed. However, the selection of features included is inherently restricted by the provided data. While we had close to 300 patients with homogeneous data that included acute imaging as well as a chronic behavioral follow-up, we must acknowledge that comparisons to the high number of combinations of features limit the sample size[56]. Additionally, the results reflect the methodologies of 24 prediction submissions, disproportionately influenced by the highest-performing team, thus not representing the full spectrum of possible outcomes. It is essential to provide additional input modalities, such as education level, and employ distinct methodological frameworks to capture a more comprehensive picture. For this, we have created an initiative that facilitates wider participation, increasing the sample size and subsequently aiming to refine the method categories into more nuanced classifications. This could unravel a finer detail of advantageous and suboptimal feature combinations and methodologies for stroke outcome prediction. Another constraint could be raised regarding the selection of neuropsychological test scores and the depth of reported results. Although the UMAP analysis was performed on the complete dataset comprising 13 score predictions (when including the subscores of MOCA), detailed results were only elaborated for five (FM, MoCA, IST, HAD-A, and HAD-D). This was motivated by the comprehensive interpretability of these selected scores across motor, cognitive, and neuropsychological (emotional) impairments and their clinical relevance backed by validated thresholds aiding in predictive model verification and comparison across the field.

Although the study presents certain limitations in quantifying the methodological approaches, it is a pioneering systematic effort to incorporate complete methodologies, exhausting all available variables. This research lays the groundwork for future investigations to enhance the accuracy of stroke outcome predictions. Having demonstrated the viability of out-of-sample predictions, we actively encourage contributions to refine our collective understanding and predictive accuracy for stroke outcomes.

This consortium has established a new benchmark in out-of-sample stroke outcome prediction. We successfully compared the long-term (one-year post-stroke) forecasts across three distinct domains, surpassing prior documented results. Our findings reveal that a universal prediction strategy for stroke outcome is less effective than employing tailored approaches for each domain or score to achieve the most accurate predictions, holding the promise to improve healthcare for stroke survivors.

Funding

This work is supported by HORIZON- INFRA-2022 SERV (Grant No. 101147319), “EBRAINS 2.0: A Research Infrastructure to Advance Neuroscience and Brain Health”, by the European Union’s Horizon 2020 research and innovation programme under the European Research Council (ERC) Consolidator grant agreement No. 818521 (DISCONNECTOME), the University of Bordeaux’s IdEx ‘Investments for the Future’ program RRI ‘IMPACT’, and the IHU ‘Precision & Global Vascular Brain Health Institute – VBHI’ funded by the France 2030 initiative (ANR-23-

IAHU-0001), and OHBM (Organisation for Human Brain Mapping). CR and AT were supported by the NIH of the USA (RF1-MH133701 and P50-DC01466).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol* **20**, 795–820 (2021).
2. Oxbury, J. M., Greenhall, R. C. & Grainger, K. M. Predicting the outcome of stroke: acute stage after cerebral infarction. *BMJ* **3**, 125–127 (1975).
3. Vogt, G., Laage, R., Shuaib, A. & Schneider, A. Initial Lesion Volume Is an Independent Predictor of Clinical Stroke Outcome at Day 90. *Stroke* **43**, 1266–1272 (2012).
4. Zhao, L. *et al.* Strategic infarct location for post-stroke cognitive impairment: A multivariate lesion-symptom mapping study. *J Cereb Blood Flow Metab* **38**, 1299–1311 (2018).
5. Talozzi, L. *et al.* Latent disconnectome prediction of long-term cognitive-behavioural symptoms in stroke. *Brain* **146**, 1963–1978 (2023).
6. Kwakkel, G. & Kollen, B. J. Predicting Activities after Stroke: What is Clinically Relevant? *International Journal of Stroke* **8**, 25–32 (2013).
7. Ryu, W.-S. *et al.* Stroke outcomes are worse with larger leukoaraiosis volumes. *Brain* **140**, 158–170 (2017).
8. Pinguet, V. *et al.* Pre-existing brain damage and association between severity and prior cognitive impairment in ischemic stroke patients. *Journal of Neuroradiology* **50**, 16–21 (2023).
9. Bartolomeo, P. & Thiebaut de Schotten, M. Let thy left brain know what thy right brain doeth: Inter-hemispheric compensation of functional deficits after brain damage. *Neuropsychologia* **93**, 407–412 (2016).
10. Moore, M. J., Demeyere, N., Rorden, C. & Mattingley, J. B. Lesion mapping in neuropsychological research: A practical and conceptual guide. *Cortex* **170**, 38–52 (2024).
11. Coutureau, J. *et al.* Cerebral Small Vessel Disease MRI Features Do Not Improve the Prediction of Stroke Outcome. *Neurology* **96**, e527–e537 (2021).
12. Munsch, F. *et al.* Stroke Location Is an Independent Predictor of Cognitive Outcome. *Stroke* **47**, 66–73 (2016).
13. Celap, I., Nikolac Gabaj, N., Demarin, V., Basic Kes, V. & Simundic, A.-M. Genetic and lifestyle predictors of ischemic stroke severity and outcome. *Neurol Sci* **40**, 2565–2572 (2019).
14. Bzdok, D., Engemann, D. & Thirion, B. Inference and Prediction Diverge in Biomedicine. *Patterns (N Y)* **1**, 100119 (2020).
15. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun* **15**, 1829 (2024).

16. Kremers, F. *et al.* Outcome Prediction Models for Endovascular Treatment of Ischemic Stroke: Systematic Review and External Validation. *Stroke* **53**, 825–836 (2022).

17. Johnston, K. C., Connors, A. F., Wagner, D. P. & Haley, E. C. Predicting Outcome in Ischemic Stroke. *Stroke* **34**, 200–202 (2003).

18. Kuchcinski, G. *et al.* Thalamic alterations remote to infarct appear as focal iron accumulation and impact clinical outcome. *Brain* **140**, 1932–1946 (2017).

19. Sagnier, S. *et al.* Chronic Cortical Cerebral Microinfarcts Slow Down Cognitive Recovery After Acute Ischemic Stroke. *Stroke* **50**, 1430–1436 (2019).

20. Linck, P. A. *et al.* Neurodegeneration of the Substantia Nigra after Ipsilateral Infarct: MRI R2* Mapping and Relationship to Clinical Outcome. *Radiology* **291**, 438–448 (2019).

21. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. (American Psychiatric Association, 2013). doi:10.1176/appi.books.9780890425596.

22. Jampathong, N., Laopaiboon, M., Rattanakanokchai, S. & Pattanittum, P. Prognostic models for complete recovery in ischemic stroke: a systematic review and meta-analysis. *BMC Neurol* **18**, 26 (2018).

23. Karnath, H.-O., Rennig, J., Johannsen, L. & Rorden, C. The anatomy underlying acute versus chronic spatial neglect: a longitudinal study. *Brain* **134**, 903–912 (2011).

24. Appleton, J. P. *et al.* Imaging markers of small vessel disease and brain frailty, and outcomes in acute stroke. *Neurology* **94**, e439–e452 (2020).

25. MRI Atlas of Human White Matter. *AJNR Am J Neuroradiol* **27**, 1384–1385 (2006).

26. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

27. Matsulevits, A. *et al.* Deep Learning disconnectomes to accelerate and improve long-term predictions for post-stroke symptoms. 2023.09.12.557396 Preprint at https://doi.org/10.1101/2023.09.12.557396 (2023).

28. Foulon, C. *et al.* Advanced lesion symptom mapping analyses and implementation as BCBtoolkit. *GigaScience* **7**, giy004 (2018).

29. Kuceyeski, A., Maruta, J., Relkin, N. & Raj, A. The Network Modification (NeMo) Tool: elucidating the effect of white matter integrity changes on cortical and subcortical structural connectivity. *Brain connectivity* **3**, 451–463 (2013).

30. Griffis, J. C., Metcalf, N. V., Corbetta, M. & Shulman, G. L. Lesion Quantification Toolkit: A MATLAB software tool for estimating grey matter damage and white matter disconnections in patients with focal brain lesions. *NeuroImage: Clinical* **30**, 102639 (2021).

31. Gladstone, D. J., Danells, C. J. & Black, S. E. The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabil Neural Repair* **16**, 232–240 (2002).

32. Pendlebury, S. T., Mariz, J., Bull, L., Mehta, Z. & Rothwell, P. M. MoCA, ACE-R, and MMSE Versus the National Institute of Neurological Disorders and Stroke–Canadian Stroke Network Vascular Cognitive Impairment Harmonization Standards Neuropsychological Battery After TIA and Stroke. *Stroke* **43**, 464–469 (2012).

33. Isaacs, B. & Kennie, A. T. The Set test as an aid to the detection of dementia in old people. *Br J Psychiatry* **123**, 467–470 (1973).

34. Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta Psychiatr Scand* **67**, 361–370 (1983).

35. Wright, S. (1921). Correlation and causation. Part I Method of path coefficients. Journal of Agricultural Research, 20, 557-585.

36. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).

37. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *NeuroImage* **92**, 381–397 (2014).

38. Maier-Hein, K. H. *et al.* The challenge of mapping the human connectome based on diffusion tractography. *Nat Commun* **8**, 1349 (2017).

39. Forkel, S. J. Lesion-Symptom Mapping: From Single Cases to the Human Disconnectome. in *Encyclopedia of Behavioral Neuroscience, 2nd edition* 142–154 (Elsevier, 2022). doi:10.1016/B978-0-12-819641-0.00056-6.

40. Longitudinal Changes of White Matter Hyperintensities in Sporadic Small Vessel Disease | Neurology. https://www.neurology.org/doi/full/10.1212/WNL.0000000000201205.

41. Taylor-Rowan, M. *et al.* Pre-Stroke Frailty Is Independently Associated With Post-Stroke Cognition: A Cross-Sectional Study. *Journal of the International Neuropsychological Society* **25**, 501–506 (2019).

42. Hannan, J. *et al.* Under pressure: the interplay of hypertension and white matter hyperintensities with cognition in chronic stroke aphasia. *Brain Commun* **6**, fcae200 (2024).

43. Bahsoun, M.-A. *et al.* FLAIR MRI biomarkers of the normal appearing brain matter are related to cognition. *NeuroImage: Clinical* **34**, 102955 (2022).

44. Shelton, F. de N. A. P., Volpe, B. T. & Reding, M. Motor Impairment as a Predictor of Functional Recovery and Guide to Rehabilitation Treatment After Stroke. *Neurorehabil Neural Repair* **15**, 229–237 (2001).

45. Heddings, A. A., Friel, K. M., Plautz, E. J., Barbay, S. & Nudo, R. J. Factors Contributing to Motor Impairment and Recovery after Stroke. *Neurorehabil Neural Repair* **14**, 301–310 (2000).

46. Patel, A. T., Duncan, P. W., Lai, S.-M. & Studenski, S. The relation between impairments and functional outcomes poststroke. *Archives of Physical Medicine and Rehabilitation* **81**, 1357–1363 (2000).

47. Hackett, M. L. & Pickles, K. Part I: Frequency of Depression after Stroke: An Updated Systematic Review and Meta-Analysis of Observational Studies. *International Journal of Stroke* **9**, 1017–1025 (2014).

48. Powers, W. J. *et al.* Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* **50**, (2019).

49. Van De Port, I., Kwakkel, G., Schepers, V. & Lindeman, E. PREDICTING MOBILITY OUTCOME ONE YEAR AFTER STROKE: A PROSPECTIVE COHORT STUDY. *Journal of Rehabilitation Medicine* **38**, 218–223 (2006).

50. Puig, J. *et al.* Decreased Corticospinal Tract Fractional Anisotropy Predicts Long-term Motor Outcome After Stroke. *Stroke* **44**, 2016–2018 (2013).

51. Stinear, C. M. Prediction of motor recovery after stroke: advances in biomarkers. *The Lancet Neurology* **16**, 826–836 (2017).

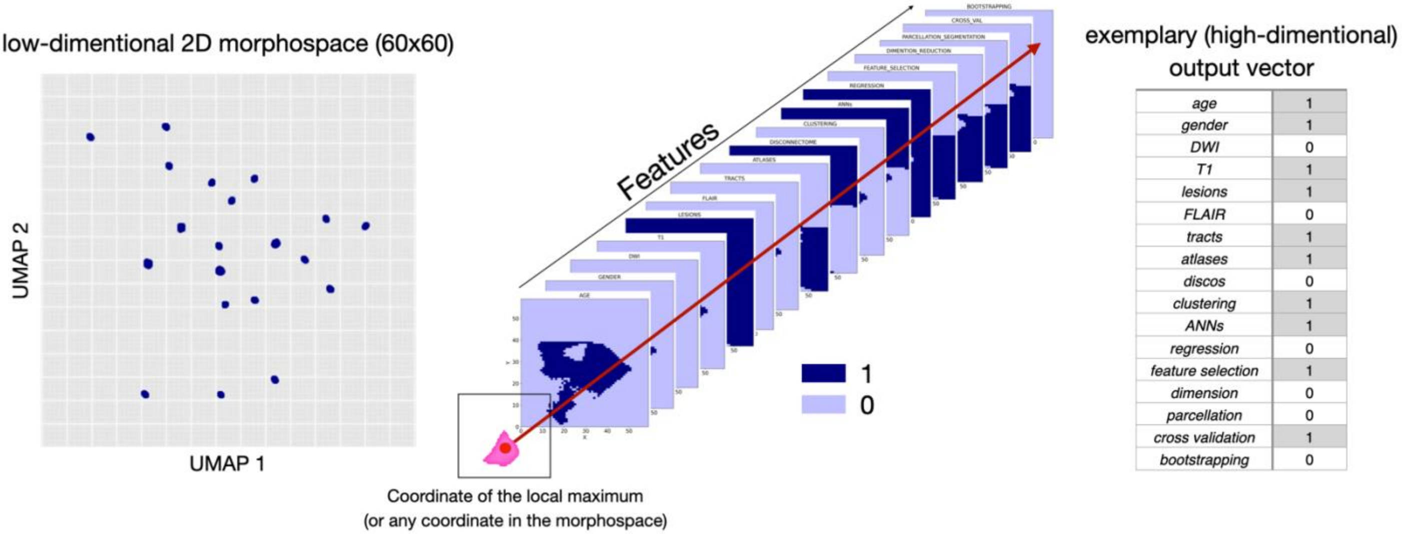
52. Espárrago Llorca, G., Castilla-Guerra, L., Fernández Moreno, M. C., Ruiz Doblado, S. & Jiménez Hernández, M. D. Post-stroke depression: an update. *Neurología (English Edition)* **30**, 23–31 (2015).

53. Towfighi, A. *et al.* Poststroke Depression: A Scientific Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* **48**, e30–e43 (2017).

54. Unsworth, D. J., Mathias, J. L. & Dorstyn, D. S. Preliminary Screening Recommendations for Patients at Risk of Depression and/or Anxiety more than 1 year Poststroke. *Journal of Stroke and Cerebrovascular Diseases* **28**, 1519–1528 (2019).

55. Pohjasvaara, T., Vataja, R., Leppävuori, A., Kaste, M. & Erkinjuntti, T. Depression is an independent predictor of poor long-term functional outcome post-stroke. *European Journal of Neurology* **8**, 315–319 (2001).

56. Bourached, A. *et al.* Scaling behaviours of deep learning and linear algorithms for the prediction of stroke severity. *Brain Communications* **6**, fcae007 (2024).



age	1
gender	1
DWI	0
T1	1
lesions	1
FLAIR	0
tracts	1
atlases	1
discos	0
clustering	1
ANNs	1
regression	0
feature selection	1
dimension	0
parcellation	0
cross validation	1
bootstrapping	0

Figure 1: Visualization of the process for obtaining the theoretically optimal feature combination for predictions: inversing the *randomise* analysis yields heatmaps for each analyzed feature. After binarizing these maps, we investigated the overlap of the local maxima of the clinical tests with the binarized feature maps (1 representing the presence of the feature, and 0 representing the absence of a feature in the final combination 'optimal recipe').

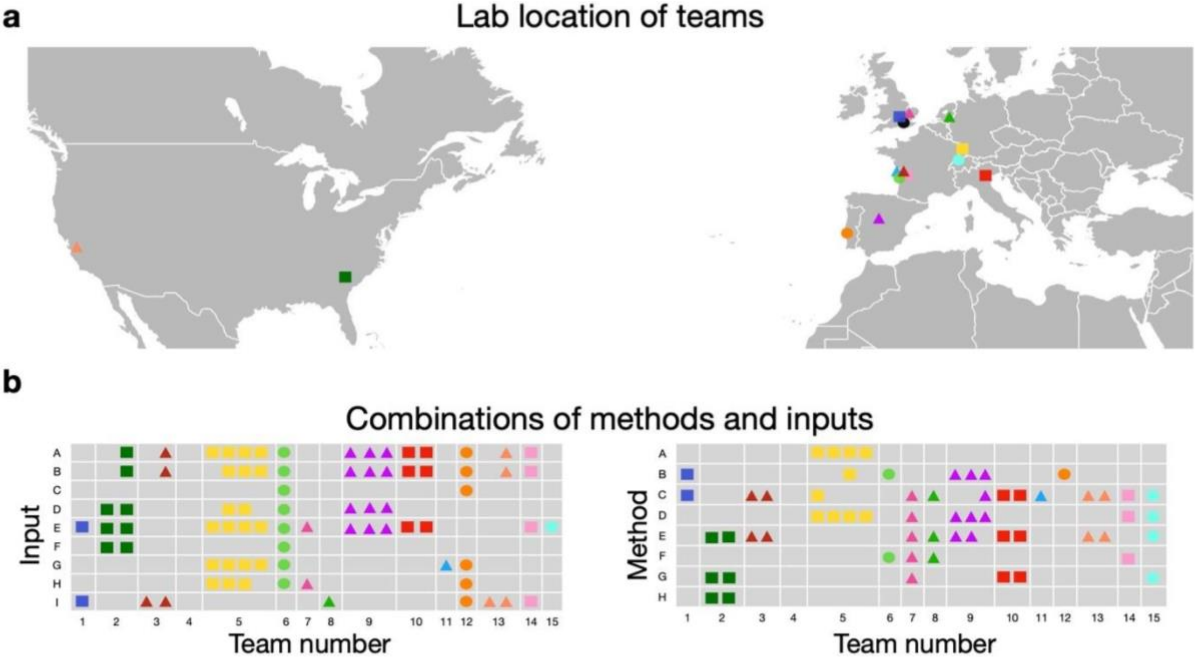


Figure 2: Summary of participating teams and the approaches taken for all predictions. **a** Locations of the teams' affiliated labs. **b** Summary of different inputs (A: age, B: gender, C: DWI, D: T1, E: lesion, F: FLAIR, G: tracts, H: atlases, I: disconnectome) and methods (A: clustering, B: artificial neural networks, C: regression, D: feature selection, E: dimensionality reduction, F: parcellation, G: cross-validation, H: bootstrap) used for each prediction. (Figure modified from[38]).

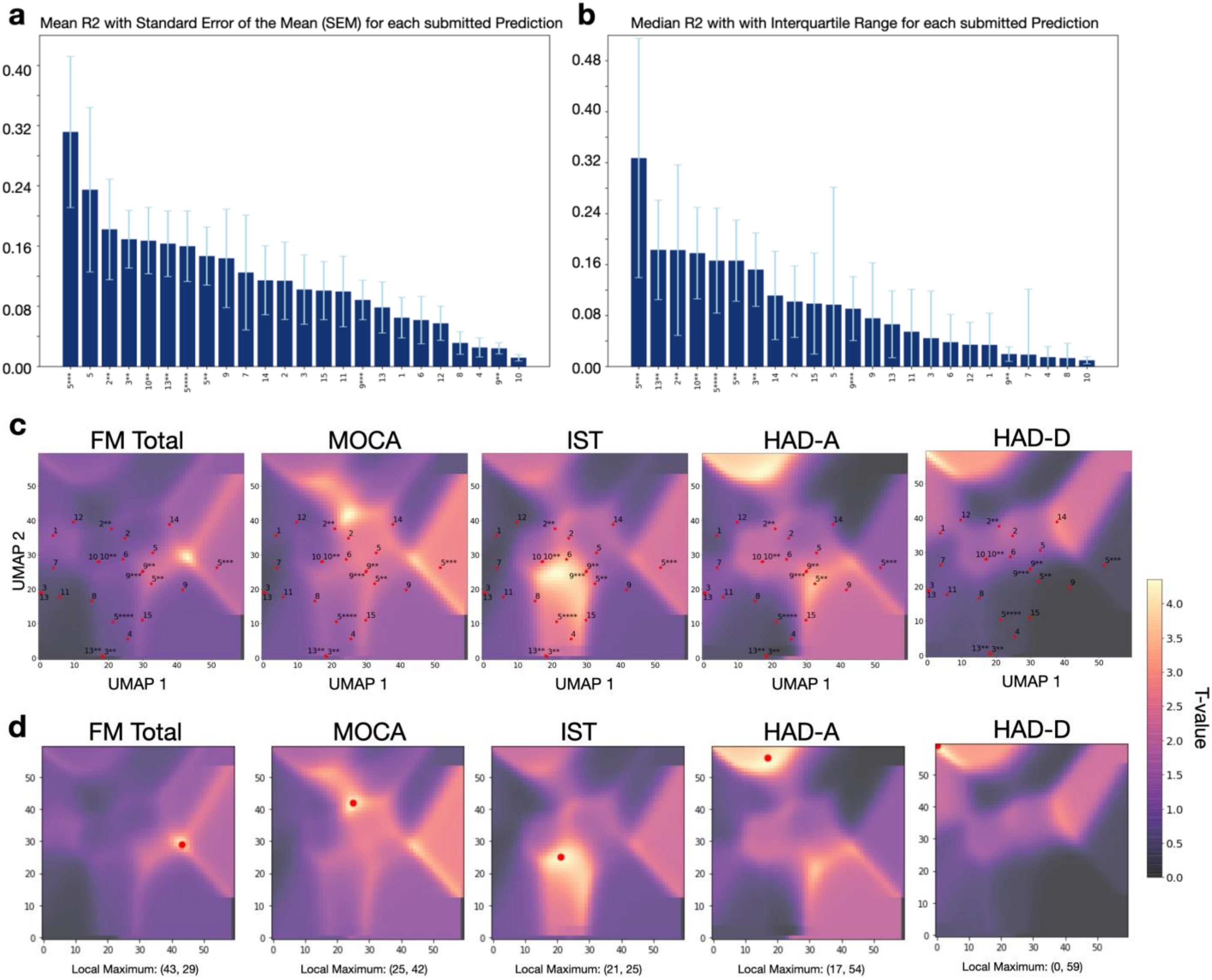


Figure 3: **a** Mean R^2 comparison for all submitted predictions (motor, cognitive, and psychological outcomes) of five neuropsychological scores (FM total, MoCA, IST, HAD-A, HAD-D) sorted ascendingly from the highest score to the lowest score across all teams whose number is indicated on the x-axis. The stars indicate the prediction number, the whiskers indicate the standard error of the mean (SEM). **b** Median R^2 comparison for all submitted predictions of the same scores, sorted ascendingly with whiskers indicating the interquartile range (IRQ). **c** T-statistic maps for each clinical outcome test, obtained from the FSL *randomise* analysis. The yellow regions indicate a significant t-value, the purple regions indicate a non-significant t-value. The UMAP distribution of all teams is plotted on the t-statistic maps. **d** Local maximum of the t-statistic map for each analyzed outcome score.