BRAIN
REVIEW ARTICLE

Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence

Anna K. Bonkhoff[1] and Christian Grefkes[2,3,4]

Stroke ranks among the leading causes for morbidity and mortality worldwide. New and continuously improving treatment options such as thrombolysis and thrombectomy have revolutionized acute stroke treatment in recent years. Following modern rhythms, the next revolution might well be the strategic use of the steadily increasing amounts of patient-related data for generating models enabling individualized outcome predictions. Milestones have already been achieved in several health care domains, as big data and artificial intelligence have entered everyday life.

The aim of this review is to synoptically illustrate and discuss how artificial intelligence approaches may help to compute single-patient predictions in stroke outcome research in the acute, subacute and chronic stage. We will present approaches considering demographic, clinical and electrophysiological data, as well as data originating from various imaging modalities and combinations thereof. We will outline their advantages, disadvantages, their potential pitfalls and the promises they hold with a special focus on a clinical audience. Throughout the review we will highlight methodological aspects of novel machine-learning approaches as they are particularly crucial to realize precision medicine. We will finally provide an outlook on how artificial intelligence approaches might contribute to enhancing favourable outcomes after stroke.

1 J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
2 Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Juelich, Juelich, Germany
3 Department of Neurology, University Hospital Cologne, Cologne, Germany
4 Medical Faculty, University of Cologne, Cologne, Germany

Correspondence to: Professor Christian Grefkes, MD
Institute of Neuroscience and Medicine - Cognitive Neuroscience (INM-3)
Forschungszentrum Juelich
52425 Juelich, Germany
E-mail: c.grefkes@fz-juelich.de

Keywords: stroke recovery; outcome; precision medicine: machine learning; big data

Abbreviations: AI = artifical intelligence; CST = corticospinal tract; MEP = motor-evoked potentials; mRS = modified Rankin Scale; NIHSS = National Institutes of Health Stroke Scale; PREP = predict recovery potential algorithm; SAFE = shoulder abduction and finger extension; SVM = support vector machine

Received March 13, 2021. Revised November 02, 2021. Accepted November 21, 2021. Advance access publication December 16, 2021

Introduction

Why precision medicine in stroke?

In spite of over 10 million yearly strokes worldwide and a global lifetime risk of 25% to suffer a stroke,[1,2] each of these strokes is a unique and very personal experience, leaving each stroke survivor with his or her very own story. Imagine being that one particular patient: you are female, 65 years old and have arterial hypertension as known comorbidity, yet are otherwise a healthy and independent person. You have noticed a weaker left-sided grip strength for 1 h and now cannot lift your left arm against gravity, your speech is slurred. Your symptom severity corresponds to a National Institutes of Health Stroke Scale (NIHSS) of 5 (maximum: 42). Initial MRI indicates ischaemia in the right internal capsule with an acute onset and no evidence for a large vessel occlusion. Would you choose a treatment and outcome prediction based on the 'average' stroke patient having a comparable NIHSS score, time constellation and imaging findings? Or would you rather prefer a more personalized version that takes into account (i) your individual constitution with respect to the potential to recovery; and/or (ii) response to a certain treatment, and has the potential to produce individualized predictions? The second, more complex choice, considering high-dimensional information, may be rendered more and more possible when merging artificial and human intelligence, as we will outline more in depth in this review.

In well-developed countries, stroke outcome has been steadily improving in recent years: These advancements have been achieved by highly effective recanalizing therapies for acute treatment, such as thrombolysis and thrombectomy,[3,4] high-quality imaging, the stratified extension of therapeutic time windows[5,6] and standardized care for dedicated stroke units. Intense rehabilitation programs[7,8] and secondary prevention, such as anticoagulants and statins,[9,10] are further examples in the subacute and chronic phases. However, most of these post-stroke treatment options require a high number of patients needed to treat to prevent an unfavourable outcome. Therefore, the optimal and most effective treatment decisions for an individual may not necessarily be derived from population averages.

These insights are not limited to stroke, but pertain to healthcare in general. They have prompted a new focus on individualizing treatments in recent years and ignited increasing numbers of precision medicine endeavours.[11,12] Ever since, more and more optimization aims focus on individuals rather than population averages to increase the efficacy in healthcare.

The role of artificial intelligence for precision medicine

Modern artificial intelligence (AI) practices offer the great opportunity to realize the vision of precision medicine.[13–15] AI can be formally defined as 'the capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this' (*Oxford English Dictionary*, see also Matheny *et al.*[16]). Of note, the term AI was introduced already about 70 years ago.[17,18] However, since then, AI has also experienced several periods of reduced interest ('AI winter') after falling short of expectations. Early AI implementations successfully completed tasks that are usually difficult for humans by applying a sequence of logical rules.[19] Examples may be seen in expert systems that imitate human decision-making processes.[20] These same implementations, however, failed to tackle tasks easy to complete for humans, such as image recognition. With the recent coincidence of growing amounts of data, exponentially increasing computational power, affordable computing and storing resources, as well as a broad software availability,[21] techniques such as machine learning and deep learning have begun to remedy these previous shortcomings. In general, both machine and deep learning have led to groundbreaking innovations, such as intelligent software to understand language[22] and images[23] or, as a very recent biological example, the prediction of protein structures based on their amino acid sequence (AlphaFold).[24] Machine and deep learning approaches, as modern branches of AI, excel in automatically detecting patterns in data and leveraging those pattern to predict future data (see Box 1 for examples of individual algorithms).[25–27] Deep learning is special in the way that it leverages artificial neural networks with multiple ('deep') levels of representations that facilitate the acquisition of particularly complex functions.[28]

Notably, AI is not a new idea in healthcare,[29] as expert-guided, rule-based medical approaches were already introduced in the 1970s, for example featuring the automated interpretation of ECGs.[18,30] Once again, machine and deep learning have recently enabled substantial improvements and demonstrated performances comparable with highly trained physicians, especially in the fields of radiology, dermatology and ophthalmology. For example, Gulshan and colleagues[31] demonstrated the feasibility of automatically detecting diabetic retinopathy in retina fundus photographs. Esteva and colleagues[32] predicted skin cancer type as accurately as dermatologists, and Hannun and colleagues[33] constructed a deep learning model that could accurately classify computerized echocardiograms into 12 rhythm classes. These successful AI implementations hold several promises in the longer term, such as predicting future disease manifestations based on routinely collected healthcare data,[34] or automated screening for certain cancer types in imaging data.[35] In the shorter term, AI-based individualized predictions on clinical outcomes could provide essential information for healthcare professionals, as well as patients, their families and friends.[16]

To foster the potential of machine and deep learning, it will be of particular importance to acquire large datasets, comprising subject-level information on hundreds to thousands of patients. Only then will these datasets have the potential to adequately represent interindividual variability in the presentation of the disease, comorbidities and predisposition,[36,37] and allow for an advantageous performance of AI models. Recent years have already seen the advent of big medical data initiatives, mostly within the framework of population studies that are not only impressive in the number of participants (number of participants >500 000), but also their data depth (number of variables >1000) (e.g. UK Biobank,[38] NIH All of Us research programme in the USA[39] and the Rhineland Study in Germany[40]). First examples of similar developments in stroke research can be observed as well: the virtual international stroke trial archive (VISTA) contains clinical data, such as the NIHSS, comorbidities or laboratory results of 82 000 patients.[41] However, 'big' imaging datasets of stroke patients are still at least an order of magnitude smaller (e.g. 2800 structural scans in the MRI-GENIE study,[42,43] 2950 scans of in Meta VCI map consortium,[44] 1800 scans in ENIGMA[45] or 1333 scans in an unicentre study[46,47]). All in all, there have been calls to accumulate and exploit regularly obtained clinical, imaging and genetic stroke patient data in a collaborative fashion.[48–50]

Article structure

In the following sections, we will specifically illustrate single-subject prediction scenarios within stroke outcome research in

Box 1 Supervised learning

The machine-learning algorithms highlighted in this review fall into the category of supervised learning algorithms. This scenario assumes that each predictor (= input) variable is linked to a response. Responses can be quantitative (i.e. taking on a numerical value, such as a patient's Fugl–Meyer score) or qualitative (i.e. categorical, such as motor symptoms versus no motor symptoms), resulting in the formulations of regression or classification problems, respectively.[83] Overall, supervised learning stands in contrast to unsupervised learning, where we have observations of measurements but no associated responses. Hence, instead of formulating regression or classification problems, the main aim of unsupervised learning approaches is to understand relationships between observations, which can, for example be achieved via clustering.[83] Further examples of unsupervised learning are dimensionality reduction techniques, such as principal component analysis (PCA) or non-negative matrix factorization. Classical examples for supervised learning algorithms for both regression and classification are linear regression models (regularized or unregularized), tree-based and nearest-neighbour algorithms, SVMs and deep neural networks:

Linear regression: In linear regression, one (simple linear regression) or more input variables (multiple linear regression) are linked to a response via a linear function. A typical application scenario for linear regression in stroke recovery research is the modelling of Fugl–Meyer follow-up scores based on initial Fugl–Meyer scores.[97] Model parameters are commonly fitted using the least square approximation or a penalized version for regularized variations (ridge regression: L^2-norm penalty; lasso: L^1-norm penalty). In case of regularization, coefficient estimates are shrunk towards or to zero, which can be particularly helpful in the case of highly variable least squares estimates that often arise when the number of explanatory variables is almost as large as the sample size, i.e. the number of observations.[83] In these situations, estimates may differ widely between different samples.

Tree-based algorithms: The simplest tree-based algorithm is a decision tree. An exemplary application in recovery research is the PREP algorithm.[128] Other tree-based algorithms are, for example, random forest[210] and gradient boosting algorithms.[122] Regression and classification are achieved by finding sequences of splitting rules that segment the space of input variables into simple regions. While being very transparent and interpretable, decision trees usually cannot compete with other algorithms with respect to prediction performance. However, modifications, such as bagging, boosting and random forests, that introduce different ways of combining multiple decision trees (ensemble learning), have been shown to enhance prediction performance substantially.[164] Interpretation is less straightforward in case of ensemble learning approaches due to their complexity (combination of trees). However, it is still possible to extract the importance of input variables for generating predictions, which facilitates their interpretability.

Nearest-neighbour algorithms: These algorithms accomplish solutions to regression or classification problems by finding *k* closest observations for any given observation and then creating average responses or majority votes.[211] Therefore, the predicted response is the average value of responses of all neighbours in regression scenarios. In classification scenarios, the predicted category is the majority class of nearest neighbours. Overall, it is appreciated that nearest-neighbour algorithms can find very complex patterns in data, which, however, comes at the cost of increased computation demand and decrease in interpretability.[165]

Support vector machines: SVMs are generalizations of so-called maximal margin classifiers.[87] SVMs are frequently used in multivariate lesion-symptom studies relying on neuroimaging data.[159] For example, a classification problem with two linearly separable classes: In this case, many straight lines can entirely separate the two classes; an SVM finds the one straight line with the widest margin. The observations closest to this separating line with the widest margin are then called support vectors. In reality, classes may not be perfectly separable, and the objective might rather be to accept misclassification in few instances to allow for a better classification performance in general, i.e. higher generalizability. While 'classic' SVMs are linear models, they can be rendered non-linear by introducing a 'kernel' that maps the input variables to an even higher dimensional space. SVMs are comparably less computationally expensive and more interpretable, but more limited in the complexity of patterns that they can fit.[165]

Deep learning algorithms: Deep learning algorithms, in particular, have gained attraction in recent years, not least due to concurrently increasing dataset sizes and available computational power. They constitute exceptionally flexible methods that combine multiple stacked layers and non-linear transformations when passing on information from one layer to the next. While each building block is comparably simple, their combination has been shown to be capable of automated feature selection and representation of complex pattern. As Goodfellow and colleagues[19] phrase it: 'Deep learning allows the computer to build complex concepts out of simpler concepts'. For example, deep learning algorithms have premiered in stroke outcome prediction scenarios based on clinical data.[113,114]

All of the outlined algorithms have their unique strengths and weaknesses. It may be particularly instrumental to compare them with respect to their transparency and complexity[212] (Fig. 6).

the acute, subacute and chronic stage. Additionally, we will highlight important considerations with respect to methodological approaches in line with the aim of this review. We first address general aspects of motor outcome research after stroke ('Motor impairment after stroke' section). Then, we will summarize the statistical foundations necessary to understand the basic principles of AI in healthcare ('Statistical background for precision medicine: inference versus prediction' section). Afterwards, we present and discuss recent studies on stroke outcome research with a special focus on those using prediction models, organized depending on the type of data, i.e. clinical data ('Stroke prognostic scales based on clinical data only' section), neurophysiological data, and combinations of clinical, neurophysiological and basic imaging data ('Neurophysiology and combination of biomarkers in individual data' section), as well as more detailed structural ('Structural imaging' section) and functional imaging data ('Functional imaging' section). Given their prime importance for the realization of precision medicine, we will outline essential methodological aspects at the beginning and end of each section. Finally, we will present a synopsis of methods as employed in concrete scenarios in motor outcome research post-stroke ('Overview of employed algorithms' section), their general advantages and promises ('General advantages and promises' section), as well as disadvantages and pitfalls ('Disadvantages and pitfalls' section). All in all, our review complements previous reviews on the use of AI in stroke, for example with a focus on clinical decision support in the acute phase,[51] acute stroke imaging,[52,53] stroke rehabilitation[54] and prognostic scales on clinical outcomes and mortality[55,56] (see the Supplementary material for our literature research strategy and selection criteria).

Motor impairment after stroke

A substantial amount of stroke patients finds themselves affected by some degree of motor impairment. Studies[7,57] report frequencies as high as 80% and 50%, respectively. The enormous burdens associated with motor impairments with regard to economic costs,[58] rehabilitation need[59] and disability-adjusted life years[60] necessitate optimizing acute and chronic stroke care. While acute stroke treatment has been considerably advanced leading to both reduced mortality and morbidity in the past decades, it may now be the restorative therapy after stroke that needs to see the same progress.[61] This focus on the subacute-to-chronic post-stroke phase may be of particular importance since only a relatively small fraction of patients presenting with acute ischaemic stroke are eligible for acute treatment options (e.g. 15.9% for thrombolysis and 5.8% for mechanical thrombectomy in Germany in 2017,[62] with comparable numbers in various other countries[63,64]).

Providing accurate outcome predictions has always been a central goal in stroke research. More specifically, predictions may point at the most suitable short and long-term treatment goals: should the focus of treatment be on true recovery or rather compensation, when significant behavioural restitution is unlikely?[65] True recovery requires neural repair to allow for an at least partial return to the pre-stroke repertoire of behaviours, e.g. the same grasping movement pattern as present prior to cerebral ischaemia. Compensation implies the substitution of pre-stroke behaviours by newly learned pattern without the necessity of neural repair, e.g. compensatory movements of the shoulder to account for extension deficits of the hand.[61] During rehabilitation, patients often show both phenomena, i.e. a partial recovery, which is complemented by compensatory behaviours. In this context, rehabilitation refers to the entire process of care after brain injury and an 'active change by which a person who has become disabled acquires the knowledge and skills needed for optimum physical, psychological and social function'.[66] The availability of predictions may help patients and their proxies to be informed about what to expect in the future and plan accordingly. Furthermore, predicting spontaneous recovery after stroke may be crucial to evaluate the effect of intervention studies. Using this information to stratify patients into control and treatment groups could decrease the overall number of patients needed to be recruited, thereby not only rendering significantly more studies feasible in terms of design and financial costs, but also yielding faster results.[67] Last, outcome models could also target the prediction of response to specific therapies, such as non-invasive brain stimulation, and thus support the identification of probable responders before the start of the therapy.[68] In the same vein, Stinear and colleagues[54] previously defined several prerequisites for rehabilitation prediction tools that may be useful in clinical practice. Accordingly, prediction tools should forecast an outcome that is meaningful for individual patients at a specific time point in the future.

Statistical background for precision medicine: inference versus prediction

Classical inference statistics, such as *F*- or *t*-tests, comprise a powerful tool kit to evaluate research hypotheses, and offer explainable results. Null hypotheses testing represents a frequently used example, which is linked to resulting *P*-values and ensuing statistical significance statements.[69,70] Importantly, these classical statistical instruments were invented almost a hundred years ago, in an era of rather limited data availability and hardly any computational power.[71] In regard to biomedical research, insights were previously commonly gleaned from either observational descriptions of single patients (e.g. Pierre-Paul Broca's patient Mr Leborgne, called 'Tan'),[72] or group comparisons. This situation, however, is changing nowadays.

The perception of statistical significance will most probably experience a redefinition in times of emerging big data scenarios. On the one hand, extensive datasets will more frequently lead to statistical significance of effects with (clinically) negligible effect sizes.[73,74] For example, Miller and colleagues conducted 14 million individual association tests between MRI-derived brain phenotypes, e.g. brain volumes or functional connectivity strength between two brain areas, and sociodemographic, neuropsychological or clinical variables in 10 000 UK Biobank participants.[75] These tests resulted in many statistically significant associations, yet these associations sometimes explained less than a percentage point of variance, which, thus, questions their relevance.[76] On the other hand, the default use and interpretation of *P*-values has been challenged frequently in recent years. This process was triggered by increasing reports on low reproducibility of research findings.[77] When trying to reproduce the findings of 100 psychological research studies, replication studies produced significant results in only 36%, while original studies reported significant results in 97% of cases.[78] In response to these findings, Benjamin and colleagues suggested a lower level of significance, i.e. $P < 0.005$, for the discovery of new effects to increase the robustness of findings.[79] Amrhein and colleagues went a step further yet and recommended to relax the over-reliance on *P*-values by completely abandoning dichotomous decisions.[80] These suggestions have prompted vital discussions: While generally being supported widely—the call by Amrhein was accompanied by >800 signatures of international researchers—other statisticians have been more cautious, for example stressing the positive effect of statistical significance as gatekeeper.[81]

It is also important to realize that statistically significant group differences, as indicated by low *P*-values, do not generally imply good single-subject level prediction performances, as measured by out-of-sample generalization (Fig. 1). The latter, however, is the idea of precision medicine.[37,82–84] In contrast to the previous focus on inference and explanation, recent years have seen an upsurge of AI and, more specifically, machine-learning techniques, that predominantly target prediction performance of single-subject outcomes. Examples of these machine-learning models include, e.g. regularized regression, (deep) neural networks, nearest-neighbour algorithms,[85] random forests[86] or kernel support vector

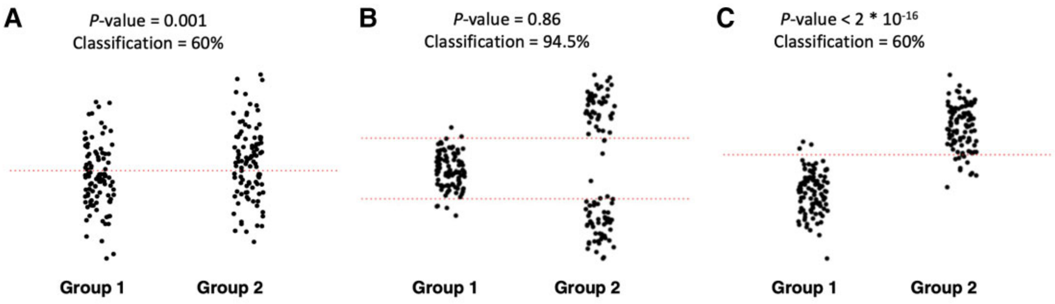

Figure 1 Three scenarios to compare group difference and classification analyses. Data is simulated, differences between groups 1 and 2 are determined via two-sample *t*-tests, classification via linear methods into groups 1 and 2 is achieved via thresholding (indicated by red dotted lines). (**A**) A significant group difference is found despite a poor classification performance. (**B**) Groups do not differ significantly, but classification accuracy is very high. (**C**) A significant group difference goes along with high classification accuracy. Overall, these three scenarios illustrate that neither significant group differences automatically lead to high classification accuracies, nor high classification accuracies to significant group differences. Adapted from Arbabshirani *et al.*,[37] with permission.

machines (SVMs) (Box 1).[87] Given multiple input variables, such as age, sex, initial stroke severity and comorbidities, these models are trained to predict some specific individual outcome, such as a motor score 3 months after stroke, based on a weighted combination of these input variables, with the highest achievable prediction performance. This performance can be quantified by various established measures, such as explained variance, accuracy, sensitivity, specificity and area under the curve. As they are evaluated by their generalization capability to previously unseen, i.e. new data samples, they are well suited to ensure accurate predictions of individual future outcomes. At the same time, these models may not typically and reliably be able to explain their predictions any further and allow for inferences on particular biological mechanisms. This characteristic has prompted the denotation black-box model.[88]

Stroke outcome studies

Stroke prognostic scales based on clinical data only

The initial level of impairment induced by the stroke lesion is a well-known explanatory variable of the neurological outcome several months later.[89–91] A number of studies aimed to explain recovery patterns through linking motor impairments at initial and follow-up time points by means of linear regression models (63–211 patients).[92–95] In these scenarios, the change between initial and follow-up motor impairment (i.e. a continuous change = follow-up − initial scores) represented the output, i.e. the dependent variable Y. This output was computed based on the recovery potential, i.e. the maximum score minus the initial motor impairment as input, i.e. the independent variable X. Motor impairment was most frequently captured as Fugl–Meyer score of the upper limb.[96] The typically obtained performance measure here was the explained variance in form of the in-sample R^2-value. This R^2-value, also called the coefficient of determination, indicates how much of the variance in the dependent variable can be explained by one or more independent variables.

These modelling endeavours resulted in the proportional recovery rule.[97] Stroke patients with mild to moderate motor impairments usually regain a certain amount of their lost motor function within the first months after their stroke. In one of the largest stroke recovery studies, considering 211 patients, initial motor impairment apparently explained up to 94% of the variance in motor recovery based on the proportional recovery rule.[93] However, recent re-evaluations of the statistics underlying the proportional recovery rule suggest that previous estimates of explained variance were inflated. This inflation occurred due to statistical confounds, such as measurement noise, ceiling effects and a phenomenon called mathematical coupling.[98–101] Mathematical coupling here describes a situation where the input and output variables are not independent—which is the case when recovery is defined as the difference between an initial score and the follow-up score, and this change score is then correlated to the very same initial score that was used to compute the change score. Thus, the assumption of no relationship between input and output is void. Simulations have shown that significant relationships between initial and change score can occur, when, in fact, there is no significant link between initial and follow-up scores.[99,100] We recently introduced a Bayesian hierarchical modelling regimen to combine patient data from six recovery studies (n = 385) and demonstrated that reducing analyses to the subset of only severely to moderately affected patients could successfully mitigate the effects of ceiling and mathematical coupling.[102] Notably, after addressing confounds, the initial impairment was shown to explain only a small amount of the variance in recovery, reaching a maximum of 32% explained variance only.[102] Therefore, proportional recovery may occur, however, to a considerably smaller degree than originally claimed.

Importantly, these recovery studies highlight the distinction between inference and prediction (see the ‘Statistical background for precision medicine: inference versus prediction’ section). In the studies mentioned before, the relationship between initial impairment and recovery was primarily investigated in-sample. In-sample here means that the performance of linear regression models was estimated relying on exactly the same data that was used for model training. Therefore, models had already seen all parts of the data that they were subsequently tested on and could optimally adapt to them. This strategy is particularly helpful, if the main study aim is to identify significant explanatory variables of the outcome and to obtain interpretable models.[82,84] If several studies then independently point to the same association, this association might be considered more stable and reliable, since it was validated. The estimates of prediction performance have, however, not been validated by these means. When training and test data do not differ, algorithms are also prone to overfitting, i.e. they might capture the characteristics of the data sample at hand very well by explaining a high percentage of the variance observed for this particular sample, but at the same time perform relatively poorly when tested on an independent dataset. Conversely, generalization capability—conceptually underlying precision medicine—is commonly tested by measuring a model's prediction performance for unseen, novel data-points, i.e. those that have not been used in the training phase. The developed model is therefore validated out-of-sample.[83,103] Therefore, it is generally crucial to know whether prediction performance estimates were obtained in-sample or out-of-sample (Supplementary Tables 1–3).

Given that the studies highlighted in the previous section mainly used a within-sample approach, they could well infer the significance of input variables (i.e. the initial motor score) and interpret their coefficients (i.e. 70%) at the group-level. In contrast, the aim of the studies presented in the following section is the accurate training of prognostic models that can predict a categorical functional outcome at the level of an individual patient[104] (for recent reviews see Fahey *et al.*[55] and Drozdowska *et al.*[56] and 'Statistical background for precision medicine: inference versus prediction' and 'General advantages and promises' sections for details on the distinction between inference on group-level versus prediction on individual-patient-level). Most of these prognostic prediction endeavours feature similar methodological steps. First, the outcome is not represented by a change score, as above, but by a binary, categorical (0–1) follow-up outcome, for example, favourable versus unfavourable functional outcome [e.g. modified Rankin Scale (mRS) of ≤2 versus >2, no-mild versus moderate-severe disability]. Most often, the model of choice is a logistic regression model[105] that considers sociodemographic and clinical information as input variables. Training and testing, or in other words developing and validating, is commonly performed in separate datasets. Importantly, it is this separate training and testing approach that enables conclusions on the generalization performance of a model to unseen data of individual patients. Prediction performance itself is frequently quantified as area under the receiver operating characteristic (AUROC), or in short area under the curve (AUC), which considers the true positive rate (i.e. sensitivity) as well as the false positive rate (i.e. 1—specificity) across various thresholds. While a value of 0.5 represents the level of chance, an AUC of 1 signals the best possible performance.[83]

Using data from 10 777 patients included in the clinical trials archive VISTA as an additional validation (test) dataset, Quinn and colleagues[106] compared the predictive capacities of eight well recognized prognostic models to predict favourable outcome post-stroke (90-day mRS ≤ 2).[107–112] The model abbreviated to ASTRAL (Acute Stroke Registry and Analysis of Lausanne)[107] provided the highest prediction fidelity of all models and achieved an AUROC of 0.78. As each model had originally been trained in a different dataset, relying on anywhere between 1645 and 12 262 patients from different countries and continents, each model included a marginally varying number and collection of input variables. However, most considered age and stroke severity, as well as pre-stroke comorbidities as input (Table 1). More recently, two studies explored the capability of deep learning algorithms to enhance the prediction of functional outcomes based on clinical information. Heo and colleagues[113] compared the performances of deep neural networks, random forest classification and logistic regression to the established ASTRAL score to predict favourable outcomes (90-day mRS ≤ 2) in 2604 patients. Their deep learning model based on 38 clinical variables, such as demographics, stroke severity and stroke subtype, was the only one to significantly outperform the ASTRAL score. Li and colleagues,[114] on the other hand, used deep neural networks, an SVM, random forest classification, a gradient boosting algorithm and logistic regression to predict unfavourable outcomes (mRS > 2) 6 months post-stroke in 1735 patients using information on clinical, demographic and laboratory characteristics. Neither of their prediction models performed clearly better. Of note, the studies by both Heo and colleagues and Li and colleagues used a test set, thus their estimates can be regarded as out-of-sample.

Table 1 Integer-based prognostic ASTRAL score for the calculation of probability of unfavourable outcome in patients with acute ischaemic stroke (1645 patients in total)

Covariates	Score
Age: for every 5 years	1
Severity: for every NIHSS point	1
Time delay from onset to admission <3 h	2
Range of visual field defect	2
Acute glucose >7.3 or <3.7 mmol/l	1
Level of consciousness decreased	3

Higher scores indicate less favourable outcomes.[108]

Further studies evaluated modified scenarios as they focused on stroke patients admitted to rehabilitation institutes and specifically strived for modelling outcomes after rehabilitation.[115–117] Brown and colleagues asserted that the motor subscore of the Functional Independence Measure (FIM),[118] age and walking distance at admission explained most variance in the FIM-based recovery (i.e. change), length of stay and discharge destination (148 367 patients).[115] Since the authors derived their results only in-sample, the generalization performance to out-of-sample, i.e. new, patients remains to be elucidated. Scrutinio and colleagues[116] developed a prediction model of the motor subscore of the FIM after rehabilitation and considered multiple available variables as predictors during model training. They eventually chose five of them based on forward stepwise logistic regression: age, time from stroke occurrence to rehabilitation admission and unilateral neglect were predictive of higher motor impairment at discharge, while lower admission motor and cognitive impairment predicted lower motor impairment at follow-up. After model development, they then tested for their algorithm's capacity to generalize to new patients and obtained a validation sample prediction performance of AUC = 0.866.

In general, objectives of these prediction model endeavours were to provide additional information to augment a doctor's judgement on the risk of favourable or unfavourable outcome and assist in (fast) clinical decision making. Most of these studies translated the original logistic model to an integer-based score or offered online calculator for a more intuitive and faster outcome calculation (Table 1 and e.g. https://goo.gl/fEAp81 for Scrutinio *et al.*'s prediction models). Indeed, some of these automated predictions were shown to outperform the intuitions of medical doctors in several datasets.[119–121] However, any one of these scores has yet to be implemented into clinical routine and several challenges remain to be addressed (see the 'Disadvantage and pitfalls' section).

Neurophysiology and combination of biomarkers in individual data

The studies in the following section make use of predictors that are closer to the neurobiology of the brain, i.e. data obtained by neuroimaging or neurophysiological recordings. Such surrogate-based predictions might yield higher prediction accuracies than those based on clinical or behavioural information as the former may better capture interindividual differences of lesion-induced disturbances in neuronal function as well as the mechanisms driving functional recovery. While some of the authors instrumentalized stepwise logistic regression to identify critical parameters for recovery or future motor performance, others demonstrate broader model type considerations that go beyond linear methods, for example, decision-tree-based algorithms (Box 1). Among others, these model types have the ability to exploit non-linear relationships and interactions of input variables automatically.[122] Conversely, when working in a linear regression framework as before, interactions have to be inserted manually and thus intentionally, probably based on previous knowledge that the researcher has. Altogether, more flexible models like tree-based algorithms as well as further non-parametric models, such as nearest-neighbour algorithms and those applying kernels, may be more capable to 'let data speak for themselves'.[123] They can—at best—uncover complex, predictive patterns in data automatically. However, these models require a lot of data to successfully do so. In view of their flexibility, these models are otherwise at the risk of overfitting and poor generalization to new data due to too close adaptation to the data at hand in case of data scarcity.

As first examples of broader biomarker consideration: Koh and colleagues[124] used stepwise linear regression to evaluate 19 variables comprising information on clinical and also imaging parameters to build a prediction model of motor recovery, i.e. the change between admission and follow-up upper limb movement capacity, in a sample of 140 severely affected stroke patients. Four variables were ultimately identified to hold the most explanatory information: 'baseline upper extremity score' (positive association with impairment) and 'baseline NIHSS score' (negative), as well as the imaging-related variables 'haemorrhagic stroke' and 'cortical lesion excluding primary motor cortex' (both positive). Model performance, however, was capped at 35% of total variance explained in-sample. This low value signalled a generally limited explanatory capacity of the considered initial input variables. Nonetheless, more complex and less predictable recovery patterns after severe stroke are a frequently described finding,[125] which renders the results of Koh and colleagues[124] less surprising. In another study comprising data of 160 acute stroke patients, mRS-based functional outcome 3 months post-stroke was significantly associated with the clinical variables 'left-sided lesions', 'stroke severity at admission' (both negative association with favourable outcome) and the 'presence of motor-evoked potentials (MEPs) on TMS of the ipsilesional motor cortex' (positive association).[126]

Going yet a step further, several studies employed the combination of clinical, imaging and neurophysiological markers to optimize outcome predictions. The Predict Recovery Potential (PREP) algorithm runs through a sequence involving all of these measures to stratify patients into four recovery groups based on their follow-up Action Research Arm Test (ARAT) score.[127],[128] It first divides patients into two groups based on a commonly conducted clinical test of upper limb function 72 h after stroke (SAFE = sum of the shoulder abduction and finger extension grades based on the Medical Research Council muscle scale). Subsequently, it considers information on transcranial magnetic stimulation (TMS)-obtained MEPs in upper limb muscles. MEP-positivity is here thought to indicate functional integrity of corticospinal pathways. Last, the PREP algorithm incorporates an MRI parameter that represents the structural integrity of the cortico-spinal tract fibres in the posterior limb of the internal capsule (fractional anisotropy asymmetry index). The PREP algorithm was designed in a decision-tree-like fashion relying on expert knowledge, i.e. the sequence and nature of tests was manually chosen and not automatically computed from data to reflect a setting that can be readily implemented into the clinical routine.[127] In general, decision-tree-based algorithms create classifications by finding sequences of splitting rules that segment the space of input variables into simple regions and, as such, are very transparent and interpretable (Box 1). The PREP decision-tree-algorithm was subsequently validated in a dataset of 40 stroke patients.[128] Especially the outcome 'full recovery' could be predicted with high positive predictive power (88%), negative predictive power (83%), specificity (88%) and sensitivity (73%). Only slightly lower prediction accuracies of correct outcome classifications (80%) were found when testing the PREP on an independent dataset of 157 patients, underlining its reliable generalization performance.[129] Furthermore, the PREP algorithm was successively refined to the PREP2 algorithm by means of a more, yet still only partly, automated classification and regression tree (CART) approach in 207 patients, thus in-sample (Fig. 2).[130] In contrast to the original PREP algorithm, the authors defined a lower SAFE score cut-off at the first decision point, i.e. it was now reduced to 5 instead of 8 points, with 5 points indicating a higher motor impairment. As a result, this change required to assess MEPs only in low SAFE score patients without any loss in prediction performance (sensitivity of 75% in comparison to 73% before). Furthermore, the PREP2 algorithm did not rely on MRI data anymore, thereby facilitating its clinical implementation. A study comparing the lengths of rehabilitation stays suggested a real-world relevance of the PREP predictions: patients in an intervention group as well their therapists were disclosed their PREP outcome predictions at the beginning of the rehabilitation stay.[129] Patients in this group could then be discharged a week sooner than the patients in the control group lacking information on the additional PREP estimate. Of note, this finding was controlled for upper limb impairment, age, sex and comorbidities (implementation group: $n = 110$, 11 days, control group: $n = 82$, 17 days). Furthermore, there were no adverse effects on later functional outcomes. The authors explained the shorter rehabilitation period by an increase in the therapists' confidence and modification of therapy content in view of the outcome prediction. Importantly, classification accuracy has been shown to be decreased in case of measuring initial performance 2 weeks post-stroke instead of within the first 72 h (91 patients).[131]

Structural imaging

Structural neuroimaging is well implemented in routine diagnostic pathways for treating stroke patients. Therefore, using this information for outcome prediction seems particularly feasible in a clinical setting, and indeed, several studies have already provided clear links between motor outcome and structural markers, such as imaging parameters reflecting pre-stroke brain health or lesion location for example with respect to fibre tracts.[132–139]

Several studies pursued a hypothesis-driven approach focusing on particular anatomical structures such as the corticospinal tract (CST)—the most important motor output tract of the brain. For example, the amount of damage to the CST, as estimated by the spatial overlap between stroke lesion and tract volume, is strongly associated with the level of motor impairment in the chronic phase post-stroke (50 participants, in-sample $R^2 = 0.71$).[133] Furthermore, Feng and colleagues[135] presented evidence that CST lesion and the initial motor score performed on par when explaining the final motor outcome 3 months after stroke. To demonstrate the generalizability of findings, analyses were conducted in two separate

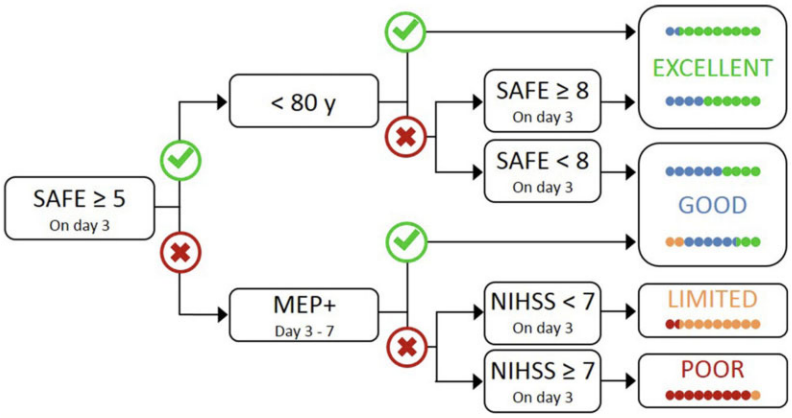

Figure 2 Prediction of ARAT score-based upper limb recovery potential via the PREP2 algorithm. The PREP2 algorithm combines several assessments in a decision-tree-like fashion considering the SAFE score, age, NIHSS and MEPs. The first decision step is based on the SAFE score, which captures the ability of shoulder abduction and finger extension, using the Medical Research Council grades (0: no palpable muscle activity, to 5: normal power) within the first 3 days after stroke onset. In the case of a SAFE score of 5 or above, the next decision is based on the patient's age. If younger than 80 years, outcome is predicted to be excellent. If older than 80 years, it is once again the SAFE score that differentiates between outcomes: The algorithm predicts excellent outcome in case of a score of 8 or higher and good outcome, if lower than 8. If, however, the patient achieves a SAFE score below 5, the next decision step considers the presence or absence of MEPs on transcranial magnetic stimulation (TMS) of the ipsilesional motor cortex on Days 3–7 after symptom onset. If MEPs are present, the patient is assigned to the second-best outcome group, i.e. a good outcome. Absent MEPs, in contrast, prompt the consideration of the NIHSS on Day 3: a score below seven leads to the prediction of limited outcomes, while an NIHSS score of 7and above results in the prediction of the lowest, i.e. poor outcome. Adapted from Stinear and colleagues,[130] with permission.

datasets, considering 37 patients for a training cohort and further 39 for a validation cohort. In the validation cohort, CST-lesion load and the initial motor score explained 69% and 62% (out-of-sample) variance of the final Fugl–Meyer score. Interestingly, CST-lesion load was also significantly associated with realized recovery, albeit to a limited degree, explaining ~20% of the in-sample variance (48 patients, with realized recovery = [follow-up − initial motor score] / [maximum score − initial score]).[140] Likewise, studies based on diffusion tensor imaging (DTI) suggest significant associations between measures of CST integrity and long-term functional outcomes as reflected by the mRS 3–12 months after stroke.[141–144]

In contrast to the low-dimensional data underlying CST-lesion-based predictions of stroke outcomes, multivariate approaches have the capacity to capture high-dimensional whole-brain lesion patterns. Thereby, they can consider specific spatial distributions in high granularity, e.g. a particular combination of lesioned voxels. For example, Forkert and colleagues[145] leveraged a multivariate SVM to predict favourable versus unfavourable functional follow-up outcome in 68 patients (favourable 30 days mRS ≤ 2). The authors found that a favourable outcome could be predicted with a cross-validated accuracy of 85% when considering detailed information on lesion location as derived from MRI-FLAIR images. Two recent studies demonstrated the feasibility of using convolutional neural networks (CNNs) for a similar prediction task, i.e. the prediction of favourable outcomes (90-day mRS ≤ 2). More specifically, Bacchi and colleagues[146] trained CNNs in combination with deep neural networks on information originating from non-contrast CT and clinical data (e.g. age, sex, stroke severity and comorbidities) to generate predictions for 204 patients, and achieved a test set accuracy of 74%. Nishi and colleagues[147] relied on diffusion-weighted MRI (DWI) of 324 patients to predict favourable outcomes, and demonstrated superior test set performance of their deep learning model compared to simpler baseline models (deep learning model: AUC = 0.81 versus ASPECTS: AUC = 0.63). Several further studies considered imaging data from a database of a 132 first-time ischaemic and haemorrhagic stroke patients.[148,149] Two weeks after stroke, DWI-derived lesion volume itself only explained a small amount of variance (cross-validated R^2 < 20%) in any of the evaluated functional domains (motor, language, attention, memory, vision). However, explained variances increased when information on lesion location was added (between 25% and 54% for motor deficits, language and attention/visual field biases). Only in case of verbal and spatial memory explained variance still totalled <20%, which probably reflects their less localized representation in the brain compared to the other functional domains.[148] These analyses relied on a pipeline comprising ridge regression and leave-one-out cross-validation. To reduce the high-dimensionality of voxel-wise lesion location information, lesion maps were also embedded in lower dimensional space via principal component analyses (PCA) before regression analyses.

The aim of a subsequent study relying on the same dataset was to predict the domain-specific 3-month outcomes (in contrast to 2-week outcomes)[150]: lesion size, age, educational attainment, hours of therapy and domain-specific scores obtained in the subacute post-stroke phase could explain in-sample variances between 42% for attention at the lower and 70% for language impairments at the higher end in a linear regression framework. When PCA-transformed lesion location information was added to the models, prediction performance significantly increased for models explaining language, motor and attention impairments (4.0–13.0% increase in explained variance). However, explained variance remained unchanged in case of the verbal and spatial memory domains, suggesting once again that there is no one-size-fits-all solution, and some deficits may not be straightforwardly explained by lesion location alone. Furthermore, it is important to note that the inclusion of hours of therapy as input variable in a prediction algorithm of stroke outcomes may be problematic, given that this value is not necessarily known in advance and hence cannot easily be entered into a prediction algorithm at the beginning of rehabilitation.[54] Last, yet another study indicated that more sophisticated neuroimaging parameters may outperform simple lesion location ones. Accordingly, DTI-derived axial diffusion maps were shown to yield higher prediction accuracies of 3-month functional outcomes compared to simple lesion segmentation maps in a sample of 87 patients (median cross-validated accuracy: 82.8 and 76.7%, respectively).[151]

Taken together, the studies reviewed above demonstrate that subacute and chronic post-stroke impairments in several functional domains can be better explained, if information on lesion location is included. Nonetheless, the variance that could be explained varied widely for different outcomes—for example from 4% to 54% in the work by Corbetta and colleagues.[148] Thus, these results raise the question whether sample sizes larger than the ones presented here may facilitate deriving more informative low-dimensional lesion representations. Independent of sample size, it may be also necessary to increase the spatial resolution as even 1-mm isotropic voxel scans may still not capture the interindividual variability that is seen in microscopical analyses of histological brain sections, especially with respect to fibre tract anatomy.[152]

Functional imaging

In addition to structural scanning, functional MRI has become a valuable method to infer post-stroke alterations of neuronal activity and also enable individual predictions.[153–156] This technique allows to draw conclusions on neural activity non-invasively on the basis of changes of blood flow and oxygen content.[157] Functional imaging can come in two forms: task-based and resting-state functional MRI. While participants are asked to perform a specific task in the first scenario, they are required to lie motionless but awake in the scanner in the second scenario. Analyses are then either centred on activity changes in certain brain areas or functional connectivity strengths between brain areas, respectively.[158]

Sample sizes are usually considerably smaller in functional imaging studies than in other stroke outcome prediction scenarios, due to methodological challenges (longer acquisition times, low signal-to-noise ratio, signal susceptibility to head movement, MRI contraindications) and substantially higher costs. Interestingly, as functional MRI datasets can be considered high-dimensional data containing thousands of voxels, AI approaches have been frequently used to detect certain patterns of activity or connectivity that allow prediction of the functional outcomes of a single patient. Importantly, here we focus only on those functional neuroimaging studies that used rigorous cross-validation schemes to generate single-subject predictions.

Two studies have made use of functional MRI data acquired in the first days after stroke to make predictions on clinical motor impairment at the time of scanning and follow-up motor impairment 4 to 6 months post-stroke (40 and 21 stroke patients).[159,160] These studies were conducted in prediction-focused frameworks similar to the structural stroke studies described previously, e.g. by applying SVMs combined with nested leave-one-out-cross-validation. In a first study, resting-state functional MRI data were used to calculate whole-brain connectivity to a 'seed region', i.e. reference region, in the ipsilesional, yet structurally intact primary motor cortex. Subsequently, this connectivity information was instrumentalized to discriminate between stroke patients with and without acute hand motor deficits as well as healthy controls.[159] Prediction models were successively refined to tell apart stroke patients with favourable versus unfavourable motor outcome several months after stroke.[160] Notably, prediction here relied on task-based functional MRI, instead of resting-state functional MRI data. Motor deficits were measured as Motricity Index of the hand[161] in the first and grip force and ARAT score in the second study.[162] Both studies reported cross-validated prediction accuracies of >80% (82.6% motor-stroke versus non-stroke, 87.6% motor-stroke versus non-motor-stroke in the first study[159] and 86% favourable versus unfavourable motor outcome in the second study[160]). In case of the discrimination of motor-stroke versus non-stroke patients, classification performance particularly relied on interhemispheric primary motor cortex M1—M1 as well as ipsilesional M1—premotor areas connectivity profiles (Fig. 3). As the resting-state data investigated in the first study was collected during routine scanning sessions, the authors underline the clinical practicability of their approach, particularly for acute and severely affected patients.[159] Another milestone study, once again relying on the 132 stroke patients introduced in the previous section on structural scans,[148,150] compared predictive capacities of dimensionality-reduced structural lesion topography and functional connectivity via ridge regression (Fig. 4): functional connectivity allowed for more accurate cross-validated predictions in neurocognitive domains (functional connectivity: visual and verbal memory: $R^2 = 0.36$ and $R^2 = 0.42$, respectively). Nonetheless, lesion topography outperformed functional connectivity in case of predictions in sensorimotor domains (structural lesion information: vision and motor impairments: $R^2 = 0.50$ and $R^2 = 0.45$, respectively).[149] Both imaging and behavioural data were obtained on average 2 weeks post-stroke. Altogether, these rather moderate levels of explained variance also suggest that a substantial fraction of variability in outcome may originate from factors that are not yet captured and considered in current studies. The studies reviewed in this section made use of SVMs as well as ridge regression to compute predictions on behavioural outcome after stroke. These two approaches are influenced by so-called hyperparameters determining the amount of model regularization. In the case of ridge linear regression, a regularized version of linear regression, as e.g. applied in the study by Siegel and colleagues,[149] the hyperparameter lambda determines the amount of shrinkage of the regression coefficients.[25] Likewise, the parameter C defines the amount of regularization of the SVM applied in Rehme and colleagues.[159,160] However, the optimization of these hyperparameters requires some extra care to avoid overfitting. One way to achieve a safe optimization can, for example, be a nested cross-validation framework, i.e. the combination of inner and outer cross-validation loops (Fig. 5). When the computational burden is high, as in case of deep learning approaches, nested cross-validation might not be feasible and, alternatively, the entire dataset can be split in three parts: training, test and validation sets. The optimal hyperparameters can then be obtained by relying on training and test sets, while less biased performance estimates can be attained in the validation set. However, such an approach may require relatively large datasets.

General considerations

Overview of employed algorithms

Having illustrated the various data fields of motor-focused stroke outcome studies, it becomes apparent that each field may have its unique repertoire of preferably used (prediction) algorithms (for a theoretical overview on algorithms, see Box 1). 'Classic' motor recovery studies that consider the sole recovery potential, e.g. defined as maximum minus initial Fugl–Meyer score as input variable, primarily rely on relatively simple, unregularized linear regression to model quantitative recovery scores, i.e. the change

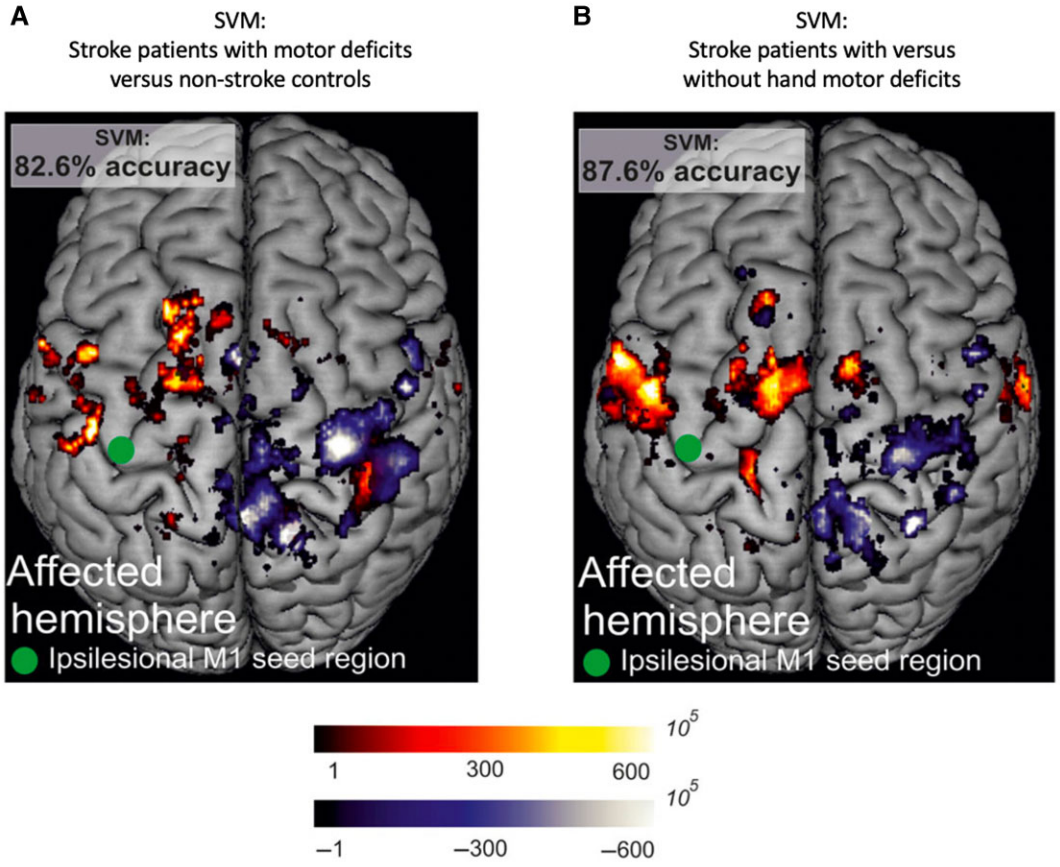
A
SVM:
Stroke patients with motor deficits
versus non-stroke controls
SVM:
82.6% accuracy
Affected
hemisphere
Ipsilesional M1 seed region
B
SVM:
Stroke patients with versus
without hand motor deficits
SVM:
87.6% accuracy
Affected
hemisphere
Ipsilesional M1 seed region
10^5
1
300
600
10^5
−1
−300
−600

Figure 3 SVM-based prediction of motor deficits after stroke. Whole-brain functional connectivity to an ipsilesional M1 seed region was computed in a voxel-wise fashion for 20 stroke patients with motor impairments, 20 stroke patients without motor impairments and 20 non-stroke controls. (**A**) Stroke patients with motor impairment could be differentiated from non-stroke controls with an accuracy of 82.6%. (**B**) Similarly, the classification of stroke patients into those with and without motor impairments resulted in an accuracy value of 87.6%. Regions coloured in blue support the prediction of non-stroke controls or stroke patients without motor impairment; their functional connectivity is enhanced in comparison to stroke patients with motor impairments. Regions coloured in red, on the other hand, indicate a higher functional connectivity in patients with motor impairment and contributed to their classification. Adapted from Rehme and colleagues,[159] with permission.

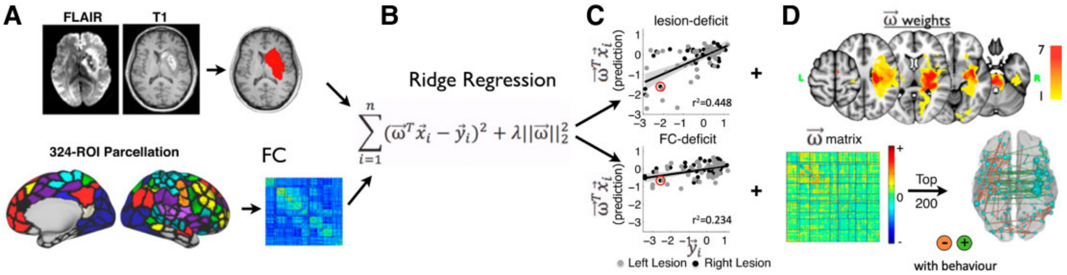

Figure 4 Overview of the analytical pipeline to predict behavioural impairments in 100 stroke patients based on structural and functional MRI. (**A**) Manual lesion segmentation in case of structural lesion information and atlas-defined region-of-interest (ROI)-based estimation of functional connectivity in case of functional data. (**B**) Structural lesion information or functional connectivity data is entered into ridge regression models to predict behavioural outcomes in a leave-one-out cross-validation. (**C**) Comparison of predicted and true behavioural scores to determine model performance. (**D**) Visualization of model weights as estimated via ridge regression. Adapted from Siegel and colleagues (Copyright 2016, National Academy of Sciences, USA).[150]

between follow-up and initial Fugl–Meyer scores. Particularly as fitted in-sample, these models have proven to be particularly easy to interpret. On the other hand, there seems to be a preference for logistic regression models within the field of prognostic studies of raw (and not change score) follow-up outcomes (binary categories: favourable versus unfavourable functional outcomes). These logistic regression analyses are often combined with stepwise procedures to select final input variables and construct a model as parsimonious as possible. While the stepwise feature selection step can indeed lead to memorable sets of predictors and allow the construction of simple, yet clinically practical point scores, there are some drawbacks to stepwise feature selection procedures; for

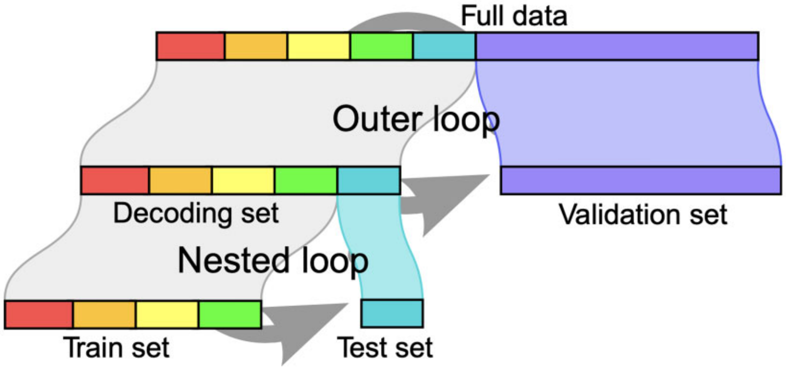

Figure 5 Schematic illustration of nested cross-validation. Two loops of cross-validation are performed, with hyperparameter optimization being performed in the inner, or nested, loop. Adapted from Varoquaux and colleagues,[103] with permission.

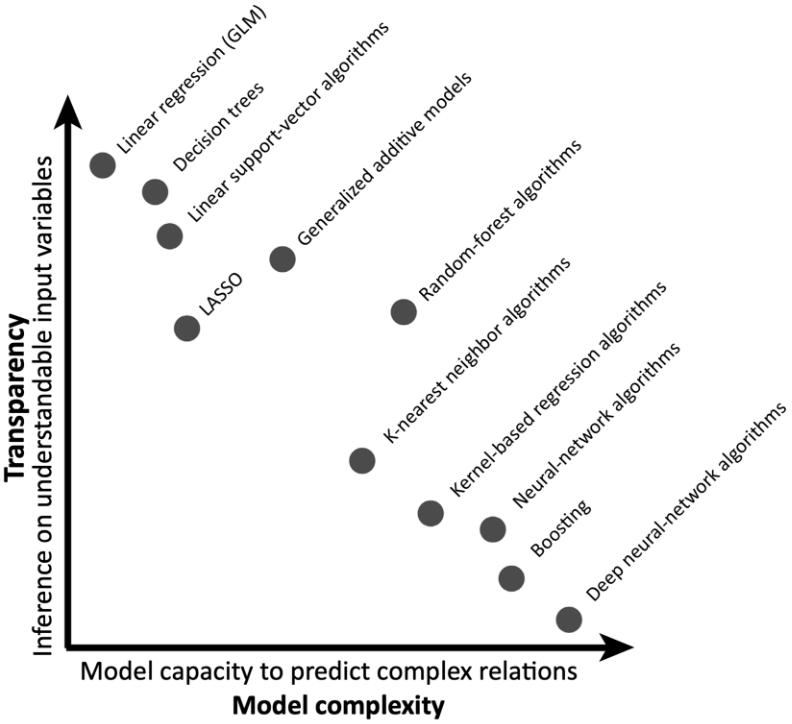

Figure 6 Comparison of various learning algorithms with respect to their model transparency and complexity. Model transparency here refers to the interpretability of input variables and thus the potential scientific insight and mechanistic understanding that can be gained. More complex models, in return, maximize the predictive power. Altogether, increased transparency may come at the cost of decreased model complexity and associated decreased predictive power and vice versa. Figure adapted from Bzdok and Ioannidis,[212] with permission.

example, neither forward nor backward feature selection are guaranteed to result in the overall best model as models are constructed and tested iteratively and not all conceivable models are considered.[83] Moreover, the performance of stepwise selection models might be overestimated, i.e. too optimistic.[163]

Decision-tree-based algorithms may be a natural choice when combining information from different sources, such as behavioural, neurophysiological and neuroimaging ones, as in case of the PREP or PREP2 algorithm.[128] Decision trees perform regression and classification tasks by finding sequences of splitting rules that segment the space of input variables into simple regions. They may excel in being transparent, easily interpretable and applicable. It has to be noted, however, that more advanced tree-based algorithms, such as random forest or gradient boosting algorithms that combine multiple individual decision trees and are thus more difficult to interpret, usually outperform simple decision-tree algorithms.[164]

Last, SVMs and regularized linear regression (e.g. ridge regression) have been frequent choices to evaluate structural or functional neuroimaging data, given that they have proved to be capable of handling high-dimensional data particularly well. While ridge regression is mostly still combined with some initial (PCA-based) unsupervised dimensionality reduction preprocessing step, SVMs have been shown to generate good predictions despite the combination of moderate sample sizes and thousands of voxels per patient.[165] More generally, SVMs can employ the 'kernel trick', i.e. map input information to high-dimensional feature spaces and by these means produce non-linear predictions, that may automatically capture complex relationships between the input and output. In contrast, ridge regression, as regularized version of linear regression, is a linear prediction model. Interestingly, deep learning has also made its entrance into several stroke outcome predictions scenarios. This update may have been incentivized by deep learning approaches' promising success in further, often machine vision-focused medical scenarios. For example, deep learning has been shown to excel when detecting skin cancer,[32] inferring genetic mutations in cancerous tissue from routine histopathology tissue slides[166] or evaluating mammography scans.[167,168] As deep learning models typically show most favourable performances when trained on particularly large samples, often $>10^5$–10^6, future studies are warranted to investigate the usefulness of deep learning approaches for stroke outcome predictions more broadly. Particularly as data sample sizes may not grow quickly enough and may not reach the standards in other non-stroke fields, successful deep learning applications might be limited to specific tasks, such as image registration.[169,170] Last, it is important to consider that the 'No Free Lunch' theorem[171] guarantees that all algorithms perform similarly on average when all possible problems are taken into account. Thus, while each field currently appears to employ a unique methodological toolset, an enhanced methodological exchange between researchers of the various displayed fields, that may motivate the application of several learning algorithms at once, may be generally beneficial.

General advantages and promises

AI approaches in stroke research have already facilitated promising developments in outcome predictions, as well as additional insights in the (neurobiological) factors and mechanisms associated with poor versus good outcomes. Importantly, we have highlighted the delicate difference between in-sample inference and out-of-sample prediction-oriented studies. The former—inference—capitalizes on the interpretability of findings, at best describing an underlying mechanism. In-sample inference, for example, focuses on estimating the importance of individual input variables in explaining the outcome of interest across an entire group (and not individual patients). In contrast, out-of-sample computations are central to prediction studies that put an emphasis on the best generalization performance possible.[83,84,172] This approach targets optimal predictions for an individual patient not only with respect to outcome, but also concerning the response to a certain treatment.

AI-based prediction approaches hold several advantages over 'classical' tools used in the field of post-stroke recovery. Most studies in stroke outcome research still apply some variant of linear or logistic regression model. Although such models are often easier to interpret, they cannot automatically exploit non-linear relationships and interactions, which can lead to poorer prediction performance. These limitations can be overcome with machine-learning algorithms, such as decision-tree-based algorithms, SVMs and neural networks. Although these techniques are computationally more demanding and the interpretation of the model parameters more complex, they might augment the prediction performance by exactly the amount that is necessary to turn an interesting prediction model into a diagnostic tool. The PREP algorithm[128,129] represents a promising example. This decision-tree-based algorithm has been shown to generate accurate predictions that are clinically beneficial: information on outcome prediction shortened rehabilitation stays without any reduction in functional outcome.[130] Nonetheless, recent non-stroke prediction-focused studies suggest that these more complex relationships, i.e. non-linearities and interactions, may not be generally present or readily exploitable in clinical datasets with for example small to moderate sample sizes ($n < 100$).[173,174] Thus, the authors of these studies caution against unrealistic expectations that the application of machine-learning algorithms instead of simple linear models will automatically enhance prediction performance.

In addition to using linear models, most of the studies highlighted in this review still relied on specifically curated datasets and considered a circumscribed list of input variables only. However, the combination of out-of-sample testing and machine-learning algorithms may allow for the consideration of a broader range of input variables—as long as data sample sizes increase in parallel. For example, it would be conceivable to jointly consider multimodal, structural and functional imaging data[175] or metabolic, demographic and mechanistic variables[176] to enhance prediction performance. Overall, it seems likely that it will be such a combination of multiple data sources, or essentially neurobiologically based biomarkers, that will facilitate the most accurate stroke outcome prediction performance at a personalized level. Future studies may hence not only explore a richer methodological toolset (see the 'Overview of employed algorithms' section) but could also plan to systematically and explicitly investigate the combination of a variety of biomarkers.

What is more, machine-learning-based prediction performance may be boosted even further when making use of unsystematically collected, but considerably bigger samples.[84] 'Unsystematically' here refers to the fact that collected variables might not have been hand-picked, but acquired without any previous hypothesis and selective inclusion and exclusion criteria. Examples for these kinds of data could be registry data, electronic health records or clinical stroke scans that have been recorded independent of specifically planned research projects. The use of general, unstructured clinical data may furthermore enable a better representation of the full spectrum of stroke patients: These prediction scenarios may also include subgroups that are often neglected in stroke outcome studies, such as very young, very old, very severely affected or multimorbid stroke patients with recurrent strokes or other interfering neurological conditions.[36,50,177,178]

A further, desirable next step to enhance current prediction scenarios is the consideration of outcome measures that go beyond coarse-grained classifications, such as favourable versus unfavourable functional outcome based on binarized or ordinal scores like the mRS. Several studies already provided evidence that the focus on detailed scales, such as the ARAT or Fugl–Meyer assessment for motor impairments of the upper limb, is feasible and instrumental.[159,160] These more detailed motor assessments could be amended by scores evaluating impairments in further functional domains, such as the cognitive or language domains, and then integrated into multi-outcome prediction algorithms.[179] Such a multi-outcome approach might represent a more holistic and hence realistic approach, as impairments are rarely limited to just one functional domain[148] and may even interact with one another during recovery (e.g. motor recovery and cognitive dysfunction).[180] In conjunction with the selection of outcome scales, it will be important to reflect on the representation of the outcome: Do we want to predict the change between follow-up and initial scores or the final, follow-up score directly? Directly predicting the final score, while of course taking into account the initial baseline score, may be more desirable for several reasons. First, it circumvents any confounds induced by mathematical coupling that arises when a change score is predicted by an initial score (see the 'Stroke prognostic scales based on clinical data only' section).

In particular, a linear regression model of raw outcome scores could be transformed into a change score model, which would then additionally allow for the interpretation of coefficients with respect to the classic proportional recovery concept.[102,181] Second, interpreting recovery solely on the basis of the change between follow-up and initial scores may mix up different patient subgroups and neurobiological mechanisms underlying different forms of functional recovery. For example, it is likely that a recovery change score of 10 points on the Fugl–Meyer assessment scale is driven by very different neurobiological processes depending on whether recovery started with an initial score of 5 (very severely affected) or 55 (almost no deficits). In turn, a patient that has recovered 20 points on the Fugl–Meyer scale, but started with an initial score of 5, is still considerably less recovered than a patient recovering 10 points but starting from 55. Therefore, follow-up scores rather than change scores seem to be better suited for recovery prediction scenarios. Complementing the increase in granularity of targeted outcomes, performance evaluation metrics could also be intelligently varied: the AUROC is the currently predominantly used score for binary prediction tasks. This one-dimensional approach could be extended to a multidimensional one by considering numerous, complimentary metrics, such as positive predictive values, sensitivity and specificity at once.[182]

Altogether, improving predictions by all these means might eventually resolve the disenchantment stemming from reports of low real-world impact as few of the prediction models are actually used in clinical routine[91,106,183] and render them more clinically useful.

Disadvantages and pitfalls

Interpretability has always been of particular importance for researchers, independent from their specific field of research.[184] However, as mentioned before, some modern learning algorithms capture and instrumentalize patterns in high-dimensional data[185] with sometimes even millions of parameters that may simply be too complex to be readily comprehensible. These characteristics have led to the denotation black box and triggered some scepticism with which these modern statistical tools are regarded.[186] Yet, these black-box characteristics may be acceptable when they produce the best prediction results including high generalization performance, as increasing interpretability—as found in simpler, e.g. linear, models—often comes at the cost of decreasing prediction performance.[88] Essentially, there is currently no consensus on what level of interpretability is required for safe deployment of prediction models.[15] However, independent of the level of model interpretability, it seems necessary that human intelligence acts together with AI.[187] In particular, due to their capacity to extract information from otherwise intractable high-dimensional data, AI or, more specifically, machine-learning approaches could represent a very effective initial step. Medical professionals, such as physicians and therapists, could then include this information into their treatment decisions to achieve an optimal outcome for their patients. Physicians, for example, tend to be too optimistic and vastly overestimate life expectancy of terminally ill patients.[188] In contrast, deep learning-based predictions were shown to generate more accurate life expectancy predictions and hence might have yielded better therapeutic decisions.[189] As outlined above (the 'Stroke prognostic scales based on clinical data only' section), some stroke outcome models have also already been shown to outperform the predictions made by physicians and/or therapists.[119–121] At the same time, for a safe implementation of machine-learning routines, physicians and therapists need to check on a regular basis whether the models established for a certain diagnostic or therapeutic scenario are still valid.[187] Incongruencies may, for example, arise in case of a 'dataset shift', i.e. when there is a mismatch between the data used during model development and the data currently used for model deployment.[190,191] As a prominent, recent example, a major US hospital had to deactivate model-based sepsis-predictions to prevent spurious alerts after patients' characteristics had substantially changed with the onset of the coronavirus disease 2019 pandemic.[191] The same sepsis-alert model has furthermore been shown to perform only poorly in independent, real-world data,[192] motivating a constant ongoing surveillance and validation of already established prediction models.

The current curriculum in medical school might thus be revised to equip physicians with the necessary toolset, e.g. in the field of health informatics.[35] Close collaborations between various disciplines might be strengthened to successfully combine statistical, computational and human perspectives.[193,194] These efforts may then increase physicians' abilities to recognize both the benefits and limitations of AI in healthcare[195] and enhance the knowledge on how to, for example, continuously quantify and validate prediction performance of used prediction tools. Recently presented checklists and guidelines for the transparent reporting of AI algorithms and interventions in medicine may represent an essential foundation.[29,196,197] In general, reliability, privacy and fairness are further important ethical aspects that need to be reconsidered and redefined in greater depth in an interdisciplinary fashion when using more machine-learning algorithms in upcoming years.[15]

Last, it will be important to warrant satisfactory data quality. Otherwise, we may be at risk of encountering the big data paradox, as Xiao-Li Meng outlines it: 'The more data, the more surely we fool ourselves'.[198] An algorithm can hardly be any better than the data that it learns from. Data acquired in the clinical routine might be noisy and biased, since, for instance, the patient moved during MRI scanning, it took too long until the blood was analysed or a junior doctor systematically misunderstood how to rate certain symptoms of the NIHSS score. Missing data, particularly those missing not at random, represent further challenges.[182] An important, somewhat trivial, but often neglected aspect is the validity of the data with respect to what can be really inferred from them. For example, DTI-based neuroimaging gives the impression of assessing anatomical fibre tracts, but they remain model-based approximations with a coarse spatial resolution when considering the nearly 1000-fold smaller diameter of axons.[152] Likewise, functional MRI is based on a haemodynamic signal, which is much slower and anatomically blurrier than true neuronal activity.[157] These issues are further complicated by strong interindividual variability, which is encountered at basically all levels of the CNS. For example, analyses of post-mortem brains have revealed that even the location of primary areas like M1 or primary visual cortex, which represent highly conserved brain regions within and across species, may vary in a centimetre range between subjects independent of anatomical landmarks.[199] Decomposing anatomical variability is technically feasible to a certain degree, but quickly meets its limits when it comes to spatial and temporal resolution issues of neurons and axons.

To proceed with any of these aspects mentioned, medical doctors, neuroscientists, statisticians, computer scientists and ethicists ideally need to work in an interdisciplinary fashion.[17,193,194] They will need to ensure that inherent biases in data are detected, react accordingly, ignite discussions and develop international standards for big data analytics.[191] In this way, it might be possible to realize data science at its best and develop clinically helpful models—as, after all: ‘All models are wrong, but some are useful’. (George E. P. Box).

Conclusion and open questions

Prediction approaches based on AI have the great potential to revolutionize medical care in general. However, it remains to be seen whether expectations can be sustainably met. It is furthermore essential to recall that machine-learning based prediction scenarios should not be mistaken as causal inference.[200] In this context, randomized clinical studies are an exemplary study type that permits conclusions on whether a specific treatment causally underlies a better outcome in the treatment group. However, we may not generally be able to decide on whether a (standard) treatment is effective or not and whether it should e.g. be stopped based on machine-learning-derived outcome predictions. To infer causal effects, we would rather have to estimate what was most likely to happen, as well as the counterfactual prediction, i.e. what would have happened, if things had been different.[201] As Wilkinson and colleagues[200] put it: We may not be able to learn this counterfactual prediction by relying on the combination of machine learning and observational data—as they do not contain any information on what would have happened given altered circumstances.

Furthermore, if it is not only the physician who is undertaking the clinical decision-taking process, but a prediction algorithm, who is responsible in case of (fatal) error? At present, it is still the physician who has to take the final responsibility and to verify that the result of a prediction algorithm complies with the current medical standards. How to deal with the situation when an effective prediction algorithm is available for a doctor but is not used, especially when the doctor’s decision was wrong and harmed the patient? How can we ensure that we comply with patients’ privacy rights and protect health data from potential cyber-attacks?[202] How can we guarantee that our prediction models are fair, i.e. that prediction performance does not vary depending on ethnicity or gender? This aspect might be of particular concern since machine-learning approaches may sometimes even enhance biases present in historical datasets that, for example, include skewed representations of people of colour, women and underserved populations.[203] Last, how do these changes affect the doctor–patient relationship and how can we unlock the potential of AI assisted healthcare to eventually enhance our physician time veridically spend on ‘caring for the patient’?[187,204]

Finally, advocating for more AI-based studies certainly does not negate the value of small data studies using inference statistics, which—especially if founded on strong theory, robust measurement and effective error variance control—can reveal systematic, functional relationships on the individual subject level[205] and may thus help to take a more mechanistic perspective on the development of therapeutic approaches for stroke recovery.[206] We have also not considered any Bayesian approaches here that hold great promise of capturing essential characteristics of stroke recovery.[207–209] In the very end, conclusions originating from different methodological approaches may be merged to maximize patients’ well-being and we may particularly embrace novel prediction techniques to augment our human performance as medical doctors.

Acknowledgements

We are grateful to our colleagues at the Department of Neurology, University Hospital Cologne & Medical Faculty, University of Cologne; UCL Queen Square Institute of Neurology, University College London; and the J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School for valuable support and discussions.

Funding

A.K.B. is supported by a Massachusetts General Hospital (MGH) Executive Committee on Research (ECOR) Fund for Medical Discovery (FMD) Clinical Research Fellowship Award. C.G. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 431549029, SFB 1451 (projects B05, C05, Z03), and Project-ID 491111487.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Feigin VL, Nguyen G, Cercy K, *et al.* Global, regional, and country-specific lifetime risks of stroke, 1990–2016. *N Engl J Med*. 2018;379(25):2429–2437.
2. Feigin VL, Krishnamurthi RV, Parmar P, *et al.* Update on the global burden of ischemic and hemorrhagic stroke in 1990–2013: The GBD 2013 study. *Neuroepidemiology*. 2015;45(3):161–176.
3. Hacke W, Kaste M, Bluhmki E, *et al.* Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N Engl J Med*. 2008;359(13):1317–1329.
4. Berkhemer OA, Fransen PS, Beumer D, *et al.* A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med*. 2015;372(1):11–20.
5. Thomalla G, Simonsen CZ, Boutitie F, *et al.* MRI-guided thrombolysis for stroke with unknown time of onset. *N Engl J Med*. 2018;379(7):611–622.
6. Nogueira RG, Jadhav AP, Haussen DC, *et al.* Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med*. 2018;378(1):11–21.
7. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: A systematic review. *Lancet Neurol*. 2009;8(8):741–754.
8. Stinear CM, Lang CE, Zeiler S, Byblow WD. Advances and challenges in stroke rehabilitation. *Lancet Neurol*. 2020;19:348–360.
9. Johnston SC, Easton JD, Farrant M, *et al.* Clopidogrel and aspirin in acute ischemic stroke and high-risk TIA. *N Engl J Med*. 2018;379(3):215–225.
10. Amarenco P, Kim JS, Labreuche J, *et al.* A comparison of two LDL cholesterol targets after ischaemic stroke. *N Engl J Med*. 2020;382:9.
11. National Research Council. *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press; 2011.
12. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793–795.
13. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309(13):1351–1352.
14. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317.

15. Wiens J, Saria S, Sendak M, *et al*. Do no harm: A roadmap for responsible machine learning for health care. *Nat Med*. 2019; 25:1337–1340.
16. Matheny M, Israni ST, Ahmed M, Whicher D. *Artificial intelligence in health care: The hope, the hype, the promise, the peril*. NAM Special Publication. National Academy of Medicine; 2019:154.
17. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag*. 2006;27(4):12.
18. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719–731.
19. Goodfellow I, Bengio Y, Courville A. *Deep learning*. MIT Press; 2016.
20. Liao S-H. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Syst Appl*. 2005;28(1):93–103.
21. Manyika J, Chui M, Brown B, *et al*. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute; 2011.
22. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Mag*. 2012;29(6):82–97.
23. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2012; 60(6):84–90.
24. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.
25. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*, vol. 1. Springer Series in Statistics. Springer; 2001.
26. Murphy KP. *Machine learning: A probabilistic perspective*. MIT Press; 2012.
27. Deo RC. Machine learning in medicine. *Circulation*. 2015;132-(20):1920–1930.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521-(7553):436–444.
29. Norgeot B, Quer G, Beaulieu-Jones BK, *et al*. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320–1324.
30. Kundu M, Nasipuri M, Basu DK. Knowledge-based ECG interpretation: A critical review. *Pattern Recognit*. 2000;33(3):351–373.
31. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
32. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115.
33. Hannun AY, Rajpurkar P, Haghpanahi M, *et al*. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65.
34. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–1358.
35. Jha S, Topol EJ. Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA*. 2016;316-(22):2353–2354.
36. Adolphs R. Human lesion studies in the 21st century. *Neuron*. 2016;90(6):1151–1153.
37. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*. 2017;145:137–165.
38. Collins R. What makes UK Biobank special? *Lancet*. 2012;379-(9822):1173–1174.
39. Hudson K, Lifton R, Patrick-Lake B. *The precision medicine initiative cohort program—Building a research foundation for 21st century medicine*. Precision Medicine Initiative (PMI) working group report to the advisory committee to the director. National Institutes of Health; 2015.
40. Breteler MM, Stöcker T, Pracht E, Brenner D, Stirnberg R. MRI in the Rhineland study: A novel protocol for population neuroimaging. *Alzheimers Dement*. 2014;10(4):P92.
41. Ali M, Bath PMW, Curram J, *et al*. The virtual international stroke trials archive. *Stroke*. 2007;38(6):1905–1910.
42. Schirmer MD, Dalca AV, Sridharan R, *et al*. White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts–The MRI-GENIE study. *NeuroImage Clin*. 2019;23:101884.
43. Bretzner M, Bonkhoff AK, Schirmer MD, *et al*. MRI radiomic signature of white matter hyperintensities is associated with clinical phenotypes. *Front Neurosci*. 2021;15:691244.
44. Weaver NA, Kuijf HJ, Aben HP, *et al*. Strategic infarct locations for post-stroke cognitive impairment: A pooled analysis of individual patient data from 12 acute ischaemic stroke cohorts. *Lancet Neurol*. 2021;20:448–459.
45. Liew S-L, Zavaliangos-Petropulu A, Jahanshad N, *et al*. The ENIGMA stroke recovery working group: Big data neuroimaging to study brain-behavior relationships after stroke. *Hum Brain Mapp*. 2022;43(1):129–148. doi: 10.1002/hbm.25015
46. Mah Y-H, Husain M, Rees G, Nachev P. Human brain lesion-deficit inference remapped. *Brain*. 2014;137(9):2522–2531.
47. Bonkhoff AK, Xu T, Nelson A, *et al*. Reclassifying stroke lesion anatomy. *Cortex*. 2021;145:1–12.
48. Rostanski SK, Marshall RS. Precision medicine for ischemic stroke. *JAMA Neurol*. 2016;73(7):773–774.
49. Liebeskind DS, Malhotra K, Hinman JD. Imaging as the nidus of precision cerebrovascular health: A million brains initiative. *JAMA Neurol*. 2017;74(3):257.
50. Grefkes C, Fink GR. Noninvasive brain stimulation after stroke: It is time for large randomized controlled trials! *Curr Opin Neurol*. 2016;29(6):714–720.
51. Bivard A, Churilov L, Parsons M. Artificial intelligence for decision support in acute stroke — current roles and potential. *Nat Rev Neurol*. 2020;16(10):575–585.
52. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. *Stroke*. 2020;51:2573–2579.
53. Murray NM, Unberath M, Hager GD, Hui FK. Artificial intelligence to diagnose ischaemic stroke and identify large vessel occlusions: A systematic review. *J Neurointerv Surg*. 2020;12: 156–164.
54. Stinear CM, Smith M-C, Byblow WD. Prediction tools for stroke rehabilitation. *Stroke*. 2019;50(11):3314–3322.
55. Fahey M, Crayton E, Wolfe C, Douiri A. Clinical prediction models for mortality and functional outcome following ischemic stroke: A systematic review and meta-analysis. *PLoS ONE*. 2018;13(1):e0185402.
56. Drozdowska BA, Singh S, Quinn TJ. Thinking about the future: A review of prognostic scales used in acute stroke. *Front Neurol*. 2019;10:274.
57. Kelly-Hayes M, Beiser A, Kase CS, Scaramucci A, D'Agostino RB, Wolf PA. The influence of gender and age on disability following ischemic stroke: The Framingham study. *J Stroke Cerebrovasc Dis*. 2003;12(3):119–126.
58. Luengo-Fernandez R, Violato M, Candio P, Leal J. Economic burden of stroke across Europe: A population-based cost analysis. *Eur Stroke J*. 2020;5(1):17–25.
59. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396:10267.

60. Feigin VL, Abajobir AA, Abate KH, *et al*. Global, regional, and national burden of neurological disorders during 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet Neurol*. 2017;16(11):877–897.
61. Bernhardt J, Hayward KS, Kwakkel G, *et al*. Agreed definitions and a shared vision for new standards in stroke recovery research: The Stroke Recovery and Rehabilitation Roundtable taskforce. *Intl J Stroke*. 2017;12(5):444–450.
62. Weber R, Krogias C, Eyding J, *et al*. Age and sex differences in ischemic stroke treatment in a nationwide analysis of 1.11 million hospitalized cases. *Stroke*. 2019;50(12):3494–3502.
63. Kuhrij LS, Wouters MW, van den Berg-Vos RM, de Leeuw F-E, Nederkoorn PJ. The Dutch acute stroke audit: Benchmarking acute stroke care in the Netherlands. *Eur Stroke J*. 2018;3(4): 361–368.
64. Gattringer T, Ferrari J, Knoflach M, *et al*. Sex-related differences of acute stroke unit care: Results from the Austrian stroke unit registry. *Stroke*. 2014;45(6):1632–1638.
65. Karnath H-O, Sperber C, Rorden C. Mapping human brain lesions and their functional consequences. *NeuroImage*. 2018; 165:180–189.
66. British Society of Rehabilitation Medicine. *Rehabilitation following acquired brain injury: National clinical guidelines*. Physicians RCo; 2003.
67. Ward NS. Restoring brain function after stroke—bridging the gap between animals and humans. *Nat Rev Neurol*. 2017;13(4):244.
68. Ovadia-Caro S, Khalil AA, Sehm B, Villringer A, Nikulin V, Nazarova M. Predicting the response to non-invasive brain stimulation in stroke. *Front Neurol*. 2019;10:302.
69. Fisher RA, Mackenzie WA. Studies in crop variation. II. The manurial response of different potato varieties. *J Agric Sci*. 1923;13(3):311–320.
70. Neyman J, Pearson ES. IX. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans R Soc Lond Ser A, Contain Papers Math Phys Character*. 1933;231(694–706):289–337.
71. Bzdok D, Nichols TE, Smith SM. Towards algorithmic analytics for large-scale datasets. *Nat Mach Intell*. 2019;1(7):296–306.
72. Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA. Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*. 2007;130(5):1432–1441.
73. Wasserstein RL, Lazar NA. The ASA's statement on p-values: Context, process, and purpose. *Am Stat*. 2016;70(2):129–133.
74. Ioannidis JPA. Meta-research: Why research on research matters. *PLoS Biol*. 2018;16(3):e2005468.
75. Miller KL, Alfaro-Almagro F, Bangerter NK, *et al*. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523–1536.
76. Smith SM, Nichols TE. Statistical challenges in 'big data' human neuroimaging. *Neuron*. 2018;97(2):263–268.
77. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
78. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716.
79. Benjamin DJ, Berger JO, Johannesson M, *et al*. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6.
80. Amrhein V, Greenland S, McShane B. *Scientists rise up against statistical significance*. Nature Publishing Group; 2019.
81. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: Do not abandon significance. *JAMA*. 2019;321:2067–2068.
82. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25(3):289–310.
83. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. Vol. 112. Springer; 2013.
84. Bzdok D, Engemann D, Thirion B. Inference and prediction diverge in biomedicine. *Patterns*. 2020;1:100119.
85. Fix E. *Discriminatory analysis: Nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine; 1951.
86. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1. IEEE; 1995:278–282.
87. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297.
88. Holm EA. In defense of the black box. *Science*. 2019;364(6435): 26–27.
89. Newman M. The process of recovery: After hemiplegia. *Stroke*. 1972;3(6):702–710.
90. Kwakkel G, Kollen B, Lindeman E. Understanding the pattern of functional recovery after stroke: Facts and theories. *Restor Neurol Neurosci*. 2004;22(3–5):281–299.
91. Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW. Early prediction of outcome of activities of daily living after stroke: A systematic review. *Stroke*. 2011;42(5):1482–1488.
92. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity. *Ann Neurol*. 2015;78(6):848–859.
93. Winters C, van Wegen EEH, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. *Neurorehabil Neural Repair*. 2015;29(7):614–622.
94. Feng W, Wang J, Chhatbar PY, *et al*. Corticospinal tract lesion load: An imaging biomarker for stroke motor outcomes: CST lesion load predicts stroke motor outcomes. *Ann Neurol*. 2015;78(6):860–870.
95. Guggisberg AG, Nicolo P, Cohen LG, Schnider A, Buch ER. Longitudinal structural and functional differences between proportional and poor motor recovery after stroke. *Neurorehabil Neural Repair*. 2017;31(12):1029–1041.
96. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance. *Scand J Rehabil Med*. 1975;7(1): 13–31.
97. Prabhakaran S, Zarahn E, Riley C, *et al*. Inter-individual variability in the capacity for motor recovery after ischemic stroke. *Neurorehabil Neural Repair*. 2008;22(1):64–71.
98. Krakauer J, Marshall R. The proportional recovery rule for stroke revisited: The proportional recovery rule for stroke revisited. *Ann Neurol*. 2015;78(6):845–847.
99. Hope TM, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. *Recovery after stroke: Not so proportional after all?* Oxford University Press; 2018.
100. Hawe RL, Scott SH, Dukelow SP. Taking proportional out of stroke recovery. *Stroke*. 2019;50(1):204–211.
101. Bowman H, Bonkhoff A, Hope T, Grefkes C, Price C. Inflated estimates of proportional recovery from stroke: The dangers of mathematical coupling and compression to ceiling. *Stroke*. 2021;52:1915–1920.
102. Bonkhoff AK, Hope T, Bzdok D, *et al*. Bringing proportional recovery into proportion: Bayesian modelling of post-stroke motor impairment. *Brain*. 2020;143:2189–2206.
103. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*. 2017; 145:166–179.
104. Quinn TJ, Drozdowska BA. Stroke prediction and the future of prognosis research. *Nat Rev Neurol*. 2019;15(6):311.
105. Hosmer DW, Lemeshow S. *Applied logistic regression*, 2nd edn. Wiley; 2000.

106. Quinn TJ, Singh S, Lees KR, Bath PM, Myint PK. Validating and comparing stroke prognosis scales. *Neurology*. 2017;89(10):997–1002.
107. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score. *Neurology*. 2012;78:1916–1922.
108. Flint AC, Faigeles BS, Cullen SP, *et al*. THRIVE score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in VISTA. *Stroke*. 2013;44(12):3365–3369.
109. O'Donnell MJ. The PLAN score: A bedside prediction rule for death and severe disability following acute ischemic stroke. *Arch Intern Med*. 2012;172(20):1548.
110. Kwok CS, Potter JF, Dalton G, *et al*. The SOAR stroke score predicts inpatient and 7-day mortality in acute stroke. *Stroke*. 2013;44(7):2010–2012.
111. Saposnik G, Kapral MK, Liu Y, *et al*. IScore: A risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation*. 2011;123(7):739–749.
112. Sung S-F, Chen Y-W, Hung L-C, Lin H-J. Revised iScore to predict outcomes after acute ischemic stroke. *J Stroke Cerebrovasc Dis*. 2014;23(6):1634–1639.
113. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning–based model for prediction of outcomes in acute stroke. *Stroke*. 2019;50(5):1263–1265.
114. Li X, Pan X, Jiang C, *et al*. Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning. *Front Neurol*. 2020;11:1464.
115. Brown AW, Therneau TM, Schultz BA, Niewczyk PM, Granger CV. Measure of functional independence dominates discharge outcome prediction after inpatient rehabilitation for stroke. *Stroke*. 2015;46(4):1038–1044.
116. Scrutinio D, Lanzillo B, Guida P, *et al*. Development and validation of a predictive model for functional outcome after stroke rehabilitation: The Maugeri model. *Stroke*. 2017;48(12):3308–3315.
117. Saito J, Koyama T, Domen K. Long-term outcomes of FIM motor items predicted from acute stage NIHSS of patients with middle cerebral artery infarct. *Ann Rehabil Med*. 2018;42(5):670–681.
118. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the functional independence measure. *Arch Phys Med Rehabil*. 1994;75(2):127–132.
119. Saposnik G, Cote R, Mamdani M, *et al*. JURaSSiC: Accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology*. 2013;81(5):448–455.
120. Ntaios G, Gioulekas F, Papavasileiou V, Strbian D, Michel P. ASTRAL, DRAGON and SEDAN scores predict stroke outcome more accurately than physicians. *Eur J Neurol*. 2016;23(11):1651–1657.
121. Reid JM, Dai D, Delmonte S, Counsell C, Phillips SJ, MacLeod MJ. Simple prediction scores predict good and devastating outcomes after stroke more accurately than physicians. *Age Ageing*. 2017;46(3):421–426.
122. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29:1189–1232.
123. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521(7553):452.
124. Koh C-L, Pan S-L, Jeng J-S, *et al*. Predicting recovery of voluntary upper extremity movement in subacute stroke patients with severe upper extremity paresis. *PLoS ONE*. 2015;10(5):e0126857.
125. Stinear CM. Prediction of motor recovery after stroke: Advances in biomarkers. *Lancet Neurol*. 2017;16(10):826–836.
126. Zhang X, Ji W, Li L, *et al*. The predictive value of motor-evoked potentials and the silent period on patient outcome after acute cerebral infarction. *J Stroke Cerebrovasc Dis*. 2016;25(7):1713–1720.
127. Stinear C. Prediction of recovery of motor function after stroke. *Lancet Neurol*. 2010;9(12):1228–1232.
128. Stinear CM, Barber PA, Petoe M, Anwar S, Byblow WD. The PREP algorithm predicts potential for upper limb recovery after stroke. *Brain*. 2012;135(8):2527–2535.
129. Stinear CM, Byblow WD, Ackerley SJ, Barber PA, Smith M-C. Predicting recovery potential for individual stroke patients increases rehabilitation efficiency. *Stroke*. 2017;48(4):1011–1019.
130. Stinear CM, Byblow WD, Ackerley SJ, Smith M-C, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. *Ann Clin Transl Neurol*. 2017;4(11):811–820.
131. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the upper limb prediction algorithm PREP2 applied 2 weeks poststroke: A prospective longitudinal study. *Neurorehabil Neural Repair*. 2021;35(1):68–78.
132. Stinear CM, Barber PA, Smale PR, Coxon JP, Fleming MK, Byblow WD. Functional potential in chronic stroke patients depends on corticospinal tract integrity. *Brain*. 2006;130(1):170–180.
133. Zhu LL, Lindenberg R, Alexander MP, Schlaug G. Lesion load of the corticospinal tract predicts motor impairment in chronic stroke. *Stroke*. 2010;41(5):910–915.
134. Wang LE, Tittgemeyer M, Imperati D, *et al*. Degeneration of corpus callosum and recovery of motor function after stroke: A multimodal magnetic resonance imaging study. *Hum Brain Mapp*. 2012;33(12):2941–2956.
135. Feng W, Wang J, Chhatbar PY, *et al*. Corticospinal tract lesion load: An imaging biomarker for stroke motor outcomes. *Ann Neurol*. 2015;78(6):860–870.
136. Liou L-M, Chen C-F, Guo Y-C, *et al*. Cerebral white matter hyperintensities predict functional stroke outcome. *Cerebrovasc Dis*. 2010;29(1):22–27.
137. Kang H-J, Stewart R, Park M-S, *et al*. White matter hyperintensities and functional outcomes at 2 weeks and 1 year after stroke. *Cerebrovasc Dis*. 2013;35(2):138–145.
138. Cheng B, Forkert ND, Zavaglia M, *et al*. Influence of stroke infarct location on functional outcome measured by the modified Rankin scale. *Stroke*. 2014;45(6):1695–1702.
139. Wu O, Cloonan L, Mocking SJ, *et al*. Role of acute lesion topography in initial ischemic stroke severity and long-term functional outcomes. *Stroke*. 2015;46(9):2438–2444.
140. Lin DJ, Cloutier AM, Erler KS, *et al*. Corticospinal tract injury estimated from acute stroke imaging predicts upper extremity motor recovery after stroke. *Stroke*. 2019;50(12):3569–3577.
141. Puig J, Pedraza S, Blasco G, *et al*. Acute damage to the posterior limb of the internal capsule on diffusion tensor tractography as an early imaging predictor of motor outcome after stroke. *Am J Neuroradiol*. 2011;32(5):857–863.
142. Puig J, Blasco G, Daunis-I-Estadella J, *et al*. Decreased corticospinal tract fractional anisotropy predicts long-term motor outcome after stroke. *Stroke*. 2013;44(7):2016–2018.
143. Bigourdan A, Munsch F, Coupé P, *et al*. Early fiber number ratio is a surrogate of corticospinal tract integrity and predicts motor recovery after stroke. *Stroke*. 2016;47(4):1053–1059.
144. Kwon YH, Jeoung YJ, Lee J, *et al*. Predictability of motor outcome according to the time of diffusion tensor imaging in patients with cerebral infarct. *Neuroradiology*. 2012;54(7):691–697.
145. Forkert ND, Verleger T, Cheng B, Thomalla G, Hilgetag CC, Fiehler J. Multiclass support vector machine-based lesion mapping predicts functional outcome in ischemic stroke patients. *PLoS ONE*. 2015;10(6):e0129569.

146. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes. *Acad Radiol*. 2020;27(2): e19–e23.
147. Nishi H, Oishi N, Ishii A, *et al*. Deep learning–derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion. *Stroke*. 2020;51(5):1484–1492.
148. Corbetta M, Ramsey L, Callejas A, *et al*. Common behavioral clusters and subcortical anatomy in stroke. *Neuron*. 2015; 85(5):927–941.
149. Siegel JS, Ramsey LE, Snyder AZ, *et al*. Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proc Natl Acad Sci USA*. 2016;113(30): E4367–E4376.
150. Ramsey LE, Siegel JS, Lang CE, Strube M, Shulman GL, Corbetta M. Behavioural clusters and predictors of performance during recovery from stroke. *Nat Hum Behav*. 2017;1(3):0038.
151. Moulton E, Valabregue R, Lehéricy S, Samson Y, Rosso C. Multivariate prediction of functional outcome using lesion topography characterized by acute diffusion tensor imaging. *NeuroImage Clin*. 2019;23:101821.
152. Takemura H, Palomero-Gallagher N, Axer M, *et al*. Anatomy of nerve fiber bundles at micrometer-resolution in the vervet monkey visual system. *eLife*. 2020;9:e55444.
153. Grefkes C, Fink GR. Reorganization of cerebral networks after stroke: New insights from neuroimaging with connectivity approaches. *Brain*. 2011;134(5):1264–1276.
154. Grefkes C, Fink GR. Connectivity-based approaches in stroke and recovery of function. *Lancet Neurol*. 2014;13(2):206–216.
155. Ovadia-Caro S, Margulies DS, Villringer A. The value of resting-state functional magnetic resonance imaging in stroke. *Stroke*. 2014;45(9):2818–2824.
156. Baldassarre A, Ramsey LE, Siegel JS, Shulman GL, Corbetta M. Brain connectivity and neurological disorders after stroke. *Curr Opin Neurol*. 2016;29(6):706–713.
157. Ogawa S, Lee T-M, Kay AR, Tank DW. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci USA*. 1990;87(24):9868–9872.
158. Friston KJ. Functional and effective connectivity in neuroimaging: A synthesis. *Hum Brain Mapp*. 1994;2(1–2):56–78.
159. Rehme AK, Volz LJ, Feis D-L, *et al*. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. *Cereb Cortex*. 2014;25(9):3046–3056.
160. Rehme AK, Volz LJ, Feis D-L, Eickhoff SB, Fink GR, Grefkes C. Individual prediction of chronic motor outcome in the acute post-stroke stage: Behavioral parameters versus functional imaging. *Hum Brain Mapp*. 2015;36(11):4553–4565.
161. Demeurisse G, Demol O, Robaye E. Motor evaluation in vascular hemiplegia. *Eur Neurol*. 1980;19(6):382–389.
162. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res*. 1981;4(4):483–492.
163. Harrell Jr FE. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer; 2015.
164. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput*. 2018;23:192–203.
165. Bzdok D, Krzywinski M, Altman N. Points of significance: Machine learning: Supervised methods. *Nat Meth*. 2018;15(1): 5–6.
166. Kather JN, Heij LR, Grabsch HI, *et al*. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1(8):789–799.
167. McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature*. 2020; 577(7788):89–94.
168. Lotter W, Diab AR, Haslam B, *et al*. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27(2):244–249.
169. Dalca AV, Guttag J, Sabuncu MR. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018:9290–9299.
170. Dubost F, de Bruijne M, Nardin M, *et al*. Multi-atlas image registration of clinical data with automated quality assessment using ventricle segmentation. *Med Image Anal*. 2020;63:101698.
171. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput*. 1997;1(1):67–82.
172. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect Psychol Sci*. 2017;12(6):1100–1122.
173. Christodoulou E, Jie MA, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
174. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open*. 2020;3(1):e1918962.
175. Koch PJ, Hummel FC. Toward precision medicine: Tailoring interventional strategies based on noninvasive brain stimulation for motor recovery after stroke. *Curr Opin Neurol*. 2017; 30(4):388–397.
176. Price CJ, Hope TM, Seghier ML. Ten problems and solutions when predicting individual outcome from lesion site after stroke. *Neuroimage*. 2017;145:200–208.
177. Rutten-Jacobs LC, Arntz RM, Maaijwee NA, *et al*. Long-term mortality after stroke among adults aged 18 to 50 years. *JAMA*. 2013;309(11):1136–1144.
178. Sanossian N, Ovbiagele B. Prevention and management of stroke in very elderly patients. *Lancet Neurol*. 2009;8(11):1031–1041.
179. Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G. Joint prediction of multiple scores captures better individual traits from brain images. *Neuroimage*. 2017;158:145–154.
180. Lin DJ, Erler KS, Snider SB, *et al*. Cognitive demands influence upper extremity motor performance during recovery from acute stroke. *Neurology*. 2021;96(21):e2576–e2586.
181. Bonkhoff AK, Hope T, Bzdok D, *et al*. Recovery after stroke: the severely impaired are a distinct group. *J Neurol Neurosurg Psy*. 2021.
182. Rose S. Machine learning for prediction in electronic health data. *JAMA Netw Open*. 2018;1(4):e181404.
183. Meehl PE. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press; 1954.
184. Lipton ZC. The mythos of model interpretability. *Queue*. 2018; 16(3):31–57.
185. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101–1102.
186. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981.
187. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA*. 2018;319(1):19–20.

188. Christakis NA, Smith JL, Parkes CM, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: Prospective cohort study commentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition. *BMJ*. 2000;320(7233): 469–473.
189. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. *BMC Med Inform Decis Mak*. 2018;18(4):55–64.
190. Subbaswamy A, Saria S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345–352.
191. Finlayson SG, Subbaswamy A, Singh K, *et al*. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3): 283–286.
192. Wong A, Otles E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181:1065–1070.
193. Krumholz HM. Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Aff*. 2014;33(7):1163–1170.
194. Blei DM, Smyth P. Science and data science. *Proc Natl Acad Sci USA*. 2017;114(33):8689–8692.
195. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30–36.
196. Rivera SC, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363.
197. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *BMJ*. 2020; 370:m3164.
198. Meng X-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2):685–726.
199. Uylings HBM, Rajkowska G, Sanz-Arigita E, Amunts K, Zilles K. Consequences of large interindividual variability for human brain atlases: Converging macroscopical imaging and microscopical neuroanatomy. *Anat Embryol*. 2005;210(5–6):423–431.
200. Wilkinson J, Arnold KF, Murray EJ, *et al*. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health*. 2020;2:e677–e680.
201. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. *CHANCE*. 2019;32(1):42–49.
202. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287–1289.
203. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med*. 2020;26(1):16–17.
204. Peabody FW. The care of the patient. *J Am Med Assoc*. 1927;88-(12):877–882.
205. Smith PL, Little DR. Small is beautiful: In defense of the small-N design. *Psychon Bull Rev*. 2018;25(6):2083–2101.
206. Krakauer JW, Carmichael ST. *Broken movement: The neurobiology of motor recovery after stroke*. MIT Press; 2017.
207. van der Vliet R, Selles RW, Andrinopoulou E-R, *et al*. Predicting upper limb motor impairment recovery after stroke: A mixture model. *Ann Neurol*. 2020;87:383–393.
208. Bonkhoff AK, Lim J-S, Bae H-J, *et al*. Generative lesion pattern decomposition of cognitive impairment after stroke. *Brain Commun*. 2021;3:fcab110.
209. Bonkhoff AK, Schirmer MD, Bretzner M, *et al*. Outcome after acute ischemic stroke is linked to sex-specific lesion patterns. *Nat Commun*. 2021;12(1):3289.
210. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
211. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat*. 1992;46(3):175–185.
212. Bzdok D, Ioannidis JPA. Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci*. 2019;42(4): 251–262.