Prediction of Stroke Outcome in Mice Based on Non-Invasive MRI and Behavioral Testing

Felix Knab#[1,2], Stefan Paul Koch#[1-3], Sebastian Major[1,2], Tracy D. Farr[1-4], Susanne Mueller[1-3], Philipp Euskirchen[1], Moritz Eggers[1,2], Melanie T.C. Kuffner[1,2], Josefine Walter[1,2,5], Jens P. Dreier[1,2,6-7], Matthias Endres[1,2,6,8-10], Ulrich Dirnagl[1,2,6,8-10] Nikolaus Wenger[1,2,6,11], Christian J. Hoffmann[1,2,11], Philipp Boehm-Sturm#[1-3], Christoph Harms[1,2,6,8,9]#

[1]Charité Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Klinik und Hochschulambulanz für Neurologie, Department of Experimental Neurology, Berlin, Germany
[2]Charité Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Center for Stroke Research Berlin, Charité-Universitätsmedizin Berlin, Germany
[3]Charité Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, NeuroCure Cluster of Excellence and Charité Core Facility 7T Experimental MRIs, Berlin, Germany
[4]School of Life Sciences, University of Nottingham, UK, NG7 2UH
[5]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, QUEST Center for Transforming Biomedical Research, Berlin, Germany
[6]Einstein Center for Neuroscience, Berlin, Germany
[7]Bernstein Center for Computational Neuroscience
[8]German Center for Cardiovascular Research (DZHK), partner site Berlin
[9]NeuroCure Clinical Research Center, Charité-Universitätsmedizin Berlin, Berlin, Germany
[10]German Center for Neurodegenerative Diseases (DZNE)
[11]Berlin Institute of Health (BIH), Berlin, Germany
#denotes equal contribution

Prediction of stroke outcome in mice

Correspondence to: Christoph Harms, Charité - Universitaetsmedizin Berlin, Center for Stroke Research Berlin, Department of Experimental Neurology, Charitéplatz 1, 10117 Berlin, Germany
Fax number: +4930450 560942
Telephone Number: +4930450560631
Email address: christoph.harms@charite.de

Total word count: 7778

ABSTRACT

Background: Prediction of post-stroke outcome using the degree of subacute deficit or magnetic resonance imaging metrics is well studied in humans. While mice are the most commonly used animals in pre-clinical stroke research, systematic analysis of outcome predictors is lacking.

Methods: Data from a total of 13 studies that included 45 minutes of middle cerebral artery occlusion on 148 mice were pooled. Motor function was measured using a modified protocol for the staircase test of skilled reaching. Phases of subacute and residual deficit were defined. Magnetic resonance images of stroke lesions were co-registered on the Allen Mouse Brain Atlas to characterize stroke topology. Different random forest prediction models that either used motor-functional deficit or imaging parameters were generated for the subacute and residual deficits.

Results: We detected both a subacute and residual motor-functional deficit after stroke in mice. Different functional severity grades and recovery trajectories could be observed. We found that lesion volume is the best predictor of subacute deficit. The residual deficit can be predicted most accurately by the degree of the subacute deficit. When using imaging parameters for the prediction of the residual deficit, including information about the lesion topology increases prediction accuracy. A subset of anatomical regions within the ischemic lesion have an outstanding impact on the prediction of long-term outcome. Prediction accuracy depends on the degree of functional impairment.

Conclusions: For the first time, we identified and characterized predictors of post-stroke outcome in a large cohort of mice and found strong concordance with clinical data. In the future, using outcome prediction can improve the design of pre-clinical studies and guide intervention decisions.

NON-STANDARD ABBREVIATIONS AND ACRONYMS

CCA Common Carotid Artery
ECA External Carotid Artery
FOV Field Of View
ICA Internal Carotid Artery
IQR Interquartile Range
MCA Middle Cerebral Artery
MCAO Middle Cerebral Artery Occlusion
MedAE Median Absolute Error
PA Pterygopalatine Artery
T2w T2-weighted

Keywords

stroke, mice, outcome prediction, motor-function, staircase

Novelty and Significance

What Is Known?

- Prediction of post-stoke outcome is an increasingly important tool in clinical stroke research as it could guide treatment decisions
- Both the subacute deficit and MRI metrics have been studied extensively in humans
- Despite mice being the most important model animal in pre-clinical stroke research, evidence on stroke outcome prediction in mice is scarce

What New Information Does This Article Contribute?

- The best predictor of long-term outcome after stroke in mice is the degree of the subacute deficit
- Including lesion topology increases prediction accuracy and outperforms total lesion volume
- A small subset of grey and white matter regions within the ischemic lesion have a great impact on prediction of long-term outcome

In contrast to stroke research in humans, there are no studies analyzing the prediction of functional outcome in mice, despite rodents being the most important model animal in translational stroke research. In this study we built imaging and behavior based prediction algorithms using random forest and found that the lesion volume is the best predictor of the subacute deficit. The residual deficit can be predicted most accurately by the degree of the subacute deficit and including information about the lesion topology increases prediction accuracy. Prediction error depends on the severity grade of functional impairment. In the future, using outcome prediction can improve the design of pre-clinical studies and guide intervention decisions.

INTRODUCTION

Stroke remains a major public health challenge, as its occurrence rises in aging societies.[1] With decreasing mortality rates due to progress in acute stroke therapy, the need to develop treatments that improve functional outcome has become more pressing.[2] Understanding predictors of functional outcome after stroke both in humans and rodents is an essential step towards defining adequate treatment decisions and improving the design of pre-clinical studies.[3] Prediction of expected motor-functional outcomes in mice could potentially guide researchers in their treatment decisions in pre-clinical stroke intervention studies and support outcome-dependent stratifications. In humans, magnetic resonance imaging (MRI) is used to assess stroke volume and topology.[4] Lesion volume is most commonly used to analyze stroke severity and predict functional outcome, but the prognostic value has been shown to be better when including lesion topology measures.[5–11] This was also shown in an exploratory pre-clinical study using a porcine stroke model.[12] Additionally, the degree of the subacute deficit has been shown to be a reliable predictor of long-term outcome, especially for mild functional deficits.[13,14] In contrast to humans, little is known about predictors of post-stroke outcome in mice, which are commonly used in pre-clinical research. Although highly standardized procedures such as the filamentous middle cerebral artery occlusion (MCAO) have been developed in rodent stroke research, the lesion volume and topology of stroke differ between individual mice, for instance, because of variations in the vasculature and degree of collaterals, such as the patency of the posterior communicating artery.[15] In addition, lesion volumes are

different between mouse strains.[16] Due to the resulting variance of lesion volumes and topology as well as small sample sizes in rodent stroke research, a systematic analysis of predictors of post-stroke outcome using early non-invasive MRI is challenging. Furthermore, detecting motor-functional deficits in rodents following stroke is difficult as they show impressive recovery and have subtle, if any, long-term motor deficits in common behavioral tests.[17]

The main goal of this study was to identify and compare different predictors of functional outcome following brain ischemia in mice. We therefore aggregated imaging and behavioral data from 13 MCAO studies. Using a newly modified protocol for the staircase test of skilled reaching, we tested forepaw function before and after stroke over a period of 41 days. We defined two phases of post stroke recovery and tried to predict the achievable functional outcome after one week (subacute outcome) or three weeks (long-term outcome) using early non-invasive MRI measurements. We hypothesized that incorporating structural and behavioral data would improve the predictability of post-stroke outcomes compared to using the total lesion volume. Therefore, we compared the prognostic value of total lesion volume and stroke topology for predicting the subacute outcome after one week and additionally included the subacute deficit for predicting the long-term motor-functional outcome using machine learning on a large cohort of 148 mice. Our work opens new avenues towards a better understanding of post-stroke recovery and impairment in mice and paves the way to using prediction of functional outcome for the optimization of pre-clinical stroke studies.

METHODS

Experimental Design

Behavioral and imaging data from a total of 13 studies performed between 2016 and 2019 were pooled for joint analyses. In total, 360 mice were considered for inclusion into the study. Four exclusion criteria were applied to the dataset (Figure S1): **1)** mice did not learn the task indicated by average performance in the staircase test in the third week of testing below 20 % of the average of mice in the respective study (n = 30); **2)** no visible infarct in MRI (n = 18); **3)** death during surgery or euthanasia at humane endpoint prior to the end of the study (n = 121) and **4)** data sets were incomplete, e.g. behavioral data had not been recorded for several days (n = 25). Following these criteria, 166 mice were included in the final dataset. 148 mice received MCAO surgery (female: n = 40, male: n = 108) and 18 mice underwent sham surgery (female: n = 7, male: n = 11). A total of 15 different genotypes were used (Table S1). If activation of genotype was required (n = 50), either Tamoxifen (n = 43) or vehicle (n = 7) was used. A total of 19 mice were part of studies that included therapeutic treatment after the MCAO surgery of which 10 received an active agent and 9 received a placebo control (Table S1).

The person performing the behavioral testing was blinded to the intervention. Mice were housed in groups of five in a 12 h day cycle and were food restricted but had free access to drinking water. Mice were weighed every day to control weight. They were familiarized with the pellets used in the behavioral task and with the test apparatus within the first week of handling by putting pellets in their cage. Eight days later, training sessions began (Figure 1). Mice were initially food restricted for 20 h before training sessions; each session lasted 30 minutes. If weights stabilized around 85% of their free-feeding weights, food restriction was adjusted and increased to up to 22 h. After the training session, mice had individual access to standard lab chow diet to avoid food hoarding of the dominant mice. Mice were trained for fourteen days until average performance stabilized. We tested the mice for another seven consecutive days and the average performance during this period was used to normalize post-stroke pellet reaching. Performance was expressed as a percentage of average pre-stroke performance. Mice were randomized and allocated to either MCAO or sham surgery. 24 hours prior to the surgery mice had free access to food. After the surgery, mice had another two hours of access to food before the previously described cycle of food deprivation began. After MCAO lab chow diet pellets were homogenized using water to assist with feeding. 24 hours after the surgery MR imaging was performed. 48 hours after the surgery testing resumed for another twenty-one days at which point the animals were sacrificed.

Staircase Test of Skilled Reaching

The staircase test of skilled reaching in mice uses a testing box consisting of two stairs on the right and left side that were baited with up to two food pellets.[18] The pellets consisted of sucrose (20 mg Dustless Precision Pellets®, Bio-Serv). Only one pellet was put onto the two highest stairs as mice would use their tongue to obtain these. In one study, number of pellets per side was 8 (n = 19), in the remaining studies number of pellets per side was 14 (n = 147). We chose one experimental variable, the percentage of collected pellets per side. For optimal comparability of motor-function, the individual performance was normalized to the pre-stroke performance. Therefore, average percentage of retrieved pellets was calculated for the period of baseline testing. Performance was then calculated as a percentage of the respective baseline performance. Based on the overall dynamic of post-stroke performance, two phases were defined: a '*subacute deficit*' phase (days 2 to 6), during which performance appeared to be dynamic and was increasing over time, and a '*residual deficit*' phase (days 12 to 21), throughout which performance appeared to be stable. A gap of five days was chosen between the *subacute* and *residual deficit* to ensure the dynamic phase had ended.

Effect of Study-Related Variables

The effect of study-related concomitant variables (genotype, activation of genotype, therapeutic treatment, experiment number, sex) on performance was investigated using separate analysis of variance linear mixed-effects models for *subacute* and *residual deficit* (response variable), respectively. For *subacute deficit*, a mixed-effects ANOVA revealed no significant effect of study-related variables (genotype: $F(11, 121) = 0.917$, $p = 0.527$; activation of genotype: $F(1, 121) = 0.805$, $p = 0.371$; therapeutic treatment: $F(3, 121) = 0.328$, $p = 0.805$; experiment number: $F(5, 121) = 0.573$, $p = 0.720$; sex: $F(1, 121) = 0.193$, $p = 0.661$). For *residual deficit*, there were no significant effects (genotype: $F(11, 121) = 1.589$, $p = 0.110$; activation of genotype: $F(1, 121) = 3.521$, $p = 0.063$; therapeutic treatment: $F(3, 121) = 0.934$, $p = 0.427$; experiment number: $F(5, 121) = 1.452$, $p = 0.211$; sex: $F(1, 121) = 0.166$, $p = 0.684$).

Classification of Severity Grades for Subgroup Analysis

Alluding to the modified Rankin scale, a score regularly used in stroke patients, five different functional subgroups were defined.[19-20] Subgroups were based on the severity of the motor-functional deficit of the paretic paw: mild deficit (performance >80 %), moderate to mild deficit (60 to 79%), moderate deficit (40 to 59 %), moderate to severe deficit (20 to 39 %) and severe deficit (0 to 19 %). Both *subacute* and *residual deficit* were used to define subgroups, depending on the analysis context (Figure 2A-C). In terms of *subacute deficit*, nineteen mice had a mild, twenty-six mice had a moderate to mild, nineteen mice had a moderate, twenty-one mice had a moderate to severe and sixty-three mice had a severe *subacute deficit*. For *residual deficit*, forty-two mice had a mild, thirty-six mice had a moderate to mild, twenty mice had a moderate, twenty-nine mice had a moderate to severe and twenty-one mice had a severe *residual deficit* group.

MCAO Surgery

Mice were anaesthetized with 1.5-3.5 % isoflurane and maintained with 1.0 % to 2.5 % in a mixture of 30 % O_2 and 70 % N_2O. During the procedure, body temperature was measured and maintained at 36 to 37.5 °C using a heating blanket and rectal probe. After disinfection of the skin, a midline neck incision was made and the common carotid artery (CCA) was carefully dissected from the surrounding nerves and ligated using 7.0 suture thread. The external carotid artery (ECA) was separated and ligated and the internal carotid artery (ICA) and the pterygopalatine (PA) arteries were clipped. A 180 µm diameter Doccol® filament was introduced into the common carotid artery, secured with a suture, the clip opened, and the filament advanced. The mice were allowed to recover in a heated cage at thermoneutral temperature (30-31 °C) for the duration of the MCAO. After 45 minutes, the mice were re-anaesthetized and the filament was withdrawn. The skin was sutured. As for the sham procedure, the filament was inserted to occlude the MCA and withdrawn immediately.

MRI

T2-weighted (T2w) images were acquired 24 h post stroke surgery on a 7 T MR scanner (Bruker, Ettlingen, Germany) and a 20 mm inner diameter transmit/receive volume coil (RAPID Biomedical, Rimpar, Germany). Anesthesia was achieved using 1-2 % isoflurane in a 70:30 nitrous oxide:oxygen mixture, and body temperature and respiration rate were monitored with MRI compatible equipment (Small Animal Instruments, Inc., Stony Brook, NY, USA). A 2D RARE sequence was used (repetition time/echo spacing/effective echo time=4.2 s/12 ms/36 ms, RARE factor=8, slice thickness 0.5 mm, FOV = $(25.6\ \text{mm})^2$, image matrix=256x196 zerofilled to 256x256, 4 averages, acquisition time 6:43 min). The lesion was manually delineated by an experienced researcher using ANALYZE 5.0 software (AnalyzeDirect, Overland Park, KS, USA). MRI data were registered on the Allen Mouse Brain Atlas using the in-house developed MATLAB toolbox ANTx2 (https://github.com/ChariteExpMri/antx2).

Incidence maps expressing percentage of animals with a lesion in a voxel were plotted for each group in atlas space. *Lesion volume* in [mm^3] was measured in atlas space in order to correct for edema and to regress out differences in whole brain volume between animals.[21] Percentage of damage in a given region was calculated for each individual mouse using the overlap of delineated lesion masks and the ipsilateral region normalized to the volume of the ipsilateral region, i.e. percent damage ranged from 0% (no overlap of lesion mask with the ipsilateral region) and 100% (ipsilateral region completely lies within the lesion). The set of percent damages for all regions is referred to as *segmented MRI*.

Statistics

Behavioral data were analyzed using GraphPad Prism software, version 9.3.0. (La Jolla, CA, United States). A mixed-effects model followed by Šidák's post-hoc tests for multiple comparison were used to analyze the effect of side (right, paretic paw or left, non –paretic paw) and group (sham or MCAO) on performance. Alpha level was 0.05, mean difference and 95 % confidence intervals are reported. For the comparison of stroke volumes in different functional subgroups as well as comparison of different prediction models and their effect on the *prediction error*, either unpaired t-test or one-way ANOVA was used. ANOVA was followed by Tukey's test for multiple comparison. When the effects of the prediction model and the severity grades on the *prediction error* were analyzed, two-way-ANOVA followed by Tukey's test were performed.

Selection of Predictors

Two factors were analyzed for their predictive value for the degree of the *subacute deficit* after MCAO: 1) Total stroke *lesion volume* and 2) MRI lesion percentages measured in the 736 regions, that were registered on the Allen Mouse Brain Atlas. Throughout the manuscript, this predictor is referred to as *segmented MRI*. The dimensionality of the *segmented MRI* (736 initial regions) was reduced by discarding all regions that did not show a lesion in at least one animal. This left the *segmented MRI* with 536 regions. In addition to those two imaging-based parameters, for the prediction of the *residual deficit*, the degree of *subacute deficit* was included as a third factor.

Development of Prediction Models with Machine Learning

For the development of prediction models with machine learning we divided the whole data set into a training (2/3 of the animals) and a test group (1/3 of the animals). Since long-term outcome was the most important outcome measure in this study, we ensured a comparable distribution of the *residual deficit* in both groups. To achieve this, we first sorted the animals according to the *residual deficit*. Subsequently two animals from each consecutive triplet were randomly assigned to the training group and one to the test group, resulting in 98 animals in the training data set and 50 animals in the test data set.

Using the training data, we tested several regression methods in the regression learner app in MATLAB (version 2021a) and chose random forest, which showed best performance on training data. Given the relatively small data set we decided to train 50 independent models for each predictor additionally performing automated Bayesian hyperparameter optimization with 100 iterations for each single model. To assess the possible performance of each predictor we calculated the median absolute error (MedAE) for each single model using the test data set and report the average MedAE value as *prediction error* together with averaged first and third quartile and resulting interquartile range. The *prediction error* gives an estimate of the deviation between the predicted and the true performance. Small *prediction errors* indicate high *prediction accuracies*. First and third quartiles and interquartile ranges (IQR) of the absolute error were calculated to give an additional estimate of how the *prediction errors* deviate within a given cohort. Small IQRs indicate a lower variance of *prediction errors* within a cohort.

To better understand on which anatomical regions the accuracy of MRI-based prediction of long-term outcome depends, the impact of each region in the MRI on the prediction model was further analyzed. The out-of-bag predictor importance was estimated for each of the 50 random forest models trained on *segmented MRI* data.[22] This was followed by calculation of the median value across all models and sorting of the regions by importance in descending order.

Data Availability

Data is openly accessible on zenodo (https://doi.org/10.5281/zenodo.6534690).[23] The repository contains all raw MRI T2w images, lesion masks and the registered atlases in NIFTI format, the input for machine learning algorithms (lesion volume, segmented MRI, behavioral scores), the trained classifiers and their output (predicted behavioral scores).

RESULTS

MRI Characterization and Lesion Topology

Incidence maps were generated for both the total cohort and the functional subgroups (Figure 3A). Mean *lesion volume* for the total cohort was 25.00 mm^3 (95% CI: 21.46 to 28.54). Lesion volume was also calculated for each of the five functional subgroups. Here, residual deficit was used to define subgroups. Mean lesion volume was 14.03 mm^3 (9.82 to 18.25) in mice with a mild, 17.17 mm^3 (12.52 to 21.82) in mice with a moderate to mild, 26.29 mm^3 (16.86 to 35.73) in mice with a moderate, 33.62 mm^3 (24.04 to 43.20) in mice with a moderate to severe and 47.23 mm^3 (36.79 to 57.68) in mice with a severe residual deficit (Figure 3B). One-Way ANOVA revealed that there was a significant difference between mean lesion volumes of the different functional subgroups ($F(4, 143) = 14.14$, $p < 0.0001$). For a full statistical report, refer to Table S2. The T2-weighted MR-images were co-registered on the Allen Mouse Brain Atlas to evaluate the topology of the lesions (*segmented MRI*). Quantitative evaluation of the lesion was performed on 536 regions and lesion volume is expressed as percentage of the region that showed elevated T2-intensities. Among the affected regions were the caudoputamen and parts of various cortical structures such as the supplemental somatosensory area, the primary somatosensory area and the primary motor area.

A Modified Staircase Test Can Assess Early and Long-Term Functional Outcome in Mice after Stroke

Mice learned to reach for pellets with their forelimbs in a training period. The performance peaked after approximately 7 days (Figure 4A-B, day -14 prior MCAO). The following period from day -7 to day -1 prior stroke was termed as '*baseline*'. The average percentage of retrieved pellets during *baseline* testing was 49.26 % (47.22 to 51.31) on left (later non-paretic) side and 46.29 % (44.16 to 48.42) on the right (later paretic) side in the MCAO group. In the sham animals, *baseline* average percentage of retrieved pellets was 45.81 % (37.38 to 54.23) on the non-paretic side and 47.11 % (38.95 to 55.26) on the paretic side (Figure S2). Mixed-effects analysis showed no statistically significant interaction between factors side and group ($F(1, 164) = 1.464$, $p = 0.2281$). As described in more detail in the methods section, performance was normalized to baseline testing and is given as a percentage of pre-stroke performance. After MCAO, mean performance in the time period of *subacute deficit* was 61.60 % (56.12 to 67.08) on the non-paretic side and 37.59 % (31.92 to 43.26) on the paretic side (Figure 4C). In the sham animals, mean performance during the phase of *subacute deficit* was 78.40 % (63.46 to 93.34) on the non-paretic side and 72.23 % (57.30 to 87.15) on the paretic side. Mixed-effects analysis revealed that there was a statistically significant interaction between the effects side and group ($F(1, 164) = 4.189$, $p = 0.0423$). After correcting for multiple comparisons using Šidák's test, performance of the paretic paw was significantly lower in MCAO animals when compared to sham (least square mean difference = 34.63 p.p., $t = 4.092$, $DF = 328$, $p = 0.0001$). There was no significant difference between groups on the non-paretic side (least square mean difference = 16.80 p.p., $t = 1.985$, $DF = 328$, $p = 0.0937$). Mean performance in the period of *residual deficit* was 95.35 % (89.54 to 101.17) on the non-paretic side and 60.88 % (54.94 to 66.81) on the paretic side (Figure 4D). In the sham animals, mean performance during the phase of *residual deficit* was 92.39 % (81.80 to 102.98) on the non-paretic side and 91.56 % (78.45 to 104.66) on the paretic side. Mixed-effects analysis showed that there was a statistically significant interaction between the effects side and group ($F(1, 164) = 9.384$, $p = 0.0026$). After correcting for multiple comparisons using Šidák's test, performance of the paretic paw was significantly lower in MCAO animals when compared to sham (least square mean difference = 30.68 p.p., $t = 3.501$, $DF = 328$, $p = 0.0011$). No statistically significant difference between groups was detected on the non-paretic side (least square mean difference = 2.96 p.p., $t = 0.3377$, $DF = 328$, $p = 0.9302$).

In summary, differences between MCAO and sham were detected for both *subacute* and *residual deficit*. After subgrouping mice based on the degree of *residual deficit*, distinct trajectories of recovery were observed which heavily varied among mice. While some mice showed impressive improvement of motor-function over time, i.e. followed a trajectory from severe *subacute* to mild *residual deficit*, others showed very little to no recovery.

Prediction of Subacute Deficit from Acute MR Imaging

Next, prediction models for the *subacute deficit* were generated using random forest and data derived from early non-invasive MRI. Two predictors were analyzed for the prediction of the *subacute deficit*: 1) total *lesion volume* in mm^3 as measured in MRI and 2) lesion percentage in 536 regions of the mouse brain based on the data from the *segmented MRI*. *Prediction error* was 13.30 percentage points (p.p.) (13.00 to 13.59, Q_1: 5.98, Q_3: 31.48, IQR: 25.51) for the *lesion volume* and 16.23 p.p. (16.03 to 16.42, Q_1: 8.04, Q_3: 29.18, IQR: 21.14) for the *segmented MRI* (Figure 5A and 5B). Unpaired, two-tailed t-test revealed, that *prediction error* was significantly smaller when using the *lesion volume* as a predictor compared to the ***segmented MRI*** (mean difference: 2.93 p.p., 2.58 to 3.28, $t = 16.60$, $df = 98$, $p < 0.0001$).

In total, *lesion volume* was the best predictor of the *subacute deficit*, but the *segmented MRI* provided comparable results. Q_3 and IQR were smallest in the prediction model using the *segmented MRI*, indicating a smaller variance of prediction accuracy compared to the *lesion volume* based prediction model.

Prediction of Long-Term Outcome from Subacute Deficit and MR Imaging

For the prediction of the *residual deficit*, three predictors were analyzed: 1) *subacute deficit*, 2) *segmented MRI* and 3) total *lesion volume* in mm^3. *Prediction error* was 18.90 p.p. (18.70 to 19.09, Q_1: 8.04, Q_3: 29.09, IQR: 21.05) for the *subacute deficit*, 22.68 p.p. (22.46 to 22.90, Q_1: 14.07, Q_3: 41.84, IQR: 27.77) for the *segmented MRI* and 25.95 p.p. (25.71 to 26.29, Q_1: 12.21, Q_3: 41.25, IQR: 29.04) for the *lesion volume* (Figure 5C and 5D). One-way ANOVA revealed, that there was a statistically significant difference between predictors ($F(2, 147) = 1032$, $p < 0.0001$). After correction for multiple comparisons using Tukey's test, the *prediction error* of the *subacute deficit* was significantly smaller than *prediction error* of the *segmented MRI* (mean difference: 3.79 p.p., 4.15 to 3.42, $p < 0.0001$) and the *lesion volume* (mean difference: 7.06 p.p., 7.43 to 6.69, $p < 0.0001$). *Segmented MRI* outperformed the prediction accuracy of the *lesion volume* (mean difference: 3.28 p.p., 3.65 to 2.91, $p < 0.0001$).

In conclusion, the *subacute deficit* after MCAO provided a stable predictor of long-term outcome and outperformed imaging-based parameters. S*egmented MRI* outperformed total *lesion volume* indicating that lesion topology provides a better predictor of long-term functional outcome than mere lesion volume.

Identification of Anatomical Regions Are Associated with Long-Term Outcome

The *segmented MRI* outperformed *lesion volume* as a predictor of the *residual deficit*, suggesting a bigger influence of lesion topology on the outcome compared to lesion volume. To better understand which anatomical regions within the lesion best predict long-term outcome, the contribution of each region measured by MRI was analyzed by quantifying its importance within the random forest algorithm as described in the methods section. We found 81 regions with out-of-bag predictor importance above 0, indicating that these regions contribute more than random data (Figure 6A and 6B, Table S3). Among those regions were the caudoputamen, the primary somatosensory area, the external capsule and the corticospinal tract. In the next step, random forest models were trained using increasing number of MRI regions, starting with the most important region. The *prediction error* of those models gradually decreased with increasing number of MRI regions (Figure 6C). However, the *prediction error* quickly reached a plateau and the addition of more regions only lead to slight changes in

prediction accuracy. A first local minimum of *prediction error* could already be observed when including the first 14 anatomical regions into the prediction model (*prediction error*: 22.91 p.p.).
In summary, a small subset of anatomical regions has a major impact on the accuracy of imaging based prediction of long-term outcome after stroke.

Prediction Accuracy Depends on Severity Grade

P*rediction errors* were examined to better understand the relationship between severity and model accuracy (Figure 7). For the assessment of the prediction accuracy of the *subacute deficit* from *lesion volume* or from *segmented MRI*, severity grades were based on the *subacute deficit*. Two-Way ANOVA was performed to test whether severity grades and prediction model had an effect on the *prediction error*. Two-Way-ANOVA showed a statistically significant interaction between prediction model and severity grade ($F(4, 490) = 682.3$, $p < 0.0001$). Simple main effect analysis revealed that severity grade ($F(4, 490) = 2822$, $p < 0.0001$) and model ($F(1, 490) = 47.49$, $p < 0.0001$) had a statistically significant effect on the *prediction error*. For a full statistical report, see Table S4. For the assessment of the prediction accuracy of *residual deficit*, severity grades were based on the *residual deficit*. Besides the two imaging based parameters, the degree of *subacute deficit* was included in the analysis Two-Way-ANOVA showed a statistically significant interaction between predictor and severity grade ($F(8, 735) = 214.5$, $p < 0.0001$). Simple main effect analysis revealed that severity grade ($F(4, 735) = 1819$, $p < 0.0001$) and predictor ($F(2, 735) = 1241$, $p < 0.0001$) had a statistically significant effect on the *prediction error*. For a full statistical report, see Table S5.
In summary, *prediction errors* varied for each severity grade. For example, for the prediction of the *subacute deficit, prediction errors* were highest in mice with mild deficits. Similar trends could be observed in the prediction of the *residual deficit,* where outcome of mice with a moderate to severe, moderate or moderate to mild *residual deficit* could be predicted with high prediction accuracies and *prediction errors* for mild or severe deficits were much higher.

DISCUSSION

For the first time, we systematically compared different predictors of subacute and long-term functional outcomes after stroke in mice. We used both MRI derived parameters such as the total *lesion volume* and lesion topology, as well as early functional testing to predict the post-stroke performance in a pellet reaching task. Predictors of post-stroke outcome, such as imaging-derived parameters or functional data, have previously been extensively studied in humans.[2–6,8] While one pre-clinical study on predictors of stroke outcome in a porcine MCAO model exists, evidence on prediction of stroke outcome in mice is scarce, despite rodents being used most commonly as model animals in stroke studies.[12,24]

We modified previously published protocols for the staircase test of skilled reaching in mice, which enabled us to daily monitor motor function and to detect both subacute and long-term functional deficits.[18,25] Training the mice over a period of 21 days prior to the MCAO surgery and implementing separation during feeding hours enabled us to detect a side-specific disability of the forepaw over a period of three weeks post stroke. We observed distinct trajectories of functional recovery after stroke and the differentiation between recoverees and non-recoverees is comparable to what can be observed in humans and presents an important tool for pre-clinical rodent studies.[14] The resemblance to the human recovery process supports the predictive value of stroke modeling in mice and the case for implementing sophisticated behavioral tests that involve a longer phase of training prior to stroke for assessing long-term impairments.[26] For future applications, one has to consider the possible impact of food restriction included as part of the staircase training and testing on mortality and lesion volume.[27]

Lesion volume is the best predictor for the *subacute deficit* after stroke in mice as measured by *prediction error.* It outperforms the predictive accuracy of our newly developed imaging parameter, the *segmented MRI*, which included information about the lesion topology. It has previously been stated that lesion volume functions as a good predictor for acute stroke outcome in humans.[6] Our data suggests that this is also the case in mice. For the long-term motor-functional outcome however, the *segmented MRI* provided significantly better prediction accuracies. This indicates that for functional recovery, lesion topology is more relevant than lesion volume, which again is in line with clinical data.[12,28] It should be noted, that in the 45 minutes MCAO model, lesion volume and lesion topology are not fully independent predictors. Larger ischemic lesions typically include cortical regions while smaller lesions are restricted to subcortical structures.

Following up on the influence of lesion topology, we identified anatomical regions that drive imaging-based prediction of the long-term outcome. We found that considering only a subset of anatomical regions can already explain functional long-term outcome and discovered that prediction accuracy displayed a local maximum after 14 regions had been added to the prediction model. Among those regions were both, cortical and subcortical regions as well as grey matter and white matter regions, such as the corpus callosum and the fiber tracts. While this underlines the importance of specific grey matter structures in stroke, it also stresses the impact of white matter lesions on motor recovery. Previous studies have suggested that especially lesions of the corticospinal tract seem to be more important to both early and long-term functional outcome than lesion volume.[29–32]

Despite imaging-based predictors providing good estimates of post-stroke long-term outcome, the best predictor of the *residual deficit* was the degree of the *subacute deficit*, again confirming data from studies on predictors of post-stroke outcome in humans.[33] Lastly, we found that independent of the chosen model, the prediction accuracy for both the *subacute* and *residual deficits* depended on the degree of the deficit. In line with clinical data, we found that severe *residual deficits* are more difficult to predict than moderate-to-mild deficits.[13,14]

Study Limitations

We have restricted our prediction tools to 45 minutes of MCAO using endoluminal filament occlusion. As the data was pooled from 13 studies performed between 2016 and 2019, the dataset contained various co-variates. However, variance analysis revealed that differences in performance after stroke were not explained by sex, genotype, or treatment. We cannot exclude confounding effects of housing, lab chow diet, surgery, anesthesia, or further co-variates. Furthermore, we did not replicate our results further in a second independent cohort or with other behavioral readouts than the staircase test of skilled pellet reaching. Finally, we corrected only for multiple tests in a single test set and not for multiple comparisons between different tests or hypotheses.

CONCLUSIONS

In conclusion, we have here for the first time identified and characterized predictors of post-stroke outcome in mice. The concordance with clinical data strengthens the usability of mice as model animals in pre-clinical studies. Prediction of outcome in intervention studies using non-invasive MRI measurements or motor-functional data should be utilized to compare expected and achieved motor-function outcomes, select mice with particularly poor predicted outcomes, and guide treatment or intervention options. This will help to improve the reproducibility of pre-clinical rodent stroke research and close the translational gap.

Acknowledgments

Janet Lips coordinated experiments and contributed with MRI, handling, and behavioral evaluation; Larissa Mosch (behavior, histology), Monika Dopatka (surgeries), and Marco Foddis (surgeries, behavioral assessment, and MRI) all provided excellent technical assistance.

Sources of Funding

Funding was provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to C.J.H. and C.H. (Project number 417284923), to P.B.S. (Project number 428869206), to N.W., M.E., and C.H. (Project number 424778381 –TRR 295) and EXC NeuroCure and the German Federal Ministry of Education and Research (BMBF, Center for Stroke Research Berlin 01EO1301) to U.D., P.B.S., and C.H., and to P.B.S. by the BMBF under the ERA-NET NEURON scheme (01EW1811). This work was supported by Charité 3^{R}| Replace - Reduce – Refine and partly by the Fondation Leducq to M.E., and C.H.. F.K. and M.Eg. received a scholarship from the Berlin Institute of Health, Berlin. P.E., N.W., and C.J.H. are participants in the Charité Clinical Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health and N.W. is a Freigeist Fellow with support from the Volkswagen Foundation. In addition, this work was supported by DFG (Project number 73500270 and 413848220) and Era-Net Neuron EBio2, with funds from BMBF 01EW2004 to J.P.D.

Disclosures

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. No funding was received to assist with the preparation of this manuscript.

Ethics approval

All animal procedures were approved by and performed in concordance with local authorities (Landesamt für Gesundheit und Soziales) under license numbers G0197/12, G005/16, G0057/16, G0119/16, GG254/16, G0157/17, G0343/17.

References

1. Béjot Y, Bailly H, Durier J, Giroud M. Epidemiology of stroke in Europe and trends for the 21st century. *Press Medicale*. 2016;45:e391–e398. Available from: http://dx.doi.org/10.1016/j.lpm.2016.10.003
2. Eriksson M, Norrving B, Terént A, Stegmayr B. Functional outcome 3 months after stroke predicts long-term survival. *Cerebrovasc Dis*. 2008;25:423–429.
3. Farr TD, Wegener S. Use of magnetic resonance imaging to predict outcome after stroke: A review of experimental and clinical evidence. *J Cereb Blood Flow Metab*. 2010;30:703–717. Available from: http://dx.doi.org/10.1038/jcbfm.2010.5
4. Albers GW, Thijs VN, Wechsler L, Kemp S, Schlaug G, Skalabrin E, Bammer R, Kakuda W, Lansberg MG, Shuaib A, Coplin W, Hamilton S, Moseley M, Marks MP. Magnetic resonance imaging profiles predict clinical response to early reperfusion: The diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study. *Ann Neurol*. 2006;60:508–517.
5. Schiemanck SK, Post MWM, Kwakkel G, Witkamp TD, Kappelle LJ, Prevo AJH. Ischemic lesion volume correlates with long-term functional outcome and quality of life of middle cerebral artery stroke survivors. *Restor Neurol Neurosci*. 2005;23:257–263.
6. Lövblad KO, Baird AE, Schlaug G, Benfield A, Siewert B, Voetsch B, Connor A, Burzynski C, Edelman RR, Warach S. Ischemic lesion volumes in acute stroke by diffusion-weighted magnetic resonance imaging correlate with clinical outcome. *Ann Neurol*. 1997;42:164–170.
7. Pinto A, McKinley R, Alves V, Wiest R, Silva CA, Reyes M. Stroke lesion outcome prediction based on MRI imaging combined with clinical information. *Front Neurol*. 2018;9:1–10.
8. Cheng B, Forkert ND, Zavaglia M, Hilgetag CC, Golsari A, Siemonsen S, Fiehler J, Pedraza S, Puig J, Cho TH, Alawneh J, Baron JC, Ostergaard L, Gerloff C, Thomalla G. Influence of stroke infarct location on functional outcome measured by the modified rankin scale. *Stroke*. 2014;45:1695–1702.
9. M.R. Borich, S.M. Brodie, W.A. Gray, S. Ionta and LAB. Understanding the role of the primary somatosensory cortex: Opportunities for rehabilitation. *Neuropsychologia*. 2015;79:246–255.
10. Munsch F, Sagnier S, Asselineau J, Bigourdan A, Guttmann CR, Debruxelles S, Poli M, Renou P, Perez P, Dousset V, Sibon I, Tourdias T. Stroke location is an independent predictor of cognitive outcome. *Stroke*. 2016;47:66–73.
11. Menezes NM, Ay H, Zhu MW, Lopez CJ, Singhal AB, Karonen JO, Aronen HJ, Liu Y, Nuutinen J, Koroshetz WJ, Sorensen a. G. The real estate factor: Quantifying the impact of infarct location on stroke severity. *Stroke*. 2007;38:194–197.
12. Scheulin KM, Jurgielewicz BJ, Spellicy SE, Waters ES, Baker EW, Kinder HA, Simchick GA, Sneed SE, Grimes JA, Zhao Q, Stice SL, West FD. Exploring the predictive value of lesion topology on motor function outcomes in a porcine ischemic stroke model. *Sci Rep*. 2021;11:1–15. Available from: https://doi.org/10.1038/s41598-021-83432-5
13. Zarahn E, Alon L, Ryan SL, Lazar RM, Vry MS, Weiller C, Marshall RS, Krakauer JW. Prediction of motor recovery using initial impairment and fMRI 48 h poststroke. *Cereb Cortex*. 2011;21:2712–2721.
14. van der Vliet R, Selles RW, Andrinopoulou ER, Nijland R, Ribbers GM, Frens MA, Meskers C, Kwakkel G. Predicting Upper Limb Motor Impairment Recovery after Stroke: A Mixture Model. *Ann Neurol*. 2020;87:383–393.
15. Knauss S, Albrecht C, Dirnagl U, Mueller S, Harms C, Hoffmann CJ, Koch SP, Endres M, Boehm-Sturm P. A Semiquantitative Non-invasive Measurement of PcomA

Patency in C57BL/6 Mice Explains Variance in Ischemic Brain Damage in Filament MCAo. *Front Neurosci*. 2020;14:1–11.
16. Maeda K, Hata R, Hossmann KA. Differences in the cerebrovascular anatomy of C57Black/6 and SV129 mice. *Neuroreport*. 1998;9:1317–1319.
17. Weber RZ, Mulders G, Kaiser J, Tackenberg C, Rust R. Deep learning based behavioral profiling of rodent stroke recovery. *bioRxiv*. 2021;Available from: https://doi.org/10.1101/2021.08.11.455647
18. Baird AL, Meldrum A, Dunnett SB. The staircase test of skilled reaching in mice. *Brain Res Bull*. 2001;
19. Bamford JM, Sandercock PAG, Wariow CP, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients: To the editor. *Stroke*. 1989;20:828.
20. Farrell B, Godwin J, Richards S, Warlow C. The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: Final results. *J Neurol Neurosurg Psychiatry*. 1951;54:1044–1054.
21. Koch S, Mueller S, Foddis M, Bienert T, von Elverfeldt D, Knab F, Farr TD, Bernard R, Dopatka M, Rex A, Dirnagl U, Harms C, Boehm-Sturm P. Atlas registration for edema-corrected MRI lesion volume in mouse stroke models. *J Cereb Blood Flow Metab*. 2019;39:313–323.
22. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32. Available from: https://doi.org/10.1023/A:1010933404324
23. Knab F, Koch SP, Major S, Farr TD, Mueller S, Euskirchen P, Eggers M, Kuffner MTC, Walter J, Dreier JP, Endres M, Dirnagl U, Wenger N, Hoffmann CJ, Boehm-Sturm P, Harms C. Open data repository, Knab et al., Prediction of stroke outcome in mice based on non-invasive MRI and behavioral testing. 2022; Available from: https://zenodo.org/record/6546522
24. Casals JB, Pieri NCG, Feitosa MLT, Ercolin ACM, Roballo KCS, Barreto RSN, Bressan FF, Martins DS, Miglino MA, Ambrósio CE. The use of animal models for stroke research: a review. *Comp Med*. 2011;61:305–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22330245
25. Bravo-Ferrer I, Cuartero MI, Zarruk JG, Pradillo JM, Hurtado O, Romera VG, Díaz-Alonso J, García-Segura JM, Guzmán M, Lizasoain I, Galve-Roperh I, Moro MA. Cannabinoid type-2 receptor drives neurogenesis and improves functional outcome after stroke. *Stroke*. 2017;48:204–212.
26. Corbett D, Carmichael ST, Murphy TH, Jones TA, Schwab ME, Jolkkonen J, Clarkson AN, Dancause N, Wieloch T, Johansen-Berg H, Nilsson M, McCullough LD, Joy MT. Enhancing the Alignment of the Preclinical and Clinical Stroke Recovery Research Pipeline: Consensus-Based Core Recommendations from the Stroke Recovery and Rehabilitation Roundtable Translational Working Group ∗. *Neurorehabil Neural Repair*. 2017;31:699–707.
27. Manzanero S, Gelderblom M, Magnus T, Arumugam T V. Calorie restriction and stroke. *Exp Transl Stroke Med*. 2011;3:1–13.
28. Menezes NM, Ay H, Zhu MW, Lopez CJ, Singhal AB, Karonen JO, Aronen HJ, Liu Y, Nuutinen J, Koroshetz WJ, Sorensen AG. The real estate factor: Quantifying the impact of infarct location on stroke severity. *Stroke*. 2007;38:194–197.
29. Etherton MR, Wu O, Giese AK, Lauer A, Boulouis G, Mills B, Cloonan L, Donahue KL, Copen W, Schaefer P, Rost NS. White Matter Integrity and Early Outcomes After Acute Ischemic Stroke. *Transl Stroke Res*. 2019;10:630–638.
30. Zhu LL, Lindenberg R, Alexander MP, Schlaug G. Lesion load of the corticospinal tract predicts motor impairment in chronic stroke. *Stroke*. 2010;41:910–915.
31. Feng W, Wang J, Chhatbar PY, Doughty C, Landsittel D, Lioutas V, Kautz SA, Schlaug G. Corticospinal tract lesion load: An imaging biomarker for stroke motor

outcomes. *Ann Neurol*. 2015;78:860–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26289123

32. Lindenberg R, Renga V, Zhu LL, Betzler F, Alsop D, Schlaug G. Structural integrity of corticospinal motor fibers predicts motor impairment in chronic stroke. *Neurology*. 2010;74:280–287.

33. Frankel MR, Morgenstern LB, Kwiatkowski T, Lu M, Tilley BC, Broderick JP, Libman R, Levine SR, Brott T. Predicting prognosis after stroke: A placebo group analysis from the National Institute of Neurological Disorders and Stroke rt-PA stroke trial. *Neurology*. 2000;55:952–959.

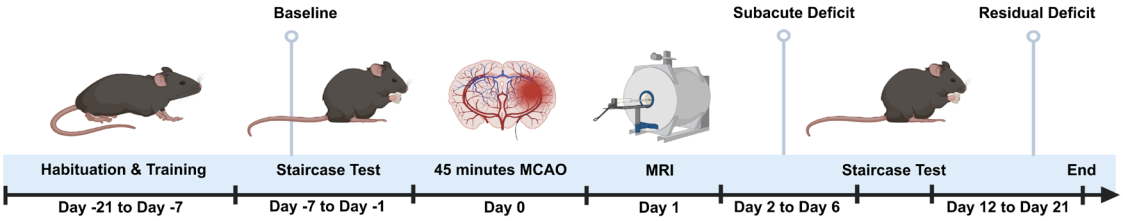

Figure 1. Experimental design
Mice were trained in pellet reaching for 21 days. Average percentage of retrieved pellets during the last week prior to MCAO was used to normalize performance (baseline). Mice received either 45 minutes of MCAO or sham surgery. Magnetic resonance imaging was performed on day 1 and mice were then tested for another 21 days (*Created with BioRender.com*).

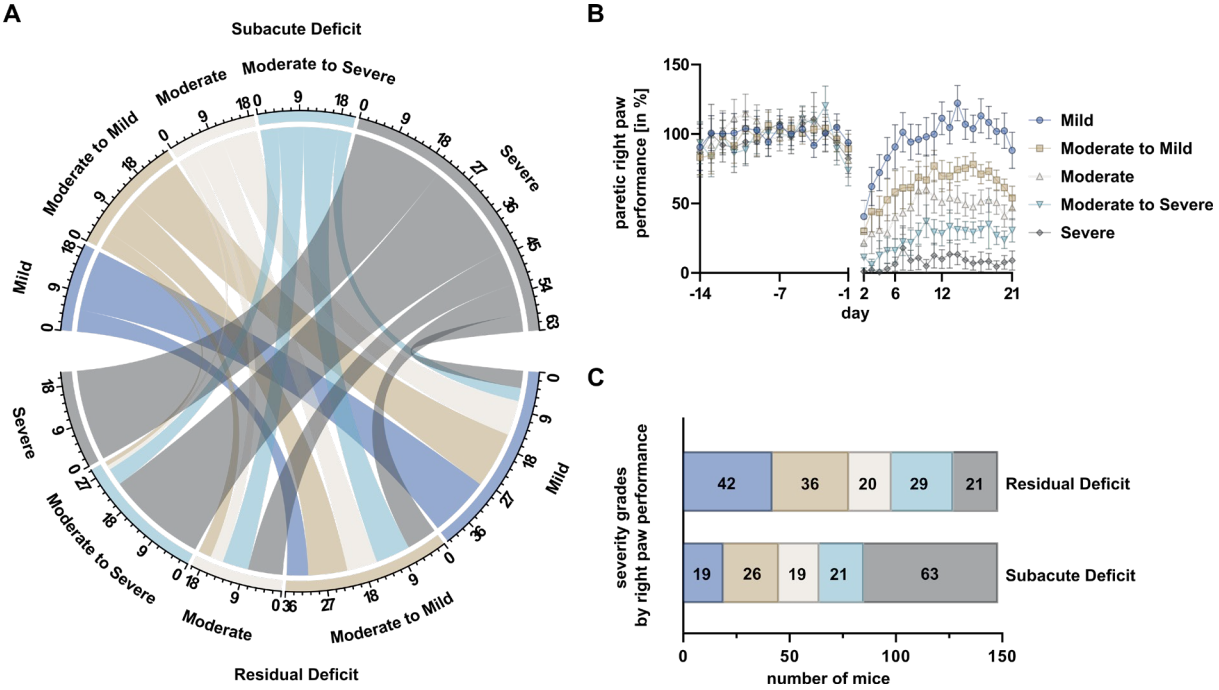

Figure 2 Identification and assessment of functional subgroups.
A, Chord diagram displaying the different trajectories of recovery. Groups on top display mice stratified by degree of *subacute deficit*, while groups on bottom display mice stratified by *residual deficit*. Numbers indicate group sizes. Interconnections visualize how many mice with a given *subacute deficit* recovered to have a given *residual deficit*, i.e. roughly a third of the mice with a severe *subacute deficit* displayed no recovery and showed a severe *residual deficit*. **B**, Performance of the paretic right paw over time for each severity grade. Error bars indicate 95 % confidence interval. Severity grades here were based on the degree of the *residual deficit*. **C**, Overview over sizes of functional subgroups. Groups are either based on the degree of *subacute* or *residual deficit*. While the majority of mice displayed a severe *subacute deficit*, due to their impressive capability for spontaneous recovery, the majority of mice displayed only mild *residual deficits*.

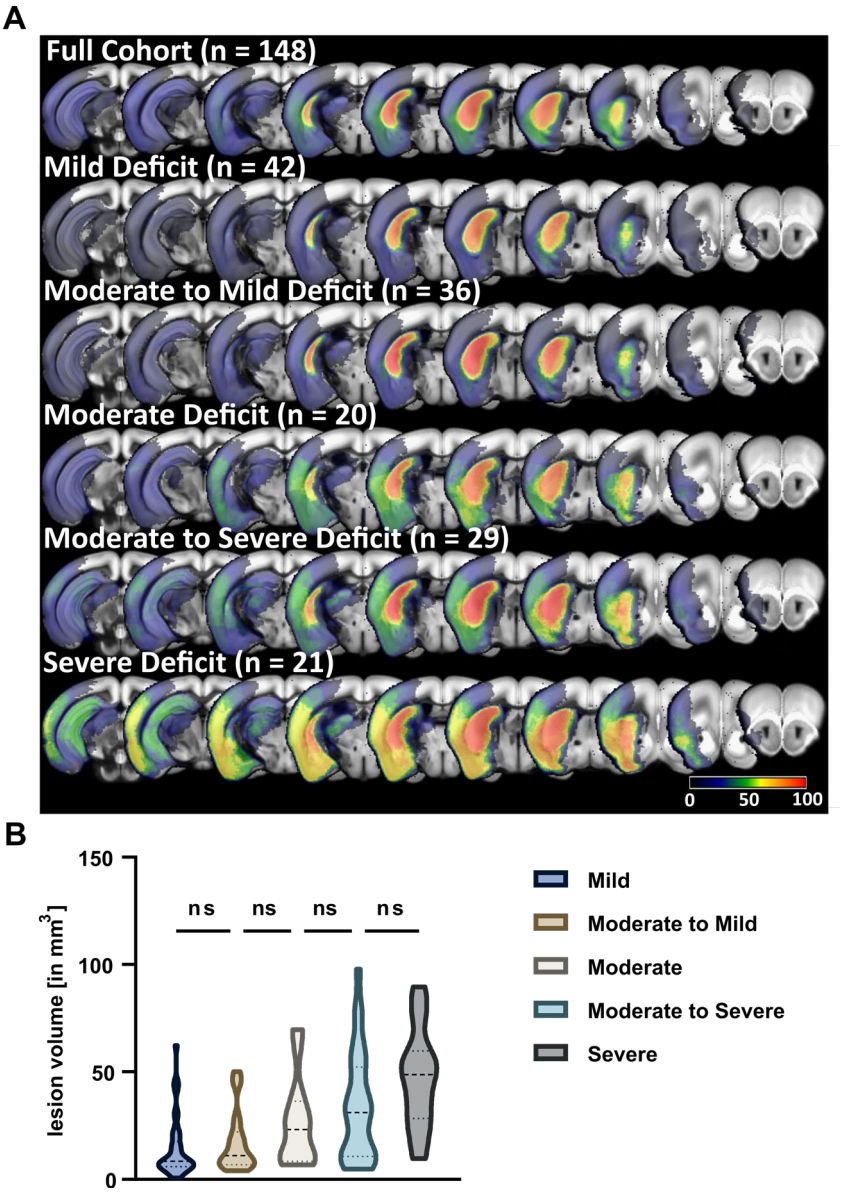

Figure 3 MRI characterization of lesion volume and location.
A, Coronal incidence maps of stroke lesions in the total cohort and the five different functional subgroups. Colours indicate percentage of mice with a lesion in a given region. Severity grades are based on degree of *residual deficit*. **B**, Total lesion volumes by severity grade (based on degree of *residual deficit*). Great variability within functional subgroups could be observed and there was no significant difference when comparing lesion volumes of consecutive severity grades, i.e. lesion volume of mice with mild or moderate to mild *residual deficits*.

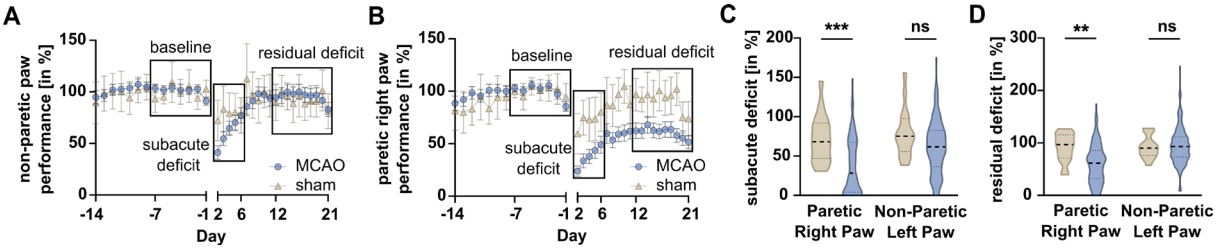

Figure 4 Assessing the sensorimotor deficit after stroke.
A, Non-paretic left and **B**, paretic right paw performance over time, bars depict 95 % confidence interval. Performance on days -7 to -1 were averaged and taken as a baseline (first black rectangle). Data from days 2 to 6 were summarized and referred to as *subacute deficit* (second rectangle) while data from days 12 to 21 are referred to as *residual deficit* (third rectangle). **C**, *Subacute* and **D**, *residual deficit* of the non-paretic and paretic paw in sham (brown plot) and MCAO (blue plot) animals. Dashed lines indicate medians, dotted line indicate quartiles. Mixed-effects model followed by Šidák's post-hoc tests for multiple comparison were used to test effect of group and side.

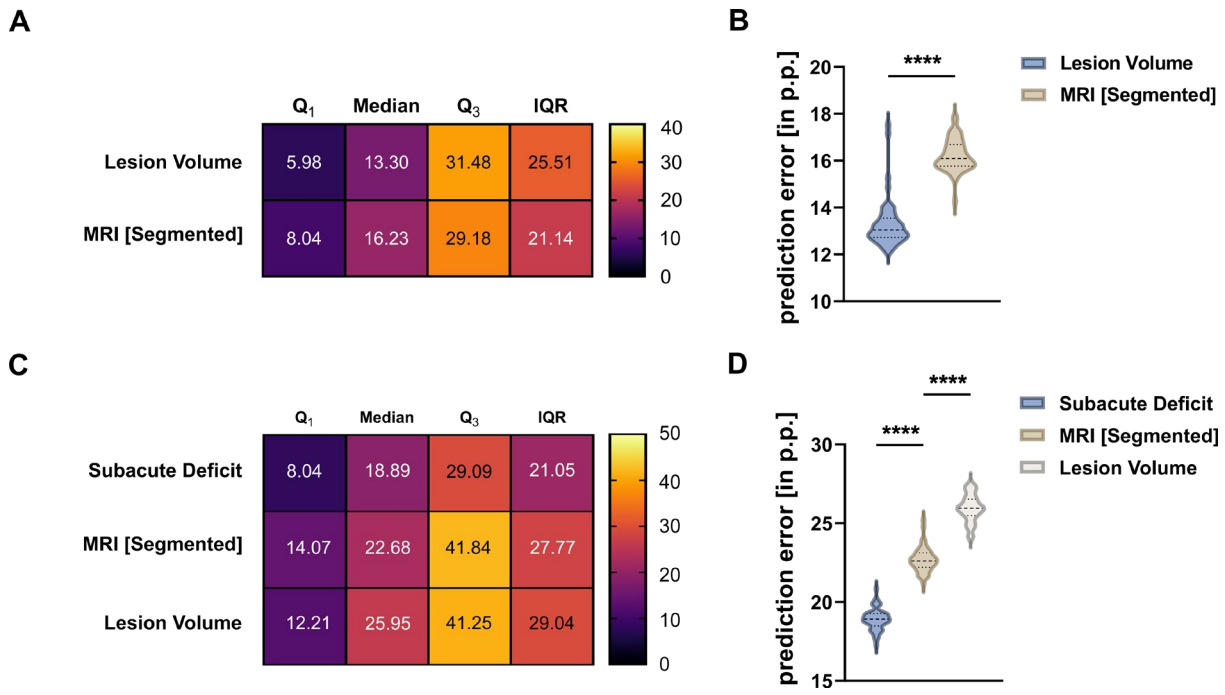

Figure 5 Comparison of predictors of the subacute and residual deficit.
A through **B**, Comparison of the prediction accuracy of the two predictors of the *subacute deficit*. Averaged Q_1, median, Q_3 and IQR were calculated using the 50 independent models for each predictor. Dashed lines indicate median, dotted lines indicate quartiles. Unpaired t-test was used to compare *prediction errors*. Small *prediction error* indicate small deviation between predicted and achieved performance. **C** through **D**, Comparison of prediction accuracy of the two imaging based parameters and the *subacute deficit* for the prediction of the *residual deficit*. Again, one-way ANOVA followed by Tukey's test for multiple comparison was used.

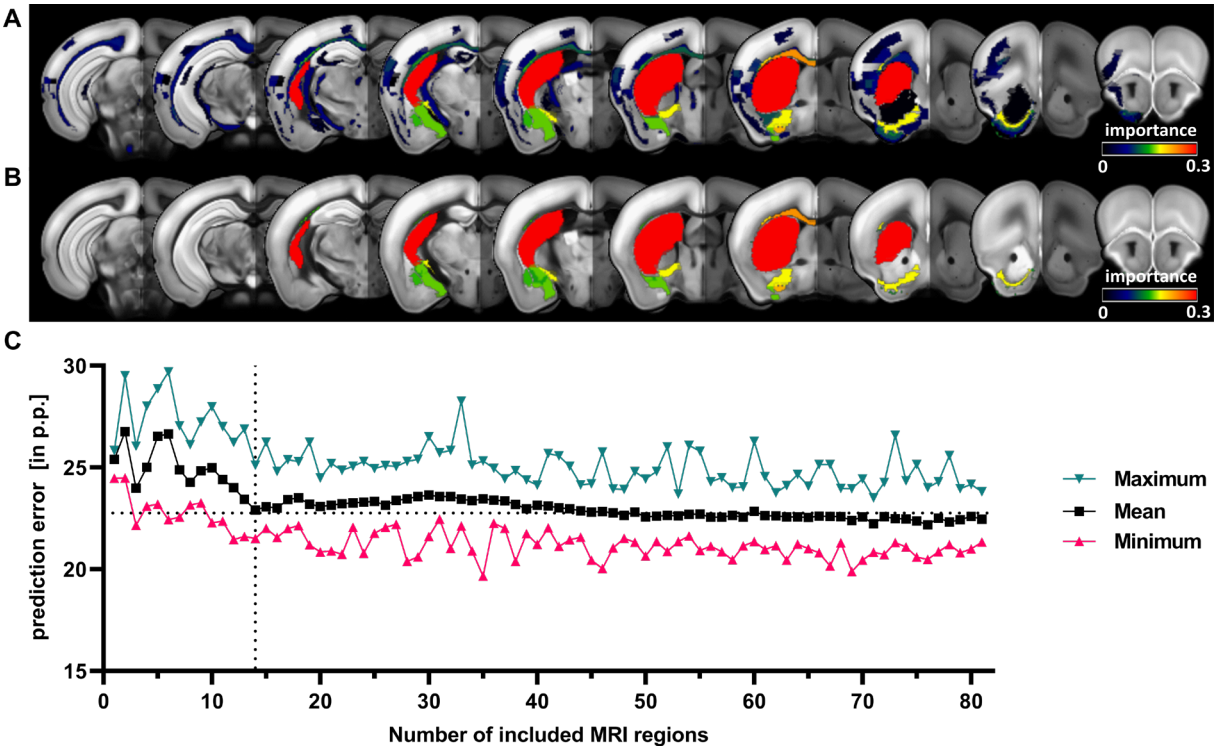

Figure 6 Identification of anatomical regions that drive imaging-based prediction of functional outcome.

A, Coronal slices depicting the importance of the 81 regions calculated from our random forest models (RF). Only regions with an importance level >0 are shown **B**, Visualization of the importance level of the 14 anatomical regions that were included in the prediction model that resulted in a first local minimum of *prediction error*. **C**, Prediction of *residual deficit* was performed using an increasing number of regions in the *segmented MRI*. *Prediction error* as well as maximum and minimum *prediction error* are depicted. Dashed lines at x = 14 and y = 21.5 denote the first local minimum of *prediction error* when including only 14 regions.

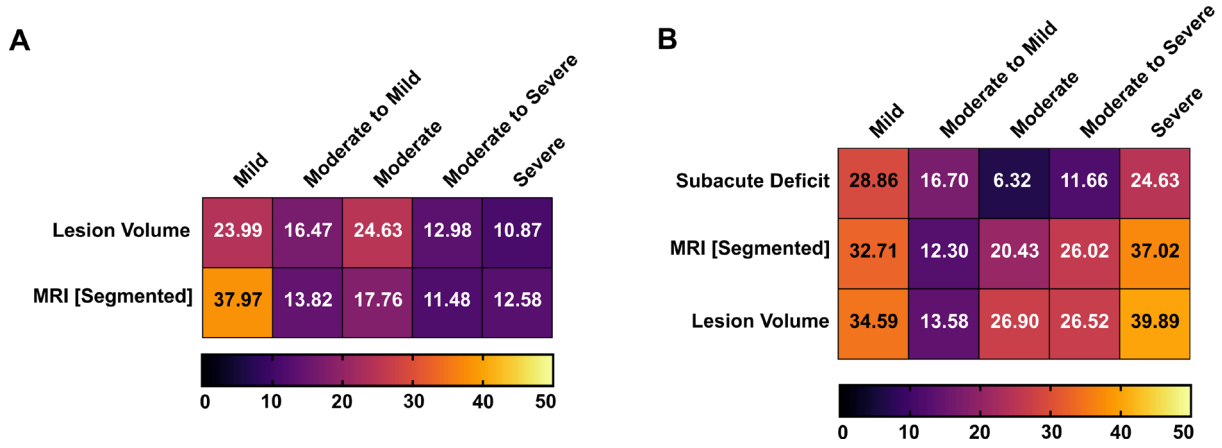

Figure 7 Comparison of severity grade dependent prediction accuracy.
A, *Prediction error* of the two predictors of the *subacute deficit* over the severity grades. Independent of the predictor, errors were higher in mice with mild compared to moderate to severe *subacute deficits*, indicating that the prediction models work especially well in those two crucial subgroups **B**, Prediction accuracy of the three predictors of the *residual deficit* over the severity grades. Again, a strong dependency of *prediction error* on the degree of deficit could be observed.