







## RESEARCH ARTICLE

# Prediction of stroke severity: systematic evaluation of lesion representations

Anna K. Bonkhoff<sup>1</sup> , Alexander L. Cohen<sup>2,3</sup> , William Drew<sup>3</sup> , Michael A. Ferguson<sup>4</sup>, Aaliya Hussain<sup>3,5</sup>, Christopher Lin<sup>3</sup>, Frederic L. W. V. J. Schaper<sup>3</sup>, Anthony Bourached<sup>1,6</sup>, Anne-Katrin Giese<sup>7</sup>, Lara C. Oliveira<sup>1</sup>, Robert W. Regenhardt<sup>1</sup> , Markus D. Schirmer<sup>1</sup>, Christina Jern<sup>8,9</sup>, Arne G. Lindgren<sup>10,11</sup>, Jane Maguire<sup>12</sup>, Ona Wu<sup>13</sup>, Sahar Zafar<sup>14</sup> , John Y. Rhee<sup>15,16</sup>, Eyal Y. Kimchi<sup>17</sup>, Maurizio Corbetta<sup>18,19</sup>, Natalia S. Rost<sup>1,†</sup>, Michael D. Fox<sup>3,†</sup>  & MRI-GENIE and GISCOME Investigators and the International Stroke Genetics Consortium

<sup>1</sup>J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

<sup>2</sup>Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA

<sup>3</sup>Center for Brain Circuit Therapeutics, Department of Neurology, Psychiatry, and Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

<sup>4</sup>Brigham and Women's Hospital, Harvard Medical School, Psychiatry, and Radiology, Boston, Massachusetts, USA

<sup>5</sup>Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

<sup>6</sup>UCL Queen Square Institute of Neurology, University College London, London, UK

<sup>7</sup>Department of Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

<sup>8</sup>Department of Laboratory Medicine, the Sahlgrenska Academy, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden

<sup>9</sup>Department of Clinical Genetics and Genomics Gothenburg, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden

<sup>10</sup>Department of Neurology, Skåne University Hospital, Lund, Sweden

<sup>11</sup>Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden

<sup>12</sup>University of Technology Sydney, Sydney, Australia

<sup>13</sup>Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts, USA

<sup>14</sup>Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA

<sup>15</sup>Center for Neuro-oncology, Department of Medical Oncology, Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA

<sup>16</sup>Division of Adult Palliative Care, Department of Psychosocial Oncology and Palliative Care, Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA

<sup>17</sup>Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

<sup>18</sup>Department of Neuroscience and Padova Neuroscience Center, University of Padova, Padova, Italy

<sup>19</sup>Venetian Institute of Molecular Medicine (VIMM), Padova, Italy

## Correspondence

Anna K. Bonkhoff, J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, 175 Cambridge St, Suite 300, Boston, MA 02114, USA. E-mail: [abonkhoff@mgb.org](mailto:abonkhoff@mgb.org)

## Funding Information

A.L.C. was supported by a Shields Research grant from the Child Neurology Foundation, the Simons Foundation Autism Research Initiative Bridge to Independence Fellowship, and the National Institute of Mental Health (K23MH120510). R.W.R. is in part supported by research grants from NIH-NINDS (R25NS065743), Society of Vascular and Interventional Neurology, and Heitman Foundation for Stroke. N.S.R. is in part supported by NIH-NINDS (R01NS082285, R01NS086905, U19NS115388). M.D.F. was

## Abstract

**Objective:** To systematically evaluate which lesion-based imaging features and methods allow for the best statistical prediction of poststroke deficits across independent datasets. **Methods:** We utilized imaging and clinical data from three independent datasets of patients experiencing acute stroke ( $N_1 = 109$ ,  $N_2 = 638$ ,  $N_3 = 794$ ) to *statistically* predict acute stroke severity (NIHSS) based on lesion volume, lesion location, and structural and functional disconnection with the lesion location using normative connectomes. **Results:** We found that prediction models trained on small single-center datasets could perform well using within-dataset cross-validation, but results did not generalize to independent datasets (median  $R^2_{N_1} = 0.2\%$ ). Performance across independent datasets improved using large single-center training data ( $R^2_{N_2} = 15.8\%$ ) and improved further using multicenter training data ( $R^2_{N_3} = 24.4\%$ ). These results were consistent across lesion attributes and prediction models. Including either structural or functional disconnection in the models outperformed prediction based on volume or location alone ( $P < 0.001$ , FDR-corrected). **Interpretation:** We conclude that (1) prediction performance in independent datasets of patients with acute stroke cannot be inferred from cross-validated results within a

supported by the Ellison/Baszucki Foundation, the Nancy Lurie Marks Foundation, the Kaye Family Research Endowment, and the National Institutes of Health (grant nos. R01MH113929, R21MH126271, R56AG069086, R01MH115949, and R01AG060987).

Received: 9 January 2024; Revised: 2 August 2024; Accepted: 8 September 2024

doi: 10.1002/acn3.52215

†Shared last-authorship.

## Introduction

Acute stroke lesions disrupt physiological brain activity, both directly at the lesion location, but also indirectly through effects on the connected brain networks, and both likely contribute to stroke deficits.<sup>1</sup>

Recent studies have used neuroimaging-based lesion features, such as *direct* lesion location data, as well as *indirect* structural and functional lesion connectivity data, to predict poststroke deficits.<sup>2–5</sup> Direct lesion location data refer to lesion segmentations derived from fluid attenuated inversion recovery (FLAIR) or diffusion-weighted imaging (DWI) abnormalities visible on the patient's MRI scan.<sup>4,6,7</sup> Indirect lesion connectivity refers to either the fiber tracks intersected by the lesion (structural disconnection [SDC])<sup>8</sup> or functional connectivity (FC) between the lesion location and other brain regions (functional disconnection [FDC]),<sup>9,10</sup> as estimated from the combination of lesion information from individual patients and connectivity data from large cohorts of healthy participants.

Recent studies have come to different conclusions as to which lesion features are most potent for the prediction of deficits.<sup>2,4,5</sup> While some studies found structural disconnection to be more potent than functional disconnection,<sup>2,4</sup> others found that functional disconnection may be more relevant for more complex deficits.<sup>5</sup> These results<sup>2,4,5</sup> were predominantly based on predicting variance within a single dataset of ~130 patients recruited from a single hospital.<sup>11</sup> While this dataset is valuable and very well phenotyped, generalization of these findings to stroke populations at large is currently unclear. In addition, previous work has focused on linear prediction models (e.g., ridge regression,<sup>2,4</sup> linear growth models,<sup>5</sup> or canonical correlation analysis<sup>12</sup>), yet non-linear algorithms may also be useful for cross-dataset predictions.<sup>13</sup>

The aim of the current study was to systematically evaluate which imaging features and methods allow for the

dataset, as performance results obtained via these two methods differed consistently, (2) prediction performance can be improved by training on large and, importantly, *multicenter* datasets, and (3) structural and functional disconnection allow for improved prediction of acute stroke severity.

best statistical prediction\* of poststroke deficits at the time of image acquisition, leveraging three large independent stroke cohorts.<sup>11,16–19</sup> We evaluated four lesion-based imaging features (lesion volume, lesion location, structural disconnection, and functional disconnection) and three methodological variables: (1) properties of the training sample (e.g., number of patients and single-center vs. multicenter collection), (2) non-linear versus linear prediction models, and (3) performance estimation using cross-validation loops versus independent test data. We altogether aimed to match our methodological approach as closely as possible to prior stroke imaging prediction studies<sup>2,4,7</sup> to allow for comparability. In all cases, our outcome metric was the prediction of acute to subacute stroke severity, measured by the NIH stroke scale (NIHSS). Although this score can be generated clinically without the need for predictions from neuroimaging, we used NIHSS as an outcome variable for three reasons: (1) it is widely used across different clinical sites allowing for cross-dataset comparisons; (2) it is measured in close temporal proximity to the brain imaging used to define the lesion features; (3) it has clinical and prognostic value, informing acute stroke treatment decisions<sup>20</sup> and correlating with stroke outcomes in the chronic phase poststroke.<sup>21,22</sup>

\*We use “predict” and “prediction” with the statistical definition in mind, that is, we want to predict a new data output based on a specific data input, individually for each patient. This goal of prediction contrasts the one of inference, that is explaining a certain output at the group level.<sup>14,15</sup> (Chapter 2.1.1 for reference 2). Importantly, “new data” can both refer to future data that did not exist yet when the prediction was generated and therefore represent a true forecast, or simply refer to data that was not used for training of a model and was therefore unseen and new to it independent of when it was obtained.

## Methods

### Patient cohorts

Our study considered three separate cohorts of stroke patients with acute stroke: A single-center<sup>†</sup> cohort of patients with first-time ischemic and hemorrhagic stroke enrolled at Barnes-Jewish Hospital and the Rehabilitation Institute, Washington University, St. Louis (WashU cohort:  $N_1 = 109$ ),<sup>11</sup> a single-center cohort of patients with acute ischemic stroke (AIS) that were part of a retrospective Massachusetts General Hospital (MGH)-based study ( $N_2 = 638$ ),<sup>16,19</sup> and multicenter cohort of patients with AIS from the international MRI–Genetics Interface Exploration (MRI-GENIE) study ( $N_3 = 794$ , 5 individual centers in four different countries: Spain, Sweden, Belgium, and the USA).<sup>17</sup> More detailed inclusion and exclusion criteria are reproduced in [supplementary materials](#). In brief, we included all those patients with available spatially normalized lesion segmentations and information on stroke severity in the acute to subacute phase, in most cases at the time of hospital admission.

All patients or their proxies of the WashU and MRI-GENIE cohorts gave written informed consent in accordance with the Declaration of Helsinki. Given the retrospective character of the MGH-based study, it was performed under a waiver of consent. The study protocols were approved by MGH's Institutional Review Board (Protocol #: 2001P001186, 2003P000836, and 2013P001024) and the Review Boards of individual sites.

### Neuroimaging data and lesion segmentation

WashU cohort: Neuroimaging scans were acquired with a 3T Siemens Tim-Trio scanner at the School of Medicine of the Washington University in St. Louis. Lesions were manually segmented onto structural MRI images obtained 1 to 3 weeks poststroke, non-linearly spatially normalized to Montreal Neurological Institute (MNI)-space and reviewed by two board certified neurologists (MC and AC).

MGH cohort: Neuroimaging scans were obtained on either a Siemens (Munich, Germany) 3T MRI or a General Electric (Fairfield, CT) 1.5T MRI machine, typically within the first 48 h after admission. Lesion segmentations were generated via an in-house deep learning-based algorithm and non-linearly spatially normalized.<sup>23</sup> Manual quality control of lesion segmentations and spatial

<sup>†</sup>While patients were recruited at two institutions, the patient population and scanning procedures were the same. We hence here loosely use the term “single-center.”

normalization occurred to guarantee a high quality of the final lesion segmentations (JR).

MRI-GENIE cohort: Neuroimaging data were recorded on various scanners depending on the recruiting site, within the first few days of hospital admission.<sup>24</sup> DWI-based lesion segmentations were created via validated automated algorithms.<sup>25</sup> DWI data and respective lesions were subsequently non-linearly spatially normalized. High quality of lesion segmentations and spatial normalization was ensured by two experienced raters (AKB and MB).

Further details on scanners and imaging parameters are stated in Tables S1 and S2.

### Computation of structural disconnection

We computed structural disconnection (SDC) maps via the BCB toolkit, in accordance with previously published methods.<sup>8</sup> 7T DWI data from 176 healthy participants from the “Human Connectome Project” were used to identify fiber tracks that passed through each lesion (age  $29.5 \pm 3.6$  years, 72 male participants; HCP7T).<sup>26,27</sup>

### Computation of functional disconnection

FDC, that is, the temporal co-activation of the lesion location with all other whole-brain voxels, was derived using local software and in accordance with previously published methods from our group.<sup>10,28,29</sup> We utilized a publicly available normative connectome dataset of 1000 healthy right-handed subjects (mean age: 21.3 (range: 18–35) years, 43% men, preprocessed in accordance with Fox et al., 2005<sup>30</sup>).<sup>31</sup> We used this connectome to compute resting-state functional connectivity between each patient's stroke lesion location and all other whole-brain voxels, creating a single lesion network map for each patient (including both positive and negative T-values).

### Prediction of stroke severity

The outcome variable of interest was stroke severity in the acute to early subacute phase poststroke<sup>32</sup> as measured by the National Institutes of Health Stroke Scale (NIHSS, 0: no symptoms, 42: death). For each dataset, we selected the stroke severity score that was obtained closest to the time of image acquisition. The different features of neuroimaging data, that is, structural (FLAIR- or DWI-derived) lesion location, as well as SDC and FDC, represented our main predictors of interest. We furthermore evaluated the predictive capacity of (log-scaled) lesion volume, by itself and in combination with each of the lesion representations, resulting in seven different inputs. We decided to include lesion volume as baseline predictor, given its already considerable predictive capacity.<sup>33</sup> In

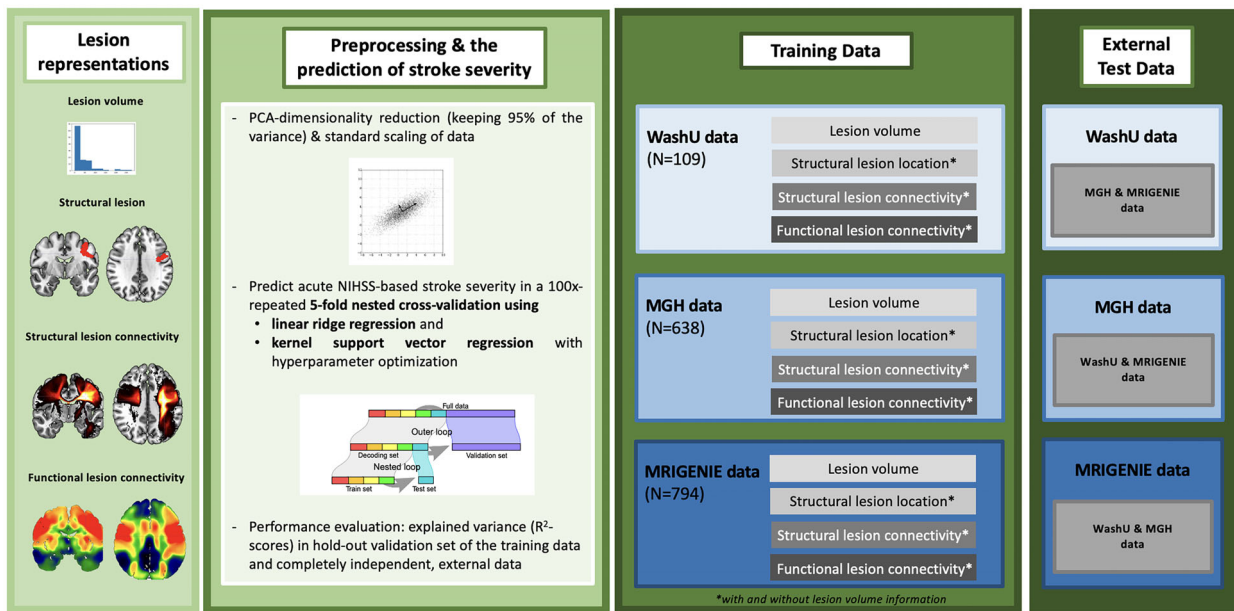
final exploratory analyses, we combined information on lesion volume, SDC and FDC.

The analysis steps described in the following were performed once with each individual cohort as the training dataset. The other two cohorts, representing the respective external data, were kept completely separate. We also randomly, repeatedly downsampled the larger two cohorts ( $N_2$  and  $N_3$ ) to the size of the smaller cohort ( $N_1$ ) and hence obtained two further training datasets (times 100 repetitions). Additionally, we repeated our prediction analyses for the first cohort ( $N_1$ ) considering those patients with ischemic stroke only. Our modeling pipeline began with an unsupervised principal component analysis (PCA)-based dimensionality reduction. This PCA step was performed separately for the three neuroimaging lesion features, that is, lesion location, SDC, and FDC. For each lesion feature, we retained as many PCA components as were necessary to explain 95% of the variance in the original data. Dimensionality-reduced lesion features and the total, log-transformed lesion volume were subsequently standard-scaled.

Next, we employed one linear and one non-linear machine learning algorithm in (100×) repeated nested

five-fold cross-validations with hyperparameter optimization to predict stroke severity. We opted for the two most frequently utilized algorithms in stroke deficit prediction studies:<sup>2,7,13,34</sup>  $l_2$ -regularized ridge regression and a support vector machine (SVM) with a radial basis function (RBF) kernel. For ridge regression, we optimized the regularization parameter  $\alpha$  [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 10000], while for support vector regression, we optimized the regularization parameter  $C$  [0.001, 0.01, 0.1, 1, 10, 100, 1000].

The performance for the prediction of each patient's stroke severity was evaluated using the *coefficient of determination*,  $R^2$ , primarily for all patients included in the external test data that was not used in any of the dimensionality reduction or prediction steps described in the previous paragraphs.<sup>35</sup> We furthermore obtained an estimate within the outer cross-validation-loop of the training cohort to allow for comparisons between cross-validated and external data results. Differences in predictive capacities between the seven inputs were evaluated via two-tailed paired  $t$ -test (level of significance:  $P < 0.05$ , FDR-corrected for multiple comparisons). Figure 1 presents a graphical overview of our methodological approach.



**Figure 1.** Prediction of stroke severity. Lesion information was captured by total lesion volume, voxel-wise structural lesion segmentations, and structural and functional lesion connectivity. For preprocessing, the voxel-wise features were each initially dimensionality reduced via principal component analysis. We kept as many components as were necessary to explain 95% of the variance in the original data. We then trained either linear ridge regression or kernel support vector regression models in a five-fold nested cross-validation to predict individual NIHSS scores. Prediction performance was evaluated as explained variance (coefficient of determination,  $R^2$ ). Training of prediction models was repeated for each of the three cohorts considered in this study: The WashU cohort with 109 patients,<sup>11</sup> an MGH-based cohort comprising 638 patients,<sup>16</sup> and the multicenter cohort of 794 MRI-GENIE patients.<sup>17</sup> For each of these cohorts, we trained prediction models considering each of the lesion information features, in isolation and in combination with lesion volume. External test data were made up of the two cohorts not involved in the training process. The nested cross-validation scheme figure is adapted from Ref. [45].



## Results

In this study, we utilized data from 1541 patients with stroke across three independent cohorts (Table 1). The overlay of stroke lesion locations was qualitatively similar for all three cohorts, primarily affecting subcortical regions in MCA territory (Figure 2). Low-dimensional representations of lesion location, FDC, and SDC are illustrated in Figure 3.

### Identifying the most useful training datasets

Prediction models that were trained using a small single-site dataset ( $N_1 = 109$ ) performed well using cross-validation within the same dataset ( $R^2 = 14.7 \pm 5.5\%$ ). However, prediction performance dropped dramatically when used to predict stroke severity in independent test datasets ( $R^2 = 0.2 \pm 5.1\%$ , Figure 6A). Models trained on different downsampled single-site or multi-site datasets showed a slightly higher, yet still very low prediction performance in absolute terms for external test data ( $N = 109$  patients randomly sampled from  $N_2$  or  $N_3$ ,  $N_2$ :  $R^2 = 7.1 \pm 5.4\%$ ,  $R^2 = 8.8 \pm 6.8\%$ , Figure 6B, C and Tables S3 and S4). Furthermore, prediction estimates for external data remained broadly the same when excluding patients with hemorrhagic stroke and taking into account only those patients with acute ischemic stroke ( $N_{1,\text{ischemic}} = 80$ , Table S5). As such, the low prediction performance in external datasets is likely due to the sample size of the training cohort, not any specific characteristics of the WashU dataset.

**Table 1.** Clinical characteristics of included cohorts.

	WashU cohort (single-site)	MGH cohort (single-site)	MRI-GENIE cohort (multi-site)
Number of participants	109	638	794
Age (years, mean, SD)	53.7 (10.6)	69.2 (14.7)	63.9 (14.8)
Sex (female)	47.7%	49.1%	38.3%
NIHSS (median, IQR)	3 (7)	4 (9)	4 (5)
Lesion volume (mL, median, IQR)	18.2 (46.9)	8.4 (36.8)	3.2 (18.3)

Patients in the WashU cohort were on average younger than the other two cohorts (54 years vs. 69 and 64 years, respectively) and had larger lesions (18.2 mL vs. 8 and 3 mL). This difference in lesion volumes could in part be due to the fact that the WashU cohort combined patients with ischemic and hemorrhagic stroke, while the MGH and MRI-GENIE cohorts exclusively focused on patients with ischemic stroke. Of note, the difference in lesion volume did not go along with comparable differences in median NIHSS scores, as the median NIHSS score was the lowest in the WashU cohort (3 vs. 4 and 4).

Prediction models trained using a larger single-site dataset ( $N_2 = 638$ ) performed very well using within-dataset cross-validation ( $R^2 = 27.5 \pm 4.6\%$ ), but again showed a reduction in explained variance when applied to independent test datasets ( $R^2 = 15.8 \pm 3.6\%$ ).

Finally, prediction models trained using our largest multicenter dataset ( $N_3 = 794$ ) performed moderately well using within-dataset cross-validation ( $R^2 = 20.6 \pm 5.6\%$ ), but were the only models that showed no decrement in prediction accuracy when applied to external test datasets ( $R^2 = 24.1 \pm 3.9\%$ ). In fact, these models predicted more variance in the independent test datasets than they did within the heterogenous training dataset.

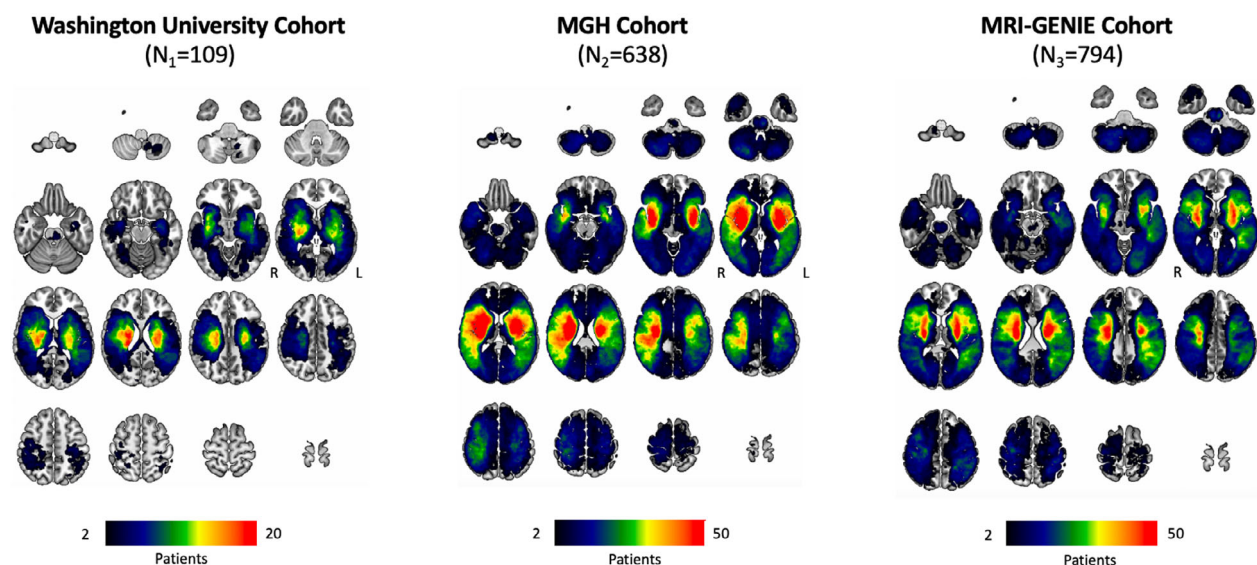
The model-wise differences between cross-validated prediction estimates and those obtained in external data can be found in Tables S6–S8 (two right-most columns). Although there were small differences depending on the model, results were largely independent of which lesion features were included in the model and whether ridge regression or support vector regression was used (Figure 4).

### Identifying the most useful lesion features

To determine which lesion attributes predicted the most variance across independent datasets, we focused on models trained using the larger multicenter data and tested the models on the other two independent datasets.

First, we found that including connectivity information in the model (SDC or FDC) consistently and significantly outperformed lesion volume and location information in the prediction of stroke severity (Figure 5). Similar results were obtained when models were trained on the other datasets (Tables S3–S8).

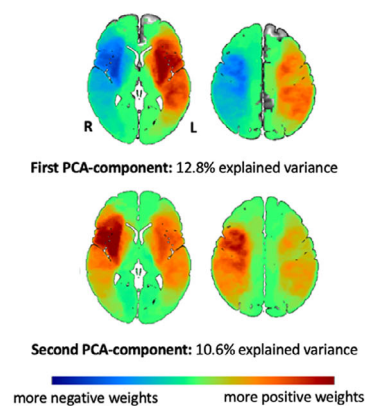
Second, we found that the ideal set of lesion attributes for predictions in test data varied depending on whether we used ridge regression or support vector regression. In case of ridge regression, the best prediction performance resulted from SDC with lesion volume ( $R^2 = 34.30 \pm 0.04\%$ ) although results were similar without lesion volume ( $R^2 = 34.05 \pm 0.02\%$ ). With support vector regression, the best prediction performance was for FDC in combination with lesion volume ( $R^2 = 31.97 \pm 0.07\%$ ). The “best” performing lesion attributes differed depending on the specific training dataset, training algorithm, and cross-validation method (Figures 6 and 7, Tables S3–S8 present a complete set of results for all individual features in all scenarios). In the case of structural disconnection, predictions were usually independent of whether lesion volume was included as a covariate. However, for functional disconnection, predictions often improved considerably when including (log-transformed) lesion volume, suggesting these variables



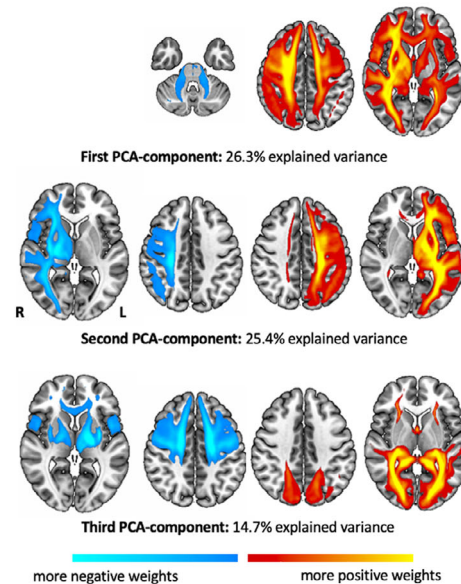
**Figure 2.** Lesion overlays of included cohorts. The maximum lesion overlap was found subcortically in the white matter in proximity to the lateral ventricles, picturing the predominance of middle cerebral artery (MCA) strokes, for all three cohorts. Lesions in MGH and MRI-GENIE additionally covered posterior circulation territories. Stroke lesions affecting the anterior cerebral artery territory were generally rare.

### Low-dimensional lesion representation

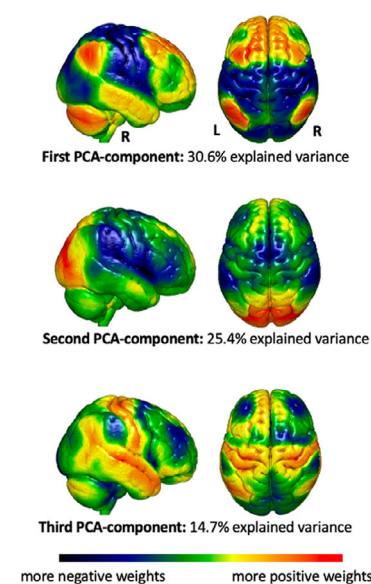
#### (A) Lesion location



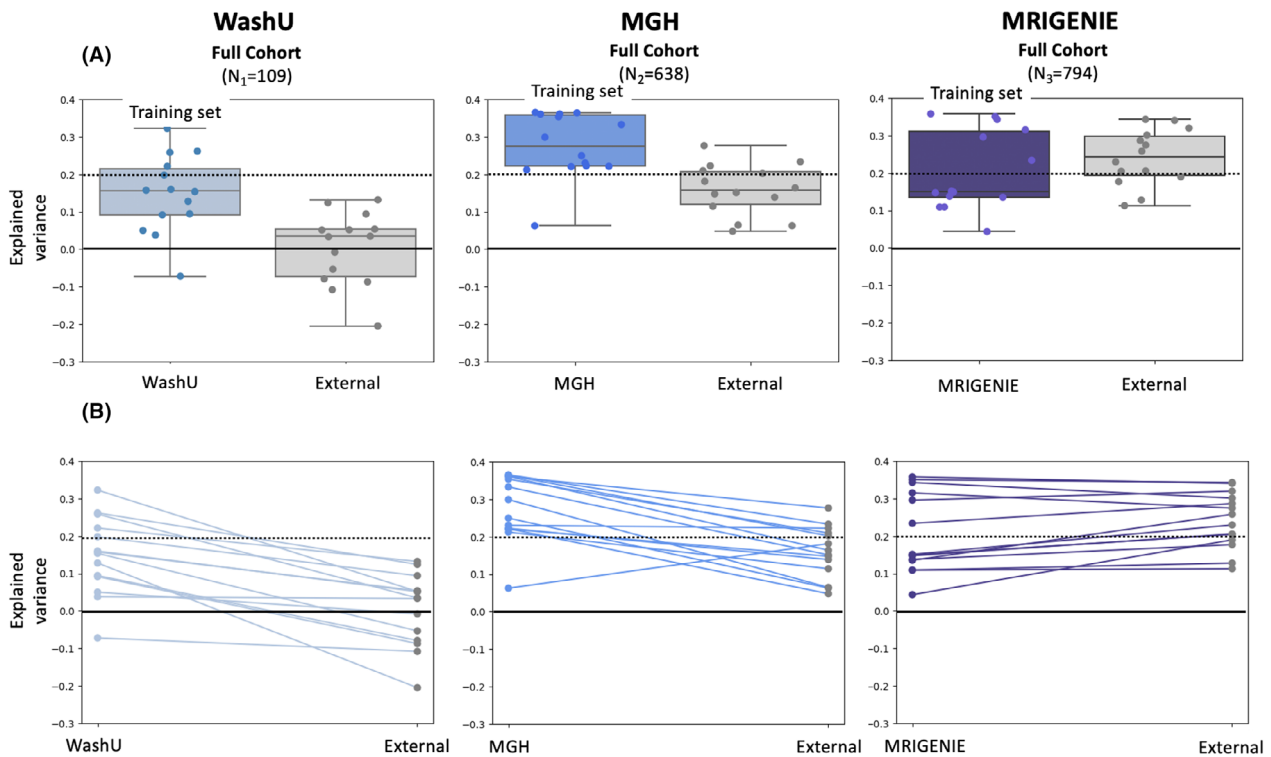
#### (B) Structural disconnection



#### (C) Functional disconnection



**Figure 3.** Low-dimensional lesion features, exemplarily illustrated for the MRI-GENIE cohort. To retain 95% of the variance of the original WashU cohort dataset, we needed 57 components for lesion location, 29 for SDC, and 11 for FDC data. For the MGH cohort, we needed 251 components for lesion location, 65 for SDC, and 13 for FDC. For the MRI-GENIE cohort, we needed 285 components for lesion location, 66 for SDC, and 14 for FDC. Going from the smallest to the largest dataset, the number of components needed to explain 95% of the variance increased substantially for lesion location (from 57 to 285), approximately doubled for SDC data (from 29 to 66 components), and changed very little for FDC data (from 12 to 14). For each of the three different sources of lesion information – lesion location, structural disconnection, and functional disconnection – we here present all the components that individually explained more than 10% of the variance in the original dataset. With respect to functional disconnection (C), the components qualitatively resembled gradients obtained via diffusion embedding;<sup>47</sup> for example, with the first components ranging from transmodal to primary sensorimotor regions.



**Figure 4.** (A) Prediction performances across training cohorts and lesion features. While the average prediction performance decreased from cross-validated estimates to estimates in external data for both small and larger single-center data, there was no such decrease observable for larger, multicenter data. The prediction performance was the highest in case of training on large and multicenter data, making it the most amenable scenario to evaluate the performance of individual lesion features. Each dot represents the average explained variance of one particular lesion feature, such as lesion location, SDC, or FDC. Explained variance was measured as the coefficient of determination. (B) Illustration of changes in explained variance from the cross-validated estimates in the training cohort to the estimates in external data. Each line represents a separate lesion feature, that is, lesion location, SDC, and FDC, each in isolation and combination with lesion volume.

explain independent variance in stroke severity. With some combinations, we did observe statistically significant differences in prediction between SDC and FDC, but the magnitude of the differences was always small and thus of uncertain clinical significance.

In final exploratory analyses, we did not observe any substantial or consistent increases in prediction performance in external data when combining lesion volume, SDC, and FDC (Tables S9–S11).

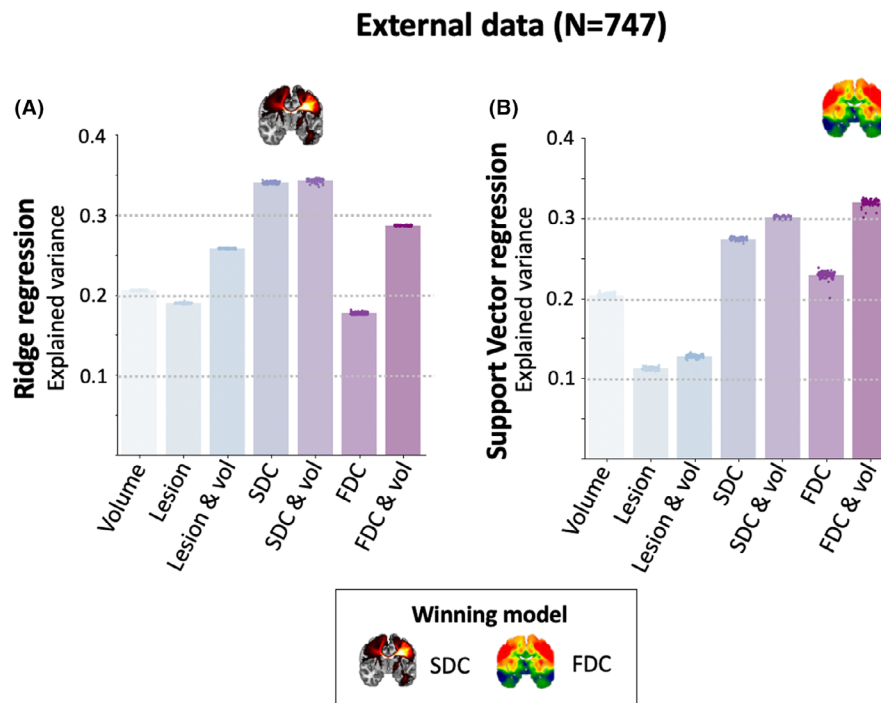
## Discussion

In this present study, we systematically evaluated the performance of various lesion-based imaging features in dependence of the characteristics of the training dataset, model, and evaluation method in predicting stroke severity at the time of imaging acquisition in independent test data. We found that (1) both structural and functional disconnection allow for improved prediction of stroke severity in independent datasets, (2) prediction performance can be improved by training on large multicenter datasets, and (3)

prediction performance in independent stroke data cannot be inferred from cross-validated results within a single dataset. We discuss most relevant results in turn.

We aimed to enhance the comparability to previous stroke imaging prediction studies by aligning our methodological approach to these studies as closely as possible.<sup>2,4,7</sup> However, we introduced one critical modification: We focused on evaluating prediction performance in independent test data. This approach stands in contrast to most prior work that obtained performance estimates using cross-validation within a single dataset from a single site.<sup>2,5–7,12</sup> Of note, some prior studies have studied the prediction of lesion-induced deficits across independent datasets.<sup>28,29,36</sup> However, these studies did not explicitly investigate differences between cross-validation estimates within a single dataset versus estimates across independent datasets.

Previous work that focused on lesion location-based stroke deficit predictions described increases in prediction performance with increasing sample sizes.<sup>16,37,38</sup> In line with this work, we saw a general increase in prediction



**Figure 5.** Prediction results of stroke severity in external data when training relied on the larger, multi-site MRI-GENIE training dataset. For both ridge regression (A) and support vector regression (B), there was a clear benefit from integrating information from indirect connectivity techniques, in case of FDC once combined with lesion volume information. The significantly highest performance was achieved by SDC in case of ridge regression and FDC with lesion volume information for support vector regression (pair-wise  $t$ -tests, level of significance  $P < 0.05$ , FWE-corrected for multiple comparisons). For the ease of interpretation of the bar graphs, the winning model is marked with a small, exemplary brain rendering representing the respective lesion representation.

performance in external data for training on larger samples as compared to smaller ones, also in case of lesion connectivity features, in addition to lesion location-based ones. However, even after training on a large single-site dataset ( $N_2$ ), there was still a drop in performance when tested in external data. Only in the case of training on large multicenter data, did we find no drop in performance between within-dataset cross-validated estimates versus those in independent data. Conceivably, this finding may illustrate that models trained on multicenter data are less prone to deteriorate in their performance due to potential data shift effects that can arise when samples of patients for training and testing models differ in essential characteristics.<sup>39</sup> This interpretation is consistent with results from other fields showing that training on multicenter neuroimaging data has the potential to reduce the biases of machine learning models with respect to the performance across different groups of subjects characterized by age, gender, or race.<sup>40</sup>

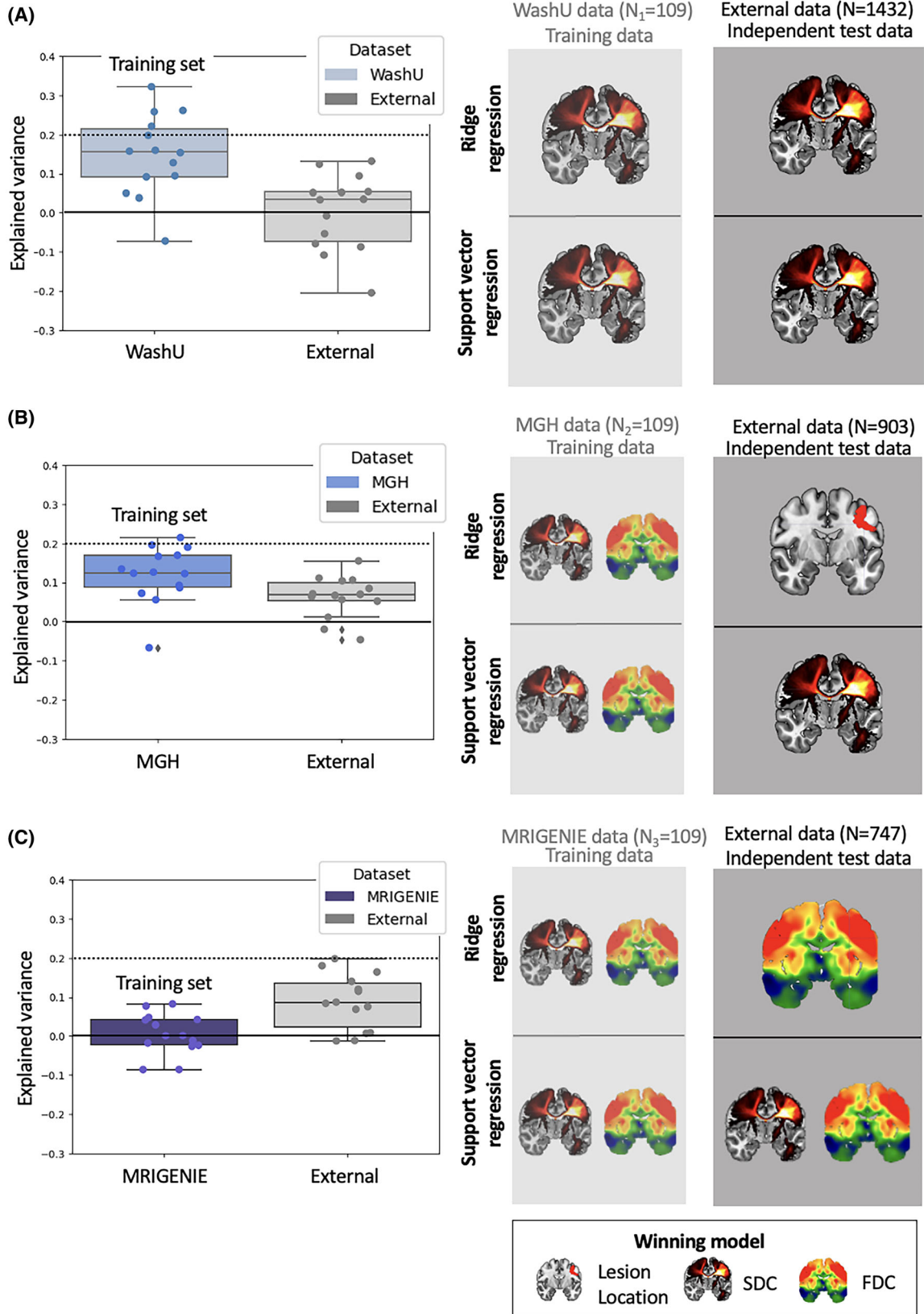
Appreciating these differences in estimates for cross-validation and external data, it occurs to us that it is of prime importance for studies *focused on predicting new data* to explicitly test the generalization performance in

external data. Otherwise, one may be adapting estimates that are too optimistic, which may be particularly relevant for conceivable later real-world applications. Similarly, conclusions about the best method for predicting variance within a dataset may not generalize to predicting variance across independent datasets. For example, we found that SDC predicted more variance within the WashU dataset than FDC, consistent with prior work.<sup>2</sup> However, in other datasets and model scenarios, FDC performed better, especially when tested in independent dataset.

## Limitations and Future Directions

One of the main limitations of our study is focusing on a single global measure of poststroke deficits (NIHSS) that was obtained at the time of imaging acquisition, rather than at a later point in time. This output variable was chosen to maximize the amount of available individual patient data since the NIHSS-based stroke severity score is one of the most frequently obtained scores (e.g., 3-month functional outcomes were not consistently available for our cohort and would have led to drastic

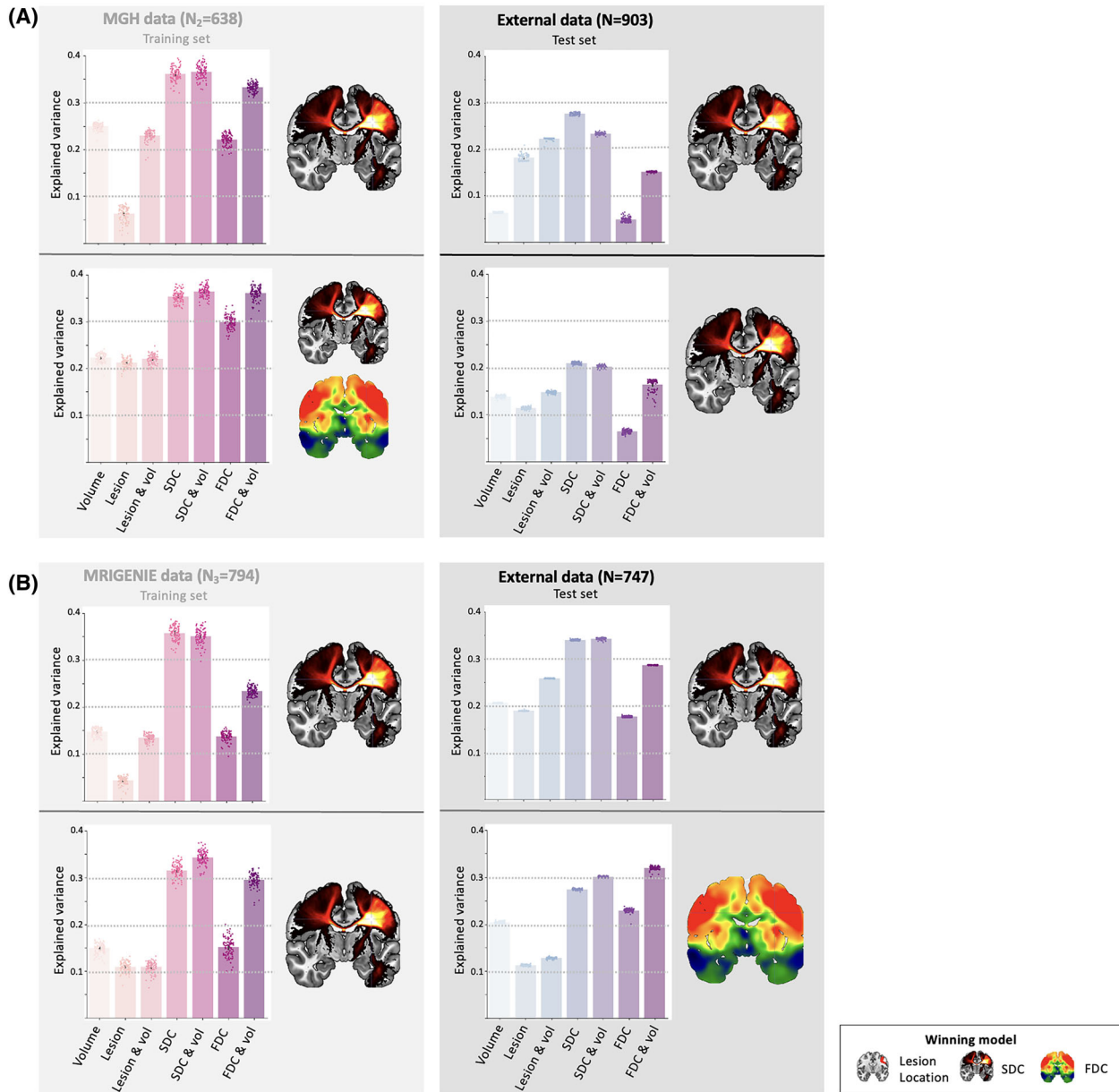




**Figure 6.** Prediction results when training relied on the WashU cohort (A) or downsampled sets of the MGH and MRI-GENIE cohorts (B, C). On the left, prediction performance across lesion features is summarized in box plots: The prediction performance obtained via cross-validation in the training datasets depended on the actual training cohort and was on average higher for the single-center cohorts, that is, WashU and MGH compared to the multicenter MRI-GENIE cohort. These differences could conceivably originate from cohort-specific patient characteristics and inclusion/exclusion criteria (given that the sample size itself was the same in all three scenarios). Of note, the overall prediction performance in external data was generally low across all three samples (on average <~9%). On the right, graphics visualize the winning lesion feature in each individual scenario: While SDC led to the significantly highest prediction performance for all prediction model and cross-validation vs. external data combinations in the WashU cohort, the situation became more complex for the further two cohorts: Here, all three lesion features, that is, lesion location, structural disconnection, and functional disconnection excelled in different scenarios. As is also the case in Figures 5 and 7, the winning representation is visually highlighted by brain renderings. The renderings themselves present examples of the respective lesion representation but cannot be interpreted with respect to voxel-wise importances.

decreases in sample size with sample size being one of the most important factors in prediction scenarios). A patient's stroke severity in the hospital can of course be determined clinically without the necessity to predict it. The value of our study may therefore be seen in the methodological insights offered – such as that large, multicenter data enabled reliably higher prediction performances in external data – rather than the clinical ones. Furthermore, the ability to predict acute stroke deficits based on lesion location or connectivity may conceivably aid prognosis and guide treatment decisions in the future. For example, candidates for invasive thrombolysis or thrombectomy are often identified based on a mismatch between stroke severity on clinical exam and neuroimaging findings.<sup>20</sup> This clinical-radiographic mismatch emerges when a patient presents with more severe stroke symptoms than might be predicted based on the lesion volume. These patients are considered good candidates for intervention as they may have brain tissue that is dysfunctional, resulting in symptoms, but which has not yet become a permanent lesion.<sup>41</sup> Future work may specifically center on refinements of this clinical-radiographic mismatch: Our study suggests that adding information originating from more advanced imaging-based measures, such as functional or structural lesion connectivity, to lesion volume information could potentially enhance the determination of this mismatch. In addition, it will also be important to explicitly test the generalization of our prediction results to outcomes in later phases poststroke, such as modified Rankin Scale Score-based functional outcomes and aim for even higher prediction performances in general. We could explain only ~one third of the variance in stroke severity, which may not suffice yet to be of clinical relevance to support optimal planning and care during recovery phases. Another technical aspect that could be examined in future work is the algorithmic choice for dimensionality reduction. While linear PCA, as used here, is one of the most commonly used strategies,<sup>7</sup> it is conceivable that other approaches, such as non-linear matrix factorization,<sup>42</sup> t-SNE,<sup>43</sup> UMAP,<sup>36</sup> and deep

learning-based techniques,<sup>37</sup> could facilitate a higher prediction performance. In prediction-focused analysis settings, it is of prime importance to mitigate the risk of data leakage.<sup>44</sup> We paid great attention to adhere to general recommendations<sup>35,45</sup> and strictly separated all operations for analyses that used individual training and test datasets. For cross-validated results within a dataset, we fitted principal components based on the entire training dataset, rather than just the inner loop of the nested cross-validation (i.e., excluding the left-out subjects and recomputing the PCA for every permutation). This decision was made to reduce the computational burden of the analyses, maintain consistency with prior work,<sup>2</sup> and because PCA is a completely unsupervised technique that relies on input data only, and thus should not result in any advantage in determining associations with outcome.<sup>15</sup> Another potential statistical limitation of our study is that we compared the prediction results originating from our various lesion representation models via two-tailed *t*-tests, inspired by methodological approaches in comparable prior work.<sup>36</sup> This approach does not make use of the variance across the estimates of the outer loops. Permutation-based tests of group means are an alternative, more rigorous statistical approach that could be used instead of *t*-tests in future studies. Finally, the differences in clinical characteristics between our individual cohorts (e.g., for age, etiology, lesion volume, and the exact time of data acquisition), as well as missing granular information on the time after stroke, that is, the exact days of time poststroke, could be seen as limitations. However, we believe they rather represent strengths, as a model's generalization performance is more convincing if it generates reliable predictions across a wide range of independent cohorts.<sup>46</sup> We also note that low prediction performances in external data were observable for all samples comprising ~100 patients and did not appear to be due to differences in clinical characteristics between the WashU cohort and other cohorts. Future studies could instrumentalize downsampling analyses that stratify for specific patient characteristics, such as age, sex, lesion size,



**Figure 7.** Prediction results when training relied on the MGH cohort (A) or MRI-GENIE cohort (B). The left columns present cross-validated estimates in the training datasets, while the right columns represent estimates for the external test data. The upper row presents ridge regression and the lower row vector regression results.

and lesion type, to inform about each factors' potential effect on prediction accuracies.

Only in case of training prediction models on large, multicenter data did we observe reliably higher prediction performances in external data. In this scenario of large, multicenter data, both structural and functional disconnection were powerful predictors of stroke deficits whose capacities exceeded that of lesion volume and location alone.

## Acknowledgments

We are grateful to our colleagues at the J. Philip Kistler Stroke Research Center, Massachusetts General Hospital and the Center for Brain Circuit Therapeutics, Brigham and Women's Hospital for valuable support and discussions. Furthermore, we are grateful to our research participants without whom this work would not have been possible. We thank Dr. Alex Carter (AC) for the review

of lesion segmentations in the WashU cohort and Dr. Martin Bretzner (MB) for quality control of lesion segmentations in the MRI-GENIE cohort.

## Author Contributions

Conception and design of the study: AKB, NSR, and MDF; acquisition and analysis of data: AKB, ALC, WD, MAF, AH, CL, FLWVJS, AB, AKG, LCO, RWR, MDS, CJ, AGL, JM, OW, SZ, JZR, EYK, and MC; drafting a significant portion of the manuscript or figures: AKB, NSR, and MDF.

## Conflict of Interest

R.W.R. has served on a DSMB for a trial sponsored by Rapid Medical and has served as site PI for studies sponsored by Penumbra and Microvention. N.S.R. has received compensation as scientific advisory consultant from Omnix, Sanofi Genzyme, and AbbVie Inc. Further authors do not have anything to disclose.

## Data Availability Statement

WashU cohort: Neuroimaging and neuropsychological data are publicly available at <https://cnda.wustl.edu/app/template/Login.vm>. MGH and MRI-GENIE cohorts: The authors agree to make the data available to any researcher for the express purposes of reproducing the here presented results and with the explicit permission for data sharing by individual sites' institutional review boards. Data for the structural connectome are available online: <http://www.humanconnectome.org/study/hcp-young-adult/>. While we cannot share the employed functional connectome due to data privacy regulations, a comparable and equivalently processed functional connectome can be found at: [10.7910/DVN/ILXIKS](https://doi.org/10.7910/DVN/ILXIKS). Prediction analyses were implemented in Python 3.9.

## References

- Carrera E, Tononi G. Diaschisis: past, present, future. *Brain*. 2014;137(9):2408-2422. doi:[10.1093/brain/awu101](https://doi.org/10.1093/brain/awu101)
- Salvalaggio A, De Filippo De Grazia M, Zorzi M, Thiebaut de Schotten M, Corbetta M. Post-stroke deficit prediction from lesion and indirect structural and functional disconnection. *Brain*. 2020;143(7):2173-2188. doi:[10.1093/brain/awaa156](https://doi.org/10.1093/brain/awaa156)
- Cohen AL, Ferguson MA, Fox MD. Lesion network mapping predicts post-stroke behavioural deficits and improves localization. *Brain*. 2021;144(4):e35.
- Salvalaggio A, De Filippo De Grazia M, Pini L, Thiebaut de Schotten M, Zorzi M, Corbetta M. Reply: lesion network mapping predicts post-stroke behavioural deficits and improves localization. *Brain*. 2021;144(4):e36.
- Bowren M Jr, Bruss J, Manzel K, et al. Post-stroke outcomes predicted from multivariate lesion-behaviour and lesion network mapping. *Brain*. 2022;145(4):1338-1353. doi:[10.1093/brain/awac010](https://doi.org/10.1093/brain/awac010)
- Smith DV, Clithero JA, Rorden C, Karnath HO. Decoding the anatomical network of spatial attention. *Proc Natl Acad Sci USA*. 2013;110(4):1518-1523.
- 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. doi:[10.1073/pnas.1521083113](https://doi.org/10.1073/pnas.1521083113)
- Foulon C, Cerliani L, Kinkingnehun S, et al. Advanced lesion symptom mapping analyses and implementation as BCBtoolkit. *GigaScience*. 2018;7(3):1-17. doi:[10.1093/gigascience/giy004](https://doi.org/10.1093/gigascience/giy004)
- Boes AD, Prasad S, Liu H, et al. Network localization of neurological symptoms from focal brain lesions. *Brain*. 2015;138(10):3061-3075. doi:[10.1093/brain/awv228](https://doi.org/10.1093/brain/awv228)
- Fox MD. Mapping symptoms to brain networks with the human connectome. *N Engl J Med*. 2018;379(23):2237-2245. doi:[10.1056/NEJMra1706158](https://doi.org/10.1056/NEJMra1706158)
- Corbetta M, Ramsey L, Callejas A, et al. Common behavioral clusters and subcortical anatomy in stroke. *Neuron*. 2015;85(5):927-941. doi:[10.1016/j.neuron.2015.02.027](https://doi.org/10.1016/j.neuron.2015.02.027)
- Jimenez-Marin A, De Bruyn N, Gooijers J, et al. Multimodal and multidomain lesion network mapping enhances prediction of sensorimotor behavior in stroke patients. *Sci Rep*. 2022;12(1):22400. doi:[10.1038/s41598-022-26945-x](https://doi.org/10.1038/s41598-022-26945-x)
- Bonkhoff AK, Grefkes C. Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. *Brain*. 2022;145(2): 457-475.
- Shmueli G. To explain or to predict? *Stat Sci*. 2010;25(3):289-310. doi:[10.1214/10-STS330](https://doi.org/10.1214/10-STS330)
- James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol 112. Springer; 2013.
- Bourached A, Bonkhoff AK, Schirmer MD, et al. Scaling behaviors of deep learning and linear algorithms for the prediction of stroke severity. *Brain Commun*. 2024;6:fcae007.
- Giese AK, Schirmer MD, Donahue KL, et al. Design and rationale for examining neuroimaging genetics in ischemic stroke: the MRI-GENIE study. *Neurol Genet*. 2017;3(5): e180.
- 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. doi:[10.1016/j.nicl.2019.101884](https://doi.org/10.1016/j.nicl.2019.101884)



19. Ryan SL, Liu X, McKenna V, et al. Associations between early in-hospital medications and the development of delirium in patients with stroke. *J Stroke Cerebrovasc Dis.* 2023;32(9):107249. doi:10.1016/j.jstrokecerebrovasdis.2023.107249
20. 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. doi:10.1056/NEJMoa1706442
21. Wouters A, Nysten C, Thijs V, Lemmens R. Prediction of outcome in patients with acute ischemic stroke based on initial severity and improvement in the first 24 h. *Front Neurol.* 2018;9:308.
22. Kwakkel G, Veerbeek JM, van Wegen EE, Nijland R, Harmeling-van der Wel BC, Dippel DW. Predictive value of the NIHSS for ADL outcome after ischemic hemispheric stroke: does timing of early assessment matter? *J Neurol Sci.* 2010;294(1-2):57-61.
23. Winzeck S, Mocking SJ, Bezerra R, et al. Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. *Am J Neuroradiol.* 2019;40(6):938-945.
24. Drake M, Frid P, Hansen BM, et al. Diffusion-weighted imaging, MR angiography, and baseline data in a systematic multicenter analysis of 3,301 MRI scans of ischemic stroke patients—Neuroradiological review within the MRI-GENIE study. *Front Neurol.* 2020;11:577.
25. Wu O, Winzeck S, Giese AK, et al. Big data approaches to phenotyping acute ischemic stroke using automated lesion segmentation of multi-center magnetic resonance imaging data. *Stroke.* 2019;50(7):1734-1741.
26. Vu AT, Auerbach E, Lenglet C, et al. High resolution whole brain diffusion imaging at 7 T for the Human Connectome Project. *NeuroImage.* 2015;122:318-331.
27. De Schotten MT, Bizzi A, Dell'Acqua F, et al. Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography. *NeuroImage.* 2011;54(1):49-59.
28. Ferguson MA, Lim C, Cooke D, et al. A human memory circuit derived from brain lesions causing amnesia. *Nat Commun.* 2019;10(1):3497. doi:10.1038/s41467-019-11353-z
29. Siddiqi SH, Schaper FLWVJ, Horn A, et al. Brain stimulation and brain lesions converge on common causal circuits in neuropsychiatric disease. *Nat Hum Behav.* 2021;5:1707-1716. doi:10.1038/s41562-021-01161-1
30. Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci USA.* 2005;102(27):9673-9678.
31. Yeo BT, Krienen FM, Sepulcre J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol.* 2011;106(3):1125-1165.
32. 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. *Int J Stroke.* 2017;12(5):444-450. doi:10.1177/1747493017711816
33. Thijs VN, Lansberg MG, Beaulieu C, Marks MP, Moseley ME, Albers GW. Is early ischemic lesion volume on diffusion-weighted imaging an independent predictor of stroke outcome? A multivariable analysis. *Stroke.* 2000;31(11):2597-2602.
34. Rehme AK, Volz LJ, Feis DL, et al. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. *Cereb Cortex.* 2014;25(9):3046-3056.
35. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry.* 2020;77(5):534-540.
36. Talozzi L, Forkel SJ, Pacella V, et al. Latent disconnectome prediction of long-term cognitive-behavioural symptoms in stroke. *Brain.* 2023;146(5):1963-1978.
37. Chauhan S, Vig L, De Filippo De Grazia M, Corbetta M, Ahmad S, Zorzi M. A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. *Front Neuroinform.* 2019;13:53.
38. Sperber C, Wiesen D, Karnath HO. An empirical evaluation of multivariate lesion behaviour mapping using support vector regression. *Hum Brain Mapp.* 2019;40(5):1381-1390. doi:10.1002/hbm.24476
39. 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. doi:10.1056/NEJMc2104626
40. Wang R, Chaudhari P, Davatzikos C. Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies. *Proc Natl Acad Sci USA.* 2023;120(6):e2211613120. doi:10.1073/pnas.2211613120
41. Butcher KS, Parsons M, MacGregor L, et al. Refining the perfusion-diffusion mismatch hypothesis. *Stroke.* 2005;36(6):1153-1159.
42. Bonkhoff AK, Lim JS, Bae HJ, et al. Generative lesion pattern decomposition of cognitive impairment after stroke. *Brain Commun.* 2021;3(2):fcab110. doi:10.1093/braincomms/fcab110
43. Bonkhoff AK, Xu T, Nelson A, et al. Reclassifying stroke lesion anatomy. *Cortex.* 2021;145:1-12.
44. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data.* 2012;6(4):1-21.
45. 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.

46. Cheng B, Chen J, Königsberg A, et al. Mapping the deficit dimension structure of the National Institutes of Health Stroke Scale. *EBioMedicine*. 2023;87:104425. doi:10.1016/j.ebiom.2022.104425
47. Margulies DS, Ghosh SS, Goulas A, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc Natl Acad Sci USA*. 2016;113(44):12574-12579. doi:10.1073/pnas.1608282113

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Table S1.** Inclusion and exclusion criteria of included cohorts.

**Table S2.** Information on neuroimaging parameters.

**Table S3.** Prediction results as cross-validated estimate and in external test data when training occurred on downsampled MGH cohort ( $N_{\text{downsampled}} = 109$ , mean explained variance, 95% confidence intervals).

**Table S4.** Prediction results as cross-validated estimate and in external test data when training occurred on downsampled MRI-GENIE cohort ( $N_{\text{downsampled}} = 109$ , mean explained variance, 95% confidence intervals).

**Table S5.** Prediction results as cross-validated estimate and in external test data when training occurred on only the patients with ischemic stroke of the WashU cohort

( $N_1 = 80$ , mean explained variance, 95% confidence intervals).

**Table S6.** Prediction results as cross-validated estimate and in external test data when training occurred on WashU cohort ( $N_1 = 109$ , mean explained variance, 95% confidence intervals).

**Table S7.** Prediction results as cross-validated estimate and in external test data when training occurred on MGH cohort ( $N_2 = 638$ , mean explained variance, 95% confidence intervals).

**Table S8.** Prediction results as cross-validated estimate and in external test data when training occurred on MRI-GENIE cohort ( $N_3 = 794$ , mean explained variance, 95% confidence intervals).

**Table S9.** Prediction results of exploratory analyses considering information from lesion volume, SDC and FDC at once: Training occurred based on WashU data ( $N_1 = 109$ , mean explained variance, 95% confidence intervals).

**Table S10.** Prediction results of exploratory analyses considering information from lesion volume, SDC and FDC at once: Training occurred based on MGH data ( $N_2 = 638$ , mean explained variance, 95% confidence intervals).

**Table S11.** Prediction results of exploratory analyses considering information from lesion volume, SDC and FDC at once: Training occurred based on MGH data ( $N_2 = 638$ , mean explained variance, 95% confidence intervals).