Mapping the deficit dimension structure of the National Institutes of Health Stroke Scale

Bastian Cheng,[a,e,*] Ji Chen,[b,c,e,**] Alina Königsberg,[a] Carola Mayer,[a] Leander Rimmele,[a] Kaustubh R. Patil,[c,d] Christian Gerloff,[a] Götz Thomalla,[a,e] and Simon B. Eickhoff[c,d,e]

[a]Klinik und Poliklinik für Neurologie, Kopf- und Neurozentrum, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[b]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China
[c]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany
[d]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

eBioMedicine
2023;87: 104425
Published Online
https://doi.org/10.1016/j.ebiom.2022.104425

Summary

Background The National Institutes of Health Stroke Scale (NIHSS) is the most frequently applied clinical rating scale for standardized assessment of neurological deficits in acute stroke in both clinical and research settings. Notwithstanding this prominent role, important questions regarding its validity remain insufficiently addressed: Investigations of the underlying dimensional structure of the NIHSS yielded inconsistent results that are largely not generalizable across studies. Neurobiological validations by linking measured deficit dimensions to brain anatomy and function are missing.

Methods We, therefore, employ advanced machine learning to identify an optimal representation of the dimensional structure of the NIHSS across two independent and heterogeneous stroke datasets (N = 503 and N = 690). Associated lesion locations are identified by multivariate lesion-deficit mapping (LDM) and their functional relevance is profiled based on *a-priori* task activation meta-data analysis, to provide an independent link to the behavioural level.

Findings A five-factor structure of the NIHSS was identified as the most robust and generalizable representation of stroke deficit dimensions across study populations, settings, and clinical phenotypes. Specifically, the identified dimensions comprised NIHSS items for (F1) left motor deficits, (F2) right motor deficits, (F3) dysarthria and facial palsy, (F4) language, and (F5) deficits in spatial attention and gaze. LDM linked four of these factors to differentially localized, eloquent neuroanatomical areas. Functional characterization of LDM results aligned with detected deficit dimensions, revealing associations with motor functions, language processing, and various functions in the perception domain.

Interpretation By cross-validating machine learning in heterogeneous multi-site stroke cohorts, we report evidence on the validity of the NIHSS: We identified an overarching structure of the NISHS containing a five-dimensional representation of stroke deficits. We provide an anatomical map of the NIHSS that is of value for future applications of individualized stroke treatment and rehabilitation.

Funding This research was supported by the National Key R&D Program of China (Grant No. 2021YFC2502200), the National Human Brain Project of China (Grant No. 2022ZD0214000)", the German Research Foundation (Deutsche Forschungsgemeinschaft), Project 178316478 (A1, C1, C2), and Project 454012190 of the SPP 2041, the Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain" and Helmholtz Imaging Platform grant NimRLS (ZT-I-PF-4-010).

Keywords: Stroke; NIHSS; Dimensionality reduction; Lesion-deficit-mapping

*Corresponding author.
**Corresponding author.
E-mail addresses: b.cheng@uke.de (B. Cheng), jichen.allen@hotmail.com (J. Chen).
[e]These authors contributed equally to this work.

Research in context

Evidence before this study

Despite the widespread use of the National Institutes of Health Stroke Scale, several psychometric properties of the scale remain understudied, specifically regarding its internal structure and associated neuroanatomy. We searched PubMed and Web of Science from the inception of the database to January 1st 2021 for studies published in English using the terms "NIHSS" AND "factor" OR "structure" OR "lesion symptom mapping" OR "lesion deficit mapping". The identified studies yielded inconsistent results regarding the internal structure of the NIHSS with various compositions of principal deficit dimensions. There were no studies linking the NIHSS to neuroanatomy and brain function via lesion-deficit inference methods.

Added value of this study

By using an unsupervised machine learning approach and systematic in- and cross-sample evaluations of stability and generalizability on large, multicenter, and heterogeneous datasets, we report a distinct structure with five dimensions best representing the clinical construct in stroke patients as assessed by NIHSS. Lesion deficit mapping revealed meaningful neuroanatomical locations. We validate the overarching structure of post-stroke deficits captured by the NIHSS and provide an anatomical map of the NIHSS.

Implications of all the available evidence

Our study improves the understanding of the internal dimensional structure of the NIHSS and its neural and functional underpinnings that will enhance prognostication of lesion locations from clinical syndromes, clinical recovery, and outcomes in individualized treatment of stroke patients. By mapping brain areas most critical for clinical impairment, our results contribute to testing individualized selection algorithms for recanalization therapy in acute stroke. We promote an approach both integrating information from eloquent brain regions at risk for infarction and known outcome predictors such as lesion volumes. Lastly, our study also opens the window for targeted use of low-dimensional versions of the NIHSS in research, clinical trials, or clinical decision-making.

Introduction

The National Institutes of Health Stroke Scale (NIHSS)[1] is one of the most frequently applied clinical rating scales in neurological practice. It is one of the rare tools for quantifying clinical deficits applied in highly diverse settings ranging from the decision-making of individualized treatment in acute stroke, measurement and adjustment of outcomes in clinical trials, to guiding regulatory agencies of medical institutions.

Despite the widespread use of NIHSS, several psychometric properties of the scale remain understudied.: Specifically, regarding the validity of the score, two research questions need to be addressed further: Does the NIHSS measure a limited number of underlying constructs (i. e. clinical phenotypes) as intended? Do the underlying clinical phenotypes reflect neurobiological properties that can be linked to brain anatomy and brain functions? A comprehensive understanding of the internal dimensional structure of the NIHSS and its neural and functional underpinnings will ultimately enhance prognostication of lesion locations from clinical syndromes, clinical recovery, and outcomes in individualized treatment of stroke patients.

Unfortunately, so far only a few studies tested the construct validity of the NIHSS in terms of its internal structure and associated neurobiological functions: Based on data from the NINDS tPA Stroke trial and using an abbreviated, 12-item-version, the NIHSS has been proposed to comprise four symptom dimensions.[2] The four dimensions were described as capturing left and right hemispheric, "motor" and "cortical" clinical deficits (supplemental Table S2).[2] However, different compositions were reported,[3,4] and the assignment of individual NIHSS items to main dimensions varied across studies with inconsistent results.[2,5] In addition, the "traditional" 4-factor structure of the NIHSS was mainly detected in data from severely affected patients from clinical trials and a limited range of stroke etiologies.[5] Patients with less severe clinical deficits (i. e. due to lacunar stroke) that represent a relevant proportion of patients in clinical practice were most likely underrepresented in these datasets.[6]

Regarding the mapping of deficit dimensions captured by the NIHSS to brain anatomy, there are currently no systematic investigations, mostly due to the lack of datasets combining both precise, MRI-based stroke lesion delineations and fine-grained clinical characterizations from individual NIHSS items. An anatomical "NIHSS map" would, however, be of high value as it could be applied to predict principal dimensions of clinical deficits in stroke patients for tailored rehabilitation approaches. Furthermore, mapping brain areas most critical for clinical impairment contributes to the aim of improving selection criteria for individualized stroke treatment: Involvement of specific eloquent brain regions is of high relevance for determining the functional outcome.[7–9] However, current practices for patient selection in stroke treatment, specifically in patients with late- or unknown time after symptom onset is based on volumetric measures of stroke lesions or salvageable brain tissue alone. It is yet unclear if a more individualized treatment selection integrating known predictors of functional outcome and anatomical information on eloquent brain regions at

risk for infarction is beneficial as suggested by recent work in a subset of patients with large ischemic cores.[10]

In the present study, we, therefore, aim to address two key questions: First, to arrive at a generalizable, optimal dimension structure of the NIHSS, and second, to detect its structural and functional underpinnings in eloquent brain areas using lesion-deficit mapping and functional decoding.

For the first aim, we plan to identify a robust, cross-validated, and interpretable factor structure of clinical stroke deficits based on single-item scores of the full-range NIHSS by implementing data-driven machine-learning with a comprehensive evaluation procedure in two large, independent datasets (total N = 1193) collected from 64 independent stroke centres. From a technical point of view, we improve previous factorial approaches, since several downsides in methodology, in particular principal component analysis (PCA) and exploratory factor analysis (EFA), have been noted. For example, the negative loadings presented on PCA and EFA factors are not intuitively interpretable, and the non-sparse factor solution renders item-to-factor assignments oftentimes non-intuitive. To ameliorate these limitations, a projective and orthonormal extension of the original non-negative matrix factorization (NMF), i.e., the orthonormal projective (OP)NMF[11,12] with a carefully designed evaluation procedure will be implemented.[13] This method has been shown to allow a robust factorization of psychometric data, and the resulting factors were found to be more reliably associated with brain network connectivity patterns than the original subscales.[11,12,14]

For the second aim, we map each of the identified NIHSS factors to anatomical brain lesion areas using multivariate regression analysis with cross-validation. Based on the concept of lesion deficit mapping (LDM), the resulting factors, i.e., deficit dimensions that represent contemporaneous neurological symptoms, causatively originate from strategically localized neuroanatomical lesion sites that are relevant to corresponding brain functions. To detect such underlying localized brain lesions, an imaging-based approach, the multivariate LDM, will be applied.[15,16] While the results from LDM inform anatomical locations for specific symptom dimensions, their functional relevance to the behaviour level requires additional tools, as the characterization through BrainMap-based functional profiling informed by previous task-evoked functional MRI (fMRI) experiments (http://brainmap.org/). This approach provides an independent link back to the phenotypic level as assessed by the NIHSS.

Methods

Sample

We analyzed two large stroke patient datasets providing single-item NIHSS scores at the time point of hospital admission. The first dataset comprised a population of patients with acute stroke lesions demonstrated by MRI and eligibility for treatment with intravenous alteplase (WAKE-UP trial).[17] In comparison, the second dataset was chosen to comprise a more heterogeneous group of stroke patients that were selected (1) from everyday hospital admissions without predefined imaging or clinical inclusion criteria regardless of acute stroke treatment (EPOS study) and (2) with predefined imaging and clinical inclusion criteria (I-Know study).

1) Patients randomized in the WAKE-UP trial, an international, multicenter, placebo-controlled trial of MRI-based intravenous thrombolysis in patients with unknown onset stroke based on MRI selection criteria (referred to as the "first dataset").[17]

2) For evaluating the generalizability of the resulting factorizations to new stroke populations, we pooled clinical data from two prospective acute stroke studies (referred to as the "second dataset"). Specifically, we selected patients from EPOS ("Outcome evaluation by patient-reported outcome measures in stroke clinical practice"), a prospective, single-centre, observational study for outcome evaluation by patient-reported outcome measures in stroke[18] and I-Know, a multicenter observational study aiming at outcome prediction based on clinical and imaging variables.[7]

In both datasets, studies included patients of all sex and gender. Sex was determined by self-report of study participants, sex and gender were not applied as exclusion criteria in any of the studies from which data was selected. From both datasets, we only included patients who have been able to carry out usual activities in their daily life without support before stroke. Detailed inclusion criteria for all studies can be found in the supplemental Table S3. A study flow diagram is illustrated in supplemental Fig. S1. In total, N = 503 patients were included for the first and N = 690 patients for the second dataset. As intended, several demographic and clinical characteristics differed significantly between datasets, providing an optimal setting for testing generalization performance across independent samples: There was a higher proportion of male patients in the first dataset (N = 325, 65% vs. N = 400, 58%; $p = 0.021$). Patients in the first dataset were younger (mean age 65.2 years, SD 11.6 years vs. mean age 71.5 years, SD 13 years; $p < 0.001$) and more severely affected by stroke (median NIHSS 6, IQR: 4–9 vs. NIHSS 4, IQR: 1–9; $p < 0.001$). In the second dataset, the severity of clinical deficits was more heterogeneous as compared to the first dataset, with a higher proportion of patients with either low or high NIHSS values (see supplemental Fig. S2). Lesion hemisphere side distributions were similar in both datasets (Chi-squared test; $p = 0.187$). Both patients with supra- and infratentorial stroke lesions were included in our study.

Stroke lesion segmentations from MRI data measured at the time point of hospital admission were only available in the first dataset (N = 503). Stroke lesions were segmented based on DWI data as described

previously based on manual segmentations following a semi-automated procedure using an apparent diffusion coefficient (ADC) threshold of 620×10^{-6} mm^2/s.[19] Lesion masks were transformed to Montreal Neurological Institute (MNI) space by linear and non-linear registrations based on FLAIR data.[20] All lesion masks were checked for correct segmentation and registration into MNI-space by two raters experienced in stroke MR imaging (A. K., B. C.). Of note, the MRI data used for LDM was collected at the same time point of NIHSS scoring used in our study.

Ethics

Written informed consent was provided according to national and local regulations by patients or their legal representatives. Approval of the local ethics committee (Ethik-Kommission der Ärztekammer Hamburg) has been obtained (PV54565, and PVN3857).

Factorization of NIHSS using OPNMF

Identification of the internal dimensional structure of the NIHSS (i.e., factorization) was conducted as reported previously.[11–13] OPNMF has several advantages over traditional factorial analyses that are particularly well suited for our aims. First, NMF produces intuitive, non-negative scores with higher values representing more severe clinical deficits. Second, owing to the projective constraint in OPNMF, the learned factors can be readily applied to the NIHSS data of new samples. Third, the enforced orthonormality constraint not only promotes a sparse, parts-based representation of the data improving interpretability but is computationally less expensive, facilitating the implementation of various cross-validation and out-of-sample generalization evaluations for deriving the optimal factor models.

The core optimization process for OPNMF is to minimize the reconstruction error measured by Frobenius norm between the input data matrix V and its estimate by iterative multiplicative updates of the basis matrix W:

$$\min \left\| V - WW^T V \right\|_F$$

$$s.t.\, W \geq 0;\; WW^T = I$$

Specifically, we applied OPNMF to decompose the NIHSS data into two non-negative matrices: 1) a basis matrix W (i.e., the dictionary) with factors (i.e., deficit dimensions) as columns and 2) a loading matrix H (formulated as W^TV due to the projective constraint) with scores representing deficit level of individual patients as expressed along these dimensions. Importantly, a non-negative singular value decomposition was employed to initialize W,[21] which enjoys several advantages over random initialization, including reduced residual error, improved convergence, and deterministic decompositions.

To derive the optimal (i.e., most robust and generalizable) factor structure of the NIHSS, we followed a comprehensive evaluation procedure.[13] In summary, we first factorized the 15 NIHSS items within the first dataset and evaluated the resulting factorizations using 5000 times repeated split-half analysis. In each split-half analysis, two indices (adjusted Rand-index [aRI] and variation of information [VI]) was employed to assess the robustness of item-to-factor assignment, based on its highest median coefficient, along with the concordance index (CI) between the dictionaries derived from the two split samples.

The idea of employing aRI and VI is based on the hard-assignment of items to specific factors (i.e., to use NMF as a natural clustering). Higher values of aRI (up to 1) indicate better correspondence of item-pair placement in the factors derived from the split samples, and lower VI indicates a higher similarity of factor-label assignment of items (i.e., more information shared between the factorizations) between the two split-samples. Since an item can be influenced by multiple dimensions and thus it may have small contributions to other factors apart from the one it is assigned to, we also employed the concordance index, which reflects the concordance of the cosine similarity for each pair of the NIHSS items between the factorizations of split-samples, on all entries of the W matrices to account for the items with multiple factor-memberships. Generalizability was assessed by out-of-sample reconstruction error calculated as how much reconstruction error is increased for one split-half sample due to the use of a dictionary from the other half data.

Two additional data perturbation strategies, bootstrapping and 10-fold cross-validation with testing in hold-out data, were then employed to corroborate the split-half findings. The same evaluation procedure was applied to the second dataset (N = 690).

Following in-sample tests, we performed between-sample testing for the optimal number of factors in the generalization from the first to the second dataset, as well as the robustness of item-to-factor assignment across the factorizations from the bootstraps of these two samples (repeated 5000 times). Results were derived from the evaluation runs to derive the optimal factor structure (referred to as factor models in the following) using both in- and between-sample testing. In addition to the main analysis, we planned for post-hoc analysis in case of evaluations of stability and generalizability pointing to *different* optimal solutions. This may happen in the presence of NIHSS items that do not fit any of the factors formed by the other (more closely expressed) items within a scale. In this case, the approach was to repeat the evaluation process after excluding the non-fitting NIHSS items indicated by poor internal consistency and within-factor inter-item correlations as described below. As a supplemental, explorative approach, we performed a conventional Principal Component Analysis (PCA)

in-sample in both datasets, methodological details and results are shown in the supplement (supplemental Figs. S9 and S10).

Internal consistency and relationship among factors

We assessed the internal consistency of the OPNMF derived optimal factor structure using Cronbach's alpha, where higher values indicate a more closely related set of items as a factor. We investigated the relationship between factors by calculating correlations between individual items, both before and after controlling for the total NIHSS score. Finally, we tested the effects of total NIHSS, age, sex, and stroke lesion volume (as independent variables) on the joint factor loadings using MANOVA followed by individual 4-way analyses of variance (ANOVAs) and bootstrap analysis (10,000 repetitions). Statistical analysis was conducted using R (Version 4.0.2) and MATLAB (MATLAB 2019b, The MathWorks, Inc., Natick, Massachusetts, United States).

Lesion-deficit mapping

Multivariate LDM was conducted in the imaging data of the first dataset based on the factor loadings from the optimally generalizable and most stable NIHSS model. We applied support vector regression (SVR) which operates on continuous variables as a regression extension of the support vector machine in use for classification.[22] LDM was conducted using a dedicated package (https://github.com/atdemarco/svrlsmgui) applying functionalities of the Statistics and Machine Learning Toolbox within MATLAB.[23] Only voxels with a minimum of five overlapping lesions were included for statistical testing. Before SVR, lesion volumes were regressed out from the factor loadings and the lesion data on a voxel-wise basis and hence their effects on LDM were controlled. We applied an epsilon-SVR using a non-linear, radial basis function (Gaussian) kernel analogous to the original publication in comparable stroke imaging datasets.[23] Specifically, the adjusted voxel-wise lesion scores for each subject were combined into a matrix and used as the features to train the SVR model with loadings on each factor individually serving as the target variable to be predicted. Three hyperparameters, *epsilon, cost,* and *sigma*, which control the behaviour of SVR, were tuned through a 20-fold cross-validated, Bayesian optimization with 200 iterations as implemented in MATLAB (bayesopt) separately for each factor. The ensuing optimal SVR hyperparameters related to our dataset with minimal cross-validation error were used to construct the final SVR model, yielding regression weights that reflect the true deficit-lesion associations. Permutation tests through shuffling the loadings randomly between subjects 10.000 times were implemented to correct for chance-level associations to derive significant voxels ($p < 0.005$) Co-localized voxels that survived at this voxel-level threshold were further grouped into clusters and an additional cluster-level family-wise error correction (FWE) of $p < 0.05$ was applied to ensure an adequate control for false positives as in prior studies.[23]

The anatomical location of all significant clusters was determined in reference to the Brainnetome (grey matter areas) and the JHU white matter diffusion tensor imaging atlases.[24,25]

Functional characterization

Finally, we aimed to provide an independent link from the anatomical locations identified by LDM to the behavioural level. Therefore, we performed a functional characterization analysis for the revealed significant clusters using the "behavioural domain" and "paradigm class" meta-data of prior task-evoked fMRI experiments as sorted in the BrainMap database (http://brainmap.org/). The behavioural domains consist of five main categories "cognition, action, perception, emotion, interoception," together with their respective subcategories. Accordingly, specific tasks that were conducted in the respective experiment are categorized into paradigm classes.

Quantitative "forward inference" and "reverse inference" were employed to characterize the functional profile of each significant cluster as previously described.[26] Specifically, the forward inference profiles a significant cluster via identifying taxonomic labels (domains or subdomains) for which the conditional probability of finding activation in a specific cluster is significantly higher than the overall chance (across the entire database) of finding activation in that particular cluster. To test the statistical significance of the forward inference, a binomial test was applied with a follow-up false discovery rate (FDR) correction for multiple comparisons at the level of $p < 0.05$. For the reverse approach, a cluster's functional profile was determined by identifying the most likely behavioural domains and paradigm classes associated with this particular cluster based on Bayes' rule. Here the statistical significance was established using a $\chi 2$ test followed by the same FDR correction strategy ($p < 0.05$) to account for multiple comparisons. In sum, forward inference assessed the probability of activation given a psychological term (i.e., task; [P(Activation|Task)]), while reverse inference assessed the probability of a psychological term given activation ([P(Task|Activation)])."

Role of funders

The funding sources for this project played no role in the study design, data collection, analysis, interpretation, writing, or editing of the manuscript.

Results

Sample characteristics

In total, data from N = 503 patients (35% female) were included in the first dataset (WAKE-UP) and N = 690

Characteristic	First dataset	Second dataset	p-value
Number of patients	503	690	–
Age, years (mean, SD)	65.2 (11.6)	71.5 (13.0)	<0.001
Sex, male (%)	325 (65%)	400 (58%)	0.021
Stroke side[a] (%)	Left 271 (54%) Right 199 (39%) Both 19 (4%)	Left 344 (50%) Right 312 (45%) Both 22 (3%)	0.187
NIHSS on admission (median, IQR)	6 (4–9)	4 (1–9)	<0.001
Stroke volume[b] (ml)	Mean 7.2 (SD 13.2) Median 1.7 (IQR 0.5–7.2)	–	–

P-values from resulting group comparisons (T-Test, Chi-squared test, and Wilcoxon rank-sum test where appropriate). Imaging data was available for the first dataset only.
[a]Data available for N = 489 patients in the first dataset and N = 678 in the second dataset. [b]Data available for N = 465 patients in the first dataset.

Table 1: **Demographic, clinical, and imaging characteristics for all stroke patient datasets analyzed in the study.**

patients (42% female) in the second dataset, reflecting recruitment in 64 stroke centres. Characteristics of both datasets are shown in Table 1 and supplemental Table S3. A sex-disaggreated demographic table can be found in the supplemental Tables S4. As already indicated, the different distributions of age and clinical severity between the datasets were intended to assess the generalization across populations and settings.

Deficit dimensions of the NIHSS

Results from in-sample and between-sample evaluations are shown in Fig. 1. In the *first dataset*, based on median evaluation indices, the most robust and generalizable in-sample factor model indicated the presence of *four* dimensions within the NIHSS. Detailed item-to-factor assignments are illustrated in Fig. 1A. The 4-factor model broadly separated traditional left-from right-hemispheric clinical deficits with one factor containing both NIHSS motor items ("motor arm and leg left") and non-motor items ("extinction", "gaze", "sensory"). In the *second dataset*, robustness and generalization evaluations resulted in a factor model containing *five* dimensions to be optimal (Fig. 1E). These five factors comprised NIHSS items for (1) left motor and (2) right motor deficits, (3) "ataxia" and "facial palsy", (4) language deficits including level of consciousness responses, and (5) items related to spatial orientation and awareness.

As a key difference to the 4-factor model, the 5-factor model in the *second dataset* separated the previous factor containing NIHSS items "motor arm left", "motor leg left", "extinction", "gaze", and "sensory" into two factors containing motor items ("motor arm left", "motor leg left") and non-motor items ("gaze", "sensory", "extinction" and "visual"). Likely, the "necessity" for an additional fifth factor in the second dataset can be attributed to the heterogeneity of NIHSS scores, i.e., clinical syndromes, among the stroke populations of both datasets.

In our between-sample comparison analysis regarding generalization from the *first* to the *second* dataset, a model with *five* factors (Fig. 1B) was again identified to be optimal for *generalizability*. As illustrated in Fig. 1F, the between-sample analysis indicated a marginally improved *robustness* of item-to-factor assignment in a model with six factors. As pre-specified in our analysis plan, this finding prompted our post-hoc analysis in case of differing optimal solutions regarding both generalizability and robustness. The results of this procedure, identifying and omitting non-fitting individual items, are described in detail below.

Of note, only a few NIHSS items contributed to multiple factors. All split-half results were fully corroborated by the follow-up bootstrapping and 10-fold cross-validation experiments with testing in hold-out data (supplemental Figs. S4 and S5). Results from the between-sample evaluation were confirmed by accounting for between-dataset differences in NIHSS score, age, and sex (supplemental Fig. S6).

Internal consistency and relationship among factors

Given the optimal in-sample and between-sample generalizability, the 5-factor model was used for subsequent analyses of internal consistency and inter-item relationships. Internal consistency of each factor assessed by Cronbach's alpha revealed the highest values for factors 1 and 2 (Cronbach's alpha: 0.87 and 0.88), followed by factor 4 (0.79) and factor 5 (0.60). Poor consistency was shown for factor 3 (0.13). Correlations between NIHSS items after adjustment of the total NIHSS score are shown in Fig. 2a and supplemental Fig. S7 (unadjusted correlations). In summary, correlations within a factor showed moderate to high positive values except for NIHSS items "level of consciousness" and "sensory," which demonstrated comparatively poorer correlations with other items in the same factor. In addition, the "ataxia" item showed anticorrelations with the other two items grouped in factor 3, both before and after controlling for NIHSS sum scores (before adjustment: items 7 and 4: r = –0.21; items 7 and 10: r = –0.12; after adjustment: items 7 and 4: r = –0.15; items 7 and 10: r = –0.05). Correlations between loadings on the five factors are illustrated in Fig. 2B.

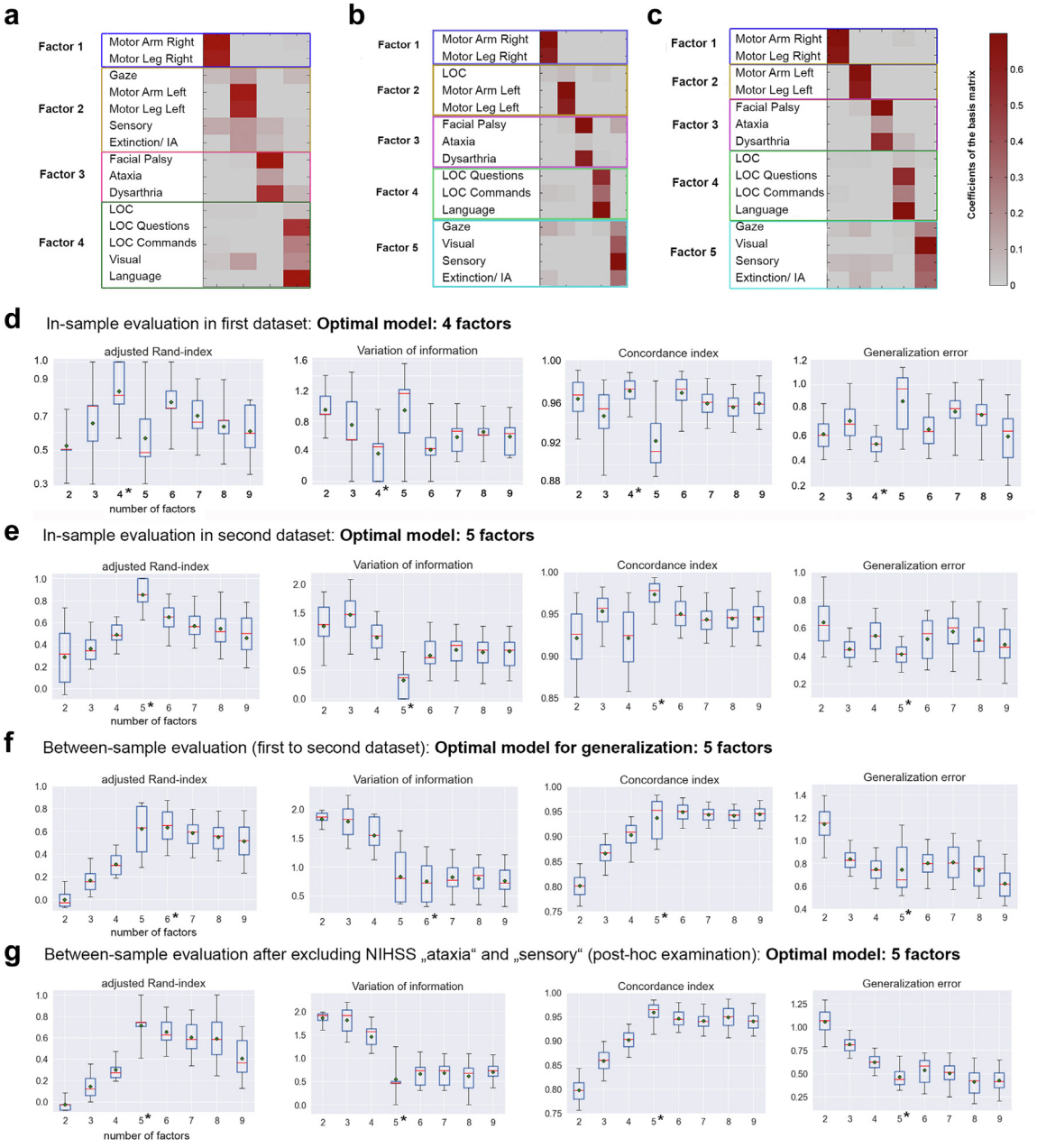

***Fig. 1:* Evaluation of within-sample and between-sample stability and generalizability of NIHSS factor models.** Results derived by orthonormal projective non-negative matrix factorization. Summary of item-to-factor assignments generated from in-sample evaluations (N = 1193): (a) the 4-factor model derived from the *first dataset*, (b) the 5-factor model derived from the *second dataset*, (c) the 5-factor model derived from the *first dataset* showing the best generalizability to the *second dataset*. The weight of an item in assigning to a specific factor (columns of the matrix) within the factor model structures is colour-coded according to the coefficients by two heat colour maps from grey (minimum) to dark red (maximum). (d) Results from the evaluation of in-sample stability and generalizability within the *first dataset*. The best model with four factors was chosen based on the median variation of information and generalization error achieving local minimum with adjusted Rand-Index and concordance index demonstrating the highest values. (e) Evaluation of in-sample stability and generalizability within the *second dataset*. (f) Evaluation of cross-sample stability and generalizability. Best generalizability from the *first* to the *second* dataset based on median out-of-sample generalization error achieving local minimum is achieved by the 5-factor model. (g) Evaluation of between-sample stability and generalizability (generalization error was assessed by projecting the factor structures from the first to the second dataset) after post-hoc examination leading to the exclusion of NIHSS items "ataxia" and "sensory" from the NIHSS. The heatmap for the 13-item abbreviated 5-factor structure resulting from post-hoc examination is illustrated in Fig. 3a. The asterisk (*) marks optimal models based on evaluation criteria as described in the methods section. Abbreviations: IA, Inattention; LOC, level of consciousness; NIHSS, National Institutes of Health Stroke Scale.

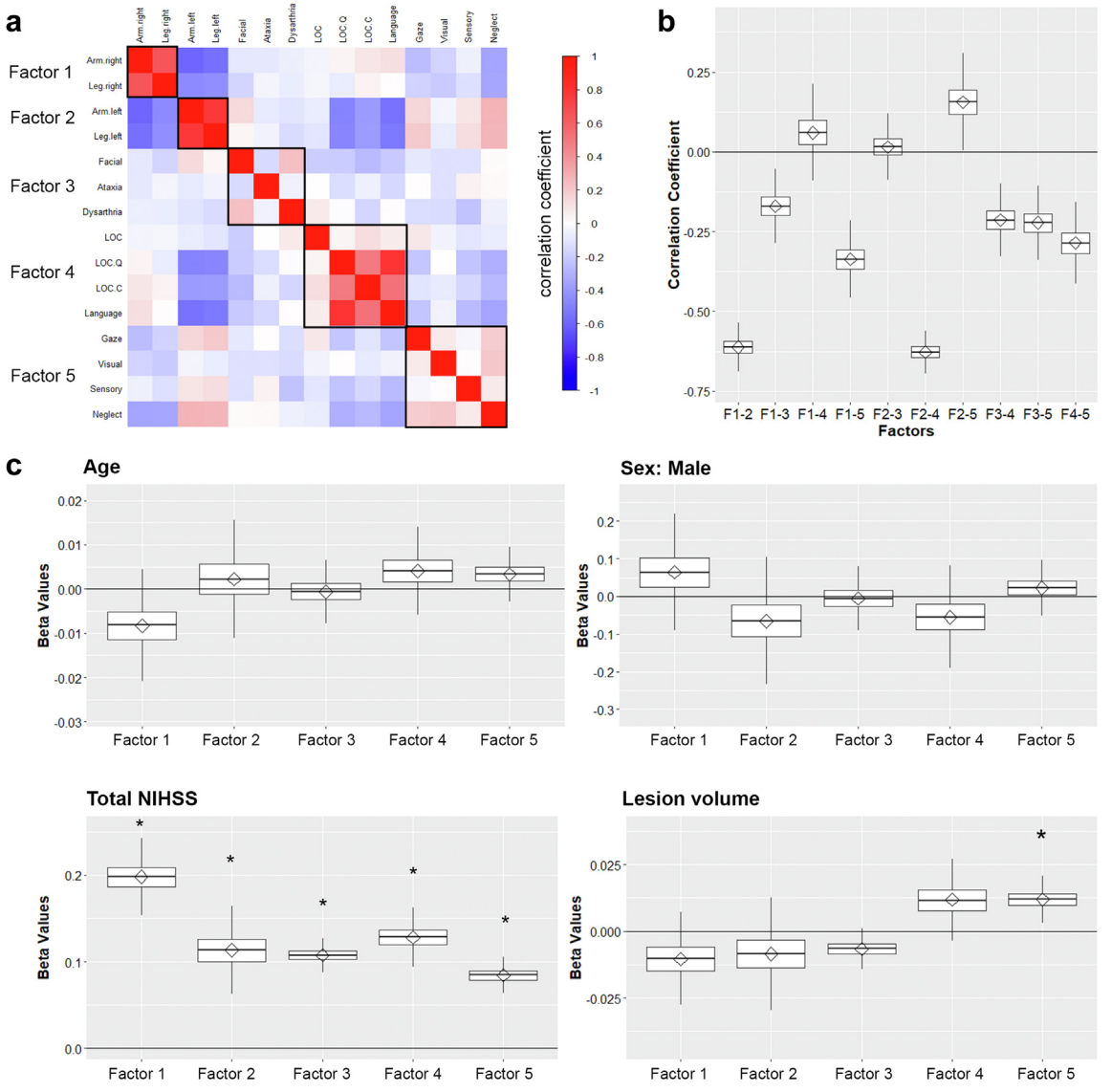

Fig. 2: **Relationship between NIHSS item scores, factor-loadings, and the association with clinical and imaging parameters.** Results for the 5-factor model from all datasets (N = 1193) are shown. (a) Heat map of single NIHSS inter-item correlations grouped by five factors (black squares) after controlling for total NIHSS score. Colour code: white to red: positive correlations, white to blue, negative correlations. (b) Boxplots of bootstrap results (repeated 10,000 times) for Pearson correlation among the loadings on the five factors after controlling for total NIHSS score. Line: median, diamond mean, whiskers 5th and 95th percentile. (c) Effect of age, sex, total NIHSS score, and lesion volume on the loadings of each of the five factors. Bootstrap results (repeated 10,000 times) for 4-way analysis of variance. Boxes refer to beta values: Line, median; diamond, mean; whiskers 5th and 95th percentile. *mean and median $p < 0.001$ (ANOVA). Abbreviations: NIHSS, National Institutes of Health Stroke Scale; LOC, Level of Consciousness.

MANOVA revealed a significant influence of total lesion volume ($F(4,495) = 11.1$; $p < 0.001$) and total NIHSS ($F(4,495) = 19.2$; $p < 0.001$) on the joint factor loadings, whereas no significant effects were detected for sex ($F(4,495) = 1.5$; $p = 0.200$) and age ($F(4,495)$, $p = 0.186$). Consecutive 4-way ANOVAs (Fig. 2C) demonstrated that lesion volume had a significant positive effect on factor 5 (median beta = 0.012; mean and median $p < 0.05$; averaged over 10,000 bootstraps). The total NIHSS score was shown to have a significant positive effect on all factors (F1: median beta = 0.198; F2: median beta = 0.113; F3: median beta = 0.108; F4: median beta = 0.128; F5: median beta = 0.085; all mean and median $p < 0.05$; averaged over 10,000 bootstraps). Due to the high collinearity of total NIHSS and lesion volumes ($r = 0.56$, $p < 0.001$), we conducted separate

consecutive 3-way ANOVAs omitting the total NIHSS score showing a significant positive effect of lesion volumes on all factors (see supplemental Fig. S8).

Post-hoc examination on factor-structure stability

Our between-sample bootstrapping evaluation pointed to a 5-factor model for optimal *generalizability* between both datasets, as demonstrated above. However, item-to-factor assignment *robustness* between the samples from the original two datasets indicated a marginally improved robustness for an alternative, albeit less *generalizable,* 6-factor model (Fig. 1F). To follow up on this finding, we performed an additional examination: Looking at the potential 6-factor model in detail (Fig. 1C), it isolated NIHSS items "ataxia" and "sensory" from the 5-factor model's factors 3 and 5 into one additional (sixth) factor that demonstrated extremely low internal Consistency (Cronbach's alpha = –0.035). Looking back at the 5-factor model, exactly these items were also characterized by poor correlations with other items in the respective factors, indicating their relatively less well-fit to their respective factors. From a clinical perspective, the "ataxia" item exclusively measures clinical deficits for infratentorial (cerebellar) stroke lesions. Therefore, it would be conceivable that it does not contribute well to the overall internal structure of the NIHSS, which was designed to capture supratentorial stroke lesion deficits.[27]

Hence, we removed the "ataxia" NIHSS item from factor 3, which considerably increased its internal consistency (Cronbach's alpha = 0.60 after and 0.13 before the removal of "ataxia"). Similarly, removing the "sensory" item from factor 5 yielded a slightly increased value of Cronbach's alpha (from 0.6 to 0.62). We re-ran the between-sample bootstrapping evaluation after excluding the two items "ataxia" and "sensory". This resulted in a 5-factor structure which was now the superior model regarding both *generalizability* and item-to-factor assignment *robustness* (Fig. 1G). In addition, running OPNMF again on the scores of the abbreviated 13-item NIHSS within the first dataset resulted in an identical factor structure to the one derived based on the full 15-item NIHSS but discarding the items "ataxia" and "sensory" (Fig. 3a). Our post-hoc examination, therefore, suggested that the items "ataxia" and "sensory" affect item-to-factor assignment *robustness*, yielding a sixth factor that is poorly defined, causing instability in the item-to-factor assignment. In other words, discarding this sixth factor resulted in the identical 5-factor structure as found before our post-hoc examination.

Neuroanatomical lesion locations and functional relevance of NIHSS deficit dimensions

We performed LDM based on factor loadings from the final 5-factor model derived from the first dataset with NIHSS items "ataxia" and "sensory" removed (Fig. 3). By this approach, the identified voxels/brain regions represent neural correlates of a low-dimensional structure of the NIHSS. That is, damage to either of the different clusters listed in Table 2 may yield similar impairments in the corresponding principal dimensions of the NIHSS. Stroke lesion masks were available for N = 451 (of N = 503) patients from the first dataset. The reasons for exclusion of N = 52 patients were insufficient data quality preventing robust lesion segmentation and registration to the MNI template (n = 38) and bilateral lesions (n = 14). The spatial distribution and frequency of stroke lesions are illustrated in Fig. 3b and supplemental Fig. S3.

Significant clusters (cluster-level FWE $p < 0.05$ corrected) from LDM were detected for all factors except for factor 3 (Fig. 3C; Table 2). Specifically, higher loadings (i.e., higher deficit severity) of factor 1 (NIHSS items "motor arm right" and "motor leg right") were associated with stroke lesions located in the left subcortical white matter (internal capsule, striatum) and left cortical motor areas (precentral gyrus, 496 voxels in total). Loadings of factor 2 ("motor arm left" and "motor leg left") were associated with lesions located in the right subcortical white matter (internal capsule, striatum, 215 voxels). Functional characterization (BrainMap) of these significant clusters showed a functional relation with motor tasks and behavioural domains of motor action execution and motor learning (Fig. 4, blue and red bars). Factor loadings for factor 4 ("language", level of consciousness commands and questions) were associated with lesions located at the left frontal cortex (caudal and ventrolateral areas of the left precentral gyrus) and left temporal cortex (left superior and middle temporal gyrus, 87 voxels in total). These clusters were found to be functionally associated with various language domains and related tasks (Fig. 4, green bars). Lastly, factor 5 ("gaze", "visual", "sensory" and "extinction") was mapped to lesions at right-hemispheric, cortical (parietal and temporal lobes), and subcortical (basal ganglia) areas (2741 voxels in total). Lesioned brain areas were associated with a variety of behavioural domains, including sensory, motor, and other perceptive functions (Fig. 4, orange bars).

The LDM and functional characterization results were largely replicated in the supplementary analyses where loadings from the 5-factor model before the post-hoc examination (i.e., including items "ataxia" and "sensory") were used (see supplemental Figs. S11 and S12, supplemental Table S6).

Discussion

In this study, we provide a comprehensive analysis of the internal structure of the NIHSS using an innovative approach with out-of-sample validation, followed by

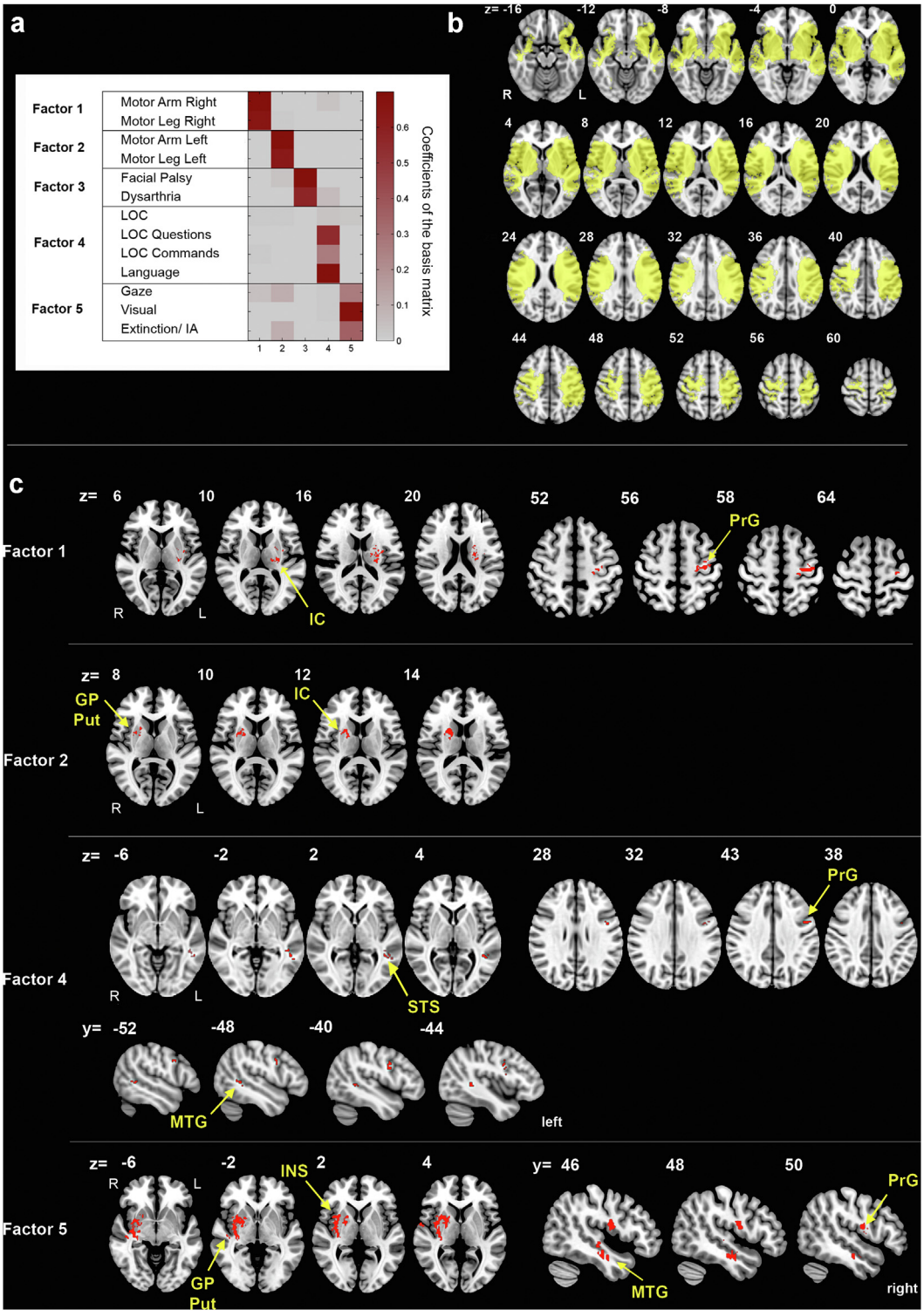

***Fig.* 3: Results of lesion-deficit mapping.** (a) Summary of item-to-factor assignments generated after post-hoc examination of the 5-factor model leading to the exclusion of NIHSS items "ataxia" and "sensory" from the NIHSS used for lesion-deficit mapping (LDM) in the first dataset (N = 451). The weight of an item in assigning to a specific factor (columns of the matrix) within the factor model structures is colour-coded according to the coefficients by two heat colour maps from grey (minimum) to dark red (maximum). (b) Brain areas with lesion coverage for statistical analysis (voxels affected by stoke lesions in N ≥ 5 patients are marked in yellow overlay). (c) Significant Clusters (cluster-level FWE

Factor	Anatomical location (Human Brainnetome Atlas and JHU ICBM-81 White matter atlas)
Factor 1 (496 voxels)	Left Internal Capsule
	Left Precentral Gyrus
	Left Globus Pallidus
	Left Putamen
	Left External Capsule
	Left Superior Corona Radiata
	Left Postcentral Gyrus
	Left Thalamus
Factor 2 (215 voxels)	Right Globus Pallidus
	Right Putamen
	Right Internal Capsule
	Right Caudate
	Right External Capsule
Factor 4 (496 voxels)	Left Superior Temporal Sulcus
	Left Precentral Gyrus
	Left Middle Temporal Gyrus
	Left Superior Longitudinal Fascicle
Factor 5 (2741 voxels)	Right Putamen
	Right Globus Pallidus
	Right Insular Cortex
	Right Internal Capsule
	Right Nucleus Accumbens
	Right External Capsule
	Right Caudate
	Right Middle Temporal Gyrus
	Right Precentral Gyrus
	Right Superior Temporal Gyrus

Results from LDM after post-hoc analysis excluding NIHSS items "ataxia" and "sensory"). The anatomical locations of significant voxel clusters (Fig. 3C) are referenced by the Human Brainnetome Atlas for grey matter areas and JHU ICBM-81 White matter atlas. Locations are listed in (descending) order ranked by the number of voxels. No significant voxels were identified for factor loadings in factor 3.

Table 2: **Results of Lesion-Deficit Mapping (LDM) for individual factor loadings of the 5-factor model.**

LDM and functional profiling. We reported two main findings. First, systematic evaluation and cross-validation in independent datasets revealed an optimal factor structure of the NIHSS, representing five symptom dimensions as the most stable, replicable, and generalizable across patient populations and study settings. Second, LDM linked deficit dimensions to anatomical regions with corresponding brain functions as informed by functional characterization, incorporating prior task-evoked fMRI experiments. Intriguingly, the characterized functional profiles of significant anatomical clusters and their associated NIHSS behaviour deficits were mutually confirmed by LDM and functional decoding by task activations. Therefore, our results converged from brain dysfunction to structure and independently back to brain function, providing comprehensive evidence supporting the construct validity of the NIHSS.

From a methodological point of view, our study highlights the impact of cross-validation and the evaluation of out-of-sample generalization performance in differing stroke populations for determining the internal structure of the NIHSS. An in-sample analysis of the NIHSS in the *first* dataset indicated the presence of *four* factors. When encountering the *second* independent dataset, a model with *five* factors was superior regarding generalizability. These differing results are explained by the distinct characteristics of both datasets, specifically the larger heterogeneity of clinical deficits of patients in the second dataset. Whereas the first dataset comprised pre-selected patients from a randomized stroke trial based on evidence of an acute stroke lesion and planned treatment by thrombolysis (WAKE-UP trial),[17] the second dataset comprised a more "general" stroke hospital population irrespective of planned thrombolytic treatment (EPOS study) complemented with a cohort of patients with large-vessel occlusions, higher NIHSS scores and larger final infarct volumes (I-Know study). Of note, the 5-factor model remained superior after matching both datasets for age, sex, and NIHSS score, indicating that none of these variables drove the emergence of the fifth factor.

Synthesizing results from NIHSS factorization, LDM, and functional decoding of associated anatomical regions, we found that the five deficit dimensions are linked to disturbed lateralized and non-lateralized brain functions. Brain regions identified by LDM represent neural correlates, i. e. network hubs supporting principal brain functions as captured by a low-dimensional structure (five factors) of the NIHSS.

Disturbed *left-hemispheric* brain functions were captured by two factors: factor 1 with NIHSS items "motor arm right" and "motor leg right" for right-sided motor deficits and factor 4 capturing speech deficits ("LOC questions", "LOC commands" and "best language"). Although LOC items do not assess language

corrected $p < 0.05$ following voxel-level $p < 0.005$ thresholding) identified through 10,000 repeated permutation tests are shown in red on representative sections of a brain template in MNI standard space oriented in radiological convention. Significant clusters for the loadings on factor 4 and 5 are illustrated in an additional sagittal orientation to illustrate the anatomical localizations on the cerebral cortex. See also Table 2 for details. LDM results based on the 5-factor model before post-hoc examination (full 15-item scale) are shown in supplemental Fig. S11. MNI z-coordinates of each section are shown. Abbreviations: IC, internal capsule; PrG, Precentral Gyrus; GP, Globus Pallidus; STS, Superior Temporal Sulcus; MTG, Middle Temporal Gyrus; Put, Putamen; MTG, Middle Temporal Gyrus; INS, Insular Cortex; R, right; L, left; LOC, Level of Consciousness.

Fig. 4: **Results from the functional characterization of anatomical locations revealed by lesion-deficit mapping.** Functional characterization based on the BrainMap database based on data from both datasets (N = 1193).[28] Functional profiles of each characterized factor were determined by forward inference (left columns), assessing the activation likelihood ratios for each significant cluster concerning a given domain or paradigm and reverse inference (right columns), assessing the probability of a domain's or paradigm's occurrence given activation in a significant cluster. Only significant functional assignments (false discovery rate corrected $p < 0.05$) are presented. Results for behavioural

directly, they evaluate the patient's comprehension abilities needed to respond to questions and simple commands. In both factors, correlations between individual items were high. As demonstrated by LDM, clinical deficits attributed to factor 1 (right-sided motor deficits) were significantly associated with stroke lesions involving left-hemispheric white matter in the internal capsule and precentral cortical areas. This finding is plausible given the presence of upper motor neurons in these areas, including both precentral, primary motor cortical areas and the corticospinal tract at the level of the internal capsule. Damage and degeneration of the corticospinal tract are known determinants of motor impairment and recovery after stroke.[29,30] Functional characterization of areas detected by our LDM analysis revealed behavioural domains of motor functions such as "execution of actions" and motor paradigm classes such as "grasping". For factor 4 (speech), significant associations were detected for two left-hemispheric clusters. The first comprised the left caudal precentral gyrus (premotor cortex), closely adjacent to Broca's area at the inferior frontal gyrus. The identified area can be attributed to language production based on findings from LDM studies after brain injury or data from electrical stimulation experiments.[31] The second involved the left superior and middle temporal cortex and adjacent portions of the superior longitudinal fascicle (SLF). The results fit well with the organization of language functions in a network of temporo-frontal cortical brain areas located in the dominant (left) hemisphere connected through long traversing white matter bundles such as the SLF, as shown by previous LDM studies in stroke patients,[5] histopathology in animal and human brain studies and functional imaging experiments in healthy participants.[31–33] Consistently, functional characterization revealed behavioural domains and experimental paradigms covering various aspects of language functions.

Deficits of *right-hemispheric* brain functions were captured by two factors: Factor 2 containing NIHSS items for left-sided motor deficits ("motor arm left" and "motor leg left") and factor 5 containing various items sensitive to deficits in lateralized attention and neglect, mainly present in patients with right-hemispheric stroke,[34–36] comprising "gaze," "extinction and inattention," "sensory" and "visual". Of note, some of these items are likely linked by potential rating misattributions rather than shared anatomical substrates. For example, ratings of sensory deficits can erroneously occur in patients with spatial neglect.[37] As a result of our post-hoc examination, the internal consistency of factor 5 was improved after removing the "sensory" item and LDM conducted on the adjusted factor. We located significant clusters in the right hemispheric subcortical and cortical areas involving the basal ganglia (caudate nucleus and putamen) as well as the right precentral, middle and superior temporal gyrus. In terms of lateralization, our LDM findings are in line with the frequent occurrence of spatial neglect with eye deviation and extinction (factor 5) in patients with right hemispheric stroke.[34–36] Regarding the specific anatomical locations, damage to right-temporal cortical areas has been shown to evoke spatial neglect due to the involvement of cortico-subcortical structural brain networks promoting lateralized attention under physiological circumstances.[38] These distributed networks also involve subcortical structures such as the putamen and caudate nucleus that were likewise detected in our LDM analysis. Functional characterization of identified brain regions revealed various brain functions involved with perception, emotions, and action execution which are at least in part affected by deficits in lateral attention. As an alternative interpretation, the significant effect of lesion volumes on factor loadings (Fig. 2c) indicates that factor 5 might more generally represent the extent of right-hemispheric stroke. In addition, we note that, that the proposed behavioural domain of factor 5 (neglect/spatial inattention) and corresponding behavioural experiments are underrepresented in the BrainMap taxonomy and database, which could explain the absence of more specific functional associations.

Apart from the lateralized deficit dimensions, our analysis revealed the presence of a previously not reported factor (factor 3) representing clinical deficits without stereotypical lateralization, specifically "dysarthria", "facial palsy", and "ataxia." Looking at individual items within this factor 3, we note that "ataxia" is the only NIHSS item exclusively capturing symptoms from infratentorial stroke.[27] Indeed, in our post-hoc examination of the 5-factor model, the internal consistency of factor 3 increased after removing "ataxia", with the remaining items consistently grouped into factor 3 among all model solutions. There are two important aspects of this result: First, it indicates that, across diverse stroke populations, the NIHSS captures a clinical syndrome of non-lateralized clinical deficits, namely "dysarthria" and "facial palsy". This finding is distinct compared to the most common understanding of the NIHSS being constructed along two axes of strictly left- or right-hemispheric clinical deficits. As an explanation, the "traditional" factor structure of the NIHSS was initially verified in data from the National Institute of Neurological Disorders and Stroke (NINDS) tPA Stroke Trial, which included a selection of severely affected

domains (a) and paradigm classes (b) are shown for each factor by colour coding. Note that since no significant anatomical regions were detected for lesion-deficit mapping of factor 3, it was omitted from functional characterization analysis.

stroke patients (median NIHSS: 14) with mainly cardioembolic or large-vessel occlusive stroke aetiology.[5] Patients with less severe clinical deficits ("minor stroke") and those with lacunar stroke lesions were most likely underrepresented in these datasets.[6] However, patients with relatively small lesion volumes and lower NIHSS scores are common in stroke populations[5] and were well represented in our study, specifically the second dataset.

Second, from a clinical point of view, factor 3 closely corresponds to a clinical syndrome found in patients with lacunar stroke caused by strategically located lesions involving the white matter motor pathways.[39–41] Interestingly, LDM did not reveal any anatomical locations associated with symptom severity for factor 3. If we consider this factor to represent lacunar syndromes, the negative LDM results could be explained by the limited power from our sample size to detect associated lacunar lesions that occur in high anatomical variability in both hemispheres and the brain stem, where lesion coverage was low in our dataset.

Analysis of relationships between factor loadings and associations with lesion volumes (Fig. 2) supports our interpretations of individual deficit dimensions. Loadings from factors capturing left- and right-hemispheric deficits were negatively correlated since bilateral motor deficits rarely occur in stroke patients. Accordingly, there was a strong positive correlation between left-hemispheric factors 1 (right motor deficits) and 4 (speech). There was a significant effect of larger lesion volumes on higher factor loadings for factor 5 (neglect) due to the incremental effect from larger, territorial stroke lesions involving cortical areas underlying brain functions of spatial awareness and orientation.

Our proposed 5-factor structure underlying the NIHSS is based on an innovative, more optimal, comprehensive, systematic evaluation in independent datasets as compared to previous reports, where a 4-factor model was suggested by EFA, capturing latent dimensions of left- and right-hemispheric, cortical, and motor symptoms.[4,2,42] Regarding the generalizability of the principal NIHSS factor structure, Lyden et al. employed CFA on an independent stroke population[42] to test for the *a-priori* hypothesis of a 2- or 4-factor structure.[5] In contrast, our fully data-driven approach evaluated multiple potential factor models across datasets. Interestingly, our findings also suggest excluding two NIHSS items ("ataxia" and "sensory"), as these would affect factor-structure stability leading to less robust factorization across independent samples. This observation converges with previous results that have either removed these items post-hoc or before analysis.[4] Therefore, our results further strengthen the argument for excluding both "ataxia" and "sensory" items for a modified, 13-item version of the NIHSS. Of note, at the brain level, the full-version 5-factor model (15 items) and the 13-item post-hoc structure gave almost identical LDM results corroborating a reliable link between lesion locations and their affected deficit dimensions.

Our study is limited by the relatively low lesion volumes in the first dataset resulting in moderate lesion coverage for LDM, owing to the stereotypical lesion distribution of stroke lesions in anterior circulation stroke. In addition, the NIHSS is biased toward symptoms resulting from supratentorial stroke lesions, further limiting the sensitivity to detect relevant associations between behaviour and anatomy at the infratentorial level (brainstem and cerebellum) in our study. We expect that in cohorts with larger territorial stroke lesions, LDM would identify additional cortical brain areas in peripheral brain regions for all factors (specifically for brain functions located in cortical brain areas). Although patients with larger final infarct volumes were represented in the second dataset, the generalization of LDM results to patient populations with territorial perfusion deficits (i. e. resulting from large vessel occlusion) was not assessed in our study. Also, independent datasets for testing the out-of-sample predictability of lesion locations for the severity of clinical deficits in specific NIHSS dimensions would be desired, given that there was no imaging data from the time point initial NIHSS scoring available in our second dataset. However, datasets comprising accurate MRI-based lesion delineations and fine-grained clinical phenotyping (i.e., single NHSS items) are scarce. Future studies, potentially from multicenter cooperation that include larger patient numbers with more severe and extensive lesion volumes, are needed.

In summary, by using an unsupervised machine learning approach of OPNMF and systematic in- and cross-sample evaluations of stability and generalizability on large, multicenter, and heterogeneous datasets, the present work revealed a structure with five dimensions best representing the clinical construct in stroke patients as assessed by NIHSS. Besides clearly lateralized clinical deficits, our results revealed an additional, non-lateralized factor, potentially capturing lacunar syndromes frequently encountered in stroke patients. Subsequent LDM revealed meaningful neuroanatomical locations corroborated by functional behavioural characterizations. Our work shows the overarching structure of post-stroke deficits captured by the NIHSS. We provide an anatomical map of the NIHSS that can be applied to predict principal dimensions of clinical deficits in stroke patients, potentially guiding individualized rehabilitation approaches. By mapping brain areas most critical for clinical impairment, our results contribute to testing individualized selection algorithms for recanalization therapy in acute stroke integrating the strategic importance of eloquent brain regions and known outcome predictors. Lastly, our study also opens the window for targeted use of low-dimensional versions of the NIHSS in research, clinical trials, or clinical decision-making.

Contributors

Each author has made a significant contribution to the manuscript and had full access to all data analyzed in our work. All authors read and approved the final version of the manuscript. Contributions in detail are listed below: Bastian Cheng: verification of underlying data, literature search, study design, data collection, data analysis, figures, data interpretation, writing, critically revising manuscript, funding acquisition. Ji Chen: verification of underlying data, literature search, study design, data analysis, figures, data interpretation, writing, critically revising manuscript, funding acquisition. Alina Königsberg: data collection, data analysis, critically revising manuscript. Carola Mayer: data collection, data analysis, critically revising manuscript. Leander Rimmele: literature search, data collection, data analysis, critically revising manuscript. Kaustubh R. Prail: verification of underlying data, literature search, study design, data analysis, critically revising manuscript. Christian Gerloff: study design, data collection, data interpretation, critically revising manuscript, funding acquisition. Götz Thomalla: verification of underlying data, study design, data collection, data analysis, figures, data interpretation, writing, critically revising manuscript, funding acquisition. Simon B. Eickhoff: verification of underlying data, study design, data collection, data analysis, figures, data interpretation, writing, critically revising manuscript, funding acquisition.

Data sharing statement

Data are available upon reasonable request from the corresponding author. Specifically, anonymized clinical data and imaging data (stroke lesion mask segmentations only) can be requested and access granted after approval of the individual trial steering committees. The code for OPNMF as applied can be accessed publicly: https://github.com/jichen-psy/OPNMFevaluation.

Declaration of interests

The authors report no competing interests.

Acknowledgements

This research was supported by the National Key R&D Program of China (Grant No. 2021YFC2502200), the National Human Brain Project of China (Grant No. 2022ZD0214000)", the German Research Foundation (Deutsche Forschungsgemeinschaft), Project 178316478 (A1, C1, C2), and Project 454012190 of the SPP 2041, the Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain" and Helmholtz Imaging Platform grant NimRLS (ZT-I-PF-4-010).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2022.104425.

References

1 Lyden P, Brott T, Tilley B, et al. Improved reliability of the NIH stroke scale using video training. NINDS TPA stroke study group. *Stroke*. 1994;25(11):2220–2226.
2 Lyden P, Lu M, Jackson C, et al. Underlying structure of the National Institutes of Health stroke scale: results of a factor analysis. *Stroke*. 1999;30(11):2347–2354.
3 Raza SA, Frankel MR, Rangaraju S. Abbreviation of the follow-up NIH stroke scale using factor analysis. *Cerebrovasc Dis Extra*. 2017;7(3):120–129.
4 Zandieh A, Kahaki ZZ, Sadeghian H, et al. The underlying factor structure of National Institutes of Health stroke scale: an exploratory factor analysis. *Int J Neurosci*. 2012;122(3):140–144.
5 Corbetta M, Ramsey L, Callejas A, et al. Common behavioral clusters and subcortical anatomy in stroke. *Neuron*. 2015;85(5):927–941.
6 Barow E, Boutitie F, Cheng B, et al. Functional outcome of intravenous thrombolysis in patients with lacunar infarcts in the WAKE-UP trial. *JAMA Neurol*. 2019;76(6):641–649.
7 Cheng B, Forkert ND, Zavaglia M, et al. Influence of stroke infarct location on functional outcome measured by the modified rankin scale. *Stroke*. 2014;45(6):1695–1702.
8 Schlemm E, Ingwersen T, Königsberg A, et al. Preserved structural connectivity mediates the clinical effect of thrombolysis in patients with anterior-circulation stroke. *Nat Commun*. 2021;12(1):2590.
9 Menezes NM, Ay H, Wang Zhu M, et al. The real estate factor: quantifying the impact of infarct location on stroke severity. *Stroke*. 2007;38(1):194–197.
10 Kerleroux B, Benzakoun J, Janot K, et al. Relevance of brain regions' eloquence assessment in patients with a large ischemic core treated with mechanical thrombectomy. *Neurology*. 2021;97(20):E1975–E1985.
11 Sotiras A, Resnick SM, Davatzikos C. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. *Neuroimage*. 2015;108:1–16.
12 Yang Z, Oja E. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Trans Neural Netw*. 2010;21(5):734–749.
13 Chen J, Patil KR, Weis S, et al. Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: an international machine learning study. *Biol Psychiatry*. 2020;87(3):282–293.
14 Chen J, Müller VI, Dukart J, et al. Intrinsic connectivity patterns of task-defined brain networks allow individual prediction of cognitive symptom dimension of schizophrenia and are linked to molecular architecture. *Biol Psychiatry*. 2021;89(3):308–319.
15 de Haan B, Karnath H-O. A hitchhiker's guide to lesion-behaviour mapping. *Neuropsychologia*. 2018;115:5–16.
16 Xu T, Jha A, Nachev P. The dimensionalities of lesion-deficit mapping. *Neuropsychologia*. 2018;115:134–141.
17 Thomalla G, Simonsen CZ, Boutitie F, et al. MRI-guided thrombolysis for stroke with unknown time of onset. *N Engl J Med*. 2018;379(7):611–622.
18 Rimmele DL, Lebherz L, Frese M, et al. Outcome evaluation by patient reported outcome measures in stroke clinical practice (EPOS) protocol for a prospective observation and implementation study. *Neurol Res Pract*. 2019;1(1):1–7.
19 Cheng B, Boutitie F, Nickel A, et al. Quantitative signal intensity in fluid-attenuated inversion recovery and treatment effect in the WAKE-UP trial. *Stroke*. 2020;51(1):209–215.
20 Mazziotta J, Toga A, Evans A, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos Trans R Soc Lond B Biol Sci*. 2001;356(1412):1293–1322.
21 Boutsidis C, Gallopoulos E. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognit*. 2008;41(4):1350–1362.
22 Zhang Y, Kimberg DY, Coslett HB, Schwartz MF, Wang Z. Multivariate lesion-symptom mapping using support vector regression. *Hum Brain Mapp*. 2014;35(12):5861–5876.
23 DeMarco AT, Turkeltaub PE. A multivariate lesion symptom mapping toolbox and examination of lesion-volume biases and correction methods in lesion-symptom mapping. *Hum Brain Mapp*. 2018;39(11):4169–4182.
24 Mori S, Oishi K, Jiang H, et al. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *Neuroimage*. 2008;40(2):570–582.
25 Fan L, Li H, Zhuo J, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb Cortex*. 2016;26(8):3508–3526.
26 Genon S, Li H, Fan L, et al. The right dorsal premotor mosaic: organization, functions, and connectivity. *Cereb Cortex*. 2017;27(3):2095–2110.
27 Lyden P. Using the National Institutes of Health stroke scale: a cautionary tale. *Stroke*. 2017;48(2):513–519.
28 Fox PT, Lancaster JL. Mapping context and content: the BrainMap model. *Nat Rev Neurosci*. 2002;3(4):319–321.
29 Zhu LL, Lindenberg R, Alexander MP, Schlaug G. Lesion load of the corticospinal tract predicts motor impairment in chronic stroke. *Stroke*. 2010;41(5):910–915.
30 Thomalla G, Glauche V, Weiller C, Röther J. Time course of wallerian degeneration after ischaemic stroke revealed by diffusion tensor imaging. *J Neurol Neurosurg Psychiatr*. 2005;76(2):266–268.
31 Petrides M. The ventrolateral frontal region. In: *Neurobiology of language*. Elsevier; 2016:25–33.
32 Poeppel D, Hickok G. Towards a new functional anatomy of language. *Cognition*. 2004;92(1–2):1–12.

33 Hickok G, Poeppel D. Processing cortical organization of speech processing. *Nat Rev Neurosci*. 2007;8(5):393–402.
34 Verdon V, Schwartz S, Lovblad K-O, Hauert C-A, Vuilleumier P. Neuroanatomy of hemispatial neglect and its functional components: a study using voxel-based lesion-symptom mapping. *Brain*. 2010;133(Pt 3):880–894.
35 Becker E, Karnath HO. Incidence of visual extinction after left versus right hemisphere stroke. *Stroke*. 2007;38(12):3172–3174.
36 Becker E, Karnath HO. Neuroimaging of eye position reveals spatial neglect. *Brain*. 2010;133(3):909–914.
37 Moore MJ, Vancleef K, Shalev N, Husain M, Demeyere N. When neglect is neglected: NIHSS observational measure lacks sensitivity in identifying post-stroke unilateral neglect. *J Neurol Neurosurg Psychiatry*. 2019;90(9):1070–1071.
38 Karnath HO, Himmelbach M, Rorden C. The subcortical anatomy of human spatial neglect: putamen, caudate nucleus and pulvinar. *Brain*. 2002;125(2):350–360.
39 Abdul-Rahim AH, Fulton RL, Sucharew H, et al. National Institutes of Health stroke scale item profiles as predictor of patient outcome: external validation on independent trial data. *Stroke*. 2015;46(2):395–400.
40 Fisher CM. Lacunar strokes and infarcts: a review. *Neurology*. 1982;32(8):871–876.
41 Tanaka K, Yamada T, Torii T, et al. Pure dysarthria and dysarthria-facial paresis syndrome due to internal capsule and/or corona radiata infarction. *BMC Neurol*. 2015;15(1):1–5.
42 Lyden P, Claesson L, Havstad S, Ashwood T, Lu M. Factor analysis of the National Institutes of Health stroke scale in patients with large strokes. *Arch Neurol*. 2004;61(11):1677–1680.