Recovery after stroke: the severely impaired are a distinct group

Anna K. Bonkhoff[1,2], Thomas Hope[3], Danilo Bzdok[4,5,6], Adrian G. Guggisberg[7], Rachel L. Hawe[8,9], Sean P. Dukelow[8], François Chollet[10], David J. Lin[11], Christian Grefkes[2,12], and Howard Bowman[13,14]

[1] J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[2] Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Juelich, Juelich, Germany
[3] Wellcome Centre for Human Neuroimaging, University College London, UK
[4] Mila – Quebec Artificial Intelligence Institute, Montreal, Canada
[5] Department of Biomedical Engineering, McConnell Brain Imaging Centre, Montreal Neurological Institute, Faculty of Medicine, McGill University, Montreal, Canada
[6] Canadian Institute for Advanced Research (CIFAR)
[7] Clinical Neuroscience, University of Geneva, Medical School, Geneva, Switzerland
[8] Department of Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Alberta, Canada
[9] School of Kinesiology, University of Minnesota, Minneapolis, Minnesota, USA
[10] Neurology Department, Centre Hospitalier Universitaire de Toulouse, Hôpital Purpan, Toulouse, France; Institut des Sciences du Cerveau de Toulouse (INSERM, CNRS, Université de Toulouse), Toulouse, France
[11] Center for Neurotechnology and Neurorecovery, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[12] Department of Neurology, University Hospital Cologne, Cologne, Germany
[13] School of Psychology, University of Birmingham, UK
[14] School of Computing, University of Kent, UK

Number of references: 39 references

Abstract count: 252 words

Word count: 4667 words

Abstract

Introduction

Stroke causes different levels of impairment and the degree of recovery varies greatly between patients. The majority of recovery studies are biased towards patients with mild-to-moderate impairments, challenging a unified recovery process framework. Our aim was to develop a statistical framework to analyze recovery patterns in patients with severe and non-severe initial impairment and concurrently investigate whether they recovered differently.

Methods

We designed a Bayesian hierarchical model to estimate three-to-six-months upper limb Fugl-Meyer(FM) scores after stroke. When focusing on the explanation of recovery patterns, we addressed confounds affecting previous recovery studies and considered patients with FM-initial-scores<45 only. We systematically explored different FM-breakpoints between severe/non-severe patients(*FM-initial*=5–30). In model comparisons, we evaluated whether impairment-level-specific recovery patterns indeed existed. Finally, we estimated the out-of-sample prediction performance for patients across the entire initial impairment range.

Results

Recovery data was assembled from eight patient cohorts(n=489). Data were best modelled by incorporating two subgroups(breakpoint: *FM-initial*=10). Both subgroups recovered a comparable constant amount, but with different proportional components: severely affected patients recovered more the smaller their impairment, while non-severely affected patients recovered more the larger their initial impairment. Three-to-six-months outcomes could be predicted with an R-squared=63.5% (95%-confidence interval=51.4%–75.5%).

Conclusions

Our work highlights the benefit of simultaneously modelling recovery of severely-to-non-severely impaired patients and demonstrates both shared and distinct recovery patterns. Our findings provide evidence that the severe/non-severe subdivision in recovery modelling is not an artifact of previous confounds. The presented out-of-sample prediction performance may serve as benchmark to evaluate promising biomarkers of stroke recovery.

Keywords

Acute ischemic stroke, recovery, Bayesian hierarchical models, Fugl-Meyer Scale

Introduction

Almost half of stroke patients are confronted with permanent impairments, such as motor weakness.[1] A comprehensive and more mechanistic understanding of recovery after stroke is thus indispensable to successfully guide clinical decision-making and neurorehabilitation treatments. This understanding may comprise two main dimensions: (i) the explanation and (ii) prediction of recovery after stroke.[2,3,4,5,6] With respect to explanation, relevant questions are: what can we infer about motor recovery patterns at a patient *group* level? In what way is the initial impairment *associated* with the outcome? Is this association the same for severely and non-severely affected patients? With respect to prediction, we may wonder: How well can we *forecast* motor recovery for *individual* patients?

The dominant perspective in the field of upper limb motor recovery is that patients should be divided into severely and non-severely affected groups,[7] when modelling stroke motor recovery in general and investigating biomarker and rehabilitative treatment development in particular.[8,9] Central to the emergence of this division between severe and non-severe groups has been the identification of a strong proportional (to lost) recovery pattern in the non-severe group: less severely affected stroke patients were repeatedly found to recover proportional to their initially lost motor function. This proportionality implied greater recovery, the more substantial the initial impairment[10,11,12,13,14,15] and was termed "proportional recovery".[7] However, these studies in non-severely affected patients were shown to considerably overestimate the explained variance, i.e., derive an excessively high estimate of how well motor recovery could be explained. Confounding effects, such as ceiling and mathematical coupling, led to the false belief that recovery could be predicted with very high certainty.[16,17] De-confounding, especially by excluding patients with mild initial impairments and a high likelihood of achieving maximum performance at follow-up ("*being at ceiling*"), resulted in a substantially lower variance explained (e.g., 94% in [13] before and 32% in [18] after addressing confounds). Interestingly, a larger explained variance of 53% could be observed when deriving a single recovery pattern across all degrees of initial impairment, i.e., severely and non-severely affected patients combined, even after addressing confounds.[18] The findings of lower explained variances for the original non-severe case in combination with the higher explained variance when combining severe and non-severe patients thus challenged the validity of the "proportional recovery" rule in regard to the explanation and prediction of stroke

recovery. These insights furthermore questioned the practice of treating severely and non-severely affected patients as distinct groups.

It is the accurate prediction of recovery for *new* patients, who have just experienced a stroke and whose outcome is yet unknown, which is of most clinical value. Such prediction of outcomes could guide decision-making for clinicians, patients and their proxies.[19] In particular, accurate predictions could allow clinicians to stratify patients for additional interventions based on their predicted potential for neurobiological recovery and could even help to demonstrate benefits of novel rehabilitative therapies. Here, effective treatments would, for example, consistently lead to a higher recovery than predicted.[9,20,21] Importantly, the predictive capacity of the initial upper limb impairment has mostly been tested in-sample, i.e., outcome models were fitted and evaluated on the same sample of patients.[7] As a result, effect sizes, such as the explained variance, were likely too optimistic.[3] In addition, this (in-sample) overestimation was probably compounded by ceiling and mathematical coupling confounds as mentioned earlier. In contrast, an out-of-sample estimate, based on a strict separation of patients used for fitting and performance evaluation, combined with procedures to mitigate ceiling and mathematical coupling, will capture the prediction performance for previously unseen patients more veridically as all of these factors can otherwise lead to inflated estimates.

In the present study, we aimed to enhance both the explanation as well as the out-of-sample prediction of motor recovery after severe and non-severe stroke. Key objectives of this paper were firstly (i) the assessment of the validity of the severe versus non-severe separation, after carefully addressing ceiling effects to mitigate the risk of any biases. That is, we evaluated whether the level of initial impairment substantially influenced the inferred recovery pattern or whether it would be the same no matter the initial impairment. To compensate for the focus on moderate-to-mildly affected patients in numerous previous recovery studies,[10,11,12,13,14,15] we additionally analyzed recovery patterns in severely affected patients. We hypothesized that we would identify impairment-specific recovery patterns, indicating biologically distinct cerebral processes after acute ischemic stroke for severe or non-severe impairment. Secondly, with the objective to progress precision neurology and individual patient-level outcome predictions, we (ii) shifted the focus from interpreting recovery patterns to obtaining an out-of-sample prediction performance.

To enhance the generalizability of our results, we assembled the largest stroke recovery data set, comprising information on the upper limb motor recovery of 489 stroke patients from eight individual studies. We chose a modelling framework that had the capacity to reflect potential similarities and differences between studies, as well as severely and non-severely affected patients. Specifically, we employed Bayesian hierarchical models[22] to model long-term motor outcome post-stroke.

Material & Methods

Participants

Analyses presented here rely on recovery data of 489 stroke patients originating from eight individual studies (**Table 1**, **Figure 1**).[10,11,12,14,15,23,24,25] The inclusion process included a literature research based on keywords "poststroke", "recovery", "motor function", "longitudinal", "Fugl-Meyer" on PubMed as well as referenced literature. Further criteria comprised: > 20 stroke patients, initial and follow-up FM Score of the upper limb; initial: acute and early subacute phase post-stroke, follow-up: 3-6 months post-stroke. Corresponding authors of eleven resulting studies were contacted, featuring longitudinal data on Fugl-Meyer (FM) scale-based motor function post-stroke were contacted. Three author teams[15,16,23] were able to share their data (c.f., **supplementary materials** for further details). Furthermore, individual-level data from [10] was openly available, and data from [24] became available after the initial literature search. Recovery data comprised individual-level Fugl-Meyer scores measuring upper limb motor impairment at two time instants: once in the early subacute stage, and once in the late subacute stage,[20] three to six months after the cerebrovascular event (FM-UL: 0: No upper limb movement including reflexes, 66: maximal score, c.f., **supplementary materials**). The majority of included patients were recruited as part of studies that considered first-ever stroke patients only.[26,12,13,14] Chollet et al.[23] and Guggisberg et al.[15] excluded patients with prior stroke in case of residual motor symptoms. Subjects provided informed consent and ethics approvals had in general been granted for all of the individual primary studies. The FLAME study[23] had been approved by the Toulouse Ethics Committee, the SMaHRT study[24] had been approved by the Institutional Review Board at Partners Healthcare, the study by Guggisberg and colleagues[15] had been approved by the Geneva Ethics Committee.

	Study 1 Zarahn et al. 2011	**Study 2** Byblow et al. 2015	**Study 3** Feng et al. 2015	**Study 4** Winters et al. 2015	**Study 5** Buch et al. 2016	**Study 6** Guggisberg et al. 2017	**Study 7** Chollet et al. 2011 Control group	**Study 8** Lin et al. 2019
Total number of subjects	30	44	57	166	25	63	56	48
Mean age in years (SD)	60.3 (9.9)	67 (range 31 – 97)	58.8 (14.0)	Fitters: 66.1 (14.1) Non-Fitters: 67.3 (14.1)	61	63.7 (12.4)	62.9 (13.4)	64.8 (1.7)
Sex	70 % male	38 % male	61 % male	Fitters: 49 % male Non-Fitters: 42 % male	56 % male	57 % male	59 % male	52 % male
Mean FM-initial score (SD)	37.8 (22.0)	26.4 (19.0)	17.5 (17.9)	25.6 (21.8)	37.1 (23.1)	31.9 (26.9)	4.44 (4.0)	26.7 (24.1)
Initial time point	First ~2 days	Within first 2 weeks	First 2-7 days	Within first 72 hours	First 2 weeks	Within first 2 to 4 weeks	Within first 5 to 10 days	First ~4 days
Mean FM-end	50.1 (17.8)	21.4 (21.4)	31.7 (23.6)	46.8 (21.9)	52.0 (20.3)	43.6 (24.6)	16.2 (16.6)	43.2 (24.7)

score (SD)								
End time point	~3 months	~3 months	~3 months	~6 months	~3 months	~3 months	90 days after initial	~3 months

Table 1. Descriptive statistics of each of the eight studies included. Patient age and sex were extracted as summary statistics from original publications, resulting in minor differences in the specification of uncertainty and fitter/non-fitter distinction. The stated number of patients, as well as mean *FM-initial* and *FM-end* scores may differ slightly from the information stated in the original publications due to our data extraction technique (c.f., **supplementary materials**).

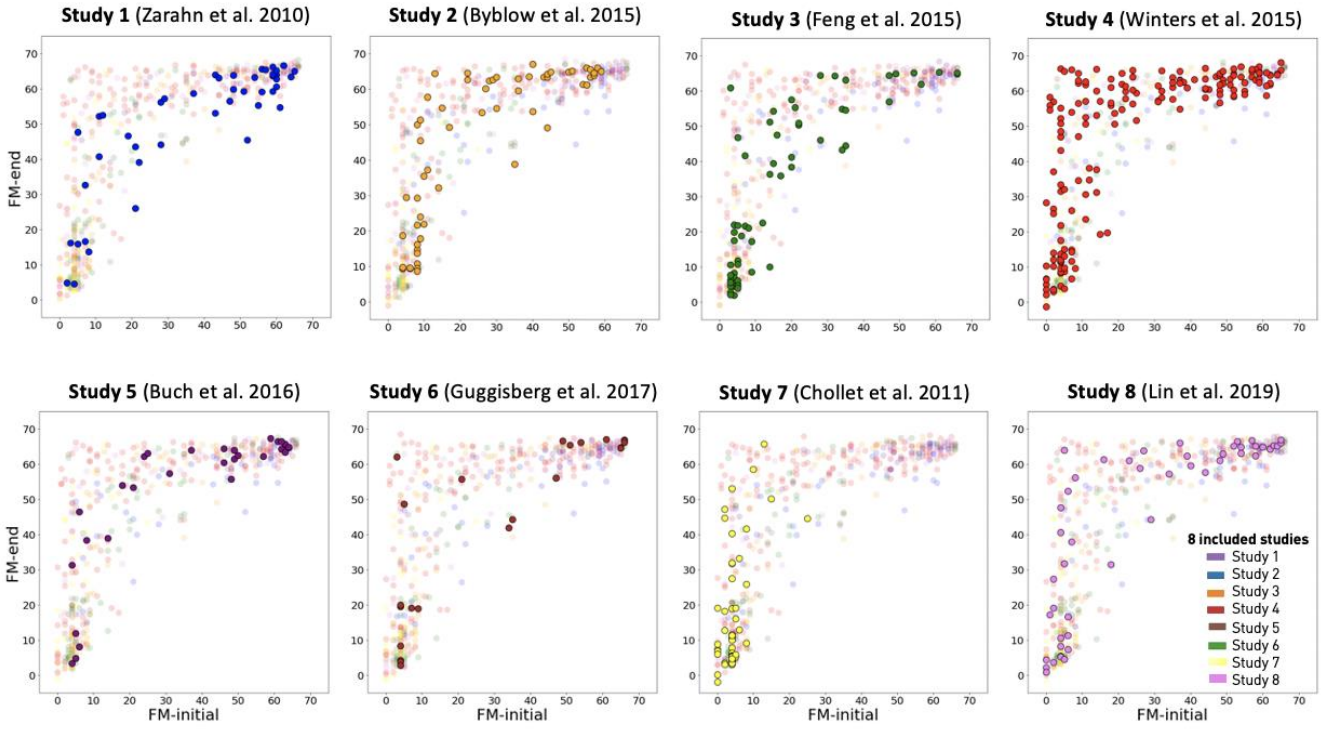

Figure 1. Motor recovery of 489 patients across the entire range of initial impairments. Plots display distributions of initial (*x-axis*) and end (*y-axis*) upper limb Fugl-Meyer Scores for all individual studies. While the respective study is highlighted in each plot, all other studies are shown in transparent colors for comparison. Altogether, we included eight studies comprising data on 489 patients in total.

Impairment-level-aware modelling of motor stroke outcome

To reduce confounding ceiling effects and facilitate the identification of more accurate recovery patterns, we employed the FM-initial cut-off established in [18] and considered all patients with a moderate to severe initial impairment only (i.e., *FM-Initial*<45, n=359). We additionally refrained from fitting the classic change models, i.e., linking initial impairment to recovery (*FM-end – FM-initial*), to avoid effects of mathematical coupling (c.f., [16,17] for in depth discussions). Instead, we predicted *FM-end* scores directly from *FM-initial* scores. Altogether, we constructed multi-level Bayesian hierarchical linear regression models with varying intercepts and slopes.[27] This hierarchical structure could address both the differences and similarities between the eight individual studies, as well as two patient groups (severely and non-severely affected patients). Thus, on the first level, we estimated full probability distributions of intercept and slope for severely and non-severely affected patients in each of the eight included studies. Therefore, we obtained 2x8 individual sets of parameters (two severity groups within eight different studies). On the intermediate level, we estimated two pairs of intercepts and slopes, this time characterizing recovery pattern from only severely and non-severely affected patients. Information originating from the various studies was thus pooled. The priors of severely and non-severely affected patients were eventually merged through joint hyperpriors for intercepts and slopes to complement the hierarchical model structure (c.f. **Figure 2B** for precise model specifications).

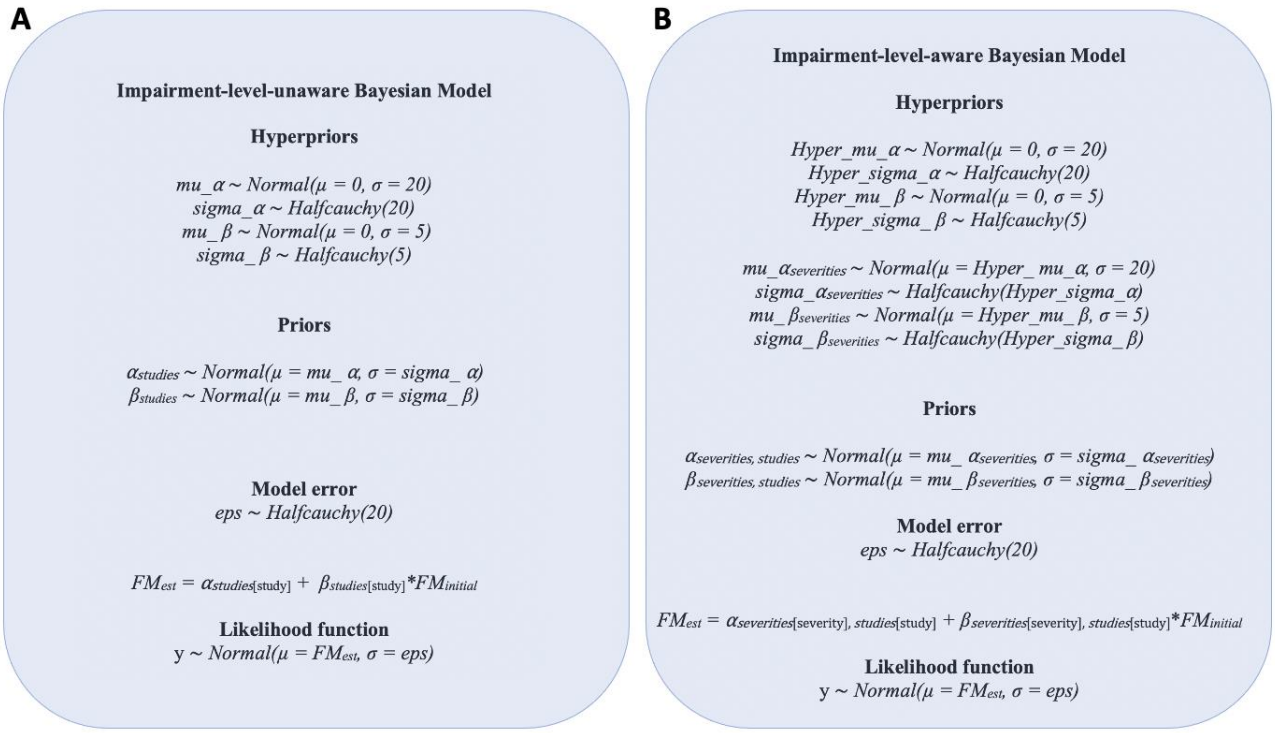

Figure 2. Bayesian hierarchical model specifications. A. Impairment-level-unaware Bayesian model. We here modelled eight pairs of intercept and slope priors, thus one pair each per integrated study. These individual priors were linked through joint hyperpriors. **B. Impairment-level-aware Bayesian models.** In contrast to the impairment-level-unaware model, we here modelled 2*8=16 pairs of intercept and slope priors, i.e., one pair for each severity group (severe vs. non-severe) for each of the eight different studies. These priors were linked to severity-specific hyperpriors, i.e., one pair of intercepts and slopes for each severity subgroup, which were eventually linked to overall hyperpriors for intercepts and slopes.

While an *FM-Initial*=10 has been chosen to differentiate between severe/non-severe patients in previous studies,[12] we here tested six breakpoints from *FM-Initial*=5 to *FM-Initial*=30 in steps of five to create patient subgroups. This approach resulted in six individual models that were fitted to the data. Additionally, we implemented an "impairment-level-unaware" model, which did not differentiate between the two groups of severely and non-severely affected patients (c.f., **Figure 2A**). We performed Bayesian model comparisons to evaluate whether recovery pattern differed between severely and non-severely affected patients or whether they were the same for stroke patients with any degree of initial impairment. Furthermore, we also assessed which breakpoint was the most suitable one to assign patients to severe/non-severe groups.

Notably, we deliberately focused on investigating whether recovery data was best explained by *only one* or *two* unique recovery patterns – especially once previous confounds such as ceiling effects and mathematical coupling were addressed. The existence of one versus two recovery patterns was the research question directly emerging from previous work that primarily separated patients into severely and less severely affected patients (i.e., *non-fitter v*s. *fitter*).[7,10,12] However, we neither tested the existence of more than two subgroups, nor whether the assumption of more than two subgroups enhanced our prediction performance further (c.f., our **Potential limitations** for further considerations).

Analyses were conducted in a nested cross-validation framework. For the outer loop, we first applied a leave-one-study-out scheme. Thus, we iteratively fitted the six impairment-level-aware models and the one impairment-level-unaware model in seven out of eight studies and performed a Bayesian model comparison to adjudicate between those seven models. The model comparison itself was based on leave-one-patient-out-cross-validation (LOOCV), which represented the inner loop.[28] The model that was assigned the highest weight in each of the eight model comparisons was lastly applied to the left-out, eighth study in each outer loop to obtain a reliable out-of-sample effect size estimate (here: R-squared). These estimates were averaged across all eight leave-one-study-out loops. It is important to note that the likelihood of overfitting was countered by employing this nested cross-validation scheme.[29]

As our focus was on the interpretation of recovery pattern in this analysis, we also re-fitted the impairment-level-unaware and the model with the highest assigned average weight to all eight studies at once and interpreted resulting posteriors of intercepts and slopes with respect to *proportional to lost, proportional to spared* and *constant* recovery (c.f. **supplementary materials**). To interpret results in this fashion, we transformed the fitted linear model that linked *FM-initial* with *FM-end*, to the previously most frequently used change form, linking *FM-initial* or *Potential (66 – FM-initial)* to recovery (*FM-end – FM-initial*) (c.f., **supplementary materials**).

Prediction of motor stroke outcome

In contrast to the first two analyses, we were here exclusively interested in the out-of-sample prediction and not in the extraction of the veridical recovery pattern, or quantification of the effect-size uncontaminated by ceiling or mathematical coupling (c.f., **Limitations** in the Discussion section for elaboration on this issue). We refrained from applying an upper boundary

on *FM-Initial,* instead including all patients (n=458), even those that were close to or at ceiling. We, once again, fitted hierarchical Bayesian linear regression models in a nested cross-validation fashion. We determined the optimal *FM-initial* breakpoint between groups of severely and non-severely affected patients in Bayesian model comparisons within inner cross-validation loops. The model ranked first in the model comparison was then fitted to the study left out in the outer cross-validation loop. The performance measure obtained was the averaged R-squared value across all eight left-out studies.

For reasons of comparison, we also determined the out-of-sample (out-of-study) performance for the "impairment-level-unaware" model that did not include a hierarchical level for severe/non-severe patients. Lastly, we computed the out-of-sample performance for the group often called "fitters",[7] i.e., we fitted our Bayesian model to all patients with an *FM-initial* score of more than 10 and calculated out-of-sample R-squared effect sizes. We excluded the control group participants by Chollet et al. 2011 in the "fitters" only analysis, since there were only three patients in the range of *FM-initial*>10.

Inference and statistical analysis

Samples from the posterior distributions of the model parameters were drawn by the No U-Turn Sampler (NUTS, draws=2000, tuning=1000).[30] The performance measure we used was the out-of-sample R-squared value.

Data availability

Data is available from the authors on reasonable request. Jupyter notebook scripts (python 3.7, predominantly pymc3)[31] will be made available: https://github.com/AnnaBonkhoff/to_be_added_upon_acceptance.

Results

Impairment-level-aware modelling of motor stroke outcome

The impairment-level-aware Bayesian model that differentiated patient subgroups (n=359 in total) based on an *FM-initial*=10, was assigned the overall highest model weight in the eight LOOCV-based model comparisons (average weight=0.56, 95% confidence interval (CI)=0.44 – 0.68). This model was furthermore ranked first in six out of eight LOOCV-based model comparisons. Additionally, if not ranked first, the estimate from the best performing model was always within the confidence interval of the difference of deviances (**Figure 3)**. Taking all of these results together, *FM-initial*=10 was the best breakpoint to model stroke outcome.

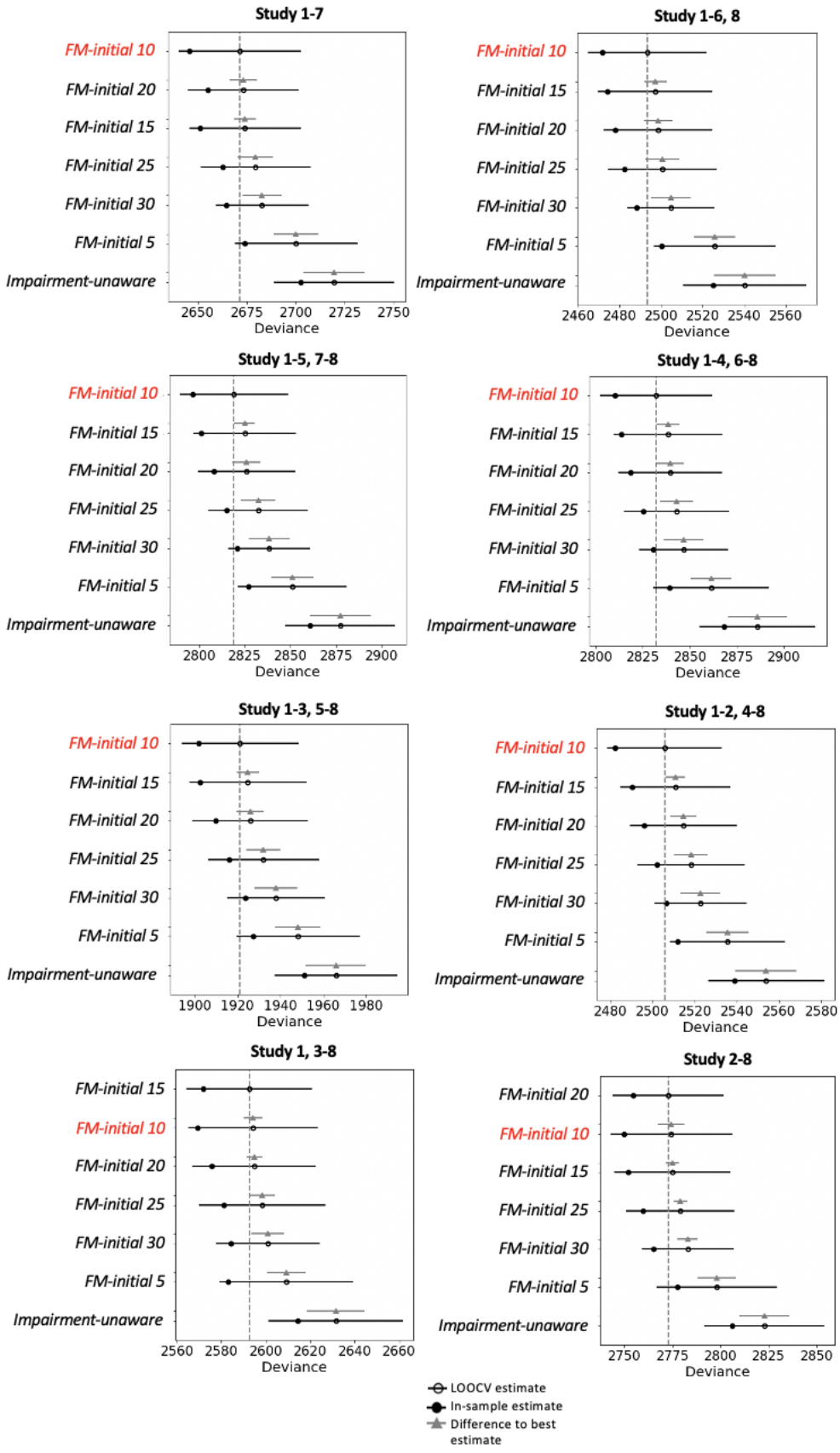
Study 1-7
FM-initial 10
FM-initial 20
FM-initial 15
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2650 2675 2700 2725 2750
Deviance
Study 1-6, 8
FM-initial 10
FM-initial 15
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2460 2480 2500 2520 2540 2560
Deviance
Study 1-5, 7-8
FM-initial 10
FM-initial 15
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2800 2825 2850 2875 2900
Deviance
Study 1-4, 6-8
FM-initial 10
FM-initial 15
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2800 2825 2850 2875 2900
Deviance
Study 1-3, 5-8
FM-initial 10
FM-initial 15
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
1900 1920 1940 1960 1980
Deviance
Study 1-2, 4-8
FM-initial 10
FM-initial 15
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2480 2500 2520 2540 2560 2580
Deviance
Study 1, 3-8
FM-initial 15
FM-initial 10
FM-initial 20
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2560 2580 2600 2620 2640 2660
Deviance
Study 2-8
FM-initial 20
FM-initial 10
FM-initial 15
FM-initial 25
FM-initial 30
FM-initial 5
Impairment-unaware
2750 2775 2800 2825 2850
Deviance
LOOCV estimate
In-sample estimate
Difference to best estimate

Figure 3. Bayesian model comparisons testing various initial Fugl-Meyer scale breakpoints: The model differentiating patients based on an initial Fugl-Meyer score of 10 was chosen as the most suitable model in the majority of cases and was also assigned the highest average model weight across all eight model comparisons. Model comparisons themselves were based on leave-one-out-cross-validation. The model that did not differentiate between patient groups with severe or non-severe initial impairment, but fitted a single regression line across all levels of impairment, consistently performed the worst.

Second and third highest ranked models were the ones relying on *FM-initial* breakpoints of 15 and 20, achieving average model weights of 0.19 (95%CI=0.12 – 0.26) and 0.16 (95%CI=0.08 – 0.24), respectively. The model that consistently performed the worst, i.e., being ranked last in all eight model comparisons, was the impairment-level-unaware Bayesian model. The cross-validated out-of-sample effect size, estimated for the highest ranked models in the left-out studies, was R-squared=47.4% (95%-CI=34.6% – 60.2%). Taken together, models that took the level of initial impairment into account, specifically with a breakpoint of *FM-initial*=10, performed best for modelling upper limb motor outcome after stroke across all eight studies. Overall, there were 214 patients with FM-initial scores below and 145 above a score of 10.

We re-fitted the impairment-level-unaware Bayesian model to all eight studies at once (**Figure 4A**) and could infer a combination of *constant* and *proportional to spared* recovery. Importantly, the interpretation with respect to constant and proportional recovery was possible only after transforming all formulations to their change form, i.e., exchanging *FM-end* as outcome with the change *FM-end – FM-initial* (c.f., **supplementary materials** for details). Nonetheless, in view of the model comparison results, the assumption of a fixed recovery pattern across patients of all initial impairment levels did not lead to competitive out-of-sample performance of motor outcome prediction. When instead interpreting the parameters of the winning model, i.e., the impairment-level-aware model with the *FM-initial*=10 breakpoint, we observed similarities, yet also marked differences in recovery patterns for patients with *FM-initial* scores below ten and those with scores above ten. Both patient groups were characterized by comparable amounts of constant recovery (8-9 points on the FM-scale). In patients with the highest degree of initial impairment (*FM-initial*≤10), this *constan*t recovery was however combined with a *proportional to spared* function recovery component (**Figure 4B**). Among these severely impaired patients, this pattern implied that reacquisition of function was greater the more function was preserved in the

first place. In contrast, patients with a more moderate initial impairment (*FM-initial*>10) featured a combination of *constant* and *proportional to lost* function recovery (c.f. **supplementary** material for explicit transformations to the change formulation). In their case, recovery was greater the higher their initial impairment (**Figure 4B**). Of note, we addressed potential biases, such as ceiling effects and mathematical coupling, by modelling *FM-end* instead of a change score and only considering those subjects with *FM-initial*<45.

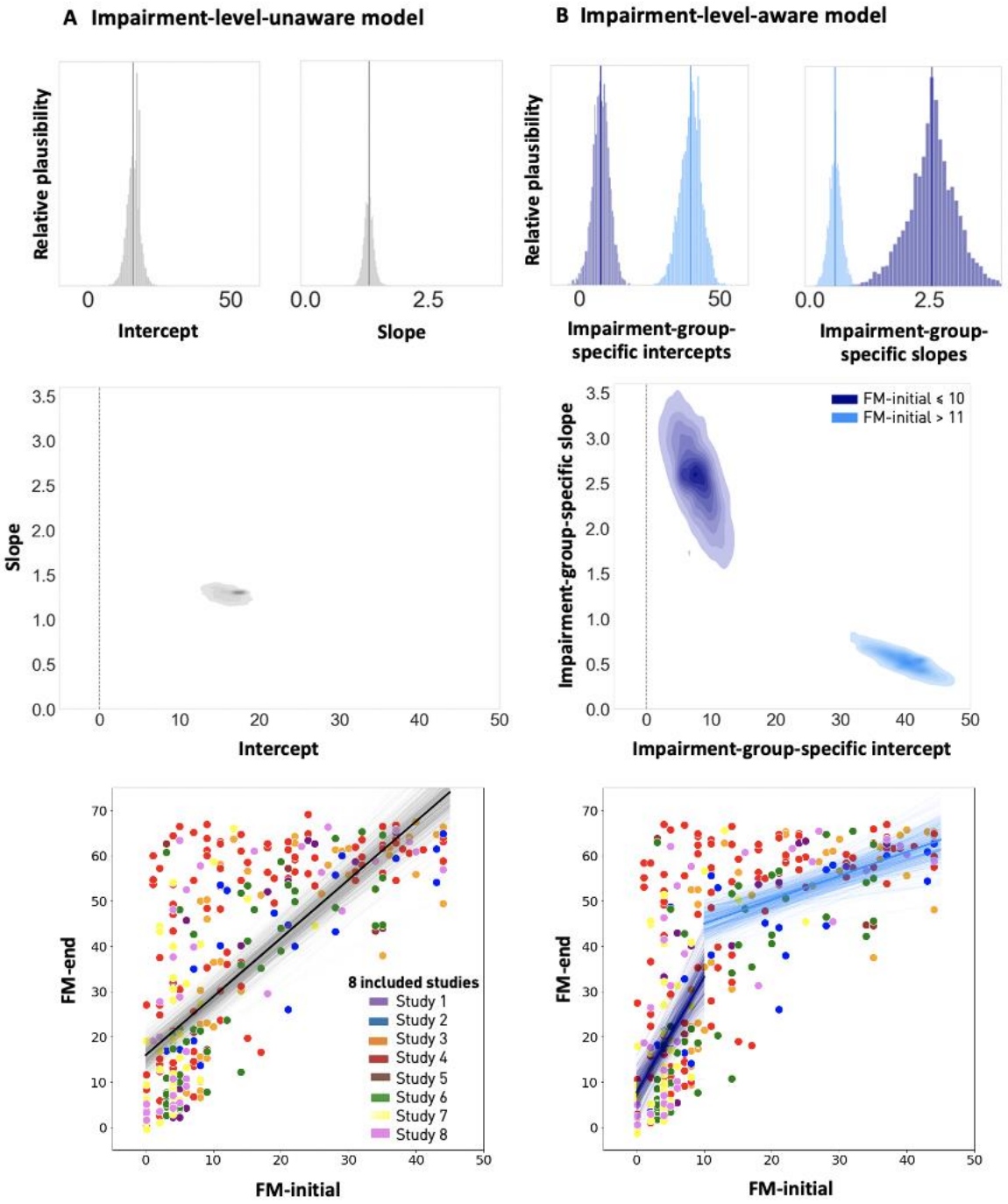

Figure 4. Motor stroke outcome is best explained by initial-impairment-dependent recovery patterns. A. Impairment-level-unaware model. The interpretation of intercepts and slopes of this model led to the conclusion of a combination of *constant* and *proportional to spared* recovery for the entirety of patients. However, this model was consistently outperformed by all impairment-level-aware models, indicating the benefit of assuming two subgroups within the entire patient cohort. **B. Impairment-level-aware model employing a breakpoint of *FM-initial*=10 to differentiate between severely and non-severely affected patients.** We here observed two distinct recovery patterns:

The patient group of initially severely impaired patients (FM-initial≤10, *dark-blue*) recovered according to a combination of *constant* and *proportional to spared* recovery pattern. However, the group with initial FM-scores above ten (*light-blue*) featured a combination of *constant* and *proportional to lost* recovery (c.f., **supplementary material** for transformations to the change formulation and clarification of combinations of recovery patterns).

In both A and B, the *upper row* illustrates the individual posterior distributions of intercepts and slopes, the *middle row* presents the joint density of both parameters, intercepts and slopes, and the *bottom row* highlights the actual fitted linear regression line (*thick line:* mean, *thinner lines:* 500 randomly sampled marginal posterior parameter fits). Individual patient data points, color-coded for the study of origin and slightly jittered on the vertical axis to reduce overlap, are shown in the background.

Prediction of motor outcome after stroke

Considered models (six impairment-level-aware models: *FM-initial* breakpoints: 5, 10, 15, 20, 25, 30; impairment-level-unaware model) were first fit to seven out of the eight studies, without excluding patients with mild initial impairments. Hence, we did not account for ceiling effects. The ensuing Bayesian model comparisons ranked the impairment-level-aware models with breakpoints *FM-initial*=15 five times and *FM-initial*=10 three times as the best performing ones. The highest ranked model after each of these model comparisons was then applied to the left-out eighth study to obtain an out-of-sample effect size. By these means, we could predict the individual Fugl-Meyer score in the chronic phase in the entire group of stroke patients (*FM-initial*=0-66) with an out-of-sample explained variance of 63.5% (R-squared, 95%CI: 51.4% - 75.5%).

In contrast, when instead exclusively focusing on the impairment-level-unaware model, the out-of-sample performance to predict the *FM-end* score of all patients across the entire impairment range amounted to only 32.1% explained variance (R-squared, 95%-CI: 15.2% - 49%). Since 95%-confidence intervals did not overlap, the impairment-level-aware model significantly outperformed the impairment-level-unaware model. When considering the group of *fitters* (*FM-initial*>10) only and fitting an impairment-level-unaware model (i.e., the typically used model, assuming the same recovery pattern for all mildly to moderately affected patients),[7] we obtained an R-squared out-of-sample prediction performance of 28.4% (95%-CI: 19.1% - 37.7%). These estimates have to be regarded with some caution, since the decrease in variability due to ceiling can underly an overestimation of the R-squared estimate of explained variance (c.f., **Limitations**).

Discussion

Motor recovery of moderately-to-severely affected patients was best modelled by assuming two distinct subgroups of patients. When simultaneously modelling and interpreting recovery trajectories for these severely and non-severely affected stroke patients, we found clear similarities and differences in recovery patterns. Both patient groups recovered a comparable constant amount in FM performance of their affected upper limb. However, severely affected patients (*FM-initial*≤10) additionally exhibited a *proportional to spared* component and non-severely affected patients (*FM-initial*>10) featured an additional *proportional to lost* recovery component. The explained variance of motor outcomes in unseen data was 49% after adjusting for confounds, such as ceiling effects. In further analyses, we predicted chronic motor impairment for patients across the entire spectrum of initial impairment and thus accepted the probable presence of ceiling effects due to the inclusion of patients with mild initial impairments. However, we could here generate predictions for all patients. We observed an out-of-sample prediction performance of an explained variance of 64%. These presented out-of-sample prediction performances may serve as benchmarks to evaluate additional, novel biomarkers of stroke recovery.

Interpretation of motor recovery pattern

Previous studies on recovery pattern had derived a "70% proportional recovery" for non-severely affected patients.[10,11,12,13,14,15] Nonetheless, these studies were markedly affected by confounds, such as ceiling effects and mathematical coupling.[16,17] Addressing these confounds in a prior study led to a substantial decrease in explained variance for the proportional recovery pattern in non-severely affected patients to only 32%.[18] In contrast, this study also showed that concurrently modelling severely and non-severely affected patients resulted in an (in-sample) explained variance of 53%.[18] This increase indicated a benefit of extracting recovery patterns across all degrees of initial impairments, instead of only non-severe stroke patients, which were the primary focus in previous studies.[10,11,12,13,14,15] However, this finding conflicted with the long-standing notion of distinctly varying recoveries for severely and non-severely affected stroke patients in view of assumed differences in biological starting positions.[9,19] In fact, our current analysis suggests that it is precisely the combination of both ideas: simultaneous modelling of severe and non-severe stroke patients, while allowing for impairment-level-dependent recovery

patterns, provides the best fit to upper limb motor impairment recovery data. The obtained out-of-sample R-squared of 49% in the current severe-non-severe analysis surpasses the reported 32% for non-severe patients.[18] Also, Bayesian model comparisons consistently suggested separate recovery patterns for severely and non-severely affected patients. Of note, we considered only moderately-to-severely affected patients in these analyses (FM<45) to ensure that subgroups did not artificially emerge due to ceiling effects and paid great attention to additionally avoid confounds due to mathematical coupling. With respect to the nature of recovery patterns after addressing these confounds affecting previous studies: severely and non-severely affected patients were both characterized by a similar constant amount in FM recovery, which we here determined to be ~8-9 points. However, they differed in the additional proportional recovery contribution. A *proportional to spared* component to recovery in severely affected patients was contrasted with a *proportional to lost* component to recovery in non-severely affected patients. *Proportional to spared* recovery implied a greater recovery in case of more preserved original motor function. Conceivably, for patients with initially severe motor impairments, any residual abilities to move limbs could facilitate the reacquisition of further motor abilities. The opposite relationship was true for *proportional to lost*, in which a patient recovered more, the greater the initial impairment. For patients with moderate motor impairments, neural injury is likely to be milder. Likewise, functional MRI data indicated categorical differences in cerebral organization after severe and non-severe stroke, which furthermore suggests the existence of impairment-specific recovery processes.[32] The varying types of proportional recovery that our analyses uncover (i.e. *proportional to spared* and *proportional to lost) remain to be explained*; we speculate that they might reflect the action of biologically distinct recovery processes. Our results would also be in line with the assumption that impairment-adapted neurorehabilitation intervention strategies may be most effective: an individual with severe motor impairment may benefit from a different neurorehabilitation strategy than an individual with more moderate impairment. Our present work does not allow for direct conclusions on which specific kinds of neurorehabilitation treatment may be more fruitful for the more or less severe initial impairment category. Rather, our findings generally motivate the inclusion of both more and less severely affected patients in rehabilitation trials *and,* importantly, an initial impairment-specific analysis. We would hypothesize that some specific (e.g., pharmaceutical) therapies could have a substantially stronger effect in only severe or only non-severe stroke, perhaps improving recovery beyond that predicted by the proportional

recovery rule. In other words, initial impairment severity may well confound the measurement of therapy effects. Importantly, our insights were gleaned from analyses that enabled and exploited the inclusion of severely affected patients, which could eventually decrease the excess of conducted research on only non-severely affected patients.[33]

Prediction

Accurate models and predictions of recovery promise both better information for patients to plan for the future, and more efficient trials of therapies that might accelerate recovery.[9] However, to date, FM-based recovery studies mainly report in-sample effect sizes.[10,11,13] Thus, models were fitted and evaluated by considering exactly the same sample of patients. Effect sizes originating from these evaluations are likely to be too high and estimates cannot be taken as faithful predictions for unseen, new patients.[3] A recent study of recovery after stroke represents an important exception, as cross-validated, i.e., out-of-sample, performance estimates were computed.[34] This study deviates from the classic recovery work[7] by incorporating not only a single initial motor impairment measurement, but several measurements, i.e. a time series, over the first few weeks to predict the final motor outcome. The consideration of multiple baseline scores can conceivably augment the overall prediction performance by increasing the available information per patient. However, we here chose to rely on an approach that is closer to the original proportional recovery studies[7] and likely clinically more feasible by requiring only one initial motor impairment score. The prediction models presented here relied on only a single, first FM measurement as input and did not necessitate repeated examinations (other than to obtain the final FM score).

Altogether, we present two such prediction models and their out-of-sample performance estimates of long-term motor outcomes. On the one hand, we excluded mildly affected patients to reduce the confounding impact of ceiling. We therefore built prediction models in a subset of the entire data sample. We obtained an out-of-sample explained variance of 49%. On the other hand, we accepted the presence of confounding effects and repeated these analyses in the entire dataset to generate an out-of-sample estimate for all patients across the entire range of initial impairments. Explained variance was at 64%. Our cross-validation scheme to obtain out-of-sample estimates of explained variance addressed potential inflation due to models overfitting to data in both cases. However, the increase in explained variance, from 49% in the first to 64% in the second prediction

model, may be attributed, at least partially, to ceiling-induced inflation.[18,35] Eventually, both of our performance estimates may represent helpful baselines to which models employing novel biomarkers can be compared.

Potential limitations

A first limitation can be seen in the prior decision to only test two subgroups of stroke patients, rather than to extend this evaluation to an arbitrary number of patient subgroups or even estimate the number of subgroups from the data themselves. We, however, opted for two impairment-severity defined groups, instead of three or more, for predominantly two reasons: Firstly, most of the previous literature differentiates between two groups of *fitters* and *non-fitters* (i.e., non-severely and severely affected patients).[7,10,12] Additionally, electrophysiological evaluations of corticospinal tract integrity suggest precisely two recovery-relevant subgroups of transcranial magnetic stimulation positive and negative patients.[36] Hence, our resulting primary aim was to explicitly test whether we could find more than one recovery pattern – even after addressing confounds affecting the majority of previous recovery studies – and if so, what breakpoint between subgroups explained our recovery data the best. On the other hand, investigating whether our recovery data could be explained still better by assuming more than those two subgroups was beyond the scope of the current work. Future work might aim to combine the mixture model approach, as recently presented in studies by van der Vliet and colleagues, as well as Selles and colleagues,[34,37] and our hierarchical linear regression approach, as presented here. In this way, it may become possible to both automatically estimate the optimal number of subgroups[34,37] and to interpret recovery patterns with respect to constant and proportional recovery contributions. Another limitation can be seen in the decision to explicitly focus on the FM scale, an impairment-based scale of motor synergies. While the FM scale is a clinically frequently employed and thus impactful scale, it has to be additionally investigated whether results generalize to further measures of motor function, such as the Action Research Arm, Box and Block or Wolf Motor Function tests. Performance in several of these tests may even be combined to more faithfully capture the full range of individual patients' motor abilities and decrease ceiling effects. Such a strategy was already shown to be effective in the case of aphasia,[38] where some clinical tests appear to be affected by ceiling effects as well.[39] Similarly, it may be beneficial to extend the granularity and dynamic range of measures by adopting dynamic staircase methodologies. Given

that we considered continuous outcome scores in a linear regression scenario, our results are also not immediately comparable to classification analyses, that have, for example, tested the capacity of non-linear algorithms.[36] The move away from linearity assumptions is also important, since, as is evident in scatter plots of *FM-initial* to *FM-end*, the data being fit is heteroscedastic.[7] Therefore, future studies are warranted to directly contrast linear and non-linear modelling approaches. We chose to model *FM-end* instead of the more classic *Change*, *FM-end – FM-initial*, as our outcome variable. While the intercept and slope coefficients, as estimated in our linear regression framework, cannot directly be interpreted in the framework of proportional recovery, transformation to the *Change* formula is straight-forward and possible without any loss of information (c.f., **supplementary materials**). Lastly, while we here optimized the dataset size to increase the generalizability of our findings, the heterogeneity of our patient collective – that combines eight different studies that featured endpoints between three and six months and thus different time frames to recover – may be considered a limitation of the current study. It may thus be a future aim to conceptualize large-scale recovery studies that harmonize sample characteristics and data acquisition.

Conclusion

We here present a novel Bayesian hierarchical modelling framework for concurrently predicting motor outcome in severely and non-severely affected stroke patients. This approach of including both severely and non-severely affected patients in the same modelling framework can potentially help to decrease the current excess of research conducted only on non-severely affected patients by motivating further similar modelling analyses.[33] All in all, we here i) inferred that there really are two distinct recovery patterns for stroke patients with different initial motor impairment severity levels. Thus, the frequent practice of viewing severely and non-severely impaired patients as distinct groups with individual recovery trajectories is supported, even when previous confounds are addressed. We furthermore ii) established *out-of-sample* motor outcome prediction for unseen ischemic stroke patients for the simplest case of recovery models relying on only one initial FM-score.

Funding

D.B. was supported by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada, National Institutes of Health (NIH R01 AG068563A), the Canadian Institute of Health Research (CIHR 438531), the Healthy Brains Healthy Lives initiative (Canada First Research Excellence fund), Google (Research Award), and by the CIFAR Artificial Intelligence Chairs program (Canada Institute for Advanced Research). A.G.G. was supported by the Swiss National Science Foundation, CRSII5-170985. C.G. is in part funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 431549029 – SFB 1451 projects B05 and C05.

Conflicts of interest

None.

References

1. Kelly-Hayes M, Beiser A, Kase CS, Scaramucci A, D'Agostino RB, Wolf PA. The influence of gender and age on disability following ischemic stroke: the Framingham study. *Journal of Stroke and Cerebrovascular Diseases*. 2003;12(3):119-126. doi:10.1016/S1052-3057(03)00042-9

2. Shmueli G. To Explain or to Predict? *Statistical Science*. 2010;25(3):289-310. doi:10.1214/10-STS330

3. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol 112. Springer; 2013.

4. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage*. 2017;145:137-165.

5. Bzdok D, Varoquaux G, Steyerberg EW. Prediction, not association, paves the road to precision medicine. *JAMA psychiatry*. Published online 2020.

6. Bzdok D, Engemann D, Thirion B. Inference and Prediction Diverge in Biomedicine. *Patterns*. Published online October 2020:100119. doi:10.1016/j.patter.2020.100119

7. Prabhakaran S, Zarahn E, Riley C, et al. Inter-individual Variability in the Capacity for Motor Recovery After Ischemic Stroke. *Neurorehabilitation and Neural Repair*. 2008;22(1):64-71. doi:10.1177/1545968307305302

8. Winters C, Heymans MW, van Wegen EE, Kwakkel G. How to design clinical rehabilitation trials for the upper paretic limb early post stroke? *Trials*. 2016;17(1):468.

9. Boyd LA, Hayward KS, Ward NS, et al. Biomarkers of stroke recovery: consensus-based core recommendations from the stroke recovery and rehabilitation roundtable. *International Journal of Stroke*. 2017;12(5):480-493.

10. Zarahn E, Alon L, Ryan SL, et al. Prediction of Motor Recovery Using Initial Impairment and fMRI 48 h Poststroke. *Cerebral Cortex*. 2011;21(12):2712-2721. doi:10.1093/cercor/bhr047

11. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity: Proportional Recovery After Stroke. *Annals of Neurology*. 2015;78(6):848-859. doi:10.1002/ana.24472

12. Feng W, Wang J, Chhatbar PY, et al. Corticospinal tract lesion load: An imaging biomarker for stroke motor outcomes: CST Lesion Load Predicts Stroke Motor Outcomes. *Annals of Neurology*. 2015;78(6):860-870. doi:10.1002/ana.24510

13. Winters C, van Wegen EEH, Daffertshofer A, Kwakkel G. Generalizability of the Proportional Recovery Model for the Upper Extremity After an Ischemic Stroke. *Neurorehabilitation and Neural Repair*. 2015;29(7):614-622. doi:10.1177/1545968314562115

14. Buch ER, Rizk S, Nicolo P, Cohen LG, Schnider A, Guggisberg AG. Predicting motor improvement after stroke with clinical assessment and diffusion tensor imaging. *Neurology*. 2016;86(20):1924-1925. doi:10.1212/WNL.0000000000002675

15. Guggisberg AG, Nicolo P, Cohen LG, Schnider A, Buch ER. Longitudinal structural and functional differences between proportional and poor motor recovery after stroke. *Neurorehabilitation and neural repair*. 2017;31(12):1029-1041.

16. Hawe RL, Scott SH, Dukelow SP. Taking Proportional Out of Stroke Recovery. *Stroke*. 2019;50(1):204-211.

17. Hope TM, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. *Recovery after Stroke: Not so Proportional after All?* Oxford University Press; 2018.

18. Bonkhoff AK, Hope T, Bzdok D, et al. Bringing proportional recovery into proportion: Bayesian modelling of post-stroke motor impairment. *Brain*. Published online June 29, 2020. doi:10.1093/brain/awaa146

19. Stinear CM. Prediction of motor recovery after stroke: advances in biomarkers. *The Lancet Neurology*. 2017;16(10):826-836.

20. 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. *International Journal of Stroke*. 2017;12(5):444-450. doi:10.1177/1747493017711816

21. Ward NS. Restoring brain function after stroke—bridging the gap between animals and humans. *Nature Reviews Neurology*. 2017;13(4):244.

22. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press; 2006.

23. Chollet F, Tardy J, Albucher J-F, et al. Fluoxetine for motor recovery after acute ischaemic stroke (FLAME): a randomised placebo-controlled trial. *The Lancet Neurology*. 2011;10(2):123-130. doi:10.1016/S1474-4422(10)70314-8

24. Lin DJ, Cloutier AM, Erler KS, et al. Corticospinal tract injury estimated from acute stroke imaging predicts upper extremity motor recovery after stroke. *Stroke*. 2019;50(12):3569-3577.

25. Lin DJ, Erler KS, Snider SB, et al. Cognitive Demands Influence Upper Extremity Motor Performance During Recovery From Acute Stroke. *Neurology*. Published online 2021.

26. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity: Proportional Recovery After Stroke. *Annals of Neurology*. 2015;78(6):848-859. doi:10.1002/ana.24472

27. Gelman A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*. 2006;48(3):432-435.

28. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 2017;27(5):1413-1432.

29. Hosseini M, Powell M, Collins J, et al. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*. 2020;119:456-467. doi:10.1016/j.neubiorev.2020.09.036

30. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*. 2014;15(1):1593-1623.

31. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*. 2016;2:e55.

32. Bonkhoff AK, Espinoza FA, Gazula H, et al. Acute ischaemic stroke alters the brain's preference for distinct dynamic connectivity states. *Brain*. Published online May 1, 2020. doi:10.1093/brain/awaa101

33. Hayward KS, Schmidt J, Lohse KR, et al. Are we armed with the right data? Pooled individual data review of biomarkers in people with severe upper limb impairment after stroke. *NeuroImage: Clinical*. 2017;13:310-319.

34. van der Vliet R, Selles RW, Andrinopoulou E-R, et al. Predicting upper limb motor impairment recovery after stroke: a mixture model. *Annals of Neurology*. Published online 2020.

35. Bowman H, Bonkhoff A, Hope T, Grefkes C, Price C. Inflated Estimates of Proportional Recovery From Stroke: The Dangers of Mathematical Coupling and Compression to Ceiling. *Stroke*. Published online 2021:STROKEAHA. 120.033031.

36. Stinear CM, Byblow WD, Ackerley SJ, Smith M-C, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. *Annals of clinical and translational neurology*. 2017;4(11):811-820.

37. Selles RW, Andrinopoulou E-R, Nijland RH, et al. Computerised patient-specific prediction of the recovery profile of upper limb capacity within stroke services: the next step. *J Neurol Neurosurg Psychiatry*. Published online January 21, 2021:jnnp-2020-324637. doi:10.1136/jnnp-2020-324637

38. Schumacher R, Bruehl S, Halai AD, Lambon Ralph MA. The verbal, non-verbal and structural bases of functional communication abilities in aphasia. *Brain Communications*. 2020;2(2):fcaa118. doi:10.1093/braincomms/fcaa118

39. Conroy P, Sotiropoulou Drosopoulou C, Humphreys GF, Halai AD, Lambon Ralph MA. Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain*. 2018;141(6):1815-1827. doi:10.1093/brain/awy087