# Bringing Proportional Recovery into Proportion: Bayesian Modelling of Post-Stroke Motor Impairment

Anna K. Bonkhoff[1,2,3], Thomas Hope[4], Danilo Bzdok[5,6,7], Adrian G. Guggisberg[8], Rachel L. Hawe[9], Sean P. Dukelow[9], Anne K. Rehme[1], Gereon R. Fink[1,2], Christian Grefkes[1,2], and Howard Bowman[10, 11]

[1] Department of Neurology, University Hospital Cologne, Cologne, Germany

[2] Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Juelich, Juelich, Germany

[3] Queen Square Institute of Neurology, University College London, London, UK

[4] Wellcome Centre for Human Neuroimaging, University College London, UK

[5] Mila – Quebec Artificial Intelligence Institute, Montreal, Canada

[6] Department of Biomedical Engineering, McConnell Brain Imaging Centre, Montreal Neurological Institute, Faculty of Medicine, McGill University, Montreal, Canada

[7] Canadian Institute for Advanced Research (CIFAR)

[8] Clinical Neuroscience, University of Geneva, Medical School, 1202 Geneva, Switzerland

[9] Department of Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Alberta, Canada

[10] School of Psychology, University of Birmingham, UK

[11] School of Computing, University of Kent, UK

## Abstract

Accurate predictions of motor impairment after stroke are of cardinal importance for the patient, clinician, and health care system. More than ten years ago, the proportional recovery rule was introduced by promising just that: high-fidelity predictions of recovery following stroke based only on the initially lost motor function, at least for a specific fraction of patients. However, emerging evidence suggests that this recovery rule is subject to various confounds and may apply less universally than previously assumed.

Here, we systematically revisited stroke outcome predictions by applying strategies to avoid confounds and fitting hierarchical Bayesian models. We jointly analyzed $n$=385 post-stroke trajectories from six separate studies – one of the currently largest overall datasets of upper limb motor recovery. We addressed confounding ceiling effects by introducing a subset approach and ensured correct model estimation through synthetic data simulations. Subsequently, we used model comparisons to assess the underlying nature of recovery within our empirical recovery data.

The first model comparison, relying on the conventional fraction of patients called *fitters*, pointed to a combination of proportional to lost function and constant recovery. Proportional to lost here describes the original notion of proportionality, indicating greater recovery in case of a more severe initial impairment. This combination explained only 32% of the variance in recovery, which is in stark contrast to previous reports of >80%. When instead analyzing the complete spectrum of subjects, *fitters and non-fitters*, a combination of proportional to spared function and constant recovery was favoured, implying a more significant improvement in case of more preserved function. Explained variance was at 53%.

Therefore, our quantitative findings suggest that motor recovery post-stroke may exhibit some characteristics of proportionality. However, the variance explained was substantially reduced compared to what has previously been reported. This finding motivates future research moving beyond solely behavior scores to explain stroke recovery and establish robust and discriminating single-subject predictions.

## Keywords

Motor outcome post-stroke, Proportional recovery, Bayesian hierarchical models, Bayesian model comparison, learning from data

## Abbreviations

FM – Fugl-Meyer

PPC – Posterior predictive check

## Introduction

The science of clinical recovery after stroke, began with comprehensive yet mainly anecdotal descriptions of patients' trajectories (Twitchell, 1951; Newman, 1972; G. Broeks, *et al.*, 1999). It then moved to increasingly larger studies aiming to create robust prediction models for individual outcome. Initial impairment status crystallized as one of the most predictive features, providing the foundation of the proportional recovery rule (Prabhakaran *et al.*, 2008). According to this widespread rule, the majority of stroke patients, considered *fitters* to the rule, recover about 70% of the initially lost function[1] within the first few months after the initial event. *Fitters* and *non-fitters* to this proportional (to lost function) recovery rule have been defined in several ways in previous studies. For example, by choosing a discrete initial cut-off score, clustering initial and follow-up scores, or utilizing measures of corticospinal tract integrity. The proportional recovery rule, developed initially for Fugl-Meyer (FM) assessment scores of the upper limb (Kundert *et al.*, 2019), has since been extended to various functional domains. Numerous studies on recovery post-stroke claim to confirm proportional (to lost function) recovery of the upper limb (Zahran et al., 2011; Byblow *et al.*, 2015), the lower limb (Smith *et al.*, 2017), language (Marchi *et al.*, 2017), and neglect (Winters *et al.*, 2017). Collectively, these studies consistently report high values of

[1] "Function" here refers to the term "body function" defined as "physiological functions of body systems" in the International Classification of Functioning, Disability and Health (ICF) (Organization, 2001). "Lost function" therefore describes the motor impairment on a scale such as the Fugl-Meyer, with decreasing body function implying increasing impairment.

explained variance in cumulatively more than 500 participants, even as high as 94% (Winters *et al.*, 2015).

Very recently, doubt has been placed on these estimates for explaining the variance observed in recovery post-stroke (Hawe *et al.*, 2019a; Hope *et al.*, 2019). The concerns relate to the problem of mathematical coupling, when correlating an initial score and the amount of change (Lord, 1956; Hayes, 1988; Chiolero *et al.*, 2013). This coupling confound may occur when the second (*End*) measurement has considerably less variability than the first (*Initial*) measurement, leading to a small ratio of *End* to *Initial* variabilities. Such small *variability ratios* of *End* to *Initial* measurements arise naturally in stroke recovery data based on the Fugl-Meyer assessment (Gladstone *et al.*, 2002), as ceiling effects predominantly occur at the *End* time-point and cause a reduction in variability.

Importantly, if this *variability ratio* of *End* to *Initial* measurements is small, the true relationship between *Initial* and *End* is irrelevant – we will, without question, find overwhelming evidence for a strong correlation between *Initial* and change measurements (**Figure 1[A]**). More concretely, correlations between the *Initial* measurement and the corresponding change score will automatically be high, as the change score is dominated by the *Initial* measurement in case of low variability of the *End* measurement due to ceiling effects. For example, let us assume three patients with *Initial* FM scores of 10, 35 and 50. All of them recover completely and we measure *End* FM scores of 66 in the chronic phase. Thus, there is no variability in their *End* score. Their change between *End* minus *Initial* measurements equals 56, 31 and 16 and correlates perfectly ($r$=-1.0) with the *Initial* measurement. Hence, this implies that testing proportional (to lost function) recovery in case of small *variability ratios* is tautological; it will always hold. Because it is central to our argument, we name the two confounds of mathematical coupling and ceiling effects, induced by small *variability ratios* and the impact of concentration of data towards ceiling, *compression enhanced coupling*.

Having identified this *compression enhanced coupling* confound, it is essential to consider whether it can be circumvented to enable an accurate assessment of the proportional recovery question. This is what we aimed to do in this article. Our logic was as follows.

1) The nature of recovery post-stroke cannot be meaningfully evaluated when there is a substantial ceiling effect at the second time-point (causing *compression enhanced coupling*).

2) By reducing data at ceiling, we can increase the *variability ratio* and address *compression enhanced coupling*. We additionally make sure that we do not incur any new confounds, when decreasing ceiling.

3) Once this confound has been handled, we can fit various recovery models and determine which one explains the data best, in order to assess the underlying mechanisms of stroke recovery.

Specifically, we first showed that estimates of explained variance for recovery were inflated, when models were fit to the entire sample of *fitters* (**Figure 1[B]**). This is the approach currently used in the literature. This inflation was expected on the basis of Hawe *et al.* (2019a) and Hope *et al.* (2019). We then reduced ceiling effects and *compression enhanced coupling* by creating *subsets*, i.e., we excluded a varying range of stroke participants with the highest scores at the *Initial* time-point (**Figure 1[C,D]**). This procedure reduced the ceiling effect at the *End* time-point and therefore also increased the *variability ratio*. Critically, we validated this subset procedure in synthetic data experiments (Gelman and Hill, 2006). That is, we generated data with known ground truth and simulated three candidate explanations of recovery post-stroke: A) *proportional to lost function*; B) *proportional to spared function*; and C) *constant* recovery (**Figure 2[A,B,C]**). *Proportional to lost function* is the familiar pattern (referred to simply as proportional recovery in the literature), where a more severe initial impairment implies a more significant recovery. *Proportional to spared function* is the opposite pattern, in which individuals recover more if they have more preserved function at the *Initial* time-point. *Constant* recovery formalizes the idea that initial severity has no impact on recovery, which is the same size, whatever the initial impairment. These three patterns are different in all but one case, which is when there is no recovery (i.e., Y=X). We, then, imposed a ceiling in our synthetic data simulations and successively created subsets by excluding subjects with the highest *Initial* scores. Since the ground truth was available in these simulations, we were able to ensure our subset approach really reduced *compression enhanced coupling* (**Figure 1[A]**).

Subsequently, we assessed recovery patterns in empirical stroke data. Therefore, we aggregated a substantial body of data, i.e., 385 individual post-stroke recoveries, across a range of representative studies focused on upper limb deficits measured as Fugl-Meyer scores (Buch *et al.*, 2016; Byblow

*et al.*, 2015; Feng *et al.*, 2015; Guggisberg *et al.*, 2017; Winters *et al.*, 2015; Zarahn *et al.*, 2011). We employed state-of-the-art hierarchical Bayesian models, enabling us to incorporate data from these various sources, while accounting for inter-study variability and fully modelling the uncertainty in the data. Crucially, these models also permitted conducting overall model comparisons, enabling us to assess the evidence in the data for each model. We considered the three change models mentioned before and shown in **Figure 2[A,B,C]** – *proportional to lost*, *proportional to spared*, and *constant recovery*. We also included a completely unconstrained standard-form regression, the most general of the models, which determined whether linear relationships outside our three candidate models explained the data any better (**Figure 2[G,H,I]**). While we first only considered patients who adhere to the conventional proportional recovery rule, i.e., *fitters*, we later extended the analyses to the full spectrum of *fitters & non-fitters* **(Figure 1[C,D]).**

Consequently, the core objective of this paper was to respond to the confounded nature of assessments of behavioral recovery from stroke, particularly from upper-limb impairments. We did so by applying a subset approach to reduce confounding effects. We then employed Bayesian models and model comparisons to answer the substantive scientific question: what mechanisms best explain the data on recovery of upper-limb impairment after stroke and with what explained variance?

## Material & Methods

### Participants and clinical data

The analyses of post-stroke upper limb impairment were based on a sample of 385 acute stroke participants originating from six different studies on stroke recovery (Buch *et al.*, 2016; Byblow *et al.*, 2015; Feng *et al.*, 2015; Guggisberg *et al.*, 2017; Zarahn *et al.* 2011). Details on data acquisition are given in the **supplementary material, section 3**. In brief, we used anonymized data available from Zarahn *et al.* (2011) and Guggisberg *et al.* (2017) and combined it with secondary data from (Hawe *et al.*, 2019*a*). Therefore, we had individual-level information on Fugl-Meyer (FM) scores assessing upper limb motor impairment in the acute as well as chronic stage

(three to six months after the event; Nakayama *et al.*, 1994). A minimum score of 0 implies no preserved and 66 maximal body function (Fugl-Meyer *et al.*, 1975). In line with previous research (Feng *et al.*, 2015), we split the stroke subject samples into *fitters* and *non-fitters* to the classic proportional recovery rule (Prabhakaran *et al.*, 2008) based on their initial scores (Non-Fitters: *FM-Initial*≤10 points, Fitters: *FM-Initial*>10 points, **Figure 1[B]**). The first set of analyses were focused exclusively on *fitters* (n=243), and subsequent analyses highlighted findings on the entire sample, i.e., *fitters* and *non-fitters* (n=385). As all of the data has been published previously, ethics approvals had been granted for all individual primary studies.

## Bayesian hierarchical modelling of motor stroke outcome

A hierarchical Bayesian framework was employed to allow for the balanced incorporation of data from various sources and facilitate model comparisons. More precisely, we built Bayesian multilevel (hierarchical) linear regression models with varying intercepts and slopes (Gelman, 2006). Therefore, each of the six considered studies was characterized by estimated full probability distributions of intercept and slope parameters – rather than simple best-fit, maximum likelihood parameter estimates as usually employed in recovery studies. Retaining study-specific information in this statistical way was essential to addressing potential differences between studies. These differences could arise from independent data collection, involving study sites in different countries, and likely minor variations in therapy regimens. Nonetheless, given that each of the included studies considered similar measures at similar time points from broadly similar participants, information was also pooled across the various studies and a set of hyperparameters, i.e., across-study intercept and slope, was derived (Bzdok *et al.*, 2020). Thus, intercepts and slopes had two levels that carefully captured across-study versus individual variation in the eligible stroke studies.

The outcome variables that we sought to predict were either the raw *FM-end* score or *Change* (i.e., *FM-end – FM-initial*). We thus created a likelihood function for the outcome and linked it to the priors of our predictor variables, i.e., either the unaltered *FM-initial* score or *Potential* (*FM-maximum – FM-initial*), through one of five different models:

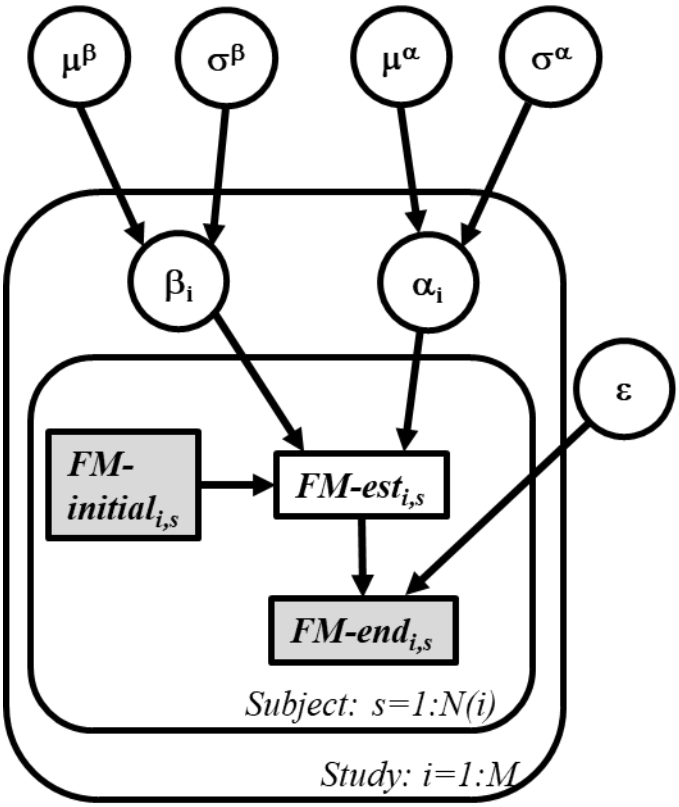
**The (classical) standard-form regression model:** *FM-end = b.FM-initial + a* (with intercept *a* and slope *b*)

**The change model:** *Change = Potential.B + A*, with *Change = FM-end – FM-initial* and *Potential = FM-maximum – FM-initial*, with FM-maximum = 66.

The change model was more precisely framed in three different ways expressing various conceivable recovery models, which we highlighted in the Introduction (c.f. panels [A,B,C] of **Figure 2**): The **classical proportional to lost function recovery model:** *Change = Potential.B*, a **proportional to spared function recovery model:** *Change = FM-initial.B* (which takes the raw initial score), and a **constant recovery model:** *Change = A*. Once fitted, we determined whether the obtained models were truly *proportional to lost, spared*, or *constant* recovery by assessing the parameter settings, where $0 \leq B < 1$ for the *proportional to lost* or *spared function* recovery and where $0 \leq A < Max \wedge Max = 66$ for the *constant* recovery model. We used lower case *a* and *b* to denote intercepts and slopes in standard-form regression models, which featured *FM-end* as outcome. Conversely, upper case A and B represented intercept and slope in change models, which had *Change (FM-end – FM-initial)* as outcome.

A critical step in any Bayesian analysis is the specification of prior beliefs. We attenuated the effects of priors and simultaneously increased the influence of the actual data by choosing simple, weakly informative Gaussian priors (i.e., with large standard deviations) for slope and intercept (hyper-)parameters and half-Cauchy priors for corresponding variance terms.

The full Bayesian model is specified in **Figure 5A**: Dependencies between variables are indicated with arrows, observed variables are in grey boxes, and the distributions defining variables are shown on the right. *M* denotes the number of studies analyzed and *N(i)* the number of subjects in each study.




Hyper-priors (across-studies)

$\mu^{\beta} \sim Normal(0,5)$ slope

$\sigma^{\beta} \sim Halfcauchy(5)$ std

$\mu^{\alpha} \sim Normal(0,20)$ intercept

$\sigma^{\alpha} \sim Halfcauchy(20)$ std

Priors (study-specific)

$\beta_i \sim Normal(\mu^{\beta}, \sigma^{\beta})$ slopes

$\alpha_i \sim Normal(\mu^{\alpha}, \sigma^{\alpha})$ intercepts

Model error (standard deviation)

$\varepsilon \sim Halfcauchy(20)$

Regression Equation

$FM\text{-}est_{i,s} = FM\text{-}initial_{i,s} \times \beta_i + \alpha_i$

Likelihood function

$FM\text{-}end_{i,s} \sim Normal(FM\text{-}est_{i,s}, \varepsilon)$

std: standard deviation

*Inference*: The analytical derivation of posterior distributions is either computationally very expensive and challenging or not possible, as it requires the integration over thousands of unknown parameters. We thus deployed a recent Monte Carlo Markov Chain algorithm, the No U-Turn Sampler (NUTS), that does not compute the posterior distribution directly, yet draws samples from it in a stochastic way (Hoffman & Gelman, 2014, setting: draws=2000, n_init=1000; for quality assurance and to check for convergence: initially 4 chains, then 1 chain for final analyses). Marginal posteriors are given as mean and 95%-credible intervals. Posterior predictive checks were run to analyze the model performance, i.e., we predicted *FM-end* or *Change* scores based on parameter drawings from the posterior. In this way, we could assess whether data originating from our fitted hierarchical model resembled data from the true underlying distribution. We compare predicted means to the actual sample means and finally compute R-squared values as a measure of explained variance.

## Synthetic data simulation experiments

Before performing Bayesian model comparisons, we first conducted data simulations (Gelman & Hill, 2006, Chapter 8, p. 155) and synthetically generated data based on ground truth models. These simulations enabled us to test strategies to ensure correct model estimation despite the effects of noise and ceiling.

We proceeded in the following way: We selected *proportional to lost*, *proportional to spared function* and *constant* recovery as "true" models. These models were re-arranged to obtain the standard-form classical regression, directly linking $X$ and $Y$ (**Figure 2[D,E,F])**.

The transformation to standard-form enabled us to generate $Y$-values from those of $X$. To consider different degrees of recovery, we assessed these in 10% steps from 10–90% of the proportional recoveries and in steps of 5 from 5–50 points for constant recovery (**Figure 2[D,E,F]**). We then entered the empirical *FM-initial* scores of all *fitters* (n=243) in one of the "true" models, added noise, and enforced ceiling (for details on these procedures, c.f., **supplementary materials, section 4**). The final and critical step then was to fit a new linear regression model to the simulated data and compare the estimated parameters for intercept and slope to the given "true" parameters to answer whether it was still possible to estimate the "true" model after alterations by noise and ceiling. Also, we tracked the ratio of the standard deviations *FM-end/FM-initial*, Pearson correlations of *FM-initial* & *FM-end*, as well as *FM-initial* & *Change* and the number of simulated subjects at absolute ceiling, i.e., at an FM of 66 (maximum score).

Aiming to reduce ceiling and thus its confounding effect, we implemented a subset approach by limiting the data simulations to specific *FM-initial* ranges, i.e., subset 1) *FM-initial* 10–60 including subjects with initial scores between 10 and 60 (n=206), 2) *FM-initial* 10–50 (n=153), 3) *FM-initial* 10–45 (n=118), and 4) *FM-initial* 10–40 (n=92)[2], and evaluated its effect on subsequent model estimation. This enabled us to assess in synthetic data, which subset approach gave us the best trade-off between retrieval of correct model and parameter settings, and size of the remaining data. For each of the described scenarios, simulations were repeated 1000 times. A typical simulation process is illustrated as the black annotations in **Figure 1[A].** We refer to

[2] In case of *FM-initial* 10-45 and *FM-initial* 10-40, we additionally excluded studies that had less than 10 subjects in the respective range (*FM-initial* 10-45: Zarahn *et al.* (2011), Buch *et al.* (2016); *FM-initial* 10-40: Zarahn *et al.* (2011), Buch *et al.* (2016), Byblow *et al.* (2015)).

**supplementary materials, sections 5** and **7** for further details on these simulation experiments and intuitive examples.

## Final model comparisons

### Fitters only

We initially focused on the *fitters* (*FM-initial>10*) portion of the data. Relying on the simulation results, we constructed Bayesian hierarchical models for the standard-form regression, the *proportional to lost function*, and *proportional to spared function* as well as the *constant* recovery models in the subset *FM-initial* 10–45 and conducted a Bayesian model comparison. We focused on this subset, as it represented an optimal compromise between mitigating ceiling effects and retaining as many subjects in the analysis as possible (*FM-initial* 10-45: n=118 out of 385 subjects, 31%. Please note that we excluded the data originating from Zarahn *et al.* (2011) and Buch *et al.* (2016), as both datasets had less than 10 subjects within the range of *FM-initial* 10-45). Results for the subsets *FM-initial* 10–40 and *FM-initial* 10–50 are provided in **supplementary materials, section 10**.

Despite our dataset being as large as currently possible, a potential limitation is that, for some of the studies included, we relied upon values extracted from published figures. This process missed 68 subjects because multiple points sat on top of one another in scatter plots. To account for these missing values and determine an upper bound of R-squared for the winning model in our model comparison, we repeatedly (1000 times) took 68 random draws from the available *FM-initial* distribution. We excluded values not in the range 10–45 and placed the remaining values on the predicted linear fit (i.e., assuming perfect prediction by the standard-form model). By these means, we obtained an average R-squared value, which can be considered an *upper bound* when correcting for missing values.

### Fitters and non-fitters

In the final analyses, we jointly investigated data on *fitters* (*FM-initial*>10) and *non-fitters* (*FM-initial*≤10) by fitting the four competing models outlined before. Once again, we ran analyses using the subset approach, employing a decreased upper limit for *FM-initial* scores to prevent confounding by ceiling (*FM-initial* 0-45: n=270 out of 385 subjects, 70%).

### Statistical analyses

The main analyses were conducted in a Bayesian hierarchical framework. The central inferential question is a model comparison. Specifically, we determined the models best describing the data based on their Leave-One-Out-Cross-Validation (LOOCV) (Vehtari *et al.*, 2017), with LOOCV being a critical out-of-sample test of model fit – indicating whether our findings are likely to hold up in future studies. Model comparisons based on the widely applicable information criterion (WAIC) are also given in **supplementary materials, section 8**. The WAIC represents a principled means of weighing goodness-of-fit against model complexity (i.e., number of effective parameters) (Watanabe, 2013). Additionally, we report R-squared values, the standard measure of the effectiveness of models at explaining within-sample variability.

### Data availability

The recovery data as well as jupyter notebooks (python 3.7, primarily software package pymc3, (Salvatier *et al.*, 2016)) employed in this study are available from the authors on reasonable request.

## Results

### Descriptive statistics

Individual studies as well as the joint distributions of *Initial* and *End* FM scores, median values, and quartiles, are illustrated in **Figure 3**. With regard to subtle differences between studies: Feng *et al.* (2015) only considered stroke subjects with *FM-initial* scores lower than 60. Guggisberg *et al.*, (2017) included more subjects with lower *Initial* scores and scheduled the second assessment sooner on average, which likely underlies the more widespread and less skewed distribution of *FM-end* scores. Further characteristics, such as size, mean age, and sex of each study are summarized in **Supplementary table 1**.

### Bayesian posterior distributions for stroke recovery prediction

The hierarchical standard-form regression and change models built on the entirety of *fitters* (n=243) could be well estimated, as indicated by the convergence of four independently sampled (Monte Carlo Markov) chains for the posterior estimates of model parameters. Additionally, the predicted posterior mean was evenly distributed around the actual sample mean, indicating that the model can reproduce patterns occurring in the real data (**Supplementary Figure 2**). **Figure 4** illustrates the marginal posteriors for the across-study and study-specific intercept and slope parameters, arising from one final chain. Furthermore, **Figure 5** highlights the joint posterior densities for intercept and slope parameters. A striking finding for both models was the dispersion of the individual studies' intercepts and slopes. For the change model, the across-study mean for the slope was 0.64, thus specifying a proportional recovery of 64% for all six studies combined. However, the individual posterior means for slopes fell in the range between 54% (Guggisberg *et al.*, 2017) and 70% (Feng *et al.*, 2015), reflecting different patient mixes and evaluation time points. The six slopes also followed two general patterns, with three studies featuring lower and three studies higher proportional recovery amounts. As expected, the explained variance of the change model markedly surpassed that of the standard-form regression model (Predictive posterior check (PPC): R-squared: 70.8% vs. 42.7%), demonstrating the problematic inflation due to mathematical coupling highlighted in Hawe *et al.* (2019a) and Hope *et al.* (2019). Also, this inflation of explained variance coincided with a small ratio of standard deviations *FM-end/FM-initial*, totaling 0.57. This small ratio at least partially resulted from the number of subjects reaching absolute ceiling at follow-up: 37 (15.2%). In sum, these are the canonical properties of *compression enhanced coupling*.

## Synthetic data simulation experiments

Synthetic data simulations in the sample of *fitters* (Gelman and Hill, 2006) facilitated the detailed study of confounding effects of noise and ceiling as well as hypothetical conclusions when assuming three conceivable ground truth models: *proportional to lost function, proportional to spared function*, as well as *constant* recovery.

Detailed descriptions as well as tables of the data simulations are given in **supplementary materials section 7**. In sum, the inclusion of noise did not impede the correct model and parameter estimation. The situation ~~of correct model and parameter estimation~~ changed markedly with the

introduction of a ceiling: correct estimation deteriorated in parallel to the increase of subjects at absolute ceiling and with growing amounts of recovery (e.g., going from 10% of proportional recovery to 20%). This scenario also demonstrated the effects of *compression enhanced coupling*, since tracked Pearson correlations of *FM-initial* & *Change* increasingly became more extreme than those of *FM-initial* & *FM-end* after enforcing ceiling (**Figure 6[A]**). Importantly, the *variability ratio*, computed as *σ(FM-end)/σ(FM-initial),* decreased in parallel (**Figure 6[B]**). **Figure 6[A]** visualizes these courses: In case of *proportional to spared function* and *constant* recovery, the tracked Pearson correlations of *FM-initial* & *Change* before and after enforcing ceiling diverged dramatically (c.f., yellow and green lines, before and after ceiling). Crucially, trajectories after introducing ceiling closely resembled those of *proportional to lost function* recovery. Hence, when considering all *fitters* and disregarding any potential ceiling effects, there was only one possible conclusion: Data would follow a *proportional to lost function* recovery regardless of the real mechanism driving recovery.

However, estimation performance gradually improved again, when running the simulations in subsets of *fitters*, i.e., considering only those below a certain cut-off of *FM-initial*. The most stringent subset of *FM-initial* 10–40 (92 subjects, out of 385, 24%) performed the best in terms of estimating the true intercepts and slopes for all models. However, in order to choose an appropriate subset range for the intended model comparisons, we tried to find the optimal balance of reducing possible confounds, while also retaining as many subjects in the analysis as possible. The subset *FM-initial* 10–50, containing 153 subjects (out of 385 in all studies, 40%), was only capable of retrieving *proportional to spared function* up to a proportion of 30% and 15 points of *constant* recovery, which we did not judge to be sufficient. On the other hand, we could only keep three studies and 24% of all subjects in the subset *FM-initial* 10–40, which seemed an inefficient use of hard-won empirical data and reduction to too few patients. Therefore, we established a further subset *FM-initial* 10–45 to combine the advantages of *FM-initial* 10–50 and 10–40. Based on the subset *FM-initial* 10–45 (118 subjects, out of 385, 31%), we were able to estimate the entire range of *proportional to lost function*; up to 40% of *proportional to spared function,* and up to 20 points of *constant* recovery. Besides, we were able to recover the true "space" of the generating model for all *proportional to spared function* models (**section 5 of supplementary material)**. Hence, we

decided to focus upon the subset *FM-initial* 10–45 for model comparisons on the human data (results for *FM-initial* 10–50 and 10–40 are presented in **section 10 of supplementary materials**).

## Final model comparisons (on human data)

### Fitters in the subset of FM-initial 10 – 45

The studies by Zarahn *et al.* (2011) and Buch *et al.* (2016) were excluded from these analyses since they had fewer than ten subjects in the range of *FM-initial* 10–45, which would lead to a substantial deterioration in accuracy when sampling from the posterior. Relying on the remaining 118 subjects (out of 385, 31%; 6 subjects were at absolute ceiling, 5%; *variability ratio* was 1.12), we successfully sampled posteriors for the (unconstrained) standard-form regression model, and change-form versions *of proportional to lost function*, *proportional to spared function*, and *constant-recovery* models. Resulting distributions for the marginal posteriors are displayed in **Figure 7**.

The mean of the across-study slope-parameter in the *proportional to lost function* model equaled 0.65 (95% credibility interval 0.39–0.90), thus indicating an across-cohort recovery of a little less than 70%. In contrast to the model on the entire dataset, the explained variance came to just 21.3%. Notably, this value was lower than the explained variance based on the (unconstrained) standard-form regression model (PPC: R-squared: 31.5%). Across studies, subjects had a marginal posterior *constant* recovery of 26 points, ranging from 25 for Byblow *et al.* (2015) and Guggisberg *et al.* (2017) to 30 points for Winters *et al.* (2015). The explained variance amounted to only 5.8%. Explained variance dropped even further in case of *proportional to spared function* (PPC: R-squared: -0.153, slope=0.85), with the negative value signaling the unsuitability of this model. Since these fittings put us on the boundary of correct parameter retrieval for *proportional to spared* and *constant* recovery, we provide further justification for our conclusions in the **supplementary materials section 9**.

As the reported R-squared values are only comparable to a certain extent, since a model's inherent degrees of freedom (i.e., flexibility) are not quantified in this measure, we performed a Bayesian model comparison based upon leave-one-out-cross-validated deviance values. The standard-form regression model, as well as the *proportional to lost* change model, had the lowest deviance and

were thus top-ranked. The standard-form regression model gave a fit that can be seen as a combination of *proportional to lost function* and *constant* recovery and could thus be viewed as liberal *proportional to lost* (**Supplementary Materials, section 1)**. The *constant* recovery model followed these two models. *Proportional to spared function* performed the worst. Non-overlapping confidence intervals for the differences in deviance increased confidence in the two winning models (**Figure 7[E**]; c.f., McElreath, 2018, Chapter 6.5, for a more in depth discussion on model comparisons**).** We refer to **Supplementary Figure 3** for WAIC-based results, which yielded similar results and indicated equidistant differences between in-sample and out-of-sample estimates (horizontal distance between filled and unfilled circles in results panels), rendering a pronounced overfitting of the LOOCV-based models unlikely. Results for the additional subsets *FM-initial* 10–40 and 10–50 are broadly comparable to the subset *FM-initial* 10–45 (**Supplementary Figures 4 & 5)**.

When adjusting for missing values in our dataset in additional analysis, we determined the *upper bound* of the R-squared value for the winning, standard-form regression model to be 44.7%.

### Fitters & Non-Fitters in the subset of FM-initial 0 – 45

Merging *non-fitters* and *fitters* increased the total sample size to 385 subjects, out of which 39 reached maximum values of 66 at follow-up (10%). Further characteristics, such as the ratio of standard deviations, are given in **Supplementary table 1**. Once again, we employed our subset approach and only considered subjects with *FM-initial* scores lower than 45, restricting our analysis to 270 subjects (out of 385, 70%; eight at absolute ceiling at follow-up, 3.0%). Posteriors of the various models' parameters could be reliably sampled, as indicated by converging chains. Evaluating the standard-form regression model first: All of the individual slopes' marginal posterior distributions had a mean in between 1.17 and 1.34, yet included 1 in their credibility intervals (across-cohort slope: 1.24; 95% credibility interval 0.99 – 1.52, **Figure 8[D]**). Therefore, they indicated a mixture of *constant* and *proportional to spared function* recovery (PPC: R-squared: 52.8%), which is similar to the pattern in **Figure 2[I]**. Two of the change models, i.e., *proportional to lost function* and *proportional to spared function*, provided a very poor (negative) in-sample explained variance (PPC: R-squared: -0.13 and –0.51 for *proportional to lost* and *proportional to spared function* recovery, respectively, **Figure 8[A,B]**). Only the *constant*

recovery model could capture some positive variance (PPC: R-squared: 7.4%, **Figure 8[C]**). The final model comparison revealed the (unconstrained) standard-form regression model, indicating a mixture of *proportional to spared function* and *constant* recovery, and *constant* recovery as the winning models (**Figure 8[E]**).

## Discussion

Current analyses of proportional recovery after stroke are subject to various confounds. We here proposed a subset approach to minimize a key confound, *compression enhanced coupling*, which we validated in synthetic data experiments. We furthermore employed hierarchical Bayesian models to analyze one of the largest, compiled dataset of upper limb recovery post-stroke (n=385) and evaluate various conceivable patterns of stroke recovery in overall model comparisons.

We first carried out the subset approach focussing on those patients considered to be *fitters* to the proportional recovery rule. Thus, we considered all 118 participants with an *Initial* Fugl-Meyer (FM) score of at least ten to exclude *non-fitters* (Feng *et al*., 2015) and a score of less than 45 to decrease *compression enhanced coupling* **(Figure 1[C]).** In this case, model comparison pointed in the direction of either *proportional to lost function*, with a recovery proportion of 65%, or a combination of *proportional to lost function* and *constant* recovery (**Figure 2[H]**) as the underlying relationships. These findings were, therefore, generally in line with previous assumptions of *proportional to lost function* recovery post-stroke (Prabhakaran *et al.*, 2008). However, the pure *proportional to lost function* recovery model could only explain 21% of the variance in recovery, a value drastically reduced in comparison to earlier studies, reporting up to 94% (Winters *et al.*, 2015). Given the likely confounds by *compression enhanced coupling* in these earlier studies, the current estimate of explained variance may be considered more accurate. As the standard-form regression directly linked initial and follow-up FM scores, it is important to note that this model was not prone to the confounds due to mathematical coupling.

Of note, these conclusions substantially depended on the exclusion of patients with very low FM-initial scores, so-called *non-fitters*: a completely different picture arose when employing the subset approach to the entire spectrum of subjects, i.e., *fitters* and *non-fitters* combined (*FM-initial* 0-45

to decrease *compression enhanced coupling*, n=270, **Figure 1[D]).** The model comparison led to the selection of a composite of *proportional to spared function* and *constant* recovery as the winning model (**Figure 2[I]**); the explained variance was 53%. In contrast to *proportional to lost function*, *proportional to spared function* recovery suggests that patients with greater preservation of function have a higher capacity to improve, presumably because basic abilities to move limbs could enable the re-acquisition of more sophisticated movement patterns more easily. Indeed, functional neuroimaging data have shown that a higher degree of residual motor function is associated with a more physiological and thus lateralized motor network architecture (Rehme *et al.*, 2011). This lateralized architecture in turn has been shown to constitute a strong predictor for good motor recovery (Grefkes and Fink, 2014). Altogether, considering this recovery pattern could have important implications for the conceptualization of recovery trajectories.

The pressing questions currently are: Are we in need of more studies jointly analyzing *fitters* and *non-fitters*, particularly given the higher explained variance for the whole sample? Are 32% or even 53% explained variance sufficient to justify a recovery *rule* that may guide individual predictions in a future of precision neurology?

**Clinical importance and prediction of recovery post-stroke**

Overestimation of the proportional recovery rule becomes particularly problematic when it impacts clinical practice, e.g., prompts the assumption of a spontaneous recovery process that exclusively depends on initial motor impairment. In particular, such a conclusion may limit the allocation of valuable therapy sessions to some stroke subjects and generate a negative prior expectation towards tailored therapies (Byblow *et al.*, 2015; Hawe *et al.*, 2019*b*). Putting its suitability for single-participant prediction aside, Byblow and Stinear (2019) and Kundert *et al.* (2019) recently underscored the proportional recovery rule's purpose for explaining the recovery process in stroke populations in general. From this perspective, it may be that *proportional to lost function* recovery explains *fitters'* trajectories better than other change models, but the low level of explained variance suggests a need for further and better predictors**.**

Because of the increase in the explained variance of motor recovery when jointly considering *fitters* and *non-fitters* – from 32% to 53% – we may also need to rethink conventional analysis

approaches and increasingly shift the focus on severely affected subjects by including *non-fitters* more often. This might be particularly important in view of the number of severely affected patients, e.g., 37% of *non-fitters* in our study and generally increasing numbers of patients with severe arm impairments (Hayward *et al.*, 2017). Additionally, recent studies have provided evidence that certain practices to classify subjects into *fitters* and *non-fitters* may be biased, as they lead to increased estimates of explained variance and potentially erroneous conclusions. This situation may particularly arise when dividing patients into *fitters* and *non-fitters* on results based on clustering *Initial* and change scores (Hawe *et al*., 2019; Kundert *et al.*, 2019). Nevertheless, even 53% of explained variance can be considered rather low, suggesting that recovery is influenced by more factors than mere initial motor impairment as measured by the Fugl-Meyer scale. In this respect, our finding of higher explained variance when estimating parameters for *fitters* and *non-fitters* combined does not stand against the observation that there are grossly different recovery patterns across patients, which may necessitate differing therapeutic (rehabilitative) approaches for differently impaired patient subgroups.

**Likert-like scales and ceiling effects**

It is also essential to be aware of a particular score's characteristics. The FM score is a Likert-like scale, thus a summary of multiple Likert-like items, comprising ordinal data (Likert, 1932). It is this combination of multiple items that renders the parametric statistical approaches applied here feasible (Norman, 2010; Harpe, 2015). Positively, reported test-retest and inter-rater reliabilities for FM scores are $r>0.95$ over repeated measurements (Gladstone *et al.*, 2002). However, as previously emphasized, the FM assessment is highly susceptible to ceiling effects (Gladstone *et al.*, 2002). We here further dissected these ceiling effects and highlighted the induction of *compression enhanced coupling* with change formulation models that link initial scores and recovery (Lord, 1956; Hawe *et al.*, 2019a; Hope *et al.*, 2019). Amongst our total sample size of 385 subjects, 39 (10%) reached maximum scores at follow-up, and many more were likely compressed towards, but not at ceiling. Our synthetic data experiments showed that this degree of ceiling effect was sufficient to impede correct conclusions: Independent of the simulated ground truth mechanism, we would always discover *proportional to lost function* recovery.

Here, we relied upon the logic of subsets to decrease confounding effects by ceiling. That is, we focused on lower ranges of initial motor performance scores, which are less likely to lead to

maximum *End* motor performance scores. The relationships identified in the ceiling-reduced subsets could then be extended to the entire sample, as generalization was ensured by the assumed constant relationship between *Initial* and *End* scores inherent to linear regression. Importantly, our synthetic data experiments additionally demonstrated that we did not incur any new confounds, which would affect conclusions when defining subsets on initial scores.

**Bayesian hierarchical models**

Potentially insufficient numbers of subjects could endanger successful subset analyses relying on the data of just one study. This may be particularly the case as it may not be feasible to increase the size of individual studies, as high-quality data acquisition is time-consuming and costly. Here, the subset approach was only rendered possible due to our Bayesian hierarchical framework that facilitated the fusion of multiple datasets with individual-level data (Gelman and Hill, 2006). In this way, we could maximize the number of included subjects, while retaining as much information on each study's characteristics as possible and modelling uncertainty explicitly (McElreath, 2018). This combination of merging studies and preserving individual features was particularly appealing, since it addressed both similarities and dissimilarities between individual studies. On the one hand, we had similar scores from similar patients at similar time-points and yet considered various study sites and likely minor variations in therapies on the other hand. Also, our Bayesian hierarchical models were capable of effectively handling diverging sample sizes in the six studies considered (McElreath, 2018). Lastly, as anticipated in (Hope *et al.*, 2019), they allowed for the evaluation of various generative models on the nature of recovery – *proportional to lost function*, *proportional to spared function*, and *constant recovery* – through model comparisons.

**Limitations and future directions**

We decreased distorting ceiling effects by limiting analyses to subsets of initial scores. However, we acknowledge that there are drawbacks to this approach, such as the exclusion of substantial portions of the entire sample. Also, it does not represent definitive handling of the ceiling problem. The FM assessment is based on several single items. For example, it asks whether a patient is able, partially able or unable to move the hand from the ipsilateral ear to the contralateral knee. This multitude of items could potentially result in multiple sub-ceilings. These may remain present even in case of excluding data at the scale's maximum. Therefore, one viable strategy to circumvent

these kinds of ceiling effects could be the increased use of behavioral and clinical assessments with continuous scales, for example, muscle forces, movement speeds or other kinematic parameters. Another strategy could be the construction of more elaborate scoring systems that allow for the detection of even very subtle variation, especially at the top of the scale. Nevertheless, motor impairment may be maximally recovered and effectively indistinguishable from a healthy pre-stroke level in some cases. As a result, some natural ceiling would occur and require special attention, primarily concerning statistical procedures. Future research may thus utilize our subset analysis or further approaches for censored data, such as frequentist Tobit models (Tobin, 1958) and Bayesian counterparts (Gelman and Hill, 2006).

Furthermore, we did not attempt to differentiate between potential non-linearities of the FM scale, e.g. is a ten-point gain from 0 to 10 the same as a ten point gain from 50 to 60? Especially when considering the involvement of different functional domains and their interactions, a linear recovery pattern seems rather unlikely. For example, motor recovery might also be influenced by recovery from visuospatial neglect. Such non-linearities might thus not be detectable by the linear regression models we, and the majority of the field, have so far focused on. Therefore, our results encourage research into other model types, for example, non-linear models, such as decision tree-like algorithms (Stinear *et al.*, 2012) and exponential recovery functions (van der Vliet *et al.*, 2020). Additionally, we may need to refocus on a variety and multivariate combination of indicators of stroke recovery, such as behavioral, physiological, and imaging biomarkers, which have already shown promise (Stinear, 2017; Ward, 2017; Findlater *et al.*, 2019).

Our study highlights the opportunity for novel insights to be gleaned by Bayesian hierarchical modelling, as it facilitates model comparisons and the creation of large datasets, thereby increasing the generalizability of obtained inferences. Therefore, they are likely to become common in stroke research, as well as in other clinical fields. Indeed, the strategies outlined here may inspire and guide future studies, raise awareness of the better handling of ceiling and change models, as well as the pernicious nature of *compression enhanced coupling*; especially, as these effects may frequently occur in biomedical data.

In this present study, we relied on a relatively large number of 385 subjects. Nonetheless, we are still in need of larger stroke recovery datasets. The individual studies, that we here combined, all specified upper limb motor impairment as inclusion criterion and primarily recorded FM scores. Some studies even explicitly excluded patients with communication or memory deficits (Winters

*et al.*, 2015) or patients with concomitant posterior or cerebellar artery stroke (Byblow *et al.*, 2015). To further explore the influence of non-motor impairments on the recovery of motor impairments and vice versa, more ambitious data-rich studies in future research (c.f., Bzdok *et al.*, 2019) will need to simultaneously record a multitude of stroke symptoms, such as motor impairments of upper and lower limbs, aphasia, neglect, apraxia and hemianopia. In particular, the field needs to take advantage of collaborative data collection and move beyond behavioral scores for both predictor and outcome variables. In this way, a range of potential explanatory and predictive variables could be incorporated to reliably increase explained variance and accuracy of out-of-sample prediction of stroke recovery - in case of *fitters* and also *non-fitters (*Grefkes and Fink, 2016; Boyd *et al.*, 2017; Bernhardt *et al., 2017).*

**Conclusion**

Our Bayesian approach to systematically revisit post-stroke motor performance revealed only weak signs of *proportional to lost function* recovery for those defined to be *fitters* to the proportional recovery rule. Variance in recovery could only be explained by up to 32%, which is less than 50% of that previously reported. Additionally, a combination of *proportional to spared function* and *constant* recovery emerged as a likely relationship for the recovery of the entirety of stroke subjects – at a higher explained variance of 53%. Importantly, these estimates were obtained after de-confounding effects of mathematical coupling and ceiling by means of subset analyses (Hawe *et al.*, 2019a; Hope *et al.*, 2019). In summary, these lower levels of explained variance may motivate research moving beyond behavioral measures and the consideration of combinations of various biomarkers, such as demographic, clinical, imaging, and physiological. Ultimately, our findings may also pave the way for more common use of Bayesian hierarchical analyses. In this way, we may distill and accumulate evidence resting upon merged clinical datasets and efficiently ensure reliable generalization performance and modelling of uncertainty.

## Acknowledgements

We thank Michael Moutoussis for valuable observations and discussions.

## Funding

AKB’s clinician scientist position is supported by the dean's office, Faculty of Medicine, University of Cologne. GRF gratefully acknowledges support by the Marga and Walter Boll foundation.

## Competing interests

None.

## References


Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Research 2019; 4

Boyd LA, Hayward KS, Ward NS, Stinear CM, Rosso C, Fisher RJ, et al. Biomarkers of stroke recovery: consensus-based core recommendations from the stroke recovery and rehabilitation roundtable. International Journal of Stroke 2017; 12: 480–493.

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: 1924–1925.

Byblow WD, Stinear CM. Letter by Byblow and Stinear Regarding Article" Taking Proportional Out of Stroke Recovery". Stroke 2019: STROKEAHA118024595–STROKEAHA118024595.

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: 848–859.

Bzdok D, Floris DL, Marquand AF. Analyzing Brain Circuits in Population Neuroscience: A Case to Be a Bayesian. arXiv preprint arXiv:190902527 2019

Bzdok D, Floris DL, Marquand AF. Analysing brain networks in population neuroscience: a case for the Bayesian philosophy. Philosophical Transactions of the Royal Society B 2020; 375: 20190661.

Bzdok D, Nichols TE, Smith SM. Towards algorithmic analytics for large-scale datasets. Nature Machine Intelligence 2019; 1: 296–306.

Chiolero A, Paradis GP, Rich BD, Hanley JP. Assessing the relationship between the baseline value of a continuous variable and subsequent change over time. Frontiers in public health 2013; 1: 29.

Cronbach LJ, Furby L. How we should measure ‘change’: Or should we? Psychological Bulletin 1970; 74: 68–80.

Feng W, Wang J, Chhatbar PY, Doughty C, Landsittel D, Lioutas V-A, 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: 860–870.

Findlater SE, Hawe RL, Mazerolle EL, Al Sultan AS, Cassidy JM, Scott SH, et al. Comparing CST Lesion Metrics as Biomarkers for Recovery of Motor and Proprioceptive Impairments After Stroke. Neurorehabilitation and Neural Repair 2019: 1545968319868714.

Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scandinavian journal of rehabilitation medicine 1975; 7: 13–31.

G. Broeks, J, Lankhorst, GJ, Rumping, K, Prevo AJH. The long-term outcome of arm function after stroke: results of a follow-up study. Disability and Rehabilitation 1999; 21: 357–364.

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

Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge university press; 2006

Gladstone DJ, Danells CJ, Black SE. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. Neurorehabilitation and neural repair 2002; 16: 232–240.

Grefkes C, Fink GR. Connectivity-based approaches in stroke and recovery of function. The Lancet Neurology 2014; 13: 206–216.

Grefkes C, Fink GR. Noninvasive brain stimulation after stroke: it is time for large randomized controlled trials! Current Opinion in Neurology 2016; 29: 714–720.

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: 1029–1041.

Harpe SE. How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning 2015; 7: 836–850.

Hawe RL, Scott SH, Dukelow SP. Taking Proportional Out of Stroke Recovery. Stroke 2019; 50: 204–211.

Hawe RL, Scott SH, Dukelow SP. Response by Hawe et al to Letter Regarding Article," Taking Proportional Out of Stroke Recovery". Stroke 2019: STROKEAHA119024794–STROKEAHA119024794.

Hayes RJ. Methods for assessing whether change depends on initial value. Statistics in Medicine 1988; 7: 915–927.

Hayward KS, Schmidt J, Lohse KR, Peters S, Bernhardt J, Lannin NA, 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.

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

Hope TM, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. Recovery after stroke: not so proportional after all? Oxford University Press; 2018

Hope TMH, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. Recovery after stroke: not so proportional after all? [Internet]. bioRxiv 2018[cited 2019 Jan 29] Available from: http://biorxiv.org/lookup/doi/10.1101/306514

Kundert R, Goldsmith J, Veerbeek JM, Krakauer JW, Luft AR. What the Proportional Recovery Rule Is (and Is Not): Methodological and Statistical Considerations. Neurorehabilitation and neural repair 2019: 1545968319872996.

Likert R. A technique for the measurement of attitudes. Archives of psychology 1932

Lord FM. THE MEASUREMENT OF GROWTH. ETS Research Bulletin Series 1956; 1956: i–22.

Marchi NA, Ptak R, Di Pietro M, Schnider A, Guggisberg AG. Principles of proportional recovery after stroke generalize to neglect and aphasia. European journal of neurology 2017; 24: 1084–1087.

McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC; 2018

Nakayama H, Jørgensen HS, Raaschou HO, Olsen TS. The influence of age on stroke outcome. The Copenhagen Stroke Study. Stroke 1994; 25: 808–813.

Newman M. The process of recovery: After hemiplegia. Stroke 1972; 3: 702–710.

Norman G. Likert scales, levels of measurement and the “laws” of statistics. Advances in health sciences education 2010; 15: 625–632.

Organization WH. International classification of functioning, disability and health: ICF. 2001

Prabhakaran S, Zarahn E, Riley C, Speizer A, Chong JY, Lazar RM, et al. Inter-individual Variability in the Capacity for Motor Recovery After Ischemic Stroke. Neurorehabilitation and Neural Repair 2008; 22: 64–71.

Rehme AK, Fink GR, von Cramon DY, Grefkes C. The role of the contralesional motor cortex for motor recovery in the early days after stroke assessed with longitudinal FMRI. Cerebral cortex 2011; 21: 756–768.

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

Smith M-C, Byblow WD, Barber PA, Stinear CM. Proportional recovery from lower limb motor impairment after stroke. Stroke 2017; 48: 1400–1403.

Stinear CM. Prediction of motor recovery after stroke: advances in biomarkers. The Lancet Neurology 2017; 16: 826–836.

Stinear CM, Barber PA, Petoe M, Anwar S, Byblow WD. The PREP algorithm predicts potential for upper limb recovery after stroke. Brain 2012; 135: 2527–2535.

Tobin J. Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society 1958: 24–36.

Twitchell TE. The restoration of motor function following hemiplegia in man. Brain 1951; 74: 443–480.

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

van der Vliet R, Selles RW, Andrinopoulou E-R, Nijland R, Ribbers GM, Frens MA, et al. Predicting upper limb motor impairment recovery after stroke: a mixture model. Annals of Neurology 2020

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

Watanabe S. A widely applicable Bayesian information criterion. Journal of Machine Learning Research 2013; 14: 867–897.

Winters C, Van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the maximum proportional recovery rule to visuospatial neglect early poststroke. Neurorehabilitation and neural repair 2017; 31: 334–342.

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: 614–622.

Zarahn E, Alon L, Ryan SL, Lazar RM, Vry M-S, Weiller C, et al. Prediction of Motor Recovery Using Initial Impairment and fMRI 48 h Poststroke. Cerebral Cortex 2011; 21: 2712–2721.

## Figures

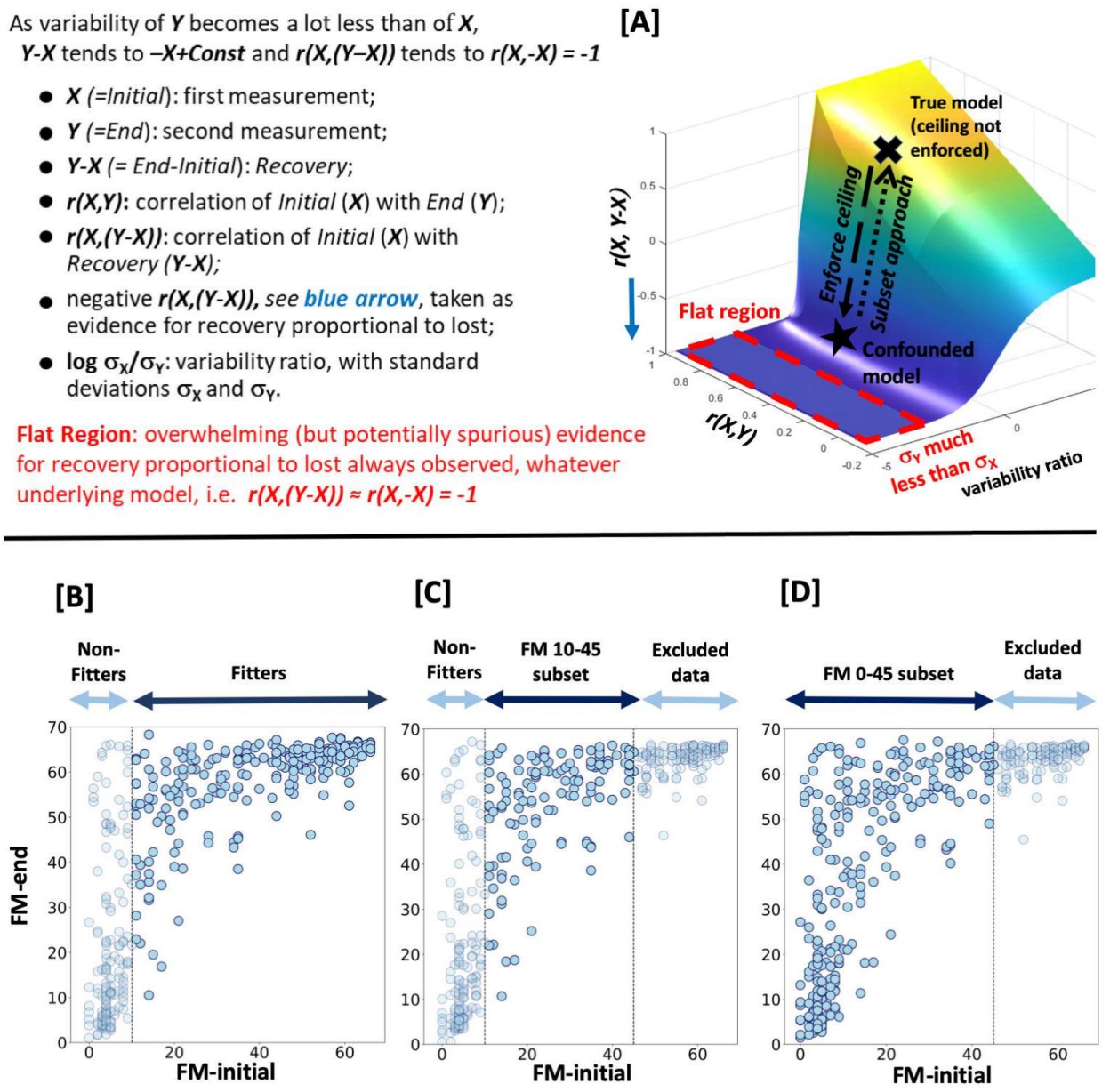



**Figure 1. The confounded nature of *proportional (to lost function)* recovery assessments and investigated subgroups. [A] Surface plot:** Depiction of the relationships between correlations of *Initial* and *End* measurements, *r(X,Y)*, correlations of *Initial* and *Change* measurements, *r(X,(Y-X))*, and the (log) ratio of *End* to *Initial* standard deviations. The logic of synthetic data simulations is also shown (black text, arrows, and symbols on surface plot, for additional intuitions c.f., **supplementary materials, section 5**). Figure modified from (Hope *et al.*, 2019). **[B, C & D] Subgroup analyses.** Recovery data are presented as *Initial* Fugl-Meyer scores in the acute phase

against *End* Fugl-Meyer scores in the chronic phase after stroke. **[B]** Conventional subgroups of *fitters* and *non-fitters* based on a cut-off of *FM-initial*=10. **[C]** Subset approach for *fitters* only: Only patients with an FM-initial between 10 and 45 are considered in order to control for ceiling effects. **[D**] Subset approach considering both fitters and non-fitters using an FM range between 0 and 45. Excluded data are indicated by lighter colour.

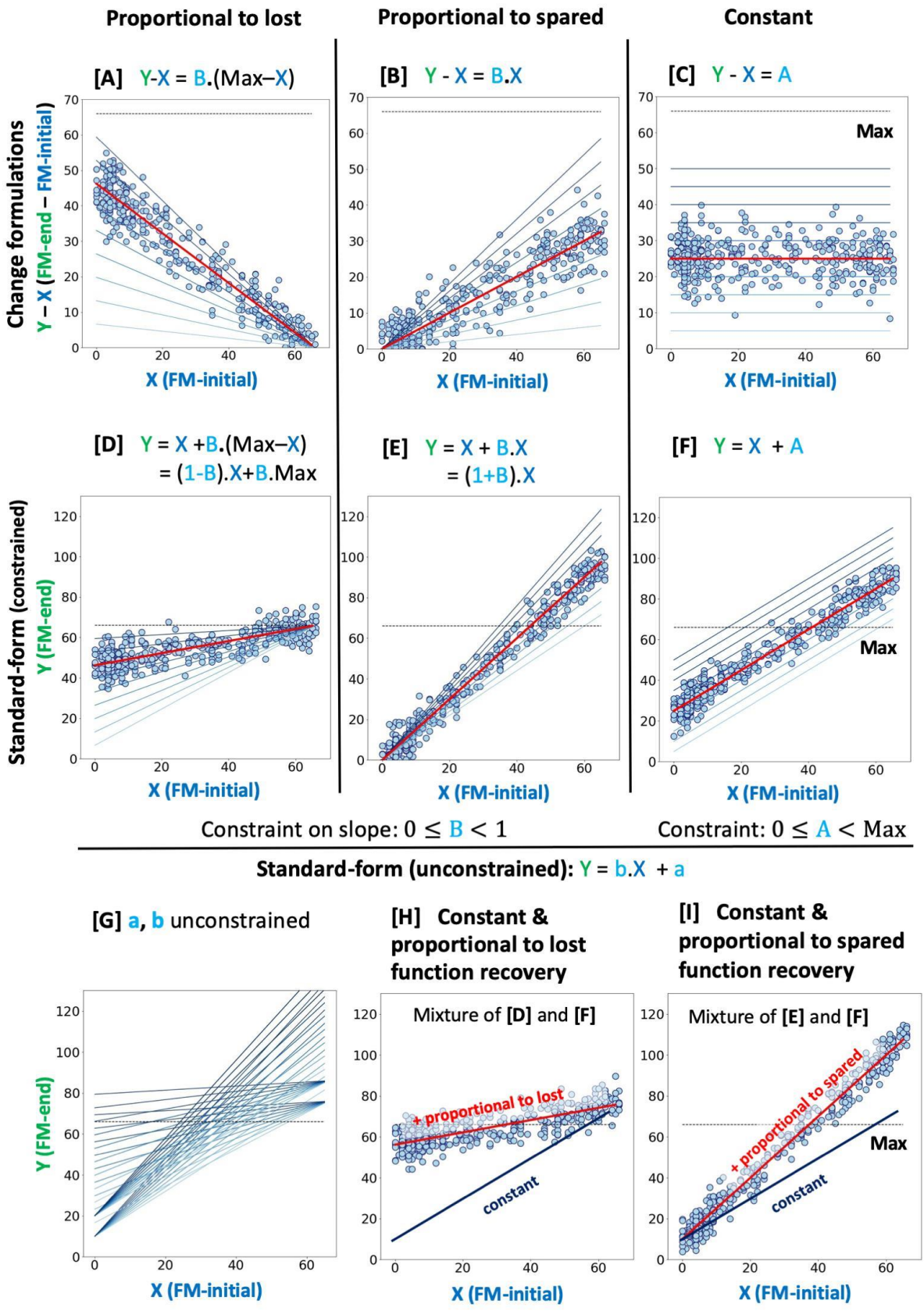

Proportional to lost
Proportional to spared
Constant
[A] Y-X = B.(Max–X)
[B] Y - X = B.X
[C] Y - X = A
Change formulations
Y – X (FM-end – FM-initial)
Max
X (FM-initial)
[D] Y = X +B.(Max–X)
= (1-B).X+B.Max
[E] Y = X + B.X
= (1+B).X
[F] Y = X + A
Standard-form (constrained)
Y (FM-end)
Constraint on slope: 0 ≤ B < 1
Constraint: 0 ≤ A < Max
Standard-form (unconstrained): Y = b.X + a
[G] a, b unconstrained
[H] Constant & proportional to lost function recovery
[I] Constant & proportional to spared function recovery
Mixture of [D] and [F]
Mixture of [E] and [F]
+ proportional to lost
+ proportional to spared
constant

**Figure 2. Various representations of recovery patterns: *Proportional to lost function* [A,D], *proportional to spared function* [B, E], *constant* recovery [C,F], and unconstrained standard-form [G,H,I].**

Performance is inspired by the Fugl Meyer (FM) assessment of the upper limb, where 66 is the maximum value, providing a ceiling. However, the depictions here show the "true" underlying recovery pattern that one would obtain if there was no ceiling (thus, we extended the scales beyond the maximum). We show a range of linear regression lines. For a more realistic visualization, simulated recovery data points are also shown, with a red line showing the best fit to them. **Top row panels [A, B & C]** depict *proportional to lost function, proportional to spared function,* and *constant* recovery to varying amounts (i.e., 10%, 20%, 30% etc. proportional recovery, each generating a different regression line) under the typical change formulations. The horizontal axis presents *Initial* scores, $X$, while the vertical axis stands for the change between *Initial* and *End* scores, $(Y\text{-}X)$. **Middle row panels [D, E & F]** depict the same linear relationships, but re-expressed as classical standard-form regressions, by merely moving the $X$ variable to the right-hand side of the equation, and then, rearranging. In contrast to the top row, the vertical axis here represents the raw *End* score, $Y$. **Bottom row panels [G, H & I]** illustrate unconstrained standard-form regression between *Initial,* X, and *End,* Y, scores. This is the most general of all models, as it fits an intercept, as well as a slope that can take any numerical values (i.e., they are unconstrained). **[G]** presents a range of possible linear relationships that can be fit. **[H]** and **[I]** visualize specific mixtures. Importantly, neither of the linear relationships in **[H]** and **[I]** is directly in the space of the basic *proportional to lost function, proportional to spared function* and *constant* recovery models, i.e., panels **[D, E & F]**. Notation: Upper case A and B represent intercept and slope in change models, while lower case *a* and *b* represent intercept and slope in standard-form regression models.

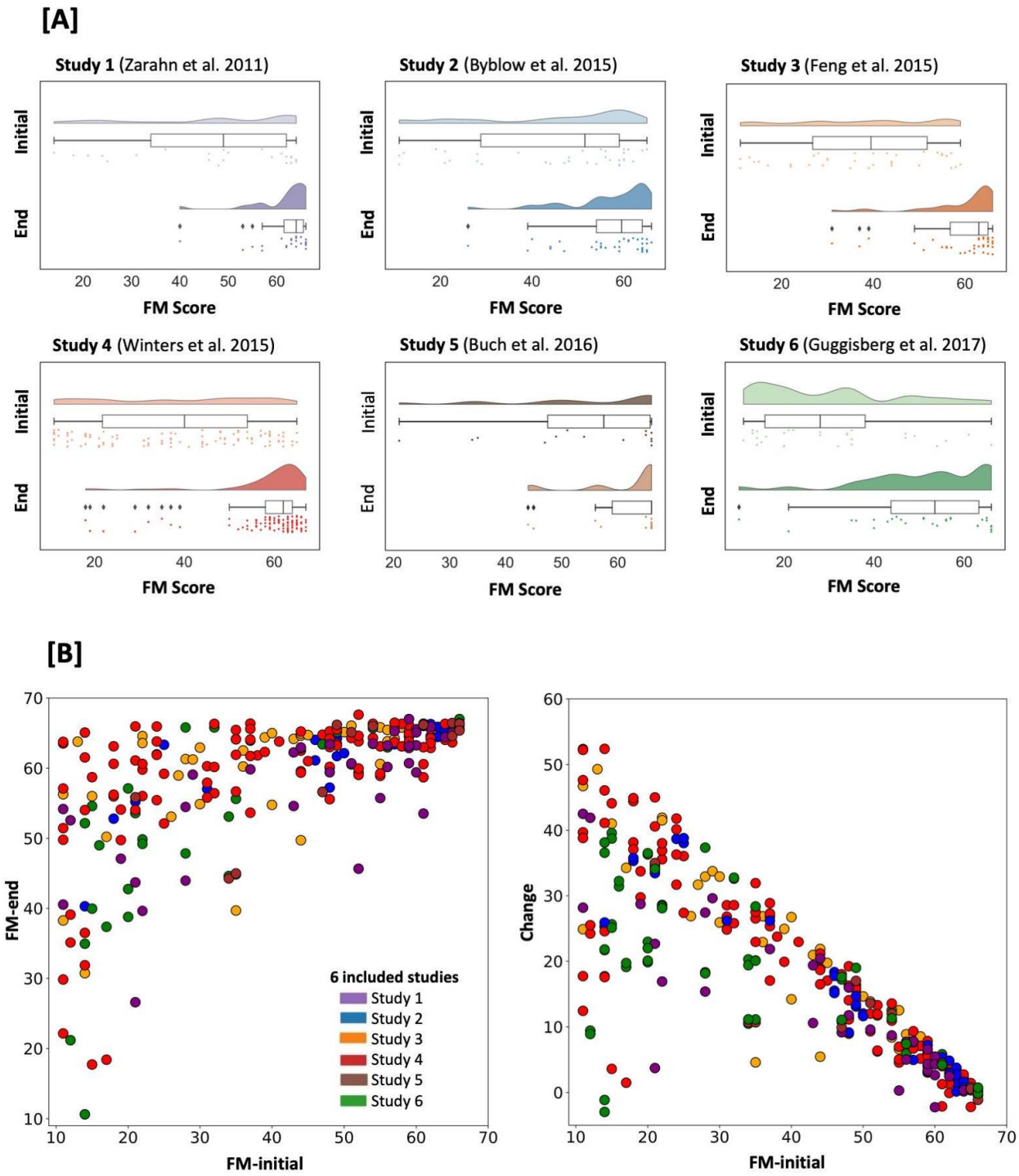



**Figure 3. Initial and follow-up upper limb motor performance, individually for each of the six studies included and aggregated across studies. [A] Raincloud plots of the Initial and End Fugl-Meyer (FM) assessment scores.** These measured the upper limb impairment post-stroke of all *fitters*, i.e., those with *FM-initial* > 10 (n=243). Each of the six studies included is displayed separately and uniquely color-coded. The upper row within each study's plot visualizes the

distribution of scores. The second row summarizes the same data in a boxplot (i.e., median, upper and lower quartiles, whiskers extending to the entire range of data, outliers indicated as separate dots). Lastly, the third row displays raw individual data points. While initial FM score distributions are more homogeneous across the entire range (i.e., more uniform), distributions at the second time point are narrower and – to varying degrees – more pronounced at the upper end of the FM assessment scale, i.e., skewed. The code for raincloud plots relies on (Allen *et al.*, 2019). [**B] Entirety of aggregated individual FM assessment scores of all stroke subjects defined as *fitters* (*FM-initial* > 10).** Scores are jittered on the vertical axis for visualization only. Left: *FM-initial* vs. *FM-end*, clearly depicting the increased density for follow-up scores close to and at ceiling, i.e., *FM-end* = 66. Right: *FM-initial* vs *Change* (*FM-end* – *FM-initial*). Included studies are color-coded as before, c.f. legend.

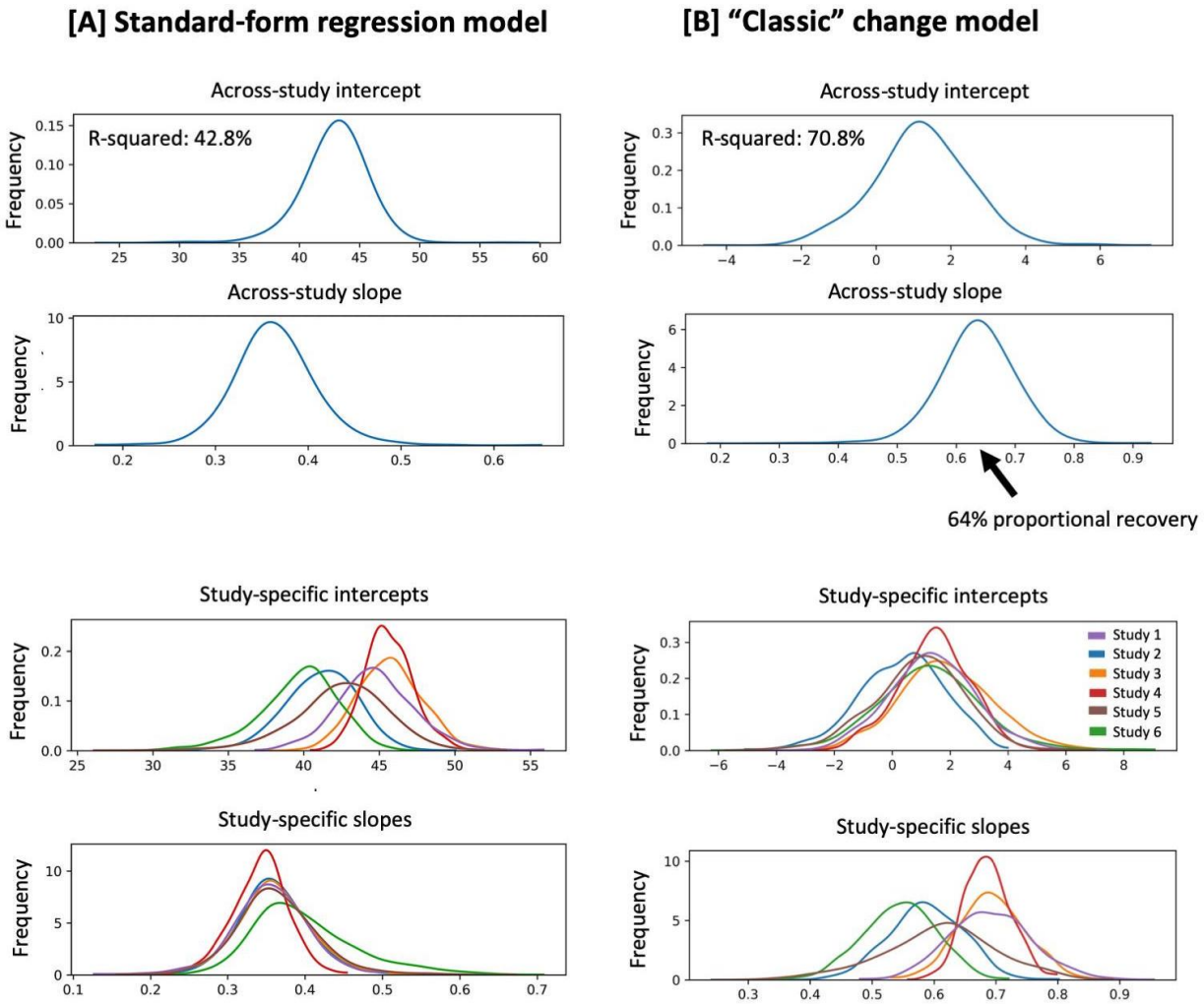


**Figure 4. Bayesian hierarchical model of all *fitters* (n=243): Marginal posteriors of intercepts and slopes.** The best parameter settings are the ones with the highest frequencies (c.f., black arrow) and the spread of the distributions indicate uncertainty associated with the parameter estimation. **[A] Standard-form regression model.** *End* FM scores are estimated according to *FM-end* = b.*FM-initial* + a. *Across-study* intercept ~~("mu_a")~~ and slope ~~("mu_b")~~ are depicted in the upper two rows. The bottom two rows visualize the lower level of the hierarchy: Varying intercepts and slopes, individually per included study (c.f. legend for study-specific color coding). [**B] "Classic" *proportional to lost function* change model.** The outcome measure of interest here is the change between the *Initial* and *End* FM scores, estimated based on *Change* = B.*Potential* + A with *Change* = *FM-end* – *FM-initial* and *Potential* = 66 – *FM-initial*. The across-study slope indicates a proportional recovery of 64% (c.f., black arrow). The inflation of explained variance due to mathematical coupling and *Ceiling enhanced Coupling* is demonstrated by an R-squared value of 70.8% for the change model that exceeds the one from the standard-form regression model by 28%. Please note that unlike our presentations of *proportional to lost function* elsewhere, we include an intercept here to maximize fit, although, as can be seen, fitting generates an intercept very close to zero.

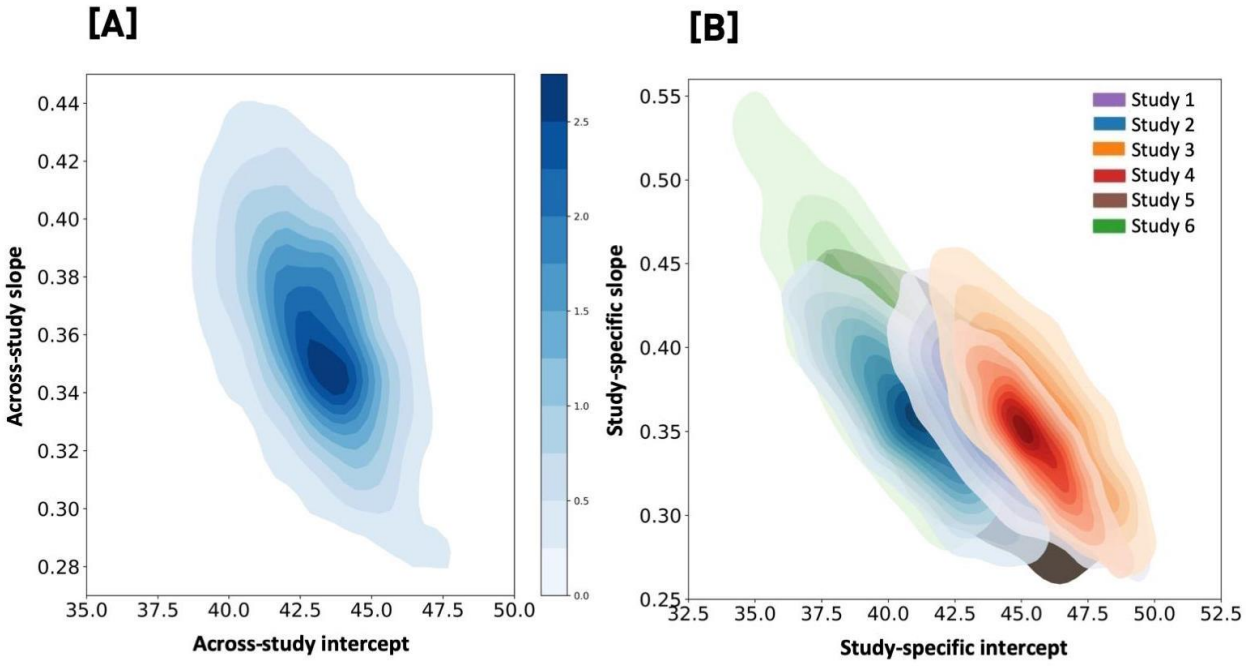



**Figure 5. Bayesian hierarchical model and model estimated within- and between-study differences in motor recovery based on the standard-form regression model. [A] Full**

**specification of the Bayesian hierarchical model. [B] Aggregated motor studies:** We display the joint distribution between the *across-study* intercept – which can be described as the average motor outcome for an *FM-initial* score of zero – and the *across-study* slope – equivalently framed as performance gain dependent on *FM-initial*. Therefore, the plot illustrates the joint posterior densities for the included hyperparameters, with the marginal posterior for the intercept ranging from 37.0 to 47.4 and from 0.28 to 0.44 for slopes (95% credibility intervals). [**C] Individual motor studies:** The figure pictures the joint density for combinations of intercepts and slopes that are plausible, given the visited data of the six included studies. It particularly highlights the relationship between sample size and width of credibility intervals, as larger studies present with narrower intervals. C.f. legend for study-specific color-coding.

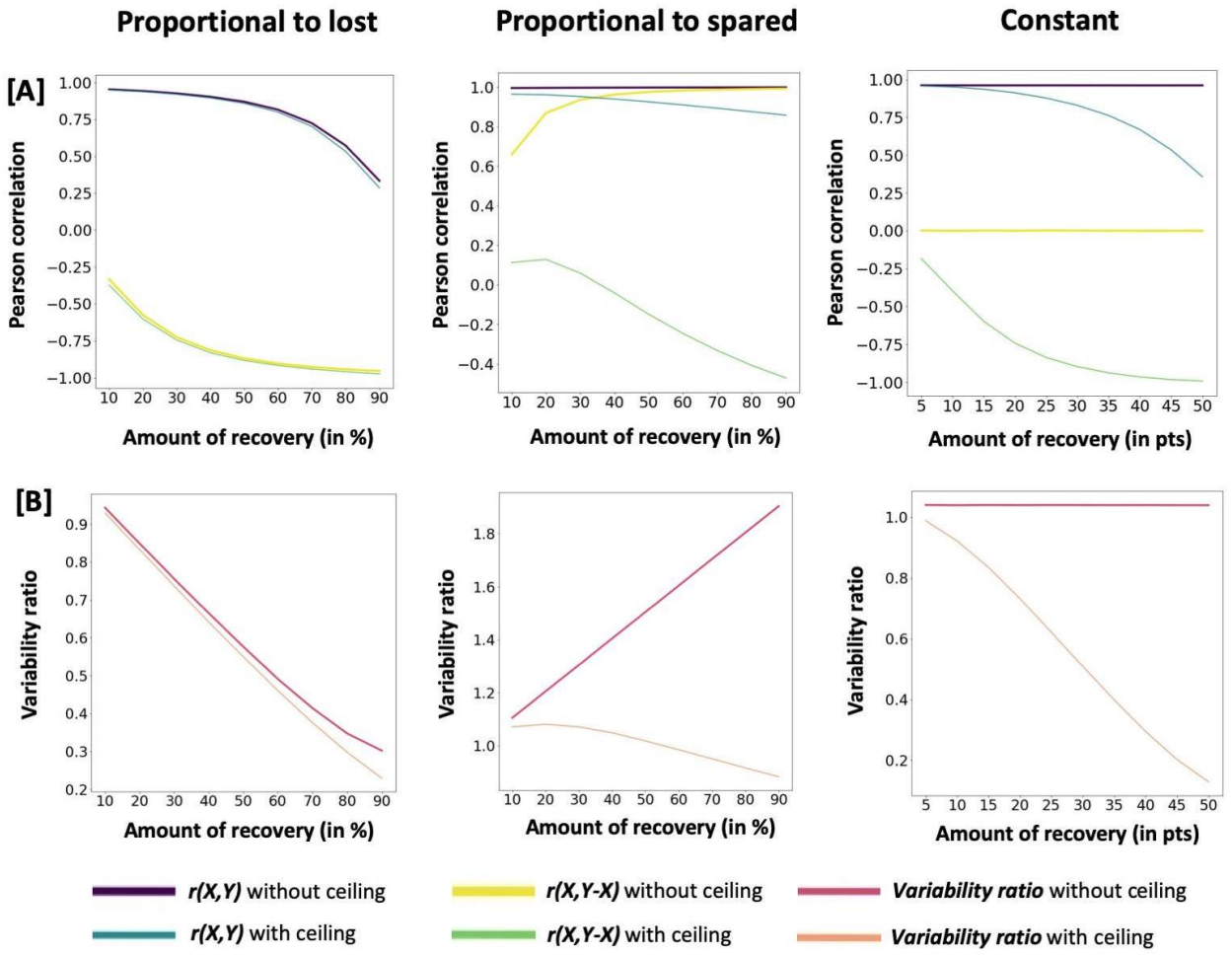



**Figure 6. Synthetic data simulations of *proportional to lost function* (left column), *proportional to spared function* (middle column), and *constant* recovery (right column) based on 243 simulated subjects. [A]. Trajectories of Pearson correlation between *FM-initial (X) & FM-***

***end* (Y) and *FM-initial (X) & Change (Y-X).*** *Proportional to lost function recovery:* Starting at almost the maximum of 1, the correlation between *FM-initial* & *FM-end* decreases the higher the amount of proportional recovery, while the correlation for *FM-initial* & *Change* becomes more negative and finally exceeds the one of *FM-initial* & *FM-end* in absolute terms, demonstrating the effect of mathematical coupling. Ceiling only exhibits a minor amplification of this effect. *Proportional to spared function recovery:* Without any ceiling, correlations of *FM-initial* & *FM-end*, as well as *FM-initial* & *Change*, are close to 1. The latter changes dramatically after enforcing ceiling: The correlations of *FM-initial* & *Change* are now decreasing monotonically, become negative in sign and are reminiscent of *proportional to lost function*. *Constant recovery:* The correlation between *FM-initial* & *FM-end* is close to 1, while *FM-initial* and *Change* are not correlated. After ceiling is enforced, patterns closely resemble the ones for *proportional to lost function*: The correlation of *FM-initial* & *FM-end* decreases monotonically, yet stays positive, the one between *FM-initial* & *Change* decreases and becomes almost -1 for high levels of constant recovery. [**B] Trajectories of *variability ratios.*** *Proportional to lost function recovery: Variability ratios* are decreasing from approx. 1 to 0.2, ceiling exhibits only minor effects. *Proportional to spared function recovery:* Trajectories differ markedly depending upon ceiling: Ratios are greater than 1 and increase before ceiling and decrease to values smaller than 1 after ceiling is enforced. *Constant recovery:* Once again, the presence of a ceiling substantially alters the trajectories of *variability ratios*: While they remain close to 1 before enforcing ceiling, they show a steep decrease after enforcing ceiling.

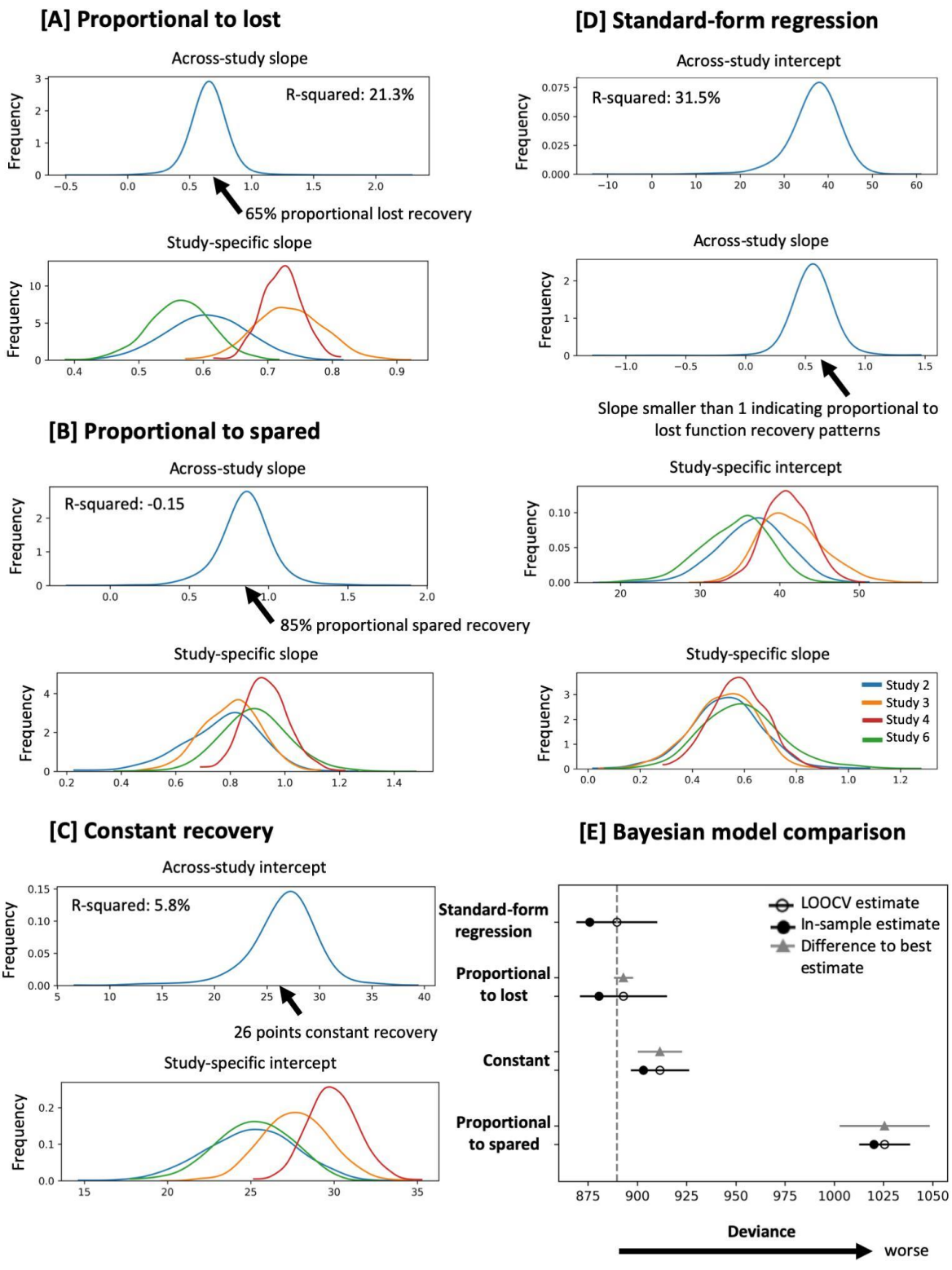

[A] Proportional to lost
Across-study slope
R-squared: 21.3%
Frequency
65% proportional lost recovery
Study-specific slope
[B] Proportional to spared
Across-study slope
R-squared: -0.15
85% proportional spared recovery
Study-specific slope
[C] Constant recovery
Across-study intercept
R-squared: 5.8%
26 points constant recovery
Study-specific intercept
[D] Standard-form regression
Across-study intercept
R-squared: 31.5%
Across-study slope
Slope smaller than 1 indicating proportional to lost function recovery patterns
Study-specific intercept
Study-specific slope
Study 2
Study 3
Study 4
Study 6
[E] Bayesian model comparison
Standard-form regression
Proportional to lost
Constant
Proportional to spared
LOOCV estimate
In-sample estimate
Difference to best estimate
Deviance
worse

**Figure 7. Alleviating confounds by ceiling effects and mathematical coupling: Bayesian hierarchical models in the subset of FM-initial 10-45 (n=118). Marginal posteriors for parameters of *proportional to lost function* recovery [A], *proportional to spared function* recovery [B], *constant* recovery [C], and *(Unconstrained) Standard-form* regression [D]. [E] Final Bayesian model comparison using leave-one-out-cross-validation (LOOCV).** As deviance increases (rightward on x-axis), the accuracy of the fit goes down. Empty circles represent the LOOCV-corrected (out-of-sample) deviance, which is the key measure we use to compare models; black error bars indicate the corresponding standard error (i.e. uncertainty) in that deviance estimate. Grey triangles are the difference to the top-ranked model and grey bars the associated standard error. The lowest (i.e., best) LOOCV-deviance value is indicated by the vertical dashed grey line. Lastly, the filled black circles mark the models' in-sample deviances, which are susceptible to overfitting and, thus, not appropriate measures of accuracy. The standard-form regression model provided the best out-of-sample performance and was ranked first in the model comparison, closely followed by the *proportional to lost function* recovery model. However, the explained variance was low at 31.5% (standard-form regression) and 21.3% (*proportional to lost function* recovery).

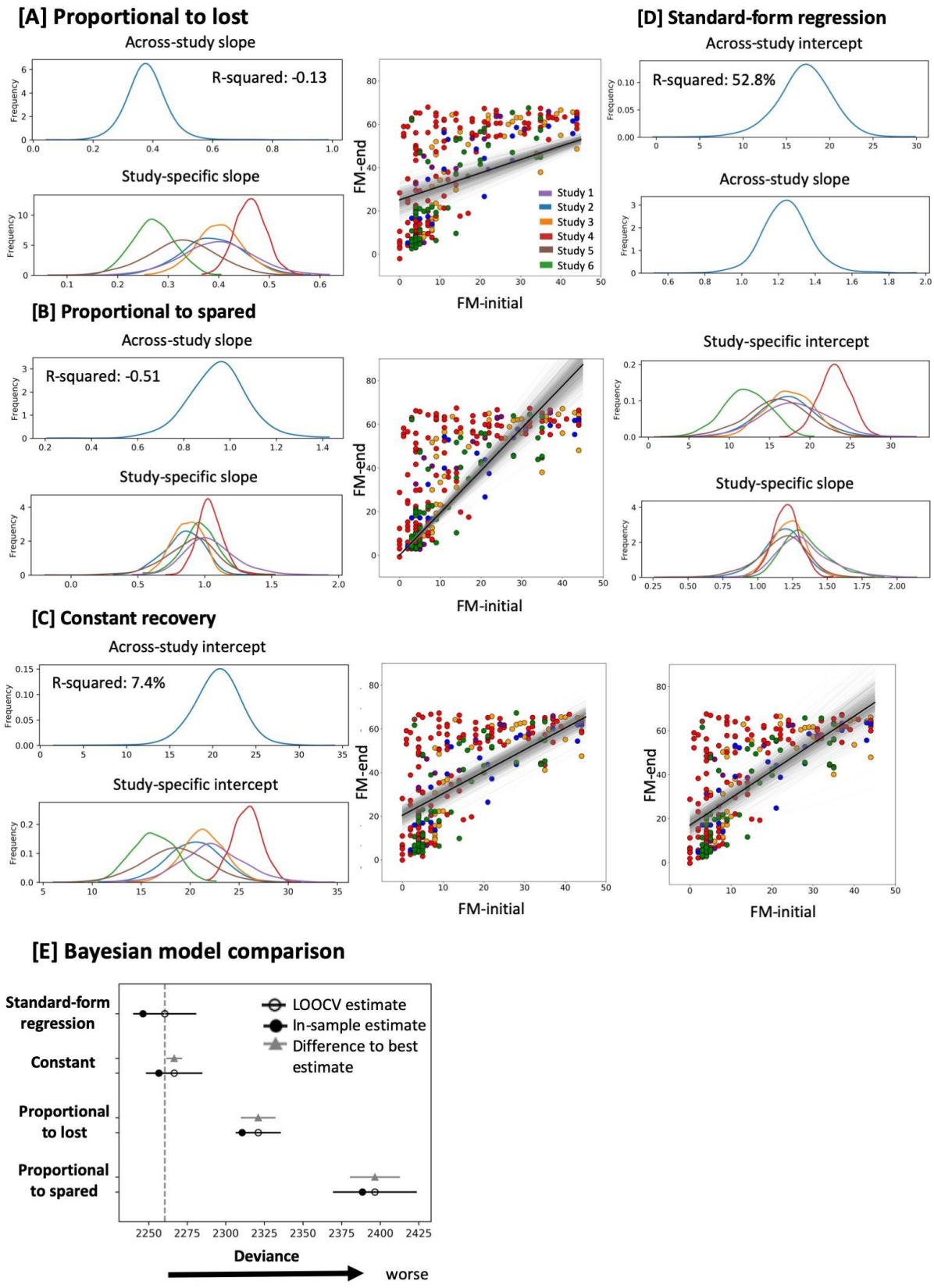

[A] Proportional to lost
Across-study slope
R-squared: -0.13
Frequency
Study-specific slope
FM-end
FM-initial
Study 1
Study 2
Study 3
Study 4
Study 5
Study 6
[D] Standard-form regression
Across-study intercept
R-squared: 52.8%
Across-study slope
[B] Proportional to spared
Across-study slope
R-squared: -0.51
Study-specific slope
Study-specific intercept
Study-specific slope
[C] Constant recovery
Across-study intercept
R-squared: 7.4%
Study-specific intercept
[E] Bayesian model comparison
Standard-form regression
Constant
Proportional to lost
Proportional to spared
LOOCV estimate
In-sample estimate
Difference to best estimate
Deviance
worse

**Figure 8. Hierarchical Bayesian analysis of *fitters* and *non-fitters* combined for FM-initial 0-45 (n=270). Recovery models: p*roportional to lost function* recovery [A], *proportional to spared function* recovery [B], *constant* recovery [C], and unconstrained standard-form regression [D].** Marginal posterior distributions are presented on the left-hand side for A – C and in the upper part of D. Distribution of *Initial* against *End* scores in conjunction with an overlay of sampled fits are added on the right-hand sides for A – C and in the lower part of D (*thick black line:* mean, *grey lines:* 2000 sampled marginal posterior parameter fits). **E. Final model comparison.** Based on leave-one-out-cross-validation, model comparison selected the standard-form regression model. Pure *constant* recovery was the best follow-up model.