Inflated Estimates of Proportional Recovery from Stroke: the Dangers of Mathematical Coupling and Compression to Ceiling

Howard Bowman (PhD)[1,2], Anna Bonkhoff (Dr med) [3], Tom Hope (PhD)[4], Christian Grefkes (Dr med)[5,6], Cathy Price (PhD)[4]

1 School of Psychology, University of Birmingham, Birmingham, UK

2 School of Computing, University of Kent, Canterbury, UK

3 J. Philip Kistler Stroke Research Center, Massachusetts General Hospital, Harvard Medical School, Boston

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

5 Department of Neurology, University Hospital Cologne, Cologne, Germany

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

Corresponding author: H. Bowman, Centre for Cognitive Neuroscience and Cognitive Systems and the School of Computing, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, UK, +44-1227-823815, email:H.Bowman@kent.ac.uk

Unstructured Abstract

The proportional recovery rule states that most survivors recover a fixed proportion (~70%) of lost function after stroke. A strong (negative) correlation between the initial score and subsequent change (outcome minus initial; i.e. recovery) is interpreted as empirical support for the proportional recovery rule. However, this rule has recently been critiqued, with a central observation being that the correlation of initial-scores with change over time is confounded in the situations in which it is typically assessed. This critique has prompted reassessments of patients' behavioural trajectory following stroke in two prominent papers. The first of these, by van der Vliet and collaborators

presented an impressive modelling of upper limb deficits following stroke, which avoided the confounded correlation of initial-scores with change.

The second by Kundert and collaborators reassessed the value of the proportional recovery rule, as classically formulated as the correlation between initial-scores and change. They argued that, while effective prediction of recovery trajectories of individual patients is not supported by the available evidence, group-level inferences about the existence of proportional recovery are reliable.

In this paper, we respond to the van der Vliet and Kundert papers by distilling the essence of the argument for why the classic assessment of proportional recovery is confounded. In this respect, we re-emphasize the role of mathematical coupling and compression to ceiling in the confounded nature of the correlation of initial-scores with change. We further argue that this confound will be present for both individual-level and group-level inference. We then focus on the difficulties that can arise from ceiling effects, even when initial-scores are not being correlated with change/ recovery. We conclude by emphasizing the need for new techniques to analyse recovery after stroke that are not confounded in the ways highlighted here.

Non-standard Abbreviations and Acronyms

FM-UE: Fugl-Meyer Upper Extremity

[Cover title: Inflated Estimates of Proportional Recovery]

Figures: 3. *Keywords*: stroke, proportional recovery from stroke, mathematical coupling, ceiling effects. *Word count*: 3977.

Inflated Estimates of Proportional Recovery from Stroke: the Dangers of Mathematical Coupling and Compression to Ceiling

The proportional recovery rule states that most survivors recover a fixed proportion (~70%) of lost function after stroke (Krakauer & Marshall, 2015 [1]). Originally developed to describe recovery from impairments of upper limb motor function (Prabhakan et al, 2008 [2]), the rule has since been applied to recovery from post-stroke impairments of lower limb function, language, and attention (see Hope et al, 2019 [3], for more discussion).

Empirical studies of the rule test patients twice: first within a couple of weeks of the stroke (initial/baseline) and then again some months later (outcome). A strong (negative) correlation between the initial score and subsequent change (outcome minus initial; i.e. recovery) is interpreted as empirical support for the proportional recovery rule. The rule is believed not to apply to a group of, so called, *non-fitters*, who have initial scores at the bottom of the Fugl-Meyer motor scale, while those above this range, to which the rule is believed to apply, are called *fitters*.

The proportional recovery rule (Prabhakan et al, 2008 [2]; Krakauer & Marshall, 2015 [1]) has recently been critiqued by Hope et al (2019) [3] and Hawe et al (2019) [4]. Central to these critiques is the observation that the correlation of initial-scores with change over time is a confounded measure in the situations in which it is typically applied (Hope et al, 2019 [3]), inflating effect sizes. These critiques have prompted reassessments of patients' behavioural trajectory following stroke. In particular, van der Vliet et al (2020) [5] present an impressive Bayesian mixture modelling of upper extremity Fugl-Meyer (FM-UE) measurements following stroke. Importantly, the authors avoid the confounded correlation of initial-scores with change by simply fitting their models to behavioural time-series, i.e. raw FM-UE scores, without involving the recovery measure. Accordingly, in a large sample of 412 patients, van der Vliet et al convincingly distinguish subgroups, including, for example, a subgroup that show very little evidence of any recovery and another one that recover rapidly.

Additionally, in another important recent piece of work, Kundert and collaborators (2019) [6] reassess the value of the proportional recovery rule, as classically formulated as the correlation between initial-scores and change. They argue that, while effective prediction of recovery trajectories of individual patients is not supported by the available evidence, group-level inferences about the existence of proportional recovery are reliable. We focus on the Kundert et al (2019) [6] publication here because it is recent and re-addresses the proportional recovery question, however, it should be emphasized that there are many previous publications that have similarly argued for the proportional recovery position, e.g. Winters et al (2015) [7]. In addition to reflecting on these two papers, we also attempt to distil the essence of the argument in Hope et al (2019) [3] in an intuitive and accessible manner. We first discuss the role of mathematical coupling in the confounded nature of the correlation of initial-scores with change. We then focus on the difficulties that arise from ceiling effects, even when initial-scores are not being correlated with change/ recovery.

Mathematical Coupling

In what follows, we denote initial-scores (typically obtained within 2 weeks of stroke) as X and outcome scores (typically obtained between 3 and 6 months after stroke) as Y. The pivotal correlation that underlies the proportional recovery hypothesis is that between initial-scores (X) and change, which corresponds to the amount recovered (i.e. outcome-scores minus initial-scores, Y-X). The confounded nature of this correlation is illustrated with an example in fig 1. This example demonstrates a phenomenon that could be called *compression enhanced coupling*.

Mathematically speaking, compression enhanced coupling arises from the fact that if the variability of X is substantially larger than the variability of Y, Y-X becomes close to –X+constant. This is just because Y contributes very little to the variability in Y-X, since its variation relative to X is very small. As a result, the correlation of X with Y-X degenerates, approaching the correlation of X with –X+constant, which, of course, is minus one, what would be interpreted as maximum evidence for proportional recovery (Fig 1).

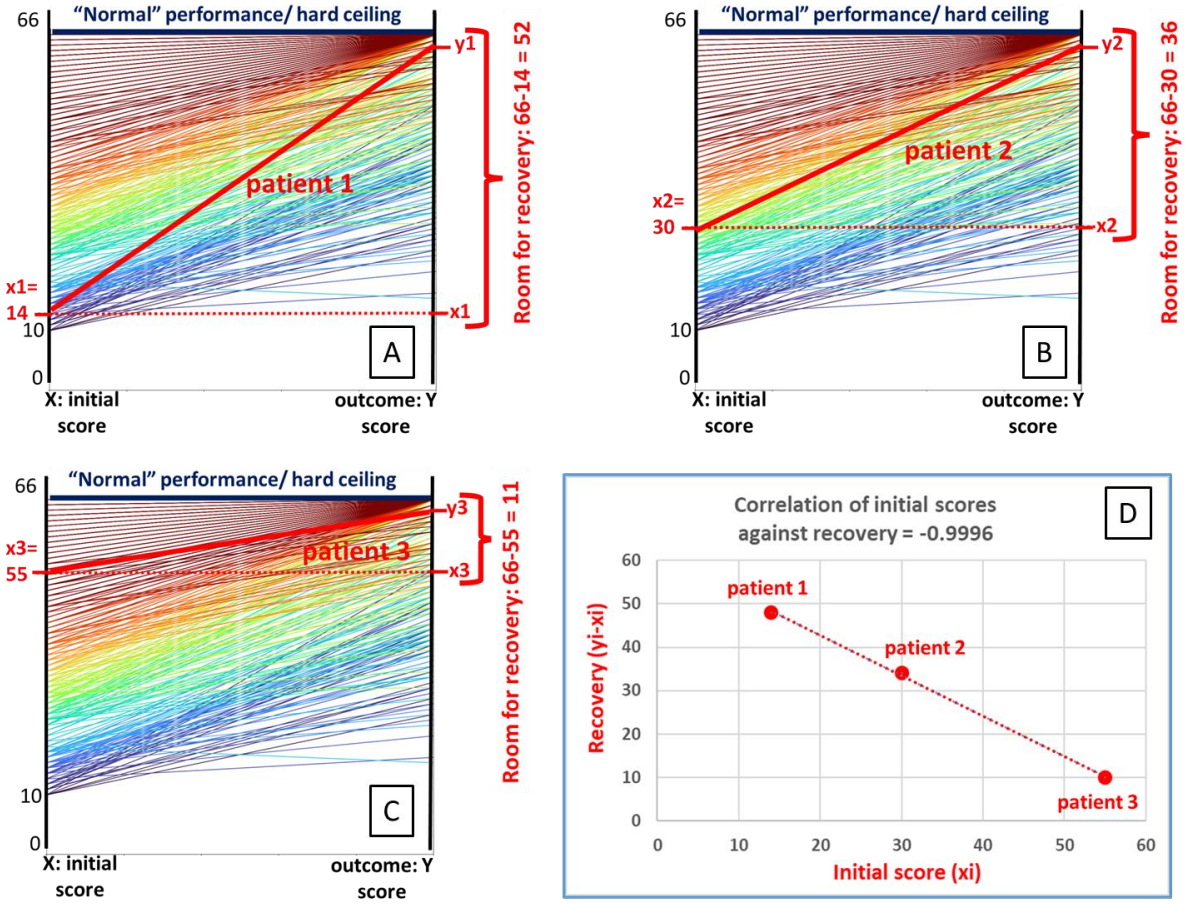

***Fig 1**: Illustration of compression enhanced coupling: typical pattern of recovery data, reproduced three times, each panel highlighting a different patient (1, 2 & 3 in panels A, B & C respectively). Data was generated from the unconstrained standard-form regression model in Bonkhoff et al (2020)* [8] *(Slope = 1.24, intercept = 17, i.e. Y = 1.24.X + 17). Outcome scores were generated from 400 randomly sampled initial scores, with Gaussian noise (SD 10) added, and values above 66 set to zero. Fugl-Meyer scores initially (X) and at outcome (Y) are shown in each panel, with non-fitters (0 to 10) removed, and the room available (i.e. potential) for recovery highlighted on the right side of each patient plot. The critical correlation that is tested is r(X,(Y-X)), with extreme negative values indicating proportional recovery. A key property that has to hold for r(X,(Y-X)) to be negative is that smaller values of X (lower on axis) should correspond to larger values of Y-X. The panels here illustrate why that cannot fail but be true, assuming the data has two properties: a) a ceiling and b) a general trend to get better. Both of these are true in this figure and invariably. Simply stated, if X is*

small, as it is for patient 1 (i.e. x1), ***there is room for recovery*** *(i.e. y1-x1 can be large), but if X is large, as it is for patient 3 (i.e. x3),* ***there is very little room for recovery*** *(i.e. y3-x3 must be small). That is, ceiling prevents patients with high initial-scores from* ***being able*** *to demonstrate their improvement: even if they have actually recovered capacities that place them well above the "Normal" level under Fugl-Meyer, their* ***measured*** *recovery will be bounded by ceiling, and thus small. In this way, a strong negative correlation would be observed, see panel D, and a proportional recovery pattern would be identified. This correlation would be due to compression to ceiling and may have little, if any, relationship to the true underlying recovery pattern in the data.*

The point about compression-enhanced ceiling is that one is not just correlating a variable with an expression containing that same variable (the definition of mathematical coupling), one is, near enough, correlating a variable with *itself* (strictly -1 times itself). To put it in the bluntest terms, if the variability of outcome scores is substantially smaller than initial-scores, there really is no need to calculate the correlation between initial-scores and change, we know exactly what it will be; it will be close to minus one. Furthermore, this change in variability is the norm in much of stroke research, since stroke samples often consist of a broad spectrum of patients ranging from mildly to severely affected patients, there is a general trend to get better and there is a ceiling on the scales employed to measure post-stroke impairment severity.

And, to be clear, the fact that a correlation approaching -1 is found does not mean proportional recovery is the true underlying relationship; it simply means that we do not know. In this situation, any underlying relationship will "look like" proportional recovery[1]. As an illustration, see fig 6A in Bonkhoff et al, 2020 [8], yellow lines (without ceiling) differ across recovery type, but green lines (with ceiling) show similar patterns for all recovery types. This raises the spectre of tautology – in other

[1] Additionally, although we focus on the correlation between X and Y-X, the same phenomenon would be observable if one regressed Y-X onto X. Also, if X is replaced with lost function, i.e. MAX – X, the polarity of effects changes, but the basic phenomenon is unchanged.

words, one cannot help but find evidence for proportional recovery, but that evidence is very often spurious.

As previously discussed, Kundert and colleagues (2019) [6] distinguish between prediction of recovery trajectories on a per-patient basis and group-level inferences. That is, “Can the outcome for an arbitrary patient be reliably predicted, as typically quantified with an R-squared effect size?” versus “Can an inference be made about the recovery pattern of the population, as quantified, for example, with a t- or F-test across the group?”. However, if compression to ceiling is present, the confound will be observed whether one is seeking to predict per-patient recovery or performing a standard group-level inference to test a hypothesis about the population. For example, if effect sizes (spuriously) increase, all else equal, group-level statistical significance will also (spuriously) increase. The inflation we demonstrated in Hope et al (2019) [3], was of the R-squared variance explained, a standard measure of effect-size, reflecting the ability to predict per-patient recovery. Inflating R-squared in this way, will typically, reduce the p-value of a group-level test (since t- or F-values would increase). This will increase the probability that a significant finding will be incorrectly reported (i.e. the null rejected, when it should not be).

Additionally, while they acknowledge the possibility that heteroscedasticity and ceiling effects could be present in some cases, Kundert et al (2019) [6] argue that classic (correlation of initial-score with change) proportional recovery data patterns can be found in several published studies. However, the observation that proportional recovery is frequently reported in the literature is consistent with our central argument that given the common presence of compression to ceiling, a range of different underlying data relationships will *look like* proportional recovery when correlating initial-score with change.

Kundert et al (2019) [6] go further to illustrate their point, by presenting a simulated canonical proportional recovery pattern (see Fig 2[A]) and argue, on the basis of visual-inspection, that similar trends can be observed in published recovery studies. As a representative example of these studies,

we focus on the largest study that Kundert et al (2019) [6] refer to, Winters et al (2015) [7], the data from which is shown in Fig 2[B]. Kundert et al (2019) [6] argue that this data shows a plausible resemblance to the canonical version.

From our perspective, there looks to be a strong compression to ceiling in Fig 2[B], which is not present in Fig 2[A], noting that, in the presentation format used in Kundert et al (2019) [6], the filled diagonal line represents ceiling. In particular, compression enhanced coupling only requires a compaction *towards* the filled line, rather than a placement of all data at absolute ceiling. Such compaction to ceiling looks to us to be a prevalent feature of the Fig 2[B] scatter plot.

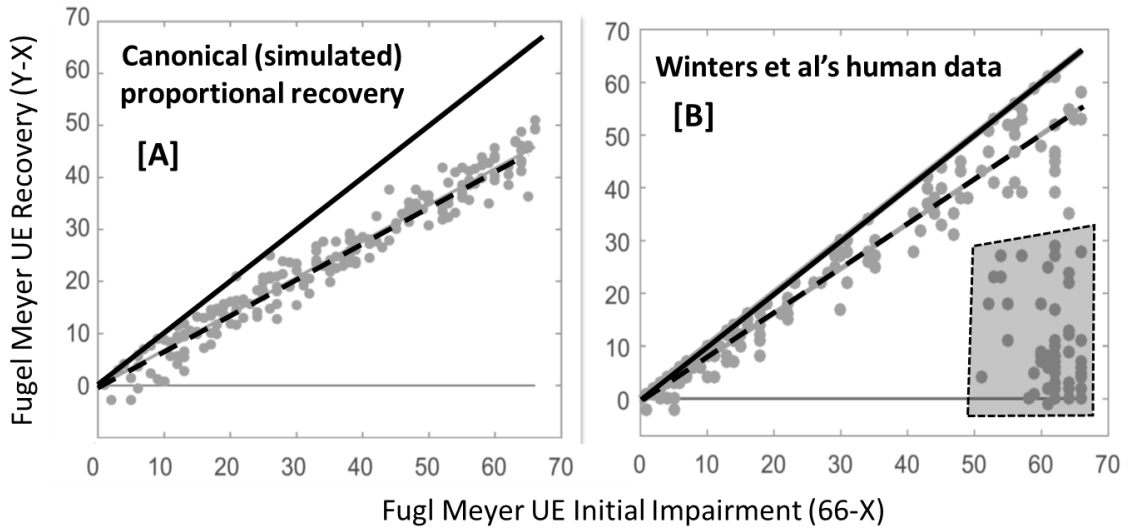

Fig 2*: proportional recovery versus ceiling, as presented in Kundert et al (2019) [6]. Performance is measured using the Fugl-Meyer Upper Extremity scale. To stay consistent with Kundert et al (2019) [6], data in both panels is plotted differently from elsewhere in this article. X-axes show initial impairment, i.e. maximum minus performance-score (66-X), while Y-axes show change, i.e. recovery (Y-X). In this form, ceiling is the filled black diagonal line and lines of best fit are in dashed black. [A] Canonical proportional recovery pattern simulated by Kundert et al (2019) [6]. [B] Observed pattern of human data from Winters et al (2015) [7], showing the compression to ceiling that can generate spurious evidence for proportional recovery. Data points in bottom right corner (overlaid by dashed*

shape with grey fill) are non-fitters and are not relevant to the current discussion. Panels are adaptations of those in Kundert et al (2019) [6].

However, more importantly, the fact that two different groups of researchers, Kundert et al. and us, can look at the same data and arrive at two different conclusions – classic proportional recovery pattern versus compression to ceiling – highlights the essential difficulty: because of the confounded nature of quantification using the initial-scores to change correlation, we are left making qualitative judgements on the basis of visual inspection. Accordingly, for the field to move forward, methods need to be devised to mitigate confounds on quantitative assessments of proportional recovery.

Ceiling Effects without Mathematical Coupling

Van der Vliet et al (2020) [5] respond to the confounded nature of correlating initial-score with change, by focussing on the patient's raw performance score at multiple time points. Specifically, they fit exponential curves to the trajectory of performance through time; see Fig 3[a]. This approach does indeed, avoid mathematical coupling, since the change variable plays no part.

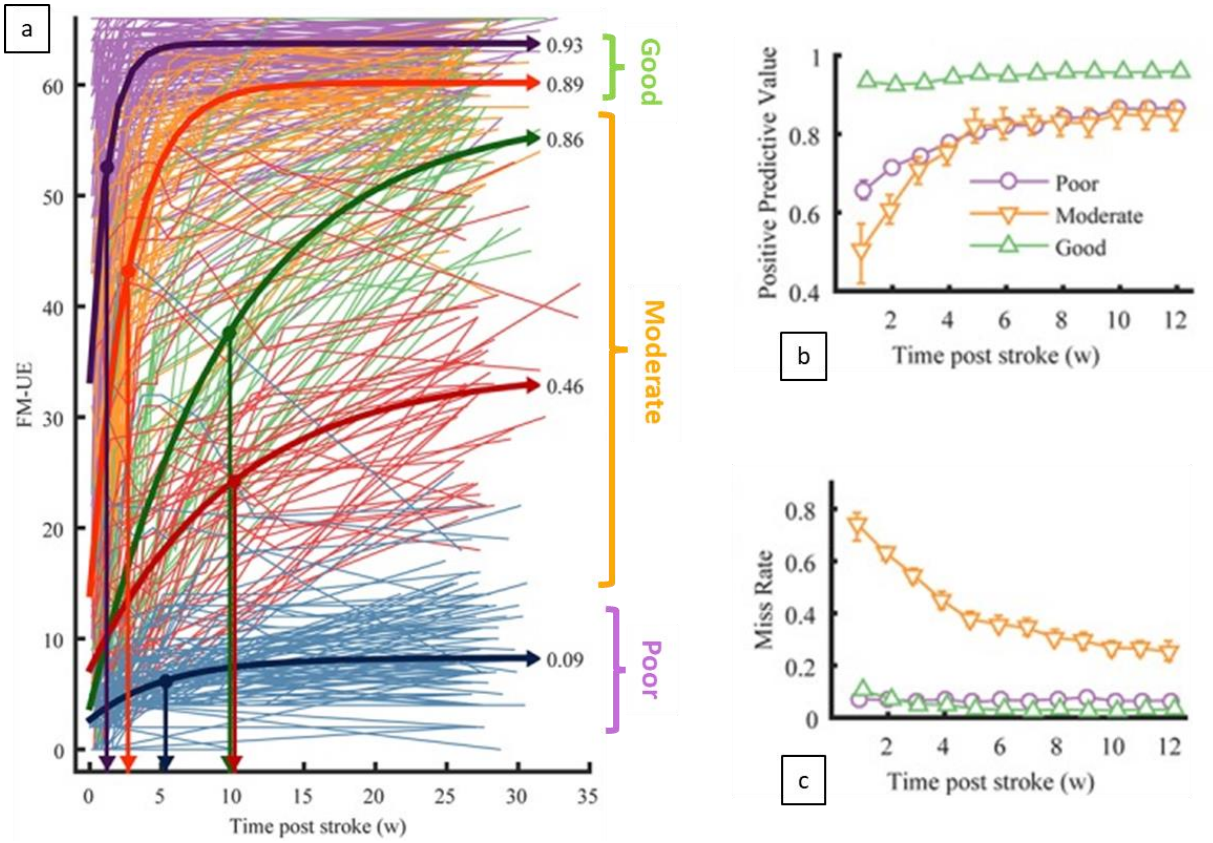

***Fig 3**: Ceiling effects in van der Vliet et al (2020) [5]: Fig 1(A) (our panel [a]) and Fig 2(H,I) (our panels [b,c]) from van der Vliet et al (2020) [5]. [a] Representation of data, with five groups identified by Bayesian mixture modelling, shown in different colours, and curves showing model fits to exponentials. Patients' performance is measured on the FM-UE scale. These five groups were combined into three, which are indicated by brackets and named **Good**, **Moderate** and **Poor** (note, colour coding changes from original groups). [b,c] Accuracy of fits. Critically, the **Good** group shows a substantial ceiling effect, while **Moderate** does not. Consistent with this, posterior predictive value (panel [b]) and misses (panel [c]) are close to perfect for **Good**, but much less for **Moderate**.*

While an important step forward, there may be one sense to which fitting curves to performance scores through time still leaves an inflated sense of the variability being explained. Avoiding the change variable in model fitting removes the problem of mathematical coupling, but it does not remove the ceiling effect.

All statistical measures, in one way or another, trade-off explained and unexplained variability, with the statistic increasing as unexplained variability decreases, relative to explained variability. A ceiling effect is, essentially, a loss of sensitivity to observe variability, which leads to a belief that there is less variability than is really the case. The particular ceiling effect present in FM-UE recovery data, whereby outcome scores are compressed at the top of the range, see Fig 1 (concentration of data points at top of Y side of plots A, B and C) and Fig 2[B], leads to the possibility that unexplained variability is underestimated. This is simply because the true variability amongst data points at ceiling cannot be observed, since all data points are compressed to a small region. This underestimation of the true variability will also make classification more straightforward, since the patients clustered at ceiling will be classified together into the *Good* group.

Van der Vliet et al (2020) [5] fit (nonlinear) exponential curves (see Fig 3[a]), which fit almost perfectly (see Fig 3[b,c]) for their *Good* group. We contend that this is largely because the *Good* group is at ceiling at the end time point. The *Moderate* group, who are not at ceiling, show much reduced accuracy.

One can view the involvement of ceiling in these fits in two contrasting ways.

1) *Scale bounded prediction*: one can accept the limitations of the FM-UE scale, particularly its limited capacity to distinguish mild impairments, and predict within its bounds. Accordingly, the nonlinear exponential models employed by Van der Vliet et al (2020) [5], do indeed, represent the path to a maximum relatively well. In this sense, they predict patients' recovery trajectories effectively. This is an impressive contribution to research on stroke recovery, with potentially significant clinical value.
2) *Inflation due to ceiling*: however, these fits do not resolve the problem that the data does not accurately represent the variability amongst a large number of patients, especially those that are only mildly affected. As a result, the exponential fits give high estimates of accuracy and variability explained, but these high estimates arise, at least in part, because the fitting

problem being solved is easier than it may at first sight seem. This is simply because data points are being compressed at ceiling. In this sense, the prediction accuracy may not be as impressive as it, initially, seems. A consequence may be an inflated assessment of the effectiveness of quantifying recovery from purely behavioural measures.

One response to these difficulties in quantifying the variability explainable by modelling behavioural recovery trajectories, is to add further steps to the analysis process. An approach that is applied in Bonkhoff et al (2020) [8], is to discard data at ceiling through a subsetting procedure, allowing an unconfounded (or, at least, less confounded) assessment of variability explained. Once this is done, only a weak proportional recovery pattern remains. Additionally, if such a subsetting procedure were applied to the van der Vliet et al (2020) [5] data, the Good group would largely be discarded, leading to a more modest model accuracy.

Van der Vliet et al (2020) [5] report out-of-sample tests of their models, i.e. they test their fitted models on unseen data. This gives considerable assurance that reported accuracies are not inflated due to over-fitting to the data. However, it is important to note that this does not protect against the confounds we are discussing here. In particular, the problem we have highlighted is not a problem of the process of model fitting, it is a problem with the data, which, in the sense discussed above, does not enable the question being asked of it to be assessed. Thus, in a statistical sense, any out-of-sample data will be confounded in the same way as the in-sample data, and fitting accuracies will be similarly inflated.

Conclusions

In summary, the question of the profile of behavioural recovery after stroke and the variability thereby explained, without benefiting from the information offered by other data modalities, such as structural MRI, remains moot. This will stay the case until the way in which behavioural data are collected and/or analysed changes in order to prevent confounding effects of 1) mathematical coupling (due to correlating initial-scores with change) and 2) compression to ceiling. From the

published literature on recovery from motor deficits following stroke, which is bedevilled by ceiling effects, we do not yet know the answer to this question. One strategy for dealing with compression to ceiling is to employ methods to discard or transform away ceiling. This, though, has to be applied with caution, since there is a substantial risk of biasing the data sample. This issue was directly addressed in Bonkhoff et al (2020) [8], where synthetic data simulations were used to justify the procedure employed to discard data at ceiling.

Alternatively, as suggested in Hope et al, 2019 [3], scales could be developed and deployed that do not induce such gross ceiling effects. The problem of ceiling is especially common for behavioral scores, for which the maximum score indicating full recovery is fixed, e.g. Fugl-Meyer assessment score, the Action Research Arm Test (Lang, Wagner, Dromerick & Edwards, 2006 [9]) or the National Institutes of Health Stroke Scale (Lyden, Lu, Levine, Brott & Broderick, 2001 [10]). Indeed, a number of researchers have noted the susceptibility of the Fugl-Meyer to ceiling effects. For example, Gladstone, Danells & Black (2002) [11] discussed the possibility of adapting the scale to make it more sensitive at its top end, in particular by making it more sensitive to different grades of muscle strength. There could also be more focus on behavioural assessments with an open upper limit, such as finger tapping frequencies or grip strengths, although, even here there are biological upper limits.

In conclusion, it may be that proportional recovery does explain some variance in the data, but we will only know that if and when the impact of compression to ceiling is resolved. Bonkhoff et al (2020) [8] provide one indicative finding, which suggests that only a weak proportional recovery pattern remains once ceiling is dealt with.

Disclosures statement: there are no disclosures to make.

References

1) Krakauer JW, Marshall RS. The proportional recovery rule for stroke revisited. Annals of neurology. 2015;78(6):845-7.

2) Prabhakaran S, Zarahn E, Riley C, Speizer A, Chong JY, Lazar RM, Marshall RS, Krakauer JW. Inter-individual variability in the capacity for motor recovery after ischemic stroke. Neurorehabilitation and neural repair. 2008;22(1):64-71.

3) Hope TM, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. Recovery after stroke: not so proportional after all? Brain 2019;142(1):15-22.

4) Hawe RL, Scott SH, Dukelow SP. Taking proportional out of stroke recovery. Stroke. 2019;50(1):204-11.

5) van der Vliet R, Selles RW, Andrinopoulou ER, Nijland R, Ribbers GM, Frens MA, Meskers C, Kwakkel G. Predicting upper limb motor impairment recovery after stroke: a mixture model. Annals of Neurology. 2020;87(3):383-93.

6) 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;33(11):876-87.

7) Winters C, van Wegen EE, 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):614-22.

8) Bonkhoff AK, Hope T, Bzdok D, Guggisberg AG, Hawe RL, Dukelow SP, Rehme AK, Fink GR, Grefkes C, Bowman H. Bringing proportional recovery into proportion: Bayesian modelling of post-stroke motor impairment. Brain. 2020;143(7):2189-206.

9) Lang CE, Wagner JM, Dromerick AW, Edwards DF. Measurement of upper-extremity function early after stroke: properties of the action research arm test. Archives of physical medicine and rehabilitation. 2006;87(12):1605-10.

10) Lyden PD, Lu M, Levine S, Brott TG, Broderick J. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials. Stroke. 2001;32(6):1310-7.

11) 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(3):232-40.