Statistical limitations on drawing inferences about proportional recovery.

Keith R. Lohse[1,2], Rachel Hawe[3], Sean Dukelow[3], & Stephen H. Scott[4]

[1.] Department of Health, Kinesiology, & Recreation; University of Utah

[2.] Department of Physical Therapy and Athletic Training; University of Utah

[3.] Department of Clinical Neurosciences; University of Calgary

[4.] Department of Biomedical and Molecular Sciences; Queen's University

Corresponding Author:
Keith Lohse, PhD
e: rehabinformatics@gmail.com
p: 801-585-7226
250 S 1850 E, Rm 258
Salt Lake City, UT, 84112

Contents: 0 Tables; 6 Figures;

Acknowledgments: The authors would like to thank Dr. Kristin Sainani and Dr. Thomas Hope for their thoughtful comments on an early draft of this manuscript.

Disclosure Statement: The authors received no funding specifically to pursue this work. SHS is the co-founder and Chief Scientific Officer of Kinarm that commercialize robotic technology for neurological assessment. All other authors have no conflicts of interest to declare.

Abstract

Background. Numerous studies have found large statistical relationships between the amount of recovery and initial impairments in people with stroke. When change scores are regressed onto initial impairments, the resulting slope is approximately ≈ 0.7 for a variety of outcomes. These findings have led to the 70% "proportional recovery rule" and the argument that proportional recovery represents a biological phenomenon. Previous studies of proportional recovery are confounded by statistical limitations that come from regressing change scores onto initial impairments in bounded scales.

Objective. Our goal is to show that data claimed as evidence for *proportional* recovery are generally consistent with *random* patterns of recovery, once statistical limitations are taken into account.

Methods. Using a pooled dataset of N = 373 Fugl-Meyer Assessment (FMA) upper extremity scores extracted from published literature, we ran simulations to illustrate three main arguments: (1) Mathematical coupling renders the traditional null-hypothesis significance test irrelevant in proportional recovery studies; (2) Proportional recovery is one of many alternative hypotheses; (3) Current evidence claimed in favor of proportional recovery is consistent with uniform random recovery.

Results. Our simulations show that if all data were included (no exclusion of "non-fitters") regressing change scores onto initial impairments in a bounded scale would lead to a slope of ≈ 0.5. Similarly, cluster analysis will spuriously identify groups of fitters and non-fitters, leading to a slope for the fitters of ≈ 0.7, when the underlying recovery is random.

Conclusions. These results cast doubt on the validity of "proportional recovery" as a population level-statistic and a biological phenomenon.

MeSH Keywords: Stroke; Rehabilitation; Methods; Data Analysis

Introduction

Recently, much ink has been spilt on the topic of the proportional recovery rule in stroke rehabilitation.[1] In its broadest sense, the proportional recovery rule posits that the amount of recovery patients are likely to have is roughly 70% of the total possible recovery they could make, on average, after the exclusion of "non-fitters" to the rule.[2,3] This relationship is usually demonstrated by regressing change scores (a terminal assessment minus the baseline assessment) onto the initial amount of impairment. Not surprisingly, severely impaired individuals show the greatest variation in their potential for recovery, and severely impaired individuals who do not recover very much are classified as "non-fitters" to the general rule. Cumulative real data to this effect are shown in Figure 1 with non-fitters shown in red. Classification of non-fitters has been based on different methods[4] that rely on either a statistical classification (e.g., outlier detection[1]) or based on physiologically relevant outside variables (e.g., cortico-spinal tract integrity[5]).

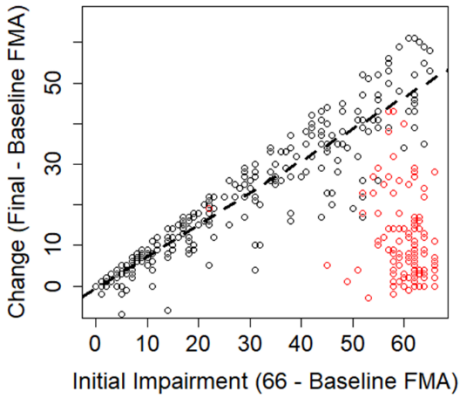


Figure 1. *Data adapted from Hawe et al. (2019). Change scores and initial impairments extracted from empirical studies have been combined to create an "overall" sense of the relationship across studies. The dashed line denotes the ordinary least squared regression line for all fitters (black points). Data-points that were identified as non-fitters in the original studies are shown as red points.*

Other manuscripts have discussed the problems with regressing change scores onto impairment and concerns with the sub-group analysis of fitters and non-fitters in some statistical

detail.[6,7] The short form of these critiques is that change scores are mathematically "coupled" to baseline impairment, because change scores contain the baseline score. Depending on the relative variance of the change scores to baseline impairment, this introduces a statistical artifact (i.e., a spurious correlation that is due strictly to the subtraction, not a true relationship) that may be quite large or relatively small. The presence of this artifact indicates that evidence for 70% proportional recovery is, at the very least, overstated. In response, Kundert et al[4] authored an article in favor of the proportional recovery rule. Kundert and colleagues' response incorporates some previous critiques and seeks to refute other criticisms in their discussion, ultimately concluding that proportional recovery is a real biological phenomenon and representative of spontaneous recovery. In their abstract, Kundert and colleagues conclude that, "existing data in aggregate are largely consistent with the [Proportional Recovery Rule] as a population level model for upper limb motor recovery; recent reports of its demise are exaggerated, as these excessively focus on the less conclusive issue of individual subject level predictions."

In this point of view, we concisely show that this assertion is mathematically incorrect. At the population level, patterns that have been claimed as evidence for proportional recovery are generally consistent with random recovery, once one accounts for artifacts that result from regressing two statistically dependent and bounded variables onto each other. Using simulations, we illustrate this problem visually, relying on as little formal mathematics as possible. We think this simulation-based approach makes the critique more intuitive and accessible to a general audience.

Below, we critique the evidence in favor of the proportional recovery rule based on three main arguments:

1. **Mathematical coupling renders the traditional null-hypothesis significance test irrelevant in current proportional recovery studies**. Simple difference scores have long been regarded as a sub-optimal method for ascertaining change over time. This is especially true when difference scores are regressed onto baseline scores, which creates a mathematical coupling.[8-10]
2. **Proportional recovery is one of many alternative hypotheses (as opposed to the null hypothesis) and researchers should directly test competing hypotheses.** Through simulation, we can show that regressing change scores onto baseline scores for any two bounded scales can lead to a population level slope of 0.5, even when change scores for individuals are random. This is because floor/ceiling effects in the data mean that change will necessarily be proportional, and the magnitude of this relationship is driven upwards if perceived "non-fitters" to the rule are excluded from the data.
3. **Current evidence claimed in favor of proportional recovery is also consistent with uniform random recovery using cluster analysis to identify "fitters"**. Again, using simulations informed by empirical data, we can test what data we should expect to observe if the underlying change is uniformly random, rather than proportional. Using hierarchical cluster analysis to identify non-fitters in our simulations, we can show that data from N=373 real stroke patients on the Fugl-Meyer Assessment is consistent with random recovery. As such, current data do not support the claim that recovery is proportional any more than that recovery is random.

In the Discussion, we focus on some of the positive evidence from the proportional recovery literature and suggest productive ways to move forward analytically. For instance, neuroanatomical differences do have strong associations with the potential for recovery at

different levels of impairment.[5,11,12] However, regressing change scores onto baseline scores is rife with statistical problems. We recommend that if researchers want to explain individual differences in recovery over time, then we should be using formal conditional longitudinal models with more data points and avoiding the statistical confounds of change scores.[13,14] Indeed, a number of researchers have started making strides in this direction, using longitudinal methods to explore trajectories of stroke recovery.[15,16] Understanding which factors explain, or better yet predict[17-19] stroke trajectories is a very important area of research. These research questions are, however, quite orthogonal to the proportional recovery rule. For instance, the finding that cortico-spinal tact integrity partially explains where a person ends up on the spectrum of recovery is very distinct from claiming that on average people recover ~70% of lost function or the further claim that this 70% pattern reflects spontaneous biological processes.

Argument 1: Mathematical Coupling Renders the Traditional Null-Hypothesis Test Irrelevant

Difference scores have been critiqued for many years in the biomedical literature as a method for capturing change.[20] The reason for this is that difference scores implicitly assume a one-to-one relationship between pre-test scores and post-test scores. This implicit assumption can be seen more explicitly if we contrast the formula for a linear regression controlling for baseline (Eq 1.) against a linear regression in which difference scores are the outcome (Eq 2.):

Eq 1. $Post_i = \beta_0 + \beta_1(Pre_i) + \epsilon_i$

Eq 2. $Diff_i = (Post_i - Pre_i) = \beta_0 + \epsilon_i \equiv Post_i = \beta_0 + (1)Pre_i + \epsilon_i$

In Eq 1., the relationship between pre-test scores and post-test scores is weighted based on the correlation between time-points observed in the data (ultimately creating the regression

coefficient β_1). When taking difference scores, however, pre-test scores are moved to the other side of the equals sign and subtracted from post-test. This subtraction implicitly assumes a one-to-one relationship between pre-test and post-test. This can be seen more clearly if we move pre-test out of the difference and back to the other side of the equals sign; taking difference scores is thus equivalent (≡) to assuming that the value of $\beta_1 = 1$, as shown in red in Eq 2. This might be a reasonable assumption when the correlation between pre-test and post-test is very high, but in general it is much better practice to control for pre-test as a covariate. First, controlling for pre-test allows for regression to the mean whereas difference scores do not. That is, random error for lower scoring participants is likely to drive their scores upward on a second measurement (and vice versa for high scoring participants). The high reliability of clinical tests makes regression to the mean much less of a concern, but it is still a problem for difference scores that is mitigated by controlling for pre-test. Second, controlling for pretest allows for β_1 to be weighted based on the correlation between pre-test and post-test, whereas taking a difference score does not. When the correlation between pre-test and post-test is very low the weight of $\beta_1$is simply reduced and there are no negative consequences for the model. Furthermore, when the correlation between pre-test and post-test is very low, taking difference scores will actually add noise to the data.[20]

This problem is related to the issues of mathematical coupling discussed by Hawe et al.[6] and Hope et al.[7]. Briefly, we illustrate the negative effects of mathematical coupling in Figure 2 using both a normal distribution (2A/B) and a uniform distribution with clear boundaries (2C/D). The point of this illustration is to show that mathematical coupling is a different effect from "boundaries" on a scale, although the two can be related when floor/ceiling effects are present. In both simulations (N = 1000 data points), the variables X and Y are totally independent (r=0.0). However, when we calculate a new variable $Z = Y - X$, we find that Z and X have a strong

negative relationship (r=-0.7). The reason for this is that Z and X are mathematically "coupled"; that is, Z contains X so they are intrinsically linked. This can be seen a little more clearly if we rearrange the terms for Z in a regression equation:

Eq. 3. $Z_i = -X_i + Y_i = \beta_0 + \beta_1(X_i) + \epsilon_i$

Recognizing that Z is just: $(-X + Y)$, it shouldn't be surprising that $-X$ and X are negatively related. In fact, the only thing distorting their relationship is Y. As Hope et al.[7] pointed out, this is why the relative variance in X and Y matters. If the variance in Y is vastly *smaller* than X, we are essentially regressing $-X$ onto X. If the variance in Y is vastly *bigger* than X, we are essentially regressing Y onto X. If we know how big these relative variances are, then we can control for the artifact we are introducing into our data. If we do not know these variances, then we are introducing an unknown artifact. In either case, we are making our analysis needlessly complex, because we did not need to regress change scores onto baseline to begin with. Note the population-level "true" correlation between X and Y also matters, but in our case where X and Y are independent, the artifact is determined by their relative variances.

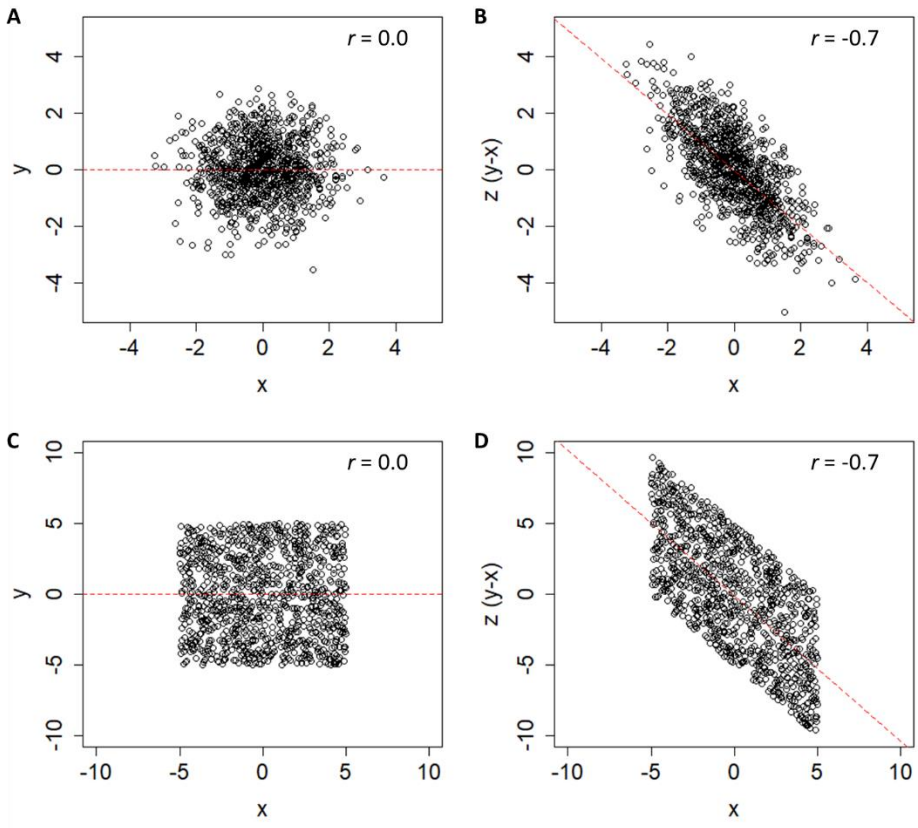


Figure 2. *N = 1000 simulated data showing two uncorrelated variables, X and Y, and third variable Z, computed from their difference. In panels A and B, these variables are based on two normally distributed, but otherwise unbounded distributions. In panels C and D, these variables are based on two bounded uniform distributions. In both cases, an artifactual negative relationship exists between X and Z, because those values are mathematically coupled.*

This mathematical coupling creates a problem for the interpretation of data on proportional recovery, where change scores are regressed onto baseline. As shown in Figure 1, we have adapted the data gathered by Hawe et al.[6] These data reflect several different available studies on proportional recovery using the Fugl-Meyer Assessment.[2,11,21-24] Regressing the N = 373 change scores onto baseline levels of impairment shows an overall slope of 0.42 when the non-fitters (as identified in past studies) are included in the data. If these non-fitters are excluded, then the slope of the regression line is shifted upward, to 0.76 (as shown by a dashed black line

in Figure 1). In either case, this slope is statistically different from zero, *p's* < 0.001, but due to mathematical coupling it is not surprising to find a statistically significant relationship in this situation. There are, however, two very different questions at work here.

One question is if there is a relationship between Change and Initial Impairment. This question could be answered by a classic null hypothesis significant test, "Assuming there is no relationship between Change and Initial Impairment, what is the probability of observing the slope (or a more extreme slope) in the sample?" Note that the null-hypothesis significance test also assumes that sampling variability is the only factor affecting the data. As we have shown briefly and as past work has explained in detail,[6,7] it is not surprising to reject the null hypothesis in this situation due to a mathematical artifact. That is, the mathematical coupling between Change and Initial Impairment scores forces a correlation onto the data. This means that random sampling is not the only factor at work, nor should one expect a "true" relationship of zero, making the traditional null-hypothesis significance test rather meaningless. However, it would still be possible to conduct a meaningful non-zero hypothesis test, if the mathematical artifact could be estimated.

The second question is if the relationship between change scores and initial impairment is truly proportional? This is a distinct question from rejecting the null hypothesis. This question requires us to choose between competing *alternative* hypotheses. As shown in Figure 3, if we want to argue that recovery is *proportional* (Figure 3A), we need to be able to distinguish it from *random* recovery (Figure 3B). Whether recovery is proportional or uniform, a relationship is already assumed. Instead, the challenge becomes distinguishing proportional recovery from random recovery when we are dealing with two bounded scales. Uniform random recovery is a reasonable alternative hypothesis in this situation, because it allows for the fact that different

strata of recovery exist,[12,25,26] but that the overall distribution of recovery covers the entire available space (e.g., as shown in animal data[27,28]). In contrast, proportional recovery asserts the distribution of change scores is biased towards a certain level and advocates of the rule argue this level is about 70% of initial impairment across various scales and measures.[4] In these simulations, our uniform randomness is epistemic and not ontological. That is, we assume that variation is due to factors that remain to be explained; we are not assuming that recovery following stroke is inherently random for each person.

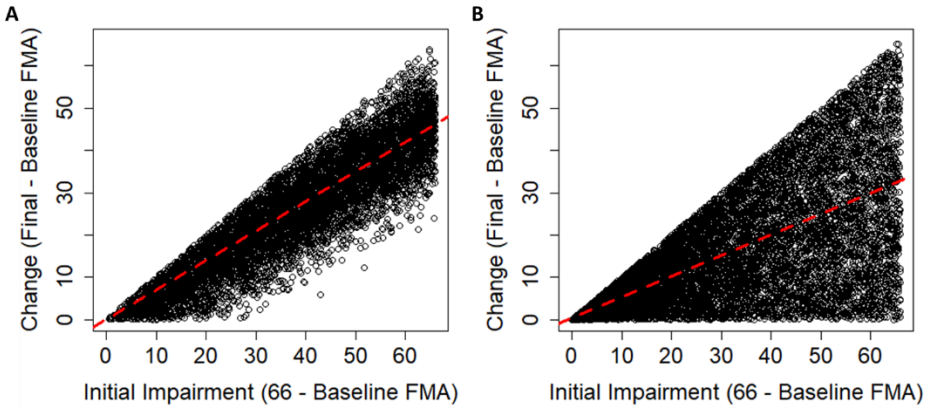


Figure 3. *(A) Simulated data in which the change in Fugl-Meyer Assessment (FMA) is proportional to recovery. These data are normally distributed around 70% of initial impairment with a standard deviation of 6 points and truncated by the ceiling of the FMA. Proportional recovery could be simulated with different parameters, but this is meant only to be an example of proportional recovery. (B) Simulated data in which the change in Fugl-Meyer Assessment (FMA) is uniformly random. Data are uniformly distributed between the initial impairment (no recovery) and maximum points allowed on the FMA (maximum possible recovery). In both cases there is an upper bound on recovery due to the nature of the FMA. Note that even when change is random, this creates a positive slope.*

Argument 2: Proportional Recovery is One of Many Alternative Hypotheses

The fact that proportional recovery is apparent across many different scales of measurement has been argued as evidence for proportional recovery being a neurobiological phenomenon.[4] First shown in the Fugl-Meyer Assessment (FMA[1]), proportional recovery has

since been shown in the Western Aphasia Battery[29] and the Letter Cancellation Test,[30] among other inventories. However, all these inventories still possess lower bounds and upper bounds. Although the individual minima and maxima are all different, the presence of these boundaries creates a real problem for interpreting the relationship between baselines scores and change scores. The code provided in the supplemental materials can be revised to demonstrate this point, but one can also consider Figure 3B in a thought experiment. Regardless of what the individual minima and maxima of these different scales are, random recovery will always lead to the bottom triangle of the possible space being filled. Fitting a regression line through that space will always have a slope of ~0.5 because (assuming a uniform distribution of change scores) there will always be an equal mass of change scores above and below the regression line. As such, it would not be unusual to see 50% "proportional recovery" in bounded scales, even when change was random.

However, as shown in Figure 4A, we can see that the story is more complicated than that because the distribution of initial impairments is not uniform. This can be seen more easily in the histogram of initial impairments in Figure 4B. There are higher densities of very low and very high impairments. Therefore, to make our simulations more realistic, we bootstrapped (i.e., repeatedly sampled) the initial impairment data from Hawe et al.[6] shown in Figure 4B, to get a new "population" of 10,000 initial impairments with a similar distribution shown in Figure 4C. (Note the relative heights of the bars is similar, but the absolute number of observations is much larger). Based on this realistic distribution of initial impairments, we created 10,000 uniformly random change scores. As shown in Figure 4D, the change is random but not proportional. This distribution of uniform change looks different from the pattern of change shown in the pooled data assembled by Hawe et al.[6] especially for participants in the middle range of initial

impairments (between 20 and 40 points). However, uniform change is a reasonable alternative hypothesis to test against, given that more uniform patterns have been observed in the animal literature[27,28] and other human studies that have shown a much more uniform pattern of recovery.[3,31] When recovery is uniformly random, there is still an overall slope of 0.50 (shown as a dashed red line) due to the constraints of the bounded scale. Thus, the slope of 0.50 is not "proportional recovery" because recovery at any given level of impairment is uniformly random. Instead, this slope is a statistical artifact created by Change scores being regressed onto Initial Impairments in a bounded scale.

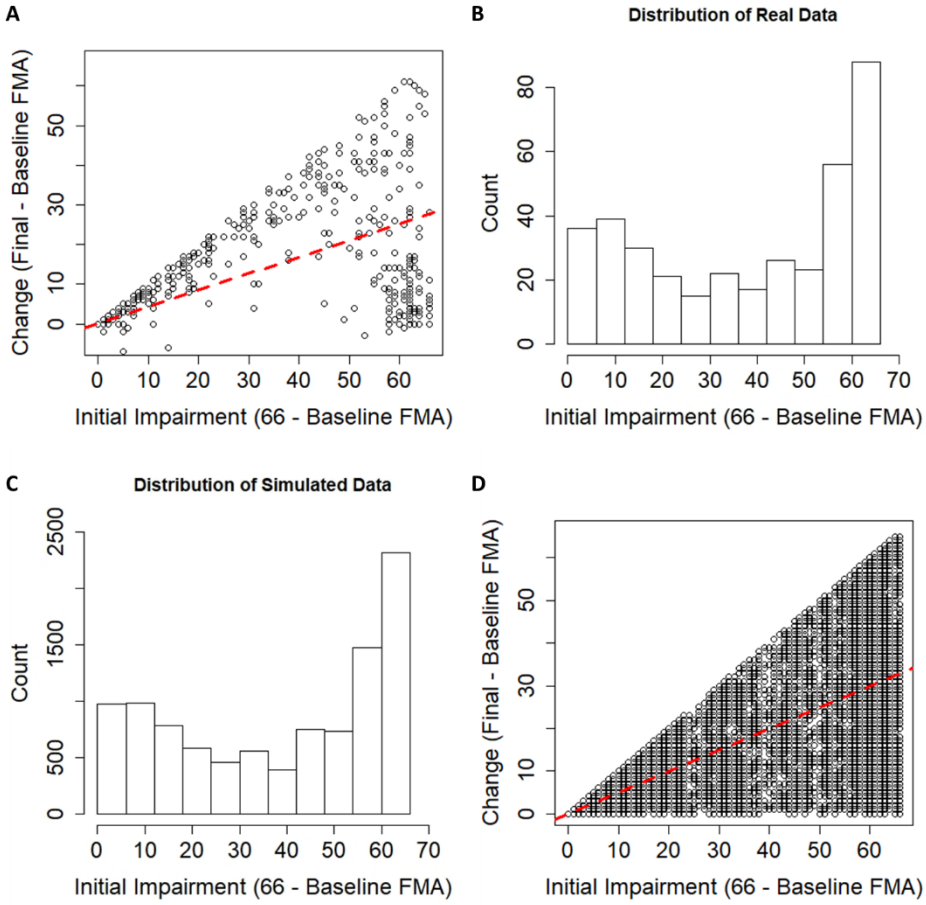


Figure 4. *(A) Real data adapted from Hawe et al. (2019). The slope of the regression line is 0.42 when all the data are included (dashed red line). (B) The univariate distribution of initial impairment scores. (C) After bootstrap resampling of the real data, we can get a simulated distribution of initial impairments with very similar properties. (D) After adding random change scores to our simulated impairment data,*

we get a simulated population with realistic initial impairments but with necessarily random change. The dashed red line is the ordinary least squares regression line with a slope of 0.50, fit to all the data.

This does not necessarily mean that any relationships between change scores and initial impairments are completely artifactual. Instead, these simulations show that regressing bounded scores onto each other will create an artifact, and we need to statistically test if our observed results are different from that artifact. As shown in Figure 5A, for instance, we can see a sample of N=30 observations randomly drawn from our population of FMA scores. The slope and intercept for this sample are 0.62 and -1.00, respectively. Conducting a traditional null-hypothesis significance test, we would conclude that this slope is statistically different from zero with $p<0.001$. As such, one might conclude that this slope is evidence for proportional recovery. However, this conclusion is incorrect because zero is not the slope against which we should test; that slope is 0.5. This can be seen in the sampling distribution of slopes in Figure 5B. If we repeatedly sample from our population of random data and fit a unique slope to each sample, we get a density of slopes centered around 0.5 (the population slope) but spreading out in either direction due to sampling variability.

As shown in the right-hand panel of Figure 5C, changing from a null hypothesis significance test against 0 (dashed black curve) to a minimal effects test against 0.5 (dashed blue curve) can drastically change our interpretation. Assuming the null hypothesis was true, the probability of observing a slope of 0.62 or greater due to sampling variability alone was $p<0.001$. However, if we center our "null" distribution around 0.5 rather than 0, we find that the probability of observing a slope of 0.62 or greater is $p=0.377$, which would not be considered statistically significant by conventional standards (df=28, standard error of the slope=0.132).

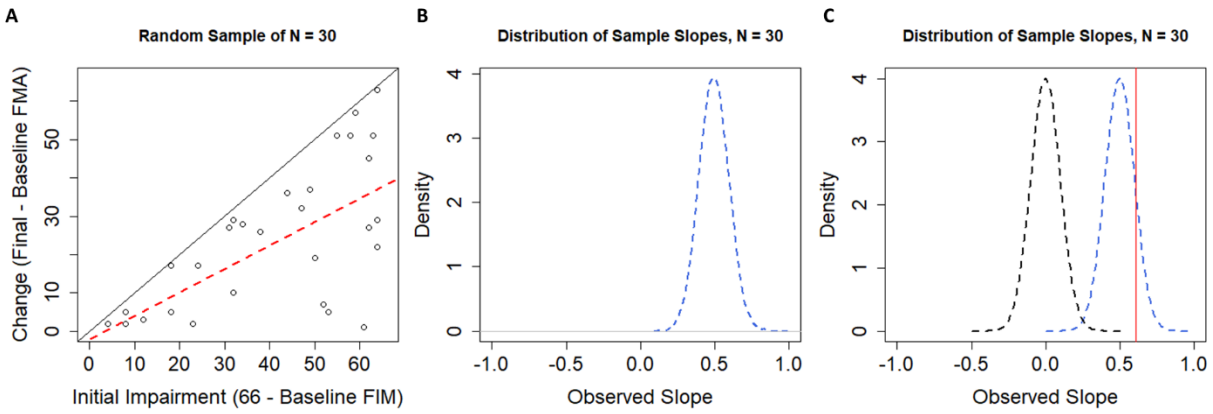


Figure 5. *(A) A single random sample of N=30 FMA scores drawn from our population. The regression line (dashed red) has a slope of 0.62, which might suggest proportional recovery were it not drawn from a sample of randomly generated data. A diagonal black line with a slope of 1 is shown for reference. (B) The sampling distribution of slopes when our simulated population was sampled 10,000 times, with replacement, at sample sizes of N=30. (C) Contrasting the distribution of sample slopes under the null-hypotheses (dashed black line) and the distribution of sample slopes from our simulated population (centered on 0.5; dashed blue line). Note that now our observed slope of 0.62 (shown as the vertical red line) is no longer statistically significant given the correct distribution (p<0.001 compared against zero; p=0.132 compared against 0.5).*

Importantly, we emphasize that this artifactual value of 0.5 emerges when all data are included. If participants are excluded for being "non-fitters" (i.e., the mostly severely impaired, least improving participants), this will bias the mean of the distribution upward. Thus, 0.5 would be a reasonable value to test against if a single regression line is fit to all of the data, but if any exclusions were made, this will introduce an upwards bias whose exact nature depends on the method of exclusion. In the next section we will deal with how "fitters" versus "non-fitters" have typically been determined and the effect this has on the data. For the moment, however, let us statistically compare the real data in Figure 4A to the simulated data in Figure 4D. For the pooled N=373 FMA scores assembled by Hawe et al.,[6] the slope of the regression line is 0.42 with a standard error of 0.018. The slope 0.42 is actually statistically smaller than what we would expect given uniform random recovery (i.e., <0.50; $t(371)=-4.60$, $p<0.001$). This suggests that average recovery following stroke is actually smaller than we would expect assuming uniform

random recovery. This reduction might arise from the fact that in reality, we do see FMA scores decrease for some participants (2.4%)[6], whereas our simulations only assumed positive change. Alternately this finding could suggest that recovery is simply lower than we anticipate across the spectrum of impairment.

Argument 3: Real Data for "Fitters" are Consistent with Uniform Random Recovery

The slope of 0.42 observed when all available FMA data are included pre-supposes that non-fitters to the proportional recovery rule belong in the same category as fitters. Clearly this violates a central assumption of the proportional recovery model. Past work on proportional recovery argues that fitters and non-fitters are two clearly discernable classifications. Further, the methods that have been used to establish fitters from non-fitters are legitimate methods whether they are data-driven methods (like hierarchical cluster analysis) or theory driven methods (such as moderator analyses using physiological data[2,12]). As mentioned above, however, our random recovery model assumes epistemic variability not ontological variability, so physiological variables can still explain individual differences in recovery but at the population-level recovery is uniform. As such, the key question is if the very high slopes found among fitters are different from what we would expect if Change scores were uniformly distributed.

To answer this question, we will use our bootstrapped distribution of Initial Impairments with uniform change scores shown in Figure 4D. We sampled from this population of individuals for 10,000 samples of N=30 (with replacement). In each sample, we used hierarchical cluster analysis[32,33] to identify clusters of participants who could be classified as fitters and non-fitters. As shown in Figure 6A, even when we are sampling from data with uniform Change scores the

clustering algorithm identifies clusters of participants who would be classified as fitters (black dots) and non-fitters (red dots). This is problematic for the proportional recovery argument, because when change is uniform, there is no proportional rule to which individuals can fit. As such, we should be concerned that fitters and non-fitters may be (at least partially) an artifactual classification.

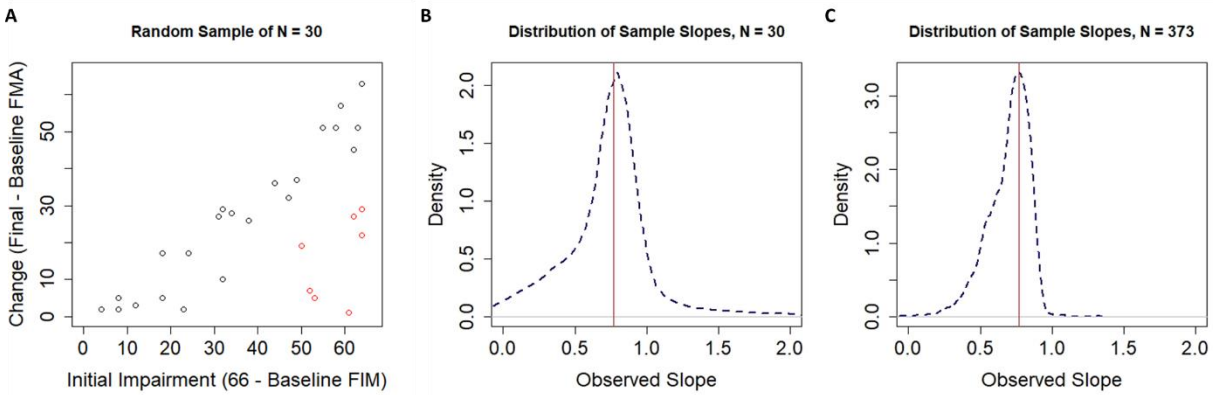


Figure 6. *(A) A single random sample of N = 30 participants drawn from a population with random Change scores. Note the clustering algorithm still classifies participants into what look like fitters and non-fitters even when there is no "rule" to which individuals can "fit". (B) The distribution of sample slopes for fitters identified by our clustering procedure when the original sample size was N=30. (C) The distribution of sample slopes for fitters identified by our clustering procedure when the original sample size was N=373.*

Current procedures described in studies of proportional recovery are not entirely clear on how their clusters were ascertained. That is, a cluster analysis can work using either a bottom-up agglomerative procedure or a top-down divisive procedure, but in general authors should have an objective criterion for where they stop in determining their clusters. Although it has been stated that a criterion has been used[23] it is not clear what the numeric value of this criterion is.[2,23,30] As such it is not clear by what criteria authors are making the decision to stop at two clusters in their

analyses. But we do know that authors are calculating Mahalanobis distances between points[34] and that these distance values are being used as input into the agglomerative clustering algorithm advocated by Ward.[33,35] In the absence of a clear criterion by which we should stop clustering, we ran cluster analyses in our simulations that always stopped at two clusters. In order to determine which cluster was the "fitters", we chose the cluster with a higher mean Change score (consistent with the central argument of proportional recovery). Otherwise, our simulations used identical methods to the published literature for the calculation of distances and determination of clusters.

As shown in Figure 6A, our clustering procedure leads to an identification of fitters and non-fitters to the proportional recovery rule consistent with past literature. Despite coming from a population of uniform random recovery, the clustering procedure spuriously identifies fitters (black dots) and non-fitters (red dots) in this random sample of N=30. Below, we show the effects of this procedure (sampling, clustering, and fitting slopes to the fitters cluster) when the total size is N=30 (to illustrate a relatively small, but common sample size) and when the total N=373, matching the sample size for the pooled data.[6]

As shown in Figure 6B, when the initial sample size was set to N=30, the distribution of sample slopes for the "fitters" had a mean of 0.770, a median of 0.781, and a negative skew. Thus, under this sampling distribution, we would not find it surprising to observe a large positive slope of 0.769 (as was observed in Hawe et al.[6]). Specifically, with a starting sample size of N=30, we'd expect a slope of ≥0.769 for the fitters about 54% of the time. As such, observing those slopes in real data should not be construed as evidence for proportional recovery; a researcher could get identical slopes if the underlying pattern of recovery was random, but they excluded data based on a cluster analysis.

We can see this conclusion most robustly if we take the pooled data for the fitters.[6] The slope for the n=254 fitters out of those initial N=373 subjects was 0.769. As shown in Figure 6C, assuming uniform random recovery, hierarchical cluster analysis with two clusters, and that fitters would be the group with higher mean Change, the mean of this sampling distribution was 0.778 and the median was 0.771. Based on this distribution we would expect to get a slope of ≥0.769 for the fitters about 54% of the time. As such, we conclude that the data invoked as evidence for the proportional recovery rule is relatively weak. These patterns are quite consistent with what one might expect if recovery was uniformly and randomly distributed.

Discussion

Patterns argued to show proportional recovery in stroke rehabilitation have been found across a wide variety of assessments and replicated in many different samples. The pattern appears to be so pervasive and the relationship so strong that it is compelling to think this pattern is a neurological rule. In the words of Carl Sagan, however, "extraordinary claims require extraordinary evidence"[36] and before we make strong deterministic claims about proportional recovery, we need to be very sure that we are ruling out alternative explanations. As we have shown, current patterns claimed to be evidence for *proportional* recovery are generally consistent with uniform *random* recovery in these measures. This does not mean that proportional recovery has been disproven, but it does mean that we do not have adequate evidence to claim proportional recovery at this time.

There is very compelling evidence that physiological variables explain individual differences in recovery,[19] but this is a very different question from claiming that recovery is

"proportional". Thus, rather than concluding that individuals with lower cortico-spinal tract integrity are more likely to be "non-fitters", we think a more appropriate conclusion is that individuals with lower cortico-spinal tract integrity are likely to show minimal recovery. As we discuss above, uniform random change scores do not mean that we are assuming that stroke recovery is a stochastic process. There are going to be individual differences in recovery and people with more similar neuroanatomy following stroke are likely to show more similar patterns of recovery. That said, it is important to consider the wide variation in recovery trajectories for neuroanatomically similar individuals.[26] Regardless, we do not have evidence that these individual differences are centered around 70% of initial impairment or that 70% recovery represents a robust biological phenomenon. There could be a non-zero and biologically meaningful relationship between initial impairment and recovery, but there is not sufficient evidence in the literature to claim proportional recovery is real.

Limitations

Our simulations used an empirical distribution of initial impairment values, but they assumed a random distribution of change scores for any given level of impairment. One could question the appropriateness of assuming uniform random change at-all, but especially in the context of the FMA upper extremity subscale. Visually comparing pooled real data in Figure 4A with our simulated data in Figure 4D, there appears to be an "island" of individuals with high levels of initial impairment and very limited levels of recovery in the real data that is not present in the simulated data. This island is likely created by nonlinearities in the FMA.[6,37] Specifically, mid-range scores are less likely to occur in the FMA upper extremity subscale. This can be seen clearly in the distribution of initial FMA scores (Figure 4B/C). As these mid-level scores are less likely, it makes sense that moderately impaired individuals will progress out of this range, but

severely impaired individuals will be more likely to either surpass it or struggle to get into it (creating the island of non-fitters). Hawe et al.[6] took this *bimodal* distribution of change scores into account in their simulations, but in the current study we explicitly chose to model change *uniformly* for three reasons.

First, the artifacts that are generated by regressing bounded change scores onto baseline scores are not a product of this nonlinearity (as shown in our simulations). Nonlinearity is a special concern for certain scales, but the problems generated by ceiling effects are more general. Second, data from both animal and human studies suggest that more uniform patterns of recovery exist for a variety of scales.[3,27,31] As such, uniform random recovery is a valid alternative hypothesis against which to test. Third, assuming uniform random recovery illustrates how using hierarchical cluster analysis with Mahalanobis distances can break down in this situation. When pairwise distances are calculated based on predicted change and actual change, these cluster analyses will spuriously identify fitters and non-fitters. In our simulations, we did sometimes identify other types of clusters, especially at small sample sizes. We explored different methods for excluding these poorly ascertained clusters to see how it would affect the distribution of sample slopes for the fitters (e.g., rejecting samples where a cluster was less than 5% of the total sample size; rejecting samples where clusters were separated based purely on initial impairment). These steps affected the tails of the sampling distribution, but in all cases the distribution was centered near 0.7. Thus, although other processing decisions could have been made, of all of the processing decisions we explored, obtaining a slope of 0.7 for the fitters' cluster was quite likely even when recovery was uniform and random.

Conclusions

Regressing change scores onto initial impairments creates a number of statistical limitations that preclude a substantive interpretation, especially in bounded scales. We recommend that future work seeking to demonstrate proportional recovery should (1) test these effects against reasonable alternative hypotheses rather than against a hypothesis of no-effect, and (2) abandon the practice of regressing change scores onto initial impairments. Rather than teasing out statistical artifacts due to coupling, we would recommend that researchers avoid these unnecessary problems entirely. Mixed-effect regression models, for instance, would allow researchers to study factors that affect change over time in a more robust way that avoids the issue of mathematical coupling.[38] Further, mixed-effect models would allow researchers to model change based on more data points, irregularly timed data points, individuals with missing data, and even model change non-linearly if needed.[39] Mixed-effect regression models are not a panacea of course, they have their own limitations and are computationally more complex, but we do think they would be a more appropriate tool teasing apart individual differences in stroke recovery. Recent work has made strides in this direction[16] and we think this work is an excellent example of how we can meaningfully investigate variation in recovery following stroke without burdening ourselves with the mathematical artifacts of difference scores.

Finally, lest one think we are waging war against the very idea of proportional recovery, we agree that this is an important area of research. Further, we agree that there could be a very real, very meaningful relationship between initial impairment and the recovery trajectory. Important strides have been made in our desire to understand recovery even if recovery turns out not to be proportional. We hope that both sides of this debate will continue to push our understanding of stroke recovery forward. At the moment, however, there is not compelling

evidence for a 70% proportional recovery rule at the population level for upper limb motor recovery. Similarly, the statistical limitations raised in this commentary would apply to change scores from any bounded scales being regressed onto initial impairments. Even if the distribution of change scores was not truly uniform (e.g., a more bimodal distribution may result from nonlinear scales), ceiling effects in the outcome measure would create the illusion of proportionality. These statistical artifacts either need to be identified and tested against (e.g., with minimal effects testing) or avoided altogether by adopting more appropriate methods for longitudinal data analysis.

References

1. 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 & Neural Repair*. 2008;22(1):64-71.

2. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity. *Annals of Neurology*. 2015;78(6):848-59.

3. Veerbeek JM, Winters C, van Wegen EE, Kwakkel G. Is the proportional recovery rule applicable to the lower limb after a first-ever ischemic stroke?. PloS One. 2018;13(1):e0189279.

4. Kundert R, Goldsmith J, Veerbeek JM, Krakauer JW, Luft AR. What the Proportional Recovery Rule Is (and Is Not): Methodological and Statistical Considerations. *Neurorehabilitation & Neural Repair*. 2019;15:1545968319872996.

5. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. Proportional motor recovery after stroke: Implications for trial design. *Stroke*. 2017;48(3):795-8.

6. Hawe RL, Scott SH, Dukelow SP. Taking Proportional Out of Stroke Recovery. *Stroke*. 2019;50(5):204-211.

7. Hope TM, Friston K, Price CJ, Leff AP, Rotshtein P, Bowman H. Recovery after stroke: not so proportional after all? *Brain*. 2019;141:15-22.

8. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chronic Dis*. 1962; 15:969–977.

9. Gill JS, Zezulka AV, Beevers DG, Davies P. Relation between initial blood pressure and its fall with treatment. *Lancet*. 1985;1:567–569.

10. Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. *Stat Med*. 2007; 26:443–457. doi: 10.1002/sim.2538

11. Feng W, Wang J, Chhatbar PY, Doughty C, Landsittel D, Lioutas VA, Kautz SA, Schlaug G. Corticospinal tract lesion load: an imaging biomarker for stroke motor outcomes. *Annals of Neurology*. 2015;78(6):860-870.

12. Smith MC, Byblow WD, Barber PA, Stinear CM. Proportional recovery from lower limb motor impairment after stroke. *Stroke*. 2017;48(5):1400-1403.

13. Long JD. *Longitudinal data analysis for the behavioral sciences using R*. Sage; 2012.

14. Singer JD, Willett JB. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press; 2003.

15. Lohse K, Bland MD, Lang CE. Quantifying change during outpatient stroke rehabilitation: a retrospective regression analysis. Archives of physical medicine and rehabilitation. 2016;97(9):1423-30.

16. van der Vliet R, Kwakkel G, Andrinopoulou E, Nijland R, Frens M, van Wegen E, Meskers C, Slaman J, Ribbers G, Selles R. Promoting Study Power of Stroke Rehabilitation Trials Using a Longitudinal Mixture Model. Abstract published in the proceedings of the *American Society for Neurorehabilitation*. T7.

17. Burke Quinlan E, Dodakian L, See J, McKenzie A, Le V, Wojnowicz M, Shahbaba B, Cramer SC. Neural function, injury, and stroke subtype predict treatment gains after stroke. *Annals of Neurology*. 2015;77(1):132-45.

18. Stinear CM, Ward NS. How useful is imaging in predicting outcomes in stroke rehabilitation? *International Journal of Stroke*. 2013 Jan;8(1):33-7.

19. Stinear, CM, Smith, MC, Byblow, WD (2019). Prediction tools for stroke rehabilitation. *Stroke.* 2019;50:3314-3322.

20. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. *British Medical Journal*. 2001;323(7321):1123-4.

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

22. Guggisberg AG, Nicolo P, Cohen LG, Schnider A, Buch ER. Longitudinal structural and functional differences between proportional and poor motor recovery after stroke. *Neurorehabilitation & Neural Repair*. 2017;31(12):1029-41.

23. Winters C, van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. *Neurorehabilitation & Neural Repair.* 2015;29(7):614-22.

24. Zarahn E, Alon L, Ryan SL, Lazar RM, Vry MS, Weiller C, Marshall RS, Krakauer JW. Prediction of motor recovery using initial impairment and fMRI 48 h poststroke. *Cerebral Cortex*. 2011;21(12):2712-21.

25. Lin DJ, Cloutier AM, Erler KS, … Cramer SC. Corticospinal tract injury estimated from acute stroke imaging predicts upper extremity motor recovery after stroke. *Stroke*. 2019;50:00-00. DOI: 10.1161/STROKEAHA.119.025898.

26. Findlater SE, Hawe RL, Mazerolle EL, … Dukelow SP. Comparing CST lesion metrics as biomarkers for recovery of motor and proprioceptive impairments after stroke. *Neurorehabilitation & Neural Repair*. 2019;1-14. DOI: 10.1177/1545968319868714

27. Jeffers MS, Karthikeyan S, Corbett D. Does stroke rehabilitation really matter? Part A: proportional stroke recovery in the rat. *Neurorehabilitation & Neural Repair*. 2018;32(1):3-6.

28. Jeffers MS, Karthikeyan S, Gomez-Smith M, Gasinzigwa S, Achenbach J, Feiten A, Corbett D. Does stroke rehabilitation really matter? Part B: An algorithm for prescribing an effective intensity of rehabilitation. *Neurorehabilitation & Neural Repair*. 2018;32(1):73-83.

29. Lazar RM, Minzer B, Antoniello D, Festa JR, Krakauer JW, Marshall RS. Improvement in aphasia scores after stroke is well predicted by initial severity. *Stroke*. 2010;41(7):1485-8.

30. Winters C, Van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the maximum proportional recovery rule to visuospatial neglect early poststroke. *Neurorehabilitation & Neural Repair*. 2017;31(4):334-42.

31. Ward NS, Brander F, Kelly K. Intensive upper limb neurorehabilitation in chronic stroke: Outcomes from the Queen Square programme. *J Neurol Neurosurg Psychiatry*. 2019;90:498–506.

32. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

33. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*. 2014;31(3):274-95.

34. McLachlan GJ. Mahalanobis distance. *Resonance*. 1999;4(06).

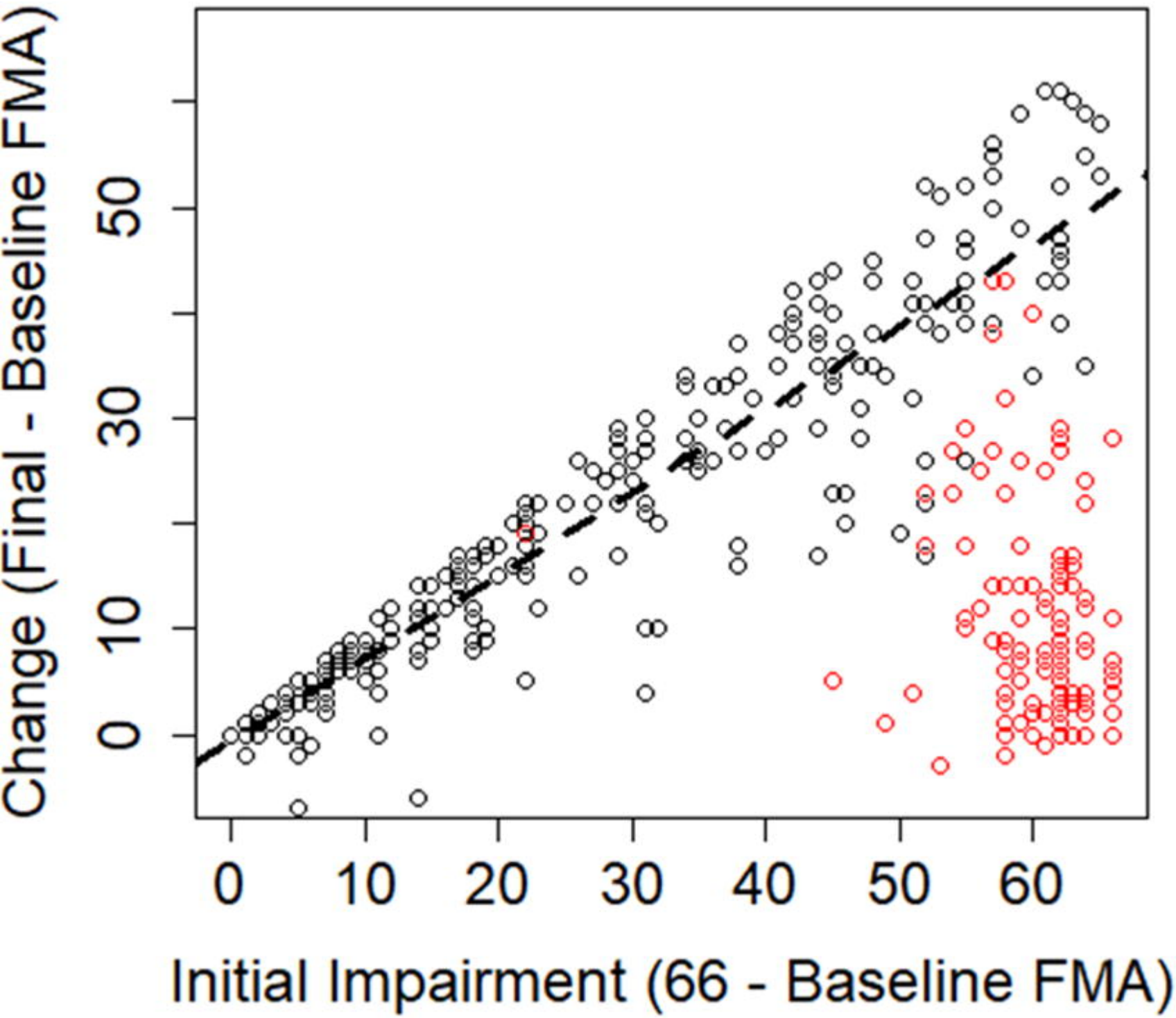
35. Ward Jr JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 1963;58(301):236-44.

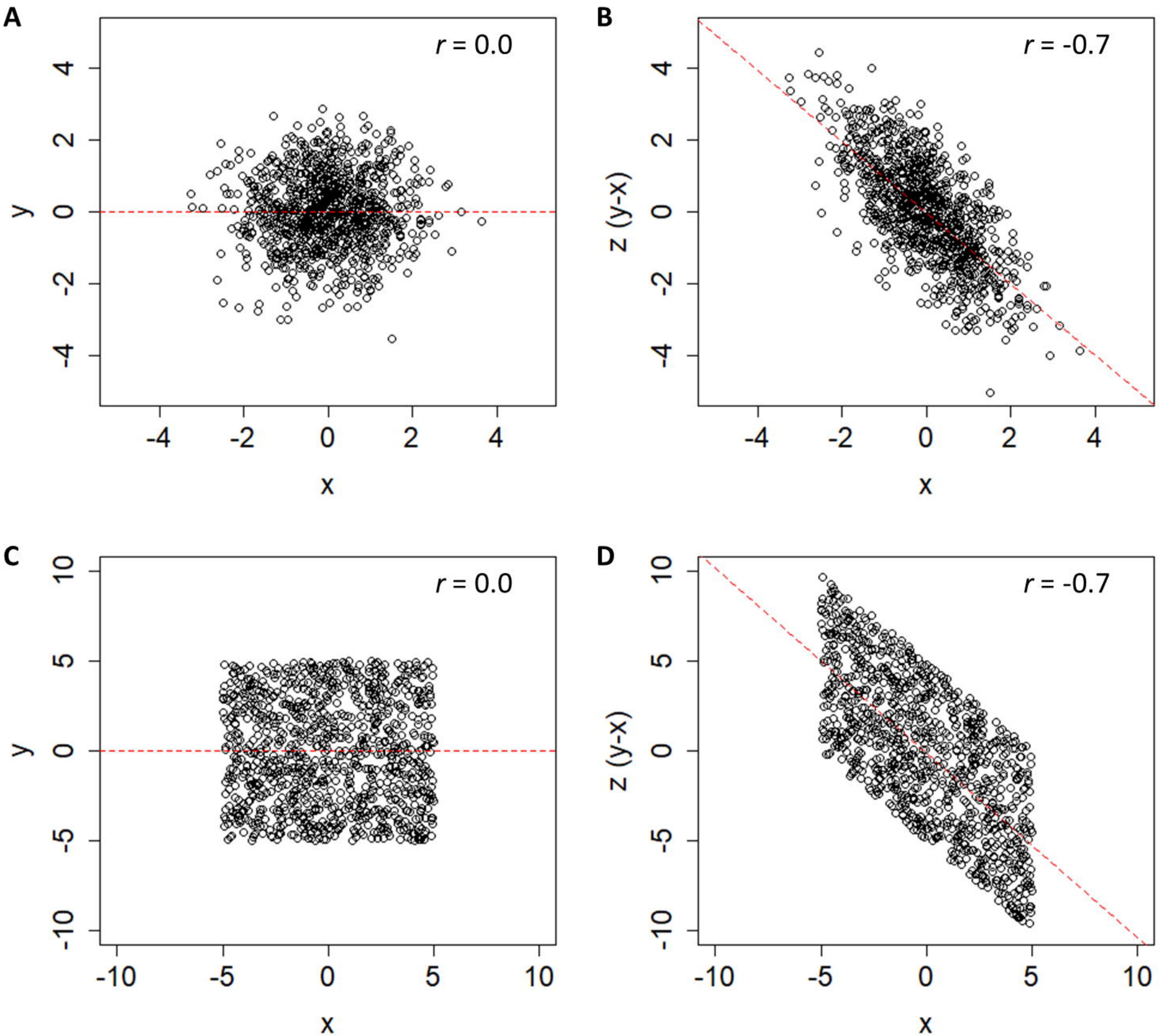
36. Sagan C. "Encyclopaedia Galactica". Cosmos: A Personal Voyage. December 14, 1980. Public Broadcasting Service.

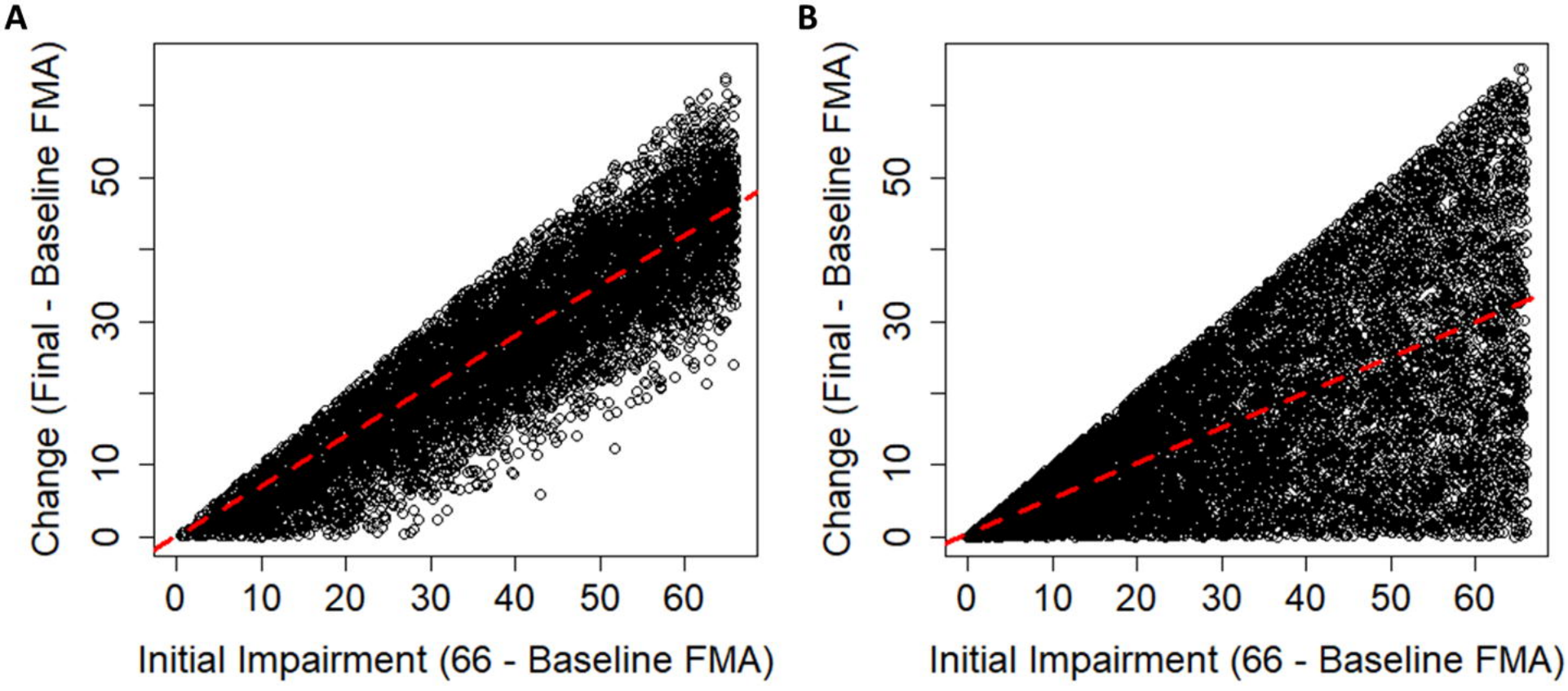
37. Senesh MR, Reinkensmeyer DJ. Breaking Proportional Recovery After Stroke. *Neurorehabilitation & Neural Repair*. 2019 Aug 16:1545968319868718.

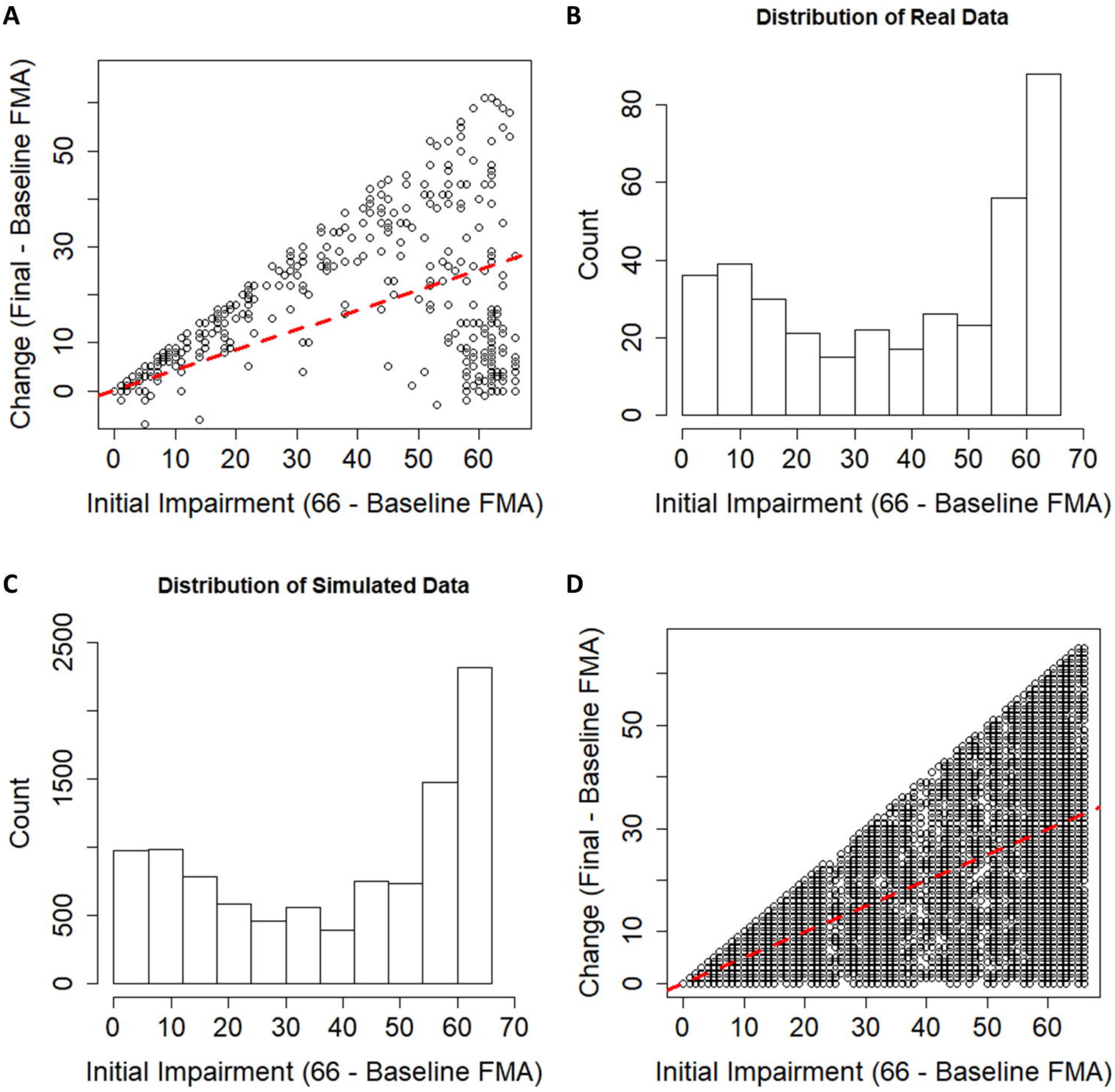
38. Blance A, Tu YK, Gilthorpe MS. A multilevel modelling solution to mathematical coupling. *Statistical Methods in Medical Research*. 2005;14(6):553-65.

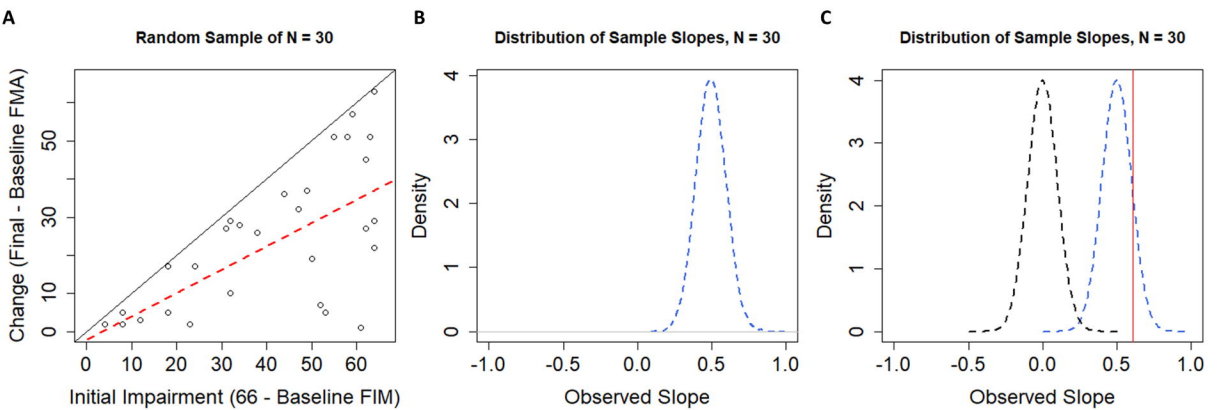
39. Garcia TP, Marder K. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current Neurology and Neuroscience Reports*. 2017;17(2):14.

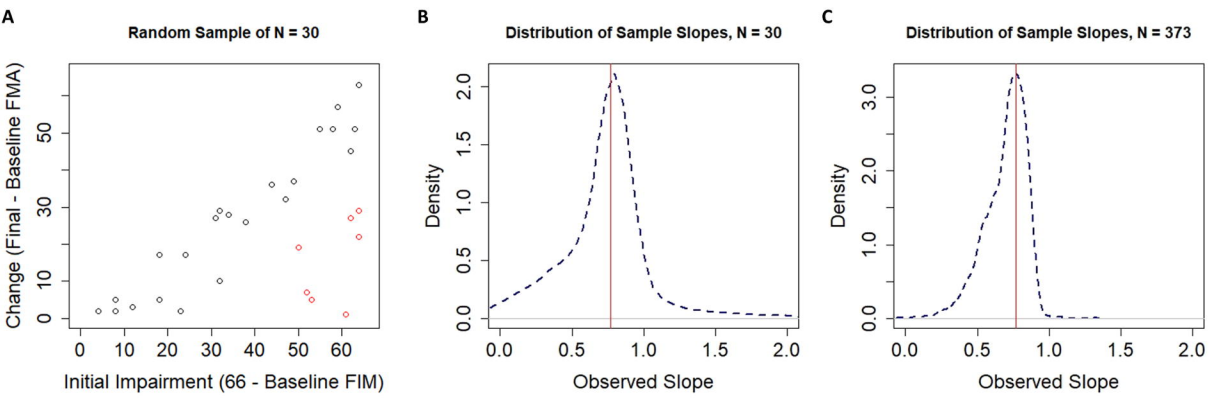

Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FMA)


A
r = 0.0
y
x
B
r = -0.7
z (y-x)
x
C
r = 0.0
y
x
D
r = -0.7
z (y-x)
x


A
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FMA)
B
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FMA)


A
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FMA)
B
Distribution of Real Data
Count
0
20
40
60
80
0
10
20
30
40
50
60
70
Initial Impairment (66 - Baseline FMA)
C
Distribution of Simulated Data
Count
0
500
1500
2500
0
10
20
30
40
50
60
70
Initial Impairment (66 - Baseline FMA)
D
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FMA)


A
Random Sample of N = 30
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FIM)
B
Distribution of Sample Slopes, N = 30
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
Observed Slope
C
Distribution of Sample Slopes, N = 30
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
Observed Slope


A
Random Sample of N = 30
Change (Final - Baseline FMA)
0
10
30
50
0
10
20
30
40
50
60
Initial Impairment (66 - Baseline FIM)
B
Distribution of Sample Slopes, N = 30
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
Observed Slope
C
Distribution of Sample Slopes, N = 373
Density
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
2.0
Observed Slope