OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1111/ejn.13847

Document Version
Early version, also known as pre-print

Link to publication record in King's Research Portal

Citation for published version (APA):
Jeffery, N. D., Bate, S. T., Safayi, S., Howard, M., Moon, L., & Jeffery, U. (2018). When neuroscience met clinical pathology: partitioning experimental variation to aid data interpretation in neuroscience. *European Journal of Neuroscience*. Advance online publication. https://doi.org/10.1111/ejn.13847

**Citing this paper**
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

**General rights**
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

**Take down policy**
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 18. Apr. 2025

*TECHNICAL SPOTLIGHT*

# When neuroscience met clinical pathology: partitioning experimental variation to aid data interpretation in neuroscience

Nick D Jeffery[1*], Simon T Bate[2], Sina Safayi[3], Matt Howard[4], Lawrence Moon[5], Unity Jeffery[6]

[1]Department of Small Animal Clinical Sciences, Texas A&M University, TX 77843, USA

[2]Statistical Sciences, GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK

[3]The University of Texas Graduate School of Biomedical Sciences at Houston, 6767 Bertner Ave, Houston, TX 77030, USA

[4]Department of Neurosurgery, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA

[5]Neurorestoration Group, Wolfson Centre for Age-Related Diseases, King's College London, University of London, London SE1 1UL, UK

[6]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX 77840

* **Corresponding author**; address as above; email: njeffery@cvm.tamu.edu

**Running title:** Analysis by reference change value

**TOTALS:**

| | | |
|---|---|---|
| Pages: | 38 | |
| Figures: | 3 | |
| Tables: | 3 | |
| Equations: | 4 | |
| Words: | a) whole manuscript: | 4739 (excluding references, abstract and figure legends) |
| | b) Abstract: | 228 |

***Keywords:*** translation, reference change value, intra-individual, partition, variability

**Abstract**

Neuroscientists typically assess the effectiveness of interventions by comparing the average response of groups of treated and untreated animals. These group comparisons provide useful insights into the effectiveness of interventions, but focusing only on group effects risks overemphasis of small, statistically significant but physiologically unimportant differences. Such differences could be created by analytical variability or by physiological within-individual variation, especially when the number of animals in each group is so small that one or two outlier values have considerable impact on the summary measures for the group.

Physicians face a similar dilemma when comparing two results from the same patient. To determine if the change between two values reflects disease progression or expected analytical and physiologic variation, the magnitude of the difference between two results is compared to the reference change value. These values are generated by quantifying analytical and within-individual variation, and differences between two results from the same patient are considered clinically meaningful only if they exceed the combined effect of these two sources of "noise".

In this article we describe how this technique can be extended into neuroscience. This form of analysis provides a measure of outcome at an individual level that complements traditional group-level statistics and therefore we hope that introduction of this technique into neuroscience will enrich interpretation of experimental data. In addition, it can also safeguard against some of the possible misinterpretations that may occur during analysis of the small experimental groups that are common in neuroscience and, by illuminating analytical error, may aid in design of more efficient experimental methods.

## Introduction

The repeated failure of neuroscience to generate novel clinical therapies has been a source of unease for many years (Garner, 2014) and this disquiet has peaked recently as deficiencies in experimental design and analysis have become more widely apparent (Begley, 2013; Button *et al.*, 2013; Steward, 2016; Academy of Medical Sciences, 2015). Many remedies for methodological shortcomings are available, focusing on both animal model relevance and experimental design (Garner, 2014, Begley, 2013; Button *et al.*, 2013; Steward, 2016; Academy of Medical Sciences, 2015; Landis, 2012; ARRIVE, 2017). Here we take another approach, by considering how data analysis strategies that are routinely used in hospital laboratories might complement traditional evaluation of neuroscience data.

In biomedical studies it is important to consider the magnitude of an effect because interventions with greater effect will likely have greater impact in the clinic. Routinely, the effect is determined by assessing an outcome measure before and after the intervention (within animal), or by comparing treated *versus* untreated groups of animals. Such analyses imply comparison, usually by statistical tests, of the population distribution and its central tendency (*i.e.* mean, median *etc*) between groups. Unfortunately, while widely used, the derived P values provide only an estimate of the consistency of the data with the null hypothesis; they do not directly provide information on the magnitude of an intervention effect and are difficult to interpret without knowledge of pre-study power calculations (Shaver, 1993; Halsey *et al.*, 2015).

In part, these problems arise from the considerable variability in data distribution, and group-level analyses do not provide the complete picture. For instance, as highlighted previously in this

journal, specific individuals may show changes in outcome that are not reflected in the summary mean or median values (Rousselet et al., 2016). In addition, a commonly overlooked problem is that random sources of variation inherent in animal experiments, such as physiological variability and intrinsic measurement inaccuracies, may combine to produce outcome effects that might erroneously be attributed to the intervention itself, especially in the small sample populations typically used in neuroscience (Button *et al.*, 2013). Currently established statistical analyses in neuroscience do not always efficiently dissect the difference between inherent random sources of variation and the effects induced by an experimental intervention. Individual-level analysis, as described below, helps to guard against drawing inappropriate conclusions from underpowered studies.

A further aim in translational biomedical research is to identify interventions that will have real-life impact for each treated individual. Whilst it is of course useful and important to show, as conventional group-level analysis can, that, on average, a treated individual will have a better outcome than a non-treated individual it is also important to examine effects at an individual level, because each individual patient wants their life to be meaningfully enhanced. The two approaches are complementary: conventional group-level analysis can provide the information as to whether, on average, one treatment is better than another and provide an estimate of how much better. Individual-level analysis provides an estimate of: a) how many individuals attain a benefit that is greater than that which can expected through variation attributable to experimental inaccuracy and physiological variation; and also, b) the extent that the improvement for each individual is beyond those limits. In this *Technical Spotlight* we explain how an individual-level

analysis can be derived and how it aids in dissecting intervention signal from experimental noise, thereby providing a complement to standard group-level analyses.

## Analysis of individual responses in clinical practice

In clinical medicine the focus is on individual patients and there is frequently a need to determine whether a patient's condition is deteriorating or whether a treatment is having a beneficial effect. Such monitoring involves obtaining serial samples, which then raises the question as to how large the difference between a pair of samples obtained from the same individual has to be before it represents a meaningful change. In clinical pathology laboratories this problem has been addressed by partitioning the various sources of variation and using this to derive boundaries that encompass all sources of extraneous variation. This can then be used to assess the results for each patient individually (Harris & Yasaka, 1983; Fraser & Harris, 1989; Walton, 2012). By comparing the changes in an individual's laboratory test results to these boundaries 'real change' can be inferred. Below we outline how the same concepts can be applied to experimental studies in neuroscience. The combination of group- and individual-level analysis together provides a more comprehensive examination of the data and may therefore aid in establishing which laboratory interventions are most likely to have sufficiently robust effects to translate into effective clinical therapy. This may also aid in design of appropriate pre-clinical functional tests.

## What is variation?

Variation refers to random fluctuations in repeated results generated for a particular sample or individual. This is distinct from systematic error, also termed bias, which describes consistent under- or over-estimation of the true value (Theodorsson *et al.*, 2014). The standard deviation

(SD), or the *coefficient of variation* (CV) (= SD/mean) provide summary measures of variation of the standard mean.

**Sources of variation**

In clinical pathology, when determining the value of a specific analyte the various possible sources of variation are identified and partitioned. There are similar sources of variation in neuroscience.

***Investigator-derived pre-analytical variation*** encompasses all the investigator-dependent sources of variation that influence the level of an analyte before measurements are made. For example, variation in how long, or in what conditions, a sample has been stored before testing may influence the final results of the experiment. In general, pre-analytical variation is not of interest to the experimenter, instead forming a source of 'noise' in data interpretation, and so should be eliminated as far as possible. In clinical pathology, investigator-dependent pre-analytical variation is minimized by strict handling protocols for sample collection, handling and storage.

***Analytical variation*** arises because every analytical technique has intrinsic sources of laboratory variability that can lead to variation in replicate results obtained from a single sample (this is also sometimes referred to as measurement, technical or instrumental error). For instance, small volume errors in loading a sample into an automated analyzer or simple variation in reaction kinetics can produce different results. While analytical variation can be reduced, for example by instrument calibration or cleaning, it cannot be completely eliminated. Nevertheless, by

158 performing multiple replicate analyses of each sample, or more generally by repeatedly
159 measuring the same physical material, it can be quantified and this estimate used when deducing
160 the sources of variation in the final results (see below).
161
162 ***Biological variation*** can be divided into intra-individual and inter-individual variation. Both
163 forms of biological variation include predictable sources of variation (*e.g.* variation in blood
164 hormone concentration over the reproductive cycle of an individual or variation in blood
165 hormone concentration between individual mice at corresponding points in the reproductive
166 cycle). Researchers typically already pay close attention to these known sources of variation and
167 standardize where possible, but unknown sources of biological variation also have important
168 implications for data interpretation.
169
170 *Intra-individual variation* arises because, as would be expected, there is some variation of
171 analyte levels within each individual between samples obtained at different times that is above
172 and beyond analytical variability. For instance, blood cholesterol concentration varies (even at
173 the same time each day) between one day and another, even in healthy individuals (Rotterdam *et*
174 *al.*, 1987). The magnitude of this variation varies between analytes: there is much less intra-
175 individual variation for some (*e.g.* chloride) than others (*e.g.* triglycerides: Nunes *et al.*, 1993).
176 Importantly, for some analytes the intra-individual variation may be very much less than the
177 inter-individual variation, implying that what might appear to be only a subtle change when
178 viewed in the context of the population as a whole may be critical to the health of a specific
179 individual. For such analytes it is critical to define limits of expected physiological change for
180 each individual and not rely on group-level assessment.

*Inter-individual variation* arises from the familiar biological observation that, even within a defined population, individuals vary. Indeed, this is the rationale for carrying out experiments using groups of animals, summarizing the outcomes across all members of the group(s) and carrying out the statistical assessment at the population level. While this approach remains valid, in view of the other sources of variation outlined above, apparent intervention effects may be obscured - or magnified - by intra- or inter-individual variation, as well as by analytical and pre-analytical variation. The risk of drawing false conclusions is greatest when the number of experimental subjects within the groups is small, as is common in neuroscience. In addition, because these random sources of variation may not be evenly distributed between groups, they may generate unreliable P-values that may give a false impression of the statistical significance of the overall difference between groups, or between repeated measures within groups (Button *et al.*, 2013) (see Figure 1, Table 1). The addition of an analysis of individual animal outcomes, as explained below, can help avoid this problem.

**How variation is dealt with in clinical pathology laboratories**

In clinical pathology laboratories it is common for an individual patient to undergo repeated testing over time. The most useful approach for determining whether there has been a meaningful change between measurements made at two different time points in a single individual is to calculate a *reference change interval* (RCI), derived from calculation of the *reference change value* (RCV) (Fraser, 2001):

$$RCI = Baseline +/- (RCV * Baseline) \quad (1)$$

where $RCV = z_p * \sqrt{2} * \sqrt{(CV_I^2 + CV_A^2)}$ *(2)*

and $CV_I$ is the intra-individual coefficient of variation and $CV_A$ is the analytical coefficient of variation. The z-score in (2), $z_p$, can be used to define how confident the experimenter wishes to be that the new result is really different from the previous value. Conventionally, a 5% false positive rate is selected, corresponding to a z-score of 1.96.

The reference change interval defines the boundaries (usually as a percentage change from a previous result) within which measurements of a single analyte might be expected to vary within a normal individual. By comparing pre- and post-intervention data, the reference change value can also put the effect of an intervention into the context of normal biological variation for an individual, thereby allowing more individualized assessment of the magnitude of the effect.

Here we show how the same approach can be applied to measures of neurologic function in laboratory animals.

## Translating clinical pathology analytical methods into neuroscience research

### *Pre-analytical variation relating to study design*

First, pre-analytical variation should be eliminated as far as possible because its reduction will increase precision. For factors other than homeostatic variation this can largely be achieved by following consistent protocols that should be fully reported and documented to aid transparency (ARRIVE guidelines, Kilkenny *et al.*, 2012). This is most straightforward when applied to familiar laboratory procedures such as immunohistochemistry, for which it is essential to use the

same tissue handling techniques and the same batches of antibodies and other reagents between compared samples. The same concept can be applied to behavioral experiments. For instance, as noted elsewhere (National Center for 3Rs, 2017), behavioral tests should be carried out at the same time every day in the same environment with the same schedules and same experimenter. Although attempts should be made to completely eliminate all possible sources of pre-analytical variation (ARRIVE guidelines), it may be impossible (*e.g*. because of extreme weather events, fire alarms *etc*.), in which case undefined sources of pre-analytical variation will usually then be incorporated into other sources of variation (see below).

***Derivation of analytical variation***

Analytical, or technical, variation is usually determined in clinical pathology by repeatedly measuring analyte levels in one sample by splitting it into multiple aliquots. This approach can easily be applied to routine laboratory techniques like ELISA or PCR, but is rarely appropriate in behavioral neuroscience because repeated examination of one individual may change measured outcomes through training. Although not ideal, it may be possible to obtain a surrogate measure of inherent variation by viewing 'aliquots' of a single segment of video footage of various individuals or repeatedly deriving values for kinematic variables from a single original source dataset (and this process is illustrated in Example 1 below).

Quantification of analytical variation can be an extremely valuable exercise in itself, because it will often provide insights into how the technique and precision of measurements can be improved. However, not all tests in behavioral neuroscience are amenable to similar methods of estimating analytical variation as described above, implying that this source of variation cannot

always be partitioned from that derived from intra-individual variation. Despite this apparent obstacle the reference change value can still be calculated (see below) and, importantly, even when the analytical component cannot be explicitly estimated, the reference change interval can still be narrowed (and thus its sensitivity to intervention-induced change increased) by taking steps to minimize analytical error. For example, eliminating any potential handling or environmental conditions that trigger 'freezing' of movement in mice (Gouveia & Hurst, 2013) reduces analytical error (and therefore the width of the reference change interval) for motor task outcomes.

***Derivation of intra-individual and inter-individual variation***

Many experiments involve repeatedly measuring outcomes in a group, or groups, of animals. The inter- and intra-individual variation can be estimated using a nested random effects ANOVA (Fraser and Harris, 1989) or repeated measures mixed models (Bate and Clark, 2014; Marechenko, 2006), which provide flexibility when modelling the correlation between repeated measurements taken on an individual. Typically, estimates of intra- and inter- animal variation are derived from pre-intervention data because it is important that variability is estimated in individuals at a steady plateau of disease or function. Fortunately, experience in clinical pathology has indicated that, for most outcomes, patients with steady-state diseases or lesions have similar variability to normal individuals (Carmen *et al.*, 2007).

However, in experimental studies an alternative approach might be feasible: instead of using control animals to define the intra-animal variability for all animals in an experiment these measures could instead be derived for each individual. Thus, before making a lesion or

implementing a therapeutic intervention it would be possible, through repeated testing, to accumulate sufficient data to calculate the intra-animal variability for each individual animal, thus providing an individualized measure of the intervention effect (Safayi *et al.*, 2015). For instance, a z-score for each outcome can be derived for each animal by re-arranging equation (2) so that:

$$z\text{-}score = \ proportional\ change\ from\ baseline\ / \sqrt{2}\ * \sqrt{(CV_I^2 + CV_A^2)} \qquad (3)$$

This can then be used to define the magnitude of intervention effect in each individual which is linked to its own biological variation. An intervention that has an effect considerably greater than biological variation is more likely to produce a meaningful improvement in an individual patient's quality of life than an intervention that has effects of similar magnitude to normal day-to-day fluctuations in performance.

***Defining the boundaries of outcome that might result from experimental and biological variation alone***

For each individual, if a new test result is to represent a 'real' change compared with a previous test result it must lie outside the range that could be expected to arise through the combination of pre-analytical, analytical and intra-individual variation alone. To determine if this is the case, the previously derived values for $CV_A$ and $CV_I$ are combined to derive the reference change value, which can then be applied, as given in equation (2), to define reference change interval boundaries.

In many neuroscience experiments it may not be possible to derive (and therefore partition) the component of variability that contributes to $CV_A$, in which case the analytical variation will become a component of $CV_I$ and equation (1) will reduce to:

$$RCV = z_p * \sqrt{2} * CV_I \qquad (4)$$

As before, the boundary values beyond which real change can be deduced to have occurred in an individual (the RCI) is calculated using equation (2).

## Examples

### *Example 1:*

*Derivation and use of measures of analytical ($CV_A$), intra-individual ($CV_I$) and inter-individual ($CV_G$) variability*

Here we use real and simulated data from an experiment to investigate the effect of a drug on the gait of sheep as they walked on a treadmill at constant velocity; the outcome of interest was hindlimb stride duration. Using a nested design analysis module (InVivoStat; http://invivostat.co.uk/) the intra- and inter-animal coefficients of variation, denoted by $CV_I$ and $CV_G$ respectively, were estimated from serial measurements made on trained normal animals before drug administration (the raw data is available as Supplementary Material).

Sheep were trained to walk on a treadmill at a steady speed until they attained a plateau of proficiency. Stride duration was then measured repeatedly on each of 4 days and on each of these days a long recording segment (~ 4 minutes) was divided into four parts, with each part considered as a separate 'trial'. This is a nested design in which trials are nested within days and days are nested within sheep; the sources of variability were apportioned into analytical

318 ('between-trial, within-day'), intra-individual ('between-day, within-sheep') and inter-individual
319 (between-sheep) to derive means and standard deviations which were used to derive the
320 respective coefficients of variation.
321
322 Inter-individual, intra-individual and analytical standard deviations were 0.039, 0.033 and 0.045
323 respectively and the overall mean stride duration was 0.862, resulting in the following
324 coefficients of variation:
325 $CV_G = 0.039 / 0.862 = 4.5\%$;
326 $CV_I = 0.033 / 0.862 = 3.8\%$;
327 $CV_A = 0.045 / 0.862 = 5.2\%$
328
329 The **reference change value** was 17.8%, calculated using equation (1) above, therefore, any
330 sheep in which the post-drug measurement was changed by more than 17.8% of the pre-drug
331 value (*i.e.* outside the upper or lower boundary of the reference change interval) can be
332 considered to have undergone 'real' change as a consequence of drug administration. It is of
333 course possible that the intervention may have had an effect that produces an outcome that does
334 not lie outside the reference change interval but such an effect can be inferred to be negligible in
335 terms of biomedical meaning (and, in clinical medicine, a drug having an effect of such small
336 magnitude would not be considered valuable).
337
338 To illustrate the different, and complementary, perspectives provided by population-level and
339 RCV within-individual analysis, see Figure 2, which depicts the difference in stride duration in
340 these same sheep before and after drug administration. A Wilcoxon matched pairs signed rank

test (post-drug population is not normally distributed) indicates a significant increase in stride duration following this intervention (p=0.023).

When the complementary individual analysis is applied to the same dataset (Table 2), in only 3 of 8 sheep does the post-drug value exceed the upper boundary of the reference change interval and indicate a real difference from pre-drug values. Furthermore, the largest drug-induced change was a 27% change from baseline (Sheep 8). When applied in equation (3) this change corresponds to a z-score of 2.97, indicating a value that is in the top 0.3 percentile of the expected distribution of baseline values (see https://measuringu.com/pcalcz/).

Therefore, although there is some evidence of an overall drug effect at the population level, this drug at this dose evoked a real (*i.e.* biomedically meaningful) change in less than 50% of subjects and the largest effect in any single individual was moderate (*i.e.* within the top 0.3 percentile of the range of values that could be attributed to biological and analytical variation). Despite the statistically significant result, there is evidence that the drug effect would require considerable augmentation to achieve a worthwhile benefit that would be discernible in a large proportion of subjects. Reporting both analyses alongside each other provides a more comprehensive picture of drug effect. Although tempting, we are not recommending comparison of the proportion of animals that respond to an intervention, because such analysis will inevitably have very low power (Snapinn & Jiang, 2010), and does not take account of the magnitude of effect beyond the reference change interval for each animal.

***Example 2:***

This example concerns bladder function in dogs in a clinical trial of a putative therapy for chronic spinal cord injury. A parallel group study was performed with dogs randomized at time 0 to receive either percutaneous intraspinal chondroitinase injection or a sham injection in which the skin was pierced with a hypodermic needle. A bladder compliance index (change in pressure with increasing volume of urine) was measured by cystometry at baseline (before intervention) and at 1, 3 and 6 months after intervention (or sham). In this example it is not possible to segregate analytical variation from the inter- and intra-animal variation (because it is not possible to repeatedly measure compliance on 1 day for each animal because this in itself would be expected to cause changes) and so this becomes incorporated into other sources of variability.

a) First, the data from the ***Control*** animals over the entire observation period was analyzed using a nested design analysis module (InVivoStat; http://invivostat.co.uk/) to segregate intra- and inter-individual variability (the raw data is available as Supplementary Material). With the mean compliance value this was used to derive the intra- and inter- animal coefficients of variation ($CV_I$ and $CV_G$ respectively), where in this example

$$\mathbf{CV_I} = 0.865 / 1.508$$

$$= 0.574$$

b) The ***reference change value*** was then calculated from the $CV_I$ using equation (4)

$$RCV = 1.96 * \sqrt{2} * CV_I$$

= 2.764 * 0.574

= 1.587

This implies that, for each dog, for a post-intervention value to be considered beyond the threshold that could be attributed to variability caused by physiological and experimental measurement variability, it would have to be more than 158.7% different from baseline.

c) We then applied this reference change value to the observed changes in dogs in the ***Intervention*** group (Table 3). There are two main findings from this analysis:

i) Two dogs (of 21) showed a change in compliance index that was outside the reference change interval, and all of these were increases. Furthermore, their z-scores were high, lying within the highest 0.5 percentile of the expected distribution, suggesting that it was extremely probable that real change in compliance had occurred during this time period in these dogs.

ii) No dog showed a reduction in compliance index. However, the analysis suggests that it is probable that this outcome measure is insensitive to that direction of change. The reference change value is large for this test (~159%) which indicates that real change can only be deduced with a large change from baseline. Many of the dogs have low values at baseline meaning that attaining values outside the lower boundary of the reference interval for those individuals is biologically extremely implausible. Therefore this analysis also reveals a floor effect (limitation) in this outcome measure.

d) This individualized analysis can be compared with conventional group-level analysis: a paired Student's t-test (datasets for baseline and month 1 [above] are both normally distributed) produces a *P* value of 0.139, conventionally regarded as not significantly different. The data from individual dogs is shown in Figure 3.

The interpretation of these data is therefore that there is no overall evidence of effect of the intervention at a group level but there is evidence that some animals exhibit changes in bladder function that cannot reasonably be attributed to experimental noise (*i.e.* physiological variation and measurement inaccuracies) alone. This might suggest that there are specific individuals that respond to the intervention, which might prompt further investigation of the specific characteristics of these individuals, for instance detailed investigation of the character of their spinal cord injuries.

However, this individual-level analysis also reveals that there is considerable variability in this outcome measure, which might be a consequence of specific physiological attributes (intra- and inter- animal variability) or of difficulties in making repeatable measurements (analytical variation), each of which will limit the ability to detect real change. Furthermore, there is evidence of a floor effect, which limits the ability to detect reduction from baseline in all animals. Both these features might suggest the need to stratify animals for entry into a study on this outcome to increase the sensitivity of the experiment to physiologically important change.

## Discussion: How can RCV-based analysis aid in neuroscience research?

### *In translational research*

Confidence in neuroscience data has been eroded recently, mainly because of the problems that have become apparent in translating apparently positive laboratory results into the clinic and because of widespread difficulties in reproducibility (Garner, 2014). In order to make a successful transition from laboratory to clinic an intervention needs to have a large enough effect to change an individual patient's life in some meaningful way but must also be effective in a suitably large proportion of treated individuals. The analysis strategy that we describe here can aid in overcoming this confidence gap because it provides a more detailed description of the magnitude of the intervention effect in each individual alongside defining the proportion of animals within a treated cohort that show a response beyond a pre-defined level.

### *In laboratory science*

The partitioning of sources of variation can be helpful in redesigning pre-clinical outcome measures and defining good laboratory practice to minimize unnecessary variation. For instance, awareness of the concept of pre-analytical variation can aid in eliminating possible causes from laboratory testing procedures, such as those associated with experimenter-centered effects, such as the time of day at which tests are performed. A recent example of the importance of these effects is the recognition that the gender of the researcher can alter mouse behavior (Sorge *et al.*, 2014). The large $CV_I$ we describe above for bladder compliance in spinal cord-injured dogs may partly be a result of an unidentified source of pre-analytical and, or, analytical variation, such as urinary tract infection or poor pre-trial management of bladder emptying.

Analysis of the various sources of variation can also help in determining whether a functional test is sufficiently sensitive to detect a change in the outcome that is being measured. As we show, the $CV_I$ for bladder compliance in spinal cord-injured dogs is large, meaning that, at a group-level, this outcome measure is relatively insensitive to intervention-evoked change. In clinical pathology the aim is for tests on patients to have an analytical component of variation that is too small to have a major impact on clinical decision-making. Although the very small analytical errors required in clinical pathology may well be unattainable for functional tests in neuroscience, tests should be redesigned with the aim to improve precision where possible. Another problem with functional testing that may be revealed is the possibility of floor or ceiling boundaries on the responses that may limit the sensitivity of the assay for the outcome of interest. This was also apparent in the bladder compliance data (example 2 above).

***Limitations***

Unfortunately, not all current outcome measures used in laboratory animals will be appropriate for the analysis strategy outlined in this article. For example, behavioral analytical methods that attribute numeric scores but are not truly numeric because the scores do not represent equally-spaced intervals on a linear scale. This implies that calculation of standard deviation (and therefore the CV) and, consequently, the RCV, can be problematic. Similarly, for some outcome measures the data for computation of the RCV may not be normally distributed. Both of these obstacles can be circumvented. First, log-transformation of positively-skewed data allows application of the techniques we describe above, followed by anti-logging to convert the mathematical answer into the appropriate clinical units. Second, measurement of test-retest

differences in a large population (>120) of individuals at a functional plateau can be used to derive the 95% reference change interval of data of any distribution through non-parametric analysis (Friedrichs et al., 2012). In practice, this method would be difficult to apply because of the need for static functional status and large animal numbers but could be sufficiently valuable to be worthwhile for some commonly applied testing methods.

## Conclusion

Analysis of sources of variation and application of reference change intervals in individual animals do not replace a conventional assessment of group-level outcomes, but can provide an additional layer of analysis that complements and extends such findings. Therefore there would appear to be considerable merit in reporting both types of analysis alongside each other. Nevertheless, the clinical effectiveness of an intervention can be usefully estimated by determining the proportion of individual outcomes that fall, or are expected to fall, outside specific z-scores and, as we show above, the same analytical technique can aid in redesign of functional tests to maximize their precision. Although we have focused on application of this analytical technique to behavioral data, because these provide the most transparent examples, the same methods can be applied to many other types of quantitative or semi-quantitative neuroscience data including analysis of cell activity or production of specific molecules.

## Acknowledgements

The International Spinal Research Trust (grant ref STR116) and University of Iowa have provided funds for work with dog and sheep models of spinal cord injury, respectively, to NDJ. LM receives funding from the European Research Council under the European Union's Seventh

Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 309731. We thank Dr Hilary Hu and Josh Bratsch-Prince for collecting raw data.

## Animal care and use

All animal studies were reviewed by the Institutional Animal Care and Use Committee and by the ethical review process at the institution where the work was performed.

## Conflict of interest

The authors declare no conflict of interest.

## Author contributions

NJ, UJ, SB designed article approach and analyzed data

SS, MH, LM collected and analyzed data

NJ, UJ, SB wrote the paper

## Data Accessibility

The primary data is available as Supplementary Material.

## References


ARRIVE Guidelines: http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-reporting-the-arrive-guidelines-for-reporting-animal-research/

Bate, S. T., & Clark, R. A. (2014). *The design and statistical analysis of animal experiments*. Section 6.14: Nested design analysis module. Cambridge University Press. p285.

Begley, C.G. (2013) Six red flags for suspect work. *Nature* **497**, 433-4.

Biering-Sørensen, F., Craggs, M., Kennelly, M., Schick, E. & Wyndaele, J.J. (2008) International urodynamic basic spinal cord injury data set. *Spinal Cord* **46**, 513-516.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S. & Munafò, M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365-76.

Carmen, R., Iglesias, N. & Garcia-Lario, J. V., *et al*. (2007) Within-subject biological variation in disease: collated data and clinical consequences. *Ann. Clin. Biochem.* **44**, 343-352.

Center for 3Rs: http://3rs.ccac.ca/en/research/reduction/experimental-design.html

549 Flatland, B., Freeman, K.P., Friedrichs, K.R., Vap, L.M., Getzy, K.M., Evans,
550 E.W. & Harr, K.E. (2010) ASVCP quality assurance guidelines: control of
551 general analytical factors in veterinary laboratories. *Vet. Clin. Path.* **39**, 264–277.
552
553 Fraser, C.G. & Harris, E.K. (1989) Generation and application of data on
554 biological variation in clinical chemistry. *Crit. Rev. Clin. Lab. Sci.* **27**, 409-437.
555
556 Fraser, C.G. (2001) Chapter 3. Changes in serial results. In Fraser, C.G. (ed) *Biological*
557 *variation: from principles to practice*. AACC Press, Washington DC, pp. 67-90.
558
559 Friedrichs, K.R., Harr, K.E., Freeman, K.P., Szladovits, B., Walton, R.M., Barnhart, K.F. &
560 Blanco-Chavez, J. American Society for Veterinary Clinical Pathology. (2012) ASVCP
561 reference interval guidelines: determination of de novo reference intervals in veterinary species
562 and other related topics. *Vet. Clin. Pathol.* **41**, 441-53.
563
564 Clinical and Laboratory Standards Institute. *Defining, Establishing, and Verifying Reference*
565 *Intervals in the Clinical Laboratory; Approved Guidelines*. 3rd ed. Wayne, PA: CLSI; 2008.
566
567 Garner, J.P. (2014) The significance of meaning: why do over 90% of behavioral neuroscience
568 results fail to translate to humans, and what can we do to fix it? *I.L.A.R. J.* **55**, 438-56
569

570 Gouveia, K. & Hurst, J.L. (2013) Reducing mouse anxiety during handling:
571 effect of experience with handling tunnels. *PLoS One* **20**, e66401.
572
573 Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. (2015) The fickle P value
574 generates irreproducible results. *Nat. Methods* **12**, 179-85.
575
576 Harris, E.K. & Yasaka, T. (1983) On the calculation of a" reference change" for
577 comparing two consecutive measurements. *Clin. Chem.* 29, 25–30.
578
579 Kilkenny C, Browne WJ, Cuthi I, Emerson M, Altman DG. (2012) Improving bioscience
580 research reporting: the ARRIVE guidelines for reporting animal research. Vet. Clin. Pathol. 41:
581 27-31.
582
583 Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal,
584 R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman,
585 H.E., Golub, R.M., Goudreau, J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells,
586 D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., Levine, M.S., Macleod,
587 M.R., McCall, J.M., Moxley, R.T. 3rd., Narasimhan, K., Noble, L.J., Perrin, S., Porter,
588 J.D., Steward, O., Unger, E., Utz, U. & Silberberg, S.D. (2012) A call for transparent reporting to
589 optimize the predictive value of preclinical research. *Nature* **490**, 187-91.
590

Marchenko, Y. (2006) Estimating variance components in Stata. *The Stata Journal* 6, 1-21.

Nunes, L.A., Brenzikofer, R. & de Macedo, D.V. (2010) Reference change values of blood analytes from physically active subjects. *Eur. J. Appl. Physiol.* **110**, 191-8.

Rotterdam, E.P., Katan, M.B., & Knuiman, J.T. (1987) Importance of time interval between repeated measurements of total or high-density lipoprotein cholesterol when estimating an individual's baseline concentrations. *Clin. Chem.* **33**, 1913-1915.

Rousselet, G.A., Foxe, J.J. & Bolam, J.P. (2016) A few simple steps to improve description of group results in neuroscience. *Eur.J. Neurosci*. **44**, 2647-2651.

Safayi, S., Jeffery, N.D., Shivapour, S.K., Zamanighomi, M., Zylstra, T.J., Bratsch-Prince, J., Wilson, S., Reddy, C.G., Fredericks, D.C., Gillies, G.T., Howard, M.A. 3rd. (2015) Kinematic analysis of the gait of adult sheep during treadmill locomotion: Parameter values, allowable total error, and potential for use in evaluating spinal cord injury. J. Neurol. Sci. **358**, 107-12.

Shaver, J.P. (2013) What Statistical Significance Testing Is, and What It Is Not. *J. Exp. Educ.* **61**, 293-316.

613 Snapinn, S.M. & Jiang, Q. (2007) Responder analyses and the assessment of a clinically relevant
614 treatment effect. Trials **8**, 31.
615
616 Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf,
617 J.S., Acland, E.L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J.C., McPhail, M., Delaney,
618 A., Wigerblad, G., Schumann, A.P., Quinn, T., Frasnelli, J., Svensson, C.I., Sternberg, W.F.
619 & Mogil, J.S. (2014) Olfactory exposure to males, including men, causes stress and related
620 analgesia in rodents. *Nat. Methods* **11**, 629-632.
621
622 Steward, O. (2016) A Rhumba of "R's": Replication, Reproducibility, Rigor, Robustness: What
623 Does a Failure to Replicate Mean? *eNeuro* **7**, 3.
624
625 The Academy of Medical Sciences Reproducibility and reliability of biomedical research:
626 improving research practice: Symposium report, October 2015.
627 https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf
628
629 Theodorsson, E., Magnusson, B. & Leito, I. (2014) Bias in clinical chemistry
630 *Bioanalysis* **6**, 2855-2875.
631
632 Walton, R.M. (2012) Subject-based reference values: biological variation,
633 individuality, and reference change values. *Vet. Clin. Pathol.* **41**, 175-181.

## Figure legends

**Figure 1:** Illustration of the relationship between 'experimental noise' and group-level change. In this theoretical example, we assume that previous analysis has revealed that the combination of experimental inaccuracy and physiological variability can be responsible for changes of up to 20% between sequential test results (methods to derive these values are explained in the text). Collectively these sources of variability are termed the reference change interval, which defines limits of experimental noise that can be expected to arise through chance; methods to derive this interval are shown in the text.
**a:** The pre-intervention test result is illustrated for each of eight experimental subjects, along with bars depicting the range of values encompassed by the reference change interval of 20% from baseline. **b:** Comparison of pre- and post-intervention results and group-level statistical comparison by paired Student's t-test, indicating a statistically significant difference. **c:** Post-intervention results for each subject. **d:** Post-intervention test results for each subject, shown alongside the pre-test result and the expected analytical variability for each test result ('RCI', bars). Note that the post-intervention result for each subject is within the reference change interval for that subject.

**Figure 2:** Hindlimb stride duration in a group of eight trained sheep walking on a treadmill before and after administration of drug X. Wilcoxon matched pairs signed rank test suggests a significant effect of drug X on stride duration ($P$=0.023).

**Figure 3:** Baseline and 1-month post-intervention measures of bladder compliance index for dogs that had received intraspinal injection of chondroitinase ABC at time 0. Group-level analysis by paired Student's t-test reveals a non-significant difference between time points.

**Supplementary Material**

The data used in Examples 1 and 2 are supplied as Excel files ('sheep treadmill data' and 'dog compliance data').

The nested design analysis module in the free-to-download statistical software package InVivoStat (http://invivostat.co.uk/) can be used to derive the analytical, intra- and inter-animal standard deviations for both sets of data.

Guidance for use of this module is available at http://invivostat.co.uk/wp-content/user-guides/Nested-Design-Analysis.pdf and also in Section 6.14 of the book '*The design and statistical analysis of animal experiments*' by Simon T. Bate and Robin A. Clark (Cambridge University Press).

**Source of data**

***Example 1 – sheep treadmill data***

Sheep were trained to walk on a treadmill moving at constant speed until they were able to maintain a constant walking gait without stopping or trotting (treadmill speed was ~4km/h). To reach a plateau of steady performance took daily training sessions for a period of about 10 days (Safayi *et al.*, 2015).

Retroflective markers were attached to the hooves and the skin overlying boney prominences on the hindlimbs and used with a motion capture system (Vicon) to capture at least 4 minutes of continuous walking gait on four individual days, each separated by a week. Each segment of motion capture data was analyzed using a custom-written Matlab (MathWorks Inc) code to extract the duration of each hindlimb step (determined by the timing to minimum hoof height).

Each 4 minute segment of motion capture data was divided into four individual segments of 1 minute each and these we termed 'trials'. Analysis of each segment was used to correspond to analysis of aliquots of a fluid sample in a clinical pathology laboratory.

The 'post-drug' data was simulated (to show specific aspects of the relationship between group-level and individual-level analyses). It was derived by multiplying the 'pre-drug' data by a restricted range of random numbers (90-130%) that would, on average, increase the 'post-drug' values by a small proportion but allow for individual variation.

Reference change value calculations were made on the pre-drug dataset and partitioned into analytical, intra-animal and inter-animal variability as described in the main text. The calculated reference change interval was then compared with 'post-drug' values. The Supplementary Dataset 'Sheep treadmill data' contains the pre-drug values. The 'post-drug' values are available in Table 2 in the main article.

***Example 2 – dog bladder compliance data***

Dogs in this study were domestic pets that had incurred severe acute spinal cord injury at least 12 weeks previously and had not regained voluntary motor function or continence. They were entered into a blinded randomized controlled trial on the efficacy of intraspinal injection of chondroitinase ABC in reversing functional deficits associated with chronic spinal cord injury (Hu *et al.*, in press). The bladder compliance was measured in all dogs at baseline (entry into the study) and then again at 1, 3 and 6 months after intervention (or sham). Dogs were randomly allocated to active intervention (chondroitinase) or sham at time 0.

Briefly, bladder compliance was measured by passing a dual lumen catheter into the bladder, draining the urine and then infusing sterile saline into the bladder at a constant rate. Infusion was

terminated when saline leaked from the urethra, or when an intravesicular pressure of greater than 40cm$H_2O$ was attained. Compliance is measured as the change in volume (from baseline) divided by the change in pressure (from baseline). The data distribution was not normally distributed and so it was log-transformed for analysis in this article; therefore it is referred to as a 'compliance index'.

To calculate the reference change interval we used the dogs that were subsequently identified as controls (in which the allocation was blinded for observers by making needle sticks in the dorsal skin). We had repeated data from each time point that allowed calculation of intra-animal and inter-animal variation in this outcome. Analytical variability could not be specifically partitioned because repeat testing of compliance would likely lead to changes in the outcome measure.

As before, the reference change interval calculated in the control dogs was then applied to the 1-month data from the active intervention dogs to determine how much 'real change' had occurred. The control dog data used to calculate the reference change value is contained in the Supplementary Dataset 'Dog compliance data' and the data from dogs in the active intervention group are shown in Table 3 in the main article.

**Figure 1**

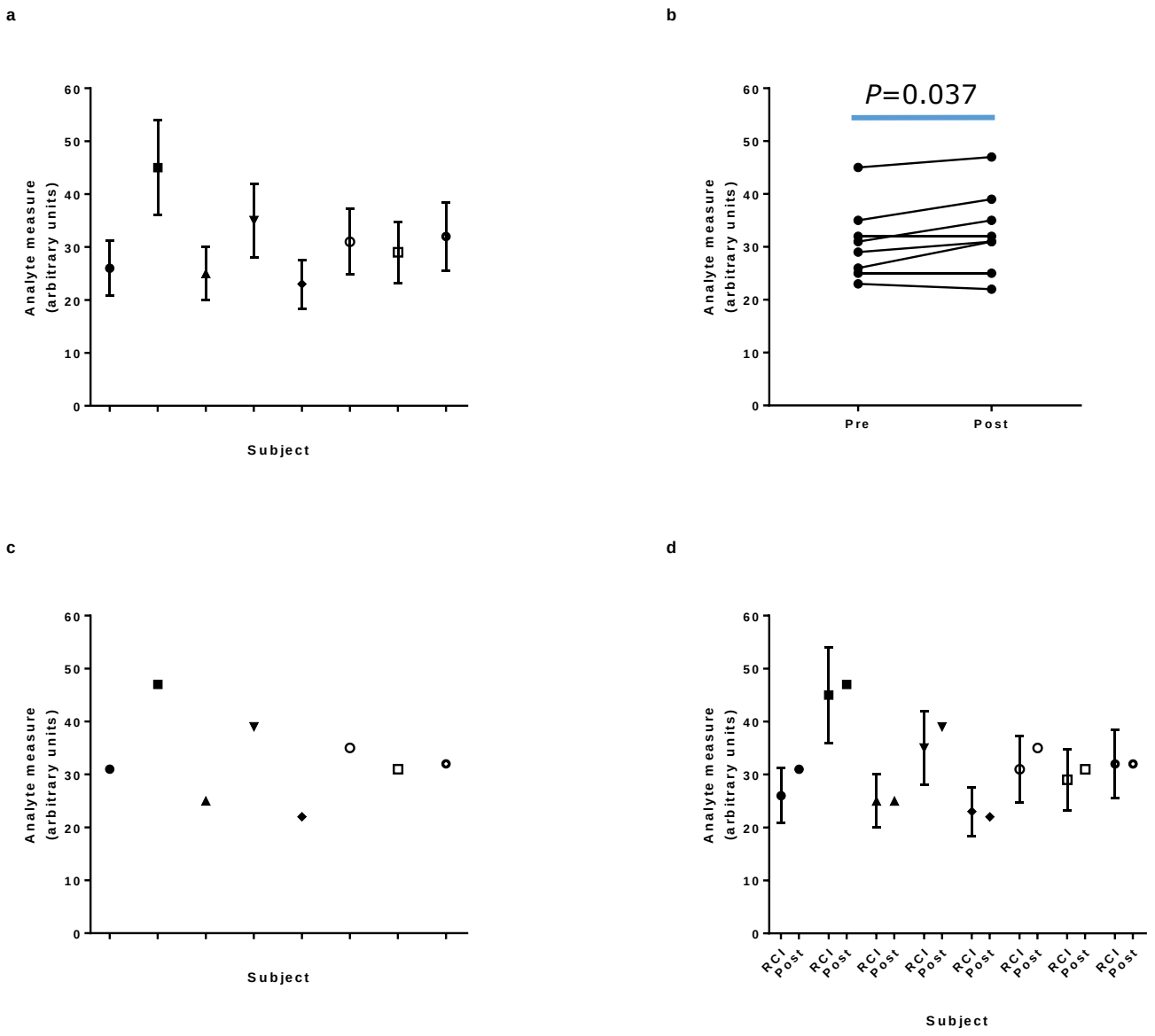
a
Analyte measure (arbitrary units)
Subject
b
P=0.037
Analyte measure (arbitrary units)
Pre
Post
c
Analyte measure (arbitrary units)
Subject
d
Analyte measure (arbitrary units)
RCI
Post
Subject

**Figure 2**

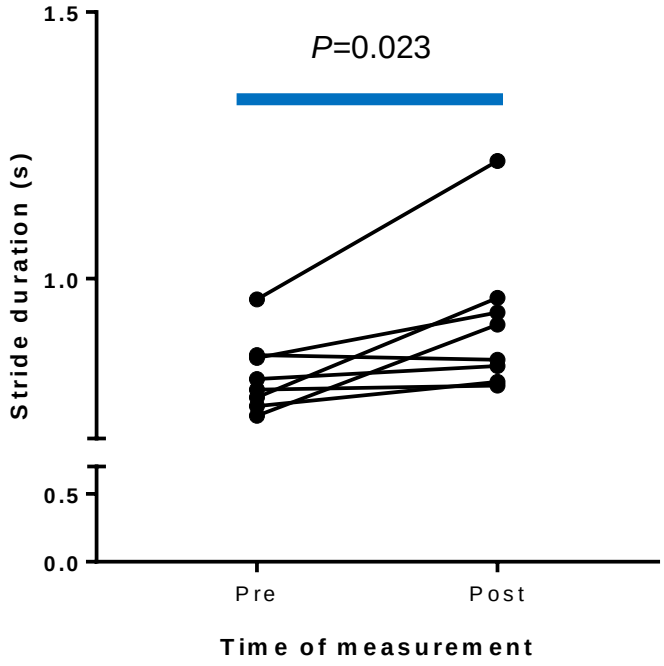
P=0.023
1.5
1.0
0.5
0.0
Stride duration (s)
Pre
Post
Time of measurement

**Figure 3**

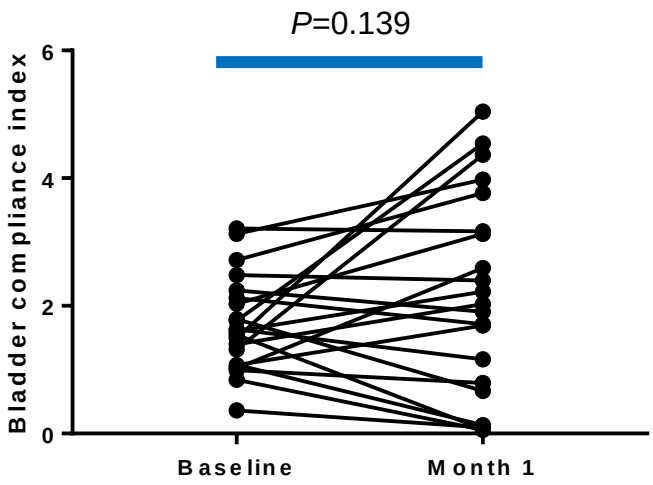

**Table 1: Baseline and repeat testing results**

| Baseline | Repeat | Repeat / Baseline % |
|---|---|---|
| 26 | 31 | 119 |
| 45 | 47 | 104 |
| 25 | 25 | 100 |
| 35 | 39 | 111 |
| 23 | 22 | 96 |
| 31 | 35 | 113 |
| 29 | 31 | 107 |
| 32 | 32 | 100 |

In this example we know that previous experiments have shown that the combination of various sources of 'experimental noise' (see text) can make retest results vary by as much as 20% from baseline values in normal, untreated animals.

In this example we use paired Student's t-test to compare baseline values with repeat test results after an experimental intervention, producing $t = 2.204$, $P = 0.037$.

Therefore population-level analysis alone suggests significant increase between baseline and repeat testing. However, inspection of the data above reveals that for none of the subjects does the retest value exceed what might be attributable to experimental noise alone.

This type of paradoxical outcome is most likely to occur in small sample size experiments as are frequently used in neuroscience.

**Table 1: Calculation of the upper limit of the reference change interval for hindlimb stride duration and comparison with measured effect of drug administration**

| Condition | Sheep | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Pre-drug** | 0.743 | 0.812 | 0.792 | 0.778 | 0.761 | 0.852 | 0.856 | 0.961 |
| **Upper RCI boundary** | 0.875 | 0.957 | 0.933 | 0.916 | 0.896 | 1.004 | 1.008 | 1.132 |
| **Post-drug** | 0.914 | 0.836 | 0.800 | 0.964 | 0.807 | 0.937 | 0.848 | 1.221 |
| **Post>upper RCI boundary?** | Yes | No | No | Yes | No | No | No | Yes |

For each sheep (columns) we show the duration of hindlimb stride before ('pre-drug') and after ('post-drug') administration of drug X. After calculating the reference change value .we have used this, in conjunction with the Pre-drug measurement to define the upper boundary of the reference change interval, see equation (2). In the bottom row we ask whether the Post-drug measurement exceeds the upper boundary of the reference change interval.

**Table 3: Analysis of change in bladder compliance index at 1 month after injection of intraspinal chondroitinase ABC**

| Baseline | Month 1 | Baseline + RCV | Baseline - RCV | Compliance increased? | Compliance decreased? | Z-score |
|---|---|---|---|---|---|---|
| 1.609 | 2.223 | 4.162 | -0.944 | No | No | 0.472 |
| 0.993 | 0.792 | 2.569 | -0.583 | No | No | -0.250 |
| 1.396 | 2.028 | 3.611 | -0.819 | No | No | 0.560 |
| 2.485 | 2.398 | 6.429 | -1.459 | No | No | -0.043 |
| 3.214 | 3.169 | 8.313 | -1.886 | No | No | -0.017 |
| 2.121 | 1.715 | 5.486 | -1.245 | No | No | -0.236 |
| **1.492** | **5.044** | **3.859** | -0.876 | **Yes** | No | **2.944** |
| 1.768 | 4.544 | 4.574 | -1.038 | No | No | 1.940 |
| 2.716 | 3.767 | 7.026 | -1.594 | No | No | 0.478 |
| 1.792 | 0.670 | 4.636 | -1.052 | No | No | -0.774 |
| 0.843 | 0.058 | 2.181 | -0.495 | No | No | -1.151 |
| 1.022 | 2.590 | 2.644 | -0.600 | No | No | 1.896 |
| 1.070 | 1.692 | 2.768 | -0.628 | No | No | 0.719 |
| 3.129 | 3.977 | 8.095 | -1.837 | No | No | 0.335 |
| 0.365 | 0.095 | 0.944 | -0.214 | No | No | -0.914 |
| 2.031 | 3.129 | 5.254 | -1.192 | No | No | 0.668 |
| 1.546 | 0.067 | 4.000 | -0.908 | No | No | -1.183 |
| 1.079 | 0.134 | 2.791 | -0.633 | No | No | -1.083 |
| 2.244 | 1.907 | 5.805 | -1.317 | No | No | -0.186 |
| 1.633 | 1.161 | 4.225 | -0.959 | No | No | -0.357 |
| **1.316** | **4.369** | **3.404** | -0.772 | **Yes** | No | **2.868** |