EJN European Journal of Neuroscience

FENS Federation of European Neuroscience Societies

European Journal of Neuroscience, pp. 1–9, 2018 doi:10.1111/ejn.13847

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] Matthew A. Howard III,[4] Lawrence Moon[5] and Unity Jeffery[6]
[1]Department of Small Animal Clinical Sciences, Texas A&M University, College Station, TX 77843, USA
[2]Statistical Sciences, GlaxoSmithKline, Medicines Research Centre, Stevenage, Hertfordshire, UK
[3]The University of Texas Graduate School of Biomedical Sciences at Houston, Houston, TX, USA
[4]Department of Neurosurgery, University of Iowa Hospitals and Clinics, Iowa City, IA, USA
[5]Neurorestoration Department, Wolfson Centre for Age-Related Diseases, King's College London, University of London, London, UK
[6]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX, USA

Keywords: intraindividual, partition, reference change value, translation, variability

Abstract

In animal experiments, neuroscientists typically assess the effectiveness of interventions by comparing the average response of groups of treated and untreated animals. While providing useful insights, focusing only on group effects risks overemphasis of small, statistically significant but physiologically unimportant, differences. Such differences can be created by analytical variability or physiological within-individual variation, especially if the number of animals in each group is small enough that one or two outlier values can 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 whether the change between two values reflects disease progression or known analytical and physiological 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 the reference change interval can be applied within neuroscience. This form of analysis provides a measure of outcome at an individual level that complements traditional group-level comparisons, and therefore, introduction of this technique into neuroscience can enrich interpretation of experimental data. 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; The Academy of Medical Sciences, 2015; Steward, 2016). Many remedies for methodological shortcomings are available, focusing on both animal model relevance and experimental design (Landis *et al.*, 2012; Begley, 2013; Button *et al.*, 2013; Bate & Clark, 2014a; Garner, 2014; The Academy of Medical Sciences, 2015; Steward, 2016; 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 and median) between groups. Unfortunately, while widely used, the derived P values provide only an estimate of the consistency of the

Correspondence: Nick D. Jeffery, as above. E-mail: njeffery@cvm.tamu.edu

Received 25 July 2017, revised 14 December 2017, accepted 15 January 2018

Edited by John Foxe. Reviewed by Stanley Lazic, AstraZeneca, UK; Cyril Pernet, University of Edinburgh, UK

All peer review communications can be found with the online version of the article.

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. While 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 the following: a) how many individuals attain a benefit that is greater than that which can be 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 the more conventional 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 meant by 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 overestimation of the true value (Theodorsson *et al.*, 2014). The standard deviation (SD) or the *coefficient of variation* (CV) (= SD/mean) provides summary measures of variation of the observed 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 analyser 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 performing multiple replicate analyses of each sample, or more generally by repeatedly measuring the same physical material, it can be quantified and this estimate used when deducing the sources of variation in the final results (see below).

Biological variation can be divided into intraindividual and interindividual variation. Both forms of biological variation include predictable sources of variation (*e.g.* variation in blood hormone concentration over the reproductive cycle of an individual or variation in blood hormone concentration between individuals at corresponding points in the reproductive cycle). Researchers typically already pay close attention to these known sources of variation and standardize where possible, but unknown sources of biological variation also have important implications for data interpretation.

Intraindividual variation arises because, as would be expected, there is some variation of analyte levels between samples obtained from the same individual at different times that are above and beyond analytical variability. For instance, blood cholesterol concentration varies (even at the same time each day) between one day and another, even in healthy individuals (Rotterdam *et al.*, 1987). The magnitude of this variation varies between analytes: there is much less intraindividual variation for some (*e.g.* chloride) than others (*e.g.* triglycerides: Nunes *et al.*, 2010).

Interindividual 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, 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 interindividual variation, as well as by analytical and pre-analytical variation. For example, for some analytes, the interindividual variation may be very much greater than the intraindividual variation, implying that what might appear to be only a subtle change when viewed in the context of the population as a whole may be critical to the health of a specific individual. For such analytes, it is critical to define limits of expected physiological change for each individual and not rely on group-level assessment. Additionally, the risk of drawing false conclusions is greatest when the number of experimental subjects within the groups is small, as is common in neuroscience. Also, 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 Fig. 1, Table 1). The addition of an analysis of individual 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 using the *reference change value* (RCV):

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

$$\text{where RCV} = z_{\mathrm{p}} * \sqrt{2} * \sqrt{\left(\mathrm{CV}_{\mathrm{I}}^2 + \mathrm{CV}_{\mathrm{A}}^2\right)} \quad (2)$$

$\mathrm{CV_I}$ is the intraindividual coefficient of variation, and $\mathrm{CV_A}$ is the analytical coefficient of variation (Fraser, 2001). 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 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 behavioural experiments. For instance, as noted elsewhere (National Center for 3Rs, 2017), behavioural 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 and fire alarms.), 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 estimated in clinical pathology by considering the random differences in the repeated measurements of the analyte level in one sample by splitting it into multiple aliquots. This approach can easily be applied to routine laboratory techniques such as ELISA or PCR, but is rarely appropriate in behavioural 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 the individual animals or repeatedly deriving values for kinematic variables from a single original source data set (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 behavioural 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 intraindividual 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 intraindividual and interindividual variation

Many experiments involve repeatedly measuring outcomes in a group, or groups, of animals. The inter- and intraindividual variation can be estimated using nested random effects ANOVA (Fraser & Harris, 1989) or repeated measures mixed models (Marchenko, 2006; Bate & Clark, 2014b), which provide flexibility when modelling the correlation between repeated measurements taken on an individual. Typically, estimates of intra- and interanimal 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

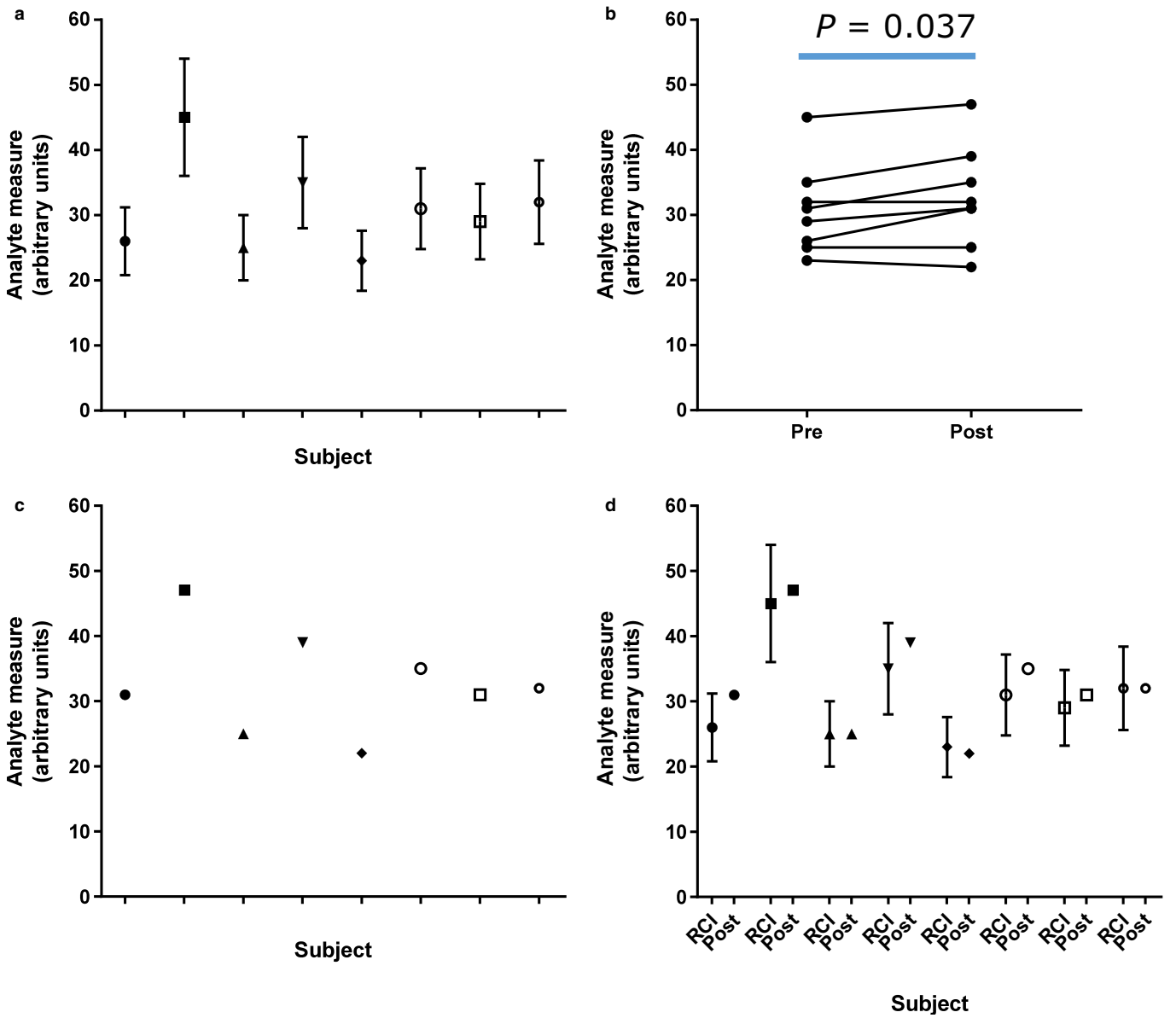

FIG. 1. Illustration of the relationship between 'experimental noise' and group-level change. In this theoretical example, we assume that previous analysis has revealed that the combination of experimental inaccuracy and physiological variability can be responsible for changes of up to 20% between sequential 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. Table 1 shows the results for each individual subject.

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 rearranging Eqn (2) so that

$$\text{z-score} = \text{proportional change from baseline}/\sqrt{2} * \sqrt{\left(\text{CV}_\text{I}^2 + \text{CV}_\text{A}^2\right)} \quad (3)$$

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

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, paired Student's *t*-test was used 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 – and the graphic representation in Figure 1 – 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.

pre-analytical, analytical and intraindividual variation alone. To determine whether 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 Eqn (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 Eqn (1) will reduce to as follows:

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

As before, the boundary values beyond which real change can be deduced to have occurred in an individual (the RCI) are calculated using Eqn (1).

Examples

Example 1

Derivation and use of measures of analytical (CV_A), intraindividual (CV_I) and interindividual (CV_G) variability

Here, we use 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 the Nested Design Analysis module within InVivoStat (version 3.7; http://invivostat.co.uk/), the analytical, intra- and interanimal coefficients of variation, denoted by CV_A, CV_I and CV_G, respectively, were estimated from serial measurements made on trained normal animals before drug administration (the raw data are 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 min) 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 ('between-trial, within-day'), intraindividual ('between-day, within-sheep') and interindividual (between-sheep) to derive means and standard deviations which were used to derive the respective coefficients of variation.

Interindividual, intraindividual and analytical standard deviations were 0.039, 0.033 and 0.045, respectively, and the overall mean stride duration was 0.862, resulting in the following coefficients of variation:

$$CV_G = 0.039/0.862 = 4.5\%;$$
$$CV_I = 0.033/0.862 = 3.8\%;$$
$$CV_A = 0.045/0.862 = 5.2\%$$

The reference change value was 17.9%, calculated using Eqn (1) above; therefore, any sheep in which the post-drug measurement was changed by more than 17.9% of the pre-drug value (*i.e.* outside the upper or lower boundary of the reference change interval) can be considered to have undergone 'real' change as a consequence of drug administration. It is of course possible that the intervention may have had an effect that produces an outcome that does not lie outside the reference change interval but such an effect can be inferred to be negligible in terms of biomedical meaning (and, in clinical medicine, a drug having an effect of such small magnitude would not likely be considered valuable).

To illustrate the different, and complementary, perspectives provided by population-level and RCV within-individual analysis, see Fig. 2, which depicts the difference in stride duration in these same sheep before and after drug administration. A paired Student's *t*-test (on log-transformed data to satisfy the parametric assumptions) indicates a significant increase in stride duration following this intervention ($P = 0.021$).

When the complementary individual analysis is applied to the same data set (Table 2), in only three of eight 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 Eqn (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, 2007) 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 were analysed using the Nested Design Analysis module in InVivoStat (version 3.7; http://invivostat.co.uk/) to estimate the intra- and interindividual variability (the raw data are available as Supplementary Material). With the mean compliance value, this was used to derive the intra- and interanimal coefficients of variation (CV_I and CV_G, respectively), where in this example,

$$CV_I = 0.865/1.508 = 0.574$$

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

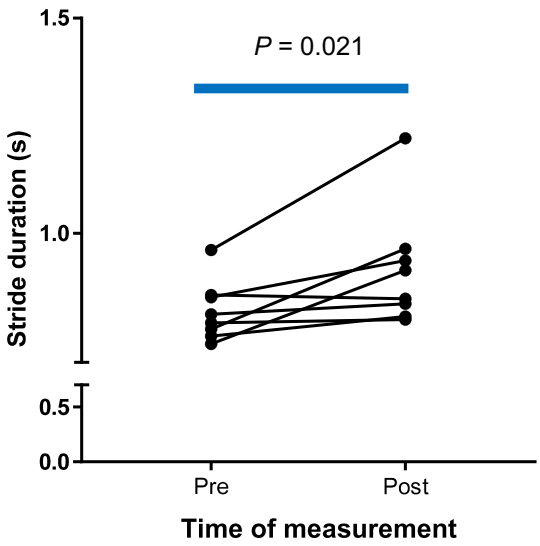

FIG. 2. Hindlimb stride duration in a group of eight trained sheep walking on a treadmill before and after administration of drug X. Paired Student's *t*-test suggests a significant effect of drug X on stride duration ($P = 0.021$).

$$RCV = 1.96 * \sqrt{2} * CV_I = 2.764 * 0.574 = 1.590$$

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 159% 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 produces a *P* value of 0.139, conventionally regarded as not statistically significantly different. The data from individual dogs are shown in Fig. 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 interanimal variability) or of difficulties in making repeatable measurements

TABLE 2. Calculation of the upper limit of the reference change interval for hindlimb stride duration and comparison with measured effect of drug administration

	Sheep							
Condition	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, this has been used in conjunction with the pre-drug measurement to define the upper boundary of the reference change interval, see Eqn (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**

Values shown in bold font indicate those for individuals in which 1 month outcomes lie outside the reference change interval.

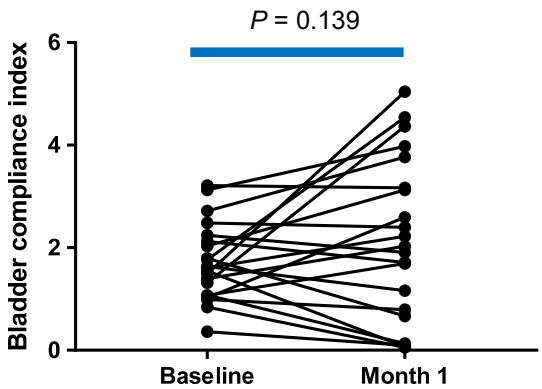

FIG. 3. Baseline and 1-month post-intervention measures of bladder compliance index for dogs that had received intraspinal injection of chondroitinase ABC at time 0. Group-level analysis by paired Student's *t*-test reveals a non-significant difference between time points.

(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 but also 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-centred 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 behaviour (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 designed 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, behavioural 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 nonparametric 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 behavioural data, because these provide the most transparent examples, the same methods can be applied to many other types of quantitative or semiquantitative neuroscience data including analysis of cell activity or production of specific molecules.

Supporting Information

Additional supporting information can be found in the online version of this article:
Appendix S1. Supplementary text describing the data acquisition and statistical analytical methods used in Examples 1 and 2.
Table S1. Sheep treadmill data for Example 1.
Table S2. Dog compliance data for Example 2.

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.

Data accessibility

The primary data are available as Supplementary Material.

Author contributions

NJ, UJ and SB designed the article approach and analysed the data. SS, MH and LM collected and analysed the data. NJ, UJ and SB wrote the article.

References

ARRIVE Guidelines (2017). http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-reporting-the-arrive-guidelines-for-reporting-animal-research/

Bate, S.T. & Clark, R.A. (2014a). *The Design and Statistical Analysis of Animal Experiments*. Cambridge University Press, Cambridge, UK.

Bate, S.T. & Clark, R.A. (2014b). *The Design and Statistical Analysis of Animal Experiments*. Repeated measures mixed models, p191. Cambridge University Press, Cambridge, UK.

Begley, C.G. (2013) Six red flags for suspect work. *Nature*, **497**, 433–434.

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**, 365–376.

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 (2017). http://3rs.ccac.ca/en/research/reduction/experimental-design.html

Fraser, C.G. (2001). Chapter 3. Changes in serial results. In Fraser, C.G. (Ed.), *Biological Variation: From Principles to Practice*. AACC Press, Washington, DC, pp. 67–90.

Fraser, C.G. & Harris, E.K. (1989) Generation and application of data on biological variation in clinical chemistry. *Crit. Rev. Cl. Lab. Sci.*, **27**, 409–437.

Friedrichs, K.R., Harr, K.E., Freeman, K.P., Szladovits, B., Walton, R.M., Barnhart, K.F., Blanco-Chavez, J. & American Society for Veterinary Clinical Pathology (2012) ASVCP reference interval guidelines: determination of de novo reference intervals in veterinary species and other related topics. *Vet. Clin. Pathol.*, **41**, 441–453.

Garner, J.P. (2014) The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *ILAR J.*, **55**, 438–456.

Gouveia, K. & Hurst, J.L. (2013) Reducing mouse anxiety during handling: effect of experience with handling tunnels. *PLoS ONE*, **20**, e66401.

Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. (2015) The fickle P value generates irreproducible results. *Nat. Meth.*, **12**, 179–185.

Harris, E.K. & Yasaka, T. (1983) On the calculation of a" reference change" for comparing two consecutive measurements. *Clin. Chem.*, **29**, 25–30.

Kilkenny, C., Browne, W.J., Cuthi, I., Emerson, M. & Altman, D.G. (2012) Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *Vet. Clin. Pathol.*, **41**, 27–31.

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B. *et al.* (2012) A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, **490**, 187–191.

Marchenko, Y. (2006) Estimating variance components in Stata. *Stata J.*, **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**, 191–198.

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. *et al.* (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**, 107–112.
Shaver, J.P. (1993) What Statistical Significance Testing Is, and What It Is Not. *J. Exp. Educ.*, **61**, 293–316.
Snapinn, S.M. & Jiang, Q. (2007) Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, **8**, 31.
Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L. *et al.* (2014) Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Meth.*, **11**, 629–632.
Steward, O. (2016) A Rhumba of "R's": Replication, Reproducibility, Rigor, Robustness: What Does a Failure to Replicate Mean?. *eNeuro*, **7**, 3.
The Academy of Medical Sciences (2015). Reproducibility and reliability of biomedical research: improving research practice: Symposium report. https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf
Theodorsson, E., Magnusson, B. & Leito, I. (2014) Bias in clinical chemistry. *Bioanalysis*, **6**, 2855–2875.
Walton, R.M. (2012) Subject-based reference values: biological variation, individuality, and reference change values. *Vet. Clin. Pathol.*, **41**, 175–181.