We need to talk about reliability: Making better use of test-retest studies for study design and interpretation

Granville J. Matheson[1]

[1]**Center for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet & Stockholm Health Care Services, Stockholm County Council, Stockholm, Sweden.**

Corresponding author:
Granville J. Matheson[1]

Email address: `granville.matheson@ki.se`

ABSTRACT

Positron emission tomography (PET), along with many other fields of clinical research, is both time-consuming and expensive, and recruitable patients can be scarce. These constraints limit the possibility of large-sample experimental designs, and often lead to statistically underpowered studies. This problem is exacerbated by the use of outcome measures whose accuracy is sometimes insufficient to answer the scientific questions posed. Reliability is usually assessed in validation studies using healthy participants, however these results are often not easily applicable to clinical studies examining different populations. I present a new method and tools for using summary statistics from previously published test-retest studies to approximate the reliability of outcomes in new samples. In this way, the feasibility of a new study can be assessed during planning stages, and before collecting any new data. An R package called relfeas also accompanies this article for performing these calculations. In summary, these methods and tools will allow researchers to avoid performing costly studies which are, by virtue of their design, unlikely to yield informative conclusions.

INTRODUCTION

In the assessment of individual differences, reliability is typically assessed using test-retest reliability, inter-rater reliability or internal consistency. Reliability is a measure of the accuracy or consistency of an outcome, the distinguishability of individual measurements, and a measure of the signal-to-noise ratio in a set of data. It is an approximation of the fraction of the total variance which is not accounted for by measurement error (i.e. random noise in the measurement). Accordingly, a reliability of 1 means that all variability is attributable to true differences and there is no measurement error, while a reliability of 0 means that all variability is accounted for by measurement error. A reliability of 0.5 means that there is equal measurement- and error-related variance: this means that an individual obtaining a score equal to the mean of the group *could* have the highest, or the lowest, underlying true value of the outcome in the group. Reliability can therefore be increased either by reducing the measurement error, or by increasing the amount of true interindividual variability in the sample such that measurement error is proportionally smaller.

In clinical research, most statistical inference is performed using the null hypothesis significance testing (NHST) paradigm. A result is considered significant when the p value is less than the prespecified alpha threshold, and the null hypothesis is rejected. In this paradigm, study design is performed by considering the risk of type I and type II errors. In practice, there is a great deal of attention given to the minimisation of type I errors, i.e. false positives. This usually takes the form of correction for multiple comparisons. Another important consideration is the minimisation of type II errors, i.e. false negatives. This takes the form of power analysis: reasoning about the number of participants to include in the study based on the effect size of interest. In contrast, comparatively little consideration is usually given to the reliability of the measures to be used in the study. The reliability of outcome measures

limits the range of effect sizes which can be expected, which is vital information for study design. Lower reliability of outcome measures leads to diminished power, or alternatively an increase in the risk of type M (magnitude) and type S (sign) errors (Gelman and Carlin 2014). In extreme cases, it can lead to what has been called type III error, defined as obtaining the right answer, but to the wrong question (Kimball 1957): if measured values are too dissimilar to the underlying 'true' values, then a statistical test will not be inferring on meaningful outcomes but simply random noise.

Assessment of reliability is critical both for study design and interpretation. However it is also dependent on characteristics of the sample: we can easily use a bathroom scale to contrast the weight of a feather and a brick. In all cases, the scale will correctly indicate that the brick weighs more than the feather. However, we cannot conclude from these results that this bathroom scale can reliably measure the weight of bricks, and proceed to use it to examine the subtle drift in weight between individual bricks produced by a brick factory. In this way, the reliability of a measure is calibrated to the inter-individual differences in that sample. In psychometrics, reliability is often assessed using internal consistency: this involves examining the similarity of the responses between individual items of the scale, compared to the total variability in scores within the sample. This means that the reliability of the scale can be estimated using only the data from a single completion of the scale by each participant. However, for most clinical/physiological measures, estimation of reliability by internal consistency is not possible, as the measurement itself cannot be broken down into smaller representative parts. For these measures, reliability can only be assessed using test-retest studies. This means that measurements are made twice on a set of individuals, and the inter- and intra-individual variability are compared to determine the reliability. If the variability of the outcome is similar in the test-retest and applied studies, it can reasonably be assumed that the measure will be equally reliable in both studies.

When examining clinical patient samples however, it often cannot be assumed that these samples are similar to that of the test-retest study. One solution is to perform a new test-retest study in a representative sample. However, for outcome measures which are invasive or costly, it is usually not feasible to perform test-retest studies using every clinical sample which might later be examined. Positron emission tomography (PET), which allows imaging of in-vivo protein concentrations or metabolism, is both invasive and costly: participants are injected with harmful radioactivity, and a single measurement can cost upwards of USD 10 000. In PET imaging, it is usually only young, healthy men who are recruited for test-retest studies. These samples can be expected to exhibit low measurement error, but may also be expected to show limited inter-individual variability. Despite reporting of reliability in test-retest studies being common practice, when the reported reliability is low, little consideration can given to these results on the basis of insufficient inter-individual variability, i.e. it is assumed that there will be more variation in clinical comparison studies and that the reliability does not accurately reflect this. This is certainly true in some circumstances. However, when it is not true, it can lead to the design of problematic studies whose ability to yield biologically meaningful conclusions is greatly limited. This is costly both in time and resources for researchers, and leads to the needless exposure of participants to radiation in PET research. It is therefore important to approximate the reliability of a measure for the sample of interest before data collection begins for studies investigating individual differences.

In this paper, I present how the reliability can be used for study design, and introduce a new method for roughly approximating the reliability of an outcome measure for new samples based on the results of previous test-retest studies. This method uses only the reliability and summary statistics from previous test-retest studies, and does not require access to the raw data. Further, this method allows for calculation of the characteristics of a sample which would be required for the measure to reach sufficient levels of reliability. This will aid in study planning and in assessment of the feasibility of new study designs, and importantly can be performed *before* the collection of any new data. I will demonstrate how these methods can be utilised by using five examples based on published PET test-retest studies. This paper is also accompanied by an R package called *relfeas* (www.github.com/mathesong/relfeas), with which all the calculations presented can easily be applied.

METHODS

Reliability

Reliability itself is defined as the proportion of the total variance which is due to true differences.

$$reliability = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} = \frac{\sigma_t^2}{\sigma_{tot}^2} \tag{1}$$

where σ^2 represents the variance due to different sources (t: true, e: error, and tot: total). Measures of internal consistency (such as Cronbach's α), and of test-retest reliability (such as the ICC), are both estimates of this same definition of reliability.

Reliability can therefore be considered a measure of the distinguishability of measurements (Carrasco et al. 2014). For example, if the uncertainty around each measurement is large, but inter-individual variance is much larger, scores can still be meaningfully compared between different individuals. Similarly, even if a measure is extremely accurate, it is still incapable of meaningfully distinguishing between individuals who all obtain the same score.

Test-Retest Reliability

Test-retest reliability is typically defined using the intraclass correlation coefficient (ICC). In measures for which there are no systematic effects of the measurement number, such as PET test-retest studies in which individuals are each measured twice by the same PET system, we use the one-way ANOVA fixed effects model (Shrout and Fleiss 1979). This is the most conservative formulation of the ICC. It is described by the following equation:

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \approx reliability \tag{2}$$

where $\mathrm{MS_B}$ represents the between subjects mean sum of squares, $\mathrm{MS_W}$ represents the within subject mean sum of squares, and k represents the number of observations (usually 2). The ICC is an approximation of the true population reliability: while true reliability can never be negative (equation 1) one can obtain negative ICC values if the $\mathrm{MS_W}$ is larger than $\mathrm{MS_B}$ (equation 2) in which case the reliability can be treated as zero.

Measurement Error

Each measurement is made with an associated error, which can be described by its standard error (σ_e). It can be estimated as the square root of the within subject mean sum of squares ($\mathrm{MS_W}$), which is used in the calculation of the ICC above (Baumgartner et al. 2018).

$$\sigma_e^2 = MS_W = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_i)^2 \tag{3}$$

where n represents the number of participants, i represents the subject number, j represents the measurement number, k represents the number of measurements per subject, y represents the outcome and $\bar{y}_i$ represents the mean outcome for that subject.

The standard error can also be estimated indirectly by rearranging equation 1, using the ICC as an estimate of reliability. This is often referred to as the standard error of measurement (SEM) (Weir 2005; Harvill 1991).

$$SEM = \sigma_t\sqrt{1 - ICC} \approx \sigma_e \tag{4}$$

in which the σ_t is the standard deviation of all measurements (both test and retest measurements for test-retest studies).

The standard error can be expressed either in the units of measurement, or relative to some property of the sample: it can be i) scaled to the variance of the sample as an estimate the reliability (ICC, Cronbach's

α), ii) scaled to the mean of the sample as an estimate the relative uncertainty (the within-subject coefficient of variation, WSCV) (Baumgartner et al. 2018), or iii) unscaled as an estimate the absolute uncertainty (σ_e or SEM).

The relative or absolute uncertainty can be used to calculate the smallest detectable difference (SDD) between two measurements in a given subject which could be considered sufficiently large that it is unlikely to have been due to chance alone (say, according to a 95% confidence interval, i.e. using $z_{(1-\alpha/2)}$=1.96 below) (Weir 2005; Baumgartner et al. 2018). This is calculated using the following equation.

$$SDD = \sqrt{2} \times z_{(1-\alpha/2)} \times \sigma_e \tag{5}$$

It can also be extended to a group level examining differences in means.

$$SDD = \sqrt{2} \times z_{(1-\alpha/2)} \times \frac{\sigma_e}{\sqrt{n}} \tag{6}$$

It is important to mention that this measure assumes that all measurements belonging to the original test-retest study and later application study exhibit exactly the same standard error. Further, this measure is not a power analysis: the SDD simply infers that a change in outcome or mean outcome of a group of measurements before and after an experimental manipulation is larger than would be expected by chance. For performing a power analysis for the within-subject change for application to a new study, one must consider the expected effect size relative to the standard deviation of the within-individual changes in the outcome measure.

Standard thresholds

In psychometrics, reliability can be calculated with one scale by examining its internal consistency. Despite being calculated from the consistency of responding between items, reliability assessed by internal consistency amounts to the same fundamental definition of reliability, namely the ratio of the true to the total variance in the data. It is considered good practice in psychometric studies to calculate the reliability of a scale using measures of internal consistency in the study sample prior to performing statistical inference. The reliability can be used to confirm that the data is sufficiently capable of distinguishing between individuals, for which Nunnally (1978), recommended 0.7 as a default lowest acceptable standard of reliability for scales used in basic research, and 0.8 as adequate. For applied settings in which important decisions are made based on measured outcomes, he suggests a reliability of 0.9 as a minimum and 0.95 as adequate, since even with a reliability of 0.9, the standard error is almost a third the size of the standard deviation.

Test-retest reliability has traditionally been defined by more lenient standards: Fleis (1986) defined ICC values between 0.4 and 0.75 as good, and above 0.75 as excellent. Cicchetti (1994) defined 0.4 to 0.59 as fair, 0.60 to 0.74 as good, and above 0.75 as excellent. These standards, however, were defined considering the test-retest (as opposed to internal consistency) reliability of psychometric questionnaires (Shrout and Fleiss 1979; Cicchetti 1994) for which changes from test to retest could be caused by measurement error, but could also be caused by actual changes in the underlying 'true' value. In PET measurement where protein concentrations are estimated, or indeed in many other physiological measures, these within-subject fluctuations can usually be assumed to be negligibly small over short time periods. For this reason, these standards cannot be considered to be directly applicable. More conservative standards have been proposed by Portney & Watkins (2015), who define values between 0.5 and 0.75 as "poor to moderate", 0.75 to 0.9 as "good", and above 0.9 as acceptable for "clinical measures". These standards correspond more closely with those defined for the internal consistency of psychometric instruments, and can be considered to be more applicable for measures for which 'true' changes are thought to be negligible.

Relation to effect size estimates

It is important to consider that reliability is related to effect sizes, and more specifically effect size attenuation. This applies both to studies examining correlations with a continuous variable, and those

making group comparisons. This relation is described by the following equation (Spearman 1904 ; Nunnally 1970).

$$r_{ObsA,ObsB} = r_{A,B} \times \sqrt{reliability_A \times reliability_B} \tag{7}$$

where $r_{A,B}$ is the true association between variables A and B, and $r_{ObsA,ObsB}$ is the observed correlation between the measured values for A and B. When comparing two or more measures, each of which are recorded with some degree of imprecision, the combined reliability is equal to the square root of their product. This defines the degree to which an association is attenuated by the imprecision of the recorded outcomes, and sets an upper bound on the strength of association between variables that can be expected (Vul et al. 2009). This attenuation is important when performing power analysis as it decreases the range of potential effect size estimates which can be expected.

When planning a study, one determines the sample size required using power analysis. Power analysis is often performed by estimating the sample size for an estimated effect size. Morey & Lakens (2016) point out that this makes studies vulnerable to the accuracy by which the effect size is approximated. It has been suggested that a better approach is to power studies for the smallest effect size of interest, such that all interesting effects will be reliably detected. More conservative still is the suggestion to power studies according to the level at which meaningful differences between individual studies can be reliably detected (Morey and Lakens 2016). These strategies would ensure that the scientific literature would produce robust findings, and that the effects in individual studies could be statistically compared respectively. However, for costly and/or invasive methods, such an ideal is not usually attainable. Rather, for studies with these constraints, we require a lower standard of evidence, with robust conclusions perhaps left for meta-analyses, and specifically meta-analyses of pre-registered studies to avoid effect size inflation due to low power (Cremers, Wager, and Yarkoni 2017) or publication bias (Ferguson and Heene 2012). To determine the feasibility of an individual study using these methods, it may be more useful not to consider the point at which an individual study is *ideal* for the scientific literature as described by Morey & Lakens (2016), but rather the point at which the study will simply *not be bad*, i.e. will at least be capable of answering the research question at hand given realistic assumptions.

For this latter aim, we can perform power analysis for study feasibility assessment in terms of a maximum realistic strength of association in biological terms. In this way, the corresponding effect size can be calculated after taking the effect size attenuation into account. If a study will be insufficiently powered even for such a strength of association, then the study is certainly miscalibrated for the detection of the effect of interest. In this way, this method is useful for deciding which research questions can be rejected out of hand given the relevant constraints on the sample size.

Extending Test-Retest Reliability

In order for the results of previous test-retest studies to be applicable for new samples with different characteristics, we need to be able to extrapolate the reliability for new samples. To this end, I propose the use of the extrapolated ICC (below). This calculation is based upon two assumptions.

Assumption 1 The absolute measurement error (σ_e) will be either similar between groups and/or studies, or the extent to which it will be larger or smaller in a new sample can be approximated.

Comment: Since test-retest studies are typically conducted using young, healthy volunteers, and not patient samples, these groups will usually provide a good estimate at least of the maximum bound of the precision of the measurement. For comparison groups for which measurement error is likely to be larger, it can either be assumed to be similar (a liberal assumption), or the extent to which it is larger can be approximated as a small multiple (a conservative assumption, e.g. 20% larger in the patient group).

Assumption 2 The absolute measurement error (SEM) is likely to be relatively stable across different ranges of the same outcome measure (i.e. an individual with low score will have the same SEM as an individual with a high score on the same outcome measure).

Comment: There is often some variation expected in the measurement error for different values of the outcome (e.g. less error for individuals with a higher value, and more error for individuals with a lower value of the outcome), however this variation is usually expected to be fairly small. If, on the other hand, this variation across the range of the outcome measure *is* expected to be substantial, then results from such a study should be interpreted with caution in any case.

Assumption Summary These assumptions are reasonable given that the goal is a rough approximation of the reliability in a new sample in order to assess feasibility in study design and interpretation. The assumptions are primarily based on the fact that most previous and current test-retest studies are performed on young, healthy participants. For planning a new study for which large deviations are expected from either of these two assumptions, whose effects cannot be approximated, a new test-retest study should probably be conducted before proceeding in any case.

The extrapolated ICC

The ICC can be expressed using the SEM and the SD (σ) of the original study from equation 4.

$$ICC = 1 - \frac{SEM^2}{\sigma^2} \tag{8}$$

Given these two assumptions above, we can approximate the ICC for a new sample given only the standard deviation of the new sample (and an approximation of the inflation extent of the SEM):

$$ICC_{NewStudy} = 1 - \frac{(\rho \times \sigma_{TRT})^2}{\sigma^2_{NewStudy}} = 1 + \frac{\rho^2 \sigma^2_{TRT}(1 - ICC_{TRT})}{\sigma^2_{NewStudy}} \tag{9}$$

where ρ represents the error inflation factor, representing the multiplicative increase in expected measurement error in the new study, for whichever reason. Using $\rho = 1$ assumes that the SEM is the same between studies, while $\rho = 1.2$ would assume a 20% increase in the measurement error. If ρ is thought to be substantially greater than 1, then a new test-retest study should probably be performed. With $\rho = 1$ or $\rho \approx 1$, the difference between the ICC of the original study and the extrapolated ICC of a new study will be determined primarily by the SD of the new study.

Studies making use of continuous independent variables (correlations) For a new study which is going to examine a correlation, reliability is simply determined by the variance of the study group. With more variation in the study sample, the reliability is therefore higher, all else being equal. Increasing sample variability can be attained by implementing wider recruiting strategies, and not simply relying on convenience sampling (see Henrich et al., 2010). This also has the advantage of increasing the external reliability of findings.

It should be noted, however, that increasing sample variability by including individuals who differ on another variable which is known to be associated with the dependent variable (for example, age) would *not* necessarily increase the reliability despite increasing the variance of the study group. In these cases, this variable should be included in the statistical test of an applied study as a covariate, and hence the variance of the unstandardised residuals *after* accounting for the covariate would therefore be a better estimation of the total variance for reliability analysis in this test-retest sample.

Studies making use of binary independent variables (t-tests) For a new study which is comparing two independent groups, one is fitting a binary regressor to the dependent outcome variable. The SD for the ICC is calculated based on the sample variability before fitting the regressor, and thus all individuals are included in the calculation of this value (Figure 1). The total SD is therefore dependent on both the within-group standard deviation, as well as the degree of difference between the two groups, which is measured by the effect size. For independent sample t tests, one can therefore calculate the effect size (Cohen's D) for which the reliability of a measure would reach a certain desired threshold.

Thus, by estimating the within-group standard deviation of each of the two groups, one can calculate the required effect size to obtain a sufficient level of reliability. The total SD of the entire sample (including both groups) is described by the following equation:

$$\sigma_{total} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + n_1\mu_1^2 + (n_2 - 1)\sigma_2^2 + n_2\mu_2^2 - (n_1 + n_2)\mu_{total}^2}{n_1 + n_2 - 1}} \tag{10}$$

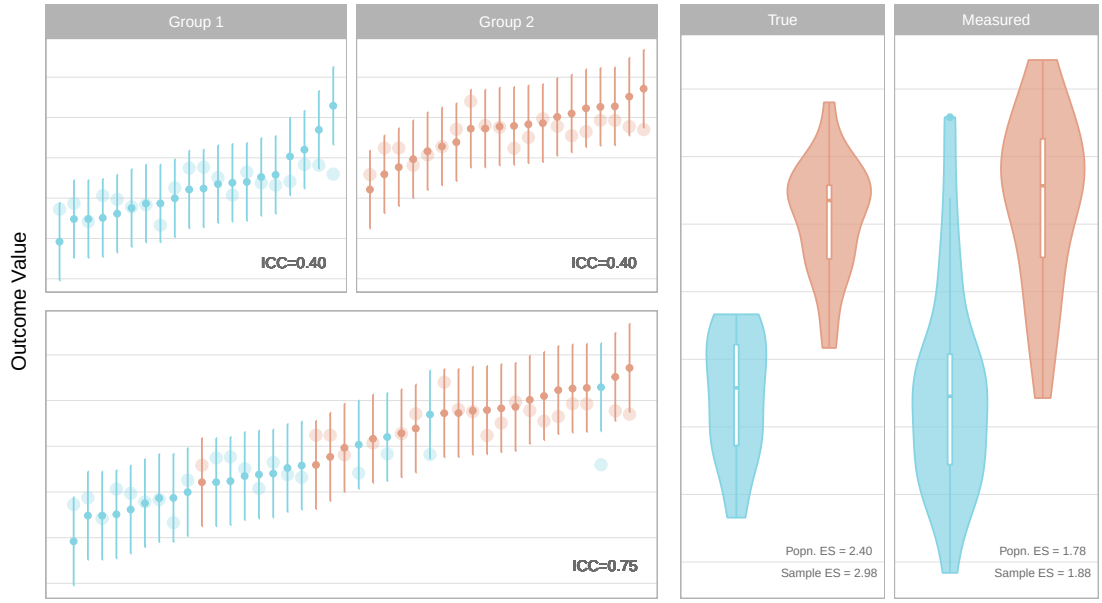

Figure 1. Left: Measured values and their 95% confidence intervals are represented by the small points and error bars. Underlying true values are represented by the larger points. Low reliability within groups is increased for the total sample in a comparison, given a sufficiently large effect size, due to the larger variance of the combined sample. Right: True underlying effect sizes (Cohen's D, left) are attenuated by measurement error (right). The population (Popn.) effect size (ES) and sample ES are of the underlying distributions from which the data are sampled and of the obtained sample respectively.

where n is the number of participants in each group, μ is the mean of each group, σ is the SD of each group, and the subscripts 1, 2 and total refer to group 1, group 2, and the combined sample. The total mean (μ_{total}) is calculated as follows.

$$\mu_{total} = \frac{1}{n_{total}}(n_1\mu_1 + n_2\mu_2) \tag{11}$$

We therefore only need to solve this equation for μ_2, the mean of the second (e.g. patient) group. This can be calculated as follows (separating the equation into three parts):

$$A = \mu_1^2 n_1 n_2 + n_1^2\sigma_1^2 - n_1^2\sigma_{total}^2 + n_1 n_2 \sigma_1^2 + n_1 n_2 \sigma_2^2 - 2n_1 n_2 \sigma_{total}^2 - n_1\sigma_1^2 \tag{12}$$

$$B = -n_1\sigma_2^2 + n_1\sigma_{total}^2 + n_2^2\sigma_2^2 - n_2^2\sigma_{total}^2 - n_2\sigma_1^2 - n_2\sigma_2^2 + n_2\sigma_{total}^2 \tag{13}$$

$$\mu_2 = \frac{1}{2n_1n_2} \times \left(2\mu_1 n_1 n_2 \pm \sqrt{4\mu_1^2 n_1^2 n_2^2 - 4n_1 n_2 (A+B)}\right) \tag{14}$$

In this way, given the group sizes and estimates of the standard deviation within both groups, we can calculate the mean difference and effect size for which the degree of variability in the total sample is sufficient to reach a specified level of reliability.

Software

All analyses in this paper were performed using R (version 3.4.3, *Kite-Eating Tree*) with the *relfeas* package (`https://github.com/mathesong/relfeas`).

RESULTS

The use of these methods will be demonstrated using examples based on two PET tracers for which test-retest studies have been published. PET imaging involves injection of a radioligand (or tracer) into the blood which binds to a target of interest, and the concentration of radioactivity in the tissue, and sometimes also the blood, are measured over time. The two tracers used as examples include [^{11}C]AZ10419369 for imaging of the serotonin 1B receptor (5-HT_{1B}) and [^{11}C]PBR28 for imaging of translocator protein (TSPO). Both of these tracers have been well-validated and have been used to make clinical comparisons. The reason for the selection of these two radioligands is that they both demonstrate properties which make the interpretation of reliability outcomes especially difficult. [^{11}C]AZ10419369 exhibits low measurement error, but also low inter-individual heterogeneity, resulting in low reliability. [^{11}C]PBR28 shows high error variance, but also high inter-individual heterogeneity, resulting in high reliability.

Outcome measures reported for these PET studies include binding potential (BP_{ND}), volume of distribution (V_T), which are defined based on compartmental models, as well as the distribution volume ratio (DVR) and standardised uptake value (SUV). For more details, see Appendix 1.

Example 1: Low reliability and low variance

[^{11}C]AZ10419369 is a radiotracer for the serotonin 1B (5-HT_{1B}) receptor, which is widely expressed in the brain and an important target in depression (Ruf and Bhagwagar 2009). In the test-retest study published with this ligand (Nord, Finnema, et al. 2014), using the frontal cortex as an example, it showed what is considered a high mean BP_{ND} (1.6), a favourable absolute variability (6.8%) and a good coefficient of variation (6-7%). However in the sample measured, the ICC was very low: 0.32. It was concluded that "[the low ICC value] can be explained by … a low between-subject variance. Thus, despite the low ICC values, it cannot be excluded that the test-retest reliability is also high in these regions" (p.304).

A followup study was published examining age-related changes in 5-HT_{1B} binding (Nord, Cselényi, et al. 2014), concluding that 5-HT_{1B} receptor availability decreases with age in cortical regions. If one were to have considered the ICC of the original study as a fixed property of this measure, then such a conclusion should be treated with caution due to the low reliability of the outcome. However, due to the differences in [^{11}C]AZ10419369 BP_{ND} across the age range examined, the standard deviation was 3.2-fold greater than that of the test-retest study. To assess the reliability of frontal cortex [^{11}C]AZ10419369 BP_{ND} for this particular application, it is necessary to take the greater variability into account.

We can calculate the reliability of the new study using the results of the test-retest study by assuming the same measurement error between the studies. In this way we obtain a reliability of this tracer and for this particular correlation equal to 0.93. From this analysis, the outcome of the study can be considered reliable in terms of measurement error, and that individuals can be easily distinguished from one another by their outcome measures. This conclusion also holds after taking into account the use of partial volume effect correction in this study (see Appendix 2).

Example 2: Reliability and power analysis

Let us assume that a new study was being planned to examine the relationship between [^{11}C]AZ10419369 BP_{ND} and a particular self-report measure. This study would be performed in young, healthy control subjects. Having observed from the test-retest study that the reliability in healthy controls was not high for most regions due to limited inter-individual variability, the investigators choose to examine the occipital cortex where reliability is highest (ICC=0.8).

The self-report scale is also associated with its own measurement error: it usually exhibits a reliability of 0.7 in healthy samples (and it is assumed that this will apply to this sample too). Using the equation7, we obtain an attenuation of the correlation coefficient of 0.75. In other words, if the scale were to explain 100% of the variance in [^{11}C]AZ10419369 BP_{ND} in the absence of any measurement error, we would expect to see a correlation of r=0.75 after taking this error into account.

For deciding whether to go ahead and conduct this study, we can consider the *maximum realistic biological* effect size for power analysis (see Methods). The investigators approximate that 30% explained variance is probably the maximum extent to which the self-report measure could realistically be explained by occipital cortex 5-HT_{1B} availability. For this effect, a power analysis (for 80% power) will reveal that such a correlation would require a sample size of 23 in the absence of measurement error. Accounting for the attenuation due to measurement error however, 44 participants would be required to observe this effect.

It can therefore be concluded that unless more than 44 participants can be examined in the planned study, then it will not be possible to draw strong conclusions from the results of this study. If significant, it would likely overestimate the true magnitude of the association as the study would only be sufficiently powered for effect sizes larger than what what was considered realistic a priori, i.e. a type M error (Gelman and Carlin 2014). If insignificant, such a result would be uninformative as the type II error rate would be too large. In this way, a costly study which had a low probability of yielding meaningful conclusions can be avoided before even its inception.

Example 3: Reliability for differences

[^{11}C]AZ10419369 is also known to be displaced by serotonin release. This can be measured by performing a baseline measurement, and then performing another measurement following administration of a serotonin-releasing agent and assessing the decrease in binding (see Appendix 3). We will assume that high doses of the drug similar to those previously administered to non-human primates (NHPs) (Nord et al. 2013) will be administered to humans, and it is estimated that this will produce a similar displacement of the radioligand in humans. The study will consider the occipital cortex, where the reliability of [^{11}C]AZ10419369 BP_{ND} is highest, and where escitalopram-induced displacement was significant in NHPs (mean=-12%, SD=10%) (Nord et al. 2013).

For this study, there are two questions to be answered with regard to its feasibility, namely the within- and between-individual effects respectively. The first is whether we can reliably measure serotonin release within individuals at all. The second question is whether the reliability of this radioligand is sufficiently high to compare the degree of radioligand displacement (i.e. ΔBP_{ND}), as opposed to binding (i.e. BP_{ND}), between individuals and hence groups (patients vs controls).

The first question can be answered using the smallest detectable difference (SDD) metric as well as power analysis. The SDD for individual scores is 14.3% (equation 5). To detect a difference in group means, the SDD is 10.1% for 2 individuals (equation 6). This means that with two or more individuals, a change of 12% is greater than that which would be expected by chance based on the reliability of [^{11}C]AZ10419369 BP_{ND}. We also perform a power analysis by considering the size of the expected effect (-12% change) relative to the standard deviation of test-retest variability (7.3%), i.e. a Cohen's D of 1.64. This means that 4 individuals would be required for 80% power for a one-sided test of this effect size, assuming that it is accurate.

The second question requires that we reassess the reliability of the radioligand for the change in [^{11}C]AZ10419369 BP_{ND}. We must consider that there are two measurements, both of which have error terms: the error (σ_e) term is therefore doubled for ΔBP_{ND} as the measurement error is additive. The total variance term is also no longer determined by the range of BP_{ND} values, but by the range of ΔBP_{ND} values. This can be approximated from the displacement statistics as 10% of the mean BP_{ND} value before treatment. The ICC obtained for ΔBP_{ND} would therefore be -0.06. This is so low due to the low variability in the outcome: there is large error variance, and a small true variance in the ΔBP_{ND} outcome measure. However this does not take the expected differences between the controls and the patients into consideration.

In planning this study, the investigators estimate that there should be 2.5 times as much serotonin release in the patient group. This will result in a larger degree of interindividual differences, and thus higher reliability. With two groups of equal size (20) and standard deviation, with this effect size, the ICC grows to 0.41. This means that the true effect (Cohen's D=1.8) will be attenuated to 0.95 when measured due to the measurement error, and a much larger sample size (308% for unpaired; 179% for paired) will be required to detect a difference of this size as significant with 80% power (with a one-sided test). Further, since differences between individuals in ΔBP_{ND} values will be mostly determined by random noise, even if the study were to produce a significant outcome, it would still be unclear to what extent this was due to noise, and we would remain uncertain about its true magnitude. It can therefore be concluded that this study is probably not worth investing in, as its results will not be greatly informative.

Example 4: High reliability and high variance

[^{11}C]PBR28 is a second generation tracer for translocator protein (TSPO), which is expressed in glial cells including immune cells, and represents an important target for studies of immune activation in various clinical disorders. Due to genetic effects (leading to high- and low-affinity subtypes of the protein) and high intra- and inter-individual variability, several different strategies have been proposed for quantification of this tracer.

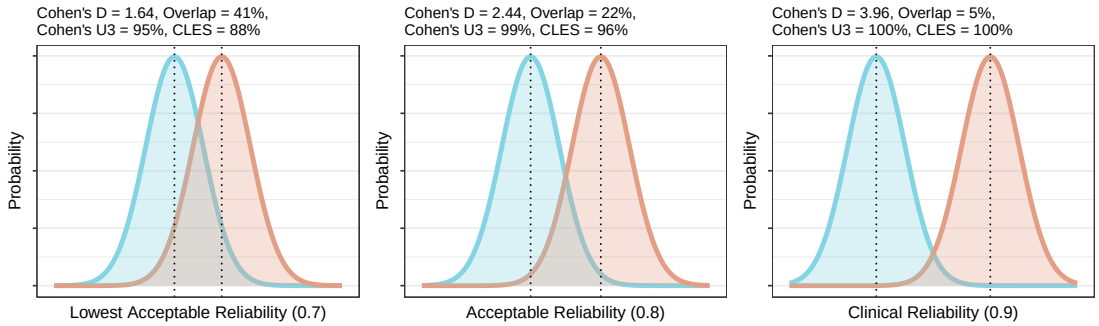

Figure 2. Overlap and effect size required for DVR differences between groups to reach different levels of reliability. The overlap is the distributional overlap. Cohen's U3 is the fraction of one group which does not overlap with the mean of the other. The Common Language Effect Size (CLES) is the chance that a person picked at random from the red group will have a higher value than a person picked at random from the blue group.

While the absolute variance of [^{11}C]PBR28 V_T appears very high ($\approx$ 20%), its test-retest reliability is also very high ($\approx$ 0.9) (Collste et al. 2016; Matheson et al. 2017) in both high-affinity binders (HABs) and medium affinity binders (MABs) (for more information, see Owen et al., 2012) due to the several-fold variability (COV $\approx$ 40%) between individuals (Collste et al. 2016; Matheson et al. 2017). As such, clinical studies comparing patients and controls using V_T are unlikely to suffer from poor reliability. However, a large effect size (Cohen's D = 0.8) corresponds to a mean difference of over 30% in [^{11}C]PBR28 V_T, while an equivalent large effect size for [^{11}C]AZ10419369 frontal cortex BP_{ND} is less than 6%. Put differently, to detect a 10% difference in binding between groups, 278 (HAB) participants would be required per group for [^{11}C]PBR28, compared to 7 per group for [^{11}C]AZ10419369. Large numbers of participants will be required to observe even relatively large proportional differences in [^{11}C]PBR28 binding. However, large proportional differences are also more likely for [^{11}C]PBR28 due to the large differences between individuals.

Example 5: Variance reduction strategies

Attempts have been made to account for this variability in [^{11}C]PBR28 V_T, both within and between genotype groups, with suggestions of simplified ratio-based approaches such as the distribution volume ratio (DVR) despite the absence of a reference region (Lyoo et al. 2015). We previously reported high interregional associations ($R \geq 0.98$), including denominator regions, resulting in poor reliability for this outcome measure (mean ICC=0.5) (Matheson et al. 2017). While we advised caution in the interpretation of results making use of ratio methods, it is still theoretically possible that ratio methods might show good reliability for the detection of very large effects in specific neurological conditions which would not be otherwise detectable with small samples due to the large inter-individual variability (Lyoo et al. 2015). Using this method of extrapolating reliability, we can calculate what size of an effect between groups would be required for the ratio methods to begin to show reasonable reliability for informative conclusions.

For this, we first calculate the required total SD for each of our specified levels of reliability. We then calculate the effect size which would produce this SD. The methods are calculated for a planned comparison of 20 healthy controls and 20 patients with equal standard deviation between groups, assuming that the SD of the test-retest sample is representative of that for the healthy group. For reliabilities of 0.7, 0.8 and 0.9, one would require Cohen's D effect sizes of 1.6, 2.4 and 4 respectively. Considering that a Cohen's D of 0.8 is considered a *large* effect size (Cohen 1988), these effects are unrealistically large. These effect sizes can be plotted as overlapping distributions to provide a visual reference for understanding these effect sizes (Figure 2).

From this information, it is clear that unless one is expecting *very* large effects, the use of DVR is unlikely to be useful due to its poor reliability. Due to the strong associations reported between numerator and denominator regions, DVR effect sizes will even further underestimate actual differences in V_T in addition to the attenuation of effect sizes due to its poor reliability. V_T therefore appears to be statistically

superior to DVR for detection of any effects, without the need to be concerned about issues of reliability.

DISCUSSION

This paper has demonstrated how the results of previously published test-retest studies can be utilised for the rough approximation of the reliability in new samples based only on commonly reported summary statistics. Test-retest studies are usually, and should usually, be performed to validate new PET tracers and quantification strategies. However, while reliability is an important consideration for the design and interpretation of studies, measures of reliability reported in test-retest studies are not currently used to their full potential. This paper outlines a method for how reliability can be extrapolated to new samples, provides examples of how these methods can be applied in practice for both study design and interpretation, and is accompanied by an open-source R package *relfeas* with functions for all calculations.

In this way, researchers can avoid performing costly studies whose results are unlikely to be replicable or sufficiently informative with regard the relevant biological question being asked, and can interpret previously published studies whose reliability might be questionable with an appropriate degree of caution. This method is analogous to the philosophy of "fail fast" from software development (Shore 2004): systems should be designed such that bugs which may lead to system failure will cause the system to fail at an early stage of operation, rather than unexpectedly at a late stage. In the same way, this method allows for studies to fail early, before even reaching the the data collection stage, leading to savings in cost and time, and with PET, avoiding exposure of participants to radioactivity.

The importance and neglect, of reliability has come up before, perhaps most notably in the study of Vul et al. (2009) examining fMRI studies of emotion, personality and social cognition. This study found 'puzzlingly high correlations', in which, given the reliability of the measures, over 100% of the biological variance was being explained. It was concluded that "a disturbingly large, and quite prominent, segment of fMRI research … is using seriously defective research methods … and producing a profusion of numbers that should not be believed." (p. 285). The failing of researchers to sufficiently acknowledge considerations of reliability has been related to the historical separation of correlational (i.e. individual differences research) and experimental (i.e. within-subject) approaches to scientific inquiry: outcome consistency with one approach does not necessarily apply to the other (Hedge, Powell, and Sumner 2017). It was shown that several well-validated cognitive tasks (e.g. Stroop, Go/No-go), despite demonstrating robust within-individual effects, exhibited very low reliability for inter-individual comparisons. This was due primarily to low inter-individual variation. It is noted that it is likely this same characteristic that both makes these measures robust in experimental research, but unreliable for correlational research (Hedge, Powell, and Sumner 2017).

The results from the examples above can be interpreted in this context: [^{11}C]AZ10419369 BP_{ND} appears to be an excellent outcome for the assessment of within-individual changes, such as in PET blocking studies where the receptor is measured before and after pharmacological blockade. However, its utility is limited for the assessment of between-individual differences. The opposite is true of [^{11}C]PBR28 V_T: this measure differentiates between individuals very well, but may be less useful for assessing within-individual changes. [^{11}C]PBR28 DVR may be more useful for within-individual effects, provided that equivalence is shown in the denominator region (Lakens 2017). For between-subjects designs, it is important to emphasise that measures whose precision is low can still be useful when differences between individuals are very large (e.g. [^{11}C]PBR28 V_T), while measures whose precision is very high can be practically uninformative when the differences between individuals are extremely small (e.g. frontal cortex [^{11}C]AZ10419369 BP_{ND}). Since reliability is indirectly proportional to the amount of inter-individual variability in a sample, reliability can be increased by recruiting more diverse participant samples. In contrast, relying on convenience sampling results in study samples which are overwhelmingly comprised of university students from privileged backgrounds (WEIRD) (Henrich, Heine, and Norenzayan 2010), who are unlikely to be representative of the broader population (i.e. showing low external validity).

Failure to replicate is a substantial problem in science (Begley and Ellis 2012; Open Science Collaboration et al. 2015). Questionable research practices (QRPs) (John, Loewenstein, and Prelec 2012; Simmons, Nelson, and Simonsohn 2011) and underpowered studies (Button et al. 2013; Morey and Lakens 2016) have been proposed as some of the primary reasons for this failure. It has even been suggested that this may threaten the very notion of cumulative science (Morey and Lakens 2016; Elk et al. 2015). Consideration of the reliability is another important, and related, factor. Studies based on unreliable outcomes are unlikely to even produce results representative of any underlying reality. Studies

can be determined as not worth being performed if outcomes are unreliable, or if maximal feasible 'true' effects after effect size attenuation is deemed to be too small to measure. This can provide enormous savings in both time and resources. A utopian science would be one in which i) analysis plans were fully pre-registered or, better yet, peer-reviewed prior to study initiation such as in Registered Reports (Nosek, Spies, and Motyl 2012; Chambers et al. 2015), ii) studies were sufficiently powered to provide meaningful effect size estimates and comparison between studies (Morey and Lakens 2016); and iii) all studies utilised outcome measures whose reliability was sufficient for the specific research question in the study sample such that they would yield informative estimates of magnitude. By providing methods such that the reliability in study samples can at least be roughly approximated and planned for, and tools by which to make this easily implementable, this paper hopes to aid researchers in this last goal.

ACKNOWLEDGEMENTS

I would like to say an enormous thank you, in alphabetical order, to Simon Cervenka, Lieke de Boer, Vincent Millischer, Pontus Plavén-Sigray, Björn Schiffler, Jonas Svensson and William Hedley Thompson for their helpful comments and discussions about the manuscript.

REFERENCES

Baumgartner, Richard, Aniket Joshi, Dai Feng, Francesca Zanderigo, and R Todd Ogden. 2018. "Statistical evaluation of test-retest studies in PET brain imaging." *EJNMMI Research* 8 (1). EJNMMI Research: 13. doi:10.1186/s13550-018-0366-8.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483 (7391): 531–33. doi:10.1038/483531a.

Button, Katherine S, John P a Ioannidis, Claire Mokrysz, Brian a Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews. Neuroscience* 14 (5). Nature Publishing Group: 365–76. doi:10.1038/nrn3475.

Carrasco, Josep L., Alejandro Caceres, Georgia Escaramis, and Lluis Jover. 2014. "Distinguishability and agreement with continuous data." *Statistics in Medicine* 33 (1): 117–28. doi:10.1002/sim.5896.

Chambers, Christopher D., Zoltan Dienes, Robert D. McIntosh, Pia Rotshtein, and Klaus Willmes. 2015. "Registered Reports: Realigning incentives in scientific publishing." *Cortex* 66: 1–2. doi:10.1016/j.cortex.2015.03.022.

Cicchetti, Domenic V. 1994. "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological Assessment* 6 (4): 284–90. doi:10.1037/1040-3590.6.4.284.

Cohen, J. 1988. "Statistical power analysis for the behavioral sciences." doi:10.1234/12345678.

Collste, K., A. Forsberg, A. Varrone, N. Amini, S. Aeinehband, I. Yakushev, C. Halldin, L. Farde, and S. Cervenka. 2016. "Test–retest reproducibility of [11C]PBR28 binding to TSPO in healthy control subjects." *European Journal of Nuclear Medicine and Molecular Imaging* 43 (1): 173–83. doi:10.1007/s00259-015-3149-8.

Cremers, Henk R., Tor D. Wager, and Tal Yarkoni. 2017. "The relation between statistical power and inference in fMRI." *PLoS ONE* 12 (11): 1–20. doi:10.1371/journal.pone.0184923.

Elk, Michiel van, Dora Matzke, Quentin F. Gronau, Maime Guan, Joachim Vandekerckhove, and Eric-Jan Wagenmakers. 2015. "Meta-analyses are no substitute for registered replications: a skeptical perspective on religious priming." *Frontiers in Psychology* 6. doi:10.3389/fpsyg.2015.01365.

Ferguson, Christopher J., and Moritz Heene. 2012. "A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null." *Perspectives on Psychological Science* 7 (6): 555–61. doi:10.1177/1745691612459059.

Fleiss, Joseph L. 1986. *Reliability of Measurement*.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–51.

doi:10.1177/1745691614551642.

Harvill, LM. 1991. “Standard error of Measurement.” *Instructional Topics in Educational Measurement* 1991 (1): 33–41. doi:10.1177/0272989X10380925.

Hedge, Craig, Georgina Powell, and Petroc Sumner. 2017. “The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences.” *Behavior Research Methods*, July. doi:10.3758/s13428-017-0935-1.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The weirdest people in the world?” *Behavioral and Brain Sciences* 33 (2-3): 61–83. doi:10.1017/S0140525X0999152X.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. “Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling.” *Psychological Science* 23 (5): 524–32. doi:10.1177/0956797611430953.

Kimball, A. W. 1957. “Errors of the Third Kind in Statistical Consulting.” *Journal of the American Statistical Association* 52 (278): 133–42. doi:10.1080/01621459.1957.10501374.

Lakens, Daniël. 2017. “Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses.” *Social Psychological and Personality Science* 8 (4): 355–62. doi:10.1177/1948550617697177.

Lyoo, C H, M Ikawa, J S Liow, S S Zoghbi, C L Morse, V W Pike, M Fujita, R B Innis, and W C Kreisl. 2015. “Cerebellum Can Serve As a Pseudo-Reference Region in Alzheimer Disease to Detect Neuroinflammation Measured with PET Radioligand Binding to Translocator Protein.” *Journal of Nuclear Medicine* 56 (5): 701–6. doi:10.2967/jnumed.114.146027.

Matheson, Granville J., Pontus Plavén-Sigray, Anton Forsberg, Andrea Varrone, Lars Farde, and Simon Cervenka. 2017. “Assessment of simplified ratio-based approaches for quantification of PET [11C]PBR28 data.” *EJNMMI Research* 7 (1): 58. doi:10.1186/s13550-017-0304-1.

Morey, Richard, and Daniël Lakens. 2016. “Why most of psychology is statistically unfalsifiable.” doi:10.5281/zenodo.838685.

Nord, Magdalena, Zsolt Cselényi, Anton Forsberg, G??ran Rosenqvist, Mikael Tiger, Johan Lundberg, Andrea Varrone, and Lars Farde. 2014. “Distinct regional age effects on [11C]AZ10419369 binding to 5-HT1B receptors in the human brain.” *NeuroImage* 103: 303–8. doi:10.1016/j.neuroimage.2014.09.040.

Nord, Magdalena, Sjoerd J Finnema, Christer Halldin, and Lars Farde. 2013. “Effect of a single dose of escitalopram on serotonin concentration in the non-human and human primate brain.” *The International Journal of Neuropsychopharmacology / Official Scientific Journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)* 16 (7): 1577–86. doi:10.1017/S1461145712001617.

Nord, Magdalena, Sjoerd J. Finnema, Martin Schain, Christer Halldin, and Lars Farde. 2014. “Test-retest reliability of [11C]AZ10419369 binding to 5-HT 1B receptors in human brain.” *European Journal of Nuclear Medicine and Molecular Imaging* 41 (2): 301–7. doi:10.1007/s00259-013-2529-1.

Nosek, Brian A., J. R. Spies, and M. Motyl. 2012. “Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability.” *Perspectives on Psychological Science* 7 (6): 615–31. doi:10.1177/1745691612459058.

Nunnally, Jum C. 1970. “Introduction to psychological measurement.” McGraw-Hill.

———. 1978. *Psychometric Theory*.

Open Science Collaboration, Open Science, C. Hempel, C. Hempel, P. Oppenheim, P. E. Meehl, J. R. Platt, B. A. Nosek, et al. 2015. “Estimating the reproducibility of psychological science.” *Science (New York, N.Y.)* 349 (6251): aac4716. doi:10.1126/science.aac4716.

Owen, David R., Astrid J. Yeo, Roger N. Gunn, Kijoung Song, Graham Wadsworth, Andrew Lewis, Chris Rhodes, et al. 2012. “An 18-kDa Translocator Protein (TSPO) polymorphism explains differences in binding affinity of the PET radioligand PBR28.” *Journal of Cerebral Blood Flow and Metabolism*

32 (1): 1–5. doi:10.1038/jcbfm.2011.147.

Portney, Leslie G, and Mary P Watkins. 2015. *Foundations of clinical research: applications to practice*. FA Davis.

Ruf, B M, and Z Bhagwagar. 2009. "The 5-HT1B receptor: a novel target for the pathophysiology of depression." *Current Drug Targets* 10 (11): 1118–38. doi:10.2174/138945009789735192.

Shore, J. 2004. "Fail fast." *IEEE Software* 21 (5): 21–25. doi:10.1109/MS.2004.1331296.

Shrout, Patrick E., and Joseph L. Fleiss. 1979. "Intraclass correlations: Uses in assessing rater reliability." doi:10.1037/0033-2909.86.2.420.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

Spearman, Charles. 1904. "The proof and measurement of association between two things." *The American Journal of Psychology* 15 (1). JSTOR: 72–101.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition1." *Perspectives on Psychological Science* 4 (3): 274–90. doi:10.1111/j.1745-6924.2009.01125.x.

Weir, Joseph P Jp. 2005. "Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM." *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 19 (1): 231–40. doi:10.1519/15184.1.

APPENDICES

Appendix 1: Outcome Measures

BP_{ND} is a measure of the binding potential: a measure of concentration of the radioligand in the specific (S) compartment of the target tissue relative to sum of free (F) and non-specific (NS) compartments (together defined as the nondisplaceable (ND) compartment). When BP_{ND} is zero, there is no S binding; and when it is 1, there is an equal equal radioligand concentration in the S and ND compartments. V_T is the total distribution volume: a measure of the total uptake in the target tissue (S + NS + F) relative to the concentration of radioligand in blood plasma. It is therefore a measure of the uptake in the target tissue relative to the exposure of that tissue to the radioligand. DVR is the distribution volume ratio, and is the ratio of V_T values from a target (numerator) and reference (denominator) region of the tissue. SUV is the standardised uptake value: a measure of the total uptake in the target tissue relative to the amount of tracer injected and the body mass. This measure is sensitive to differences in tissue exposure, due to peripheral binding for example, which is assumed to scale with body weight. These outcomes range from more to less specific in this order, and different outcomes are used for different circumstances depending on the fulfilment of certain assumptions underlying them, as well as whether arterial plasma radioactivity concentrations were recorded. More information can be found in Innis et al (2007).

Innis, R. B., Cunningham, V. J., Delforge, J., Fujita, M., Gjedde, A., Gunn, R. N., … & Iida, H. (2007). Consensus nomenclature for in vivo imaging of reversibly binding radioligands. *Journal of Cerebral Blood Flow & Metabolism*, 27(9), 1533-1539.

Appendix 2: Example 1 Caveat

There is one caveat: the correlational study of age was performed using partial volume effect (PVE) correction. This procedure is used to minimise the bias due to differences in brain volumes and usually produces higher BP_{ND} estimates. However it can also introduce more noise to the data (i.e. a higher SEM). Without access to the original test-retest data, this increase in SEM (ρ) would have to be approximated by assuming that the true between-subject variance is similar between the test-retest sample and the subset of participants from the correlational study of similar age to the test-retest sample (i.e. $\rho \approx COV_{NewStudySubset}/COV_{TRT}$). Another possibility would be to consider the reliability for possible values of ρ. In this way, it can be shown that even in the extreme case that the SEM were expected to be doubled following PVE correction, the extrapolated reliability would remain sufficiently high at 0.73.

Appendix 3: Example 3 Caveat

Radioligand displacement by endogenous or exogenous substances is common in PET studies, resulting in a decrease in binding. However, it is worth noting that clinical doses of SSRIs have been found to produce a paradoxical increase in $[^{11}C]$AZ10419369 BP_{ND} in humans (Nord et al. 2013) for reasons beyond the scope of the current paper. Large doses of SSRIs have been found to produce decreases in NHPs though (Nord et al. 2013). For convenience, we will assume that high doses similar to those administered to NHPs will be administered, and assume that it will produce the same change in humans.