Suspiciously high correlations in brain imaging research

Edward Vul [evul@ucsd.edu]; Harold Pashler [hpashler@ucsd.edu]; UCSD Psychology

In early 2005 a speaker visiting our department reported that blood oxygen-level dependent (BOLD) activity in a small region of the brain accounted for the great majority of the variance in speed with which subjects walk out of the experiment several hours later (this finding was never published, as far as we know). This result struck us as puzzling, to say the least. It made us wonder about various apparent implications, for example: Are walking speeds really so reliable that most of their variability can be predicted? Does a single, localized cortical region chiefly determine walking speeds? Are walking speeds largely predetermined hours in advance? These implications all struck us as far-fetched. Browsing the literature, we were surprised to find that similarly high correlations between fMRI data and individual differences in social behavior and other individual differences were not at all uncommon in the literature. And yet, when we tried to estimate the maximum plausible population correlation between fMRI measures and social behavior based on psychometric considerations, it seemed that the upper bound should be around 0.75 (given a generous estimate of the reliability of the behavioral and brain measures). And yet, correlations exceeding this upper bound were very common (Figure 1). Our efforts to figure out what was amiss to yield such high correlations led to a 2009 article--initially titled "Voodoo Correlations in Social Neuroscience"--which generated far more interest and controversy than we had remotely anticipated.

The source of these suspicious correlations turned out to be a fairly simple selection bias, namely, a selective analysis procedure that effectively reports the highest observed sample correlations chosen out of a very large set of candidates. The problem with such circular

Cite as: Vul & Pashler (in press) Suspiciously high correlations in brain imaging research. In Lilienfeld, S.O., & Waldman, I.D. (Eds). Psychological science under scrutiny: Recent challenges and proposed solutions. New York: Wiley.

selective analyses – namely, that they provide grossly inflated estimates of effect size -- were described as far back as 1950 by Edward Cureton in the context of practitioners using the same data to develop and validate psychological tests. Cureton too had been led to a cheeky label for the procedures he was criticizing, titling his paper "Validity, Reliability, and Baloney." By 2008, this analysis error had become prevalent not only in neuroimaging studies of individual differences in emotion, personality, and social cognition (Vul et al. 2009a), but seemed to occur in many other guises (Vul & Kanwisher, 2010), and was estimated to have played some role in 40-50% of high profile fMRI papers regardless if their substantive focus (Kriegeskorte et al. 2009).

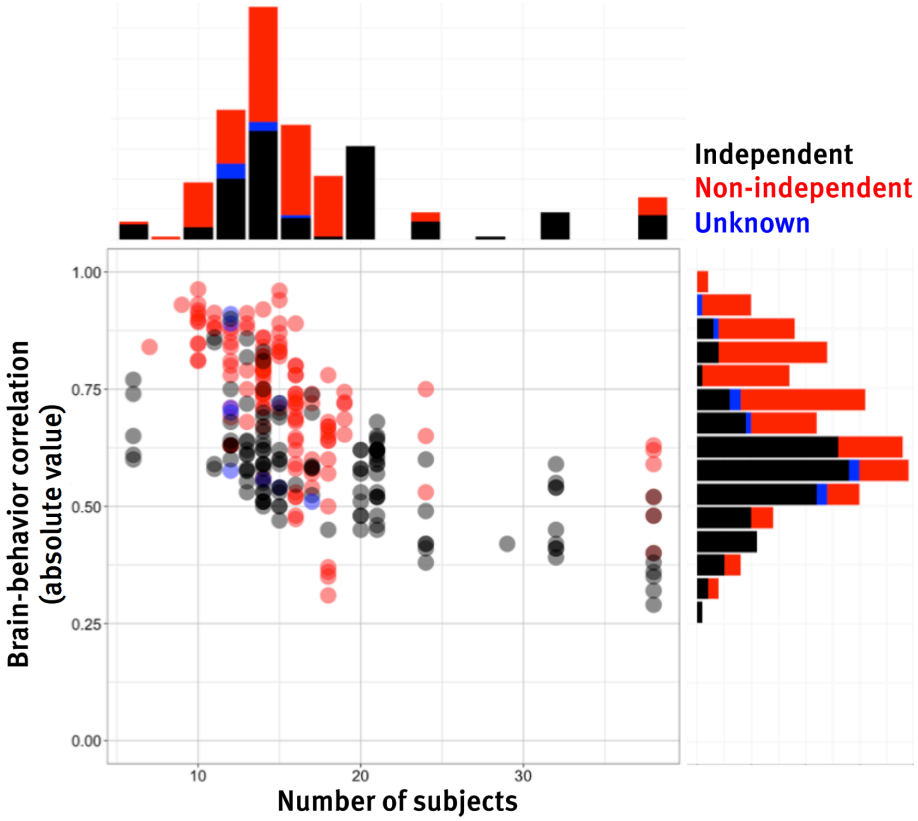

Figure 1. The set of correlations surveyed by Vul et al. (2009), showing how the absolute correlation (y) varies with sample size of the study (x), along with the marginal histograms of both sample size and absolute correlation. Individual observations are color-coded by whether a request for information from the authors revealed the analysis to be independent (black), non-independent (red), or if no response was obtained

(blue). The vast majority of the surprisingly high correlations (r>0.7) were obtained by a non-independent analysis procedure that is guaranteed to inflate effect size, especially when sample sizes are small.

The fallout of the papers describing the prevalence and gravity of the non-independence / circularity error in fMRI unearthed a great many more surprisingly common errors in fMRI analysis and interpretation. We are pleased to say that there has been general endorsement by statisticians, as well as by fMRI practitioners and consumers, of our suggestions for how to avoid committing such errors, misrepresenting results, and misinterpreting the data (Kriegeskorte et al. 2010). Indeed, the most common variants of these errors seem much reduced in prevalence. However, these errors arose in the context of the great difficulty of whole-brain, across-subject fMRI, and as this difficulty is not easily addressed, such problems will continue to arise in various guises.

1. Challenges of fMRI analysis

fMRI analysis is hard, and whole-brain across-subject fMRI is harder. All fMRI is hard because the signal-to-noise ratio for a particular task-contrast in a single subject tends to be quite low (e.g., Kong et al. 2006). This problem is exacerbated by the fact that the noise has non-homogenous magnitude and complicated correlational structure over space and time due to the underlying physiology and physics of measurement (Lund et al. 2006). Whole-brain fMRI is especially hard because the signal is presumed to be carried by a small subset of the thousands of voxels that are measured for a given subject; thus the analysis aims to not only find a small signal bobbing up and down in a maelstrom of structured noise, but also to characterize its properties. Across-subject fMRI is harder still because there are large individual differences across subjects in basic neuroanatomy, as well as in the mapping of task-relevant signals to that varying neuroanatomy (Fedorenko & Kanwisher, 2009); thus, the analysis must find which locations in the 3D grid of measurements of one subject correspond to functionally

matched locations in every other subject. Moreover, across-subject fMRI analysis must grapple with variation in signal across subjects, under severe limitations to the number of subjects that can be included in an experiment given the high cost of using a scanner. In short, whole-brain, across-subject fMRI experiments face a great many statistical and practical challenges that have rendered single whole-brain across-subject fMRI studies in some way a microcosm of the larger replication crisis in psychology.

A single whole-brain, across-subject fMRI experiment is a microcosm of the larger replication crisis in psychology because a single massively multivariate analysis is analogous to a whole field carrying out many experiments. Publication bias (Rosenthal, 1979) inflates the effect sizes in a given field by filtering many executed studies to get just those that passed a significance threshold. A non-independent analysis in fMRI operates the same way: Effect sizes are inflated by filtering many candidate voxels for those that yielded a significant signal. Similarly, given the generally very low power of fMRI studies (Yarkoni, 2009; Button et al, 2013), the set of significant findings is likely to include many false positives, as is the case when considering the set of published findings in a whole field (Ioannides, 2005). The complexity and variability of fMRI analysis offers researchers many choices during the analysis pipeline (Carp, 2012). Insofar as these choices are made in light of the data, the results may be "p-hacked" (Simmons et al. 2011) to a large degree simply by virtue of data-driven selection of analysis procedures (i.e., the "garden-of-forking paths"; Gelman, 2014). Finally, given the expense of fMRI experiments, there are few direct replications, so “conceptual” replication is instead the norm, potentially worsening publication bias for reasons described by Pashler & Harris (2012). More importantly, even a replication of a specific task-contrast in a particular brain region is necessarily only "conceptual" because exact replications at specific sets of

coordinates cannot reasonably be expected. Also, due to a lack of general methods in spatial statistics to characterize the location and uncertainty of an activated region, there is no way to formally determine whether two locations are "the same", thus whether two activations are replications of one another is often a subjective judgment by the researchers. Thus, many of the common problems underlying the replicability crisis across the set of findings in a whole field arise within the context of any given whole-brain across-subject fMRI experiment.

2. Non-independence / circularity

(For a brief introduction to fMRI analysis see the Appendix.)

The central challenge of whole-brain fMRI parallels the challenge faced by genetics and other domains in which the candidate pool of variables far exceeds the number of independent measurements. fMRI research confronts the twin challenge of both finding which of those measured variables actually carries task relevant signals, and characterizing that signal. The non-independence, or circularity, error arises when researchers try to achieve both of these goals using the same set of data.

The prototypical non-independent correlation we uncovered in 2009 was obtained as follows. A particular task contrast (say, BOLD response to happy faces vs sad faces) is calculated in each voxel for each subject. Each voxel's task contrast is correlated across subjects with some other measure on that subject (say, scores on a standard behavioral depression measure). This analysis yields one correlation per voxel in the analysis – on the order of thousands of correlations. A subset of the thousands of correlations is selected based on a search for the

cluster of voxels showing the greatest response[1]. The average, or peak, correlation from this cluster is reported as the effect size of the correlation.

This procedure is guaranteed to overestimate the size of the across-subject correlation in the population. The magnitude of this overestimation will depend on the sample size (larger samples yield less overestimation), the true underlying population correlation (larger true correlations are less overestimated), and the stringency of the statistical threshold used to select voxels (ironically, more stringent multiple comparisons correction yields greater overestimation).

[1] Typically, a cluster-size corrected statistical test is used, identifying groups of adjacent voxels that all have high enough sample correlations to pass a particular significance threshold, and are plentiful enough to pass a cluster-size threshold.

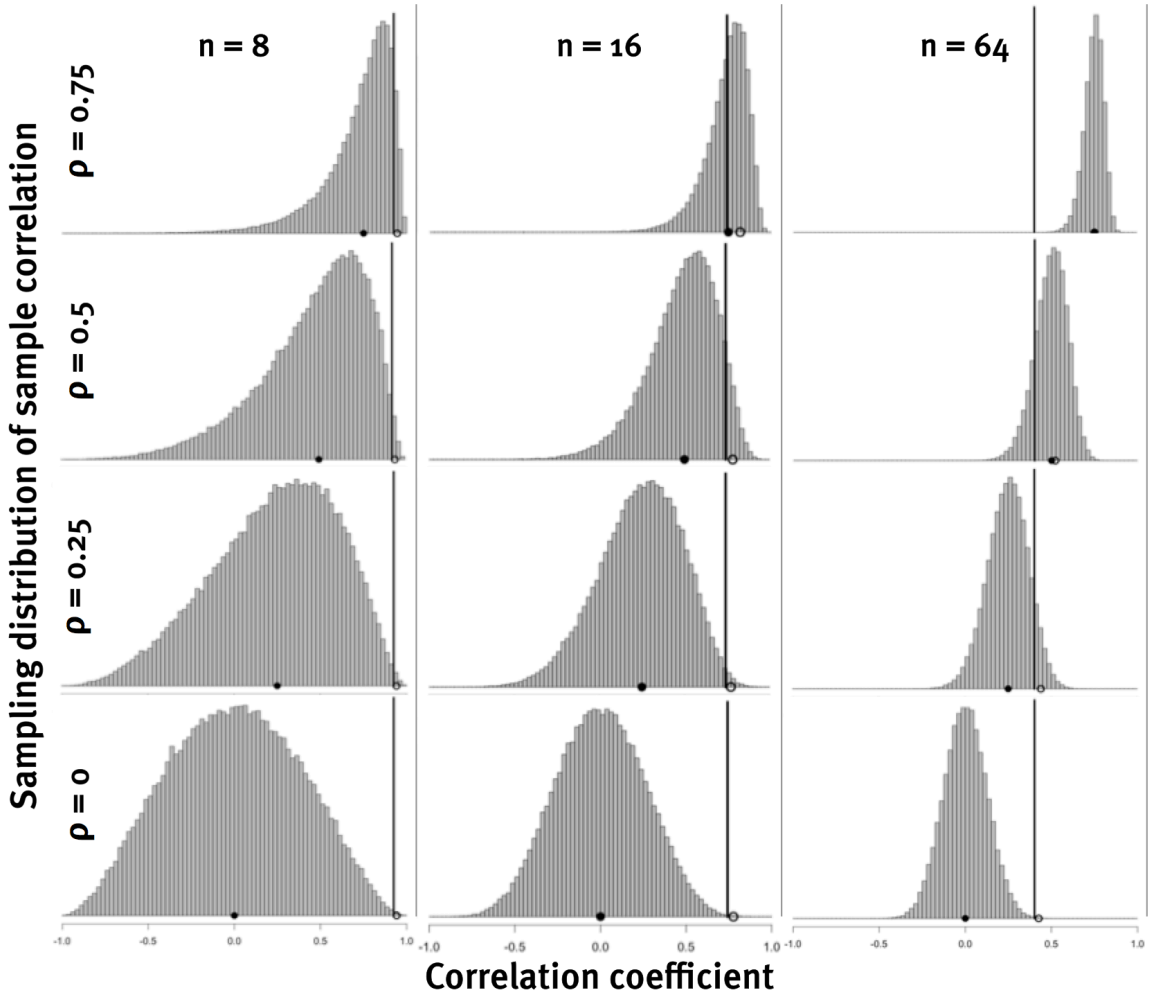

Figure 2. Illustration of the non-independence error. The sampling distribution of the sample correlation varies with the sample size (columns) and the true underlying population correlation (rows; illustrated as a solid black dot on each histogram). A statistical significance threshold (here we use the common cluster-height threshold of $p<0.001$), however, yields a constant critical correlation value for every sample size (black lines). The average sample correlation that passes the significance threshold (open circles) are much higher than their true population correlations unless the statistical power of the threshold is high (meaning that most of the sampling distribution is larger than the threshold, as in the case of $n=64$, $r=0.75$. Consequently, the selected sample correlations are very likely to be much higher than the true populations correlations.

To obtain an intuition for this bias, consider the sampling distribution of the sample correlation (Figure 2). A sample correlation will differ from the true correlation due to sampling variability and noise in the measurements, with smaller sample sizes yielding more variable sample correlations. The statistical threshold used for selection imposes a minimum sample correlation: To be significant at a particular $p<\alpha$ threshold, the sample correlation must exceed some value.

Consequently, the smallest sample correlation that would pass a given significance threshold is constant, regardless of the true correlation. This means that if the statistical power is low (meaning that a small fraction of the sample distribution is above the threshold), the average sample correlation that passes a given significance threshold is uniformly high, regardless of the underlying population correlation.

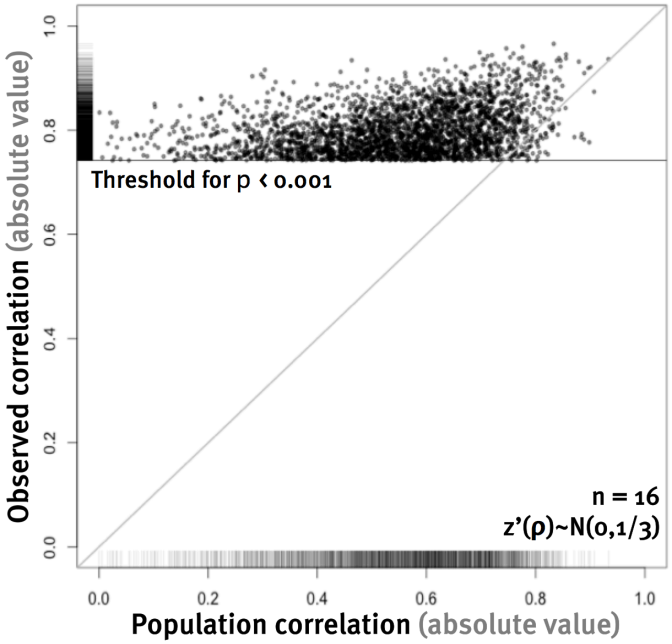

Figure 3. The mechanism of bias in non-independent analyses. Even in the presence of non-zero true correlations (x axis), the sample correlations (y) selected as exceeding a particular threshold are systematically overestimated. With a sample size of 16, the minimum sample correlation to pass a common p<0.001 whole-brain threshold is quite large, ensuring that all observed correlations will be large, even if their true population correlations are small.

This statistical thresholding of sample correlations means that the set of estimated correlation coefficients from such a circular analysis will systematically overestimate the population correlations in those voxels. Figure 3 shows the true underlying correlations as well as the observed correlations in a set of simulated voxels that passed a $p<0.001$ significance threshold. Nearly all observed correlations in this case are higher than their true underlying correlations.

In some non-independent whole-brain correlation studies, instead of reporting the average correlation of a detected cluster, the investigators instead report the "peak voxel" from that cluster. In this case, the exaggeration is even worse. Figure 4 shows the expected maximum correlation identified from different set sizes for different underlying population correlations. If the whole-brain across-subject correlation analysis with 16 subjects considers 1000 possible correlations (considerably less than the number of voxels in a whole-brain analysis), the peak correlation coefficient is expected to be about 0.75, even if the true correlation is actually 0.

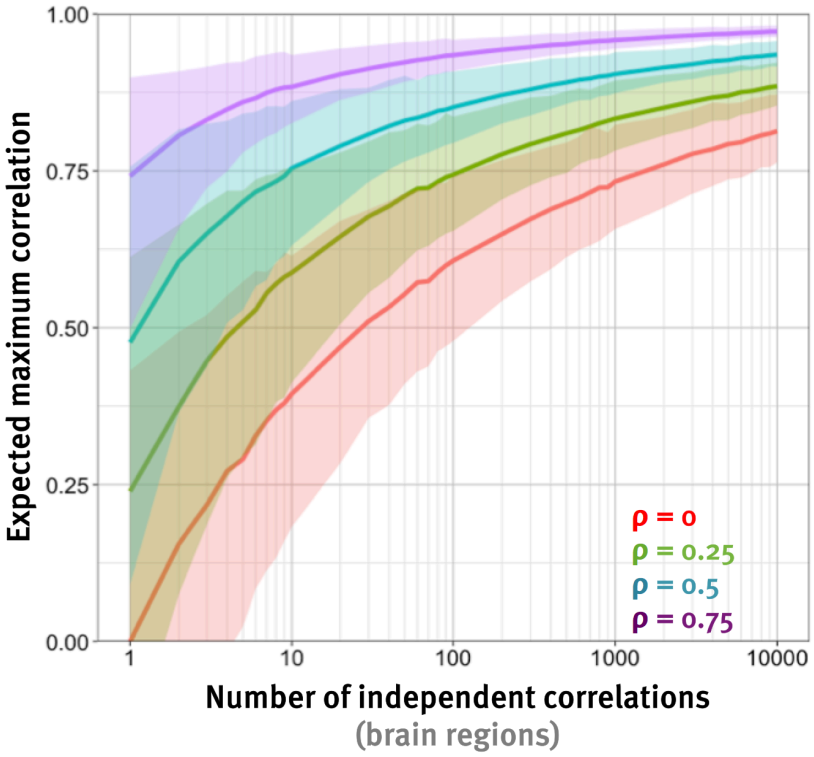

Figure 4. What is the expected value of the peak correlation reported from an analysis? The expected maximum correlation (y) increases with the number of independent brain regions it is chosen from (x), yielding large overestimates of the true correlation, regardless of its value (colors). Lines reflect the expectation, while shaded regions show the 90% interval. These calculations used a sample size of 16 subjects.

Furthermore, multiple comparisons correction, which is designed to reduce the rate of false positives in the statistical test by imposing a more stringent statistical threshold for significance, will only increase the overestimation bias (Figure 5). Although it may at first seem

counterintuitive that more stringent correction for multiple comparisons yields greater overestimation in a circular analysis, it will make sense when considering Figure 2. Greater correction for multiple comparisons increases the correlation threshold: with 16 subjects, $p<0.001$ (a correction for only 50 voxels) requires a correlation of 0.74, while $p<0.0001$ (a correction for 500 voxels) requires a correlation of 0.82. Thus, greater correction for multiple comparisons increases the minimum sample correlation needed to pass the statistical threshold, thereby exacerbating the circular overestimation bias. Of course, the reader should not interpret this point as advocacy for inadequate multiple comparisons correction. Our intent is to illustrate that multiple comparisons correction, although necessary to ensure that the signals are not merely noise, also increases the overestimation bias in non-independent analyses. Indeed, the worst combination is a non-independent analysis combined with inadequate multiple comparisons correction: Such a procedure will typically produce high correlations out of pure noise! (Vul et al., 2009b).

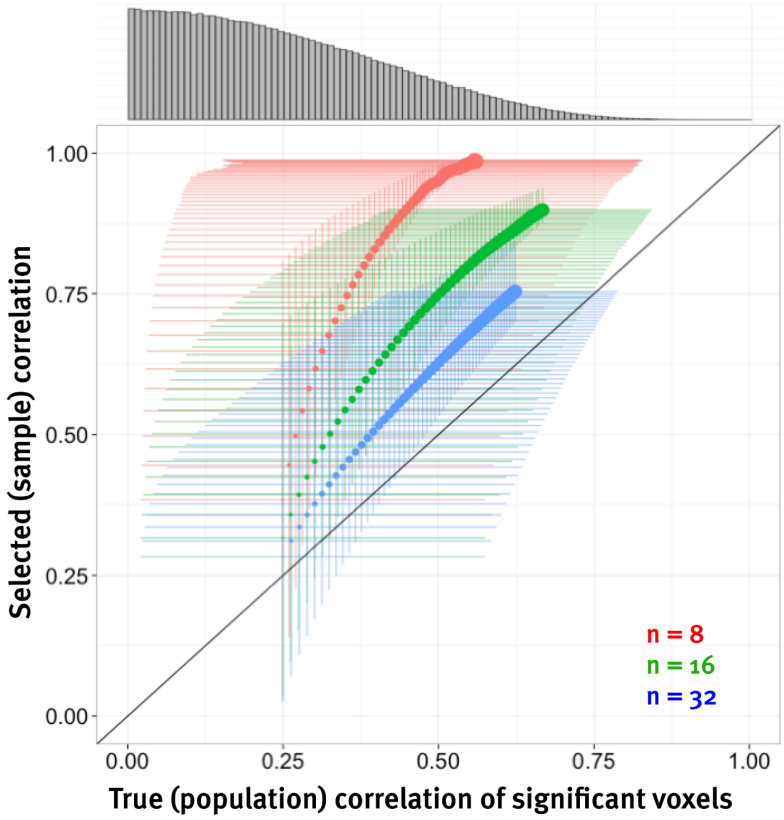

Figure 5. How does multiple comparisons correction influence the bias from non-independent analyses? We simulated how the absolute selected sample correlation (y) relates to the absolute true underlying correlation (x) for different numbers of subjects (8, 16, and 32), as we varied the statistical threshold between ($p<10^{0}$, to $p<10^{-5}$; with larger circles indicating more stringent thresholds). For each threshold we show both the average, and the 90% interval of selected and true correlations. Bias (discrepancy between selected and true correlation – y-distance above the diagonal identity line) is smaller under larger sample sizes, but increases systematically as the statistical threshold becomes more conservative. (The distribution of population correlations is pictured above in gray; this distribution captures the common assumption that there are many small correlations, and few large ones in the brain; formally, this is obtained via a truncated normal distribution with a mean of 0 and a standard deviation of 1/3 on the Fisher z' transforms of the population correlations).

As one might surmise from Figure 2, the magnitude of overestimation depends on the underlying population correlation and the sample size -- effectively the power of the statistical threshold used to select voxels. Figure 6 shows how the non-independent correlation estimate changes as a function of power. When power is high (>0.8) estimated correlations and coefficients of determination are within 10% of the true population values, so there is nearly zero bias from selective analyses. When power is low (0.2-0.4) correlations are misestimated by 14% to 50% (roughly comparable to the 25%-53% overestimation reported by Poldrack et

al., 2010); this amounts to estimating a correlation of 0.4 to be 0.6, and believing that more than twice as much variance can be accounted for than is actually the case. Of course, the most drastic overestimation happens when power is very low (e.g., below 0.2): in those cases researchers might find that they can account for nearly all the variance, when in reality the coefficient of determination (the amount of variance accounted for) is just a few percent. So overestimation is catastrophic with power below 0.2, non-existent when power is larger than 0.8, and considerable -- but perhaps tolerable -- with power as low as 0.5. What kind of power do typical across-subject correlation studies achieve? We explore this issue in a later section, but as a teaser, consider that to achieve power of 50% when the population correlation is a respectable 0.5, an appropriately corrected whole-brain correlation experiment with just 100 independent voxels (far fewer than a real whole-brain analysis) would need to consist of 44 subjects -- more than was used in any of the studies in our surveyed sample of suspicious correlations.

Thus, we see the confluence of factors required to produce a grossly overestimated correlation. An analysis must consider many possible correlation measures (e.g., a correlation per voxel in a whole-brain analysis), choose a subset based on a criterion of them being sufficiently high (e.g., passing a statistical threshold), and the statistical threshold used to select those correlations must have low power with respect to the true underlying correlation. Ironically, because multiple comparisons correction decreases power, it exacerbates the overestimation.

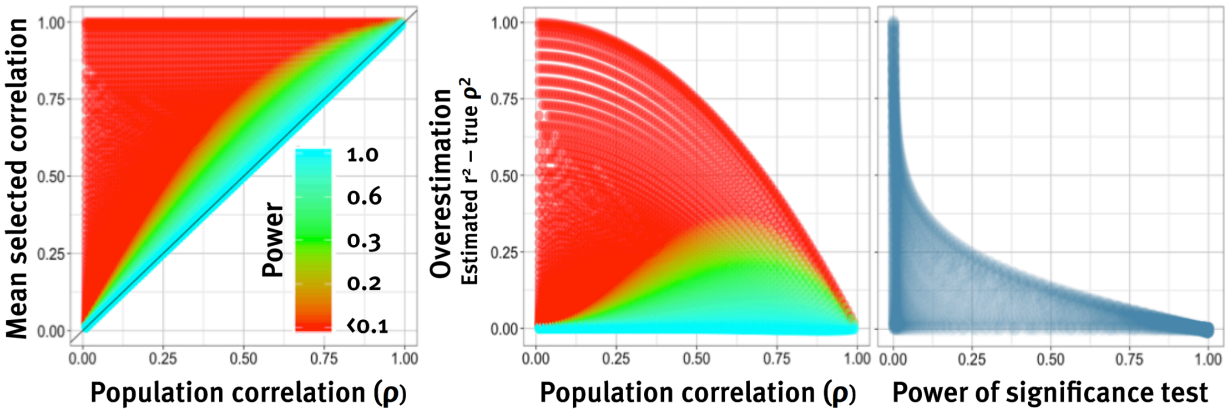

Figure 6. The influence of statistical power on overestimation from non-independent analyses. (Left) Average selected correlation (x) under different true population correlations (y); each point represents a particular sample size, with the color corresponding to the statistical power of a $p<0.001$ threshold with that sample size and true population correlation. Although the relationship is not numerically uniform across population correlations, in all cases less power means greater overestimation. (Middle) Magnitude of overestimation of the coefficient of determination (r^2): the difference between the selected sample r^2 and the population ρ^2 decreases with the power of the test. (Right) Collapsing over true population correlations, statistical power (x) seems to impose an upper bound on the magnitude of overestimation such that the maximum observed overestimate decreases as power increases.

2.1 Avoiding the non-independence error

It is critical to note that not all analyses that yield correlation coefficients or other effect size estimates from fMRI data arise from circular analyses. Indeed, it is quite simple to avoid the non-independence error; that is, to avoid estimating an effect size from a sample selected from a large set because it had a large effect size. There are three classes of strategies to avoid non-independence.

Perhaps the best tactic is to avoid whole-brain across-subject analyses altogether. This is achieved by collapsing the massively multivariate brain measurements of each subject into one, or just a few, summary statistics of the signal for that subject. This is often accomplished by defining within-subject regions of interest, and estimating the task activation therein; thus the across-subject analysis is carried out on just the aggregate signal within a particular region, and thereby avoids the complexities of analyzing thousands of candidate measurements. Such a strategy has a further advantage of allowing within-subject designs, such as looking for

correlations across trials, rather than correlations across individuals (thus gaining considerable power by avoiding across-subject variability).

Another strategy for avoiding non-independence is to use an independent source of data to select across-subject regions of interest, and limit the critical across-subject correlation analysis only to the average signal in those regions (again, avoiding a whole-brain search for significant correlations). This can be achieved by using purely anatomical definitions of regions (e.g., the anatomically-defined amygdala), or using independent across-subject contrasts to define those regions (e.g., finding a region that responds more to happy than sad faces across subjects). After that, one would aggregate some signal within that region for each subject, and estimate its correlation across subjects. Again, the critical part here is that the across-subject analysis of interest is not carried out on every voxel.

The last strategy that one might adopt is cross-validation: using one set of subjects to find clusters in a whole-brain analysis that are correlated with some measure of interest, and then using a *different* set of subjects to estimate the strength of the correlation in that identified cluster. Critically, this strategy again avoids using the same data to identify regions in a voxel-by-voxel whole-brain analysis, but in contrast to the first two strategies, it also identifies voxels based on the signal of interest. Although this approach is quite appealing (and indeed, we advocated it in Vul et al., 2009), it may not be generally advisable given the low power of most whole-brain across-subject correlations, as the consequence would be accurately estimating the magnitude of very few strongest correlations, while missing the vast majority of others.

3. Associated problems that emerged in the fallout

Our paper spurred enthusiasm for critical analysis of fMRI methods, both in the many direct commentaries that were published alongside it, and additional papers that emerged thereafter. Many of these papers uncovered other prevalent problems, some of which interact in vicious ways with the non-independence error.

3.1 Inadequate multiple comparisons correction

Because there are thousands of voxels in whole-brain fMRI, identifying which voxels carry scientifically relevant signals introduces a massive multiple comparisons problem. This problem is not often adequately addressed, with many papers reporting uncorrected whole-brain analyses thresholded at $p<0.001$. Figure 7 shows the expected probability of falsely detecting a significant signal in a whole-brain analysis thresholded at $p<0.001$: if a whole-brain analysis is carried out on ~1000 independent voxels, this uncorrected procedure will yield a significant false positive more than 60% of the time.

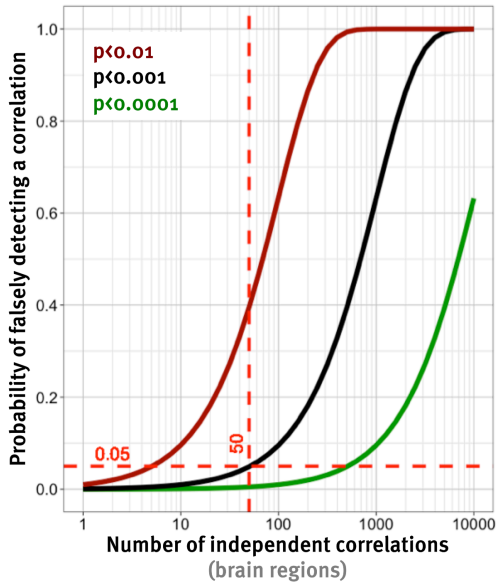

Figure 7. The importance of adequate multiple comparisons correction. As the number of independent brain regions in a whole-brain analysis increases (x), the probability of falsely detecting a correlation (or any other signal) increases if the statistical threshold is held constant. The common p<0.001 threshold is sufficient to

correct for 50 multiple comparisons to the α=0.05 level, but will yield more than 60% false positives if there are 1000 voxels in the whole-brain analysis.

If we consider not just the magnitudes, but the significance, of the correlations in our original sample, we find that although these correlations are surprisingly high, they are not highly significant, given the sample sizes with which they were observed (Figure 8). Such *p* values are to be expected of independent correlations (that need not correct for whole-brain multiple comparisons of the correlation), but are potentially worrisome for those correlations that were estimated from whole-brain analyses, as they may not have been adequately corrected.

Many of the whole-brain correlations were corrected using "cluster-size correction." Instead of correcting for multiple comparisons via classical Bonferroni (or more contemporary false discovery rate [FDR] procedures; Benjamin's & Hochberg, 2001) correction of individual voxels, fMRI analysis often aims to exploit presumed spatial structure of the signals to increase power. In cluster-size correction, a contiguous group of potentially signal-carrying voxels (a cluster) is deemed to be significant by determining what combination of cluster size and measured signal strength is unlikely to arise by chance. Thus, one might achieve an adequate level of correction in a particular 128x128x1 voxel image (per-voxel $p<0.000001$) by jointly thresholding signal strength at $p<0.005$ and cluster size at $k>10$; which should yield greater power than thresholding each voxel at $p<0.000001$ (Forman et al., 1995). The null hypothesis distribution of size-strength cluster combinations is analytically intractable, so determining an adequate correction requires Monte Carlo simulations. In practice, however, many researchers seem to forego that analysis in favor of some "standard" cluster-size correction.

Adding to the problem, many of the "standard" correction thresholds seem to offer inadequate multiple comparisons correction. In a now-classic article, Bennett et al. (2010) showed that a common "standard" cluster-size correction ($p<=0.001$, $k>=10$) can detect task-related activity in a dead salmon. Needless to say, such activity is entirely spurious. Moreover, in a separate paper they found about 30% of fMRI papers published in 2008 used this inadequate correction threshold (Bennett et al. 2009). It is exceedingly unlikely that all of those papers arrived at the same threshold in their particular setting using Monte Carlo simulations. More likely, they simply applied what they thought was a “standard” size-strength threshold indiscriminately across settings. This "standard" correction generally seems to have been borrowed from Forman et al. (1995), who showed that such thresholds are adequate for a 2D 128x128 slice with particular spatial smoothing. However, these thresholds yield much higher false positive rates when applied to 3D volumes. The reason is that each voxel has more neighbors in three dimensions than two, thus yielding greater rates of random contiguity (i.e., false positives appearing in adjacent voxels). The threshold also likely underestimates the impact of smoothing, which induces a greater correlation between adjacent voxels, again increasing the rate of random contiguity. Consequently, this correction threshold is usually (but not always) far too liberal when applied to whole-brain fMRI signals, and yields considerably higher rates of false positives than researchers report (Bennett, 2009).

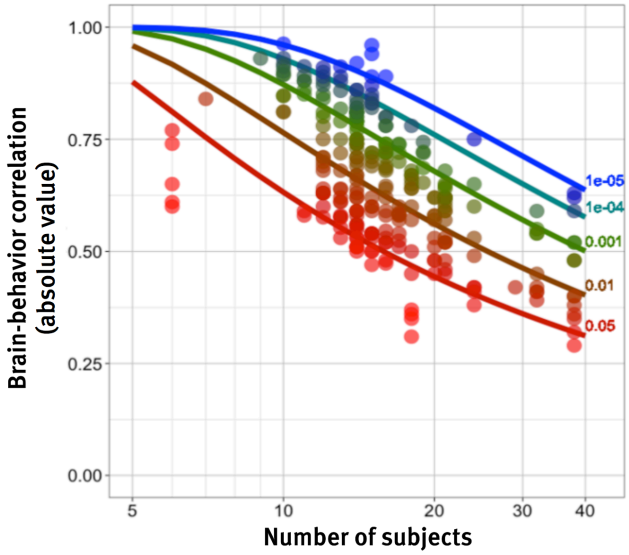

Figure 8. The correlations surveyed in Vul et al (2009), plotted as a function of the number of subjects, and the (absolute) reported correlation. Color corresponds to the (uncorrected) p value of the correlation, and lines indicate the critical correlation values at different α levels. While the reported correlations are large, they are not very significant, especially when considering that many of them arose from whole-brain analyses that would require multiple comparisons correction.

The common practices of arbitrarily choosing among different correction procedures (e.g., family-wise error, false-discovery rate, cluster-size correction), adopting inappropriate "standard" thresholds, and arbitrarily adjusting free parameters (e.g., trading off signal and size thresholds) offer many "researcher degrees of freedom" (Simmons et al., 2011) that make many reported *p*-values uninterpretable. Thus, although it is impossible to assess which of the whole-brain correlations in our sample were adequately corrected, and which adopted an inadequate heuristic procedure, it is likely that at least some of those thresholding at $p<0.005$ or $p<0.001$ and $k>10$ were doing so inappropriately.

Multiple comparisons correction interacts in two vicious ways with non-independent analyses. First, as we showed in the previous section, more stringent multiple comparisons correction during a circular analysis actually exacerbates the effect size overestimation. Second, non-independent analyses with inadequate multiple comparisons correction can produce large,

compelling effects out of pure noise because multiple comparisons correction need not be very stringent to produce a grossly biased effect size estimate (as shown in Figure 2: $p<0.001$ is more than sufficient given a small sample size). A threshold of $p<0.001$ means that the smallest correlation deemed significant with a sample size of 10 would be 0.87, which is very high indeed; however, if $p<0.001$ and its associated cluster size threshold are inadequate to correct for the whole-brain analysis, then a non-independent analysis can produce such a high $r=0.87$ correlation quite reliably, even if the population correlation is 0. Together, there is no way to avoid the problems associated with circular analyses: stringent multiple comparisons correction will exacerbate the bias of effect size overestimation, whereas inadequate multiple comparisons correction is likely to produce impressively large effect sizes from pure noise.

3.2 Low power

After our initial paper, many pointed out that low power is not only a major problem underlying the suspiciously high correlations we reported (Yarkoni, 2009; Yarkoni & Braver, 2010), but is more generally prevalent in neuroscience (Button et al, 2013). As we showed in Figure 6, statistical power is critical to the magnitude of the bias introduced by non-independent analyses. With high power, the bias is virtually zero, whereas with very low power the bias may account for nearly the entire observed effect size. How much power did the correlational studies we surveyed have for a whole-brain analysis? And how large of a sample would be necessary to detect a plausible correlation in a whole-brain across-subject correlation study?

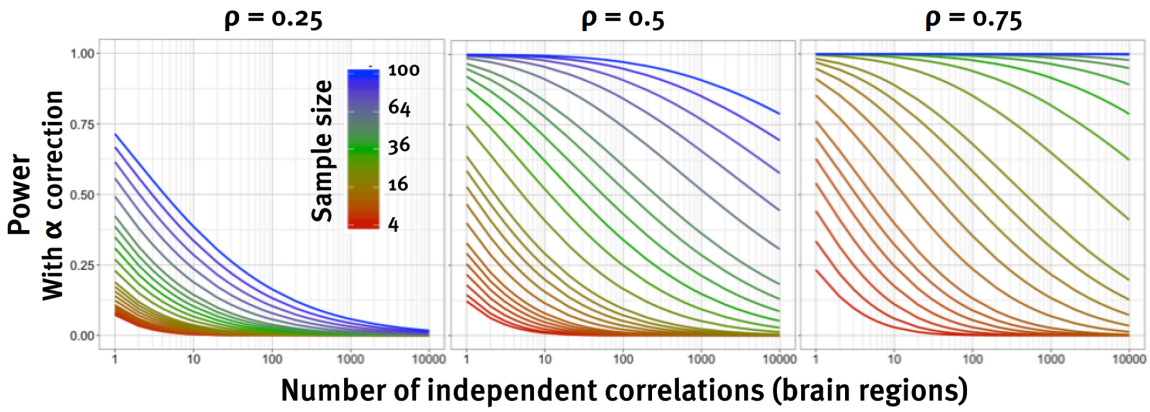

Figure 9. Statistical power (y) for Bonferroni corrected correlation tests as a function of population correlation (panels), sample size (lines), and the number of independent correlations in the analysis (x). A small population correlation (ρ=0.25; left) yields low power even with few independent correlations. In contrast large correlations (ρ=0.75; right) can be tested with high power with just 16 subjects, provided that the analysis considers only one correlation; however, a whole-brain analysis with 1000 correlations requires twice as many subjects to achieve the same level of power. A test for an optimistic, but plausible population correlation (ρ=0.5; middle) requires nearly 100 subjects to achieve a high level of power in a whole-brain analysis.

To assess power, we need to assume some population effect size – here the population correlation between an fMRI signal and a social/personality/behavioral measure. Given the reliability of fMRI signals and social/personality measures, we previously estimated the maximum theoretically possible population correlation to be 0.75 (Vul et al. 2009), and this figure assumes (rather absurdly, we would think) that were it not for measurement variability, this particular fMRI signal would account for 100% of the variance in this behavioral measure. A population correlation of 0.5 is also optimistic, but more plausible, assuming that 45% of the true variance in the behavioral measure could be explained by the noise-free fMRI signal. Finally, a population correlation of 0.25 may seem pessimistic, but strikes us as considerably more likely, as it assumes that a specific fMRI signal accounts for somewhat less than 10% of variability in behavior. Because without access to the original data we cannot assess statistical power of these published studies using permutation-based cluster-size correction, we instead consider simple Bonferroni correction, which seems to show well-calibrated correction under

low to moderate amounts of data smoothness (Nichols & Hayasaka, 2003). (For comparable analyses yielding similar results using FDR correction, see Appendix B).

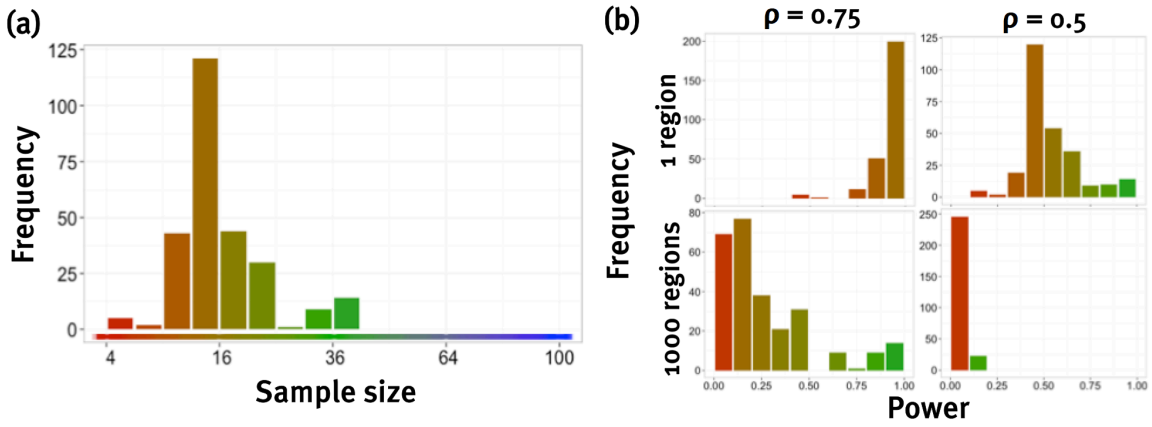

Figure 10. (a) The histogram of sample sizes from the studies surveyed in Vul et al. (2009), color coded to match the colors in Figure 9. (b) Histograms of the power these studies will have to detect a population correlation of 0.5 or 0.75 either with a single measured correlation, or with a 1000-voxel whole-brain analysis. The sample sizes used in these studies offer a lot of power for detecting an implausibly large population correlation in a univariate analysis (ρ=0.75, 1 region), but all have less than 20% power to detect a plausible (ρ=0.5) correlation in a whole-brain analysis.

Figure 9 shows the power we can expect for detecting a whole-brain, across-subject correlation varying in the number of independent brain regions, the number of subjects (sample size), and the underlying population correlation. When the true correlation is a modest 0.25, multiple comparisons correction renders sample sizes of even 100 grossly underpowered when there are more than 100 independent voxels in the whole-brain analysis. If the population correlation is 0.75 (the maximum theoretically possible), then a correlation analysis on a whole-brain analysis of 1000 voxels could achieve power of 80% with only 30 subjects. With a respectable (and more realistic) population correlation of 0.5 in a 1000-voxel brain, 30 subjects buys us only 10% power, and we would need 59 subjects just to get our power to a hardly impressive level of 50%. Considering that the largest sample size we found in our survey of published brain-behavior correlations was 38, with the median at 15, it seems that as a whole, whole-brain across-subject correlation studies are generally severely underpowered.

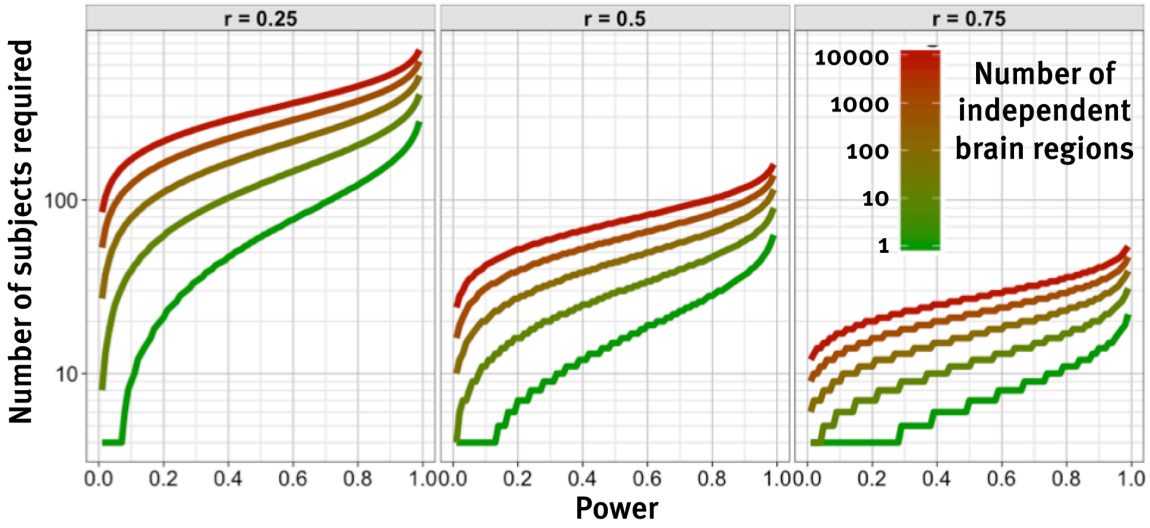

Figure 11. Sample size required (y) to achieve a certain level of power (x) as a function of the population correlation (panels), and the number of Bonferroni corrected comparisons (brain regions). A realistically small population correlation (ρ=0.25) will require hundreds of subjects in a whole-brain analysis (e.g., 1000 voxels) to achieve adequate power. However, even optimistic but plausible population correlations (ρ=0.5) require many more subjects than are commonly run in whole-brain across-subject correlation studies.

What is the expected power of the whole-brain correlation studies in the literature? Figure 10b shows a histogram of the power one might expect from the studies in our sample. Nearly all of them had adequate sample sizes to achieve 80% power for the theoretically maximal population correlation of 0.75, if they considered only a single measurement of the brain. However, a 1000-voxel whole-brain analysis with the same population correlation would yield less than 50% power for 87% of the studies, and less than 20% power for 54% of them. If we consider a more realistic population correlation of 0.5, all of the studies have less than 20% power for a whole-brain analysis, and 91% have power less than 5%. These are quite startling numbers: we showed in Figure 6 that overestimation from circular analyses is expected to be worrisome even with less than 50% power, and to be catastrophically bad with power below 20%; here it seems that the vast majority of studies that undertake whole-brain across-subject correlation analyses do so with less than 5% power for plausible effect sizes.

How large of a sample would be necessary to achieve adequate power? Figure 12 shows the sample size requirements for whole-brain analyses with different numbers of voxels. With a plausible population correlation of 0.5, a 1000-voxel whole-brain analysis would require 83 subjects to achieve 80% power. A sample size of 83 is five times greater than the average used in the studies we surveyed: collecting this much data in an fMRI experiment is an enormous expense that is not attempted by any except a few major collaborative networks.

In short, our exploration of power suggests that across-subject whole-brain correlation experiments are generally impractical: without adequate multiple comparisons correction they will have false positive rates approaching 100%, with adequate multiple comparisons correction they require 5 times as many subjects than what the typical lab currently utilizes. With the sample sizes currently used in the literature, such whole-brain correlation studies seem to have less than 5% power, meaning that if only half of the hypotheses tested in these studies are truly null (i.e., because there is no monotonic relationship between brain activity measured at the fMRI scale and individual differences), then more than half of the reported significant findings will be false positives (Pashler & Harris, 2012). This bizarre "winners' curse" outcome arises from the small fraction of voxels that will have a true signal, even when one is present: with so many candidate voxels in a whole-brain analysis, even 5% false positives will outnumber correct detections of the (generally sparse) true effects.

4. Conclusions

In our original paper, we reported that many correlations between individual differences in social and personality measures and brain activity measured by fMRI suffer from a "non-independence" error: They rely on a procedure that effectively picks out the largest of thousands

of correlations, and report those high sample correlations as estimates of the effect size (Vul et al, 2009a). This procedure is biased in that it is guaranteed to yield correlation estimates higher than the population correlation, just as publication bias (Rosenthal, 1979) will tend to yield inflated estimates of effect sizes across the published literature (Ioannidis, 2008). Thus, we implored the original authors to reanalyze their data with cross-validation procedures to obtain unbiased estimates and correct the literature. Although Poldrack and Mumford (2009) reanalyzed their own, analogous, experiment to estimate the magnitude of overestimation, as far as we know, none of the papers we reviewed were reanalyzed with unbiased methods.

It initially seemed to us that none of the authors carried out these reanalyses due to obstinacy and unwillingness to put their own "discoveries" at risk. However, our current analysis of power in whole-brain across-subject correlation studies puts their recalcitrance in a slightly more charitable light (motives aside): cross-validation, we now suspect, would rarely have surmounted the inherent problems with these severely underpowered datasets. Even using all of their subjects, these studies would probably have only 5% power for detecting plausible correlations; reducing this sample size further in a cross-validation would make the false-positive rate exceed the true positive rate, rendering any attempts at cross-validation futile.

Thus, we are now inclined to change our suggestions for how to avoid non-independence: cross-validation is clearly impractical for whole-brain across-subject correlation studies. Moreover, it seems that whole-brain across-subject correlation studies in general are impracticable, as they require sample sizes that are five times larger than typically used in fMRI experiments, likely financially impossible for all but the most well-funded research enterprises. So, unless researchers undertake enormous studies with adequate sample sizes, we are revising our suggestions: the best way to avoid non-independent effect size estimates, and false positive rates

that are comparable to true positive rates, is to avoid whole-brain across-subject correlation studies altogether. Instead, researchers should opt for within-subject region of interest approaches that obtain just a few signal estimates from each subject to avoid the many pitfalls of a voxel-by-voxel across-subject correlation.

More generally, the future of replicable research in whole-brain fMRI localization studies would be brighter with much larger sample-sizes. This will likely need to be achieved through multi-site consortia and/or comprehensive data-sharing and aggregation enterprises. Furthermore, to extract useful results from such datasets, there will need to be a concerted development of statistical methods for quantifying the variability and precision in localization estimates. That development process should provide important new insights regarding basic questions about human brain function and its variability across individuals.

References

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2010). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. Journal of Serendipitous and Unexpected results, 1(1), 1-5.

Bennett CM, Wolford GL, Miller MB. (2009) The principled control of false positives in neuroimaging. Soc Cogn Affect Neur.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews Neuroscience.

Carp, J. (2012) On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Frontiers in Neuroscience.

Churchill NW, Oder A, Abdi H, Tam F, Lee W, Thomas C, et al. (2012) Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. Human brain mapping. 2012

Cureton, E. E. (1950). Validity, Reliability, and Baloney. Educational and Psychological Measurement, 10, 94-96.

Fedorenko, E. & Kanwisher, N. (2009). Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? Language and Linguistics Compass. 3(10):1749-818.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A. and Noll, D. C. (1995), Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold. Magn Reson Med, 33: 636–647.

Friston, K.J., Rotsthein, P., Geng, J.J., Sterzer, P., Henson, R.N.A., (2006) A critique of functional localizers. NeuroImage

Gelman, A. & Loken, E. (2014) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no ``fishing expedition'' or ``p-hacking'' and the research hypothesis was posited ahead of time. Retrieved from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Ioannidis JP. (2005) Why most published research findings are false. PLoS medicine.

Ioannidis, J.P. (2008) Why most discovered true associations are inflated. Epidemiology, 19(5) 640-648.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci, 17(11), 4302-4311.

Kong, J., Gollub, R. L., Webb, J. M., Kong, J-T, Vangel, M. G., & Kwong, K. (2007). Test-retest study of fMRI signal change evoked by electroacupuncture stimulation. NeuroImage, 34, 1171-1181.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. Journal of Cerebral Blood Flow and Metabolism, 30(9), 1551-1557.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience.

Liu, P. & Hwang, J.T.G. (2007) Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics.*

Lund TE, Madsen KH, Sidaros K, Luo WL, Nichols TE. (2006) Non-white noise in fMRI: does modelling have an impact? NeuroImage. 2006

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological Methods, 9, 147–163.

Nichols, T.E., and Hayasaka, S. (2003). Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. Statistical Methods in Medical Research, 12(5), 419–446.

Poldrack RA, Mumford JA. 2009. Independence in ROI analysis: where is the voodoo? Soc Cogn Affect Neurosci. 4:208-13.

Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. (2012) Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage. 2012

Rosenthal, R. (1979), The file drawer problem and tolerance for null results, Psychological Bulletin, 86, 638–641

Simmons JP, Nelson LD, Simonsohn U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science.

Saxe, R., Brett, M., Kanwisher, N. (2006). Divide and Conquer: A defense of functional localizers. Neuroimage May 1. 30(4):1088-96

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009b). Reply to Comments on "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition". Perspectives on Psychological Science, 4(3), 319-324.

Vul E, Harris C, Winkielman P, Pashler H. (2009b) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. Perspectives on Psychological Science

Vul, E., & Kanwisher, N. (2010). Begging the question: The non-independence error in fMRI data analysis. In S. B. Hanson, M. (Ed.), Foundational Issues for human brain mapping. Cambridge, MA: MIT press.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power. Perspectives on Psychological Science, 4(3).

Yarkoni, T., & Braver, T. S. (2010). Cognitive neuroscience approaches to individual differences in working memory and executive control: Conceptual and methodological issues. In M. Gruszka, & Szymura (Ed.), Handbook of individual differences in cognition.

Appendix A: A quick tour of fMRI analysis

Individual analysis

A typical neuroimaging experiment yields massively multivariate time-series data: a 3D grid of about 10^5 “voxels” each measuring the amplitude of the blood oxygenation level dependent (BOLD) signal every few seconds throughout the experiment. Specifically, every few seconds (typically around 2s) the BOLD signal is acquired in a 3D volume of space that encompasses the brain of the participant. This image volume is divided up into ~10-30 "slices", where each slice is a square grid of voxels (commonly 64x64 or 128x128). Thus, there are roughly 10^5 or 10^6 voxels in a given imaging volume, and each voxel gets its own BOLD measurement at every acquisition. The resulting data are 4-dimensional: each of about 10^5 voxels in a 3D grid covering the imaging volume is associated with a timeseries of BOLD measurements at that point in space.

The 4-dimensional BOLD data are subject to assorted signal and image-processing procedures before any statistical analysis can be fruitfully carried out. These aim to align the 3D grid of voxels to itself over time and to anatomy (either that of the specific subject, or a group average, or a standard brain), to remove the measurements of non-brain regions (such as the skull, the ventricles, sometimes the white-matter), and to reduce the noise by filtering and smoothing the signals. Although there is considerable innovation and discussion about how to optimize these "pre-processing" procedures (Churchill et al., 2012), and some worry about possible errors that they might introduce (Power et al, 2012), that is not our focus here.

After pre-processing of BOLD data, it is analyzed via a "massively univariate" regression analysis to account for the BOLD timeseries in each voxel in terms of some number of time-varying task predictors. The timeseries of task-predictors are convolved with a "hemodynamic response function" that captures the dynamics of how BOLD signals respond to a single impulse. The resulting task-predictor time series are combined in a multiple regression (sometimes including nuisance predictors, like head-movement signals) to explain the BOLD timeseries for each voxel. This process yields a set of coefficients for each voxel, indicating how much a given task-predictor changes the BOLD activity in that voxel.

The task-predictor coefficients are typically analyzed via linear contrasts to identify the extent to which the voxel response is different for some tasks predictors than others; thus isolating the differential BOLD activation arising from a particular "task contrast". Via this process, the researcher collapses the timeseries of BOLD measurements at each voxel into a few contrast estimates, yielding one estimate per voxel per contrast of interest. For simplicity, let’s say there is only one contrast of interest in question (e.g., images of happy vs. sad faces), this yields ~10^5 contrast estimates for a given subject – one per voxel.

These 3D grids of contrast estimates can be subjected to a statistical threshold and displayed as “statistical parametric maps” for an individual subject, indicating the statistical reliability of the task contrast at each voxel for that subject. However, typically,

the aim is to make inferences about how these contrasts behave in the population, so from here analysis proceeds to the group stage that aggregates these linear contrasts across subjects in some way.

Within-subject region-of-interest vs. whole-brain analyses

Here we must distinguish two strategies for the group analysis: the *within-subject region of interest* (Saxe & Kanwisher, 2006) approach, and the *whole-brain* (Friston et al, 2006) approach.

The within-subject region of interest (ROI) approach aims to collapse the ~10^5 contrast estimates obtained for each subject (roughly one per voxel) into a few averages corresponding to specific areas of the brain. One such strategy would be to use an "anatomical localizer" to choose a region of the brain identifiable from anatomy alone (e.g., the amygdala), pick out the voxels that are in that region, and aggregate the contrast signal in that region. This would yield just a single measure of task-contrast per person: the contrast of the mean signal in the amygdala. An alternate region-of-interest strategy common in visual neuroscience relies on "functional localizers": an independent set of "localizer" data are used to define a region within an individual (e.g., contrast of faces-objects would yield a statistical parametric map that could be used to identify the fusiform face area in each subject, Kanwisher et al, 1997), then the contrast of interest would be averaged across all voxels within that functionally defined region. Just as for anatomical regions of interest, averaging task contrasts within a functionally defined region will yield just a single contrast estimate per subject (e.g., the contrast of the mean signal in the fusiform face area). By aggregating data within each subject into just one (or a few) measures, the region of interest approach avoids the statistical complications of massively multivariate data analysis, including stringent multiple comparisons correction and the non-independence error at the group analysis level.

In contrast to the region of interest approach, the whole-brain analysis does not collapse the fMRI task-contrasts of each individual into a few summary numbers for specific regions, but instead aims to assess how contrasts in each voxel behave across-subjects. This means that the whole-brain approach must grapple with massively multivariate voxel-by-voxel analysis at the group level.

Whole-brain, across-subject analysis

At the group analysis, the whole-brain across-subject study assesses how the task-contrast in each voxel varies across subjects. This might amount to assessing whether the mean contrast is sufficiently different from zero, given the across-subject reliability. Or this might mean assessing whether the magnitude of the task-contrast correlates with some other measure that varies across subjects (such as a personality score, behavioral test performance, or walking speed after the fMRI session). In either case, the whole-brain analysis now has ~10^5 across-subject statistical tests to perform: one for each voxel. Given the pre-processing (that averages signals of adjacent voxels to reduce noise while making them less independent), and some a priori constraints on which voxels are meaningful (those in the brain, rather than the skull or ventricles), the ~10^5 voxels

might yield only 10^4 or 10^3 effectively independent statistics each assessing how the task-contrast at a particular brain location varies across subjects.

Thus, the whole-brain across-subject correlation study now has 3D grids consisting of thousands of independent correlations between behavior and measures of BOLD activity. These can be (and indeed generally are) displayed as images (across-subject statistical parametric maps); however, to obtain quantitative summaries of these results, such as a correlation coefficient describing the brain-behavior relationship, investigators must somehow select a subset of voxels and aggregate correlations across them. This is generally achieved by defining a group region of interest either anatomically (e.g., all voxels in a region generally agreed to represent the amygdala, or all voxels within a certain radius of some a priori specified brain coordinates), functionally (e.g., voxels with task-contrasts that behave in a particular way across subjects), or some combination of anatomy and functional response.

A non-independent, or circular, effect size estimate requires a particular confluence of analysis decisions. First, it requires that the group analysis be carried out on a great many measures for each subject, most commonly, a whole-brain across-subject analysis. Second, it requires that the voxels over which effects are aggregated are selected based on the effect itself. Third, it requires that the same data be used to estimate the effect size as were used for selection.

Appendix B: Power calculations with False Discovery Rate correction

Since false discovery rate (FDR) correction is known to yield greater power than familywise error control procedures, readers might wonder whether the appallingly low power estimates suggested above reflected the assumption of Bonferroni correction rather than FDR. We can also calculate power for a false-discovery rate correction, on the assumption that it is carried out for many voxels, by adapting the calculations of (Liu & Hwang, 2007) to the across-subject correlation case. Again, we need to assume some true population correlation underlying the non-null voxels, and we must also assume the prevalence, or baserate, of voxels that have this signal. What kind of baserate would be plausible? If we take the published literature at face value, it would seem that fewer than 1/100 or 1/1000 voxels in the whole brain carry any one signal (e.g., one cluster of a few dozens voxels in a 10^5-voxel whole-brain analysis).

We find that power with FDR correction remains very low for detecting reasonable population correlations (0.5), unless the baserate of signal carrying voxels is implausibly high (greater than 10%; Figure 12). Specifically, with a plausible prevalence of 1% signal-carrying voxels in the brain, 91% of the studies we sampled would have power less than 10% to detect a population correlation of 0.5 (Figure 13). Indeed, to achieve 80% power for an FDR corrected whole-brain analysis looking for an across-subject population correlation of 0.5 with 1% prevalence, a study would need 66 subjects (Figure 14). While this is fewer than Bonferroni correction, that sample size is still 4 times greater than that of the median study in our sample, and nearly twice as large as the largest: in short, also an impractical sample size.

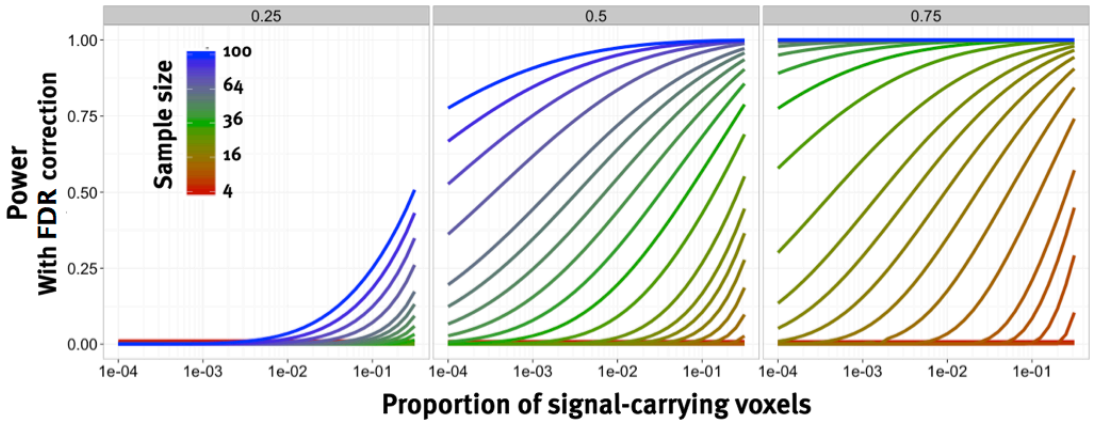

Figure 12. Statistical power (y) for FDR corrected correlation tests as a function of population correlation (panels), sample size (lines), and the proportion of voxels in the whole brain that contain the effect (x). A small population correlation (ρ=0.25; left) yields low power even when nearly 30% of brain voxels have this signal. In contrast large correlations (ρ=0.75; right) can be tested with high power with just 16 subjects, provided that 30% of the voxels contain the effect; however, if only 1/1000 voxels carry the signal, then twice as many subjects are needed to achieve the same level of power. A test for an optimistic, but plausible population correlation (ρ=0.5; middle) that is highly localized (occurring in 1/1000 voxels of the brain) requires nearly 100 subjects to achieve a high level of power.

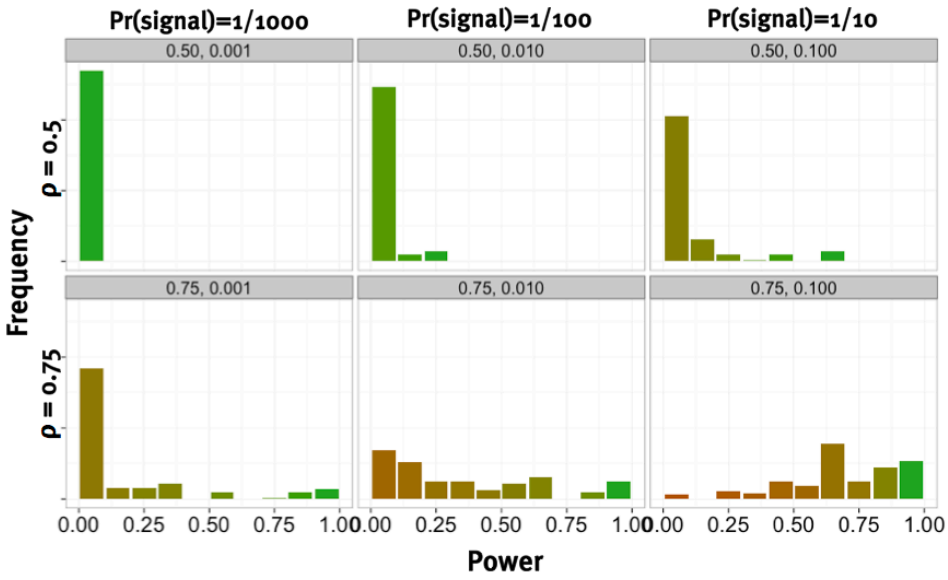

Figure 13. Histograms of the power the studies surveyed by Vul et al. (2009) will have to detect a population correlation of 0.5 or 0.75 with FDR correction, under different prevalence rates of the effect among tested voxels. 36% of the sample sizes used in these studies offer a lot of power for detecting an implausibly large and dense population correlation (ρ=0.75, prevalence=10%), but all have less than 30% power to detect a plausible (ρ=0.5) correlation with a prevalence of 1%; and less than 10% power if the prevalence is 1/1000.

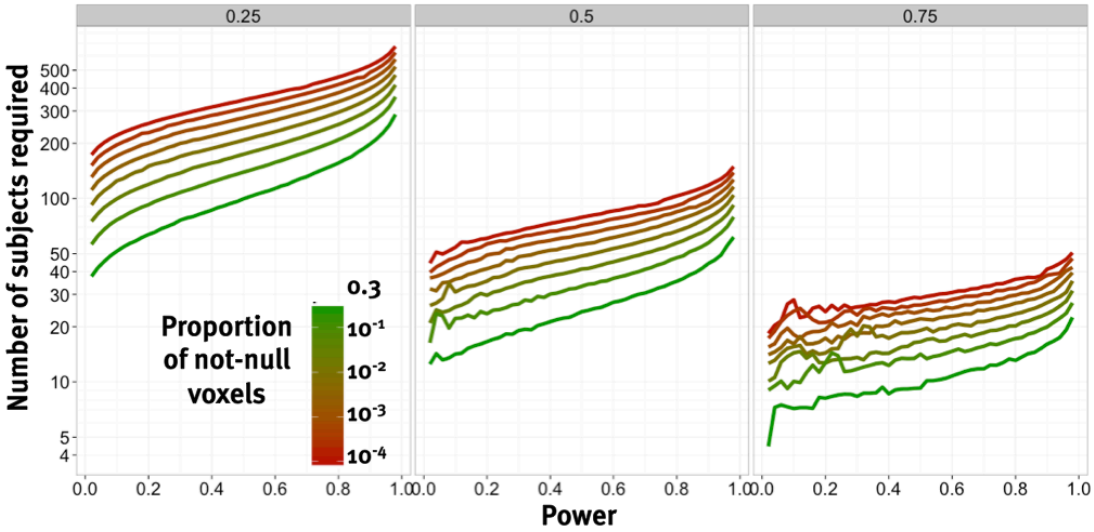

Figure 14. Sample size required (y) to achieve a certain level of power (x) as a function of the population correlation (panels), and the proportion of signal-carrying voxels in the FDR-corrected analysis. A realistically small population correlation (ρ=0.25) will require hundreds of subjects to achieve adequate power. However, even optimistic but plausible population correlations (ρ=0.5) require many more subjects than are commonly run in whole-brain across-subject correlation studies, if true effects are as sparse as reported results suggest.