Commentary on Vul et al.'s (2009) "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition"

Thomas E. Nichols[1,2,3] and **Jean-Baptist Poline**[4]

[1]*Clinical Imaging Centre, GlaxoSmithKline, London;* [2]*Centre for Functional MRI of the Brain, University of Oxford, Oxford, United Kingdom;* [3]*Department of Biostatistics, University of Michigan, Ann Arbor;* [4]*Institut d'Imagerie Biomedicale, Neurospin, CEA, Gif sur Yvette, France*

ABSTRACT—*The article "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (Vul, Harris, Winkielman, & Pashler, 2009, this issue) makes a broad case that current practice in neuroimaging methodology is deficient. Vul et al. go so far as to demand that authors retract or restate results, which we find wrongly casts suspicion on the confirmatory inference methods that form the foundation of neuroimaging statistics. We contend the authors' argument is overstated and that their work can be distilled down to two points already familiar to the neuroimaging community: that the multiple testing problem must be accounted for, and that reporting of methods and results should be improved. We also illuminate their concerns with standard statistical concepts such as the distinction between estimation and inference and between confirmatory and post hoc inferences, which makes their findings less puzzling.*

We are happy that Vul, Harris, Winkielman, and Pashler (2009, this issue) have generated such a stimulating discussion over fundamental statistical issues in neuroimaging. However, the issues raised are well known to experienced brain imaging

Address correspondence to Thomas E. Nichols, Clinical Imaging Centre, GlaxoSmithKline, Imperial College, Hammersmith Hospital, London, W12 0NN, United Kingdom; e-mail: thomas.e.nichols @gsk.com.

researchers, and the article could be distilled to two points that have already received much attention in the literature. The first one is that brain imaging has a massive multiple-testing problem (MTP) which must be accounted for in order to have trustworthy inferences, and the presence of this problem requires careful distinction between corrected and uncorrected inferences. Second, articles in neuroimaging have methods descriptions that are confusing or incomplete, which is a disservice to scientific discourse especially as neuroimaging reaches into new applied areas.

Finding solutions to the MTP has been an active area of research during the past two decades. We now have consensus methods that are widely accepted and used (see, e.g., Chapters 18–21 of Friston, 2006, or Chapter 14 of Jezzard, Matthews, & Smith, 2001). The two types of commonly used inference methods are those that control the family-wise error rate (FWE; Nichols & Hayasaka, 2003) and those that control the false discovery rate (FDR; Genovese, Lazar, & Nichols, 2002). FWE is the chance of one or more false positives, and Bonferroni and random field theory thresholds are just two methods that control FWE. A statistic image with a valid 5% FWE threshold is guaranteed to have no false positives at all with 95% confidence. FDR is a more lenient measure of false positives, and a valid 5% FDR threshold will allow as many as 5% of the suprathreshold voxels to be false positives on average. Both FWE and FDR methods can be applied voxel-wise, as a threshold on a statistic image, or cluster-wise, as a threshold on the size of clusters after applying an arbitrary cluster-forming threshold.

Whether FWE or FDR, voxel-wise or cluster-wise, corrected inferences must be used to ensure that results are not attributable to chance. Such corrected inferences are known as *confirmatory inference*, a test of a prespecified null hypothesis with calibrated false-positive risk. This is in distinction to exploratory or post hoc inference, in which no attempt is made to control false positive risk. Reporting and interpreting the voxels or clusters that survive a corrected threshold is a valid confirmatory inference and the foundation of brain imaging methodology. Complete reporting of these results usually consists of a corrected P and raw t (equivalently r) value, and in no way does the unveiling of the t value invalidate this inference.

The authors suggest that that the raw t (equivalently r) scores that survive a corrected threshold are impossible; this is incorrect because they are simply suprathreshold values that should be reported for what they are: post hoc measures of significance uncorrected for multiple testing. Crucially, as the raw scores are uncorrected measures, they are incomparable with a behavioral correlation that did not arise out of a search over 100,000 tests. In other words, those values are maximum values over a large number of comparisons.

This incompatibility issue is also related to how the authors misinterpret the reliability result (Nunnally, 1970), applying it to sample correlations when it is statement about population correlations. There is in fact substantial variation in a sample correlation about its true population value, with the approximate standard error of r being $1/\sqrt{n}$. Thus a sample correlation based on 25 subjects has an approximate 95% confidence interval of ± 0.4, and indicates that, in this setting, an r of 0.9 is entirely consistent with a ρ of 0.7.[1] Moreover, this sampling variability issue is magnified by the reporting of maximal correlations.

The second essential point of the article is that publications in neuroimaging have methods descriptions that are confusing or incomplete. Although this is a point of embarrassment for the field, it is a point that has been addressed in several publications (Carter, Heckers, Nichols, Pine, & Strother, 2008; Poldrack et al., 2007; Ridgway et al., 2008). If there is any misdeed committed by the nonindependent articles, perhaps it is that they failed to fully label the inferences as post hoc. It is self-evident that suprathreshold-selected post hoc tests give rise to greater correlations than do confirmatory tests based on one voxel or region of interest (Fig. 5 in Vul et al.) and it is not worthy of the tenor of the note. In particular, we argue that although authors have the responsibility to clearly and completely describe their methods and results, readers have the responsibility to understand the technology used and how to correctly interpret the results it generates. For example, in the field of genetics, whole-genome association analyses search over hundreds of thousands of tests for genotype–phenotype correlations and publications

routinely include plots of uncorrected P values (see, e.g., Fig. 4 in The Wellcome Trust Case Control Consortium, 2007). Yet we are unaware of any movement to suppress these plots from publication, presumably because genetic researchers understand the difference between these massive analyses and candidate single-nucleotide-polymorphism analyses in which no multiplicity is involved.

We must also take issue with the seemingly most compelling argument of the article, as explained in Figure 4 and the weather/stock-market correlation example. The problem with these examples is that they are based entirely on a null-hypothesis argument (i.e., the total noise case). However, if the articles used corrected thresholds, then the suprathreshold voxels will be mostly or entirely true positives.

As reviewed above, a 0.05 voxel-wise FWE threshold guarantees no more than a 0.05 chance that any null voxels will survive the threshold. In this case, Figure 4(a) is totally irrelevant (with 95% confidence) and the distribution of suprathreshold correlations is purely due to true positives. If, instead, the articles cited use a 0.05 FDR threshold, the suprathreshold voxels will be a mixture of true and false positives, but the fraction of false-positive voxels will be no more than 5% on average.

Finally, we find that the focus on correlation itself is problematic, as the correlation coefficient entangles estimation of effect magnitude and inference on a nonzero effect. A much more informative approach is to separately report significance and effect magnitude. That is, report significance with a corrected P value and report effect magnitude (still post hoc, of course) with a unit change in social behavioral score per unit percent blood oxygenation level dependent (BOLD) level change (as recommended in Poldrack et al., 2007). The behavioral scores have known scales and properties, and the percent BOLD change has an approximate interpretation of percent change in blood flow (Moonen & Bandettini, 2000). Reporting such measures will provide more interpretable and comparable measures for the reader.

The authors seem to be arguing that the field of neuroimaging should turn away from inference on where an effect is localized and focus instead solely on estimation of effect magnitude assuming a known location (Saxe, Brett, & Kanwisher, 2006). This is a significant shift in perspective that justifies ample and perhaps strident scientific discourse, but should not suggest that standard inferential practice is erroneous.

We would like to thank the authors for an engaging article that raises issues that apply to every neuroimaging study. However, we maintain that the community would have been better served if the alarmist rhetoric had been replaced by a measured discussion that made connections to standard statistical practice, distinguishing between estimation and inference and between confirmatory and post hoc inferences, and had simply acknowledged the incomparability of reported post hoc imaging correlations with other correlations in the psychology literature.

---

[1] More accurate confidence intervals can be computed with Monte Carlo or Fisher's Z transformation and may be shorter than the ± 0.4 approximate interval. However such intervals would still demonstrate the substantial sampling variability about the population value.

REFERENCES

Carter, C.S., Heckers, S., Nichols, T., Pine, D.S., & Strother, S. (2008). Optimizing the design and analysis of clinical fMRI research studies. *Biological Psychiatry, 64*, 842–849.

Friston, K.J. (2006). *Statistical parametric mapping: The analysis of functional brain images*. Amsterdam: Academic Press.

Genovese, C.R., Lazar, N., & Nichols, T.E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 15*, 870–878.

Jezzard, P., Matthews, P.M., & Smith, S.M. (Eds.). (2001). *Functional MRI: An introduction to methods*. Oxford, United Kingdom: Oxford University Press.

Moonen, C.T.W., & Bandettini, P.A. (2000). *Functional MRI*. Berlin, Germany: Springer.

Nichols, T.E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research, 12*, 419–446.

Nunnally, J.C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., & Nichols, T.E. (2007). Guidelines for reporting an fMRI study. *NeuroImage, 40*, 409–414.

Ridgway, G.R., Henley, S.M.D., Rohrer, J.D., Scahill, R.I., Warren, J.D., & Fox, N.C. (2008). Ten simple rules for reporting voxel-based morphometry studies. *NeuroImage, 40*, 1429–1435.

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage, 30*, 1088–1096.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*, 661–678.