doi:10.1093/scan/nsp011 SCAN (2009) 4, 208–213

Tools of the Trade

Independence in ROI analysis: where is the voodoo?

Russell A. Poldrack,[1,2] and Jeanette A. Mumford[1]

[1]Department of Psychology and [2]Department of Psychiatry & Biobehavioral Sciences, University of California, Los Angeles, CA 90095, USA

We discuss the effects of non-independence on region of interest (ROI) analysis of functional magnetic resonance imaging data, which has recently been raised in a prominent article by Vul *et al*. We outline the problem of non-independence, and use a previously published dataset to examine the effects of non-independence. These analyses show that very strong correlations (exceeding 0.8) can occur even when the ROI is completely independent of the data being analyzed, suggesting that the claims of Vul *et al*. regarding the implausibility of these high correlations are incorrect. We conclude with some recommendations to help limit the potential problems caused by non-independence.

Keywords: functional magnetic resonance imaging; region of interest analysis; bias; statistics; multiple comparisons

Rarely does a methodological review paper evoke the kind of frenzy that occurred when the paper on 'Voodoo correlations in social neuroscience' by Vul *et al.* (in press) was released as a preprint.[1] The blogosphere was soon abuzz with discussions of its implications, and authors on the 'red list' scrambled to write rejoinders to the piece and defend their methods and previous findings to editors and funding agencies. The discussion of this issue even reached the pages of *Newsweek* (Begley 2009), which reflects just how important functional magnetic resonance imaging (fMRI) has become due to its prevalence in the media.

In this article, we summarize the arguments of Vul *et al.* and discuss the strengths and weaknesses of several strategies to address the problem that their paper raises. We then evaluate the impact of using non-independent region of interest (ROI) analysis, using a published dataset that had originally included such analyses. We find that the bias due to using non-independent analysis is relatively small and does not invalidate the claims of the paper, and certainly does not support the dramatic label of 'voodoo'. We note up front that this does not necessarily imply that the same holds for other papers that have used non-independent analyses. We hope that others will also apply some of the methods discussed here in order to determine the degree of bias due to non-independence.

WHY ALL THE FUSS?

The basis for the argument by Vul *et al.* is simple and statistically incontrovertible (also see Kriegeskorte *et al.* 2009). Imagine a study in which one performs a whole-brain analysis to find a correlation between a personality test scores and brain activity across subjects, and thresholds the resulting statistical map at an uncorrected level of $P < 0.05$. However, a research assistant accidentally reordered the list of personality scores, so that they bear no true relation to brain activity. It is almost certain that some voxels will make it through this disorganized data analysis just by chance, even though there is no true relationship between brain activity and test scores. If one were to then take the signal from those surviving voxels and plot their relationship with the test scores, it might look quite impressive, but this is only because we have selected the voxels that show the best relation to the scores by chance.

Vul *et al.* motivated their review by noting that a number of studies in the social neuroscience literature reported 'implausibly high' correlations between brain activity and behavior (i.e. >0.8). They argued that it is rare for either fMRI signals or behavioral measures to have reliability above 0.8; because the maximum observable correlation coefficient is a function of the reliability of the measures being correlated, this would suggest that correlations above 0.8 are implausible. There are reasons to question the specific reliability estimates cited by Vul *et al.* [e.g. in the study by

1 The article was subsequently retitled 'Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and social Cognition'.

Received 28 February 2009; Accepted 6 March 2009

Correspondence should be addressed to Russell A. Poldrack, UCLA Department of Psychology, Wendell Jefferey and Bernice Wenzel Team Chair in Behavioral Neuroscience, Franz Hall, Box 951563, Los Angeles, CA 90095-1563, USA. E-mail: poldrack@ucla.edu

Aron *et al.* (2006) we found that 1-year test-restest reliability of fMRI signal estimates in regions of interest reached 0.99 in some cortical regions], but we will for the moment take their point at face value.

Motivated by this concern, Vul *et al.* surveyed a large set of papers from the social neuroscience literature, and then asked the authors of those papers for details regarding how the ROI analyses were performed. They then determined which papers had employed non-independent analyses; that is, analyses where the choice of voxels in the ROI analysis is made using the results from the whole-brain analysis, such as choosing the voxel with the maximum statistical value or taking the mean of a significant cluster. They compared the correlations obtained using these analyses with those obtained using independent analyses, e.g. using anatomical ROIs or independent localizer scans. Their meta-analysis showed that the studies using non-independent analyses reported correlations that were substantially higher than those reported in studies using independent ROI analyses. They conclude from this that correlations between behavioral tests and brain activity obtained using non-independent ROI analyses are not to be believed. The specifics of their meta-analysis have been called into question by Lieberman *et al.* in press, but the point that non-independent analysis can lead to bias is not in question.

INFERENCE *VS* PRESENTATION

The bias that is inherent in non-independent analyses would be deeply troubling if these analyses were the basis for the inferences reported in these papers. We suspect that this sometimes may be the case, but in most studies, inference from fMRI data is made on the basis of whole-brain voxelwise analyses. This inference can be plausible or not, depending upon the methods that are used. In particular, it is critical to employ accurate corrections for multiple tests, since a large number of voxels will generally be significant by chance if uncorrected statistics are used. An instructive example comes from Bennett *et al.* (2009). In a bit of instructive humor, these investigators scanned a dead salmon while showing it pictures of humans in social situations in a blocked design; the salmon was 'asked' to perform an emotional judgment task. Using methods that are not uncommon in the literature (i.e. an uncorrected threshold of $P < 0.001$ and 2-voxel extent threshold), they found a cluster within the salmon's brain that appeared activated, which disappeared upon using formal multiple comparison procedures. The problem of multiple comparisons is well known but unfortunately many journals still allow publication of results based on uncorrected whole-brain statistics.

There are well-developed and validated methods in the literature for multiple test correction, including family wise error (FWE) correction using Gaussian random field theory or nonparametric methods, which control the probability of having any false positives, and false discovery rate (FDR) correction, which controls the fraction of rejections that are false positives. Any statistic that passes an FWE or FDR correction when properly applied is guaranteed to be significantly different from the null value with a specific error rate, and any inferences made on the basis of those analyses are thus protected. If one then performs a non-independent ROI analysis on the significant voxels or clusters, the worst that can happen is that the observed effect size will be inflated, making the observed effect appear stronger than it actually is.[2]

Rather than using it for inference, when we have used non-independent analyses, the goal has generally been to examine the data that contributed to a significant correlation for quality control, and to convince our readers that the relationship observed in the data follows the expected functional form and is not driven by outliers. In our experience, correlations between fMRI signals and behavioral scores are notoriously riddled with outliers, which can sometimes result in very strong correlations that do not truly reflect the pattern across the group. This problem is so prevalent that we now try to use robust analyses whenever possible (e.g. Wager *et al.*, 2005; Woolrich, 2008), though there are some cases where robust analyses may not be feasible. Thus, we believe that whereas non-independent ROI analysis should play no role in inference, it can and should play a critical role as a sanity check for quality control. The lack of a visible outlier certainly does not prove that a result is robust, but the presence of a visible outlier can suggest the need for further investigation.

INDEPENDENT ROI ANALYSIS

Although we have argued that there is a place for non-independent ROI analysis, it is important to understand how much bias is introduced by those analyses, and this requires the parallel use of independent ROI analyses, in which the selection of the ROI is made with no information about the data being analyzed. As Vul *et al.* discuss, one approach to solving the problem of non-independence is to use ROIs that are either anatomically defined or defined using a completely independent localizer scan. Anatomical ROIs can certainly be useful, but they do pose some problems for analysis of functional MRI data (cf. Poldrack, 2007). First, anatomical ROIs are often large, such that the truly active voxels will make up a relatively small proportion of any anatomical region. This means that purely anatomical ROIs will almost always be biased towards the null

2 Studies often present results that are corrected using a 'small volume correction', in which the correction is much less severe because a much small number of tests is corrected for. This is legitimate if the small volume was identified completely independently of the data being analyzed. If the regions are chosen with any knowledge of the results, then there is a potential for bias. Because of the severe potential for bias, we are generally leery of the use of small volume corrections unless there is a clear regional prediction from multiple previous studies, and the small volume being corrected for must be chosen in an independent manner, e.g. using anatomically defined regions.

hypothesis. Second, if one does not have a preexisting anatomical hypothesis, then it is necessary to correct for a relatively large number of tests (e.g. 110 regions in the Harvard–Oxford Probabilistic Atlas that accompanies the FMRIB Software Library, FSL), which will also reduce sensitivity. The best solution is to obtain anatomical parcellations for each individual and use those to perform the ROI analysis; recent developments in automated anatomical parcellation (e.g. Fischl *et al.*, 2002) make this feasible, but such methods are not available in many centers and they require some degree of expertise to use successfully. Thus, anatomical ROIs may not be a suitable general solution to the problem of regional interrogation.

The functional localizer approach has been used to very good effect in visual neuroscience, and when available can be very useful. However, the use of functional localizers presupposes localization of function that is often not present, e.g. for regions such as prefrontal cortex. Thus, while very useful in some domains it also does not seem to offer a general solution.

The approach preferred by Vul *et al.* is the use of split-half or cross-validation strategies, wherein one portion of the data from each subject are used to create an ROI that is then used to interrogate the other portion of the data. Although the within-subject time series noise is independent across runs, the presence of any between-subject variance will induce a correlation between runs, making this approach non-independent. Examination of several datasets suggests that between-subject variance is present even in regions that are not activated, in which case the split half approach can still overestimate the true effect. Another alternative is to split the data across subjects, either splitting them into two groups or using more sophisticated cross-validation approaches. These approaches are in theory useful, but they can be difficult to interpret since each split will have a potentially different ROI. Additionally, both the split-runs and split-groups approaches reduce the amount of data that goes into the analysis, and thus increases the number of subjects that must be scanned to reach the same level of power (Poldrack and Mumford, 2009).

A CASE STUDY OF BIAS AND NON-INDEPENDENCE

In order to further examine the effects of non-independence, we reanalyzed a dataset that we had previously published including non-independent ROI analyses. Tom *et al.* (2007) presented subjects on each trial with 50/50 gain/loss gambles that parametrically varied the amount that could be gained or lost, and asked them to decide whether they would accept each gamble; the gambles were not resolved during scanning. Analyses estimated the parametric response in each voxel to gains and losses, and found that a set of regions (including ventromedial prefrontal cortex and ventral striatum) showed increasing activity for increasing possible gains and decreasing activity for increasing possible losses. Based on this analysis, we then computed a 'neural loss aversion' parameter that was defined as the difference in steepness between the (negative) slope of loss responses and the (positive) slope of gain responses; a positive neural loss aversion quotient reflected greater sensitivity to losses *vs* gains in that voxel. We then computed an analogous measure on behavioral data, and performed whole-brain correlation analysis (using robust regression) to identify voxels where there was a correlation across subjects between neural and behavioral loss aversion, controlling FDR at 0.05 across the entire brain. This analysis identified a set of clusters where such correlations were significant; the signal within each of these clusters was averaged for each subject, and these data were presented as scatterplots in the paper, along with correlation coefficients and *P*-values from the robust regression analysis. Thus, this was a non-independent ROI analysis, and the correlations for some regions were in the range (0.8–0.9) referred to as 'implausible' by Vul *et al.* (Table 1). In retrospect, it was a mistake to present the correlation and *P*-value numbers in the figure, as they are certainly biased for the reasons that Vul *et al.* describe. However, because we had controlled FDR at the whole-brain level, which was the basis for our inference, we had no undue concern about the true existence of that relationship. Inspired by the paper of Vul *et al.*, we wished to further investigate how badly the effect size estimates were inflated by the use of non-independent analysis.

Table 1 Original non-indepedent ROI analysis results from Supplementary Table 2 of Tom *et al.* (2007)

Correlation	Number of voxels	Anatomical location
0.9	284	L inferior/middle frontal
0.88	175	R inferior/middle frontal
0.87	104	L inferior frontal (opercular)/anterior insula
0.86	122	R inferior frontal (opercular)
0.85	332	B ventral striatum
0.83	358	R inferior parietal
0.81	110	B pre-supplementary motor area
0.46	963	L lateral occipital/cerebellum

Regions were obtained from a whole-brain analysis with FDR = 0.05 and a 100-voxel extent threshold. The first column presents the correlation between neural loss aversion and the log of the behavioral loss aversion parameter; the second column presents the size of the ROI and the third presents the anatomical location of the ROI.

Between-runs analysis

We first addressed the issue of bias by determining the ROIs from a subset of scanning runs and testing them on another subset. Because this study included three experimental runs for each subject, this was possible. There were three different

Table 2 ROI analysis using leave-one-out strategy, presented separately for each of the three left-out runs

Run 1			Run 2			Run 3		
Voxels	Test r	Train r	Voxels	Test r	Train r	Voxels	Test r	Train r
257	0.563	0.761	216	0.558	0.872	216	0.558	0.872
276	0.228	0.783	304	0.676	0.767	304	0.676	0.767
311	0.486	0.760	377	0.606	0.788	377	0.606	0.788
331	0.473	0.875	473	0.724	0.85	473	0.724	0.850
346	0.614	0.827	590	0.766	0.812	590	0.766	0.812
492	0.329	0.861	698	0.466	0.810	698	0.466	0.81
498	0.470	0.793	806	0.677	0.822	806	0.677	0.822
634	0.551	0.825	829	0.728	0.845	829	0.728	0.845
711	0.666	0.787	1341	0.510	0.748	1341	0.510	0.748
1135	0.552	0.806	2151	0.282	0.808	2151	0.282	0.808
Bias	**0.315**			**0.213**			**0.337**	

The columns include the number of voxels in the cluster, along with the correlation (Pearson r) between neural and behavioral loss aversion on the test (i.e. independent) and training (non-independent) data, respectively. Bias (presented in the bottom row) is computed by subtracting the mean correlation for test (independent) data from the correlation for training (non-independent) data across all clusters.

stimulus lists that were counterbalanced in order across the three scanning runs for each subject. For the purpose of the leave-one-run-out analysis, runs were grouped by stimulus list rather than by temporal order in the scanning session; because there were no systematic differences in the stimuli between the lists, this seemed appropriate. For each run, a statistical map was first computed by performing a whole-brain correlation analysis between behavioral and neural loss aversion measures on the other two runs. This map was thresholded at an uncorrected $t \leq 2.3$ and a cluster extent of 200 voxels; we used this uncorrected threshold because there were no voxels that passed a corrected threshold for one of the training sets, and because the split-half approach should in principle work even if the training set is analyzed using an uncorrected threshold.

For each pair of training runs, we took all of the clusters that passed this threshold and used them to create ROIs, from which we then extracted and averaged the data from the left-out run for each subject and computed the correlation between this mean signal and the behavioral loss aversion parameter. The results are presented in Table 2. These results show that the mean bias across all leave-one-out folds is 0.29; that is, the non-independent correlations are on average 0.29 higher than the independent correlations. All of the correlations in this analysis are somewhat lower than those obtained in our non-independent analyses reported in the paper, but still in a range (up to 0.77) that would suggest substantial effects. However, as mentioned above, the presence of non-zero between-subject variance can cause voxels to be correlated across runs, and therefore these values may still be biased. The next section used anatomical ROIs, which completely avoid the non-independence problem.

Table 3 Results from independent ROI analysis using anatomical ROIs

Correlation	*P*-value	Number of voxels	Anatomical region
0.772	0.05	729	L Inferior Frontal Gyrus, pars opercularis
0.793	0.027	696	L Inferior Temporal Gyrus, temporooccipital part
0.820	0.011	655	R Inferior Frontal Gyrus, pars opercularis

Results are presented for the three regions with correlations that were $P \leq .05$ after Bonferroni correction for the 111 regions tested. The first column presents the correlation between the mean neural loss aversion signal within the ROI and the behavioral loss aversion measure; the second column presents the corrected *P*-value; the third column presents the number of voxels within the ROI and the fourth presents the anatomical label.

Anatomical ROIs

Another approach to independent ROI analysis is to extract the data from anatomical ROIs, either defined by the subject's own anatomy or using an anatomical atlas. We applied this approach to the data from Tom *et al.*, using the Harvard–Oxford probabilistic anatomical atlas provided with FSL. This atlas provides probabilities that each voxel falls into a particular anatomical region across a dataset of 37 subjects. At each voxel, we assigned it to the most likely region at each voxel, as long as it had a likelihood of 25% or greater. For each subject, data were extracted from all voxels in each region, and the mean signal in these voxels was entered into a correlation analysis with the behavioral loss aversion parameter. The *P*-values were corrected using Bonferroni across all 111 regions; this is almost certainly too conservative due to correlations between regions, but we used it here to be maximally conservative.

Three regions exhibited correlations that reached significance at a Bonferroni-corrected level (Table 3).

This procedure is likely suboptimal because it will make it difficult to find correlations within large regions where the correlation only occurs in a relatively small number of voxels. Nonetheless, with a completely independent analysis it is possible to find correlations in the 0.7–0.8 range, which Vul *et al.* ruled to be implausible.

CONCLUSION AND RECOMMENDATIONS

Our analyses show that Vul *et al.* were correct that non-independent ROI analyses result in bias, but incorrect in their suggestion that correlations between behavior and imaging data in the 0.7–0.8 range are 'impossibly high'. We would hasten to note that this does not necessarily apply to other studies that have used non-independent analysis, and we would encourage authors to reanalyze their data, especially if the regions were derived using uncorrected whole-brain maps.

We have a number of recommendations that we hope will strengthen the results of any fMRI study and ensure that the resulting inferences are not impeachable on the grounds of bias:

(1) Strict control for multiple comparisons should *always* be employed. This will certainly reduce power and require larger sample sizes, but the alternative of inflated Type I error is unacceptable. There are a number of standard methods that are widely available for control of FWE, including Gaussian random field theory (Worsley *et al.* 1992) and non-parametric permutation testing (Nichols and Holmes, 2002), as well as methods for control of FDR (Genovese *et al.*, 2002) (though see Chumbley & Friston, 2009). We are generally suspicious of the use of small volume correction because of the potential for bias due to the selection of correction volumes once one has knowledge of the data. If one wishes to use a small volume correction approach, then one should choose those ROIs *before any data analysis has been performed.*

(2) Robust statistical methods should be used whenever possible, though they are not yet widely available. One notable exception is the outlier rejection method that is now available within the FSL software package (Woolrich, 2008). It is also possible to use the robust methods available in standard software such as MATLAB or R to obtain voxel-wise robust statistical estimates, but this can be cumbersome to implement.

(3) We believe that it is important to visualize the data that are driving an effect, using non-independent ROI analysis as a quality control step. Lack of an apparent artifact does not guarantee that the data are not corrupted, but problems can often be spotted by visualization of data in this manner. We recommend that these figures not be presented in publications due to their potential for misrepresenting the strength of the effect, but that they be included in Supplementary Materials for reviewers and interested readers.

(4) If one wishes to compute statistics from a ROI or present correlations in a figure within a paper, then independent analyses should be used. This can be done through the use of anatomically defined regions or though split-half analyses, in which part of the data are used to create the regions of interest through whole-brain analysis, and the other half is used to estimate the signal within those regions. In the case of split-half analyses, to completely avoid any correlations induced by the between-subject variability, the data should be split over subjects, not over runs. However, it should be noted that this approach causes a reduction in power and significant activation may be missed. Again, it is critical to point out that any anatomical regions of interest must be chosen prior to analyzing the data; otherwise, all regions should be analyzed and correction for multiple comparisons applied across those analyses.

(5) Whatever analyses are performed should be described in detail in the methods section or Supplementary Materials (cf. Poldrack *et al.*, 2008). One of the most worrisome aspects of the paper of Vul *et al.* is the difficulty that they encountered in determining how each analysis was performed; it should not be necessary to send a questionnaire to the authors in order to determine how an analysis was performed.

We believe that the paper by Vul *et al.*, despite its shortcomings, has done a service to the fMRI community by highlighting the need for methdological care and the potential for bias that can arise with some forms of analysis. We hope that the field will take these lessons to heart and ensure that fMRI results are never again open to the claim of voodoo.

SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

REFERENCES

Aron, A.R., Gluck, M.A., Poldrack, R.A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage*, *29*, 1000–6.

Begley, S. (2009). Of voodoo and the brain. *Newsweek*, **VLIII**, 25.

Bennett, C.M., Miller, M.B., Wolford, G.L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. In: *Organization for Human Brain Mapping Abstracts.*

Chumbley, J.R., Friston, K.J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage*, *44*, 62–70.

Fischl, B., Salat, D.H., Busa, E., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*, 341–55.

Genovese, C.R., Lazar, N.A., Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*, 870–8.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience*, *5*, 535–40.

Lieberman, M.D., Berkman, E.T., Wager, T.D. (in press). Correlations in social neuroscience aren't voodoo; A reply to Vul et al. *Perspectives on Psychological Science*.

Nichols, T.E., Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, *15*, 1–25.

Poldrack, R.A. (2007). Region of interest analysis for fMRI. *Social, Cognitive, and Affective Neuroscience*, *2*, 67–70.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E. (2008). Guidelines for reporting an fMRI study. *Neuroimage*, *40*, 409–14.

Poldrack, R.A., Mumford, J.A. (2009). On the proper role of non-independent ROI analysis: A commentary on Vul and Kanwisher. In: Bunzl, M., Hanson, S.J., editors. *Philosophical Foundations of Neuroimaging*. Cambridge, MA: MIT Press.

Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*, 515–8.

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

Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, *26*, 99–113.

Woolrich, M. (2008). Robust group analysis using outlier inference. *Neuroimage*, *41*, 286–301.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*, 900–18.