

Discussion of Vul et al. (2009)

Nicole A. Lazar

Discussion of “Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition” by Vul et al. (2009)

Nicole A. Lazar

*Department of Statistics, University of Georgia*

Address correspondence to Nicole A. Lazar, Department of Statistics, University of Georgia, Athens, GA 30602; e-mail: [nlazar@stat.uga.edu](mailto:nlazar@stat.uga.edu).

**ABSTRACT**—In their article, Vul, Harris, Winkielman, and Pashler (2009, this issue) raise the issue of nonindependent analysis in behavioral neuroimaging, whereby correlations are artificially inflated as a result of spurious statistical procedures. In this comment, I note that the phenomenon in question is a type of selection bias and hence is neither new nor unique to fMRI. The use of massive, complex data sets (common in modern applications) to answer increasingly intricate scientific questions presents many potential pitfalls to valid statistical analysis. Strong collaboration between statisticians and scientists and the development of statistical methods specific to the types of data encountered in practice can help researchers avoid these pitfalls.

Statistical tools were originally devised for a rather specific set of circumstances—mostly small or moderately sized data sets, collected under controlled experimental conditions—in disciplines such as agriculture, chemistry, and astronomy. Pioneering theorists of the early 20th century—Fisher, Gosset, and Pearson, for example—could not have anticipated the types of data problems that statisticians working less than a mere 100 years later would encounter. Modern science and the proliferation of data everywhere in our daily lives (their generation and collection) have posed unprecedented challenges to statisticians and scientists alike. This is an exciting time to be working with data, but also one laden with pitfalls to a “proper” (statistically valid) analysis.

Vul, Harris, Winkielman, and Pashler (2009, this issue) point out one such pitfall, which they term the *nonindependence error*, in certain types of correlation analysis common in functional and behavioral neuroimaging. I prefer to label this a *selection bias*, borrowing a phrase from survey sampling. Naming it thus highlights that the phenomenon is neither new, nor unique to the neuroimaging context described by Vul and colleagues. Selection bias is obviously well known in sampling, where, for example, individuals who choose to respond to an internet survey are likely those with stronger opinions (for or against) the matter in question, and hence are not representative of the population as a whole. Inference drawn from such a self-selecting sample will tend to exaggerate the studied effect, and these polls are rightly viewed with skepticism. Closer to home, this type of selection bias is also known as the “file drawer problem” in meta-analysis (Rosenthal, 1979), as typically only studies with significant results (that is,  $p$  values less than .05) get published, effect sizes estimated by meta-analysis, relying on the published literature alone, tend to be biased in the direction of overinflation.

As for uniqueness, Benjamini (2008) recently wrote about almost being booted off the

stage at a genomics conference, for pointing out the selection bias in the genes that are chosen for analysis. As he writes: “. . . a fifth [problem is] the set of  $p$  values reaching the stage of statistical analysis has been selected from the set originally measured. . . . this is not an innocent act.” (Benjamini, 2008, p. 25) He later describes going back to the Methods sections of several articles in the genetics literature to understand the basis on which genes were selected or deselected for further analysis. I suspect that if one were to delve into other areas where large complicated data sets are common, one would find similar selection biases arising.

So, where does this leave us as statisticians and practitioners? Vul et al. rightly note that the phenomenon in question is one that, at some level, everyone is aware of—yet, the error keeps creeping in. I contend that it is the nature of large, complicated data sets, coupled with the intricate scientific questions, that obscures the relatively clear edict not to use the same data twice (or more) for a single analysis.

Let us consider, as the authors do, analyses based on regions of interest (ROIs). Part of the difficulty—and I have encountered this many times over the years, albeit in different guises each time—is summarizing what takes place in a region, which may encompass hundreds of voxels, and using this in a further statistical analysis to answer specific questions of scientific interest. It is extremely tempting to focus on a subset of the voxels, and once this step is made, it seems natural to look at the “most activated” of those. After all, prescreening based on activation levels is common as a dimension reduction technique prior to, for example, clustering of fMRI time courses (see, for instance, Goutte, Toft, Rostrup, Nielsen, & Hansen, 1999). And those “peak voxels” express most strongly the typical behavior of the ROI (or so the reasoning would go). If this is not a valid approach, what is one to do? What is the way forward in fairly characterizing the disparate behaviors of the many voxels that make up a region? Whether ROIs

are defined anatomically or functionally (or both), this question is bound to arise; hence, simply having anatomically defined ROIs is not really sufficient to preclude the selection bias. My colleagues and I are currently exploring objective ways (that is, ways that are not based on peeking at the data first, nor on using the values of the data to include particular voxels and exclude others) of summarizing the dominant pattern among the voxels within a given ROI. This could then be passed forward to a next level of statistical analysis, and selection bias would be avoided. There is a wealth of multivariate statistical methods, as well as approaches based on functional data analysis (Ramsay & Silverman, 1997), that seem to be both promising and appropriate; see Lazar (2008) for further discussion of some of these ideas.

Vul et al. paint a rather bleak picture of the current state of one corner of the statistical neuroimaging world. Perhaps they are overly pessimistic. Researchers don't set out to perform statistically suspect analyses—data are too expensive and the implications are too great. Rather, as I argue here, I think that complicated, large data sets used to answer increasingly complex scientific questions together with the field of statistics that is still making a paradigm shift away from its roots in small, relatively structured samples (Efron, 2008) increase our liability to make errors in the direction of selection bias. Corrections for multiple testing are a start, as they explicitly recognize that looking repeatedly at the data increases the probability of making a Type I error, but they are not a panacea. Rather, there is a need to develop new methods specifically tailored for large, heavily correlated, spatiotemporal data; some such methods already exist but they are not in widespread use in the neuroimaging community.

As scientists—and statisticians—working in these disciplines become more aware of potential pitfalls, the situation will no doubt improve. Increased transparency in reporting the details of an analysis will also help. Appealing to statistical methods that sidestep the potential

for selection bias (for instance, by including all voxels in the ROI, but differentially weighting them according to their “representativeness” of the behavior in the ROI as a whole) is yet a third rail in this scheme. Finally, as mentioned above, the community should strive to develop statistical methodologies that are specific to the types of data collected and the questions that are being asked; this will most likely involve moving away from traditional linear models and correlations.

### REFERENCES

- Benjamini, Y. (2008). Comment: Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, *23*, 23–28.
- Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, *23*, 1–22.
- Goutte, C., Toft, P., Rostrup, E., Nielsen, F.A., & Hansen, L.K. (1999). On clustering fMRI time series. *NeuroImage*, *9*, 298–310.
- Lazar, N.A. (2008). *The statistical analysis of functional MRI data*. New York: Springer.
- Ramsay, J.O., & Silverman, B.W. (1997). *Functional data analysis*. New York: Springer.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.
- 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*, xx–xx.