Correlations and Multiple Comparisons in Functional Imaging – a Statistical Perspective

Martin A. Lindquist and Andrew Gelman

Department of Statistics, Columbia University, New York, NY, 10027

ADDRESS:

Martin Lindquist
Department of Statistics
1255 Amsterdam Ave, 10th Floor, MC 4409
New York, NY 10027
Phone: (212) 851-2148
Fax: (212) 851-2164
E-Mail: martin@stat.columbia.edu

To appear in *Perspectives in Psychological Science* (2009)

Abstract

Vul et al. claim in their paper that the correlations reported in fMRI studies are commonly overstated because researchers tend to report only the highest correlations, or only those correlations that exceed some threshold. Their paper has in a short time given rise to a spirited debate about key statistical issues at the heart of most functional neuroimaging studies. The debate provides a useful opportunity to discuss core statistical issues in neuroimaging and ultimately provides a chance for the field to grow and move forward. This commentary approaches the debate from a fundamentally statistical perspective. We begin by summarizing several of the key points under discussion, followed by our own commentary on these issues from a statistical point of view. We conclude our discussion by contemplating whether it may be time to move beyond the correlation and multiple comparisons framework, which is causing so much confusion, and instead represent all relevant research questions as parameters in one coherent multilevel model.

Introduction

With great interest, we have followed the spirited debate raging around the article originally entitled, "Voodoo Correlations in Social Neuroscience" by Ed Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. We are pleased that the paper has created such a stimulating discussion about key statistical issues that are at the heart of most functional neuroimaging studies. In general, we feel that the debate provides a useful opportunity to discuss core statistical issues in neuroimaging and ultimately, we hope it provides a chance for the field to grow and move forward.

Our thoughts on these issues come from a statistical perspective as our training lies primarily outside of neuroscience.[1] However, since the discussion is in essence a statistical one, we feel that perhaps we have something to add. We begin our discussion by summarizing several of the key points that have arisen in the debate so far, followed by our commentary on the issues from a fundamentally statistical point of view.

A Summary of the Debate

In their article, Vul et al. point out that the correlations reported in fMRI studies are commonly overstated because researchers tend to report only the highest correlations, or only those correlations that exceed some threshold. They suggest that these statistical problems are leading researchers, and the general public, to overstate the connections between social behaviors and specific brain patterns. They react particularly strongly to the practice of using a two-stage

[1] Gelman did his Ph.D. thesis work on medical imaging (PET scans) but has published only one article in this field, over fifteen years ago. Lindquist works on fMRI and collaborates with Tor Wager, who is a coauthor of one of the articles being discussed here. We shared drafts of earlier versions of this article with Ed Vul, Hal Pashler, Tor Wager, and others, and we do not think our affiliation with Wager has biased our assessment here.

analysis procedure where the method used to select which voxels should be tested is not independent of the tests performed on the resulting regions.

After appearing online, the paper received a great deal of attention and gave rise to multiple responses, several of which were centered on the idea that properly performed corrections for multiple comparisons allow researchers to, in large, circumvent the problems raised by Vul et al. For example, Jabbi, Keysers, Singer, and Stephan argue that "correcting for multiple comparisons eliminates the concern by Vul et al. that the voxel selection 'distorts the results by selecting noise exhibited by the effects being searched for.'" Huizenga, Winkel, Grasman and Waldorp argue that if adequate corrections for multiple comparisons are performed, it is not warranted to label high correlations as being "voodoo" and that "the correlations simply are high because they survive more conservative thresholds." Both Jabbi et al. and Huizenga et al. argue that the focus should be on the statistical testing and not on the magnitude of the correlations: as Jabbi et al. write, "the key question is often not *how strongly* the two measurements are correlated, but *whether* and *where in the brain* such correlations may exist." In a separate discussion, Nichols and Poline feel the paper discusses two key points that have already received much attention in the literature. The first is the problem of multiple testing and the second is that methods descriptions in neuroimaging papers are confusing or incomplete. Finally, they agree that the focus on correlation itself is problematic, as it entangles effect magnitude and significance.

Finally, Lieberman, Berkman, and Wager defend social neuroscience and argue that while they accept that correlations are overstated, the correlations may not be nearly as overstated as Vul et al. fear. In addition, they disagree with the implied claim that the overstated correlations have distorted scientists' understanding of social neuroscience research. They

further object to Vul et al.'s focusing on social neuroscience, given that the same statistical issues arise in all sorts of brain imaging studies. Finally, they point out some specific areas where Vul et al. mischaracterized the data-analytic methods used in this field. In particular, they react strongly to implications that researchers use a two-stage analysis procedure with inferences at both steps. Instead, they write, most researchers use a single-stage test to search for regions showing significant non-zero correlation, with a subsequent correction for multiple comparisons. While they agree with Vul et al. that massive numbers of tests with multiple comparison corrections are not a good way to provide unbiased estimates of the correlation magnitude, they claim this is not the purpose of their analysis.

Statistical Thoughts

The debate so far has raised several interesting statistical questions. The first is the validity of the so-called non-independent two-stage analysis procedure criticized by Vul et al. From a statistician's point of view it is hard to disagree with their statement that it is unsound to perform a two-stage analysis that tests the significance of nonzero correlation on voxels that were chosen simply due to that fact that they exhibited high correlation in the data. However, it is unclear how often this type of analysis is actually used in the literature and quantifying this is beyond the scope of our expertise. Lieberman et al. give a compelling argument that it is not common (at least in the studies surveyed by Vul et al.), and that most studies first conduct a test of significance and thereafter simply report an aggregate correlation value for each region deemed significant in the first test. With proper control for multiple comparisons, this second procedure will not change the underlying result that certain voxels exhibited significant nonzero correlation in the hypothesis testing framework, but the reported correlation will be radically inflated.

This leads us to the next question regarding the interpretation of the reported correlations. Researchers often do, as Vul et al. point out, use correlations to summarize their results. However, the appropriate guidelines for interpreting these results are often not provided, and even if said correlations survived a multiple-comparisons analysis, readers might interpret these at face value without understanding the selection issue. For these reasons, the practice of simply reporting the magnitude of the reported correlations is somewhat suspect. The fact that many imaging studies are underpowered adds an additional wrinkle, as estimates with relatively large standard errors are more likely to produce effect estimates that are larger in magnitude than estimates with relatively smaller standard errors; regardless of the true effect size. Indeed, it is well known that with a large enough sample size even very small effects will be statistically significant, and statisticians often warn about mistaking statistical significance in a large sample for practical importance. However, on a similar note, just because it is difficult to obtain statistically significant results in a small sample, this doesn't necessarily imply that that said effects are real and important (Gelman & Weakliem, 2009). Often large estimates simply reflect the influence of random variation. This may be disappointing to researchers, since they may indicate that even significant findings do not provide strong evidence. However, accurately identifying findings that are suggestive rather than definitive still benefits the field.

The commentary by Nichols and Poline raises an important point regarding the quality of the methods sections in neuroscience publications. It is critical to provide readers with the necessary tools needed to correctly interpret the results, and researchers should avoid trying to overstate the results in question. Similarly, while statistics provides many useful methods, the conclusions are often only as valid as the underlying model assumptions. If these assumptions fall apart, so may the validity of the conclusions being made. It is therefore important for papers

to contain careful descriptions of the assumptions required for the methods they employ and to state conclusions in the context of these assumptions. Readers can then decide on their validity and interpret the conclusions of the study in the appropriate context.

Many of the responses have centered on the multiple comparisons problem and how appropriate control of these issues allows one to circumvent the problems outlined by Vul et al. Multiple comparisons methods are designed to control the rate of false positives in a setting where true effects are zero, but one can certainly imagine situations where this may not actually be the most relevant null hypothesis. There are many factors that affect blood flow in the brain, and we probably wouldn't expect the average scans of two different groups of people to be exactly the same. Hence, if the number of subjects is large enough we would expect to see significant correlations over most of the brain, even after proper correction for multiple comparisons. For these reasons, it is not clear that the approach based on separate analyses of voxels and p-values is optimal, as rejecting the hypothesis of zero correlations may not actually be what is most interesting at the end of the day. What's really of interest is the pattern of differences in the brain, and how consistent these patterns are across persons and conditions. Related to this point is that, ultimately, when trying to understand differences in brain processing between different groups of people (or between people doing different tasks), the maximum correlation among voxels is not what you're looking for. That may be one reason why researchers summarize using regions of interest (as discussed in the Lieberman et al. article).

Vul et al. are correct to warn about overinterpretation of correlations that have been selected as the maximum, as the naive reader can see such correlations (and accompanying scatterplots) and think that certain personality traits are more predictable from brain scans than they actually are. The fact that certain correlations survive the multiple comparisons procedure

is evidence against the hypothesis of zero differences, and does not imply that these correlations can be directly interpreted.

Perhaps the way forward is to go beyond the correlation and the multiple comparisons framework, which causes so much confusion. Vul et al. and Lieberman et al. both correctly point out that classical multiple comparisons adjustments do not eliminate the systematic overstatement of correlations. Therefore, rather than correcting for problems arising from multiple significance tests, perhaps it is more appropriate to represent all relevant research questions as parameters in one coherent multilevel model. In other words, rather than correcting for a perceived problem, we should just build a more appropriate model from the start.

A multilevel Bayesian approach using some sort of mixture for the population of voxel differences, ideally modeled hierarchically with voxels grouped within regions of interest, would help here. These types of models shift estimates and their corresponding intervals toward each other through a process referred to as partial pooling (or shrinkage). In contrast, classical procedures keep the point estimates stationary and adjust for multiple comparisons by making the intervals wider. In this way, multilevel estimates make comparisons appropriately more conservative in a data-driven manner. As a result, we can say with confidence that those comparisons made with multilevel estimates are more likely to be valid. At the same time this adjustment doesn't detract from our power to detect true differences as is often the case in the multiple comparisons framework (Gelman, Hill & Yajima, 2009).

In essence, classical inference only uses information in each voxel to obtain voxel-wise effect estimates and their corresponding standard error. A multilevel model recognizes that the voxel-wise estimate is ignoring information provided by the other voxels. While still allowing for heterogeneity across voxels, the multilevel model also recognizes that since all the voxels are

measuring the same phenomenon it doesn't make sense to completely ignore what has been found in other voxels. Therefore, each voxel-specific estimate gets shrunk towards the overall estimate. The greater the uncertainty, the more it will get pulled toward the overall estimate. The less the uncertainty, the more we trust that individual estimate and the less it gets shrunk. This process leads to estimates that lie closer together than those obtained using classical analysis. Rather than inflating our uncertainty estimates, which doesn't really reflect the information we have regarding the effect size, the point estimates are shifted in ways that reflect the information we have. It has been recognized (James and Stein, 1960; Efron and Morris, 1975) that partial pooling can lead to estimates with better properties than traditional estimators. It should also be noted that partial pooling has previously been applied to fMRI time series data in the context of the multilevel general linear model (GLM) approach (e.g. Friston et al., 2002 and Friston and Penny, 2003).

At its simplest, the model would have two levels: a data-level model of the measurement of each correlation given the true underlying correlation, and a model of the distribution of correlations. For example, if $\hat{\rho}_i$ represents the measurement of the correlation at voxel *i* and ρ_i the true underlying correlation, we can write $\hat{\rho}_i \sim N(\rho_i, \sigma_\rho^2)$, and then $\rho_i \sim \lambda g_0 + (1-\lambda) g_1$, where g_0 is a distribution with a spike at zero (representing the idea that most correlations are expected to be small) and g_1 is a wider distribution representing the correlations that can appear in reality. The goal is to estimate λ and perform inference for individual ρ_i's and for the average correlations over regions of interest, all of which can be done within the Bayesian framework (Gelman, et al., 2003).

Detailing out the model above reveals many problems in its simplicity, most notably first that the measurements are not independent and second that the correlations are themselves

correlated, both spatially and also with regard to the experimental conditions. The difficulty of this sort of modeling is presumably one reason why it is not done in practice. On the other hand, if the correlations are to be analyzed, we suspect that a hierarchical mixture model would address the multiple comparisons issue (by explicitly estimating the distribution of the correlations) and also solve the crude overestimation problem that comes from selecting the maximum.

We view our suggested hierarchical model for partially pooling correlations not as a competitor to a full probability model of the data and data collection process (as in Genovese, 2000) but rather as a sort of rationalized reconstruction of existing correlation analysis to better adjust for the multiplicities in the analysis.

Final Thoughts

The motivations of Vul et al. in writing their article no doubt included frustration at too-good-to-be-true numbers which they felt led to exaggerated claims of neuro super-science. Conversely, one of the frustrations of Lieberman et al. is that they are doing a lot more than correlations and fishing expeditions--they're running experiments to test theories in psychology and trying to synthesize results from many different labs. From that perspective it must be frustrating for them to see a criticism that is so focused on correlation, which is really the least of their concerns. The frustration was no doubt exacerbated by what they saw as a mischaracterization of their analysis techniques.

It also seems that both sides were irritated by what they saw as giddy press coverage: on one side, claims of dramatic breakthroughs in understanding the biological basis of behavior and personality; on the other, claims of a dramatic emperor-has-no-clothes debunking. As scientists, most of us welcome press coverage--after all, we think our work is important and we'd like

others to know about it--but we are sensitive to uncritical press coverage of work we see as flawed. As applied statisticians, we are happy to see the discussion raised by Vul et al. and their correspondents and we hope this will lead to statistical methods that more directly address the important research questions in psychology that are being studied by Lieberman et al. and others.

References:

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70, 311-319.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G. and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.

Friston, K.J. and Penny, W. (2003). Posterior probability maps and spms. NeuroImage 19 (3), 1240–1249.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). Bayesian Data Analysis. Chapman & Hall/CRC, second edition.

Gelman, A. and Weakliem, D. (2009). Of beauty, sex, and power: statistical challenges in estimating small effects. Under revision for American Scientist.

Gelman, A., Hill, J. and Yajima, M. (2009). Why we (usually) don't have to worry about multiple comparisons. Technical Report, Department of Statistics, Columbia University.

Genovese, C. R. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data (with discussion), Journal of the American Statistical Association, 95, 691-703.

James, W., and Stein, C. (1960). Estimation with quadratic loss. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, ed. J. Neyman, 361-380. Berkeley: University of California Press.