NeuroImage 62 (2012) 945–948

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Review

Voodoo and circularity errors

Edward Vul *, Hal Pashler

Dept. of Psychology, University of California, San Diego, United States

ARTICLE INFO

Article history:
Accepted 1 January 2012
Available online 9 January 2012

Keywords:
fMRI
Data analysis

ABSTRACT

We briefly describe the circularity/non-independence problem, and our perception of the impact the ensuing discussion has had on fMRI research.

Contents

In early 2005 a speaker in our department reported that BOLD activity in a small region of the brain can account 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). The implications of this result struck us as puzzling, to say the least: Are walking speeds really so reliable that most of their variability can be predicted? Does a focal cortical region determine walking speeds? Are walking speeds largely predetermined hours in advance? These implications all struck us as far-fetched. This puzzlement, and a few other encounters with similarly remarkable correlations, prompted us a few years later to look into the matter further. We started by asking if dramatically high correlations between fMRI data and individual differences in social behavior were frequent in the literature. It turned out that they were common in all types of brain imaging studies that were looking at individual differences across people. While we suspected that something must be terribly amiss with these reported correlations, at first it was not at all apparent what that might be. Our efforts to figure this out led to a 2009 article – initially titled “Voodoo Correlations in Social Neuroscience” – which generated far more interest and controversy than we had remotely anticipated.

When we began looking into these correlations we were of course mindful of the fact that the brain contains a great many voxels, and suspicious that some sort of "voxel shopping" lay at the root of the problem. However, since the late 90s, fMRI practitioners have been well aware of the multiple-comparisons problem, and most of the papers reporting high correlations alluded to precautionary measures taken to correct for multiple comparisons.

Our interest in probing the matter was further whetted by an episode occurring a short while later: Grill-Spector (Grill-Spector et al., 2006) reported that individual voxels in face selective regions have a variety of stable stimulus preferences; in a critical commentary, Baker et al. (Baker et al., 2007) found that the analysis used to ascertain this fact implicitly built these conclusions into the method, such that the same analysis applied to noise data (voxels from the nasal cavity) revealed a similar variety of stable preferences. It occurred to us that a similar circularity might underlie the puzzlingly high correlations.

To figure out whether such a “selection bias” was lurking behind the surprisingly high correlations between social behavior and focal brain regions, we surveyed the literature reporting these correlations and sent out a survey to the authors to ascertain exactly how those correlations were computed (because most of the method sections

* Corresponding author.
E-mail address: evul@ucsd.edu (E. Vul).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.01.027

did not make this clear). The results revealed that over half of the reported correlations (and an overwhelming majority when considering only the very high correlations) were reporting correlations measured on a region picked out precisely for having very high correlations, and were thus contaminated by the same sort of bias uncovered by Baker et al. (2007).

The essence of the error is that using the same data to filter out voxels that carry relevant signal and to estimate the strength/reliability of that signal results in systematic overestimation of signal strength. When considering experiments rather than voxels, this effect is known as "publication bias" — experiments are filtered by the preference to publish significant results; thus meta-analyses reveal systematic overestimation of effect sizes via funnel plots (Egger et al., 1997). For fMRI data, where there are many voxels to consider within one experiment, this systematic overestimation can arise within individual analyses, and to a greater degree because of the large number of voxels in question.

Intuitively, when dealing with across-subject correlations, the problem arises because the threshold applied to filter out significantly correlated voxels imposes a minimum value on the admissible correlation. The correlation measured in any one voxel will necessarily be some combination of the actual signal (some correlation that will be observed in that voxel across measurements) and a contribution of noise. Sometimes the noise will decrease the measured correlation from its true value, sometimes the noise will increase the apparent correlation. The measured correlations that pass the significance threshold will disproportionately reflect those voxels where the noise happened to have increased the measured correlation (enough to pass the significance threshold). Thus, the correlations estimated only in those voxels that passed the significance threshold will systematically overestimate the underlying correlation by an unknown amount. This overestimation cannot be corrected for because its magnitude is unknown unless we independently estimate the true correlation value; thus correlations estimated in this manner are effectively uninterpretable.

The bitter irony of this phenomenon is that correcting for multiple comparisons when selecting voxels that carry signal *exacerbates* the overestimation of the magnitude of signal (Fig. 1). The multiple comparisons correction appropriately lowers the probability of falsely reporting that signal exists when no signal is present; however, it does not cure the overestimation. Instead, it raises the filtering criterion to identify voxels with meaningful signal; consequently, the more conservative the multiple comparisons correction (conservative in the sense of minimizing false alarms), the greater the overestimation of the strength of signals.

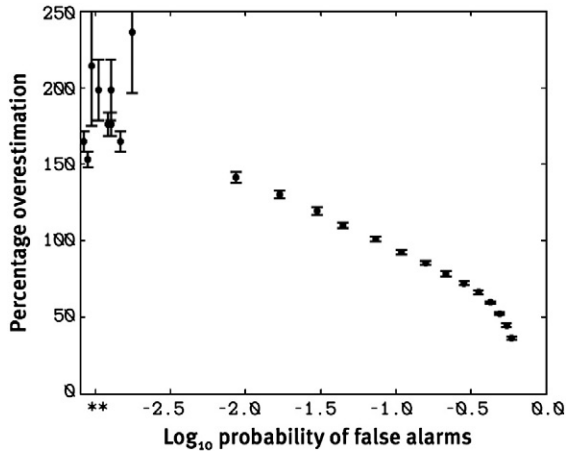

Fig. 1. Overestimation of signals as a function of false alarm rates. Higher thresholds – like those used for multiple comparisons correction – yield lower probabilities of false alarms (defined as false discover rate); and result in larger overestimation of effect sizes. This figure is modified from that in the Appendix of (Vul et al., 2009a), obtained by simulating a smooth 2D array of voxels. Asterisks on the x axis indicate simulations with no false alarms. The figure shows the qualitative effect of overestimation increasing as thresholds are increased to correct for multiple comparisons: the quantitative effects for any given study will depend on the true effect size, the prevalence of signals in the search volume, and the variability of the noise.

Variants of this problem seem to arise in every field that takes on the considerable challenge of identifying and quantifying signals found in massively multivariate data, where one cannot ascertain in advance where the signals of interest may lie. In psychometrics, Edward Cureton (Cureton, 1950) showed that when researchers use the same data to identify which test items to use in constructing a measure designed to predict a relevant behavioral outcome and also to measure the predictive power of the measure, the estimations of predictive power end up being "Baloney", as Cureton termed it. In finance, Lo and MacKinlay (Lo and MacKinlay, 1990) found that the practice of "data snooping" – using the same data to group assets into portfolios and then test asset pricing models on those portfolios – "no longer reject the null hypothesis". In epidemiology, Feinstein (Feinstein, 1988) reported that the common "data dredging" procedure by which risk factors are identified in large population surveys involves no a priori hypotheses about which factors ("diet, smoking, alcohol," etc.) will yield which outcomes ("birth defects, stroke, heart disease, cancer, death") in which demographics ("age, race, sex, socioeconomic status"); consequently, such surveys will often falsely report that some everyday behaviors are either menaces or boons to public health because they were selected and validated with the same data (see also Smith and Ebrahim, 2002). Similarly, in medical gene sequencing, Michiels, Koscielny, and Hill (Michiels et al., 2005) report that the data mining procedures used to find genetic associations with cancer outcomes from the thousands of genes available in microarray data yielded "highly unstable" results which often "did not classify [new] patients better than chance" (see also Hunter and Kraft, 2007). Ioannides (Ioannidis, 2005) summarized which fields are at greater risk for these problems: "The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true… Fields [with a] wealth of assembled and tested information, such as microarrays and other high-throughput discovery-oriented research, should have extremely low positive predictive value."

Brain mapping suffers from these risks: a whole-brain fMRI scan will yield thousands of different voxels, and for novel tasks it is difficult to specify a priori which voxels ought to show task related signal. Across-subject whole-brain correlation studies of social behavior and personality also suffer from low power due to the small numbers of subjects used in a typical experiment (Yarkoni, 2009) and low expected effect sizes given the unreliability of personality measures and fMRI data (Vul et al., 2009a). So it should be no surprise that about half of the papers in these fields report analyses that include some version of the *data-snooping/data-dredging/non-independence/circularity/double dipping* problem (Kriegeskorte et al., 2009; Vul et al., 2009a). So how serious is this problem? And how can it be avoided?

How serious is the problem?

The gravity of such selection biases in reported fMRI results varies considerably, from simple misleading plots of reported effects, to overestimated effect sizes, to significant results arising potentially out of noise (Vul and Kanwisher, 2010). Our impression is that the frequency of non-independence problems drops off with severity, and varies across subfields. Across fields the rate of any kind of circularity is between 40% and 56% (Kriegeskorte et al., 2009), and methods guaranteed to produce effect size overestimation were used in slightly over half of the studies of social behavior surveyed (Vul et al., 2009a). A more egregious problem arises when voxels are selected without appropriate multiple comparisons correction, and then the signal in those voxels is evaluated using the same

data. This procedure is likely to result not merely in overestimation of effects, but also in completely spurious correlations. Fortunately, our impression is that this aggravated form of the error has been relatively uncommon in the literature.[1]

Some authors argued that as long as false alarm rates are kept in check, the overestimation of effect sizes is not a serious problem (Nichols and Poline, 2009). Statistical analysis without effect size harks back to the period in the 1950s and 1960s when experimental psychologists often published tables of ANOVA results (F and p values) without disclosing means or measures of variability. In our view, such practices are completely wrongheaded, for reasons that have been pointed out by many statisticians (e.g., Wilkinson and Inference, 1999). They are also quite unnecessary — since valid effect size measures can be readily obtained.

How can this problem be avoided?

Fortunately, all variations of the circularity error – whether simple misrepresentation of data in graphs, overestimation of effects, or entirely spurious results – can be easily avoided by using independent data to identify signal-carrying voxels, and reporting estimates of the signal in those voxels. Most generally, this can be done via cross-validation: using one part of the data for signal identification, and a different part of the data for signal estimation. A variety of detailed proposals about how to circumvent such errors have been described in a number of recent articles (Kriegeskorte et al., 2009, 2010; Vul and Kanwisher, 2010; Vul et al., 2009a, 2009b).

Reactions to the paper

The immediate reaction to our paper was marked by great expressions of upset and hurt feelings by some of the neuroimaging investigators whose methods and results we had challenged — with particular indignation focused on the humorous title we had initially selected for our article ("Voodoo Correlations") (Lieberman et al., 2009). In hindsight, it seems possible that we might have effectively communicated our points with a more neutrally worded title, but we have some doubts about this. More generally, while the disagreement about our statistical argument has subsided, some people still question whether our paper has had a net positive effect on fMRI research. While we cannot possibly evaluate the total sociological impact of our paper, we consider its impact on the statistical practices in fMRI to be of greater significance.

In following the neuroimaging literature since our paper was published, we have been struck by three things — one very encouraging, the other two much less so:

First, it appears that the biased correlation measures that we described are now quite rare, and when mentioned in talks, are regularly accompanied by appropriate disclaimers. We have noticed a few likely exceptions in high profile journals, but the overall incidence of the non-independence error has clearly diminished. Reference to cross-validation in the neuroimaging literature now seems more common than it did before 2009. It would be interesting in a few years to compute a histogram of brain-behavior correlation magnitudes appearing subsequent to our paper, to see whether correlations above, say, .6 now appear as rare as we suggested they ought to be.

Second, on a less cheerful note: since most of the raw data in the field still exist, our paper had also advocated re-analyses using appropriate methods. Such reports have been exceedingly rare (the only one we are aware of is by Poldrack and Mumford, 2009). Perhaps this should not be surprising: even though the scientific enterprise would obviously be served by corrections, the incentives for individual researchers are aligned against reanalyzing and correcting their previously reported findings.

Third, a number of high-profile papers have revealed other common errors in fMRI analyses that are pervasive enough to undermine many results in the literature. These include the "Dead Salmon" paper pointing out that some of the multiple comparisons correction procedures used in the literature are insufficient and yield higher-than-advertised false alarm rates (and apparent social cognition in a dead salmon) (Bennett et al., 2010). Another recent paper points out that, the difference between "significant" and "not significant" is not itself statistically significant; a statistical fact that jeopardizes a number of findings that interpret different patterns of activation between groups without explicitly testing the interaction (Nieuwenhuis et al., 2011). Related interpretational issues have been raised previously, resulting in some positive change in fMRI practice (see for instance: Henson, 2005; Poldrack, 2006).

Fortunately, non-independence, inadequate multiple comparisons correction, and not testing an interaction when that's the required analysis are easy problems: there are plenty of methods available to overcome them, and we are optimistic that despite the discomfort caused by the papers that point these problems out, they will be resolved, and the literature as a whole will be improved. This leads us to some harder problems, where our suggestions must of necessity be more tentative.

Considerations for the future

The discussion about circular analyses has revealed a much more vexing problem for fMRI analysis, one that lies at or outside the capabilities of current statistical science. One not-infrequent response to the evidence of inflated effect size estimates has been the suggestion – mentioned above – that the strength of the signal in a given voxel is not of import, because the primary goal of brain imaging research should be to identify the *location* of signals. While there is disagreement as to whether or not the effect size should be of primary interest (our view was described above; for a spectrum of opinions see the responses to issue 5 in Kriegeskorte et al., 2010) the claim that localization is the only important goal draws attention to the thorny statistical challenges associated with localization. We believe that improved methods to deal with these localization challenges will prove to be a most useful methodological development in the field in the coming future.

To see why this matter is so thorny, imagine a particular task contrast reveals a cluster centered at MNI coordinates (−12,32,38); in another experiment, suppose the same task contrast activated a cluster centered at coordinates (−10, 30, 35). Does the second result constitute a successful replication of the first?

In practice, this question is usually answered by drawing on assumptions about which anatomical features of the location are relevant (and opinions about this matter will typically vary across researchers). There are no good alternatives to such subjective evaluation because appropriate statistical techniques to answer this question do not exist. To progress beyond this unsatisfactory state of affairs, we need statistical methods that can answer two critical questions.

First: How do we characterize the activation in a given subject? Common practice includes reporting the peak/center voxel, and displaying a statistically thresholded image; however, this belies the fact that task activations tend not to be punctate sources, but are rather diffuse regions of varying signal intensity — and with sufficient

[1] Lieberman and Cunningham (2009) acknowledged these errors, but rather than re-analyzing prior data, they argued that neuroimaging research might be better off if researchers were less concerned about false alarms than they currently are!

statistical power, more and more areas of the brain will be revealed to be significantly activated by a task (Yarkoni and Braver, 2010). So the activation of the brain to a task is not a thresholded region, but a continuous map. How, then, can we summarize these activation maps in a useful manner without throwing out a vast amount of important information?

Second: Once we find a useful summary of activation maps, how do we characterize the variability of these summaries across subjects while respecting the variability in anatomy and the function-to-anatomy mapping across subjects.

Both of these challenges reveal unanswered questions in spatial statistics, and unfortunately answers to both are required to properly determine whether one task activation replicates another, or to address a host of related, fundamental questions about the neural bases of task-specific processes. Some initial attempts have been made to address some of these challenges (Fedorenko et al., 2011; Xu et al., 2009; Yarkoni et al., 2011), but much remains to be done. We are optimistic that with input from the statistical community, useful methods for addressing these problems will be developed.

In closing: Massively multivariate data with a low signal-to-noise ratio is inherently difficult to analyze and invites mischaracterization. Substantial progress has already been made in developing techniques for overcoming some of these difficulties. Although the recent rounds of criticism of dubious practices in fMRI data analysis has been viewed by some as a bit embarrassing to the field of cognitive neuroscience, we are confident it has already benefited the scientific integrity of the field and in all likelihood, further improvements are to be expected.

References

Baker, C.I., Hutchison, T.L., Kanwisher, N., 2007. Does the fusiform face area contain subregions highly selective for nonfaces? Nat. Neurosci. 10 (1), 3–4.

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. JSUR 1 (1), 1–5.

Cureton, E.E., 1950. Validity, reliability, and baloney. Educ. Psychol. Meas. 10, 94–96.

Egger, M., Davey Smith, G., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. BMJ 315 (7109), 629–634.

Fedorenko, E., Behr, M., Kanwisher, N., 2011. Functional specificity for high-level linguistic processing in the human brain. Proc. Natl. Acad. Sci. U. S. A. 108, 16428–16433.

Feinstein, A.R., 1988. Scientific standards in epidemiologic studies of the menace of daily life. Science 242 (4883), 1257–1263.

Grill-Spector, K., Sayres, R., Ress, D., 2006. High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. Nat. Neurosci. 9 (9), 1177–1185.

Henson, R., 2005. What can functional neuroimaging tell the experimental psychologist? Q. J. Exp. Psychol. A 58 (2), 193–233.

Hunter, D.J., Kraft, P., 2007. Drinking from the Fire Hose - Statistical issues in genome-wide association studies. N. Engl. J. Med. 357, 436–439.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2 (8), 0696–0701.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.

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. J. Cereb. Blood Flow Metab. 30 (9), 1551–1557.

Lieberman, M.D., Cunningham, W.A., 2009. Type I and Type II error concerns in fMRI research: Re-balancing the scale. Soc. Cogn. Affect. Neurosci. 4, 423–428.

Lieberman, M.D., Berkman, E.T., Wager, T.D., 2009. Correlations in social neuroscience aren't voodoo: commentary on Vul et al. Perspect. Psychol. Sci. 4 (3).

Lo, A.W., MacKinlay, A.C., 1990. Data-Snooping biases in tests of financial asset pricing models. Rev. Financ. Stat. 3 (3), 431–467.

Michiels, S., Koscielny, S., Hill, C., 2005. Prediction of cancer outcome with microarrays: a multiple random validation study. Lancet 365 (9458), 488–492.

Nichols, T.E., Poline, J.-P., 2009. Commentary on Vul et al.'s (2009) "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition". Perspect. Psychol. Sci. 4 (3).

Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat. Neurosci. 14 (9), 1105–1107.

Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10 (2), 59–63.

Poldrack, R.A., Mumford, J.A., 2009. Independence in ROI analysis: where is the voodoo? Soc. Cogn. Affect. Neurosci. 4 (2), 208–213.

Smith, G.W., Ebrahim, S., 2002. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. BMJ 325 (1437).

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

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009a. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. Perspect. Psychol. Sci. 4 (3), 274–290.

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". Perspect. Psychol. Sci. 4 (3), 319–324.

Wilkinson, L., Inference,, T.T.F.o.S., 1999. Statistical methods in psychology journals: Guidelines and explanations. Am. Psychol. 54, 594–604.

Xu, L., Johnson, T.D., Nichols, T.E., Nee, D.E., 2009. Modeling inter-subject variability in fMRI activation location: A Bayesian Hierarchical spatial model. Biometrics 65 (4), 10410–10451.

Yarkoni, T., 2009. Big correlations in little studies: Inflated fMRI correlations reflect low statistical power. Perspect. Psychol. Sci. 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: Gruszka, M., Szymura (Eds.), Handbook of individual differences in cognition.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670.