bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Scanning the Horizon: Future challenges for neuroimaging research.

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1], Krzysztof J. Gorgolewski[1], Paul M. Matthews[3], Marcus Munafò[4,5], Thomas E. Nichols[6], Jean-Baptiste Poline[7], Edward Vul[8], Tal Yarkoni[9]

Affiliations:

1. Department of Psychology and Stanford Center for Reproducible Neuroscience, Stanford University
2. Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health
3. Division of Brain Sciences, Department of Medicine, Hammersmith Hospital, London, UK
4. MRC Integrative Epidemiology Unit at the University of Bristol
5. UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol
6. WMG & Department of Statistics, University of Warwick, Coventry, UK
7. Helen Wills Neuroscience Institute, Brain Imaging Center, University of California, Berkeley, CA, USA
8. Department of Psychology, University of California, San Diego
9. Department of Psychology, University of Texas at Austin

Abstract

Neuroimaging has transformed our ability to probe the neurobiological basis of behaviour. As neuroimaging technologies become more widely available, they are increasingly being applied by the wider neuroscience community. However, concerns have recently been raised that the conclusions drawn from many human neuroimaging studies are either spurious or not generalizable. Problems such as low statistical power, analytical flexibility, and lack of direct replication apply to many fields, but perhaps particularly to neuroimaging. In this *Opinion* article, we discuss these problems, outline current and suggested best practices to improve the quality and reproducibility of neuroimaging findings, and describe how we think the field should evolve if it is to produce the most meaningful answers to current and future research questions.

Main text

“With great power there must also come–great responsibility!” - Stan Lee, *Spiderman*

Neuroimaging, particularly using functional magnetic resonance imaging (fMRI), has become the primary tool of human neuroscience [1], and recent advances in the acquisition and analysis of fMRI data have provided increasingly powerful means to dissect brain function. These advances promise to offer important insights into the workings of the human brain, but also

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

generate the potential for a “perfect storm” of irreproducible results based on questionable research practices (QRPs). In the present review we will outline these potential problems and offer practical solutions. While we focus on fMRI, the problems outlined here are general to any field that deals in similarly large and complex datasets; many of our solutions are adapted from other fields (such as genetics) that have already dealt with some of these issues (see Box 1).

Recent years have seen intense interest in the degree to which widespread QRPs are responsible for high rates of false findings in the scientific literature[2–4]. There is growing interest in “meta-research”[5], and a corresponding growth in studies investigating factors that contribute to poor reproducibility. These factors include study design characteristics which may introduce bias, low statistical power, and flexibility in data collection, analysis, and reporting — termed “researcher degrees of freedom” by Simmons and colleagues[3]. There is clearly concern that these issues may be undermining the value of science – in the UK, the Academy of Medical Sciences recently convened a joint meeting with a number of other funders to explore these issues, while in the US the National Institutes of Health has an ongoing initiative to improve research reproducibility[6].

Perhaps one of the most surprising findings in recent work is the lack of appreciation of the QRP problem by researchers. John and colleagues[7] polled psychology researchers to determine the rate of QRPs, and asked them to rate the defensibility of a number of QRPs on a scale of 0 (indefensible) to 2 (defensible). These researchers gave surprisingly high defensibility ratings to such clearly problematic practices as stopping data collection once a desired result is found (mean rating = 1.76), reporting unexpected results as having been predicted *a priori* (mean = 1.5), and deciding whether to exclude data after looking at the effects of doing so (mean = 1.61). These results suggest that there remains a substantial need for consciousness-raising amongst researchers regarding QRPs.

Problems and solutions in neuroimaging research

Below we outline a number of potentially problematic research practices in neuroimaging, which can lead to increased risk of false or exaggerated results. For each problem, we propose a set of solutions. Most of these are in principle uncontroversial, but in reality we find that best practices are often not followed. Many of these solutions arise from the experience of other fields (such as genetics) with similar problems.

Problem: Statistical power

The analyses of Button and colleagues[8] provided a wake-up call regarding statistical power in neuroscience, particularly by highlighting the point (raised earlier by Ioannidis[2]) that low power not only reduces the likelihood of finding a true result if it exists, but also raises the likelihood that any positive result is false, as well as causing substantial inflation of observed positive effect sizes[9]. In the context of neuroimaging, Button and colleagues considered only structural MRI studies, but the situation in fMRI is no better. Figure 1 presents an analysis of sample sizes

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

and the resulting statistical power over the last 20 years. Table S1 gives intuition about observed effect sizes, demonstrating that realistic effect sizes in fMRI are quite small, and thus the average fMRI study remains very poorly powered for realistic effects. However, one “bright side” of this analysis is the rapidly increasing number of recent studies with large samples (greater than 100), suggesting that the field may be progressing towards adequately-powered research.

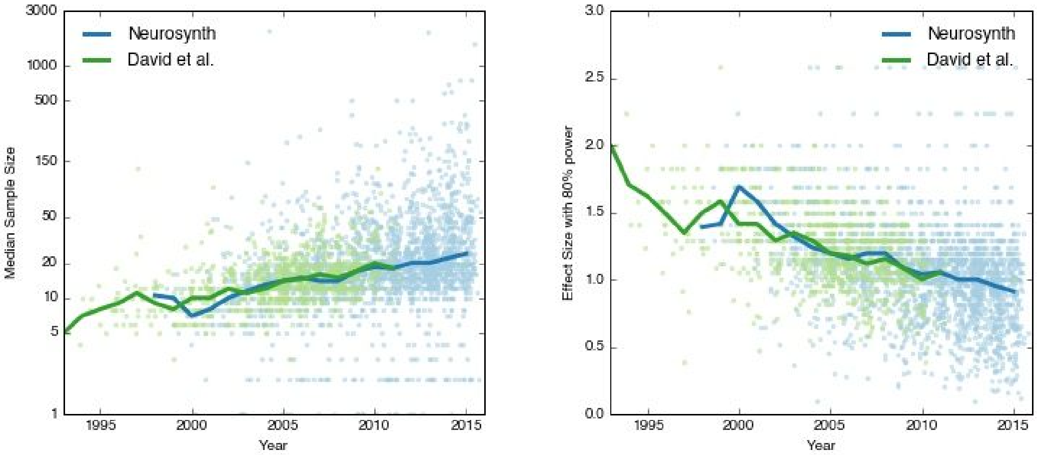

Figure 1: Sample size estimates and estimated power for fMRI studies. Left panel: Sample sizes over years obtained from two sources: manual extraction from published meta-analyses[10], and automated extraction from the Neurosynth database[11]. These data demonstrate that despite a steady increase in sample size, median sample size remained below 25 as of 2015. Right panel: Using the sample sizes from the left panel, we estimated the standardized effect size required to detect an effect with 80% power for a whole-brain linear mixed effects analysis using a voxelwise 5% familywise error rate threshold from random field theory[12] (see Supplementary Methods for details). This shows that even today, the average study is only sufficiently powered to detect very large effects; given that many of the studies will be assessing group differences or brain-behavior correlations (which will inherently have lower power), this represents an optimistic lower bound on the powered effect size. See Supplementary Materials for additional analyses. Data and code to generate these figures are available at http://nbviewer.jupyter.org/github/poldracklab/power/blob/master/Fig_power/fig_power.ipynb .

Solutions:

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

When possible, all sample sizes should be justified by an *a priori* power analysis. A number of tools are available to enable power analyses for fMRI (e.g., neuropowertools.org[13], fmripower.org[14]). When previous data are not available to support a power analysis, one can instead identify the sample size that would support finding the minimum effect size that one would find theoretically informative (see Supplementary Table 1 for example effect sizes). The use of heuristic sample size guidelines (e.g., "our sample size was based on similar published studies") is bound to result in a misuse of resources, either by collecting too many or (more likely) too few subjects.

In some cases, researchers must perform a study with an insufficient sample size, either due to resources limitations or to limitations in the specific sample (e.g., a rare patient group). In this case there are two commonly used options to improve power:

1. Engage in a consortium with other researchers in order to combine data. This approach has been highly successful in the field of genetics, where well-powered genome-wide analyses require samples far beyond the ability of any individual laboratory (see Box 1). Examples of successful consortia in neuroimaging include the 1000 Functional Connectomes/INDI project[15] and the ENIGMA consortium[16].
2. Restrict the search space using a small number of *a priori* regions of interest or an independent "functional localizer" to identify specific regions of interest for each individual. It is essential that these regions of interest (or a specific functional localizer strategy) be explicitly defined *prior* to any analyses; it is always possible to develop a *post hoc* justification for any specific ROI based on previously published papers, which results in an ROI that appears independent but is actually circular and thus results in meaningless statistics and inflated Type I errors. By analogy to the idea of HARKing (Hypothesing After Results are Known)[17], we refer to this as SHARKing ("Selecting Hypothesized Areas after Results Known"). We would only recommend the use of restricted search spaces if the exact regions of interest and hypotheses are pre-registered[18,19].

Problem: Analytic flexibility

The standard fMRI analysis workflow contains a large number of preprocessing and analysis operations, each with choices to be made about parameters and/or methods (see Box 2). Carp[20] applied more than 6,000 analysis workflows to a single dataset and quantified the resulting variability in statistical maps, showing that some regions exhibited substantial variation across the different workflows. This issue is not unique to fMRI; for example, similar issues have been raised in genetics[21]. These "researcher degrees of freedom" can lead to substantial inflation of Type I error rates[22], even when there is no intentional p-hacking[4].

Exploration is key to scientific discovery, but rarely does a research paper comprehensively describe the actual process of exploration that led to the ultimate result; to do so would render the resulting narrative far too complex and murky. As a clean and simple narrative has become

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

an essential component of publication, the intellectual journey of the research is often obscured. Instead, reports often engage in “hypothesizing after the results are known”, or HARKing[17], in which the results of exploratory analyses are presented as having been hypothesized from the beginning. Because HARKing hides the number of data-driven choices made during analysis, it can strongly overstate the actual evidence for a hypothesis. We would argue that there is a great need to support the publication of exploratory studies without forcing those studies to masquerade as hypothesis-driven science, while at the same time realizing that such exploratory findings will ultimately require validation in independent studies.

Solutions:

- We recommend pre-registration of methods and analysis plans as a default. This should include planned sample size, specific analysis tools to be used, specification of predicted outcomes, and definition of any ROIs that will be used for analysis.
- Exploratory analyses (including any deviations from planned analyses) should be clearly distinguished from planned analyses in the study description. Ideally, results from exploratory analyses should be confirmed in an independent validation dataset.

Problem: Multiple comparisons

The most common approach to neuroimaging analysis involves “mass univariate” testing of one hypothesis for each voxel, which implies that the false positive rate will be inflated if there is no correction for multiple tests. A humorous example of this was seen in the now-infamous “dead salmon” study reported by Bennett and colleagues[23], in which “activation” was detected in the brain of a dead salmon (which disappeared when the proper corrections for multiple comparisons were performed). Figure 2 presents an example in which completely random data lead to seemingly impressive results, through a combination of failure to adequately correct for multiple comparisons and circular ROI analysis. The problem of multiplicity was recognized very early, and the last 25 years have seen the development of well-established and validated methods for correction of familywise error and false discovery rate in neuroimaging data[24]. However, recent work[25] has suggested that even some very well-established methods for inference based on spatial extent of activations can produce inflated error rates (also see below under ***Software Errors***).

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

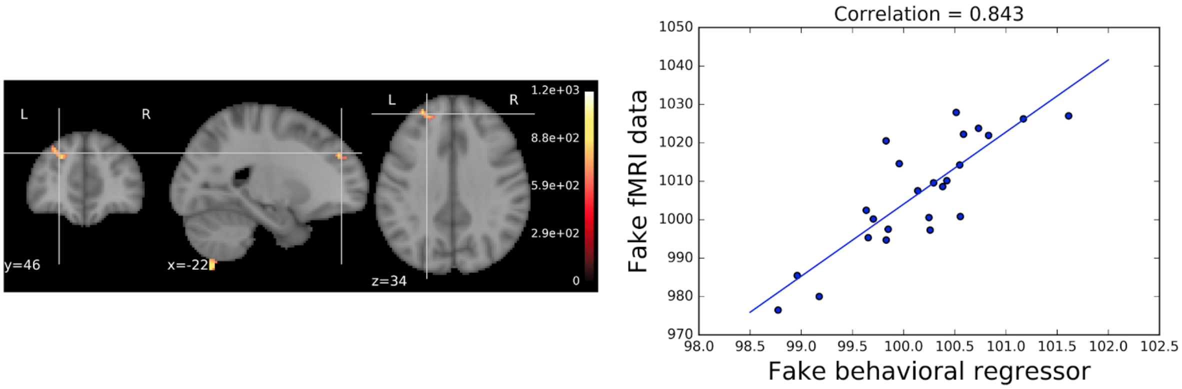

Figure 2. Seemingly impressive brain-behavior association arising from completely random data through the use of uncorrected statistics and circular ROI analysis to capitalize on the large sampling error arising from small samples (see Supplementary Methods for details). The analysis revealed a cluster in the lateral prefrontal cortex (left panel); signal extracted from that cluster (i.e., using circular analysis) showed a very strong correlation between brain and behavior (right panel; r = 0.84). A computational notebook for this example can be viewed at http://nbviewer.jupyter.org/github/poldrack/corrsim/blob/master/Correlation_simulation.ipynb.

There is an ongoing debate between neuroimaging researchers who feel that conventional approaches to multiple comparison correction are too lax and allow too many false positives[26], and those who feel that thresholds are too conservative, and risk missing most of the interesting effects[27]. In our view, the deeper problem is the inconsistent application of principled correction approaches[28]. Many researchers freely combine different approaches and thresholds in ways that produce a high number of undocumented "researcher degrees of freedom"[22], rendering reported p-values uninterpretable. To assess this more directly, we examined the top 100 results for the Pubmed query ("fMRI" AND brain AND activation NOT review[PT] AND human[MESH] AND english[la]), performed May 23, 2016; of these, 65 reported whole-brain task fMRI results and were available in full text. Only three presented fully uncorrected results, with four others presenting a mixture of corrected and uncorrected results; this suggests that corrections for multiple comparisons are now standard. However, there is evidence that researchers may be "method-shopping" for techniques that provide greater sensitivity, at a potential cost of increased error rates. Nine of the 65 papers used FSL or SPM to perform their primary analysis, but then used the AFNI alphasim/3dClustSim tool (7 papers) or other simulation-based approaches (2 papers) to correct for multiple comparisons. This is concerning, because both FSL and SPM offer well-established methods for correction of multiple comparisons using Gaussian random field theory. Given the substantial degree of extra work involved in using multiple software packages, the use of a different tool raises some concern that this might reflect analytic p-hacking. This concern is further amplified by the finding that until very recently, this AFNI program had substantially inflated Type I error rates due to a software bug[25]. Distressingly, whereas nonparametric methods are known to provide the most accurate control over familywise error rates[24,25,29], they were not used in any of these papers.

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Solutions:

- To balance Type I and Type II error rates in a principled way, we suggest a dual approach of reporting FWE-corrected whole-brain results, and sharing a copy of the unthresholded statistical map via a repository that allows viewing and downloading (e.g., Neurovault.org [30]). For an example of this practice, see ref[31] and shared data at http://neurovault.org/collections/122/.
- Any use of non-standard methods for correction of multiple comparisons (e.g., using an AFNI tool when other analyses were performed using SPM) should be justified explicitly (and reviewers should demand such justification).

Problem: Software errors

Most fMRI researchers use one of several open-source analysis packages for preprocessing and statistical analyses; many additional analyses require custom programs. Because most researchers are not trained in software engineering, there is insufficient attention to good software development practices that could help catch and prevent errors. This issue came to the fore recently, when a 15-year-old bug in the AFNI 3dClustSim program was discovered, which resulted in substantially inflated Type I error rates[25] (the bug was fixed in May 2015). The impact of such bugs is substantial; PubMed Central lists 1525 publications mentioning AlphaSim or 3dClustSim published before the bug was fixed.

Solutions:

- Whenever possible, software tools from a well-established project should be used instead of custom code. Errors are more likely to be discovered when the code is used by a larger group, and larger projects are more likely to follow better software development practices.
- Researchers should learn and implement defensive programming practices, including the judicious use of software testing and validation. Validation methodologies should be clearly defined.
- Custom analysis codes should always be shared upon manuscript submission (for an example, see[32]), and code should be reviewed as part of the scientific review process. Reviewers should request access to code when it is important for evaluation purposes.

Problem: Insufficient study reporting

Eight years ago we[33] published an initial set of guidelines for reporting the methods used in an fMRI study. Unfortunately, reporting standards in the fMRI literature remain poor. Carp[34] and Guo and colleagues[35] analyzed a large number of fMRI papers for the reporting of methodological details, and both found that many important analysis details were rarely described. Consistent with this, in 22 of the 65 papers discussed above it was impossible to identify exactly which correction technique was used (beyond generic terms such as “cluster-based correction”) because no specific method or citation was provided. The

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Organization for Human Brain Mapping has recently addressed this issue through its Committee on Best Practices in Data Analysis and Sharing (COBIDAS), which has issued a new, detailed set of reporting guidelines[36] (http://www.humanbrainmapping.org/COBIDAS).

Beyond the description of methods, claims in the neuroimaging literature are often advanced without corresponding statistical support. In particular, failures to observe a significant effect often lead researchers to proclaim the absence of an effect—a dangerous and almost invariably unsupported acceptance of the null hypothesis. “Reverse inference” claims, in which the presence a given pattern of brain activity is taken to imply a specific cognitive process, are rarely grounded in quantitative evidence[11,37]. Claims of “selective” activation in one brain region or experimental condition are often made when activation is statistically significant in one region or condition but not in others—ignoring the fact that “the difference between significant and non-significant is not itself significant”[38], and absent appropriate tests for statistical interactions[39].

Solutions:

- Authors should follow accepted standards for methods reporting (e.g., the COBIDAS standard for MRI studies), and journals should require adherence to these standards.
- Every major claim in a paper should be directly supported by appropriate statistical evidence, including specific tests for significance across conditions and relevant tests for interactions.

Problem: Lack of independent replications

There are surprisingly few examples of direct replication in the field of neuroimaging, likely reflecting both the expense of fMRI studies along with the emphasis of most top journals on novelty rather than informativeness. One study attempted to replicate 17 studies of association between brain structure and behavior; only one of the 17 studies showed stronger evidence for the original effect size than for a null effect, and 8/17 showed stronger evidence for a null effect[40,41]. This suggests that replicability of neuroimaging findings (particularly brain-behavior correlations) may be exceedingly low, similar to recent findings in other areas of science such as cancer biology[42] and psychology[43].

Solutions:

- The neuroimaging community should acknowledge replication reports as scientifically important research outcomes that are essential in advancing knowledge. One such attempt is the upcoming OHBM Replication Award for the best neuroimaging replication study.

Conclusion

We have outlined what we see as a set of problems with neuroimaging methodology and reporting, and solutions to solve them. It is likely that the reproducibility of neuroimaging

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

research is no better than many other fields, where it has been shown to be surprisingly low. Given the substantial amount of research funds currently invested in neuroimaging research, we believe that it is essential that the field address the issues raised here, lest it suffer a backlash that could badly impact on future support for research in this area.

Text Boxes:

Box 1: Lessons from Genetics

The study of genetic influences on complex traits has been transformed by the advent of whole genome methods, and the subsequent use of stringent statistical criteria, independent replication, large collaborative consortia, and complete reporting of statistical results. Previously, “candidate” genes would be selected on the basis of known or presumed biology, and a handful of variants genotyped (many of which would go unreported) and tested in small studies (typically in the low 100s). An enormous literature proliferated, but these findings generally failed to replicate[44]. The transformation brought about by whole genome methods (i.e., genome wide association studies) was partly necessitated by the simultaneous testing of several hundred thousand genetic loci (hence the need for very stringent statistical criteria in order to reach “genome wide significance”), but also an awareness that any effects of common genetic variants would almost certainly be very small (<1% phenotypic variance). The combination of these factors required very large sample sizes, in turn necessitating large-scale collaboration and data sharing. The resulting cultural shift in best practice has transformed our understanding of the genetic architecture of complex traits, and in a few years produced many hundred more reproducible findings than in the previous fifteen years. Routine sharing of SNP-level statistical results has facilitated routine use of meta-analyses, as well as the development of novel methods of secondary analysis[45].

This relatively rosy picture contrasts markedly with the situation in “imaging genomics”--a burgeoning field that has yet to embrace the standards commonly followed in the genetics literature, and remains largely focused on individual candidate gene association studies characterized by numerous researcher degrees of freedom. To illustrate, we examined the first 50 abstracts matching a PubMed search for “fMRI” and “genetics” (excluding reviews, studies of genetic disorders, and nonhuman studies) that included a genetic association analysis. Of these, the vast majority (43/50) reported analysis of a single or small number of candidate genes; only two reported a genome-wide analysis, with the rest reporting analyses using biologically inspired gene sets or polygenic risk scores. Recent empirical evidence casts doubt on the validity of candidate gene associations. A large genome-wide association study of brain structure (including whole-brain and hippocampal volume) identified two genetic associations that were replicated across two large samples of more than 10,000 individuals. Strikingly, analysis of a set of previously identified candidate genes showed no evidence for any association in this very well-powered study [46].

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Box 2: Analytic flexibility in fMRI

In the early days of fMRI analysis, it was rare to find two laboratories that used the same software to analyze their data, with most using locally-developed custom software. Over time, a small number of open-source analysis packages have gained prominence (SPM, FSL, and AFNI being the most common), and now most laboratories use one of these packages for their primary data processing and analysis. Within each of these packages, there is a great deal of flexibility in how data are analyzed; in some cases there are clear best practices, but in other cases there is no consensus regarding the optimal approach. This leads to a multiplicity of analysis options. In Table B1 we outline some of the major choices involved in performing analyses using one of the common software packages (FSL). Even for this non-exhaustive list from a single analysis package, the number of possible analysis workflows exceeds the number of papers that have been published on fMRI since its inception more than two decades ago!

It is possible that many of these alternative pipelines could lead to very similar results, though the analyses of Carp[20] suggest that many of them may lead to significant heterogeneity in the results. In addition, there is evidence that choices of preprocessing parameters may interact with the statistical modeling approach, and that the optimal preprocessing pipeline may differ across subjects[47].

Table B1: A non-exhaustive list of data processing/analysis options available within the FSL software package, enumerating a total of 69,120 different possible workflows.

Processing step	**Reason**	**Options**	**Number of plausible options**
Motion correction	Correct for head motion during scanning	Interpolation [linear vs. sinc] Reference volume [single vs. mean]	4
Slice timing correction	Correct for differences in acquisition timing of different slices	No/before motion correction/after motion correction	3
Field map correction	Correct for distortion due to magnetic susceptibility	Yes/No	2
Spatial smoothing	Increase SNR for larger activations and ensure assumptions of Gaussian random field theory	FWHM [4/6/8 mm]	3

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Spatial normalization	Warp individual brain to match a group template	Method [linear/nonlinear]	2
High pass filter	Remove low-frequency nuisance signals from data	Frequency cutoff [100, 120]	2
Head motion regressors	Remove remaining signals due to head motion via statistical model	Yes/No If Yes: 6/12/24 parameters or single timepoint “scrubbing” regressors	5
Hemodynamic response	Account for delayed nature of hemodynamic response to neuronal activity	Basis function [single-gamma, double-gamma] Derivatives [none/shift/dispersion]	6
Temporal autocorrelation model	Model for the temporal autocorrelation inherent in fMRI signals.	Yes/no	2
Multiple comparison correction	Correct for large number of comparisons across the brain	Voxel-based GRF, Cluster-based GRF, FDR, nonparameteric	4
Total possible workflows			**69,120**

Acknowledgements

RP, JD, JBP and KG are supported by the Laura and John Arnold Foundation. MRM is a member of the UK Centre for Tobacco and Alcohol Studies, a UK Clinical Research Council Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. Support from the Medical Research Council (MC UU 12013/6) is also gratefully acknowledged. CIB is supported by the Intramural Research Program of NIMH. TY is supported by NIMH (R01MH096906). PMM gratefully acknowledges personal support from the Edmond J. Safra Foundation and Lily Safra and research support from the Medical Research Council, the Imperial College Healthcare Trust Biomedical Research Centre and the Imperial EPSRC Mathematics in Healthcare Centre. TEN is supported by the Wellcome Trust (100309/Z/12/Z). Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Washington University. Thanks to Sean David for providing sample size data. JBP is supported by NIH-NIBIB P41-EB019936 and by NIH-NIDA U24-038653.

References

1. Poldrack, R. A. & Farah, M. J. Progress and challenges in probing the human brain. *Nature* **526,** 371–379 (2015).
2. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
3. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22,** 1359–1366 (2011).
4. Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* (2013).
5. Ioannidis, J. P. A., Fanelli, D., Dunne, D. D. & Goodman, S. N. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS Biol.* **13,** e1002264 (2015).
6. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505,** 612–613 (2014).
7. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23,** 524–532 (2012).
8. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).
9. Yarkoni, T. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* **4,** 294–298

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

(2009).

10. David, S. P. *et al.* Potential reporting bias in fMRI studies of the brain. *PLoS One* **8,** e70104 (2013).
11. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8,** 665–670 (2011).
12. Friston, K. J., Frith, C. D., Liddle, P. F. & Frackowiak, R. S. Comparing functional (PET) images: the assessment of significant change. *J. Cereb. Blood Flow Metab.* **11,** 690–699 (1991).
13. Durnez, J. *et al.* Power and sample size calculations for fMRI studies based on the prevalence of active peaks. *bioRxiv* 049429 (2016). doi:10.1101/049429
14. Mumford, J. A. & Nichols, T. E. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **39,** 261–268 (2008).
15. Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4734–4739 (2010).
16. Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* **8,** 153–182 (2014).
17. Kerr, N. L. HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2,** 196–217 (1998).
18. Nosek, B. A. *et al.* SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* **348,** 1422–1425 (2015).
19. Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. Registered reports: realigning incentives in scientific publishing. *Cortex* **66,** A1–2 (2015).
20. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

FMRI experiments. *Front. Neurosci.* **6,** 149 (2012).

21. Heininga, V. E., Oldehinkel, A. J., Veenstra, R. & Nederhof, E. I just ran a thousand analyses: benefits of multiple testing in understanding equivocal evidence on gene-environment interactions. *PLoS One* **10,** e0125383 (2015).
22. Simmons, J. P., Nelson, L. D. & Uri, S. False-Positive Psychology: The Way We Report Studies Privileges False Findings. *PsycEXTRA Dataset* doi:10.1037/e636412012-001
23. Bennett, C. M., Miller, M. B. & Wolford, G. L. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Neuroimage* **47,** S125 (2009).
24. Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12,** 419–446 (2003).
25. Eklund, A , Nichols, T E , Knutsson, K. Cluster failure: Why fMRI inferences for spatial extent have inflated false positive rates. *Proc. Natl. Acad. Sci. U. S. A.* (2016).
26. Wager, T. D., Lindquist, M. & Kaplan, L. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* **2,** 150–158 (2007).
27. Lieberman, M. D. & Cunningham, W. A. Type I and Type II error concerns in fMRI research: re-balancing the scale. *Soc. Cogn. Affect. Neurosci.* **4,** 423–428 (2009).
28. Bennett, C. M., Wolford, G. L. & Miller, M. B. The principled control of false positives in neuroimaging. *Soc. Cogn. Affect. Neurosci.* **4,** 417–422 (2009).
29. Hayasaka, S. & Nichols, T. E. Validating cluster size inference: random field and permutation methods. *Neuroimage* **20,** 2343–2356 (2003).
30. Gorgolewski, K. J. *et al.* NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* **9,** 8 (2015).
31. Hunt, L. T., Dolan, R. J. & Behrens, T. E. J. Hierarchical competitions subserving

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

multi-attribute choice. *Nat. Neurosci.* **17,** 1613–1622 (2014).

32. Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J. & Wagner, A. D. Frontoparietal representations of task context support the flexible control of goal-directed cognition. *J. Neurosci.* **34,** 10743–10755 (2014).

33. Poldrack, R. A. *et al.* Guidelines for reporting an fMRI study. *Neuroimage* **40,** 409–414 (2008).

34. Carp, J. & Joshua, C. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage* **63,** 289–300 (2012).

35. Guo, Q. *et al.* The Reporting of Observational Clinical Functional Magnetic Resonance Imaging Studies: A Systematic Review. *PLoS One* **9,** e94412 (2014).

36. Nichols, T. E. *et al.* Best Practices in Data Analysis and Sharing in Neuroimaging using MRI. *bioRxiv* 054262 (2016). doi:10.1101/054262

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

38. Gelman, A. & Stern, H. The Difference Between ‘Significant’ and ‘Not Significant’ is not Itself Statistically Significant. *Am. Stat.* **60,** 328–331 (2006).

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

40. Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66,** 115–133 (2015).

41. Boekel, W., Forstmann, B. U. & Wagenmakers, E.-J. Challenges in replicating brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015). *Cortex* **74,** 348–352 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/059188. this version posted June 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

42. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).

43. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).

44. Flint, J. & Munafò, M. R. Candidate and non-candidate genes in behavior genetics. *Curr. Opin. Neurobiol.* **23,** 57–61 (2013).

45. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30,** 543–552 (2015).

46. Stein, J. L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–561 (2012).

47. Churchill, N. W. *et al.* Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. *PLoS One* **7,** e31147 (2012).