**Scanning the horizon: future challenges for neuroimaging research.**

Russell A. Poldrack1, Chris I. Baker2, Joke Durnez1, Krzysztof J. Gorgolewski1, Paul M. Matthews3, Marcus Munafò4,5, Thomas E. Nichols6, Jean-Baptiste Poline7, Edward Vul8, Tal Yarkoni9

Affiliations:

1. Department of Psychology and Stanford Center for Reproducible Neuroscience, Stanford University, Stanford, CA 94305
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

Corresponding author: R.A.P. ([russpold@stanford.edu](mailto:russpold@stanford.edu)) **[Au:OK?]**

**Abstract**

Neuroimaging techniques have transformed our ability to probe the neurobiological basis of behaviour and 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, and describe how we think the field should evolve to produce the most meaningful answers to research questions.

**Main text**

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

Neuroimaging, particularly using functional MRI (fMRI), has become the primary tool of human neuroscience[1](https://paperpile.com/c/BVQBEO/gjzr), 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 generate the potential for a ‘perfect storm’ of irreproducible results based on questionable research practices (QRPs).

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](https://paperpile.com/c/BVQBEO/hul1+fC6o+9nrt). There is growing interest in ‘meta-research’[5](https://paperpile.com/c/BVQBEO/zO2a), and a corresponding growth in studies investigating factors that contribute to poor reproducibility. These factors include study design characteristics that may introduce bias, low statistical power and flexibility in data collection, analysis, and reporting — termed “researcher degrees of freedom” by Simmons and colleagues[3](https://paperpile.com/c/BVQBEO/fC6o). There is clearly concern that these issues may be undermining the value of science – in the United Kingdom, the Academy of Medical Sciences recently convened a joint meeting with a number of other funders to explore these issues, whereas in the United States the National Institutes of Health has an ongoing initiative to improve research reproducibility[6](https://paperpile.com/c/BVQBEO/xHD9).

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](https://paperpile.com/c/BVQBEO/MCeX) 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 raising the awareness of QRPs among researchers. **[Au:OK?]**

In this article we outline a number of potentially problematic neuroimaging research practices that can lead to increased risk of false or exaggerated results. For each problematic research practice, 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 with similar problems (particularly those dealing with similarly large and complex data sets, such as genetics; Box 1).

**Statistical power**

In their Analysis article[8](https://paperpile.com/c/BVQBEO/jku2), Button and colleagues[8](https://paperpile.com/c/BVQBEO/jku2) provided a wake-up call regarding statistical power in neuroscience. In particular, they highlight the point (raised earlier by Ioannidis[2](https://paperpile.com/c/BVQBEO/hul1)) 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](https://paperpile.com/c/BVQBEO/0QnQ). In the context of neuroimaging, Button and colleagues considered only structural MRI studies, but the situation in fMRI is no better.

Here, we present an analysis of sample sizes and the resulting statistical power of fMRI studies over the past 20 years. **[Au:OK?]** First, to gain a perspective of any changes in sample sizes over this time period, we obtained sample sizes from fMRI studies from two sources: manual extraction from published meta-analyses[10](https://paperpile.com/c/BVQBEO/JksF), and automated extraction from the Neurosynth database[11](https://paperpile.com/c/BVQBEO/ymU0) The data and code to reproduce Figure 1a are available here: . Figure 1a shows that, despite a steady increase in sample size over the past two decades, the median sample size for an fMRI study in 2015 remains below 25. However, one encouraging finding from this analysis is that the number of recent studies with large samples (greater than 100) is rapidly increasing, suggesting that the field may be progressing towards adequately powered research.

For each of the sample sizes shown in Figure 1a, we estimated the standardized effect size that would be required to detect an effect with 80% power for a whole-brain linear mixed-effects analysis using a voxelwise 5% familywise error (FWE) rate threshold from random field theory[12](https://paperpile.com/c/BVQBEO/rn03). These error rates are standard in neuroimaging and statistics in general. In other words, we found the minimum effect size (Cohen’s D) that would have been needed in each of these studies in order for the difference to be considered statistically significant, given the median sample size. **[Au:OK?]** The data and code to generate Figure 1b figures are available at <http://nbviewer.jupyter.org/github/poldracklab/power/blob/master/Fig_power/fig_power.ipynb>

To do this, we assumed that each study used a statistical map with T-values in an MNI (Montreal Neurological Institute) template with smoothness of three times the voxelsize, a commonly used value for smoothness in fMRI analysis. The MNI template is a freely available template, obtained from an average T1 scan for 152 subjects with a resolution of 2 millimeter, with a volume of 228483 voxels **[Au:OK? The volume needs to be included here - please specify units],** used by default in FSL (Oxford Centre for Functional MRI of the Brain Software Library) and SPM (Statistical Parametric Mapping; a suite of programmes available from MATLAB (MathWorks)) **[Au:OK?]**. We assumed that in each case there would be one active region, with voxelwise standardised effect size D. In other words, we assume that for each subject, all voxels in the active region are on average D standardised units higher than the voxels in the non-active region. In this demonstration, we assume the spatial extent of the active region to be relatively small so that there is only one local maximum in the region. Larger regions with multiple local maxima will decrease the detectable effect size.

To calculate the voxelwise statistical significance threshold for the active region in this model statistical map, we used the FSL function ptoz, which computes a FWE threshold for a given volume and smoothness using the Euler Characteristic. This approach ensures that the probability of a voxel in the non-active brain region exceeding this significance threshold is controlled at 5%. This resulting significance threshold, t, is equal to 5.12.

For each sample size in the studies selected in Figure 1a, the statistical power is computed for a range of standardised effect sizes D. The statistical power is defined as the probability that the local maximum peak of activation in the active region exceeds this significance threshold and was computed by taking the integral of the null distribution of local maxima increased with D/sqrt(n), over the interval [t,] from Cheng and Schwartzman. Lastly, the median effect size needed to exceed the significance threshold in each of the studies was found by selecting the effect size D that results in statistical power higher than 0.80 as computed in the previous step.

Figure 1b shows the median effect sizes needed to establish significance, with 80% power and alpha 0.05**[Au:OK?]**. Despite the decreases in these hypothetical effect sizes over the past 20 years, Fig. 1b shows that, even today, the average study is only sufficiently powered to detect very large effects of more than ~0.9. **[Au:OK?]** Given that many of the studies will be assessing group differences or brain activity **[Au:OK?]**–behaviour correlations (which will inherently have lower power than average group effects), this represents an optimistic lower bound on the powered effect size.

Indeed, Supplementary information S1 (box) demonstrates typical effect sizes observed in task-related BOLD imaging studies. Briefly, we analysed BOLD data from 186 individuals who were imaged doing motor, emotion, working memory and gambling tasks in the fMRI scanner as part of the Human Connectome Project. We created masks that captured the intersection between functional masks (identified from Neurosynth.org as regions consistently active in studies examining the effects of ‘motor’, ‘emotion’, ‘gambling’ and ‘working memory’ tasks) and anatomical masks (defined using the Harvard–Oxford probabilistic atlas). Within these intersection masks, we then determined the average task-related increases in BOLD signal — and the effect size (Cohen’s D) — associated with each different task. More-detailed explanation of the data are given in Supplementary information S1 (box). The table in Supplementary information S1 (box), which lists the resulting BOLD signal changes and inferred effect sizes, demonstrate that realistic effect sizes – i.e. BOLD changes associated with different tasks (for example a motor task or working memory task) - in fMRI are quite small: 90% of the voxels in the masks have a standardised effect size smaller than 1.2 for powerful tasks such as motor and emotion tasks. For more subtle tasks, such as gambling, only 10% of the voxels in our masks reflect standardised effect sizes larger than 0.5.

Thus the average fMRI study remains very poorly powered for capturing **[Au:OK?]** realistic effects.

***Solutions*.**

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 (for example, neuropowertools.org (see Further information; described in ref [13](https://paperpile.com/c/BVQBEO/8H09)) and fmripower.org (see Further information; described in ref. [14](https://paperpile.com/c/BVQBEO/MK6Z)). 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 would be theoretically informative (see Supplementary information S1 (table) for example effect sizes). The use of heuristic sample size guidelines (for example, based only on similar published studies) is likely to result in a misuse of resources, either by collecting too many or (more likely) too few subjects.

In some cases, researchers must use an insufficient sample size in a study, either owing to resources limitations or to limitations in the specific sample (for example, when studying a rare patient group). In such cases, there are two commonly used options to improve power. First, researchers may engage in a consortium with other researchers in order to combine data. This approach has been highly successful in the field of genetics, in which 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 Project and its International Neuroimaging Data-sharing Initiative (INDI)[15](https://paperpile.com/c/BVQBEO/XkzE) **[Au:OK?]** and the ENIGMA (Enhancing Neuro Imaging Genetics by Meta-Analysis) consortium[16](https://paperpile.com/c/BVQBEO/tyc9). Second, researchers may restrict the search space using a small number of *a priori* regions of interest (ROIs) or an independent ‘functional localizer’ to identify specific ROIs for each individual. It is essential that these ROIs (or a specific functional localizer strategy) be explicitly defined before any analyses. This is important because it is always possible to develop a *post hoc* justification for any specific ROI on the basis of previously published papers — a strategy that results in an ROI that appears independent but actually has a circular definition and thus leads to meaningless statistics and inflated Type I errors. By analogy to the idea of HARKing (hypothesizing after results are known; in which the results of exploratory analyses are presented as having been hypothesized from the beginning)[17](https://paperpile.com/c/BVQBEO/HNoV), we refer to this as SHARKing (selecting hypothesized areas after results are known). **[Au:OK?]** We would only recommend the use of restricted search spaces if the exact ROIs and hypotheses are pre-registered[18,19](https://paperpile.com/c/BVQBEO/I3Ag+LP9U).

**Analytic flexibility**

The typical 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](https://paperpile.com/c/BVQBEO/nvmk) applied more than 6,000 analysis workflows to a single data set and quantified the variability in resulting statistical maps. **[Au:OK?]** This revealed that some brain regions **[Au:OK?]** exhibited more substantial variation across the different workflows than did other regions. This issue is not unique to fMRI; for example, similar issues have been raised in genetics[21](https://paperpile.com/c/BVQBEO/JEbo). These “researcher degrees of freedom” can lead to substantial inflation of Type I error rates[22](https://paperpile.com/c/BVQBEO/v7fQ), even when there is no intentional p-value hacking[4](https://paperpile.com/c/BVQBEO/9nrt). **[Au:OK?]**

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 an essential component of publication, the intellectual journey of the research is often obscured. Instead, reports often engage in HARKing[17](https://paperpile.com/c/BVQBEO/HNoV). Because HARKing hides the number of data-driven choices made during analysis, it can strongly overstate the actual evidence for a hypothesis. There is arguably 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. The details to be pre-registered 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 data set.

**Multiple comparisons**

The most common approach to neuroimaging analysis involves ‘mass univariate’ testing of one hypothesis for each voxel. In such an approach, **[Au:OK?]** the false positive rate will be inflated if there is no correction for multiple tests. A humorous example of this was seen in a now-infamous study reported by Bennett and colleagues[23](https://paperpile.com/c/BVQBEO/yXaa), in which ‘activation’ was detected in the brain of a dead salmon (but disappeared when the proper corrections for multiple comparisons were performed).

Here we provide a similar example; a computational notebook for this demonstration can be viewed at <http://nbviewer.jupyter.org/github/poldrack/corrsim/blob/master/Correlation_simulation.ipynb> . Figure 2 presents an example in which random data can be analysed (incorrectly) to lead to seemingly impressive results, through a combination of failure to adequately correct for multiple comparisons and circular ROI analysis. We generated random simulated fMRI and behavioral data from a Gaussian distribution for 24 simulated subjects. The simulated fMRI data were smoothed with a 6mm Gaussian kernel. A univariate analysis was performed to assess the correlation between voxel **[Au:OK?]** activation and the simulated behavioural regressor, and the resulting statistical map was thresholded at p < 0.005 and with a 50-voxel extent threshold (that is, a heuristic correction for multiple comparisons). This approach reveals a lateral prefrontal cortex cluster of voxels in which the fMRI data apparently ‘correlate’ with the behavioural regressor (Fig. 2a). **[Au:OK?]**

The problem of multiplicity was recognized very early, and the past 25 years have seen the establishment of well-validated methods for correcting for FWE and false discovery rate in neuroimaging data[24](https://paperpile.com/c/BVQBEO/zW70). However, recent work[25](https://paperpile.com/c/BVQBEO/s4Ui) has suggested that even some very well-established methods for inference based on spatial extent of activations can produce inflated error rates (also see ‘Software errors’).

There is an ongoing debate between neuroimaging researchers who feel that conventional approaches to correcting for multiple comparisons are too loose and allow too many false positives[26](https://paperpile.com/c/BVQBEO/LeNS), and those who feel that thresholds are too conservative, and risk missing most of the interesting effects[27](https://paperpile.com/c/BVQBEO/Cvsn). In our view, the deeper problem is the inconsistent application of principled correction approaches[28](https://paperpile.com/c/BVQBEO/DtyZ). Many researchers freely combine different approaches and thresholds in ways that produce a high number of undocumented researcher degrees of freedom[22](https://paperpile.com/c/BVQBEO/v7fQ), rendering reported p-values uninterpretable.

To assess this more directly, we examined the first 100 results for the Pubmed query ("fMRI" AND brain AND activation NOT review[PT] AND human[MESH] AND english[la]), performed 23 May 2016; of these, 65 reported whole-brain task fMRI results and were available in full text. Only 3 presented fully uncorrected results, with 4 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 (Analysis of Functional NeuroImages; US National Institutes of Health and Neuroimaging Informatics Technology Initiative) **[Au:OK?]** 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 that use Gaussian random field theory to correct for multiple comparisons. **[Au:OK?]** Given the substantial degree of extra work involved in using multiple software packages, the use of a different tool solely for correcting for multiple comparisons 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 owing to a software bug[25](https://paperpile.com/c/BVQBEO/s4Ui). Distressingly, although nonparametric methods provide the most accurate control over FWE rates[24,25,29](https://paperpile.com/c/BVQBEO/zW70+s4Ui+ECfv), they were not used in any of these papers.

***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 through a repository that allows viewing and downloading (such as Neurovault.org[30](https://paperpile.com/c/BVQBEO/t34e)). For an example of this practice, see ref[31](https://paperpile.com/c/BVQBEO/bwwQ) and shared data at <http://neurovault.org/collections/122/>. Any use of non-standard methods for correction of multiple comparisons (for example, using an AFNI tool when other analyses were performed using SPM) should be justified explicitly (and reviewers should demand such justification).

**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; the bug resulted in substantially inflated Type I error rates[25](https://paperpile.com/c/BVQBEO/s4Ui), and 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](https://paperpile.com/c/BVQBEO/nukR)), 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.

**Insufficient study reporting**

Eight years ago we[33](https://paperpile.com/c/BVQBEO/xtqi) 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](https://paperpile.com/c/BVQBEO/u6Ts) and Guo and colleagues[35](https://paperpile.com/c/BVQBEO/eUaq) analysed 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 fMRI papers we found in the literature search 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 Organization for Human Brain Mapping (see Further information for link) has recently addressed this issue through its 2015–2016 **[Au:OK?]** Committee on Best Practices in Data Analysis and Sharing (COBIDAS), which has issued a new, detailed set of reporting guidelines[36](https://paperpile.com/c/BVQBEO/iuGW) (<http://www.humanbrainmapping.org/COBIDAS>).

Besides 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. In addition, ‘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](https://paperpile.com/c/BVQBEO/bPxq+ymU0). Furthermore, 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](https://paperpile.com/c/BVQBEO/jPGV), and the absence of **[Au:OK?]** appropriate tests for statistical interactions[39](https://paperpile.com/c/BVQBEO/wdIT).

***Solutions*.**

Authors should follow accepted standards for reporting methods (such as 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.

**Lack of independent replications**

There are surprisingly few examples of direct replication in the field of neuroimaging, probably 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 that had previously found associations between brain structure and behaviour. **[Au:OK?]** Only one of the 17 replication **[Au:OK?]** studies showed stronger evidence for the original effect size than for a null effect, and 8 out of 17 showed stronger evidence for a null effect[40,41](https://paperpile.com/c/BVQBEO/AT3O+skLz). This suggests that replicability of neuroimaging findings (and particularly brain activity–behaviour correlations) may be exceedingly low, similar to recent findings in other areas of science such as cancer biology[42](https://paperpile.com/c/BVQBEO/sSZm) and psychology[43](https://paperpile.com/c/BVQBEO/FVl3).

***Solutions*.**

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

**Conclusion**

We have outlined what we see as a set of problems associated with neuroimaging methodology and reporting, and solutions to solve them. It is likely that the reproducibility of neuroimaging 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 affect future support for research in this area.

**Further information**

NeuroPower: neuropowertools.org

Fmripower: fmripower.org

Organisation for Human Brain Mapping (OHBM): www.humanbrainmapping.org

Neurovault: http://neurovault.org/ **[Au:OK?]**

**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](https://paperpile.com/c/BVQBEO/uHO1). The transformation brought about by whole-genome methods (that is, genome-wide association studies) was partly necessitated by the simultaneous testing of several hundreds of thousands of genetic loci (hence the need for very stringent statistical criteria in order to reach ‘genome-wide significance’), but also by 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 has produced many hundreds more reproducible findings than in the previous fifteen years. Routine sharing of single nucleotide polymorphism (SNP)-level **[Au:OK?]** statistical results has facilitated the routine use of meta-analyses, as well as the development of novel methods of secondary analysis[45](https://paperpile.com/c/BVQBEO/NtFr).

This relatively rosy picture contrasts markedly with the situation in ‘imaging genomics’ — a burgeoning field that has yet to embrace the standards that are commonly followed in the genetics literature, and that remains largely focused on individual candidate gene association studies, which are characterized by numerous researcher degrees of freedom. **[Au: Edit OK?]** 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) 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 volumes) identified two genetic associations that were both **[Au:OK?]** replicated across two large samples each containing **[Au:OK?]** more than 10,000 individuals. Strikingly, by contrast, analysis of a set of previously identified candidate genes showed no evidence for any association in this very well-powered study[46](https://paperpile.com/c/BVQBEO/MBhT).

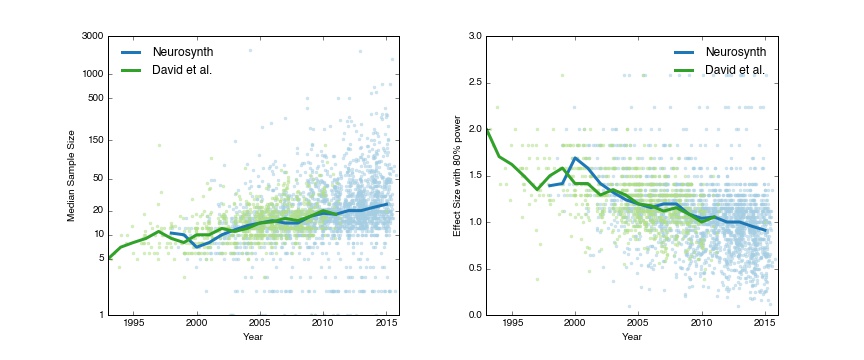
**Box 2 | Analytic flexibility in fMRI**

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 (with 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 analysed; 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 the table, 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!

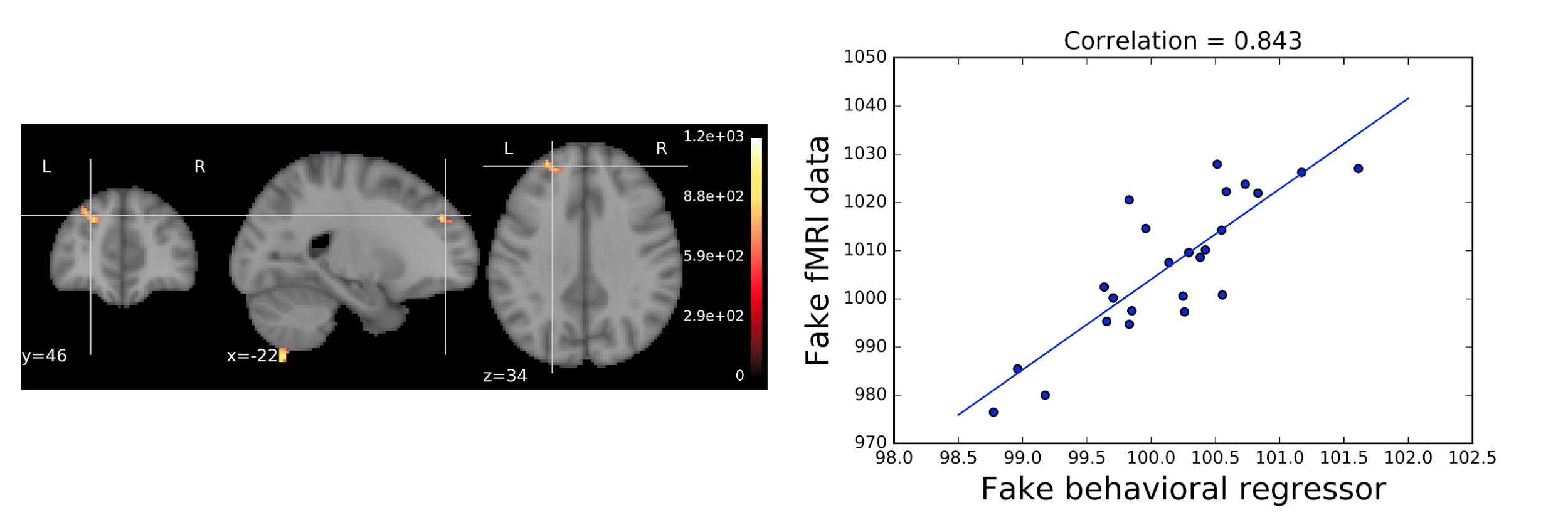
It is possible that many of these alternative pipelines could lead to very similar results, although the analyses of Carp[20](https://paperpile.com/c/BVQBEO/nvmk) suggest that many of them may lead to considerable heterogeneity in the results. In addition, there is evidence to show that choices of preprocessing parameters may interact with the statistical modelling approach, and that the optimal preprocessing pipeline may differ across subjects[47](https://paperpile.com/c/BVQBEO/gf38).

|  |  |  |  |
| --- | --- | --- | --- |
| **Processing step** | **Reason** | **Options [suboptions] [Au:OK?]** | **Number of plausible options** |
| Motion correction | Correct for head motion during scanning | * ‘Interpolation’ [‘linear’ or ‘sinc’] * Reference volume [‘single’ or ‘mean’] | 4 |
| Slice timing correction | Correct for differences in acquisition timing of different slices | ‘No’, ‘before motion correction’ or ‘after motion correction’ | 3 |
| Field map correction | Correct for distortion due to magnetic susceptibility | ‘Yes’ or ‘no’ | 2 |
| Spatial smoothing | Increase SNR for larger activations and to ensure assumptions of Gaussian random field theory | FWHM [‘4mm’, ‘6mm’ or ‘8mm’] | 3 |
| Spatial normalization | Warp individual brain to match a group template | Method [‘linear’ or ‘nonlinear’] | 2 |
| High pass filter | Remove low-frequency nuisance signals from data | Frequency cutoff [‘100’ or ‘120’] | 2 |
| Head motion regressors | Remove remaining signals related to head motion through a statistical model | ‘Yes’ or ‘no’; if ‘yes’, opt from ‘6’, ‘12’ or ‘24’ parameters or single-timepoint ‘scrubbing’ regressors **[Au:OK?]** | 5 |
| Haemodynamic response | Account for delayed nature of haemodynamic response to neuronal activity | * Basis function [‘single-gamma’ or ‘double-gamma’] * Derivatives [‘none’, ‘shift’ or ‘dispersion’] | 6 |
| Temporal autocorrelation model | Model for the temporal autocorrelation inherent in fMRI signals | ‘Yes’ or ‘no’ | 2 |
| Multiple comparison correction | Correct for large number of comparisons across the brain | ‘Voxel-based GRF’, ‘cluster-based GRF’, ‘FDR’ or ‘nonparameteric’ | 4 |
| **Total possible workflows** |  |  | **69,120** |

FDR, false discovery rate; FWHM, full width at half maximum; GRF, Gaussian random field; SNR, signal-to-noise ratio.



**Figure 1** **| Sample size estimates and estimated power for fMRI studies.** a | Sample sizes over more than 20 years obtained from two sources: manual extraction from published meta-analyses[10](https://paperpile.com/c/BVQBEO/JksF), and automated extraction from the Neurosynth database[11](https://paperpile.com/c/BVQBEO/ymU0). These data demonstrate that despite a steady increase in sample size, median sample size remained below 25 as of 2015. b | 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](https://paperpile.com/c/BVQBEO/rn03) (see main text for details). 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>.



**Figure 2 |** Seemingly impressive brain activity–behaviour association arising from random data through the use of uncorrected statistics and circular region-of-interest (ROI) analysis to capitalize on the large sampling error arising from small samples (see main text for details). The analysis revealed a cluster in the lateral prefrontal cortex (left panel); signal extracted from that cluster (that is, using circular analysis) showed a very strong correlation between the (fake) BOLD signal change in this region and the (fake) behavioural measure (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>.

**Acknowledgements**

R.P., J.D., J.B.P. and K.G. are supported by the Laura and John Arnold Foundation. M.R.M. 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 (MRC), and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. Support from the MRC (MC UU 12013/6) is also gratefully acknowledged. C.I.B. is supported by the Intramural Research Program of the US National Institute of Mental Health (NIMH). T.Y. is supported by the NIMH (R01MH096906). P.M.M. gratefully acknowledges personal support from the Edmond J. Safra Foundation and Lily Safra and research support from the MRC, the Imperial College Healthcare Trust Biomedical Research Centre and the Imperial EPSRC Mathematics in Healthcare Centre. T.E.N. is supported by the Wellcome Trust (100309/Z/12/Z). Data were provided (in part) by the Human Connectome Project, Washington University– University of Minnesota (WU-Minn) Consortium (principal investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 US National Institutes of Health (NIH) and NIH Centers that support the NIH Blueprint for Neuroscience Research; **[Au:OK?]** and by the McDonnell Center for Systems Neuroscience at Washington University. Thanks to S. David for providing sample size data. J.B.P. 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).](http://paperpile.com/b/BVQBEO/gjzr)

2. [Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).](http://paperpile.com/b/BVQBEO/hul1)

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).](http://paperpile.com/b/BVQBEO/fC6o)

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).](http://paperpile.com/b/BVQBEO/9nrt)

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).](http://paperpile.com/b/BVQBEO/zO2a)

6. [Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505,** 612–613 (2014).](http://paperpile.com/b/BVQBEO/xHD9)

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).](http://paperpile.com/b/BVQBEO/MCeX)

8. [Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).](http://paperpile.com/b/BVQBEO/jku2)

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 (2009).](http://paperpile.com/b/BVQBEO/0QnQ)

10. [David, S. P. *et al.* Potential reporting bias in fMRI studies of the brain. *PLoS One* **8,** e70104 (2013).](http://paperpile.com/b/BVQBEO/JksF)

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).](http://paperpile.com/b/BVQBEO/ymU0)

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).](http://paperpile.com/b/BVQBEO/rn03)

13. [Durnez, J. *et al.* Power and sample size calculations for fMRI studies based on the prevalence of active peaks. *bioRxiv* 049429 (2016). doi:](http://paperpile.com/b/BVQBEO/8H09)[10.1101/049429](http://dx.doi.org/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).](http://paperpile.com/b/BVQBEO/MK6Z)

15. [Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4734–4739 (2010).](http://paperpile.com/b/BVQBEO/XkzE)

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).](http://paperpile.com/b/BVQBEO/tyc9)

17. [Kerr, N. L. HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2,** 196–217 (1998).](http://paperpile.com/b/BVQBEO/HNoV)

18. [Nosek, B. A. *et al.* SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* **348,** 1422–1425 (2015).](http://paperpile.com/b/BVQBEO/I3Ag)

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).](http://paperpile.com/b/BVQBEO/LP9U)

20. [Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6,** 149 (2012).](http://paperpile.com/b/BVQBEO/nvmk)

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).](http://paperpile.com/b/BVQBEO/JEbo)

22. [Simmons, J. P., Nelson, L. D. & Uri, S. False-Positive Psychology: The Way We Report Studies Privileges False Findings.](http://paperpile.com/b/BVQBEO/v7fQ) *[PsycEXTRA Dataset](http://paperpile.com/b/BVQBEO/v7fQ)* [doi:](http://paperpile.com/b/BVQBEO/v7fQ)[10.1037/e636412012-001](http://dx.doi.org/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).](http://paperpile.com/b/BVQBEO/yXaa)

24. [Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12,** 419–446 (2003).](http://paperpile.com/b/BVQBEO/zW70)

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).](http://paperpile.com/b/BVQBEO/s4Ui)

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).](http://paperpile.com/b/BVQBEO/LeNS)

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).](http://paperpile.com/b/BVQBEO/Cvsn)

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).](http://paperpile.com/b/BVQBEO/DtyZ)

29. [Hayasaka, S. & Nichols, T. E. Validating cluster size inference: random field and permutation methods. *Neuroimage* **20,** 2343–2356 (2003).](http://paperpile.com/b/BVQBEO/ECfv)

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).](http://paperpile.com/b/BVQBEO/t34e)

31. [Hunt, L. T., Dolan, R. J. & Behrens, T. E. J. Hierarchical competitions subserving multi-attribute choice. *Nat. Neurosci.* **17,** 1613–1622 (2014).](http://paperpile.com/b/BVQBEO/bwwQ)

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).](http://paperpile.com/b/BVQBEO/nukR)

33. [Poldrack, R. A. *et al.* Guidelines for reporting an fMRI study. *Neuroimage* **40,** 409–414 (2008).](http://paperpile.com/b/BVQBEO/xtqi)

34. [Carp, J. & Joshua, C. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage* **63,** 289–300 (2012).](http://paperpile.com/b/BVQBEO/u6Ts)

35. [Guo, Q. *et al.* The Reporting of Observational Clinical Functional Magnetic Resonance Imaging Studies: A Systematic Review. *PLoS One* **9,** e94412 (2014).](http://paperpile.com/b/BVQBEO/eUaq)

36. [Nichols, T. E. *et al.* Best Practices in Data Analysis and Sharing in Neuroimaging using MRI. *bioRxiv* 054262 (2016). doi:](http://paperpile.com/b/BVQBEO/iuGW)[10.1101/054262](http://dx.doi.org/10.1101/054262)

37. [Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10,** 59–63 (2006).](http://paperpile.com/b/BVQBEO/bPxq)

38. [Gelman, A. & Stern, H. The Difference Between ‘Significant’ and ‘Not Significant’ is not Itself Statistically Significant. *Am. Stat.* **60,** 328–331 (2006).](http://paperpile.com/b/BVQBEO/jPGV)

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).](http://paperpile.com/b/BVQBEO/wdIT)

40. [Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66,** 115–133 (2015).](http://paperpile.com/b/BVQBEO/AT3O)

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).](http://paperpile.com/b/BVQBEO/skLz)

42. [Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).](http://paperpile.com/b/BVQBEO/sSZm)

43. [Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).](http://paperpile.com/b/BVQBEO/FVl3)

44. [Flint, J. & Munafò, M. R. Candidate and non-candidate genes in behavior genetics. *Curr. Opin. Neurobiol.* **23,** 57–61 (2013).](http://paperpile.com/b/BVQBEO/uHO1)

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).](http://paperpile.com/b/BVQBEO/NtFr)

46. [Stein, J. L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–561 (2012).](http://paperpile.com/b/BVQBEO/MBhT)

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).](http://paperpile.com/b/BVQBEO/gf38)