Scanning the Horizon: challenges and solutions 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, Stanford, CA, 94305, USA
2. Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, MD, 20892, USA
3. Division of Brain Sciences, Department of Medicine, Hammersmith Hospital, London, London W12 0NN, UK
4. MRC Integrative Epidemiology Unit at the University of Bristol, BS8 1BN, UK
5. UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, BS8 1TU, UK
6. Department of Statistics & WMG, University of Warwick, Coventry, CV4 7AL, UK
7. Helen Wills Neuroscience Institute, 132 Barker Hall 210S, Henry H. Wheeler Jr. Brain Imaging Center, University of California, Berkeley, 94720-3192, CA, USA
8. Department of Psychology, University of California, San Diego; San Diego, CA, 92093, USA.
9. Department of Psychology, University of Texas at Austin, Austin, TX, 78712, USA.

Corresponding author: R.A.P. (russpold@stanford.edu)

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 some 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. Here 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 neuroscientific questions.

Main text

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. The most common form of fMRI (known as “blood oxygen level dependent” or BOLD fMRI) measures

brain activity indirectly through localized changes in blood oxygenation that occur in relation to synaptic signaling [2]. These signal changes provide the ability to map activation in relation to specific mental processes, characterize neural representational spaces [3], and decode or predict mental function from brain activity [4,5]. 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. In particular, the high dimensionality of fMRI data, relatively low power of most fMRI studies, and the great amount of flexibility in data analysis all potentially contribute to a high degree of false positive findings.

Recent years have seen intense interest in the degree to which widespread "questionable research practices" (QRPs) are responsible for high rates of false findings in the scientific literature, particularly within psychology but also more generally[6–8]. There is growing interest in "meta-research"[9], 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[7]. 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[10].

Perhaps one of the most surprising findings in recent work is the lack of appreciation of the QRP problem by researchers. John and colleagues[11] 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.

In this article we outline a number of potentially questionable research practices in neuroimaging 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 as is evident from the discussion below, it is not always clear whether best practices have been 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

The analyses of Button and colleagues[12] provided a wake-up call regarding statistical power in neuroscience, particularly by highlighting the point (raised earlier by Ioannidis[6]) 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[13]. In the context of neuroimaging, Button and colleagues considered only structural MRI studies. In order to assess the current state of statistical power in fMRI studies, we performed an analysis of sample sizes and the resulting statistical power of fMRI studies over the past 20 years.

To gain a perspective on how sample sizes have changed over this time period, we obtained sample sizes from fMRI studies using two sources. First, manually annotated sample size data were obtained from from published meta-analyses[14]. Second, sample sizes were automatically extracted from the Neurosynth database [15] for fMRI studies published between 2011 and 2015 (by searching for regular expressions reflecting sample size, e.g. “13 subjects”, “n=24”) and then manually annotated to confirm automatic estimates and identify single-group versus multiple-group studies. (Data and code to generate all figures in this paper are available from the Open Science Framework at https://osf.io/spr9a/.) Figure 1a shows that sample sizes have steadily increased over the past two decades, with the median estimated sample size for a single-group fMRI study in 2015 at 28.5. A particularly encouraging finding from this analysis is that the number of recent studies with large samples (greater than 100) is rapidly increasing (from 8 in 2012 to 17 in 2015, in the studied sample), suggesting that the field may be progressing towards adequately powered research. On the other hand, the median group size in 2015 for fMRI studies with multiple groups was 19 subjects, which is below even the absolute minimum sample size of 20 per cell proposed by Simonsohn et al.[7].

In order to assess the implications of these results for statistical power, for each of the 1131 sample sizes shown in Figure 1a we estimated the standardized effect size that would be required to detect an effect with 80% power (the standard level of power for most fields) for a whole-brain linear mixed-effects analysis using a voxelwise 5% familywise error (FWE) rate threshold from random field theory[16] (a standard thresholding level for neuroimaging studies). In other words, we found the minimum effect size that would have been needed in each of these studies in order for the difference to be considered statistically significant, given the sample size. We quantify the standardised effect size using Cohen’s D, computed as the average effect divided by the standard deviation for the data.

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 voxel size, 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 millimeters and a volume within the brain mask of 228483 voxels, used by default in most fMRI analysis software. We assume that in each case there would be one active region, with voxelwise standardised effect size D; that is, we assume that for each subject, all voxels in the active region are on average D standardised units higher in their activity than the voxels in the non-active region, and that the active region is 1,600 mm^2 (200 voxels). To calculate the voxelwise statistical significance threshold for the active region in this model statistical map, we

used the function *ptoz* from the FSL[17] software package, which computes a FWE threshold for a given volume and smoothness using the Euler Characteristic derived from Gaussian random field theory [18]. This approach ensures that the probability of a voxel in the non-active brain region exceeding this significance threshold is controlled at 5%; the resulting significance threshold, t_{α}, is equal to 5.12.

--- Figure 1 about here ---

The statistical power defined as the probability that the local maximum peak of activation in the active region exceeds this significance threshold. This probability was computed using a shifted version of the null local maximum distribution, with shift of D*sqrt(n) to reflect a given effect size and sample size. 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. Despite the decreases in these hypothetical required effect sizes over the past 20 years, Fig. 1b shows that in 2015 the median study is only sufficiently powered to detect relatively large effects of greater than ~0.75. Given that many of the studies will be assessing group differences or brain activity–behaviour correlations (which will inherently have lower power than average group activation effects), this represents an optimistic lower bound on the powered effect size.

Indeed, the analysis presented in Box 2 demonstrates that typical effect sizes observed in task-related BOLD imaging studies fall well below this level. Briefly, we analysed BOLD data from 186 individuals who were imaged using fMRI while performing motor, emotion, working memory and gambling tasks as part of the Human Connectome Project[20]. Assessing effect sizes in fMRI requires the definition of an independent region of interest that captures the expected area of activation within which the effect size can be measured. To acheive this, we created masks that captured the intersection between functional activation (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[21], based on the published regions of interest from the HCP)[22]. 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. Additional details are provided in Box 2. The figure in Box 2, which lists the resulting BOLD signal changes and inferred effect sizes, demonstrates that realistic effect sizes – i.e. BOLD changes associated with a range of cognitive tasks - in fMRI are surprisingly small: even for powerful tasks such as the motor task which evokes median signal changes of greater than 4%, 75% of the voxels in the masks have a standardised effect size smaller than 1. For more subtle tasks, such as gambling, only 10% of the voxels in our masks demonstrated standardised effect sizes larger than 0.5. Thus the average fMRI study remains poorly powered for capturing 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 [23]) and fmripower.org (see Further information; described in ref. [24]). 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 (e.g. based on results from Box 2). The use of heuristic sample size guidelines (for example, based on a small number of previously 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, due 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)[25] and the ENIGMA (Enhancing Neuro Imaging Genetics by Meta-Analysis) consortium[26]. Second, researchers may restrict the search space using a small number of *a priori* regions of interest (ROIs) or an independent 'functional localizer' (a separate scan used to identify regions based on their functional response, such as retinotopic visual areas or face-responsive regions) 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)[27], we refer to this as SHARKing (selecting hypothesized areas after results are known). We would only recommend the use of restricted search spaces if the exact ROIs and hypotheses are pre-registered[28,29].

Problem: 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 3). Carp[30] applied 6,912 analysis workflows (using the SPM[31] and AFNI[32] software packages) to a single data set and quantified the variability in resulting statistical maps. This revealed that some brain regions 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[33]. These “researcher degrees of freedom” can lead to substantial inflation of Type I error rates[7], even when there is no intentional “p-hacking”[8].

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[27]. 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.

Problem: Multiple comparisons

The most common approach to neuroimaging analysis involves “mass univariate” testing in which a separate hypothesis test is performed for each voxel. In such an approach, 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[34], 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 a similar 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 (mean±standard deviation = 1000±100 for fMRI data, 100±1 for behavioral data) for 28 simulated subjects (based on the median sample size found in the analysis of Figure 1 for studies from 2015). For the fMRI data, we simulated statistical values at each voxel for a comparison of activation and baseline conditions for each of the simulated subjects within the standard MNI152 mask, and then spatially smoothed the image with a 6mm Gaussian kernel, based on the common smoothing level of 3 times the voxel size. A univariate analysis was performed using FSL to assess the correlation between activation in each voxel and the simulated behavioural regressor across subjects, and the resulting statistical map was thresholded at $p < 0.001$ and with a 10-voxel extent threshold (which is a common heuristic correction shown by Eklund et al.[35] to result in highly inflated

levels of false positives). This approach revealed a cluster of false positive activation in the superior temporal cortex in which the simulated fMRI data are highly correlated with the simulated behavioural regressor (Fig. 2a).

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[36]. However, recent work[35] has suggested that even some very well-established inferential methods based on spatial extent of activations can produce inflated Type I error rates (also see below under ***Software Errors***).

-- Figure 2 about here--

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[37], and those who feel that thresholds are too conservative, and risk missing most of the interesting effects[38]. In our view, the deeper problem is the inconsistent application of principled correction approaches[39]. Many researchers freely combine different approaches and thresholds in ways that produce a high number of undocumented researcher degrees of freedom[7], 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 (full list of papers and annotations available at https://osf.io/spr9a/). 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 engage in “method-shopping” for techniques that provide greater sensitivity, at a potential cost of increased error rates. Nine of the 65 papers used the FSL or SPM software packages to perform their primary analysis, but then used the alphasim or 3dClustSim tools from the AFNI software package (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 or nonparametric analyses to correct for multiple comparisons. Given the substantial degree of extra work (e.g. software installation, file reformatting) 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 [35]. Distressingly, whereas nonparametric (randomization/permutation) methods are known to provide the more accurate control over familywise error rates compared to parametric methods[36,40], 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[41]). For an example of this practice, see ref[42] and shared data at http://neurovault.org/collections/122/. Any use of non-standard methods for correction of multiple comparisons (for example, using tools from different packages for the main analysis and the multiple comparison correction) 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 was discovered in the AFNI program 3dClustSim (and the older AlphaSim), which resulted in slightly inflated Type I error rates [35] (the bug was fixed in May 2015). The impact of such bugs could be widespread; for example, PubMed Central lists 1362 publications mentioning AlphaSim or 3dClustSim published prior to 2015 (query [("alphasim" OR "3DClustSim") AND 1992:2014[DP]] performed July 14, 2016). Similarly, the analyses presented in a preprint of the present article contained two software errors that led to different results being presented in the final version of the paper. This led us to perform a code review and to include software tests in order to reduce the likelihood of remaining errors.

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 good programming practices, including the judicious use of software testing and validation. Validation methodologies (such as comparing with existing implementation or using simulated data) should be clearly defined. Custom analysis codes should always be shared upon manuscript submission (for an example, see[43]), 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[44] 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[45] and Guo and colleagues[46] analyzed 100 and 241 fMRI papers respectively for the reporting of methodological details, and both found that some important analysis details (e.g. interpolation methods, smoothness estimates) were rarely described. Consistent with this, in 22 of the 65

papers discussed above it was impossible to identify exactly which multiple comparison 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 Committee on Best Practices in Data Analysis and Sharing (COBIDAS), which has issued a new, detailed set of reporting guidelines[47] (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 (e.g., “the anterior insula was activated, suggesting that subjects experienced empathy”), are rarely grounded in quantitative evidence[15,48]. 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”[49] and in the absence of appropriate tests for statistical interactions[50].

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.

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[51,52] attempted to replicate 17 studies that had previously found associations between brain structure and behaviour. Only one of the 17 replication attempts showed stronger evidence for an effect as large the original effect size rather than for a null effect, and 8 out of 17 showed stronger evidence for a null effect. 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[53] and psychology[54].

Solutions:
The neuroimaging community should acknowledge replication reports as scientifically important research outcomes that are essential in advancing knowledge. One such attempt is the OHBM

Replication Award to be awarded in 2017 for the best neuroimaging replication study in the previous year.

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 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, so as to ensure that public funds are spent effectively and in a way that maximizes our understanding of the human brain.

Further information

Fmripower: fmripower.org
Human Connectome Project: https://www.humanconnectome.org/
Organisation for Human Brain Mapping (OHBM): www.humanbrainmapping.org
NeuroPower: neuropowertools.org
Neurosynth: http://neurosynth.org/
Neurovault: http://neurovault.org/

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[55]. 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 (generally <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[56]. Routine sharing of

single nucleotide polymorphism (SNP)-level statistical results has facilitated routine use of meta-analyses, as well as the development of novel methods of secondary analysis[57].

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 broader genetics literature, and remains largely focused on individual candidate gene association studies, which are 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) which included a genetic association analysis (for list of search results, see https://osf.io/spr9a/). Of these, the vast majority (43/50) reported analysis of a single or small number (5 or fewer) of candidate genes; only 2/50 reported a genome-wide analysis, with the rest reporting analyses using biologically inspired gene sets (3/50) or polygenic risk scores (2/50). Recent empirical evidence also casts doubt on the validity of candidate gene associations in imaging genomics. A large genome-wide association study of brain structure[58] (including whole-brain and hippocampal volume) identified two genetic associations that were both replicated across two large samples each containing more than 10,000 individuals. Strikingly, analysis of a set of candidate genes previously reported in the literature showed no evidence for any association in this very well-powered study[58].

Box 2: Effect-size estimates for common neuroimaging experimental paradigms.
The aim of this analysis is to estimate the magnitude of typical effect sizes of blood oxygen level-dependent changes in fMRI signal associated with common psychological paradigms. We focus on four experiments administered by the Human Connectome Project (HCP)[59]: an emotion task, gambling task, working memory task and motor task (detailed below). We chose data from the HCP for its large sample size, which allows computation of stable effect size estimates, and its diverse set of activation tasks. The data and code used for this analysis are available at https://osf.io/spr9a/ .

Briefly, the processing of data from the Human Connectome Project was carried out in 4 main steps:

1. **Subject Selection**: The analyses are performed on the 500 subjects release of the HCP data, freely available at www.humanconnectome.org. We selected 186 independent subjects from the HCP data on the bases that (1) all subjects have results for all four of the tasks and (2) there are no genetically related subjects in the analysis.

2. **Group Analyses**: The first-level analyses, which summarise the relation between the experimental design and the measured timeseries for each subject, were obtained from the Human Connectome Project. The processing and analysis pipelines for these analyses are shared together with the data. Here we perform second-level analyses — that is, an assessment of the average effect of the task on BOLD signal over subjects — using the FSL program flame1[17] which performs a linear mixed-effects regression at each voxel, using

generalized least squares with a local estimate of random effects variance. This analysis averages over subjects, while separating within-subject and between-subject variability to ensure control of unobserved heterogeneity.

The specific contrasts that were tested are:

- Motor: average BOLD response for tongue, hand and foot movements versus rest
- Emotion: viewing faces with a fearful expression versus viewing neutral faces
- Gambling: monetary reward versus punishment
- Working memory: a contrast between conditions in which the participants indicate whether the current stimulus matches the one from 2 trials earlier ("2-back"), versus a condition where the participants indicate whether the current stimulus matches a specific target ("0-back")

3. **Create Masks**: The masks used for the analyses are the intersections of anatomical and *a priori* functional masks for each contrast. The rationale behind this is to find effect sizes in regions that are functionally related to the task, but restricted to certain anatomical regions.

- Functional: We created masks using www.neurosynth.org[60]. To do this, we performed forward inference meta-analysis using the respective search terms "Motor","Emotion","Gambling","Working memory" for each of the tasks, with false discovery rate (FDR) control at 0.01, the default threshold on neurosynth. The resulting mask identifies voxels consistently found to be activated in studies that mention each of the search terms in their abstract.
- Anatomical: We have used Harvard-Oxford probabilistic atlas[21] at p>0. Regions were chosen for each task based on the published a priori hypothesized regions from the HCP[22]. The size of the masks was assessed by the number of voxels in the mask.

Task	**Anatomical Mask**
Motor	• Precentral gyrus • Supplementary motor cortex • Left putamen • Right putamen
Working memory	Middle frontal gyrus
Emotion	• Left amygdala • Right amygdala
Gambling	• Left accumbens • Right accumbens

4. **Compute Effect Size**: The intersection masks created above were used to isolate the regions of interest in the second-level-analysed BOLD signal data. From these mask-isolated data sets, the size of the task-related effect (Cohen's D) were computed for

each relevant region(see Figure B2 below). FSL's Featquery computes for each voxel the % BOLD change in the data within the masks.

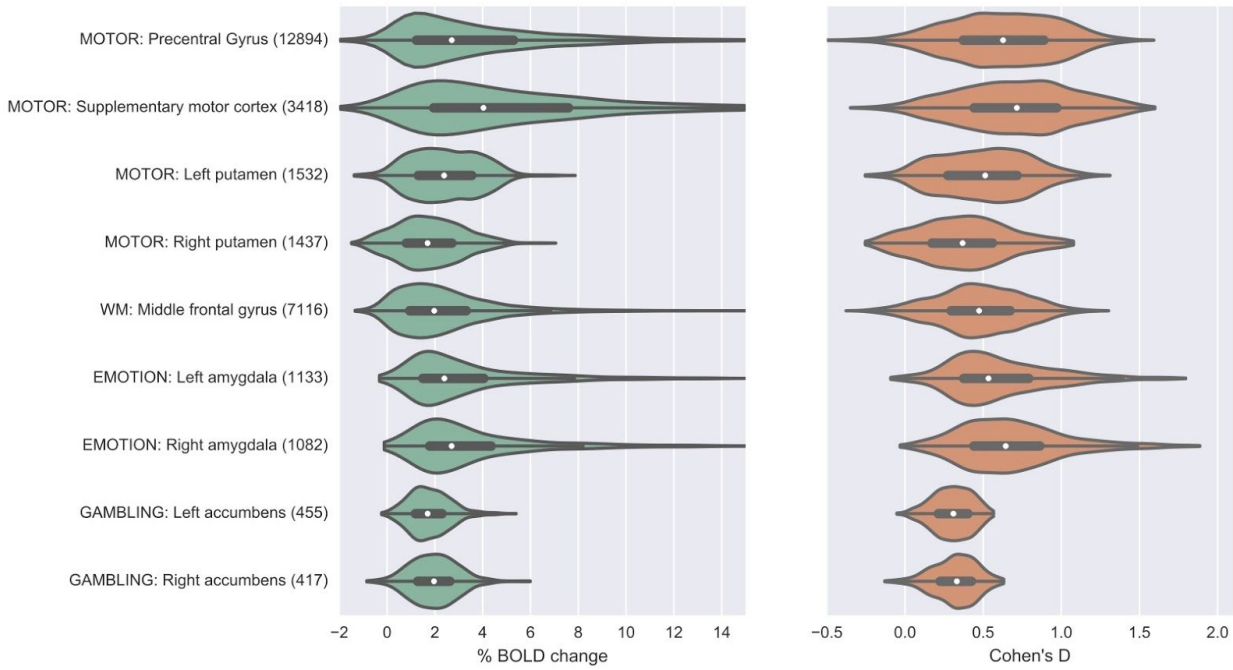

Figure B2: The distributions of the observed effect size estimates and BOLD signal change estimates for common experimental paradigms. The boxplot inside the violins represent the inter-quartile range (first quartile to third quartile) and the white dot shows the median value.

Box 3: 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[30] 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 (e.g., interactions between head motion modeling and physiological noise correction), and that the optimal preprocessing pipeline may differ across subjects (e.g. interacting with the amount of head motion)[61].

Table B3: 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 [suboptions]**	**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

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 secs, 120 secs]	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**

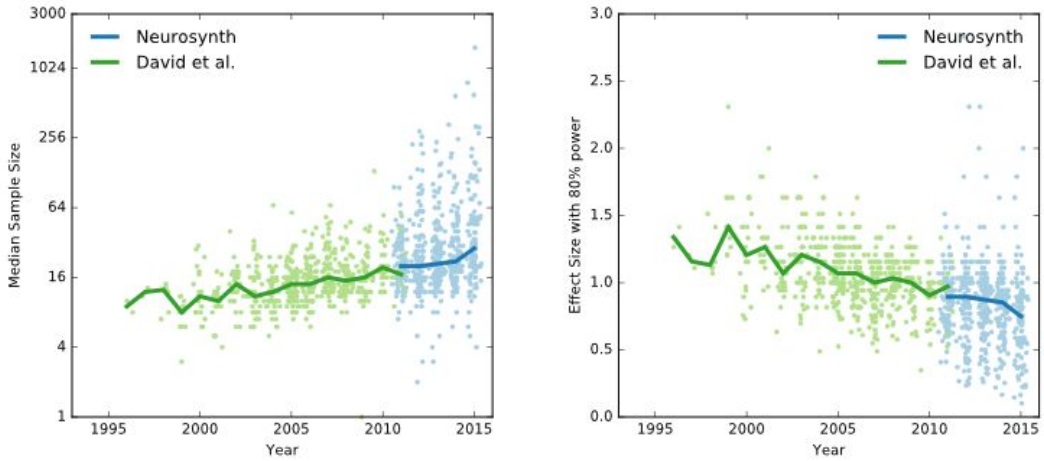

Figure 1 | Sample size estimates and estimated power for fMRI studies. a | 904 sample sizes over more than 20 years obtained from two sources: 583 sample sizes by manual extraction from published meta-analyses[14], and 548 sample sizes obtained by automated extraction from the Neurosynth database[15] with manual verification. These data demonstrate that sample sizes have steadily increased over the last two decades, with a median estimated sample size of 28.5 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[16] (see main text for details). Median effect size for which studies were powered to find in 2015 was 0.75. Data and code to generate these figures are available at https://osf.io/spr9a/

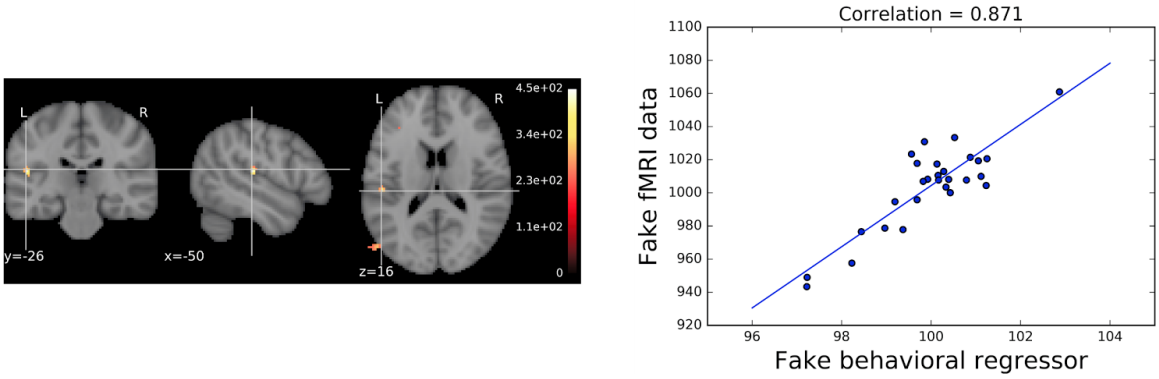

Figure 2: **Small samples can produce misleadingly large effects.**
Seemingly impressive brain-behavior association can arise from completely random data through the use of uncorrected statistics and circular ROI analysis to capitalize on the large sampling error arising from small samples. The analysis revealed a cluster in the superior temporal gyrus (left panel); signal extracted from that cluster (i.e., using circular analysis) showed a very strong correlation between brain and behavior (right panel; r = 0.87). See main

text for details of the analysis. A computational notebook for this example is available at https://osf.io/spr9a/.

Acknowledgements

RP, JD, JBP and KG are supported by the Laura and John Arnold Foundation. MRM is supported by the MRC (MC UU 12013/6) and 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. 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). JBP is supported by NIH-NIBIB P41-EB019936 and by NIH-NIDA U24-038653. 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 Institutes and Centers of the US National Institutes of Health (NIH) that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Thanks to Joe Wexler for performing annotation of Neurosynth data, Sean David for providing sample size data, and Robert Cox and Paul Taylor for helpful comments on a draft of the manuscript.

References

1. Poldrack, R. A. & Farah, M. J. Progress and challenges in probing the human brain. *Nature* **526,** 371–379 (2015).
2. Logothetis, N. K. What we can do and what we cannot do with fMRI. *Nature* **453,** 869–878 (2008).
3. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60,** 1126–1141 (2008).
4. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10,** 424–430 (2006).

5. Poldrack, R. A. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* **72,** 692–697 (2011).
6. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
7. 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).
8. 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).
9. 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).
10. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505,** 612–613 (2014).
11. 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).
12. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).
13. 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).
14. David, S. P. *et al.* Potential reporting bias in fMRI studies of the brain. *PLoS One* **8,** e70104 (2013).

15. 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).
16. 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).
17. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62,** 782–790 (2012).
18. Worsley, K. J. *et al.* A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4,** 58–73 (1996).
19. Cheng, D. & Schwartzman, A. Distribution of the Height of Local Maxima of Gaussian Random Fields. *Extremes* **18,** 213–240 (2015).
20. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *Neuroimage* **80,** 62–79 (2013).
21. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31,** 968–980 (2006).
22. Barch, D. M. *et al.* Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80,** 169–189 (2013).
23. 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
24. Mumford, J. A. & Nichols, T. E. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **39,** 261–268 (2008).
25. Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4734–4739 (2010).

26. Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* **8,** 153–182 (2014).

27. Kerr, N. L. HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2,** 196–217 (1998).

28. Nosek, B. A. *et al.* SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* **348,** 1422–1425 (2015).

29. Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. Registered reports: realigning incentives in scientific publishing. *Cortex* **66,** A1–2 (2015).

30. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6,** 149 (2012).

31. Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J. & Nichols, T. E. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. (Elsevier Science, 2011).

32. Cox, R. W. AFNI: what a long strange trip it's been. *Neuroimage* **62,** 743–747 (2012).

33. 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).

34. 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).

35. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1602413113

36. Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12,** 419–446 (2003).

37. 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).

38. 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).

39. 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).

40. Hayasaka, S. & Nichols, T. E. Validating cluster size inference: random field and permutation methods. *Neuroimage* **20,** 2343–2356 (2003).

41. 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).

42. Hunt, L. T., Dolan, R. J. & Behrens, T. E. J. Hierarchical competitions subserving multi-attribute choice. *Nat. Neurosci.* **17,** 1613–1622 (2014).

43. 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).

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

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

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

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

48. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends*

Cogn. Sci. **10,** 59–63 (2006).

49. Gelman, A. & Stern, H. The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant. *Am. Stat.* **60,** 328–331 (2006).
50. Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E.-J. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* **14,** 1105–1107 (2011).
51. Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66,** 115–133 (2015).
52. 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).
53. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
54. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).
55. Flint, J. & Munafò, M. R. Candidate and non-candidate genes in behavior genetics. *Curr. Opin. Neurobiol.* **23,** 57–61 (2013).
56. Ioannidis, J. P. A., Tarone, R. & McLaughlin, J. K. The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology* **22,** 450 (2011).
57. 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).
58. Stein, J. L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–561 (2012).
59. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview.

Neuroimage **80,** 62–79 (2013).

60. 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).

61. 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).