ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *The Year in Cognitive Neuroscience*
REVIEW

Progress toward openness, transparency, and reproducibility in cognitive neuroscience

Rick O. Gilmore,[1] Michele T. Diaz,[1,2] Brad A. Wyble,[1] and Tal Yarkoni[3]

[1]Department of Psychology, the Pennsylvania State University, University Park, Pennsylvania. [2]Social, Life, & Engineering Sciences Imaging Center, the Pennsylvania State University, University Park, Pennsylvania. [3]University of Texas at Austin, Austin, Texas

Address for correspondence: Rick O. Gilmore, Ph.D., Associate Professor, Department of Psychology, the Pennsylvania State University, University Park, PA 16802. rogilmore@psu.edu

Accumulating evidence suggests that many findings in psychological science and cognitive neuroscience may prove difficult to reproduce; statistical power in brain imaging studies is low and has not improved recently; software errors in analysis tools are common and can go undetected for many years; and, a few large-scale studies notwithstanding, open sharing of data, code, and materials remain the rare exception. At the same time, there is a renewed focus on reproducibility, transparency, and openness as essential core values in cognitive neuroscience. The emergence and rapid growth of data archives, meta-analytic tools, software pipelines, and research groups devoted to improved methodology reflect this new sensibility. We review evidence that the field has begun to embrace new open research practices and illustrate how these can begin to address problems of reproducibility, statistical power, and transparency in ways that will ultimately accelerate discovery.

Keywords: open science; reproducibility; data sharing

Introduction

Most cognitive neuroscientists seek answers to questions about what patterns of neural activity underlie perception, thinking, memory, and action, among other topics. In answering these questions, we marshal evidence from studies of human and animal behavior, nervous system structure and activity, the effects of endogenous and exogenous substances, patterns of disorder and disease, and trajectories of change across the life span. Our common aim is to reveal reliable, reproducible, and useful facts about the relationship between mind and brain. These facts depend crucially on the tools we deploy to collect and evaluate data and on how we report what we do or do not find. Here, we review the degree to which our field meets the scientific ideals of reproducibility, transparency, and openness.

Rigorous self-reflection and self-criticism about methodology have been core values in cognitive neuroscience for some time.[1–3] Efforts to foster widespread data sharing[4–6] and other open research practices have long histories. What strikes us as new and important enough to merit reviewing them in 2017 are developments that likely cheer the pessimist and the optimist alike. On the one hand, accumulating evidence suggests that many findings in psychological science may be difficult to reproduce;[7] statistical power in brain imaging studies is low[8–10] and has not improved[11] over time; software errors in analysis tools are common and can go undetected for many years;[12] and, a few large-scale studies and databases notwithstanding,[4,13,14] the open sharing of data, code, and materials are rare. On the other hand, we see a renewed focus on reaffirming reproducibility, transparency, and openness as essential core values in psychological science and related fields.[7,15–19] This reinvigorated focus has begun to provide greater clarity about what these values mean in practice.[20] We find that the emergence and rapid growth of data archives, meta-analytic tools, software pipelines, and research groups devoted to improved methodology are genuine reasons for optimism about the future of an open, transparent, and reproducible cognitive neuroscience.

doi: 10.1111/nyas.13325

In the sections that follow, we discuss definitions of open science practices and why they might be important for the field. We then review some of the history of these practices, discuss a range of recent developments, and speculate about what the near future might hold.

History of open science practices in cognitive neuroscience

What are open science practices? What does it mean to reproduce or replicate a study? Most researchers agree that discovering robust and generalizable findings is central to the scientific enterprise,[1,2,16,19,21] but what evidence determines success or failure in meeting the ideal? In a previous paper in *Annals of the New York Academy of Sciences*, Bennett and Miller[1] sought to assess the reliability of functional magnetic resonance imaging (fMRI) results and provided data about the diverse measures used to assess reliability in the functional neuroimaging literature of that time. In summarizing the results of a large sample of studies reporting measures of test/retest reliability, Bennett and Miller[1] observed that no agreement exists about what constitutes acceptable reliability, nor was there consensus on what measure or measures should be used to evaluate it. Half a decade later, Goodman *et al.*[20] argued that uncertainty and disagreement about the meaning of these concepts[22] persists and that misunderstanding impedes progress toward solutions. In response, Goodman *et al.* suggest three new terms that we adopt here: methods reproducibility, results reproducibility, and inferential reproducibility. *Methods reproducibility* means that a different investigator is able to obtain the same results when applying the same tools and analytical procedures used in a study to the same (i.e., original) data set. *Results reproducibility* means that a new study with new data, collected following the original procedures as closely as possible, yields the same outcomes. *Inferential reproducibility* occurs when independent researchers come to similar conclusions about what patterns of data mean, based on their own replication study or a reanalysis of a prior study.[20] For example, Goodman *et al.*[20] suggest that competing views about the implications of a recent high-profile study of replicability in psychology[7] stem, at least in part, from a disagreement at this level.[23,24]

Clearly, to achieve methods reproducibility, research practices that accurately and precisely capture essential details about methods, data, and workflows must be deployed; to achieve results reproducibility, those elements must be made openly and freely available to the scientific community; and achieving inferential reproducibility requires, among other developments, the capacity to accumulate, analyze, and interpret large quantities of data[25,26] in consistent ways. Thus, openness and transparency relate directly to reproducibility of all three kinds. Reflecting this sensibility, a diverse array of behavioral scientists have begun arguing that achieving the scientific ideals of a free and open exchange of information requires the widespread adoption of open and transparent communication practices.[15,16,27–30] How well has the field of cognitive neuroscience measured up to these ideals?

Methods reproducibility

Much of cognitive neuroscience research is computationally intensive, so the extent to which the field's methods are reproducible depends on whether complex computational workflows can be reliably regenerated. Whether measuring task- or non-task–related nervous system activity using electroencephalography (EEG) or fMRI or brain structure using MRI, computed tomography, or positron emission tomography, cognitive neuroscience studies regularly generate spatially and temporally dense data streams. Seemingly minor choices made at each step of an analysis pipeline—including experimental design, data acquisition, preprocessing, analysis, and reporting—can ramify and have important implications for reproducibility.

The complexity of the typical neuroimaging pipeline is visible from the earliest stages of data acquisition. While there are only three major manufacturers of MRI scanners, the machines run different pulse sequences, and even scanners from the same manufacturers do not often run the same software. At the preprocessing stage—even before statistical analysis—researchers face a bewildering array of options when considering when and how (or even whether) to account for subject movement, signal spikes, differences in brain anatomy, physiological confounds, and any number of standard concerns. Statistical analysis is no less complicated, as researchers must decide what kind of analyses to conduct (mass univariate, multivariate pattern classification, etc.), what search space to use (whole brain, specific regions of interest), what statistical

contrasts and multiple comparisons correction procedures to apply, and so on. The sheer magnitude of variation in analytical approaches underscores why computational reproducibility is so critical in cognitive neuroscience and why it has historically seemed so daunting. Put simply, without the ability to understand precisely what steps a research group took, it is doubtful that anyone else could ever reproduce the procedures.

In EEG, the diversity of methods is arguably even larger than in fMRI, with numerous manufacturers and a corresponding variety of technologies using different kinds of electrodes, amplifier settings, cap configurations, and software packages.[31] There have been some efforts at standardization of analysis methods through the release of software packages, such as the Matlab-based EEGLAB[32] and ERPLAB.[33] These packages have the advantage of allowing researchers to explore the data using a graphical interface while simultaneously generating an executable history script that records most of the analysis decisions. The BigEEG Consortium (www.bigeeg.org), an offshoot of the EEGLAB initiative, seeks to develop and promote data and metadata standards for EEG-based research that may eventually facilitate large-scale analysis and meta-analysis. But, by and large, EEG data collection and analysis involves equipment and workflows that vary considerably from laboratory to laboratory.

Of course, the complexity of neuroimaging data analysis is not itself the enemy. It is not the raw number of methodological and analytical choices per se that creates barriers to reproducibility; rather, the challenge lies in encoding those degrees of freedom in a standardized (and ideally, machine-readable) way. Fortunately, over time, the brain imaging community has converged on recommendations about what parameters should be reported and how.[34,35] Moreover, at least in fMRI, imaging data analysis software shows a significant degree of standardization. From the earliest days of human brain imaging, leading research groups in the United States and the United Kingdom wrote and freely distributed analysis software. This led to the widespread adoption of common tools with similar, although not identical, algorithms. Concerns about the inferential consequences of using one tool over another have been largely alleviated by findings from Gold[36] and Morgan,[37] and questions about the reliability of workflows using one or more tools have been addressed by Strother *et al.*[38,39] and others (but see Refs. 2,12, and 40). All of the major tools in common use—SPM, FSL, AFNI, and BrainVoyager—enable researchers to write scriptable workflows, built on internal engines (BrainVoyager), widely available commercial software (SPM-MATLAB), or free/open source software languages (Linux/unix shell, python, C/C++ for FSL and AFNI).

Naturally, there are some important caveats to this seemingly rosy picture. One concern is that, while existing software supports relatively standardized and highly processing workflows in principle, whether researchers actually take advantage of those features in practice is a separate matter. The number of SPM, AFNI, and BrainVoyager users who commonly rely exclusively on automated scripting in their analysis workflows, as opposed to using more user-friendly but inherently irreproducible graphical interfaces, is not known. We speculate that it is small. Moreover, even in laboratories that conduct fully automated analyses, the sharing or publication of the corresponding scripts or data processing pipelines remains rare.[40] While the differences between pipelines can be subtle,[40] the margin for error is also small (many published results only barely survive statistical correction), so a lack of full reporting can severely impair reproducibility.

A second caveat is that perfect reproducibility may be impossible to achieve even when a researcher is armed with all of the original data and scripts used to generate an analysis. Operating system differences, untracked differences in implicit software dependencies, and other factors can sometimes produce numerical discrepancies that, while initially small, may magnify as they cascade through a workflow to the point of introducing qualitative differences in results.[41] We discuss potential solutions to this problem (e.g., containerization) later; for present purposes, we note that acknowledging the intrinsic limits of methodological reproducibility does not grant researchers license to ignore best practices in automation and code sharing.

Importantly, brain imaging data are only part of the reproducibility story in cognitive neuroscience. It is also critical to understand how to reproduce the psychological components of cognitive neuroscience studies—most notably, the experimental design and its intended relationship to the latent constructs of interest. Here, the prospects for full reproducibility have historically seemed

less promising. Most experimental tasks involve the presentation of sequences of visual or auditory events and the collection of participants' behavioral responses—button presses, mouse movements or clicks, vocalizations, or eye movements—using computer programs that instantiate tasks custom tailored by a research team to address particular questions of interest. There have been researcher-initiated efforts to develop controlled vocabularies that describe the range of cognitive tasks deployed in the literature.[42,43] The National Institutes of Health (NIH) has spearheaded the creation of a standard toolbox of easily deployable tasks[44] and the development of data repositories designed to capture metadata about behavioral tasks and their variants[45] in standardized and searchable forms. However, such efforts notwithstanding, most cognitive neuroscience researchers employ customized tasks built using a variety of software and scripting environments (e.g., E-Prime, the Matlab-based psychophysics toolbox, PsychoPy, and DMDX). Tasks use customized image and sound components, and researchers rarely share the code, image, or sound files used in experimental tasks.[40] These practices limit the reproducibility of behavioral measures used in cognitive neuroscience and psychology as a whole.[7] Of course, the rigid standardization of tasks and materials has its own significant flaws, including the possibility of stifling innovation and slowing progress. We suggest that more widespread and open sharing of behavioral tasks, code, and materials provides a constructive middle ground.

Results reproducibility

Assuming that independent researchers are able to reproduce the methods of each other's studies, how closely do the findings generated converge? The answer is: It depends. In principle, even differences as basic as the brand of MRI scanner could undermine the ability to compare results across studies,[46] so considerable effort has gone into standardizing techniques that allow multisite imaging studies to be carried out in rigorous and reproducible ways. Fortunately, the viability of the basic technology is no longer in any serious doubt; abundant evidence demonstrates that all major brain imaging techniques are at least capable of producing highly convergent results across different sites and experimental procedures. Perhaps the best-known effort to demonstrate the basic robustness of results—not only in neuroimaging, but in other biomedical fields—is the Biomedical Informatics Research Network (BIRN; https://www.nitrc.org/projects/birn/). BIRN is a multisite, collaborative research consortium that strives to advance understanding of brain research and brain disease through the principles of data sharing and collaboration. There were several different BIRN initiatives, including the morphology BIRN, the mouse BIRN, and the function BIRN (fBIRN), among others. Although the fBIRN's disease focus was schizophrenia, considerable effort went into developing generalizable models for multisite data collection, best practices for research, and methods to facilitate the use of standardized processes across sites. One of fBIRN's biggest contributions was software that enabled the systematic investigation of how fMRI activation signals vary across sites, field strengths, and scanner platforms. The project also developed methods to control for these differences.[47] Scientists from fBIRN developed an automated quality assurance procedure based on a standard MRI phantom, and the team released freely available software that could be easily incorporated into any service center's data-transfer pipeline.[48] The fBIRN also provided leadership in modeling intersite reliability.[49] The same 18 participants were scanned at four different scanning sites. These analyses revealed that intersubject variability was 10 times greater than intersite variability; activation in many brain regions showed fair to good reliability; and measures of reliability increased with more runs of data.

More generally, the ability of techniques, such as fMRI, to produce robust and replicable findings is demonstrated by the rapid canonization of many initially surprising neuroimaging findings. For example, the tendency of a spatially conserved frontoparietal "task-positive" brain network to increase activity when participants engage in effortful cognitive activity has been replicated so often with fMRI over the past two decades[50–52] that the result is now often treated as a *de facto* manipulation check in new experiments. Large-scale meta-analyses of hundreds or even thousands of neuroimaging studies at a time further demonstrate a marked degree of convergence on stable neural correlates for most major psychological processes, from pain perception to episodic memory to language production.[26,53,54]

Of course, it is one thing to establish that neuroimaging methods can consistently reveal broad mappings between cognitive processes and distributed brain networks, and quite another to establish that the specific pattern of findings generated by any single study can be reproduced with a high degree of fidelity in another study. Unfortunately, as previous commentators have observed,[1,55] it is unclear whether neuroimaging findings meet this criterion. Arguably, the central problem is not that results reproducibility is particularly low, but that it has been difficult to quantify, leaving open the question of how much faith one should have in the results of any given published study. We focus on two critical barriers to the comprehensive assessment of results reproducibility in cognitive neuroscience, emphasizing fMRI (though the same concerns apply to other commonly used methods, such as EEG, MEG, and transcranial magnetic stimulation).

A first major challenge is that careful comparison of results across independent sites and studies typically requires that the full results be openly shared between sites, yet initiatives promoting neuroimaging data sharing have historically met with limited success. An early pioneer in data sharing was the fMRI Data Center (fMRIDC) at Dartmouth College, founded in 1999.[6,56,57] Around the same time, several journals, most notably the *Journal of Cognitive Neuroscience (JOCN),* tried to implement mandatory open data sharing—including deposition of raw data files (e.g., blood oxygen level–dependent (BOLD) time series, anatomical images)—as a requirement for publication. These efforts to foster increased transparency, while laudable, sparked controversy and backlash from the community. Opponents raised concerns about practical issues—technology, data formats, time and money constraints, and privacy—and cultural ones—the possibly negative impact of open sharing on individual scientific careers and advancement, questions about data ownership, and whether data sharing should be mandatory or optional.[57,58] The backlash eventually led *JOCN* to backstep on the data sharing requirement, and when funding to maintain the archive ran out, the fMRIDC stopped accepting new data. The fMRIDC's architects argue that, despite the setbacks, the fMRIDC should be viewed as a successful pioneer in open fMRI data sharing[56] whose experiences shaped the next generation of repositories, like the 1000 Functional Connectomes Project (FCP; fcon_1000.projects.nitrc.org/) and its International Neuroimaging Data Initiative (INDI),[58] the OpenfMRI project (openfmri.org) and NeuroVault (neurovault.org), the Human Connectome Project (humanconnectomeproject.org), and the National Institute of Mental Health–based National Database for Autism Research (ndar.nih.gov). More broadly, fMRIDC helped fuel interest in, recognition of, and support for the essential role that information infrastructure (neuroinformatics) plays in making widespread data sharing and reuse possible.

A second barrier to the evaluation of results reproducibility is a lack of consensus about what quantitative measures should be used.[1] One area of contention is whether the magnitude or spatial extent of task-related activation (or both) should be assessed. Measures of magnitude and spatial extent depend on criteria for determining which voxels are active, of course. Bennett and Miller[1] argued that the plurality of measures of reliability reported within individual studies made it challenging to ask about the reliability of findings across studies. They reported that the reliability of group-level results, using individual participants tested at different times, varied depending on the temporal gap between the tests, the specific tasks employed (sensory/motor versus cognitive), design factors (block versus event-related), the magnitude of activations, and other interindividual factors. Importantly, Miller *et al.*[59] and Costafreda *et al.*[60] found that, like the difference between within- and between-site variability found by the fBIRN team,[49] variability within participants across testing sessions was lower than variability between participants. Differences in tasks, the degree of selectivity of active voxels to those tasks, and subject motion appeared to be the biggest contributors to intersubject variability.[61] Nevertheless, Bennett and Miller[1] noted that all of the studies reporting test/retest reliabilities had small sample sizes, foreshadowing concerns about limited statistical power that others raised in the intervening years.[8,11]

In sum, we believe that, perhaps surprisingly given the size of the primary literature, the jury is still out on the degree to which researchers should expect individual neuroimaging findings to replicate when repeated under similar conditions. There is little doubt that methods like fMRI can produce highly replicable results, and that many canonical findings are indeed highly robust; however, as

the low-hanging fruit are plucked and researchers increasingly turn to subtler phenomena, it becomes more important for researchers to share data and results openly. Only in doing so can we progress toward consensus on criteria for evaluating the reproducibility of results.

Inferential reproducibility

The challenge of generating reproducible inferences—where independent researchers come to similar conclusions about what patterns of data mean—has been a central concern in the field for many years. Numerous published reviews have highlighted conceptual and statistical problems that threaten common neuroimaging inferences.[25,62,63] One source of concern is that the statistical power of most fMRI studies is well below conventionally adequate levels;[8–11] in a recent review based on over 1100 samples, Poldrack *et al.*[62] found that the median fMRI study in 2015 was underpowered to detect anything but relatively large effects (Cohen's d of ~ 0.75) even when using relatively high-powered procedures (i.e., a one-sample t-test). This observation is worrisome not only because low power implies a high false-negative rate and inefficient resource expenditure, but because it frequently leads to incorrect interpretations—the notion that effects are stronger and better localized than they actually are.[64,65] This increases the false-positive rate[8] across the literature as a whole.

A second set of concerns arises at the analysis stage. As the Bennett *et al.*[66] well-known "dead salmon" illustration showed, insufficiently stringent multiple corrections procedures can easily inflate the false-positive rate—an observation echoed by numerous studies that have highlighted limitations with common correction methods.[12,67–69] Moreover, such analyses all assume a best-case scenario under which researchers are not (inadvertently) capitalizing on the many "researcher degrees of freedom" available in a typical fMRI pipeline.[40] If one could formally account for P-hacking (i.e., data-dependent selection of analysis procedures), it is likely that the false-positive rate would rise, perhaps substantially.[11,70]

Last, even if one sets aside the statistical issues involved in the generation of cognitive neuroscience findings and assumes for the sake of argument that most published findings are fundamentally sound, it does not follow that researchers will agree about how to interpret such findings. Indeed, trenchant concerns have been raised about some of the most common assumptions researchers make when interpreting neuroimaging results, ranging from basic questions about what the BOLD signal reflects to what kind of information is actually extracted from multivariate pattern analysis.[71–74] Poldrack[75] flagged the problem of *reverse inference* as a particularly serious challenge, noting that the widespread approach of inferring mental function on the basis of the pattern of observed brain activity results runs a high risk of failure—unless it is supported by an appropriate Bayesian analysis that directly estimates the probability of a given task or state occurring conditional on an observed pattern of activity that is based on a reasonable prior distribution.[26]

Of course, science is a difficult enterprise, and it is easy to find serious methodological or statistical problems with virtually any piece of scientific research. The key question is what steps are researchers taking to address inferential concerns and to ensure that research findings continue to improve in reliability over time. To this end, we consider more recent initiatives aimed at improving the reproducibility of cognitive neuroscience research.

Recent initiatives

The focus on problems of reproducibility in scientific research as a whole has accelerated in the last several years,[16,19,20] and its scope extends well beyond psychological and neural science. As a result, cognitive neuroscience is both a beneficiary of new tools that promise to improve reproducibility and a contributor to them. We show that, fortunately, our field has already begun to embrace new, open, and transparent research practices that promise to mitigate or even eliminate many of the serious problems of methods, results, and inferential reproducibility.

Methods reproducibility

Concern about reproducible workflows and practices across the computational sciences has sharpened in similar ways.[76,77] While the specific practices that make computations reproducible vary from one field to another, Sandve *et al.*[78] summarized a set of steps that have broad applicability to cognitive neuroscientists. These include avoiding manual data manipulation steps (using scripts, not graphical user interfaces); keeping careful track of the provenance

(history) of all data, including derived results; tracking versions of all software and data; and providing public access to all code, outputs, and data.

Several data analysis tools have been developed based on free open-source languages, like R (RStudio; rstudio.com) and Python (Jupyter; jupyter.org). These tools support the creation of interactive electronic notebooks that combine data manipulation and analysis code along with graphic visualizations and text-based commentary. The tools can be used with version-control environments like git or mercurial, allowing the history of a project's data analysis to be captured. Version-control software can be used to store and share software, analyses, manuscripts, and documents written in virtually any language (both human and computer). Coupled with web-based repositories like GitHub (github.com), BitBucket (neurovault.org), or the Open Science Framework (OSF; osf.io), version-control systems enable researchers to share the histories and current status of all project materials and data. Some researchers concerned about computational reproducibility have gone even further, creating full software environments that can run a particular analysis and packaging them in a specialized "containerized" environment (e.g., Docker; www.docker.com) that can be distributed for others' use across a wide range of computer platforms. The use of electronic notebooks, version-control software, and web-based open data repositories has begun to enable cognitive neuroscience researchers to produce open and transparent workflows that can be readily reproduced. The authors use many of these techniques in their own research workflows.

Other efforts focus on methods reproducibility across study teams. One such initiative is the development of the Brain Imaging Data Structure (BIDS; bids.neuroimaging.io), a new open data format designed to facilitate the storage and sharing of data from brain imaging studies.[79] The BIDS attempts to achieve an easily implementable file directory and data structure that captures critical data and metadata about brain imaging studies and some data about the behavioral tasks performed by participants. The BIDS arose out of the work involved in creating the OpenFMRI (openfmri.org) data repository,[34] designed to allow researchers to openly share raw BOLD imaging data sets with sufficient information to permit re- or meta-analysis. The BigEEG project mentioned earlier represents a similar data format standardization initiative targeted at the EEG community.

On the data sharing side, modern platforms have picked up where pioneers like the fMRIDC left off, making it ever easier for researchers to distribute large neuroimaging data sets in a readily usable form. A major initiative focused on methods and results reproducibility is the Stanford Center for Reproducible Neuroscience (CRN; reproducibility.stanford.edu) formed in 2015 by Russell Poldrack and colleagues. The CRN is developing data repositories for both raw neuroimaging data sets (an upcoming successor to the OpenFMRI platform) and whole-brain statistical maps (NeuroVault.org).[80] A long-term goal of the CRN is not only to facilitate sharing, but also to provide containerized, modular, and fully reproducible cloud-based tools that can be easily executed via a graphical web interface. This will bring reproducible state-of-the-art neuroimaging data analysis within reach of researchers who lack the resources to deploy their own pipelines locally.[81]

One of us[82] has argued that many problems in reproducing the methods of behavioral studies could be ameliorated if video of all experimental procedures was more widely recorded and shared with researchers. Text-based methods sections with restrictive page or word limits simply cannot convey sufficiently detailed information about a study's methods so that it can be reproduced by another researcher. Sharing video can pose privacy risks, but the Databrary (databrary.org) digital library, a repository specialized for storing and sharing video, has developed a policy framework to share identifiable data with participant permission. Like the OSF, the Databrary has begun to serve as a web-based home for researchers to store and share data, metadata, and materials about the nonimaging-related portions of a study, including videos of experimental procedures, images, audio recordings, or displays. The Databrary largely focuses on developmental and learning science research now, but may expand in the future.

In sum, the field is making rapid strides to improve the reproducibility of methods, with the emergence of new tools, practices, centers, web-based data management systems, and data repositories.

Results reproducibility

Despite the acknowledged lack of consensus about how to measure and thereby evaluate results reproducibility[1] and the noted significant problems with statistical power, we find a number of encouraging developments concerning the reproducibility of results. Researchers continue to take seriously the effort to systematically measure the factors that influence test/retest reliability of responses across time and tasks, and these sorts of studies are increasingly common.[83–88] Other research programs focus on addressing questions about the long-term within-subject stability of responses[89,90] and how the accurate assessment of within-participant differences might address questions about individual differences.[91–93] There is increasing support for conducting and publishing the results of confirmatory studies,[94,95] thereby rectifying some existing biases that often favor the publication of new, novel results over confirmatory ones.

Several large-scale cross-site imaging studies whose results were designed to be widely shared with the research community have been undertaken (e.g., the Human Connectome Project and the U.K. Biobank Project). Findings from these studies are beginning to appear,[95] with results that both confirm and extend current understanding. Perhaps equally important is the extent to which planning for the large-scale sharing of these sorts of data has led to publication of extensive details about processing pipelines[96] and careful planning about how to make shared components useful to other researchers.

Policy makers and publishers have taken a renewed interest in how the context in which scientific research is conducted and results are shared can influence reproducibility. The Consortium for Reliability and Reproducibility has developed best-practice guidelines for the use of resting-state fMRI data available through the INDI archive,[96] and the Organization for Human Brain Mapping has created a Committee on Best Practice in Data Analysis and Sharing.[97] Following on the success of the ArXiv preprint service, increasing numbers of cognitive neuroscientists have begun to deposit article preprints in the BioRxiv preprint service (biorxiv.org/neuroscience). An effort specific to psychological science (PsyArxiv; osf.io/view/psyarxiv/) has begun with support from the Center for Open Science (cos.io) and the newly formed Society for the Improvement of Psychological Science (improvingpsych.org). High-profile generalist and topic-specific journals are adopting data sharing requirements reminiscent of those *JOCN* attempted to implement 15 years ago; there are new journals, such as Nature Publishing's *Scientific Data*, focused on creating citable, scholarly homes for well-curated data sets; and some journals (e.g., *Cortex*) have adopted a new publication format, the preregistered report, that conducts a review of the methods and analysis plan before data collection in exchange for a commitment to publish the results regardless of the findings.

There have also been recent developments to improve appropriate usage and reporting of statistical tests by means of automated tools, such as *statcheck* (statcheck.io),[98,99] which looks for elementary errors in the reporting of individual statistical tests. A related tool, *P-curve* (www.p-curve.com),[100] uses the complete set of statistical results from a body of work to estimate the evidentiary strength in favor of a hypothesis. While largely focused on the psychological science literature, these initiatives warrant close attention from cognitive neuroscientists, as they illustrate how the standardization of reporting practices can lead to insights about the quality of research practices and the strength or weakness of evidence across a broad published literature.[100] In the case of *statcheck*, the system depends on the fact that most experimental psychology papers report statistical analyses in ways that allow pertinent parameters to be automatically extracted from the published texts. Clearly, the diversity of efforts focused on bolstering the reproducibility of cognitive neuroscience results has considerable forward momentum.

Inferential reproducibility

Since the 2010 Bennett and Miller review on replicability, new tools and practices that promise to bolster the reproducibility of inferences have been created and are being adopted at an accelerating rate. We highlight three: meta-analysis, improved statistical practices, and machine learning.

For meta-analysis to succeed, the statistical effects from a large number of disparate studies must be collected, normalized, and reported in standardized ways.[101] The variability in analysis and reporting practices across the cognitive neuroscience literature can make meta-analysis challenging. As a result, the creation and curation of large-scale brain

imaging databases have been essential for the growth of meta-analysis as an inferential tool. One of the oldest such systems devoted to supporting meta-analytic data sets and software is the BrainMap project (www.brainmap.org) project.[5] As of late fall 2016, BrainMap consisted of data from more than 100,000 individual participants from nearly 4000 papers. The BrainMap data and meta-analysis tools have been used and cited more than 600 times since 1992, with more than 125 citations in 2016 alone.

Neurosynth (neurosynth.org)[26] takes an alternative approach to meta-analysis in which the raw data are (1) activation (*x,y,z*) coordinates mined from the text of imaging papers published in HTML on the web, combined with (2) word frequencies from the same papers. In this way, Neurosynth aims to automate and thereby standardize and accelerate the process of meta-analysis. By combining information about activation coordinates with term frequencies derived from calculating distributional statistics from the published articles, Neurosynth enables the analyst to interactively determine the extent of evidence for a relationship between a specific term of interest and a set of brain coordinates. For example, the analyst could visualize either the probability of a given voxel's activation given the existence of a specific term in the system's database of papers or the probability of a target term appearing in papers that report a particular voxel as active. The system allows users to view 3D maps of the conditional probabilities online, to download the maps for further analysis, and to create customized sets of searches. As of early 2017, Neurosynth contained data from more than 11,000 imaging studies, and it provides users with downloadable interactive meta-analyses from more than 3000 terms. The system has been cited almost 600 times, with more than 180 citations in 2016 alone. Of course, Neurosynth only supports meta-analysis on a subset of the published literature—older papers that were not published in easily parsable HTML formats and unpublished findings fall outside its scope.

Beyond meta-analysis, cognitive neuroscience research continues to push for new statistical procedures and the wider adoption of long-standing but more robust ones. For space reasons, we do not elaborate extensively here, but among the issues under active discussion are the appropriate handling of main effects and interaction tests,[102] the applicability of linear mixed-effects modeling techniques,[103] and the ongoing need to guard against the risks of false-positive results[3,104] even when using well-established, vetted, and widely used analysis software.[12] Still others suggest that the standard practice of treating stimulus effects as fixed and not random may undermine the generalizability of findings across studies.[105] An emergent theme is the ongoing need for vigorous and rigorous methodological reevaluation combined with a commitment to more open software publication practices. In the recent case of Eklund *et al.*,[12] the discovery of an error in the algorithm for controlling for cluster-wise fMRI activation effects quickly led to changes in the widely-used AFNI package.[106,107] The episode highlights the corrective, collaborative nature of open-source software development while also underscoring the uncomfortable reality that, at present, very few people who use open-source software packages actually bother to read the underlying code (the AFNI bug had previously gone undetected for many years).

In many other areas of social and computational science, progress has been facilitated by borrowing ideas and techniques from the field of *machine learning*. Philosophically, machine learning researchers tend to emphasize their ability to quantitatively predict key outcomes and pay less attention to traditional forms of scientific explanation.[108] This philosophy has led to the rapid proliferation of thousands of predictive modeling techniques—a number of which (e.g., support vector machines) are deployed regularly in cognitive neuroscience. Many fMRI studies are now framed as predictive problems in classification or regression, where the goal is to build a model that successfully discovers a mapping between a set of predictor variables and a set of discrete (in the case of classification) or continuous (in the case of regression) outcomes. For example, the distributed activation pattern of a large number of voxels in an fMRI data set can be used to predict successful versus unsuccessful attempts to recognize a stimulus. The resultant classifier can then be used to predict outcomes "out-of-sample" (i.e., in new data sets) and potentially also to aid in the interpretation of which voxels were likely to have played a role in processing relevant information. This gives rise to applications such as the "mind-reading" of representations present in the visual cortex during movie viewing[109,110] or revealing participants' semantic maps activated during narrative comprehension.[111]

Machine learning and related Big Data techniques have provided entirely new approaches to analysis. They allow researchers to capitalize on neural information patterns that may be too subtle or complex to be easily discovered using more conventional summary statistic approaches (e.g., brain activation maps or event-related potentials).[112] Of course, prediction-oriented approaches are not a panacea for standard concerns about the seeming ease with which researchers can fool themselves and unwittingly generate false or exaggerated findings. In the context of machine learning, the term *overfitting* is used to describe a case in which the predictions of an analysis have been inadvertently contaminated by noise in the data that were used to develop the analysis. As a consequence of overfitting, favorable results obtained when analyzing the same data set that was used to develop and calibrate the analysis will not be obtained when examining other, equivalent data sets. Some researchers who use machine learning take the notion of overfitting more seriously than others.[113] The cautious deploy methods, such as cross-validation (i.e., training and testing a model on independent subsets of the data), should in principle guard against overfitting. However, machine learning pipelines allow for considerably more analytical flexibility than conventional analyses. Researchers often have a choice among literally hundreds of different estimation approaches, each of which may have its own free parameters that require tuning to perform optimally. Whereas there is widespread awareness of the need to cross-validate results once a model has been selected, there is much less recognition that overfitting can still occur through the optimization of an analysis (for further discussion, see Refs. 114 and 115). Thus, as our field increasingly adopts machine learning techniques, it will be important to borrow established best practices from fields that have been using similar approaches for longer periods of time.[112,115]

The future

As Van Horn and Gazzaniga[56] observed, "the reality remains that very little of the neuroimaging data gathered each day in the field have been made available to those who could help provide much needed understanding." While we agree that this assessment still holds, we see other evidence that points toward a very different future. There is increasing recognition that greater openness and transparency, reflected in data, materials, and code sharing, offers individual investigators and the field as a whole far more benefits than risks.[28,58,108,116] While significant challenges remain in developing technology and workflow practices that make open and transparent workflows easy to generate and data readily shareable, that progress is being made. It is increasingly clear that there are substantial scientific rewards in analyzing or reanalyzing large-scale shared data sets beyond improving statistical power. Accordingly, while we take seriously the concerns many have raised recently about the methods, results, and inferential reproducibility of our field, we encourage our colleagues to embrace the newly emerging open science practices with an optimistic mindset,[56,117] as there is so much more to gain than to lose.

At the same time, it is essential that the field identify barriers that stand in the way of a more open, transparent, and reproducible neuroscience of cognition. One clear gap is the difficulty of capturing and reporting reproducible information about tasks, displays, and analysis procedures, although new data and materials repositories like the OSF and the Databrary, emerging data standards (BIDS; BigEEG), and pipelines (EEGLAB; nipype) can play constructive roles. Another concerns the need to forge community consensus around a set of principles about the culture in which cognitive neuroscience research is carried out—how to seek permission to share, when data and materials should be shared, how to measure and report individual scholarly contributions to large-scale studies, how to weigh the impact of analyses conducted on secondary data relative to the collection of new data, and how to ensure that the transition to more open science practices does not unduly harm the careers of the next generation of researchers. Providing answers to these questions goes beyond our scope, but we urge continued dialogue focused on achieving community consensus.

A vital question for which there remains no satisfying answer is what entity will pay for the curation, support, maintenance, and long-term storage of cognitive neuroscience data and materials. Data repositories, both past and present, have been funded either by short-term (3- to 5-year duration) NIH or National Science Foundation (NSF) research grants or private foundation funders. Thus, despite increasingly strong encouragements from granting

agencies to share data and materials or even mandates to do so, there is as yet no long-term commitment from the agencies for funding devoted to long-term data preservation. Data curation, storage, and preservation are not inexpensive, and the problem of how to sustain research infrastructure that benefits the entire research community will neither solve itself nor go away. Nevertheless, on the basis of the success of other fields like astronomy, high-energy physics, and the geosciences, we think that a strong case can be made for enduring federal and private donor support for research infrastructure that empowers cognitive neuroscientists to openly share data, materials, and methods.

Fundamentally, we think that investments in the future of cognitive neuroscience infrastructure will generate big payoffs. Fostering the widespread adoption of open, transparent, and reproducible research practices coupled with innovations in technology that enable the large-scale analysis of our particular store of Big Data will accelerate the discovery of generalizable, robust, and meaningful findings about the nature and origins[118] of human cognition.

Acknowledgments

R.O.G. acknowledges support from NSF BCS-1147440, NSF BCS-1238599, and NICHD U01-HD-076595. B.A.W. acknowledges support from NSF BCS-1331073.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Bennett, C.M. & M.B. Miller. 2010. How reliable are the results from functional magnetic resonance imaging? *Ann. N.Y. Acad. Sci.* **1191:** 133–155.
2. Carp, J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* **63:** 289–300.
3. Vul, E., C. Harris, P. Winkielman & H. Pashler. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* **4:** 274–290.
4. Biswal, B.B., M. Mennes, X.-N. Zuo, *et al.* 2010. Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* **107:** 4734–4739.
5. Fox, P.T., S. Mikiten, G. Davis & J.L. Lancaster. 1994. BrainMap: a database of human functional brain mapping. In *Functional Neuroimaging: Technical Foundations.* R.W. Thatcher, M. Hallett, T. Zeffiro, E.R. John & M. Huerta, Eds.: 95–105. Orlando: Academic Press, Inc.
6. Van Horn, J.D. & M.S. Gazzaniga. 2002. Databasing fMRI studies—towards a "discovery science" of brain function. *Nat. Rev. Neurosci.* **3:** 314–318.
7. Open Science Collaboration. 2015. Psychology. Estimating the reproducibility of psychological science. *Science* **349:** aac4716.
8. Button, K.S., J.P.A. Ioannidis, C. Mokrysz, *et al.* 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14:** 365–376.
9. David, S.P., J.J. Ware, I.M. Chu, *et al.* 2013. Potential reporting bias in fMRI studies of the brain. *PLoS One* **8:** e70104.
10. Ioannidis, J.P.A., M.R. Munafò, P. Fusar-Poli, *et al.* 2014. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.* **18:** 235–241.
11. Szucs, D. & J.P. Ioannidis. 2016. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *bioRxiv* 071530.
12. Eklund, A., T.E. Nichols & H. Knutsson. 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* **113:** 7900–7905.
13. Dolgin, E. 2010. This is your brain online: the functional connectomes project. *Nat. Med.* **16:** 351.
14. Van Essen, D.C., S.M. Smith, D.M. Barch, *et al.* 2013. The WU-Minn Human Connectome Project: an overview. *NeuroImage* **80:** 62–79.
15. Adolph, K.E., R.O. Gilmore, C. Freeman, *et al.* 2012. Toward open behavioral science. *Psychol. Inq.* **23:** 244–247.
16. Nosek, B.A. & Y. Bar-Anan. 2012. Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* **23:** 217–243.
17. Pashler, H. & C.R. Harris. 2012. Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* **7:** 531–536.
18. Poldrack, R.A. & J.-B. Poline. 2015. The publication and reproducibility challenges of shared data. *Trends Cogn. Sci.* **19:** 59–61.
19. Munafò, M.R., B.A. Nosek, D.V.M. Bishop, *et al.* 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* **1:** 0021.
20. Goodman, S.N., D. Fanelli & J.P.A. Ioannidis. 2016. What does research reproducibility mean? *Sci. Transl. Med.* **8:** 3.
21. Nosek, B.A., G. Alter, G.C. Banks, *et al.* 2015. Promoting an open research culture. *Science* **348:** 1422–1425.
22. Patil, P., R.D. Peng & J.T. Leek. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* **11:** 539–544.
23. Gilbert, D.T., G. King, S. Pettigrew & T.D. Wilson. 2016. Comment on "Estimating the reproducibility of psychological science." *Science* **351:** 1037.
24. Etz, A. & J. Vandekerckhove. 2016. A Bayesian perspective on the reproducibility project: psychology. *PLoS One* **11:** e0149794.
25. Yarkoni, T., R.A. Poldrack, D.C. Van Essen & T.D. Wager. 2010. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* **14:** 489–496.

26. Yarkoni, T., R.A. Poldrack, T.E. Nichols, *et al.* 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8:** 665–670.
27. Nosek, B.A., J.R. Spies & M. Motyl. 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7:** 615–631.
28. Poldrack, R.A. & K.J. Gorgolewski. 2014. Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* **17:** 1510–1517.
29. Poline, J.-B., J.L. Breeze, S.S. Ghosh, *et al.* 2012. Data sharing in neuroimaging research. *Front. Neuroinform.* **6:** 9.
30. Wicherts, J.M., M. Bakker & D. Molenaar. 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One* **6:** e26828.
31. Luck, S.J. 2014. *An Introduction to the Event-Related Potential Technique.* MIT Press.
32. Delorme, A. & S. Makeig. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *J. Neurosci. Methods* **134:** 9–21.
33. Lopez-Calderon, J. & S.J. Luck. 2014. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **8:** 213.
34. Poldrack, R.A., D.M. Barch, J. Mitchell, *et al.* 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* **7:** 12.
35. practiCal fMRI: the nuts & bolts. 2013. A checklist for fMRI acquisition methods reporting in the literature. Accessed February 21, 2017. https://practicalfmri.blogspot.com/2013/01/a-checklist-for-fmri-acquisition.html.
36. Gold, S., B. Christian, S. Arndt, *et al.* 1998. Functional MRI statistical software packages: a comparative analysis. *Hum. Brain Mapp.* **6:** 73–84.
37. Morgan, V.L., B.M. Dawant, Y. Li & D.R. Pickens. 2007. Comparison of fMRI statistical software packages and strategies for analysis of images containing random and stimulus-correlated motion. *Comput. Med. Imaging Graph.* **31:** 436–446.
38. Strother, S.C., J. Anderson, L.K. Hansen, *et al.* 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage* **15:** 747–771.
39. Strother, S., S. La Conte, L. Kai Hansen, *et al.* 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage* **23**(Suppl. 1): S196–S207.
40. Carp, J. 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Brain Imag. Methods* **6:** 149.
41. Glatard, T., L.B. Lewis, R. Ferreira da Silva, *et al.* 2015. Reproducibility of neuroimaging analyses across operating systems. *Front. Neuroinform.* **9:** 12.
42. Turner, J.A. & A.R. Laird. 2011. The cognitive paradigm ontology: design and application. *Neuroinformatics* **10:** 57–66.
43. Poldrack, R.A., A. Kittur, D. Kalar, *et al.* 2011. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinform.* **5:** 17.
44. Gershon, R.C., D. Cella, N.A. Fox, *et al.* 2010. Assessment of neurological and behavioural function: the NIH toolbox. *Lancet Neurol.* **9:** 138–139.
45. Hall, D., M.F. Huerta, M.J. McAuliffe & G.K. Farber. 2012. Sharing heterogeneous data: the National Database for Autism Research. *Neuroinformatics* **10:** 331–339.
46. Casey, B.J., J.D. Cohen, K. O'Craven, *et al.* 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage* **8:** 249–261.
47. Friedman, L., H. Stern, G.G. Brown, *et al.* 2008. Test–retest and between-site reliability in a multicenter fMRI study. *Hum. Brain Mapp.* **29:** 958–972.
48. Friedman, L. & G.H. Glover. 2006. Report on a multicenter fMRI quality assurance protocol. *J. Magn. Reson. Imaging* **23:** 827–839.
49. Brown, G.G., D.H. Mathalon, H. Stern, *et al.* 2011. Multisite reliability of cognitive BOLD data. *NeuroImage* **54:** 2163–2175.
50. Braver, T.S., J.D. Cohen, L.E. Nystrom, *et al.* 1997. A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage* **5:** 49–62.
51. Dosenbach, N.U.F., K.M. Visscher, E.D. Palmer, *et al.* 2006. A core system for the implementation of task sets. *Neuron* **50:** 799–812.
52. Duncan, J. 2010. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14:** 172–179.
53. Smith, S.M., P.T. Fox, K.L. Miller, *et al.* 2009. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.* **106**: 13040–13045.
54. Vigneau, M., V. Beaucousin, P.Y. Hervé, *et al.* 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *NeuroImage* **30:** 1414–1432.
55. Yarkoni, T. & T.S. Braver. 2010. Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological issues. In *Handbook of Individual Differences in Cognition.* A. Gruszka, G. Matthews & B. Szymura, Eds.: 87–107. New York: Springer.
56. Van Horn, J.D. & M.S. Gazzaniga. 2013. Why share data? Lessons learned from the fMRIDC. *NeuroImage* **82:** 677–682.
57. Van Horn, J.D., J.S. Grethe, P. Kostelec, *et al.* 2001. The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356:** 1323–1339.
58. Mennes, M., B. Biswal, F.X. Castellanos & M.P. Milham. 2013. Making data sharing work: the FCP/INDI experience. *NeuroImage* **82:** 683–691.
59. Miller, M.B., C.-L. Donovan, J.D. Van Horn, *et al.* 2009. Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *NeuroImage* **48:** 625–635.
60. Costafreda, S.G., M.J. Brammer, R.Z.N. Vêncio, *et al.* 2007. Multisite fMRI reproducibility of a motor task using

identical MR systems. *J. Magn. Reson. Imaging* **26:** 1122–1126.
61. Duncan, K.J., C. Pattamadilok, I. Knierim & J.T. Devlin. 2009. Consistency and variability in functional localisers. *NeuroImage* **46:** 1018–1026.
62. Poldrack, R.A., C.I. Baker, J. Durnez, *et al.* 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18:** 115–126.
63. Thirion, B., P. Pinel, S. Mériaux, *et al.* 2007. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. *NeuroImage* **35:** 105–120.
64. Yarkoni, T. 2009. Big correlations in little studies: inflated fMRI correlations reflect low statistical power—commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* **4:** 294–298.
65. Kriegeskorte, N., M.A. Lindquist, T.E. Nichols, *et al.* 2010. Everything you never wanted to know about circular analysis, but were afraid to ask. *J. Cereb. Blood Flow Metab.* **30:** 1551–1557.
66. Bennett, C., M. Miller & G. Wolford. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. *NeuroImage* **47:** S125.
67. Bennett, C.M., G.L. Wolford & M.B. Miller. 2009. The principled control of false positives in neuroimaging. *Social Cogn. Affect. Neurosci.* **4:** 417–422.
68. Woo, C.-W., A. Krishnan & T.D. Wager. 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage* **91:** 412–419.
69. Smith, S.M. & T.E. Nichols. 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* **44:** 83–98.
70. Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Med.* **2:** e124.
71. Logothetis, N.K. 2008. What we can do and what we cannot do with fMRI. *Nature* **453:** 869–878.
72. Cole, D.M., S.M. Smith & C.F. Beckmann. 2010. Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Front. Syst. Neurosci.* **4:** 8.
73. Poldrack, R.A. 2010. Subtraction and beyond: the logic of experimental designs for neuroimaging. In *Foundational Issues in Human Brain Mapping.* S.J. Hanson & M. Bunzl, Eds. MIT Press.
74. Etzel, J.A., J.M. Zacks & T.S. Braver. 2013. Searchlight analysis: promise, pitfalls, and potential. *NeuroImage* **78:** 261–269.
75. Poldrack, R.A. 2006. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10:** 59–63.
76. Peng, R.D. 2011. Reproducible research in computational science. *Science* **334:** 1226–1227.
77. Stodden, V. 2012. Reproducible research: tools and strategies for scientific computing. *Comput. Sci. Eng.* **14:** 11–12.
78. Sandve, G.K., A. Nekrutenko, J. Taylor & E. Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9:** e1003285.
79. Gorgolewski, K.J., T. Auer, V.D. Calhoun, *et al.* 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3:** 160044.
80. Gorgolewski, K.J., G. Varoquaux, G. Rivera, *et al.* 2015. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* **9:** 8.
81. Gorgolewski, K.J., F. Alfaro-Almagro, T. Auer, *et al.* 2016. BIDS Apps: improving ease of use, accessibility and reproducibility of neuroimaging data analysis methods. *bioRxiv* 079145.
82. Gilmore, R.O. & K.E. Adolph. Open sharing of research video—breaking down the boundaries of the research team. In *Advancing Social and Behavioral Health Research through Cross-Disciplinary Team Science: Principles for Success.* K. Hall, R. Croyle & A. Vogel, Eds. Springer. In press.
83. Jann, K., D.G. Gee, E. Kilroy, *et al.* 2015. Functional connectivity in BOLD and CBF data: similarity and reliability of resting brain networks. *NeuroImage* **106:** 111–122.
84. Jovicich, J., M. Marizzoni, B. Bosch, *et al.* 2014. Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *NeuroImage* **101:** 390–403.
85. Koolschijn, P.C.M.P., M.A. Schel, M. de Rooij, *et al.* 2011. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test–retest reliability from childhood to early adulthood. *J. Neurosci.* **31:** 4204–4212.
86. Liao, X.-H., M.-R. Xia, T. Xu, *et al.* 2013. Functional brain hubs and their test–retest reliability: a multiband resting-state functional MRI study. *NeuroImage* **83:** 969–982.
87. Madhyastha, T., S. Mérillat, S. Hirsiger, *et al.* 2014. Longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging. *Hum. Brain Mapp.* **35:** 4544–4555.
88. Marchitelli, R., L. Minati, M. Marizzoni, *et al.* 2016. Test–retest reliability of the default mode network in a multi-centric fMRI study of healthy elderly: effects of data-driven physiological noise correction techniques. *Hum. Brain Mapp.* **37:** 2114–2132.
89. Choe, A.S., C.K. Jones, S.E. Joel, *et al.* 2015. Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years. *PLoS One* **10:** e0140134.
90. Poldrack, R.A., T.O. Laumann, O. Koyejo, *et al.* 2015. Long-term neural and physiological phenotyping of a single human. *Nat. Commun.* **6:** 8885.
91. Dubois, J. & R. Adolphs. 2016. Building a science of individual differences from fMRI. *Trends Cogn. Sci.* **20:** 425–443.
92. Boekel, W., E.-J. Wagenmakers, L. Belay, *et al.* 2015. A purely confirmatory replication study of structural brain–behavior correlations. *Cortex* **66:** 115–133.
93. Muhlert, N. & G.R. Ridgway. 2016. Failed replications, contributing factors and careful interpretations: commentary on Boekel *et al.*, 2015. *Cortex* **74:** 338–342.
94. Glasser, M.F., T.S. Coalson, E.C. Robinson, *et al.* 2016. A multi-modal parcellation of human cerebral cortex. *Nature* **536:** 171–178.
95. Glasser, M.F., S.N. Sotiropoulos, J.A. Wilson, *et al.* 2013. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80:** 105–124.
96. Zuo, X.-N., J.S. Anderson, P. Bellec, *et al.* 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* **1:** 140049.

97. Nichols, T.E., S. Das, S.B. Eickhoff, *et al.* 2016. Best practices in data analysis and sharing in neuroimaging using MRI. *bioRxiv* 054262.
98. Epskamp, S. & M.B. Nuijten. 2016. statcheck: extract statistics from articles and recompute *p* values. http://CRAN.R-project.org/package=statcheck.
99. Nuijten, M.B., C.H.J. Hartgerink, M.A.L.M. van Assen, *et al.* 2015. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* **48:** 1–22.
100. Simonsohn, U., L.D. Nelson & J.P. Simmons. 2014. *P*-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143:** 534.
101. Wager, T.D., M. Lindquist & L. Kaplan. 2007. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* **2:** 150–158.
102. Nieuwenhuis, S., B.U. Forstmann & E.-J. Wagenmakers. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* **14:** 1105–1107.
103. Chen, G., Z.S. Saad, J.C. Britton, *et al.* 2013. Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage* **73:** 176–190.
104. Kriegeskorte, N., W.K. Simmons, P.S.F. Bellgowan & C.I. Baker. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12:** 535–540.
105. Westfall, J., T. Nichols & T. Yarkoni. 2016. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *bioRxiv* 077131.
106. Cox, R.W., R.C. Reynolds & P.A. Taylor. 2016. AFNI and clustering: false positive rates redux. *bioRxiv* 065862.
107. Eickhoff, S.B., A.R. Laird, P.M. Fox, *et al.* 2016. Implementation errors in the GingerALE software: description and recommendations. *Hum. Brain Mapp.* **38:** 7–11.
108. Breiman, L. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16:** 199–231.
109. Kay, K.N., T. Naselaris, R.J. Prenger & J.L. Gallant. 2008. Identifying natural images from human brain activity. *Nature* **452:** 352–355.
110. Nishimoto, S., A.T. Vu, T. Naselaris, *et al.* 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21:** 1641–1646.
111. Huth, A.G., W.A. de Heer, T.L. Griffiths, *et al.* 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532:** 453–458.
112. Varoquaux, G., P.R. Raamana, D. Engemann, *et al.* 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* **145:** 166–179.
113. Yarkoni, T. & J. Westfall. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* In press.
114. Cawley, G.C. & N.L.C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11:** 2079–2107.
115. Skocik, M., J. Collins, C. Callahan-Flintoft, *et al.* 2016. I tried a bunch of things: the dangers of unexpected overfitting in classification. *bioRxiv* 078816.
116. Bandrowski, A.E. & M.E. Martone. 2016. RRIDs: a simple step toward improving reproducibility through rigor and transparency of experimental methods. *Neuron* **90:** 434–436.
117. Ascoli, G.A. 2006. The ups and downs of neuroscience shares. *Neuroinformatics* **4:** 213–215.
118. Gilmore, R.O. 2016. From big data to deep insight in developmental science. *Wiley Interdiscip. Rev. Cogn. Sci.* **7:** 112–126.