

Reproducibility in Neuroimaging Analysis: Challenges and Solutions

Rotem Botvinik-Nezer and Tor D. Wager

ABSTRACT

Recent years have marked a renaissance in efforts to increase research reproducibility in psychology, neuroscience, and related fields. Reproducibility is the cornerstone of a solid foundation of fundamental research—one that will support new theories built on valid findings and technological innovation that works. The increased focus on reproducibility has made the barriers to it increasingly apparent, along with the development of new tools and practices to overcome these barriers. Here, we review challenges, solutions, and emerging best practices with a particular emphasis on neuroimaging studies. We distinguish 3 main types of reproducibility, discussing each in turn. Analytical reproducibility is the ability to reproduce findings using the same data and methods. Replicability is the ability to find an effect in new datasets, using the same or similar methods. Finally, robustness to analytical variability refers to the ability to identify a finding consistently across variation in methods. The incorporation of these tools and practices will result in more reproducible, replicable, and robust psychological and brain research and a stronger scientific foundation across fields of inquiry.

<https://doi.org/10.1016/j.bpsc.2022.12.006>

The last decade has marked a prominent shift in focus toward reproducibility across many fields. Converging evidence from multiple explorations and large-scale collaborations has indicated that many published findings are potentially false positives (1–6). Consequently, a “renaissance” [as termed by (7)] was sparked, with researchers re-examining the scientific endeavor, identifying problems—from lack of methodological rigor to misaligned incentive structures—and working on solutions. This renaissance included heated debates alongside calls for fundamental changes in how research is performed and evaluated. Critical, constructive steps include the ongoing development of new tools and approaches to increase the reproducibility of scientific findings (8).

The abundance of new tools requires substantial adaptations from researchers, who are now expected to remain just as productive while implementing practices that often require more time and a broader set of skills than ever before. Thus, it is understandable that the adoption of these tools and practices is relatively slow (9–11). Nevertheless, increasing the reproducibility of research is vital to its advancement and to the ability to translate findings into integrative theories and clinical interventions.

The reproducibility of scientific findings is fundamental to their validity and utility in guiding further research, technological development, and treatment development. Reproducibility is therefore central to the return on investment of public money dedicated to scientific research. For example, results that cannot be reproduced using the same data and methods used to derive them are likely invalid. Publishing irreproducible results is worse than not publishing at all, because it is more difficult to eliminate an idea than it is to introduce it (12–14). Spurious results could mislead other researchers who conduct

follow-up investigations or try to integrate findings into broader theories. Moreover, practices such as data sharing increase the value of datasets by allowing their reuse (15,16).

An important development has been to increasingly align incentives with reproducibility. Researchers are increasingly recognized for reproducibility efforts through awards and citations (17). Concomitantly, funding agencies increasingly require researchers to implement reproducibility practices (18). For example, since 2019, the U.S. National Institute of Mental Health requires funded projects to share their data via the National Institute of Mental Health Data Archive (<https://nda.nih.gov/about/policy.html>).

In this review, we aim to facilitate researchers' transition toward increasingly reproducible research, with a focus on neuroimaging in psychiatric research. We provide an overview of challenges, potential solutions, and best practices related to 3 main types of reproducibility (see Figure 1): 1) the ability to reproduce the same results using the original data and methods (analytical reproducibility); 2) the ability to replicate findings with the same methods but new data (replicability); and 3) the ability to reproduce similar or converging results with the same data and hypotheses but different methods (robustness to analytical variability) (19). These complement other important properties of useful scientific research, including generalizability, interpretability, and translatability, which are also critical for future progress.

ANALYTICAL REPRODUCIBILITY: SAME DATA, SAME METHODS

The minimal requirement of reproducible research is that results will be precisely reproduced when the same data and

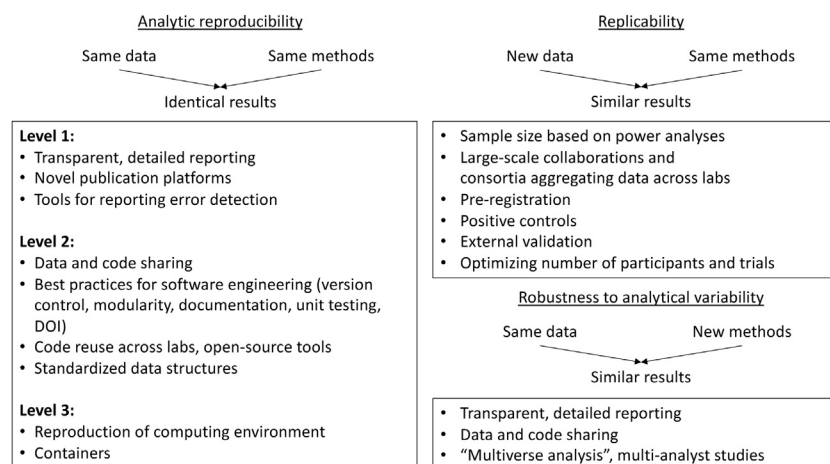


Figure 1. The 3 types of reproducibility covered in this review, along with a summary of the main solutions. DOI, digital object identifier.

methods are used. While this may seem like an obvious feature, it has been shown to be much more challenging than expected. For example, Hardwicke *et al.* (20) attempted to reproduce results from 25 published papers that publicly shared their data and code (and thus might be particularly likely to be reproducible). They found substantial numerical discrepancies between reported statistical values and values obtained from reproduction attempts in 64% of these papers (or 28%, following input from the original authors). Such discrepancies do not necessarily imply that the original studies' conclusions are wrong, but they do limit confidence in their accuracy and raise questions about how they were produced.

This example highlights the fact that reproducibility requires a complete description of the methods used to collect, process, and analyze data. This is a fundamental challenge, as textual descriptions are often vague or insufficient to describe complex procedures—for example, stating that “a linear model was fitted” is not enough. Unless all variables included in the model are described, including covariates, interaction terms, and coding of categorical variables, both interpretability and reproducibility are compromised. The Reproducibility Project: Cancer Biology illustrated the extent of this problem by attempting to replicate 193 experiments from 53 papers in preclinical cancer biology (21). All 193 studies were missing important details in the published description and therefore could not be replicated based on the published paper alone. Attempted replication was possible in only 50 of the original 193 studies, even after attempting to get further information from the original authors. Other attempts to reproduce published papers have encountered similar challenges (22–24).

This issue is further complicated by the uncertainty regarding which details are ignorable. For example, the model of the analysis computer and the operating system used are widely considered ignorable but have been shown to affect results (25,26). Here, we describe 3 levels of defense that enable increasingly precise reproduction of published results.

The first level is standardization of reporting practices. Many societies have produced consensus documents on methods and reporting (27,28), and journals increasingly require detailed methods checklists, though notably these guidelines are rarely

enforced. Novel publication platforms allow publication of objects beyond static text and two-dimensional visualizations [e.g., *Aperture* journal and NeuroLibre preprint server (29)]. A related, promising development is the emergence of automatic tools for error detection. For example, *statcheck* (<http://statcheck.io/>) helps automatically detect inconsistencies between statistical values, degrees of freedom, and corresponding *p* values and has already been implemented by some journals (30).

The second level is data and code sharing. Many repositories are available for sharing data and code (e.g., Github, Open Science Framework). Open sharing enables readers to reproduce the analysis, understand precisely what was done, and reuse methods and data in future studies. However, sharing does not guarantee utility, and useful sharing will require a shift in scientific training and allocation of time and resources. There are currently few standards in place for code readability, reusability, and error checking. Most scientists were not formally trained in software engineering and have not dedicated years to perfecting their code—nor is it practical for them to do so. The result, however, is that code and datasets are often not sufficiently clear and well documented to allow reproducibility. Even if results are reproducible, code errors may make the original results incorrect.

Several good practices can help reduce these problems. First, while scientists must learn many things beyond software coding, it is increasingly essential for trainees to obtain some proficiency in coding and data science. Trainees and established researchers alike can benefit from adopting best practices in code writing (31–34), including version control (35), code modularity, code documentation (36,37), unit testing (38), and versioned, documented code releases to clearly track which version was used to produce published results (e.g., with Zenodo).

Reuse of code across user groups is another crucial advantage, as it increases the chances that errors will be discovered. For example, the machine learning toolbox *scikit-learn* (39) is used by thousands of groups worldwide. Along with its extensible design and clear structure, it provides a robust analysis platform. In neuroimaging, there is substantial

Reproducible Neuroimaging Analysis

convergence on popular free, open-source toolboxes. SPM (40), the FMRIB Software Library (41), and AFNI (42) packages are used by numerous groups. Many other open-source toolboxes complement these tools or provide specialized functions. For example, the Cognitive and Affective Neuroscience Lab neuroimaging analysis tools (canlab.github.io) is an object-oriented toolbox that enables complex analytic processes to be run with simple commands, resulting in readable and reusable analysis scripts. The Group ICA of fMRI Toolbox (GIFT; <https://trendscenter.org/software/gift/>) performs multi-variate analysis functions not covered in other packages. Both reuse SPM functions for image reading and writing, building off of a heavily vetted code base for basic functions.

As for data sharing, this key practice has become much more common with the emergence of sharing platforms and regulations by some funding agencies and journals (15,43,44). Publications are now expected to be accompanied by openly shared data, unless there are specific ethical constraints. Importantly, enforcement of these requirements is still lacking, which has slowed progress. Free platforms are available for neuroimaging data sharing, such as OpenNeuro (45).

Standardized data structures are making it easier for researchers to understand and use shared data. For example, the National Institute of Mental Health Data Archive uses common data elements that identify the same type of data across studies (e.g., the same question in a questionnaire). This facilitates making data findable, accessible, interoperable, and reusable (46). Many of the most important data standardization efforts are community-driven. For example, the brain imaging data structure (BIDS) (47), a widely accepted standard for organizing neuroimaging data, has greatly facilitated sharing and reuse. This has allowed the development of BIDS Apps that perform analytic processes on these data structures (48). Open tools for research data management, such as the Neuroimaging Data Model (49,50) and DataLad (51), further help researchers organize, manage, track, and share their data (10,52).

The third level comprises tools for managing the computing environment. As mentioned above, even minor deviations from the original computing environment (e.g., a different software version) could result in meaningful differences in results (25,26,53,54). Thus, tools that reproduce the computing environment do not affect the quality of results but are crucial for analytic reproducibility. For example, R's checkpoint (<https://github.com/RevolutionAnalytics/checkpoint>) and Groundhog (<https://groundhogr.com>) packages install other packages with the versions that were available on a specific chosen date, thus preventing discrepancies due to changes in code over time. Containers go a step further. They are standalone executable software packages that include an operating system, code, and all dependencies (e.g., required packages and software, system tools, and settings, etc.) to run an analysis. Popular container platforms are Docker and Singularity (55) (see also <https://www.repronim.org/neurodocker/>).

In summary, a wide variety of practices, tools, and standards for improving analytical reproducibility have emerged over the past decade and come into widespread use. This is encouraging, as reproducibility is the most basic expectation for published results and a gateway to replicability and robustness. We turn to these next.

REPLICABILITY: NEW DATA, SAME METHODS

Replicability is the ability to reproduce findings using the same or similar methods in new samples. The first study to provide direct large-scale evidence on replicability of findings across a field was the Reproducibility Project: Psychology. The Open Science Collaboration attempted to replicate 100 prominent findings in psychology (2). Strikingly, only 36 of the 100 studies were successfully replicated, with effect sizes about half the size of the original effect sizes, on average. Similar projects in different fields followed, revealing a replication crisis that spans many scientific fields, from social sciences (56) to economics (57), experimental philosophy (58), and preclinical cancer biology (21,59).

Another series of large-scale collaborations, the Many Labs studies, focused on providing broader evidence on the replicability of findings in the social sciences and testing potential moderating variables. Many Labs I successfully replicated 10 of 13 effects (findings) across 36 independent samples and provided evidence that replicability depends more on the effect being studied than on the specific sample or settings (e.g., online vs. in-lab settings) (60). Many Labs II went bigger, attempting to replicate 28 effects, with about 60 preregistered, peer-reviewed protocols per effect. The study successfully replicated 15 effects (54%). Like Many Labs I, replicability depended mainly on the effect, rather than the sample or context (61). One of the main claims against the findings of the Reproducibility Project: Psychology was that some effects failed to replicate because of lack of adherence to expert review or low power. To test this, Many Labs 5 conducted multiple additional replication attempts of each of 10 of these effects, with preregistered, peer-reviewed protocols (62). These revised replication attempts produced effect sizes similar to those of the Reproducibility Project: Psychology, providing evidence against this claim.

Similar projects are emerging in various fields, such as developmental psychology and electroencephalography (63–65). An exciting development is the emergence of scientific organizations that promote cross-lab replication. In psychology, the Many Labs projects have led to the foundation of the Psychological Science Accelerator, a globally distributed network of psychology labs that coordinates data collection for democratically selected studies (66).

Sample Size, Effect Size, and Replicability

Replicability depends on statistical power. Underpowered studies become a big problem when combined with a publication bias toward positive findings (67). Notably, sample sizes in neuroscience are often underpowered (68,69). For example, a recent study concluded that replicable brainwide association study–associations between individual differences in brain structure or function and complex cognitive or mental health phenotypes often have very small effects (e.g., $r < 0.15$) and thus require thousands of participants for high replicability, far more than the typical sample sizes of dozens of participants (70). Moreover, the reliability of phenotypic measurements common in psychiatric research are typically low, limiting the ability to develop replicable and reliable markers, irrespective of sample size and the type of data used to predict these phenotypes (71).

This is a sobering analysis, but there is cause for hope. Effect sizes from multivariate predictive models are often several times larger than univariate brainwide association studies (e.g., $r \approx 0.4$). This confers a dramatic increase of power (72). In addition, the brainwide association study analyses did not consider within-person effects, requiring much fewer participants for high replicability. Finally, many of the limitations in replicability stem from the need to replicate many small effects in individual brain regions. If multivariate analyses are used to define integrated measures that aggregate across brain regions, effect sizes are larger (73,74) and the multiple testing problem is eliminated (75,76). Nevertheless, while such machine learning approaches address some issues (e.g., power), they raise new challenges, including data leakage, sample variability, and the need for sufficiently large test sets for robustness, depending on the research aim, level of analysis, potential confounding factors, effect sizes, and more (77–83).

At the same time, large collaborations such as the Human Connectome Project (84), UK Biobank (85), and Adolescent Brain Cognitive Development (ABCD) Study (86) are generating mega large databases that could be used to achieve more replicable findings. However, these projects mostly include well-studied tasks and cannot replace smaller studies of more novel effects, rare populations, or specific experimental designs. To fill this gap, consortia have formed to aggregate data across labs into large samples [e.g., (87)], although they face the challenge of data harmonization across studies and sites in multisite studies (88–90).

An overall lesson is that researchers should design their studies to be statistically powered (see <https://brainpower.readthedocs.io/en/latest/index.html> for resources), considering the effect size, type of association (between or within-person), and analysis (univariate or multivariate) (91). Notably, power analysis for neuroimaging data is particularly challenging, in part because brain regions are intercorrelated in complex ways, and the current standard is study-specific multiple comparison corrections that accommodate the characteristics of the individual dataset. Furthermore, these standards are also being debated and revised (92–94). Finally, few unbiased a priori estimates of effect sizes are available in most areas of study (95). One approach is to pick a minimal effect size of interest and fixed multiple comparisons threshold in advance. This enables straightforward power analysis (e.g., calculating power to detect an effect of a given size, sample size required for a specific power level, or minimal detectable effect size given a fixed sample size). In addition, a minimal effect size of interest could later serve in equivalence and Bayes factor tests providing evidence also for the absence, not just the presence, of an effect (96). Nevertheless, defining a minimal effect size of interest is often challenging, particularly in nonclinical mechanistic neuroimaging studies.

Other methodological decisions are also key. For example, testing external validation with independent samples (82) and optimization of data acquisition and measurement (97) are important for creating replicable biomarkers (98). Another important, but often overlooked, source of statistical power is the number of trials collected from each participant (99). Importantly, the optimal number of participants and trials depends on the ratio between the within-participant and between-participant variance (100,101). Additional approaches

include real-time within-participant optimization of experimental designs or artifact minimization (102–104).

Analysis Flexibility and Preregistration

If researchers test multiple analyses and report only the one that yielded significant results [i.e., p-hacking (105,106)], or change their hypotheses after the results are known [HARKing (107)], they introduce a selection bias that increases the rate of false positives and decreases the likelihood of independent replication. Despite the increased awareness of this issue, such questionable practices are still prevalent (108–110).

Preregistration is a partial solution to these issues. The experimental design, sample size, hypotheses, and analysis plan are registered prior to data collection (or at least prior to observing the outcomes), distinguishing between confirmatory and exploratory analyses (111,112). This limits p-hacking and HARKing, depending on the completeness of the preregistered plan and the closeness with which it is followed. Guides and templates for preregistration have been developed for several fields and study types [e.g., (113–115)]. A related promising publication format that has already been implemented in hundreds of journals is Registered Reports, in which studies are peer reviewed prior to data collection, and once accepted (in principle) they are executed and later published in the journal, irrespective of the outcome (116). It has already been shown to mitigate publication bias with much higher rates of null findings than traditional publications (17), and guidance for writing Registered Reports is available (117).

Preregistration comes with many challenges that are actively being discussed. Many analytic choices depend (and should depend) on the characteristics of the data, which are difficult or impossible to know in advance, particularly with novel research questions. Thus, the plan that is preregistered is likely to be suboptimal. In addition, best practices are constantly developing. It is common for methods to develop after a study is registered. Consequently, deviations from the preregistered analysis plan are very common.

One way to address the uncertainties inherent in preregistration is to include a range of options that will be explored and optimized based on positive control effects, which are effects that are expected but are independent from the outcome of interest (e.g., motor and visual responses when the outcome of interest is an emotional response). Then, the optimized analysis will be used once on the effect of interest. For example, we often use the neurological pain signature (118)—a multivariate pattern whose effect size, sensitivity, and specificity have been extensively examined (73)—as a positive control. Importantly, the effect of interest (e.g., meditation effects) must be independent of the effect that is optimized (e.g., neurological pain signature responses to pain vs. rest). Such domain-specific positive controls are common in many biological assays, and the principle can be used broadly across research domains. Another approach is to use independent data (rather than an independent effect) to optimize analysis and test models and then apply them once on the test data, as is common in machine learning. Preregistered plans can specify the range of options tested and how the final choices will be made. If followed assiduously, these practices allow researchers to accommodate data characteristics and optimized methods

Reproducible Neuroimaging Analysis

while retaining the ability to test a hypothesis of interest in an unbiased fashion.

ROBUSTNESS TO ANALYTICAL VARIABILITY: SAME DATA, DIFFERENT METHODS

Data analysis requires many analytical decisions. As discussed above, these decisions can lead to false positive results (105,119). Critically, they also increase the uncertainty of any given single result. For example, Carp (120) compared brain maps from nearly 7000 analysis pipelines based on the same data and found substantial variability. More recently, 70 independent analysis teams tested 9 prespecified hypotheses using the same task–functional magnetic resonance imaging (fMRI) dataset (121). Strikingly, the 70 teams chose 70 different analysis pipelines, and this variation affected the statistical maps and conclusions drawn about the hypotheses tested. Similar effects of analytical variability have since been shown in resting-state fMRI (122), diffusion MRI (123), structural MRI (124,125), positron emission tomography (126), and electroencephalography (127) and also in psychology (128–130) and the social sciences (131,132), more broadly. Even in the absence of selection bias (e.g., p-hacking), such variation raises critical questions about which choices lead to correct conclusions and how robust findings are to arbitrary choices (133).

One solution that has been suggested is multiverse analysis (134,135), in which a range of reasonable analysis pipelines are tested and reported. This concept has been used across fields, sometimes under different names [e.g., specification curve analysis (136,137)]. Such analyses can be performed by collaborative teams in multianalyst studies (121,128,132) [for guidelines, see (138)] or with a single analyst (Figure 2). Knowing whether a finding is robust to variability helps calibrate confidence in its accuracy and generalizability. In addition, the inference across models is likely to be more accurate than any single model.

Multiverse analysis comes with challenges that must be addressed for it to become widely adopted. First, it requires the specification of a range of valid choices. Some choices may better fit particular datasets, and this will be largely unknown in advance when testing novel hypotheses. Furthermore, results will depend on the multiverse of pipelines chosen, and multiverse analysis can be p-hacked just like

standard analyses (139). Therefore, tools and practices must be developed to guide researchers through this process. For example, the multianalyst fMRI study discussed above (121) identified several key factors contributing to results variability, including data smoothness, the analysis software used, and parametric versus nonparametric statistical tests. Future studies could be designed to identify such field-specific key choices with more certainty. Alternatively, machine learning approaches can inform the choice of a subset of analysis pipelines (140).

A second challenge is that multiverse analysis requires extensive computational resources. Tools to increase computational efficiency could enable researchers to run them with widely available resources. Sharing of preprocessed data and derivatives further decreases computational barriers. Infrastructures and tools for multiverse analysis in neuroimaging are starting to be developed (122,141), but there is still a long way to go before they are ready for broad use.

Third, multiverse analysis is complex, making it challenging to integrate, visualize, and summarize findings. Diverse approaches to visualization and reporting have been developed (134,136,142,143). One approach to integrate results of multiverse analyses in fMRI is a consensus analysis, which is a type of meta-analysis adapted to account for the dependency across pipelines that are based on the same data (121,144).

All in all, multiverse analysis is a developing framework and is currently more feasible with simpler datasets and analyses (136,145,146) than with the multidimensional data and complex pipelines in neuroimaging. Extensive efforts are being made and new tools are expected in the near future. Meanwhile, researchers can perform more limited sensitivity analyses, in which key parameters are varied (e.g., inclusion of covariates) and their effects assessed. Importantly, any analysis pipeline that was run and observed should be reported. And, finally, multiverse analyses can themselves be preregistered, facilitating the examination of robustness to methodological choices.

CONCLUSIONS

Psychiatric disorders such as depression, anxiety, substance use, and even chronic pain are increasingly understood to result from biological dispositions interacting with environmental stressors that are altering the brain. Yet, unlike other types of

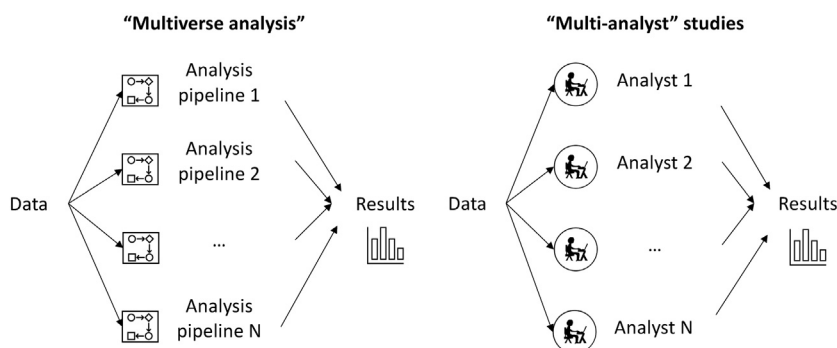


Figure 2. Illustration of multiverse analysis and multianalyst studies.

medicine, both diagnosis and treatments in psychiatry are mostly based on symptoms or self-reports, which are subjective and context dependent (147–149). Effective treatments for diabetes, for example, would probably not be available if they were developed based on patients' subjective reports rather than their blood sugar levels. Thus, psychiatric biomarkers are urgently needed to provide mechanistic targets.

For such biomarkers to be translated to clinical practice, they must be reproducible. Numerous new tools and practices aiming to increase reproducibility have been developed, but their adoption is still relatively slow. Furthermore, as issues are being solved, new challenges are being revealed. Thus, the quest toward reproducibility is an ongoing endeavor. We hope this review, together with other resources [e.g., a review including a comprehensive table of tools and resources (8)], helps by informing researchers in the field about the main challenges and introducing some new solutions. Along with other developments, increasing reproducibility in psychiatric research will advance the field toward developing these long-needed biomarkers.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the National Institute of Mental Health (Grant No. R01 MH076136 [to TDW]). RB-N is an awardee of the Weizmann Institute of Science - Israel National Postdoctoral Award Program for Advancing Women in Science.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire (RB-N, TDW).

Address correspondence to Rotem Botvink-Nezer, Ph.D., at rotemb9@gmail.com.

Received Jul 15, 2022; revised Nov 27, 2022; accepted Dec 11, 2022.

REFERENCES

- Ioannidis JPA (2005): Why most published research findings are false [published correction appears in *PLoS Med* 2022;19:e1004085]. *PLoS Med* 2:e124.
- Open Science Collaboration (2015): PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349:aac4716.
- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013): Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376.
- Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP (2014): Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends Cogn Sci* 18:235–241.
- Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. (2017): A manifesto for reproducible science. *Nat Hum Behav* 1:0021.
- Houtkoop BL, Chambers C, Macleod M, Bishop DVM, Nichols TE, Wagenmakers E-J (2018): Data sharing in psychology: A survey on barriers and preconditions. *Adv Methods Pract Psychol Sci* 1:70–85.
- Nelson LD, Simmons J, Simonsohn U (2018): Psychology's renaissance. *Annu Rev Psychol* 69:511–534.
- Niso G, Botvink-Nezer R, Appelhoff S, De La Vega A, Esteban O, Etzel JA, et al. (2022): Open and reproducible neuroimaging: From study inception to publication. *NeuroImage* 263:119623.
- Paret C, Unverhau N, Feingold F, Poldrack RA, Stirner M, Schmahl C, Sicorello M (2022): Survey on open science practices in functional neuroimaging. *Neuroimage* 257:119306.
- Borghi JA, Van Gulick AE (2018): Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLoS One* 13:e0200562.
- Hardwicke TE, Thibault RT, Kosie JE, Wallach JD, Kidwell MC, Ioannidis JPA (2022): Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspect Psychol Sci* 17:239–251.
- Piller C (2021): Disgraced COVID-19 studies are still routinely cited. *Science* 371:331–332.
- Bucci EM (2019): On zombie papers. *Cell Death Dis* 10:189.
- Nissen SB, Magidson T, Gross K, Bergstrom CT (2016): Publication bias and the canonization of false facts. *eLife* 5:e21451.
- Jwa AS, Poldrack RA (2022): The spectrum of data sharing policies in neuroimaging data repositories. *Hum Brain Mapp* 43:2707–2721.
- Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, et al. (2018): Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* 9:2818.
- Allen C, Mehler DMA (2019): Open science challenges, benefits and tips in early career and beyond [published correction appears in *PLoS Biol* 2019;17:e3000587]. *PLoS Biol* 17:e3000246.
- de Jonge H, Cruz M, Holst S (2021): Funders need to credit open science. *Nature* 599:372.
- Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, et al. (2022): Replicability, robustness, and reproducibility in psychological science. *Annu Rev Psychol* 73:719–748.
- Hardwicke TE, Bohn M, MacDonald K, Hembacher E, Nuijten MB, Peloquin BN, et al. (2021): Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science: An observational study*. *R Soc Open Sci* 8:201494.
- Errington TM, Denis A, Perfito N, Iorns E, Nosek BA (2021): Challenges for assessing replicability in preclinical cancer biology. *eLife* 10:e67995.
- Stodden V, Seiler J, Ma Z (2018): An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc Natl Acad Sci USA* 115:2584–2589.
- Obels P, Lakens D, Coles NA, Gottfried J, Green SA (2020): Analysis of open data and computational reproducibility in Registered Reports in psychology. *Adv Methods Pract Psychol Sci* 3:229–237.
- Hardwicke TE, Mathur MB, MacDonald K, Nilsson G, Banks GC, Kidwell MC, et al. (2018): Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R Soc Open Sci* 5:180448.
- Gronenschild EH, Habets P, Jacobs HI, Mengelers R, Rozendaal N, van Os J, Marcelis M (2012): The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS One* 7:e38234.
- Glatard T, Lewis LB, Ferreira da Silva R, Adalat R, Beck N, Lepage C, et al. (2015): Reproducibility of neuroimaging analyses across operating systems. *Front Neuroinform* 9:12.
- Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, et al. (2017): Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci* 20:299–303.
- Pernet C, Garrido MI, Gramfort A, Maurits N, Michel CM, Pang E, et al. (2020): Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research. *Nat Neurosci* 23:1473–1483.
- Karakuzu A, DuPre E, Tetrel L, Bermudez P, Boudreau M, Chin M, et al. (2022): NeuroLibre: A preprint server for full-fledged reproducible neuroscience. Available at: <https://osf.io/h89js/>. Accessed March 3, 2023.
- Nuijten MB, Polanin JR (2020): “statcheck”: Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res Synth Methods* 11:574–579.
- Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013): Ten simple rules for reproducible computational research. *PLoS Comput Biol* 9:e1003285.

Reproducible Neuroimaging Analysis

32. Balaban G, Grytten I, Rand KD, Scheffer L, Sandve GK (2021): Ten simple rules for quick and dirty scientific programming. *PLoS Comput Biol* 17:e1008549.
33. Eglen SJ, Marwick B, Halchenko YO, Hanke M, Sufi S, Gleeson P, *et al.* (2017): Toward standard practices for sharing computer code and programs in neuroscience. *Nat Neurosci* 20:770–773.
34. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017): Good enough practices in scientific computing. *PLoS Comput Biol* 13:e1005510.
35. Blischak JD, Davenport ER, Wilson G (2016): A quick introduction to version control with git and GitHub. *PLoS Comput Biol* 12:e1004668.
36. Lee BD (2018): Ten simple rules for documenting scientific software. *PLoS Comput Biol* 14:e1006561.
37. Riquelme JL, Gjorgjieva J (2021): Towards readable code in neuroscience. *Nat Rev Neurosci* 22:257–258.
38. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, *et al.* (2014): Best practices for scientific computing. *PLoS Biol* 12:e1001745.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2011): Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.
40. Flandin G, Friston K (2008): Statistical parametric mapping (SPM). *Scholarpedia* 3:6232.
41. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012). *FSL. Neuroimage* 62:782–790.
42. Cox RW (1996): AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29:162–173.
43. Breeze JL, Poline JB, Kennedy DN (2012): Data sharing and publishing in the field of neuroimaging. *GigaScience* 1:9.
44. Poldrack RA, Gorgolewski KJ (2014): Making big data open: Data sharing in neuroimaging. *Nat Neurosci* 17:1510–1517.
45. Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, *et al.* (2021): The OpenNeuro resource for sharing of neuroscience data. *eLife* 10:e71774.
46. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, *et al.* (2016): The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018.
47. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, *et al.* (2016): The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* 3:160044.
48. Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, *et al.* (2017): BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput Biol* 13:e1005209.
49. Keator DB, Helmer K, Steffener J, Turner JA, Van Erp TGM, Gadde S, *et al.* (2013): Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82:647–661.
50. Maumet C, Auer T, Bowring A, Chen G, Das S, Flandin G, *et al.* (2016): Sharing brain mapping statistical results with the neuroimaging data model. *Sci Data* 3:160102.
51. Halchenko Y, Meyer K, Poldrack B, Solanky D, Wagner A, Gors J, *et al.* (2021): DataLad: Distributed system for joint management of code, data, and their relationship. *J Open Source Softw* 6:3262.
52. Borghi JA, Van Gulick AE (2021): Promoting open science through research data management. *arXiv* <https://doi.org/10.48550/arXiv.2110.00888> version 2, <https://arxiv.org/abs/2110.00888v2>.
53. Kiar G, Chatelain Y, de Oliveira Castro P, Petit E, Rokem A, Varoquaux G, *et al.* (2021): Numerical uncertainty in analytical pipelines lead to impactful variability in brain networks. *PLoS One* 16:e0250755.
54. Kiar G, de Oliveira Castro P, Rioux P, Petit E, Brown ST, Evans AC, Glatard T (2020): Comparing perturbation models for evaluating stability of neuroimaging pipelines. *Int J High Perform Comput Appl* 34:491–501.
55. Kurtzer GM, Sochat V, Bauer MW (2017): Singularity: Scientific containers for mobility of compute. *PLoS One* 12:e0177459.
56. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, *et al.* (2018): Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav* 2:637–644.
57. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, *et al.* (2016): Evaluating replicability of laboratory experiments in economics. *Science* 351:1433–1436.
58. Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, *et al.* (2021): Estimating the reproducibility of experimental philosophy. *Rev Philos Psychol* 12:9–44.
59. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA (2021): Investigating the replicability of preclinical cancer biology. *eLife* 10:e71601.
60. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, *et al.* (2014): Investigating variation in replicability: A “many labs” replication project. *Soc Psychol* 45:142–152.
61. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, *et al.* (2018): Many labs 2: Investigating variation in replicability across samples and settings. *Adv Methods Pract Psychol Sci* 1:443–490.
62. Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, Chartier CR, *et al.* (2020): Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Adv Methods Pract Psychol Sci* 3:309–331.
63. Frank MC, Bergelson E, Bergmann C, Cristia A, Floccia C, Gervain J, *et al.* (2017): A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy* 22:421–435.
64. Pavlov YG, Adamian N, Appelhoff S, Arvaneh M, Benwell CSY, Beste C, *et al.* (2021): #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex* 144:213–229.
65. Coles NA, March DS, Marmolejo-Ramos F, Larsen JT, Arinze NC, Ndukaihe ILG, *et al.* (2019): A multi-lab test of the facial feedback hypothesis by the many smiles collaboration. *Nat Hum Behav* 6:1731–1742.
66. Moshontz H, Campbell L, Ebersole CR, Ijzerman H, Urry HL, Forscher PS, *et al.* (2018): The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Adv Methods Pract Psychol Sci* 1:501–515.
67. Algermissen J, Mehler DMA (2018): May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them? *J Neurophysiol* 119:2114–2117.
68. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, *et al.* (2017): Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* 18:115–126.
69. Szucs D, Ioannidis JP (2020): Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* 221:117164.
70. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, *et al.* (2022): Reproducible brain-wide association studies require thousands of individuals. *Nature* 603:654–660.
71. Nikolaidis A, Chen AA, He X, Shinohara R, Vogelstein J, Milham M, Shou H (2022): Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv* <https://doi.org/10.1101/2022.07.22.501193>.
72. Spisak T, Bingle U, Wager T (2022): Replicable multivariate BWAS with moderate sample sizes. *bioRxiv* <https://doi.org/10.1101/2022.06.22.497072>.
73. Han X, Ashar YK, Kragel P, Petre B, Schelkun V, Atlas LY, *et al.* (2022): Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *Neuroimage* 247:118844.
74. Reddan MC, Lindquist MA, Wager TD (2017): Effect size estimation in neuroimaging. *JAMA Psychiatry* 74:207–208.
75. Zunhammer M, Bingle U, Wager TD, Placebo Imaging Consortium (2018): Placebo effects on the neurologic pain signature: A meta-analysis of individual participant functional magnetic resonance imaging data. *JAMA Neurol* 75:1321–1330.

76. Lindquist KA, Satpute AB, Wager TD, Weber J, Barrett LF (2016): The brain basis of positive and negative affect: Evidence from a meta-analysis of the human neuroimaging literature. *Cereb Cortex* 26:1910–1922.
77. Flint C, Cearns M, Opel N, Redlich R, Mehler DMA, Emden D, *et al.* (2021): Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology* 46:1510–1517.
78. Belov V, Erwin-Grabner T, Gonul AS, Amod AR, Ojha A, Aleman A, *et al.* (2022): Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures. *arXiv* <https://doi.org/10.48550/arxiv.2206.08122> version 3, <http://arxiv.org/abs/2206.08122v3>.
79. Nielsen AN, Barch DM, Petersen SE, Schlaggar BL, Greene DJ (2020): Machine learning with neuroimaging: Evaluating its applications in psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging* 5:791–798.
80. Poldrack RA, Huckins G, Varoquaux G (2020): Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry* 77:534–540.
81. Davatzikos C (2019): Machine learning in neuroimaging: Progress and challenges. *Neuroimage* 197:652–656.
82. Woo CW, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: Brain models in translational neuroimaging. *Nat Neurosci* 20:365–377.
83. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* 180:68–77.
84. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, *et al.* (2012): The Human connectome Project: A data acquisition perspective. *Neuroimage* 62:2222–2231.
85. Miller KL, Alfaro-Almagro F, Bangarter NK, Thomas DL, Yacoub E, Xu J, *et al.* (2016): Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 19:1523–1536.
86. Feldstein Ewing S, Luciano M (2018): The Adolescent Brain Cognitive Development (ABCD) consortium: Rationale, aims, and assessment strategy. *Dev Cogn Neurosci* 32:1–164.
87. Schmaal L, Pozzi E, Ho CT, van Velzen LS, Veer IM, Opel N, *et al.* (2020): ENIGMA MDD: Seven years of global neuroimaging studies of major depression through worldwide data sharing. *Transl Psychiatry* 10:172.
88. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, *et al.* (2018): Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum Brain Mapp* 39:4213–4227.
89. Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, *et al.* (2017): Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161:149–170.
90. Bayer JMM, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, *et al.* (2022): Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Front Neurol* 13: 923988.
91. Button KS, Munafò MR (2017): Powering reproducible research. In: Lilienfeld SO, Waldman ID, editors. *Psychological Science Under Scrutiny*. Hoboken, NJ: John Wiley & Sons, Inc., 22–33
92. Noble S, Scheinost D, Constable RT (2020): Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. *Neuroimage* 209:116468.
93. Noble S, Mejia AF, Zalesky A, Scheinost D (2022): Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proc Natl Acad Sci USA* 119:e2203020119.
94. Eklund A, Nichols TE, Knutsson H (2016): Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:7900–7905.
95. Lakens D (2022): Sample size justification. *Collabra Psychol* 8:33267.
96. Lakens D, McLatchie N, Isager PM, Scheel AM, Dienes Z (2020): Improving inferences about null effects with Bayes factors and equivalence tests. *J Gerontol B Psychol Sci Soc Sci* 75:45–57.
97. Finn ES (2021): Is it time to put rest to rest? *Trends Cogn Sci* 25:1021–1032.
98. Rosenberg MD, Finn ES (2022): How to establish robust brain-behavior relationships without thousands of individuals. *Nat Neurosci* 25:835–837.
99. Fröhner JH, Teckenrump V, Smolka MN, Kroemer NB (2019): Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *Neuroimage* 195:174–189.
100. Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Taylor PA, Haller SP (2022): Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. *Neuroimage* 247: 118786.
101. Baker DH, Vilidate G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ (2021): Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychol Methods* 26:295–314.
102. Lorenz R, Monti RP, Violante IR, Anagnostopoulos C, Faisal AA, Montana G, Leech R (2016): The Automatic Neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI. *Neuroimage* 129:320–334.
103. Lorenz R, Johal M, Dick F, Hampshire A, Leech R, Geranmayeh F (2021): A Bayesian optimization approach for rapidly mapping residual network function in stroke. *Brain* 144:2120–2134.
104. Dosenbach NUF, Koller JM, Earl EA, Miranda-Dominguez O, Klein RL, Van AN, *et al.* (2017): Real-time motion analytics during brain MRI improve data quality and reduce costs. *Neuroimage* 161:80–93.
105. Simmons JP, Nelson LD, Simonsohn U (2011): False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366.
106. Simonsohn U, Nelson LD, Simmons JP (2014): P-curve: A key to the file-drawer. *J Exp Psychol Gen* 143:534–547.
107. Kerr NL (1998): HARKing: Hypothesizing after the results are known. *Pers Soc Psychol Rev* 2:196–217.
108. John LK, Loewenstein G, Prelec D (2012): Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci* 23:524–532.
109. Gopalakrishna G, Ter Riet G, Vink G, Stoop I, Wicherts JM, Bouter LM (2022): Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the Netherlands. *PLoS One* 17:e0263023.
110. Xie Y, Wang K, Kong Y (2021): Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Sci Eng Ethics* 27:41.
111. Simmons J, Nelson L, Simonsohn U (2021): Pre-registration: Why and how. *J Consum Psychol* 31:151–162.
112. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018): The pre-registration revolution. *Proc Natl Acad Sci USA* 115:2600–2606.
113. Paul M, Govaert GH, Schettino A (2021): Making ERP research more transparent: Guidelines for preregistration. *Int J Psychophysiol* 164:52–63.
114. Beyer F, Flannery J, Gau R, Janssen L, Schaare L, Hartmann H, *et al.* (2021): A fMRI preregistration template. *Psycharchives*. <https://doi.org/10.23668/PSYCHARCHIVES.5121>.
115. Crüwell S, Evans NJ (2021): Preregistration in diverse contexts: A preregistration template for the application of cognitive models. *R Soc Open Sci* 8:210155.
116. Chambers CD, Tzavella L (2022): The past, present and future of Registered Reports. *Nat Hum Behav* 6:29–42.
117. Henderson EL, Chambers CD (2022): Ten simple rules for writing a Registered Report. *PLoS Comput Biol* 18:e1010571.
118. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E (2013): An fMRI-based neurologic signature of physical pain. *N Engl J Med* 368:1388–1397.
119. Gelman A, Loken E (2013): 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* 348:1–17.

Reproducible Neuroimaging Analysis

120. Carp J (2012): On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments. *Front Neurosci* 6:149.
121. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, *et al.* (2020): Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582:84–88.
122. Li X, Ai L, Giavasis S, Jin H, Feczko E, Xu T, *et al.* (2021): Moving beyond processing and analysis-related variation in neuroscience. *bioRxiv* <https://doi.org/10.1101/2021.12.01.470790>.
123. Schilling KG, Rheault F, Petit L, Hansen CB, Nath V, Yeh FC, *et al.* (2021): Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset? *Neuroimage* 243:118502.
124. Zhou X, Wu R, Zeng Y, Qi Z, Ferraro S, Yao S, *et al.* (2021): Location, location, location– choice of voxel-Based morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. *bioRxiv* <https://doi.org/10.1101/2021.03.09.434531>.
125. Bhagwat N, Barry A, Dickie EW, Brown ST, Devenyi GA, Hatano K, *et al.* (2021): Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience* 10:giaa155.
126. Nørgaard M, Ganz M, Svarer C, Frokjaer VG, Greve DN, Strother SC, Knudsen GM (2020): Different preprocessing strategies lead to different conclusions: A [¹¹C]DASB-PET reproducibility study. *J Cereb Blood Flow Metab* 40:1902–1911.
127. Clayton PE, Baldwin SA, Rocha HA, Larson MJ (2021): The data-processing multiverse of event-related potentials (ERPs): A road-map for the optimization and standardization of ERP processing and reduction pipelines. *Neuroimage* 245:118712.
128. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, *et al.* (2018): Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Adv Methods Pract Psychol Sci* 1:337–356.
129. Schweinsberg M, Madan N, Vianello M, Sommer SA, Jordan J, Tierney W, *et al.* (2016): The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J Exp Soc Psychol* 66:55–67.
130. Landy JF, Jia ML, Ding IL, Viganola D, Tierney W, Dreber A, *et al.* (2020): Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull* 146:451–479.
131. Breznau N, Rinke EM, Wuttke A, Nguyen HHV, Adem M, Adriaans J, *et al.* (2022): Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc Natl Acad Sci USA* 119:e2203150119.
132. Schweinsberg M, Feldman M, Staub N, van den Akker OR, van Aert RCM, van Assen MALM, *et al.* (2021): Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organ Behav Hum Decis Process* 165:228–249.
133. Wagenmakers EJ, Sarafoglou A, Aczel B (2022): One statistical analysis must not rule them all. *Nature* 605:423–425.
134. Hall BD, Liu Y, Jansen Y, Dragicevic P, Chevalier F, Kay M (2022): A survey of tasks and visualizations in multiverse analysis reports. *Comput Graph Forum* 41:402–426.
135. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W (2016): Increasing transparency through a multiverse analysis. *Perspect Psychol Sci* 11:702–712.
136. Simonsohn U, Simmons JP, Nelson LD (2020): Specification curve analysis. *Nat Hum Behav* 4:1208–1214.
137. Simonsohn U, Simmons JP, Nelson LD (2015): Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Journal*. <https://doi.org/10.2139/ssrn.2694998>.
138. Aczel B, Szasz B, Nilsson G, van den Akker OR, Albers CJ, van Assen MA, *et al.* (2021): Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* 10:e72185.
139. Del Giudice M, Gangestad SW (2021): A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Adv Methods Pract Psychol Sci* 4:1–15.
140. Dafflon J, Da Costa FP, Váša F, Monti RP, Bzdok D, Hellyer PJ, *et al.* (2022): A guided multiverse study of neuroimaging analyses. *Nat Commun* 13:3758.
141. Markiewicz CJ, De La Vega A, Wagner A, Halchenko YO, Finc K, Ciric R, *et al.* (2021): Poldracklab/fitlins, v.0.9.2. Available at: <https://zenodo.org/record/5120201#.Y51aaFFBzIU>. Accessed March 3, 2023.
142. Dragicevic P, Jansen Y, Sarma A, Kay M, Chevalier F (2019): Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery, 1–15.
143. Liu Y, Kale A, Althoff T, Heer J (2021): Boba: Authoring and visualizing multiverse analyses. *IEEE Trans Vis Comput Graph* 27:1753–1763.
144. Bowering A, Nichols TE, Maumet C (2022): Isolating the sources of pipeline-variability in group-level task-fMRI results. *Hum Brain Mapp* 43:1112–1128.
145. Lonsdorf TB, Gerlicher A, Klingelhöfer-Jens M, Krypotos AM (2022): Multiverse analyses in fear conditioning research. *Behav Res Ther* 153:104072.
146. Donnelly S, Brooks PJ, Homer BD (2019): Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychon Bull Rev* 26:1122–1147.
147. Kapur S, Phillips AG, Insel TR (2012): Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry* 17:1174–1179.
148. Insel TR, Cuthbert BN (2015): Medicine. Brain disorders? Precisely. *Science* 348:499–500.
149. Davis KD, Aghaeeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, *et al.* (2020): Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: Challenges and opportunities. *Nat Rev Neurol* 16:381–400.