Best Practices in Data Analysis and Sharing in Neuroimaging using MRI

Thomas E. Nichols[1,*], Samir Das[2], Simon B. Eickhoff[3], Alan C. Evans[2], Tristan Glatard[2], Michael Hanke[4], Nikolaus Kriegeskorte[5], Michael P. Milham[6], Russell A. Poldrack[7], Jean-Baptiste Poline[8], Erika Proal[9], Bertrand Thirion[10], David C. Van Essen[11], Tonya White[12], B.T. Thomas Yeo[13]

[1]University of Warwick; [2]McGill University; [3]Heinrich-Heine University Düsseldorf; [4]Otto-von-Guericke-University Magdeburg; [5]MRC Cognition and Brain Sciences Unit; [6]Child Mind Institute; [7]Stanford University; [8]University of California, Berkeley;
[9]Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz & Neuroingenia; [10]Inria, Paris-Saclay University; [11]Washington University in St. Louis; [12]Erasmus University Medical Center; [13]National University of Singapore.
*Correspondence: t.e.nichols@warwick.ac.uk.

Abstract

Neuroimaging enables rich noninvasive measurements of human brain activity, but translating such data into neuroscientific insights and clinical applications requires complex analyses and collaboration among a diverse array of researchers. The open science movement is reshaping scientific culture and addressing the challenges of transparency and reproducibility of research. To advance open science in neuroimaging the Organization for Human Brain Mapping created the Committee on Best Practice in Data Analysis and Sharing (COBIDAS), charged with creating a report that collects best practice recommendations from experts and the entire brain imaging community. The purpose of this work is to elaborate the principles of open and reproducible research for neuroimaging using Magnetic Resonance Imaging (MRI), and then distill these principles to specific research practices. Many elements of a study are so varied that practice cannot be prescribed, but for these areas we detail the information that must be reported to fully understand and potentially replicate a study. For other elements of a study, like statistical modelling where specific poor practices can be identified, and the emerging areas of data sharing and reproducibility, we detail both good practice and reporting standards. For each of seven areas of a study we provide tabular listing of over 100 items to help plan, execute, report and share research in the most transparent fashion. Whether for individual scientists, or for editors and reviewers, we hope these guidelines serve as a benchmark, to raise the standards of practice and reporting in neuroimaging using MRI.

Contents

1. Introduction 3
    1.1. Approach 4
    1.2. Scope 5
2. Experimental Design Reporting 5
    2.1. Scope 5
    2.2. General Principles 6
    2.3. Lexicon of fMRI Design 6
    Box 2.1. fMRI Terminology 6
    2.4. Design Optimization 7
    2.5. Power Analysis 7
    2.6. Subjects 8
    2.7. Behavioral Performance 8
3. Acquisition Reporting 8
    3.1. Scope 8
    3.2. General Principles 8
    3.3. Device Information 9
    3.4. Acquisition-Specific Information 9
    3.5. Format for sharing 9
4. Preprocessing Reporting 9
    4.1. Scope 9
    4.2. General Principles 10
    4.3. Software Issues 10
5. Statistical Modeling & Inference 11
    5.1. Scope 11
    5.2. General Principles 11
    5.3. Assumptions 11
    5.4. Software 12
    5.5. Mass Univariate Modelling 12
    5.6. Connectivity Analyses 13
    5.7. Other Resting-State Analyses 14
    5.8. Multivariate Modelling & Predictive Analysis 15
6. Results Reporting 16
    6.1. Scope 16
    6.2. General Principles 16
    6.3. Mass Univariate Modelling 16
    6.4. Functional Connectivity 18
    6.5. Multivariate Modelling & Predictive Analysis 18
7. Data Sharing 19
    7.1. Scope 19
    7.2. General Principles 19
    7.3. Planning for Sharing 20
    7.4. Databases 22
    7.5. Documentation 23
    7.6. Ethics 24
8. Reproducibility 24
    8.1. Scope 24
    8.2. Documentation 25
    8.3. Archiving 25
    8.4. Citation 26
9. Conclusions 26
References 27
Appendix A. COBIDAS Membership & Acknowledgements 35
Appendix B. Defining Reproducibility ... 37
    Table B1. Levels of Reproducibility 39
Appendix C. Short descriptions of fMRI models 40
    C1. Task fMRI 40
    C2. Single-Modality ICA 41
    C3. Multi-Modality ICA 42
Appendix D. Itemized lists of best practices and reporting items 43
    Table D.1. Experimental Design Reporting 44
    Table D.2. Acquisition Reporting 48
    Table D.3. Preprocessing Reporting 53
    Table D.4. Statistical Modeling & Inference 59
    Table D.5. Results Reporting 67
    Table D.6. Data Sharing 69
    Table D.7. Reproducibility 71

1. Introduction

In many areas of science and in the public sphere there are growing concerns about the reproducibility of published research. From early claims by John Ioannidis in 2005 that "most published research findings are false" [Ioannidis2005] to the recent work by the Open Science Collaboration, which attempted to replicate 100 psychology studies and succeeded in only 39 cases [OpenScienceCollaboration2015], there is mounting evidence that scientific results are less reliable than widely assumed. As a result, calls to improve the transparency and reproducibility of scientific research have risen in frequency and fervor.

In response to these concerns, the Organization for Human Brain Mapping (OHBM) released "OHBM Council Statement on Neuroimaging Research and Data Integrity"[1] in June 2014, at the same time creating the Committee on Best Practices in Data Analysis and Sharing (COBIDAS). The committee was charged with (i) identifying best practices of data analysis and data sharing in the brain mapping community, (ii) preparing a white paper organizing and describing these practices, and (iii) seeking input from the OHBM community before (iv) publishing these recommendations.

COBIDAS focuses on data analysis and statistical inference procedures because they play an essential role in the reliability of scientific results. Brain imaging data is complicated because of the many processing steps and a massive number of measured variables. There are many different specialised analyses investigators can choose from, and analyses often involve cycles of exploration and selective analysis that can bias effect estimates and invalidate inference [Kriegeskorte2009, Carp2012].

Beyond data analysis, COBIDAS also addresses best practices in data sharing. The sharing of data can enable reuse, saving costs of data acquisition and making the best use of scarce research funding[2] [Macleod2014]. In addition, data sharing enables other researchers to reproduce results using the same or different analyses, which may reveal errors or bring new insights overlooked initially (see, e.g., [LeNoury2015]). There is also evidence that data sharing is associated with better statistical reporting practices and stronger empirical evidence [Wicherts2011]. In short, data sharing fosters a scientific culture of transparency.

While many recent publications prescribe greater transparency and sharing of data (see, e.g., a pair of editorials in *Science* & *Nature* [Journals2014,McNutt2014]), such works are general to all of science or do not focus on human neuroimaging specifically (though see [Poline2012,Poldrack2014]). Thus the purpose of this paper is to elaborate some principles of open and reproducible research for the areas of practice relevant to the OHBM community. To

[1] http://www.humanbrainmapping.org/OHBMDataIntegrity2014.
[2] The Lancet's series on "Research: increasing value, reducing waste", http://www.thelancet.com/series/research.

make these principles practical and usable, we created explicit lists of items to be shared (Appendix D).

Working closely with OHBM Council, a "Proposed" version of this document was prepared by COBIDAS and first posted for comment in October 2015. Comments were collected via email and the COBIDAS blog[3] and a revised document[4] presented to the membership for an up/down vote in May 2016; 96% positive ballots were received. We note that while best practice white papers like this are not uncommon (see, e.g., [Alsop2014,Kanal2013,Gilmore2013]), they are generally authored by and represent the consensus of a small committee or at most a special-interest section of a larger professional body. Hence we are excited to present this work with the explicit participation and support of the OHBM membership.

1.1. Approach

There are different responses to the perceived crisis of reproducibility, with some simply letting the problem `self-correct' as reviewers and readers become more aware of the problem, to more transformative measures like using blinded analyses or preregistration. In a blinded analysis all preprocessing, modelling and results generation is conducted with experimental labels hidden or scrambled, only being revealed after the analysis is fixed [MacCoun2015], while in preregistration all research hypotheses and analysis plans are published before data are collected, [Nosek2014].

The pragmatic approach behind this report is to increase the transparency of how research has been executed. Such transparency can be accomplished by comprehensive sharing of data, research methods and finalized results. This both enables other investigators to reproduce findings with the same data, better interrogate the methodology used and, ultimately, makes best use of research funding by allowing reuse of data.

The reader may be daunted by the sheer scale and detail of recommendations and checklists in this work (Appendix D). However we expect that any experienced neuroimaging researcher who has read a paper in depth, and been frustrated by the inevitable ambiguity or lack of detail, will appreciate the value of each entry. We do not intend for these lists to become absolute, inflexible requirements for publication. However they are the product of extensive deliberation by this panel of experts, and represent what we considered most effective and correct; hence, deviations from these practices may warrant explanation. Finally, we hope these lists can serve as tools for reviewers, e.g. as a reference for the importance of these items, and for editors and publishers, who can consider how this necessary detail can be conveyed without stringent word limits.

[3] See http://www.humanbrainmapping.org/cobidas.
[4] The present document differs from that voted on only in formating changes, typographical corrections, and the update of this particular sentence.

1.2. Scope

While the OHBM community is diverse, including users of a variety of brain imaging modalities, for this effort we focus exclusively on MRI. This encompasses a broad range of work, including task-based and resting-state functional MRI (fMRI), analyzed voxel-wise and on the surface, structural and diffusion MRI, but excludes other widely used methods like PET, EEG & MEG. We found that practice in neuroimaging with MR can be broken into seven areas that roughly span the entire enterprise of a study: (1) experimental design reporting, (2) image acquisition reporting, (3) preprocessing reporting, (4) statistical modeling, (5) results reporting, (6) data sharing, and (7) reproducibility.

Reproducibility has different and conflicting definitions (see Appendix B), but in this work we make the distinction between reproducing results *with the same data* [Peng2011] versus replicating a result *with different data* and possibly methods. Hence while this entire work is about maximizing replicability, the last section focuses specifically on reproducibility at the analysis-level.

This paper is structured around these areas, and for each we explore both general principles of open and reproducible research, as well as specific recommendations in a variety of settings. As the respective titles imply, for experimental design, data acquisition and preprocessing, studies are so varied that we provide general recommendations without recommending particular practices. Thus these sections focus mostly on thorough reporting and little on best practice. In contrast, for statistical modeling there are areas like task fMRI where mature methodology allows the clear identification of best practices. Likewise for the areas of data sharing, replication and reproducibility we focus exactly on those emerging practices that need to become prevalent.

We ask that authors challenge themselves: “If I gave my paper to a colleague, would the text and supplementary materials be sufficient to allow them to prepare the same stimuli, acquire data with same properties, preprocess in a similar manner and produce the same models and types of inferences as in my study?” This is an immense challenge! The purpose of this work is to guide researchers towards this goal and to provide a framework to assess how well a study meets this challenge.

2. Experimental Design Reporting

2.1. Scope

In this section we consider all aspects of the planned and actual experimental manipulation of the subject. This includes the type and temporal ordering of stimuli, feedback to be recorded and any subject-adaptive aspects of the experiment. It also encompasses basic information on

the experiment such as duration, number of subjects used and selection criterion for the subjects. It is impossible to prescribe the “right” design for all experiments, and so instead the focus is on the complete reporting of all facets of the design.

2.2. General Principles

For experimental design, the goal of reproducible research requires the reporting of how the subjects were identified, selected, and manipulated. This enables a critical reader to evaluate whether the findings will generalize to other populations, and facilitates the efforts of others to reproduce and replicate the work.

2.3. Lexicon of fMRI Design

While other areas of these guidelines, like MRI physics and statistical modeling, have rather well defined terminology, we find there is substantial variation in the use of experimental design terms used in fMRI publications. Thus Box 2.1 provides terminology that captures typical use in the discipline. Since the analysis approach is dependent on the fMRI design, providing accurate and consistent characterization of the design will provide greater clarity.

There is often confusion between block and mixed block/event designs [Petersen2012], or block designs composed of discrete events. Thus we recommend reserving the term “block design” for paradigms comprised of continuous stimuli (e.g. flashing checkerboard) or unchanging stimuli presented for the entire length of a block (generally at least 8 seconds). All other designs comprise variants of event-related designs and must have their timing carefully described.

Box 2.1. fMRI Terminology

Session. The experimental session encompasses the time that the subject enters the scanner until they leave the scanner. This will usually include multiple scanning runs with different pulse sequences, including structural, diffusion imaging, functional MRI, spectroscopy, etc.

Run. A run is a period of temporally continuous data acquisition using a single pulse sequence.

Volume. A volume (or alternatively “frame”) is single 3-dimensional image acquired as part of a run.

Condition. A condition is a set of task features that are created to engage a particular mental state.

Trial. A trial (or alternatively “event”) is a temporally isolated period during which a particular condition is presented, or a specific behavior is observed.

Event. The term “trial” and “event” are often interchangeable. However, in certain situations of ‘compound-trials,’ a trial will consist of multiple subunits; for example, a working memory task may contain subunits of encoding, delay, and retrieval. In these cases the subunits are labeled as “events” and the “trial” is defined as the overarching task.
Block. A block (or alternatively “epoch”) is a temporally contiguous period when a subject is presented with a particular condition.

2.4. Design Optimization

Especially with an event-related design with multiple conditions, it can be advantageous to optimize the timing and order of the events with respect to statistical power, possibly subject to counterbalancing and other constraints [Dale1999, Wager2003]. It is essential to specify whether the target of optimization is detection power (i.e. ability to identify differences between conditions) or estimation efficiency (i.e. ability to estimate the shape of the hemodynamic response, which requires jittering) [Liu2001]. It is likewise advisable to optimize your designs to minimize the correlation between key variables. For example, in model-based or computational fMRI experiments, variables such as reward, prediction error and choices will usually be highly correlated unless the design has been tuned to minimise this dependence. Be sure to include all possible covariates in a single statistical model to ensure variance is appropriately partitioned between these variables.

2.5. Power Analysis

The positive predictive value—the probability that an alternative hypothesis is true given a significant test—depends on the power of the study [Ioannidis2005], and underpowered studies have been found to been endemic in neuroscience as a whole [Button2013]. Power analysis for imaging is difficult as the outcome is typically the entire brain image, and not a single univariate measure [Mumford2012]. However there are power analysis tools available to account for intra- and inter-subject fMRI variance at each voxel [Mumford2008][5], as well as tools that account for the spatial structure of the signal [Joyce2012, Durnez2014][6]. Researchers hence can and should provide a realistic power computation that corresponds to the primary analysis of your study.

[5] http://fmripower.org.
[6] https://sourceforge.net/projects/powermap & https://neuropower.shinyapps.io/neuropower respectively.

2.6. Subjects

The population from which the subjects are sampled is critical to *any* experiment, not just those with clinical samples. Be sure to note any specific sampling strategies that limited inclusion to a particular group (e.g. laboratory members, undergraduates at your university).

One should take special care with defining a “Normal” vs. “Healthy” sample. Screening for lifetime neurological or psychiatric illness (e.g. as opposed to “current”) could have unintended consequences. For example, in older subjects this could exclude up to 50% of the population [Kessler2005] and this restriction could induce a bias towards a ‘super healthy,’ thus limiting the generalization to the population.

2.7. Behavioral Performance

The successful execution of a task is essential for interpreting the cognitive effects of a task. Report behavioral measures in and out of the scanner, measures that are appropriate for the task at hand (e.g. response times, accuracy). For example, provide statistical summaries over subjects like mean, range and/or standard deviation. Note any pre-scan training, the setting (e.g. mock scanner vs. bench) and any training goals.

3. Acquisition Reporting

3.1. Scope

This section concerns everything relating to the manner in which the image data is collected on each subject. Again we do not attempt to prescribe best MRI sequences to use, but focus on the reporting of acquisition choices.

3.2. General Principles

Research can only be regarded as transparent when the reader of a research report can easily find and understand the details of the data acquisition. This is necessary in order to fully interpret results and grasp potential limitations. For the work to be reproducible, there must be sufficient detail conveyed to actually plan a new study, where data collected will have, e.g., similar resolution, contrast, and noise properties as the original data.

More so than many sections in this document, MRI acquisition information can be easily organized in ‘checklist’ form (see Appendix D). Thus in the remainder of this section we only briefly review the categories of information that should be conveyed.

3.3. Device Information

One fundamental aspect of data is the device used to acquire it. Thus every study using MRI should report basic information on the scanner, like make and model, field strength, and details of the coil used, etc.

3.4. Acquisition-Specific Information

Each acquisition is described by a variety of parameters that determine the pulse sequence, the field of view, resolution, etc. For example, image type (gradient echo, spin echo, with EPI or spiral trajectories; TE, TR, flip angle, acquisition time), parallel imaging parameters, use of field maps, and acquisition orientation are all critical information. Further details are needed for functional acquisitions (e.g. volumes per run, discarded dummy volumes) and diffusion acquisitions (e.g. number of directions and averages and magnitude and number of b-values [Jones2012]).

3.5. Format for sharing

While there is some overlap with Section 7. Data Sharing, there are sufficient manufacturer- and even model-specific details that we consider here related to data format. When providing acquisition information in a manuscript keep in mind that readers may use a different make of scanner, and thus you should minimize the use of vendor-specific terminology. To provide comprehensive acquisition detail we recommend exporting vendor-specific protocol definitions or “exam cards” and provide them as supplementary material.

When primary image data are being shared, a file format should be chosen that provides detailed information on the respective acquisition parameters (e.g. DICOM). If it is impractical to share the primary image data in such a form, retain as much information about the original data as possible (e.g. via NIfTI header extensions, or “sidecar” files). However, sensitive personal information in the acquisition metadata should be carefully removed through appropriate anonymization procedures before sharing (see Section 6. Data Sharing).

4. Preprocessing Reporting

4.1. Scope

This section concerns the extensive adjustments and “denoising” steps neuroimaging data require before useful information can be extracted. In fMRI, the two most prominent of these preprocessing steps are head-motion correction and intersubject registration (i.e., spatial

normalisation), but there are many others. In diffusion imaging, motion correction, eddy current correction, skull stripping, and fitting of tensors (least squares, ROBUST, etc.) or more complex diffusion models are also key steps.

4.2. General Principles

As with other areas of practice, good reporting here requires authors to clearly detail each manipulation done to the data before a statistical or predictive model is fit. This is essential for reproducibility, as the exact outcome of preprocessing is dependent on the exact steps, their order and the particular software used. Moreover, the vast array of reprocessing options gives ample opportunity for p-hacking [Simmons2011] and thus it is vital to constrain choices and establish fixed preprocessing protocols whenever possible.

4.3. Software Issues

Software versions. Different tools implementing the same methodological pipeline, or different versions of the same tool, may produce different results [Gronenschild2012]. Thus ensure that the exact name, version, and URL of all the tools involved in the analysis are accurately reported. It is essential to provide not just the major version number (e.g., SPM12, or FSL 5.0) but indicate the exact version (e.g. SPM12 revision 6225, or FSL 5.0.8). The version of software interpreters, e.g. Matlab or Python, should also be included as well as compilation conditions when known. To avoid ambiguities on the tool name, consider adding a Research Resource Identifier (RRID[7]) [Bandrowski2015] citation for each tool used in addition to reporting the version. RRIDs index everything from software to mouse strains, and provide a consistent and searchable reference.

In-house pipelines & software. When using a combination of software tools, be sure to detail the different functions utilized from each tool (e.g., SPM's realign tool followed by FreeSurfer's boundary-based registration; see Reproducibility section for more on pipelines). In-house software should be described in detail, giving explicit details (or reference to peer-reviewed citation with such details) for any processing steps/operations carried out. Public release of in-house software through an open code repository is strongly recommended (e.g. Bitbucket or Github).

Quality control. Quality control criteria, such as visual inspection and automated checks (e.g., motion parameters), should be specified. If automated checks are considered, metric and criteria thresholds should be provided. If data has been excluded, i.e., due to scrubbing or other denoising of fMRI time series or removal of slices or volumes in diffusion imaging data, this should be reported.

[7] http://www.force11.org/group/resource-identification-initiative & https://scicrunch.org/resources.

Ordering of steps. The ordering of preprocessing steps (e.g., slice time correction before motion correction) should be explicitly stated.

Handling of exceptional data. Sometimes individual subjects will have problems, e.g. with brain extraction or intersubject registration. Any subjects that require unique preprocessing operations or settings should be justified and explained clearly, including the number of subjects in each group for case-control studies.

5. Statistical Modeling & Inference

5.1. Scope

This section covers the general process of extracting results from data, distilling down vast datasets to meaningful, interpretable summaries. Usually this consists of model fitting followed by statistical inference or prediction. Models relate the observable data to unobservable parameters, while inference quantifies the uncertainty in the estimated parameter values, including hypothesis tests of whether an observed effect is distinguishable from chance variation. Inference can also be seen as part of making predictions about unseen data, from the same or different subjects. Note that while we make a clean distinction between preprocessing and modeling, there is some overlap (e.g. movement can be a preprocessing "correction" or part of a model) and they can in fact interact [Strother2002].

5.2. General Principles

For statistical modeling and inference, the guiding principle of openness dictates that the reader of published work can readily understand what statistical model was used to draw the conclusions of the paper. Whether accidental or intentional (i.e. for brevity), omission of methodological details needed to reproduce the analyses violates these principles. For maximal clarity, be sure to describe all data manipulation and modeling in the methods section [Gopen1990]. For maximal transparency, report all regions of interest (ROIs) and/or experimental conditions examined as part of the research, so that the reader can gauge the degree of any HARKing (Hypothesizing After the Results are Known) [Kerr1998]; to absolutely minimize HARKing, register your hypothesis and analysis before exploring the data.

5.3. Assumptions

Every modelling and inference method described below makes assumptions about the data analyzed. Take the time to understand assumptions and the implications on your results. Just to name a few for linear models: The correct model (no missing variables), additive model (no missing interactions), normality of errors (no outliers), etc. [Luo2003]. In group analyses, pay

special attention to reversal paradoxes where an effect can be flipped if an important covariate is omitted [Tu2008]. And always take care when removing outliers, as informative observations may be discarded. Consider using the analysis that has the weakest assumptions while still providing the needed inference, e.g. tools like permutation or bootstrap (e.g. [Eklund2015,Bellec2010].

5.4. Software

See the previous section for details on how to describe the exact software and pipeline used.

5.5. Mass Univariate Modelling

A simple univariate model fit to each voxel or surface element is known as a mass univariate modelling approach, and is an essential tool for everything from task fMRI, structural MRI measures like Voxel Based Morphometry, scalar diffusion measures like Fractional Anisotropy or even resting state fMRI, when measured with low frequency variance (see *Other Resting-State Analyses* below). Regardless of the type of data, a mass-univariate linear model is specified by five types of information: Dependent variables, independent variables, model, estimation method and inference method (where inference refers to quantification of uncertainty of estimated parameters and hypothesis testing).

While the *dependent variable* (or response) may be unambiguous (e.g. for T2* BOLD), be sure to identify it in any nonstandard analysis. Itemize each *independent* (or explanatory) *variable* in each model used. In a first level fMRI model, this includes the usual condition effects, as well as motion regressors added to explain nuisance variation. In a second level or group model, independent variables include the group assignment (e.g. patient vs. control) or other between-subject effects that may or may not be of interest (e.g. age or sex). Report non-trivial contrasts, linear combinations of independent variables, that are used to interrogate the experimental effect of interest. Variables generally do not need to be centered[8], but do indicate how and if this was done.

While software may make the *model* and *estimation method* seem 'automatic', a short description is needed for a complete scientific report. See Appendix C for examples of short descriptions of commonly used task fMRI models. Beyond the mass univariate model, there is growing use of other types of models, including local multivariate, whole-brain multivariate, etc. Regardless of the model, be sure to note the essential details of the estimation procedure.

The *inference method* is used to flag some voxels or elements as "active" or "different", as distinguished from background noise, and is a crucial final step. In brain imaging, inference

[8] http://mumford.fmripower.org/mean_centering.

usually amounts to a thresholding procedure, though if ROIs are used, it should also include computation of confidence or credible intervals. ROIs can be defined based on anatomy or previous literature [Poldrack2007], or based on functional results with great care and attention to circularity (see *Extracted Data* below).

While the dangers of multiple testing are generally well understood, one review found only 59% of studies used a correction for multiple testing [Carp2012]. While there are domains where stringent corrections are not suitable (e.g. presurgical fMRI, where false positive risk has to be balanced against the dangers of missing eloquent tissue), every study needs to address the issue, using a standard correction method or defending the lack of correction. Thus clearly state the type of inference and the manner of multiple-testing correction or reasons for no correction. Note that the inference method description "5% cluster wise inference" doesn't specify the cluster-forming threshold nor the multiple-testing correction measure (e.g. familywise or false discovery rate). Also describe the volume, sub-volume, or surface domain for which multiple-testing correction has been performed, and whether multiple volumes/ROIs have been corrected for.

5.6. Connectivity Analyses

Functional and effective connectivity encompass a broad range of methods, from data-driven multivariate or clustering methods on high resolution voxel-wise data, to highly structured physiological-based models on a small number of regions. Methods are still evolving for resting-state fMRI in particular, but careful execution of a study requires considering topics similar to task fMRI modeling: response variables, model, estimation method and inference method.

The goal of most connectivity analyses is to understand the relationships among multiple *response (dependent) variables*. These variables can be defined by regions-of-interest (ROIs), in which case be sure to report the number of ROIs and how the ROIs are defined (e.g. citable anatomical atlas; auxiliary fMRI experiments). State whether analyses were carried out as a voxelwise whole-brain analysis or by using cortical surfaces or CIFTI 'grayordinates' (surface vertices + subcortical gray matter voxels [Glasser2013]). For seed-based analyses, or small-scale (e.g. Bayes Net) methods, provide the rationale and method for selecting the particular ROIs. Carefully describe how time series were attributed to each ROI (e.g. averaging, median, or eigenvariate), and detail any additional (temporal or spatial) filtering or transformations (e.g. into wavelet coefficients) used, or nuisance variables (e.g. motion parameters) 'pre-regressed' out of the data.

A number of exploratory multivariate methods are used to understand high-dimensional fMRI data in a lower dimensional space. These include Principal Component Analysis, Multidimensional Scaling, Self Organizing Maps, and Independent Component Analysis (ICA), of which ICA is probably the most widely used. For any such method report the model variant

(e.g. spatial or temporal ICA), the estimation method (i.e. algorithm) and the number of dimensions or components used and, crucially, how this number was selected. ICA fitting and interpretation depends on choices about scaling (a.k.a. variance normalization), both to the data before fitting and as a constraint between spatial and temporal components; describe the type of scaling (i.e. variance normalization) applied to data and extracted components. Always report how the components were sorted and selected for analysis, whether for the primary analysis or as part of a denoising procedure; if using a post-hoc task regression model supply associated task model details (see above).

As any nuisance variation jointly influencing multiple voxels/regions can be mistaken for brain connectivity, it is essential that careful preprocessing has been applied, including artifact removal (See Section 3).

For many connectivity analyses the *model* is nothing more than the summary measure of dependence, e.g. Pearson's (full) correlation, partial correlation, mutual information, etc. However, be sure to note any further transformations (e.g. Fisher's Z-transform, regularization of partial correlation estimates). For seed-based analyses, describe the voxel-wise statistic or regression model (and other covariates) used. For regression-based group ICA analyses ("dual regression", or "PCA-based back-reconstruction"), clearly describe how the per-subject images are created. As with task-fMRI, any group analysis should be described in terms of dependent variables, independent variables, model, estimation method, and inference method. For graph analysis methods based on binary connection matrices, state how thresholding was done and consider the sensitivity of your results to the particular threshold used. While lag-based methods like Granger have been criticised for fMRI [Smith2012], they remain suitable for EEG and MEG.

For functional connectivity, *inference* typically focuses on making statements comparing two or more groups of subjects or assessing the impact of a covariate. Ensure that it is clear what is the response being fed into the group model. For some connectivity analyses, like Structural Equation Modelling or Dynamic Causal Modelling, the inference concerns selecting among a set of models. Be sure to justify and enumerate the models considered and how they were compared; describe how evidence for model selection was aggregated over multiple subjects. Discuss the prior distributions used and their impact on the result. For graph-based analyses, detail the construction of adjacency matrices (i.e. what was binarized and how), or if using weighted measures, how the weights are computed. Note the problems of comparing networks of different size or overall connection density [vanWijk2010].

5.7. Other Resting-State Analyses

The analysis of resting state data does not necessarily incorporate connectivity. Methods like Amplitude of Low Frequency Fluctuations (ALFF) [Zang2007] and fractional ALFF (fALFF) [Zou2008] summarise brain activity with absolute (ALFF) or relative (fALFF) BOLD variance, and Regional Homogeneity (ReHo) [Zang2004] measures local consistency of signals. These

methods produce a map per subject that can be analyzed with a mass univariate model (see above).

5.8. Multivariate Modelling & Predictive Analysis

Predictive methods focus on estimating an outcome for each experimental trial, block or subject, often using multivariate models. Multivariate methods exploit dependencies between many variables to overcome the limitations of mass univariate models, often providing better explanatory or predictive models. In brain imaging, predictive methods are often called decoding or multi-voxel pattern analyses [Norman2006]; an example of a multivariate analysis is representational similarity analysis [Kriegeskorte2008]. A complete description should include details of the following: Target values, features, predictive model, and training and validation method.

The *target values* are the outcomes or values to be predicted, which may be discrete or continuous. It should be made clear exactly what is being predicted, and what are the relative frequencies of this variable (e.g. proportions in each group, or a histogram for a continuous target). Unbalanced group sizes are not a problem but require appropriate evaluation measures, as described in the next section.

The *features* are the variables used to create the prediction, and often are not the raw data themselves but derived quantities. In addition, some features may be discarded in the process of feature selection. If feature selection is based on the target values there will be a tendency to over-fit, and then feature selection must be embedded in the validation framework (see below). It is essential that the analysis pipeline is described in sufficient detail to capture the definition of each element of the feature, any feature selection that precedes model-training, and any feature transformations (including possible standardization).

The *predictive model* is the type of method used to map features to targets. Typical examples include linear discriminant analysis and support vector machines. The model is distinct from the algorithm or training procedure used to optimize the parameters of the method (i.e. usually to minimize prediction error on held-out data). Be sure to clearly identify the model used and (if used) the specific machine learning library used.

Finally, the *training and validation method* is perhaps the most important facet of a predictive analysis. This comprises the algorithm used to build the predictive model and the framework used to evaluate the model. Training may be nothing more than fitting a regression model, but more typically consists of a complex algorithm that depends on the tuning of hyper-parameters. In the validation step the model's predictive performance (e.g. accuracy) is assessed using an independent dataset or a cross-validation framework (e.g. leave-one-out, k-fold cross-validation, stratified cross validation). Clearly specify the algorithm used, what objective function was optimized, how the algorithm's convergence was established (for iterative methods), and any

post-processing of the fitted model. Be sure to clearly describe how hyper-parameters were estimated, including the choice of the hyper-parameter grid, figure-of-merit optimized, the type of validation scheme used, and the use of an averaging strategy to produce a final classifier. In particular, identify which hyper-tuning parameters were optimized outside vs. inside a cross-validation loop: The reported accuracy is evidently valid only if *all* estimated hyper-parameters are optimized inside the loop as part of a nested cross-validation procedure with three-way splits providing disjoint training, testing and validation datasets, and no information from the test data enters the optimization of any of the parameters.

6. Results Reporting

6.1. Scope

The reporting of statistical results is inextricably tied to the statistical modeling and inference procedures of the previous section. However, a scientific investigation invariably requires dozens of analyses, inferences and views of the data, and thus any published report typically contains a subset of all output of every statistical procedure completed. Thus we feel that results reporting deserves its own section here, providing guidance on how authors should select and present the outcomes of the modeling process.

6.2. General Principles

Transparency of published research requires that the reader can easily interpret the results shown and, crucially, what results were considered but then not shown. Unreported selective inference (a type of file drawer effect) inflates the significance of results shown and will stymie efforts to replicate a finding.

6.3. Mass Univariate Modelling

For reporting single univariate outcomes, like average BOLD response in an ROI or global mean FA, there is a wealth of best practice guidelines available[9] [Altman2008]. For mass univariate models, there are four general classes of information that need to be carefully described: Effects tested, tables of brain coordinates, thresholded maps, parcellated maps, and extracted data.

A complete itemization of the *effects tested* must be presented, identifying the subset that are presented. This is necessary to understand the true magnitude of the multiplicity involved and the potential danger of selection biases. For example, if a study has a multifaceted design

[9] http://www.equator-network.org.

allowing various main and interaction effects to be considered, effects tested and omitted should be enumerated, including references to previously published results on the current dataset. A full sense of how extensively the data has been explored is needed for the reader to understand the strength of the results.

Tables of coordinates historically have often been the *only* quantification of the results, and now should be complemented with sharing of full statistic images (see, e.g., NeuroVault[10]). If coordinates are reported, each table or sub-portion of a table should be clearly labeled as to what contrast / effect it refers to (nature of the contrast, individual versus group result, group size), and should have columns for: Anatomical region, X-Y-Z coordinate, T/Z/F statistic, and the P-value on which inference is based (e.g. voxel-wise FWE corrected P; or cluster-wise FDR corrected P); if cluster-wise inference is used, the cluster statistic (e.g. size, mass, etc) should be included. Avoid having multiple columns of results, e.g. multiple XYZ columns, one for increases, one for decreases, or one for left hemisphere, one for right hemisphere.

The table caption should clearly state (even if in repetition of the body text) the significance criterion used to obtain these coordinates, and whether they represent a subset of all such significant results (e.g. all findings from whole-brain significance, or just those in a selected anatomical region). If T or F statistics are listed, supply the degrees of freedom. Whenever possible, provide effects sizes at the selected coordinates together with 95% confidence intervals. Finally, the space (i.e., Talairach, MNI, fsaverage) of the coordinate system should be noted.

The *thresholded map figures* perhaps garner the most attention by readers and should be carefully described. In the figure caption clearly state the type of inference and the correction method (e.g. “5% FWE cluster size inference with P=0.001 cluster-forming threshold”), and the form of any sub-volume corrections applied. For small volume or surface ROI corrections, specify whether or not the ROI was identified prior to any data analysis and how it was defined. Always annotate threshold maps with a color bar for the statistic values; when showing multiple maps, use a common color bar when feasible; and always indicate right and left. Avoid common fallacies in interpreting maps; e.g. an activation in region A but not region B doesn’t mean A is significantly more active than B [Poldrack2008], and lack of activation is not evidence of no activation. Most important, publicly share the original statistic images, unthresholded and thresholded, so readers can explore the maps themselves in 3D (see Data Sharing below).

Extracted data from images aids the interpretation of the complex imaging results, and is presented as effect magnitudes, bar plots, scatter plots or activation time courses. Computed from a single voxel/vertex, or an average or principal component of a set of voxels/vertices, they however present a great risk for “circularity” [Vul2009; Kriegeskorte2009]. Specifically, when the voxels summarized are selected on the basis of a statistic map, they are biased estimates of the effect that map describes. Thus it is essential that every extracted summary clearly address the

[10] http://neurovault.org.

circularity problem; e.g. “derived from independently-formed ROI”, or “values based on voxels in a significant cluster and are susceptible to selection bias”. When working with single regions and uncorrected P-values, consider the current discussions on the limitations of P-values [Wasserstein2016] and in particular how P=0.05 can amount to very weak evidence of an effect [Nuzzo2014].

6.4. Functional Connectivity

The critical issues when reporting functional connectivity differ between types of approaches, for example exploratory multivariate vs. seed-based correlation methods, which provide whole maps, versus confirmatory multivariate methods for a handful of regions.

When reporting multivariate decomposition methods like ICA or PCA, state how the number of components were selected. With either ICA or seed-based analyses, when conducting inference on multiple networks, be sure to account for multiplicity when searching over the networks. For example, if testing for patient vs. control differences in the default mode, attentional, visual and motor networks, the inference must account for not only the voxels within networks, but additionally for searching the four IC maps for significance.

6.5. Multivariate Modelling & Predictive Analysis

While it may appear that predictive analyses are trivial to report (“Accuracy was X%, p(X> chance)<0.001”), there are in fact two broad types of information to convey: Evaluation & interpretation.

Evaluation refers to the assessment of a fitted classifier on out-of-sample data. As shown in the tabular listing, there are several measures of classifier performance that should be reported aside from overall accuracy (percentage of correct predictions). For example, when group sizes are unequal, be sure to also report average or balanced accuracy (accuracy per group, averaged).

Do not make claims of “above chance accuracy” unless based on confidence intervals or some appropriate formal test, ideally a permutation test [Combrisson2015]. For regression report prediction R^2, though be aware this may be negative when the explained variance is low (but is not necessarily truly zero). Avoid using a correlation coefficient as an evaluation metric (computed between actual and held-out-predicted continuous values) as this is susceptible to bias [Hastie2011, Ch7].

Interpretation of the fitted classifier allows potential insights to brain function or structure that drives prediction, though must be done with care (see e.g. [Haufe2014]). In particular, be sure not to over-interpret whole brain weight maps as localizing the source of decoding information,

as the very multivariate nature of the method means it is impossible to isolate a single region as being responsible for classification. Voxels or vertices containing significant information may receive small or zero weight if a regularisation penalty is used in fitting. Conversely, voxels/vertices with high absolute weight may contain little predictive signal, but may mostly serve to cancel correlated noise, thus improving classifier performance. The same caveats apply in the context of an encoding model that predicts brain responses from various experimental features; e.g. a predictor with large weight may be cancelling effects of other predictors and may not by itself contain any information about the voxel in question. Solutions to this problem include adding relevant (e.g. smoothness, continuity) priors to the multivariate model to improve its interpretability, and using resampling techniques like *stability selection* to enhance the reliability of the estimated classifier weights [Varoquaux2012]. Mapping procedures that conduct the same analysis at every location, such as multivariate searchlight mapping, can also outline regions that are predictive in isolation of activity elsewhere and thus complement whole-brain classification methods.

Finally, just as mass univariate analyses can be weakened by 'data dredging' through scores of contrasts, a predictive analysis is also less meaningful if it is the (say) 10th analytical approach tried on a single dataset. It is essential to itemize the analyses attempted, both to convey what doesn't work and the size of the model space considered.

7. Data Sharing

While previous sections have largely described good practice that is (more or less) prevalent in the community, this and the next section concerns practices that are currently scarce. Thus these sections are necessarily more prescriptive, providing explicit suggestions on ways to change how we conduct studies, meeting the challenges of making neuroimaging science as transparent and reproducible as possible.

7.1. Scope

Neuroimaging, relative to other disciplines like genetics and bioinformatics, has lagged behind in widespread acceptance of data sharing. This section outlines the practicalities of sharing of data and results, including issues related to the use of software infrastructure, data repositories and the details surrounding retrieval of data.

7.2. General Principles

Data sharing is one of the cornerstones of verifiable and efficient research, permitting others to reproduce the results of a study and maximizing the value of research funds already spent. However, to fully realize this value, data should not just be "available on request", but shared in

a data repository that is well organized, properly documented, easily searchable and sufficiently resourced as to have good prospects for longevity. In this respect, we support the FAIR data guiding principles according to whom data should be Findable, Accessible, Interoperable and Re-usable[11]. To fulfill these goals, there are four key elements to a successful data sharing effort: planning, databases, documentation, & ethics.

7.3. Planning for Sharing

Data sharing is most onerous when done as an afterthought [Halchenko2015]. Instead, if data sharing is considered when a study is planned and initiated as part of a complete data management plan, the additional effort required will be minimal. A data sharing plan should establish a viable path for any qualified researcher to gain access to the data. Planning efforts should begin with ethics (consent forms) and funding agencies, by creating realistic funding roadmaps for long-term stewardship of data. Having realistic workflows and a proper technical infrastructure are important prospective steps. In addition, there are a number of considerations and hurdles for publishers, each of which have their own policies for data sharing.

A key to an effective data sharing is the use of a strict naming structure for files and directories. This regularity brings a number of benefits, including greater ease in finding errors and anomalies. But most valuable, organized data facilitates extensive use of scripting and automation, reducing time needed for analysis and quality control (QC). Best practice is to use a standardized data structure; for example, the recently developed BIDS standard[12] provides a detailed directory hierarchy for images and a system of plain text files for key information about a study's data. This structure is used by OpenfMRI[13], making it easy to upload data to that repository. Whatever the system, arranging your data in a regular structure will simplify all efforts to manipulate and—specifically—share your data.

Another essential decision to make early in a study is exactly what kinds of data are to be shared. The exact data shared must be consistent with the ethics of the study (see below, Ethics). But once suitably anonymised, there are still the various versions of image data to choose from: DICOM files from the scanner for each subject; "raw" converted data (e.g. NIFTI), free of any preprocessing; ready-to-model fMRI data for each subject, having all of the basic processing completed; per-subject summary maps, e.g. one effect/contrast image per subject in fMRI; per-study statistic maps. Sharing raw data gives more options to other users (DICOM being the rawest), while sharing preprocessed images makes it easier for others to immediately start analyzing your data. Sharing of extensively processed data, such as (unthresholded) statistical maps and underlying structural data (e.g., volumes and cortical surfaces of individuals and/or group averages) can be very valuable, enabling readers of an article to access much more information than can be conveyed in a static image in a publication. Finally, share as

[11] https://www.force11.org/node/6062.
[12] http://bids.neuroimaging.io.
[13] http://openfmri.org.

many QC measures as possible as well as providing a PASS/FAIL summary for each dataset, allowing easy selection of usable data while allowing users to revisit QC decisions.

Decide at the outset with whom the data is to be shared and at what stage, as it may be useful to share data with collaborators prior to publication, then more freely after publication. We support the widest sharing of data possible, but in certain (e.g. clinical) circumstances this may not be possible without additional protective measures, such as an explicit Data Use Agreement (DUA), or the use of "data on request" services (e.g. Clinical Study Data Request[14]) that enable assessment of compliance with such terms. Again consistent with ethics, have a data management plan that clearly specifies whether data can be freely distributed, or under exactly what constraints it can be shared. For example, in large-scale databases, data may be freely shared within a project, with some limits to other related projects, or with yet more constraints to the general public. Establishing these limits before a single subject is scanned will save many headaches down the road. Instead of setting the exact rules for data use yourself, consider using an established license, like from the Creative Commons[15] or Open Data Commons[16], saving yourself time and making the terms of use clear to users.

For large-scale, multi-site studies, the greater effort put into harmonization of experimental paradigms, data acquisition, analysis and modeling, the easier it will be to amalgamate the data later. If separate databases are used, then an ontological standardization is important, establishing how to map data fields and the data dictionaries between sites.

Another facet to consider is the sharing of data analysis pipelines scripts and any provenance traces. These are generally free of ethical concerns (unless protected information like subject names creeps into a script!) and there is great value in allowing others to recreate your results and apply your methods to new data. This is discussed in greater detail below (see Documentation).

Finally, whenever possible use publication as the milestone for sharing. The longer you wait the harder it becomes to assemble all the pieces (data, scripts, etc), plus the article can then have the DOI/URL reference to the data.

In short, a comprehensive data management plan—that involves all authors, collaborators, funding agencies, and publishing entities—is essential no matter what is shared and should be considered from the outset of a study. Without such planning, in a jumble of folders and after a graduate student or post-doc has moved on, data can effectively be lost.

---

[14] http://clinicalstudydatarequest.com.
[15] http://creativecommons.org/licenses.
[16] http://opendatacommons.org/licenses.

7.4. Databases

While a highly organized arrangement of data in a folder hierarchy is prerequisite for good data management, it does not in itself constitute a database. A database, in addition to organizing data, is searchable and provides access controls. Databases for imaging data may include non-imaging data and allow direct entry of data. There are a number of imaging-oriented databases, ranging in scale, complexity, features and, crucially, effort needed to install and maintain them. As individual users are unlikely (and not advised!) to create imaging databases, we review the considerations when choosing a database.

Consider access control options, and exactly who and which types of users should be allowed to enter data, and access the data. There may be some types of data (e.g. sensitive behavioral tests or essential personal information) that require special, restricted access. The ability to modify existing data should be highly restricted, ideally with a form of audit control that records the nature of the changes. A public repository must of course provide access for external users.

Comprehensive search functionality is important, especially for large scale, multi-project databases. Useful features include being able to select subsets of data of interest, e.g. finding subjects that have a certain age range, IQ and a clinical diagnosis, with two different imaging modalities. Once a selection is made, some systems may only let you download data, while others may provide quick visualization or extensive analysis options. Especially when working with large repositories, the availability of a scriptable query interface can be handy for complex queries.

Consider the ability of a system to handle heterogeneous data. Most imaging databases will accommodate the most basic demographic information, but may not accept more than one modality (e.g. both MRI and EEG) or other types of essential data, like clinical evaluations or batteries of psychological tests. Consider carefully all the data that comprises your studies and whether it can all be stored in one unified system. Some systems allow staff to directly enter subject information, and even conduct batteries of psychological tests on subjects, eliminating double entry and reducing the risk of errors.

Finally, assess the complexity of installation and maintenance of a system. At a single site, the system must be easy to install and maintain, while a database for a multi-site study will necessarily be more complex and require adequate expertise to manage. As part of this, ensure there is detailed documentation for maintainers, as well for end users on how to navigate the resource. And if serving as a repository, a database must additionally possess a long-term preservation plan.

Now, with a variety of mature imaging databases available, building a de novo home-grown database cannot be recommended. For example, IDA [Mueller2005], XNAT [Marcus2007], COINS [Scott2011], and LORIS [Das2012] are four established and well-resourced systems for

longitudinal, multi-modal, web-based data storage and querying, with proper user control. Some of these tools interface to high performance computing platforms for mass processing (e.g. IDA to LONI [Dinov2009] or LORIS to CBRAIN [Das2016]) and can be an important element in reproducibility (see *Reproducibility* section).

While these established databases are becoming easier to install and maintain, we acknowledge that in low resource environments they may be impractical. In these settings, the use of highly structured storage of imaging data (see BIDS above) and extensive use of scripting is the best approach, and facilitates a transition later to a formal database. In most research environments, however, informatics support should be regarded as a necessity and funded accordingly, if for no other reason than to obtain the maximal value of the data collected, now and for years to come.

7.5. Documentation

Even an organized and searchable database is of no use, unless users have access to information describing what is actually stored in the repository. Clear documentation on the studies within a repository, the data acquisition and experimental paradigm detail are all examples of information that are needed to make use of information in a database. If processed data and results are stored, details on the preprocessing and models fit are also essential. The documentation should be written for a wide audience, including members from multiple disciplines. The extensive documentation for the Human Connectome Project[17] provides a great example of how to describe data (unprocessed and minimally preprocessed) as well as the acquisition and preprocessing methods in a large and complex database.

A form of self-documentation is provenance, i.e. recording exactly what happened to data through preprocessing and modeling. These “provenance traces” can help track-down problems and provide invaluable reference for others who want to replicate previous studies. While provenance is not usually recorded, the AFNI BRIK[18] and MINC[19] formats have forms of provenance tracking, and the NIDM project[20] is developing a framework to save this information in a standard format. Pipeline software like LONI Pipeline[21] or Nipype[22] explicitly provide such provenance records.

[17] http://humanconnectome.org/documentation.
[18] http://afni.nimh.nih.gov/afni/doc/faq/39.
[19] http://www.bic.mni.mcgill.ca/software/minc.
[20] http://nidm.nidash.org.
[21] http://loni.usc.edu/Software/Pipeline.
[22] http://nipy.org/nipype.

7.6. Ethics

There are ethical issues that promote and constrain data sharing. In favor of data sharing is the gain in knowledge from re-used data that can improve health in patient populations, not to mention how sharing can make better use of the public funds spent on research. But the data to be shared pertains humans whose rights and wishes must be respected. Hence suitably crafted ethics and consent documents are essential for data sharing. While in the United States de-identified data is currently not "protected health information" and should be able to be shared, regulations differ between countries and institutions and are subject to change. In the United Kingdom, a separate set of data protection laws exist and must be complied with. Thus be sure to consult your ethics or institutional review board, and any data protection office, before acquiring data with the intent of sharing, as well as before releasing data. The Open Brain Consent project[23] can also be of use, providing sample forms written specifically to account for later sharing of data. Some level of anonymization will be required, ensuring all sensitive personal information is withheld or suitably coarsened or obscured (e.g. reporting only age in years at scan time instead of birth date), and/or applying a "de-facing" procedure to anatomical MRI images. Careful 'scrubbing' (e.g. removing subject names from DICOM files, or analysis pathnames) is required to ensure no personal information is discussed.

8. Reproducibility

We make the distinction articulated by [Peng2011] and others that *reproducible* results can be recreated by others using the same data and software as shared by the original authors, while a *replication* is the traditional scientific goal of independent researchers using independent data and possibly distinct methods to arrive at the same scientific conclusion (see Appendix B). While some have argued that reproducibility is secondary, and that "one should replicate the result not the experiment" [Drummond2009], recent failures to replicate high-impact results and occasional but acutely concerning examples of outright fraud have made the case for the importance of reproducibility.

8.1. Scope

We focus on analysis-level reproducibility, i.e. the ability to recreate the results of a well-defined analysis using the same data. All of the recommendations of this paper are in the service of the clear, unambiguous reporting of design, data and analysis workflow as recommended by the FAIR principles[24] [Wilkinson2016], EQUATOR Network[25] [Altman2008], Reward Alliance[26] and

[23] http://open-brain-consent.readthedocs.org.
[24] https://www.force11.org/node/6062.
[25] http://www.equator-network.org.
[26] http://researchwaste.net.

ARRIVE guidelines[27] [Kilkenny2010]. To further make your analysis as reproducible as possible, ensure it is *documented*, *archived* and *citable*.

8.2. Documentation

As even the operation system version can influence the exact results obtained [Glatard2015], be sure to cite the computational infrastructure as well as versions of software used.

When the analysis involves multiple tools, some formal description of the workflow connecting these tools should be provided. Tools such as BrainVISA [Cointepas2001], LONI pipeline [Rex2003], NiPype [Gorgolewski2011], PSOM [Bellec2012], Automatic Analysis Pipeline [Cusack2015], SPM batch [Penny2006] and CBRAIN [Das2015] may help structure and describe workflows. myExperiment [DeRoure2010] can be used to share and run workflows online (see for instance this FSL fMRI workflow from the LONI Pipeline environment[28]).

Any additional information on provenance will aid in efforts to reproduce your analysis. For example, tools like NiPype & the LONI Pipeline Processing environment [MacKenzie-Graham2008] records an exact "provenance trace" of the analysis, and the MINC[29] and AFNI BRIK formats also store histories of analysis commands used to create a file. The Neuroimaging Data Model (NIDM [Keator2013]) is being actively developed to describe all steps of a data analysis in analysis-program-independent fashion.

Even when the data and workflow used in an analysis are properly documented, it may not be easy to reproduce the exact same data, for instance figures, as presented in a publication. Consider the use of literate programming tools such as iPython notebooks (used for instance in [Waskom2014]), or R-based Sweave [Leisch2002]. Another example involves 'scene' files that store all of the information (including links to the associated data files) that is needed to exactly reproduce a published figure. Currently, scene files are supported by the Connectome Workbench [Marcus2013] and Caret [VanEssen2001] software platforms.

8.3. Archiving

The analysis documentation should be archived in a long-term accessible location on the web. Of course, even with excellent documentation resources may disappear, become inaccessible, or change, further challenging reproducibility.

Open-source software is more likely to be available long term and is thus recommended. Whenever available, report on the availability of tools in repositories such as the INCF software

[27] https://www.nc3rs.org.uk/arrive-guidelines.
[28] http://www.myexperiment.org/workflows/2048.
[29] http://en.wikibooks.org/wiki/MINC/Reference/MINC2.0_Users_Guide.

center[30], the NITRC Resource Registry[31], or in software suites such as NeuroDebian [Halchenko2012] or Lin4Neuro [Nemoto2011].

A good way to facilitate reproducibility is to create and release a virtual machine (VM) or a container with the software and pipelines used in the analysis. A good starting point is the NeuroDebian VM[32] that can be further customized for a particular use case. Examples of other practical solutions that demonstrate this approach are the Nipype vagrant box and the NITRC Computational Environment[33] (used e.g. in [Ziegler2014]), both NeuroDebian-based VMs, and Niak[34] (available on DockerHub[35]). Of course licensing may prevent creating comprehensive VM. With Matlab code, consider using the Matlab Compiler to create standalone applications or free alternatives such as GNU Octave.

8.4. Citation

URLs tend to "decay" over time [Habibzadeh2013], making them inappropriate to cite online material permanently. Instead, Digital Object Identifiers (DOIs) provide a persistent way to index digital data. Various platforms are now available to host your data and workflows and create DOIs for them, such as Zenodo[36], figshare[37] or the Harvard Dataverse system[38] (see examples in [Tustison2014], [Soelter2014] & [Holmes2015]). Data hosting projects may also register DOIs directly to organizations such as DataCite[39] or CrossRef[40] .

9. Conclusions

In this work we have attempted to create an extensive (but not comprehensive) overview of reporting practices and, to a lesser extent, the practices themselves needed to maximize the openness and replicability of neuroimaging research. We have focused exclusively on MRI, but many of the suggestions and guidelines will easily translate to other areas of neuroimaging and related fields.

---

[30] http://software.incf.org.
[31] http://www.nitrc.org.
[32] http://neuro.debian.net or https://hub.docker.com/_/neurodebian.
[33] http://www.nitrc.org/plugins/mwiki/index.php/nitrc:User_Guide_-_NITRC_Computational_Environment.
[34] http://simexp.github.io/niak.
[35] http://hub.docker.com.
[36] http://zenodo.org.
[37] http://figshare.com.
[38] http://dataverse.harvard.edu.
[39] http://www.datacite.org.
[40] http://www.crossref.org.

This document is inevitably dated by the current technology and means of reporting scientific results. As these evolve this document will need to be updated and revised. Updates and the current version of these guidelines will be available at the COBIDAS website[41].

A reaction to the extensive checklists (Appendix D) could be "What human can put all that into their paper!?", and our response is that ideally no human should, that is it should be a computer's job. Many of the elements to report exist in some machine readable form, but in countless different forms. Thus the next important work to be done is to align these checklists to a controlled vocabulary, e.g. from NIH's Common Data Elements[42]. Once terms are set like this, they are more easily entered into a format like ISA-Tab[43], a table-based system for recording machine-readable metadata (now used, e.g., by the journal *Scientific Data*), and the stage is then set to develop the tools to automatically extract and export such vital meta-data.

Acknowledgements

Please see Appendix A.

References

Alsop, D. C., Detre, J. a, Golay, X., Günther, M., Hendrikse, J., Hernandez-Garcia, L., … Zaharchuk, G. (2014). Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: A consensus of the ISMRM perfusion study group and the european consortium for ASL in dementia. *Magnetic Resonance in Medicine*. doi:10.1002/mrm.25197

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: reporting guidelines for health research. *The Lancet*, 371(9619), 1149–1150. doi:10.1016/S0140-6736(08)60505-X

Bandrowski, A., Brush, M., Grethe, J. S., Haendel, M. a., Kennedy, D. N., Hill, S., … Vasilevsky, N. (2015). The Resource Identification Initiative: A cultural shift in publishing [version 1; referees: 2 approved]. *F1000Research*, 4:134. doi:10.12688/f1000research.6555.1

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, 51(3), 1126–39. doi:10.1016/j.neuroimage.2010.02.082

Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P., & Evans, A. C. (2012). The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Frontiers in Neuroinformatics*, *6*(April), 7. doi:10.3389/fninf.2012.00007

[41] http://www.humanbrainmapping.org/cobidas.
[42] https://www.nlm.nih.gov/cde.
[43] http://www.isa-tools.org.

Button, K. S., Ioannidis, J. P. a., Mokrysz, C., Nosek, B. a., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, 14(5), 365–76. doi:10.1038/nrn3475

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage*, 63(1), 289–300. doi:10.1016/j.neuroimage.2012.07.004

Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., & Benali, H. (2001). BrainVISA: software platform for visualization and analysis of multi-modality brain data. *NeuroImage*, *13*(6), 98. doi:10.1016/S1053-8119(01)91441-7

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 1–11. doi:10.1016/j.jneumeth.2015.01.010

Cusack, R., Vicente-Grabovetsky, A., Mitchell, D. J., Wild, C. J., Auer, T., Linke, A. C., & Peelle, J. E. (2015). Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. *Frontiers in Neuroinformatics*, 8(January), 90. doi:10.3389/fninf.2014.00090

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, 8(2-3), 109–114. doi:10.1002/(SICI)1097-0193(1999)8:2/3<109::AID-HBM7>3.0.CO;2-W

Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., & Evans, A. C. (2012). LORIS: a web-based data management system for multi-center studies. *Frontiers in Neuroinformatics*, 5(January), 1–11. doi:10.3389/fninf.2011.00037

Das, S., Glatard, T., MacIntyre, L. C., Madjar, C., Rogers, C., Rousseau, M.-E., … Evans, A. C. (2016). The MNI data-sharing and processing ecosystem. *NeuroImage*, 124(Pt B), 1188–95. doi:10.1016/j.neuroimage.2015.08.076

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., … Newman, D. (2010). Towards open science: the myExperiment approach. *Concurrency and Computation: Practice and Experience*, *22*(17), 2335–2353. doi:10.1002/cpe.1601

Dinov, I. D., Van Horn, J. D., Lozev, K. M., Magsipoc, R., Petrosyan, P., Liu, Z., … Toga, A. W. (2009). Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline. *Frontiers in Neuroinformatics*, 3(July), 22. doi:10.3389/neuro.11.022.2009

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop 26th International Conference for Machine Learning* (p. 4). Montreal.

Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, 84, 45–64. doi:10.1016/j.neuroimage.2013.07.072

Eklund, A., Nichols, T., Andersson, M., & Knutsson, H. (2015). Empirically investigating the statistical validity of spm, fsl and afni for single subject fmri analysis. In *IEEE 12th International Symposium on Biomedical Imaging* (pp. 1376–1380). doi:10.1109/ISBI.2015.7164132

Gilmore, C. D., Comeau, C. R., Alessi, A. M., Blaine, M., El Fakhri, G. N., Hunt, J. K. E., … Wenzel-Lamb, N. (2013). PET/MR imaging consensus paper: a joint paper by the Society of Nuclear Medicine and Molecular Imaging Technologist Section and the Section for Magnetic Resonance Technologists. *Journal of Nuclear Medicine Technology*, 41(2), 108–113. doi:10.2967/jnmt.113.123869

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., … Evans, A. C. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, *9*(April), 1–14. doi:10.3389/fninf.2015.00012

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson T, Fischl B, Andersson J, Xu J, Jbabdi S, Webster M, Polimeni J, Van Essen DC, Jenkinson M (2013) The minimal preprocessing pipelines for the Human Connectome Projects. *Neuroimage* 80: 105-124. 10.1016/j.neuroimage.2013.04.127

Gopen, G. D., & Swan, J. A. (1990). The Science of Scientific Writing. *The American Scientist*, 550–558.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*. doi:10.3389/fninf.2011.00013

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., 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*(6), e38234. doi:10.1371/journal.pone.0038234

Habibzadeh, P. (2013). Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals. *Applied Clinical Informatics*, *4*, 455–64. doi:10.4338/ACI-2013-07-RA-0055

Halchenko, Y. O., & Hanke, M. (2012). Open is not enough. Let's take the next step: an integrated, community-driven computing platform for neuroscience. *Frontiers in Neuroinformatics*, *6*. doi:10.3389/fninf.2012.00022

Halchenko, Y. O & Hanke, M. (2015). Four aspects to make science open "by design" and not as an after-thought. *GigaScience* 2015, 4:31. doi:10.1186/s13742-015-0072-7

Hastie, T., Tibshirani, R. J., & Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2nd ed.). Springer-Verlag.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi:10.1016/j.neuroimage.2013.10.067

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., … Buckner, R. L. (2015). Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific Data*, 2, 150031. doi:10.1038/sdata.2015.31

Inglis, B. (2015). A checklist for fMRI acquisition methods reporting in the literature. *The Winnower*. doi:10.15200/winn.143191.17127

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. 10.1371/journal.pmed.0020124

ISO. (2006). *Statistics - Vocabulary and symbols. Part 2: Applied statistics.* ISO 3534–2 (Second.). Geneva: ISO.

Jones, D. K., Knösche, T. R., & Turner, R. (2012). White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *NeuroImage*. doi:10.1016/j.neuroimage.2012.06.081

Journals unite for reproducibility. (2014). *Nature*, 515(7525), 7. doi:10.1038/515007a

Joyce, K. E., & Hayasaka, S. (2012). Development of PowerMap: a Software Package for Statistical Power Calculation in Neuroimaging Studies. *Neuroinformatics*, 10(4), 351–65. doi:10.1007/s12021-012-9152-3

Kanal, E., Barkovich, a J., Bell, C., Borgstede, J. P., Bradley, W. G., Froelich, J. W., … Hernandez, D. (2013). ACR guidance document on MR safe practices: 2013. *Journal of Magnetic Resonance Imaging : JMRI*, 37(3), 501–30. doi:10.1002/jmri.24011

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., … Nichols, B. N. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *NeuroImage*, 82, 647–61. doi:10.1016/j.neuroimage.2013.05.094

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217. doi:10.1207/s15327957pspr0203_4

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6), 593–602. doi:10.1001/archpsyc.62.6.593

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, 8(6), e1000412. doi:10.1371/journal.pbio.1000412

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November), 4. doi:10.3389/neuro.06.004.2008

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–40. doi:10.1038/nn.2303

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat* (pp. 575–580). Springer. doi:10.1007/978-3-642-57489-4_89

Le Noury, J., Nardo, J. M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., & Abi-Jaoude, E. (2015). Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of

major depression in adolescence. *BMJ (Clinical Research Ed.),* 351, h4320. doi:10.1136/bmj.h4320

Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. *NeuroImage*, 13(4), 759–773. doi:10.1006/nimg.2000.0728

Luo, W.-L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, 19(3), 1014–1032. doi:10.1016/S1053-8119(03)00149-6

Mackenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., & Toga, A. W. (2008). Provenance in neuroimaging. *NeuroImage*, *42*(1), 178–95. doi:10.1016/j.neuroimage.2008.04.186

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., … Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, 383(9912), 101–104. doi:10.1016/S0140-6736(13)62329-6

MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth. *Nature*, 526, 187–189. doi:10.1038/526187a

Marcus, D. S., Olsen, T. R., Ramaratnam, M., & Buckner, R. L. (2007). The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*, 5(1), 11–34. doi:10.1385/NI

Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., … Van Essen, D. C. (2013). Human Connectome Project informatics: Quality control, database services, and data visualization. *NeuroImage*, 80, 202–219. doi:10.1016/j.neuroimage.2013.05.077

McNutt, M. (2014). Journals unite for reproducibility. *Science*, 346(6210), 679–679. doi:10.1126/science.aaa1724

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., … Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's and Dementia*, 1(1), 55–66. 10.1016/j.jalz.2005.06.003

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39(1), 261–8. doi:10.1016/j.neuroimage.2007.07.061

Mumford, J. a. (2012). A power calculation guide for fMRI studies. *Social Cognitive and Affective Neuroscience*, 7(6), 738–42. doi:10.1093/scan/nss059

Nemoto, K., Dan, I., Rorden, C., Ohnishi, T., Tsuzuki, D., Okamoto, M., … Asada, T. (2011). Lin4Neuro: a customized Linux distribution ready for neuroimaging analysis. *BMC Medical Imaging*, *11*(1), 3. doi:10.1186/1471-2342-11-3

Norman, K. A, Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–30. doi:10.1016/j.tics.2006.07.005

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45(3), 137–141. doi:10.1027/1864-9335/a000192

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, 506(7487), 150–152. doi:10.1038/506150a

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi:10.1126/science.aac4716

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–7. doi:10.1126/science.1213847

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

Petersen, S. E., & Dubis, J. W. (2012). The mixed block/event-related design. *NeuroImage*, 62, 1177–1184. doi:10.1016/j.neuroimage.2011.09.084

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2(1), 67. doi:10.1093/scan/nsm006

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, *40*(2), 409–14. doi:10.1016/j.neuroimage.2007.11.048

Poldrack, R. a, & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17(11), 1510–1517. doi:10.1038/nn.3818

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., … Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6(April), 1–13. doi:10.3389/fninf.2012.00009

Rex, D. E., Ma, J. Q., & Toga, A. W. (2003). The LONI pipeline processing environment. *Neuroimage*, *19*(3), 1033–1048. doi:10.1016/S1053-8119(03)00185-X

Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., … Calhoun, V. D. (2011). COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Frontiers in Neuroinformatics*, 5(December), 33. doi:10.3389/fninf.2011.00033

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–66. doi:10.1177/0956797611417632

Smith, S. M., Bandettini, P. A., Miller, K. L., Behrens, T. E. J., Friston, K. J., David, O., … Nichols, T. E. (2012). The danger of systematic bias in group-level FMRI-lag-based causality estimation. *NeuroImage*, 59(2), 1228–9. doi:10.1016/j.neuroimage.2011.08.015

Stark, D. E., Margulies, D. S., Shehzad, Z. E., Reiss, P., Kelly, a. M. C., Uddin, L. Q., … Milham, M. P. (2008). Regional variation in interhemispheric coordination of intrinsic hemodynamic fluctuations. *The Journal of Neuroscience*, 28(51), 13754–13764. doi:10.1523/JNEUROSCI.4544-08.2008

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., … Siditis, J. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, 15, 747–771. doi:10.1006/nimg.2001.1034

Tu, Y.-K., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon--the reversal paradox. *Emerging Themes in Epidemiology*, 5, 2. doi:10.1186/1742-7622-5-2

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., … Gee, J. C. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*, *99*, 166–179. doi:10.1016/j.neuroimage.2014.05.044

Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., & Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association*, 8(5), 443–459. doi:10.1136/jamia.2001.0080443

van Wijk, B. C. M., Stam, C. J., & Daffertshofer, A. (2010). Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory. *PLoS ONE*, 5(10), e13701. doi:10.1371/journal.pone.0013701

Varoquaux, G., Gramfort, A., & Thirion, B. (2012). Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 1375–1382.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage*, 18(2), 293–309. doi:10.1016/S1053-8119(02)00046-0

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *The Journal of Neuroscience*, *34*(32), 10743–10755. doi:10.1523/JNEUROSCI.5282-13.2014

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 1305(April), 00–00. doi:10.1080/00031305.2016.1154108

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. doi:10.1038/sdata.2016.18

Wicherts, J. M., Bakker, M., & Molenaar, D. (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(11), 1–7. doi:10.1371/journal.pone.0026828

Zang, He, Zhu, Cao, Sui, Liang, … Wang. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain & Development*, 29(2), 83–91. doi:10.1016/j.braindev.2006.07.002

Zang, Y., Jiang, T., Lu, Y., He, Y., & Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *NeuroImage*, 22(1), 394–400. doi:10.1016/j.neuroimage.2003.12.030

Ziegler, E., Rouillard, M., André, E., Coolen, T., Stender, J., Balteau, E., … Garraux, G. (2014). Mapping track density changes in nigrostriatal and extranigral pathways in Parkinson's disease. *NeuroImage*, *99*, 498–508. doi:10.1016/j.neuroimage.2014.06.033

Zou, Q. H., Zhu, C. Z., Yang, Y., Zuo, X. N., Long, X. Y., Cao, Q. J., … Zang, Y. F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *Journal of Neuroscience Methods*, 172(1), 137–141. doi:10.1016/j.jneumeth.2008.04.012

Appendix A. COBIDAS Membership & Acknowledgements

Upon creating COBIDAS in June 2014, Dr. Nichols was named as chair and subsequently invited nominations from the OHBM membership. From over 100 nominees Dr. Nichols selected a dozen experts from the membership that reflected the diversity of OHBM, with the final list approved by Council. The different constituencies considered included: Researchers focusing in cognitive applications, clinical applications, methods and database developers; different geographic areas; sex; representation of junior researchers; and, to facilitate communication within OHBM leadership, at least one member from Council and one member from the OHBM Program Committee.

The full list of members is as follows (in alphabetical order).

Simon Eickhoff, Department of Clinical Neuroscience and Medical Psychology Heinrich-Heine University Düsseldorf, Düsseldorf, Germany.
Alan Evans, Montreal Neurological Institute, McGill University, Montreal, Canada.
Michael Hanke, Otto-von-Guericke-University Magdeburg, Germany.
Nikolaus Kriegeskorte, MRC Cognition and Brain Sciences Unit, Cambridge, UK
Michael Milham, Child Mind Institute, New York City, USA.
Thomas Nichols (chair), University of Warwick, UK.
Russell Poldrack, Stanford University, Stanford, USA.
Jean-Baptiste Poline, University of California, Berkeley, Berkeley, USA.
Erika Proal, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz & Neuroingenia, Mexico City, Mexico.
Bertrand Thirion, Inria, Paris-Saclay University, France.
David Van Essen, Washington University, Division of Biology and Biomedical Sciences, St. Louis, USA.
Tonya White, Erasmus MC, Rotterdam, Netherlands.
BT Thomas Yeo, National University of Singapore, Singapore.

We are deeply grateful to guest members Tristan Glatard & Samir Das, informatics experts who contributed greatly to data sharing and reproducibility sections, and (then) OHBM President Karen Berman and former OHBM president Peter Bandettini, who participated in a number of calls.

We are indebted to Ben Inglis for allowing us to use his fMRI acquisition methods reporting checklist [Inglis2015] as a template for our Acquisition Reporting checklist. We are also grateful to the following for input on different specific sections: Ben Inglis, Doug Noll, Robert Welsh, Jörg Stadler, Derek Jones and Stamatios Sotiropoulos on the Data Acquisition section; Tim Behrens on the Experimental Design section; Vince Calhoun and Christian Beckmann on Statistical Modeling & Inference; Ged Ridgway for Preprocessing and Statistical Modeling & Inference; for help with Ting Xu for Statistical Modeling & Inference.

We would like to thank the members of the brain imaging community who took time to comment on the original draft of this work. Through email comments, marked up manuscripts and blog entries, we received over 100 comments, each of which has been considered and, with only a few exceptions, integrated into the present manuscript. In addition to anonymous participants in the blog, these contributors included:

David Abbott, Carsten Allefeld, John Ashburner, Mary Askren, Peter Bandettini, Andreas Bartsch, Patrick Bellgowan, Saskia Bollmann, Matthew Brown, Karen Davis, Simon Davis, Ekaterina Dobryakova, Simon Eskildsen, Satrajit Ghosh, Matt Glasser, Tristan Glatard, Enrico Glerean, Douglas Greve, Luis Hernandez-Garcia, Emilia Iannilli, Tom Johnstone, Jorge Jovicich, Varsha Khodiyar, Vesa Kiviniemi, Thomas Liu, Martin Lotze, Dan Lurie, David Macfarlane, Clare Mackay, Paul Mazaika, Donald McLaren, Janaina Mourao Miranda, Jeanette Mumford, Helmut Nebl, Cyril Pernet, Aina Puce, Ged Ridgway, Alard Roebroeck, Tom Ross, Keith Schneider, David Soltysik, Stamatios Sotiropoulos, Stephen Strother, Paul Summers, Karsten Tabelow, Jussi Tohka, Robert Turner, Ilya Veer, Tor Wager, Alle Meije Wink, Tal Yarkoni, and Yu-Feng Zang.

Thanks also to Alex Bowring for assistance in final preparation of this manuscript.

Appendix B. Defining Reproducibility

A number of terms with overlapping meaning are used to refer to the merits of scientific findings, including reproducibility, replicability, and reliability. Here we attempt to set the terminology and clarify their meaning as used this report.

Replication is a cornerstone of the scientific method. A replication, where independent researchers use independent data and possibly distinct methods to arrive at the same original conclusion, is the ultimate standard for validating a scientific claim.

Roger Peng [Peng2011] suggested a specific notion of *reproducibility* in the computational sciences. He articulated a kind of reproducibility where independent researchers use the exact same data and code to arrive at the original result. Within this there is a spectrum of reproducibility practice, ranging from a publication sharing only code, or code and data, to the best case, where detailed scripts and code and data are shared that produces the results reported in the paper when executed.

The US Food & Drug Administration also has definitions to describe the precision of measurements, as codified by terms from the International Standards Organization (ISO), “repeatability” and “reproducibility” [ISO2006].

ISO *repeatability* (ISO 3534-2:2006 3.3.5) is defined as precision under “conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in the same test or measuring facility by the same operator using the same equipment within short intervals of time”.

ISO *reproducibility* (ISO 3534-2:2006 3.3.10) is defined as precision under “conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in different test or measurement facilities with different operators using different equipment”

While these definitions are motivated by laboratory use, in a setting where the “test item” is more likely to be a Petri dish culture than a human subject, they still offer a useful sharp definition. In the neuroimaging setting, we find these terms too narrow and unable to capture the range of types of consistency that can be considered. Specifically, not just consistency of the imaging device, or consistency of analysis process, we can consider an expansive concept of consistency, a generalizability of results over different classes of experimental stimuli and context.

In Table A1 we present an incomplete taxonomy of different possible types of consistency, from the ISO repeatability to the widest senses of scientific generalizability. While this taxonomy is

useful to see the multifarious nature of “reproducibility”, in the body of this work we use “reproducibility” (without qualification) to mean Roger Peng’s “computational” notion.

Table B1. Levels of Reproducibility

This table provides an incomplete taxonomy of types of consistency of neuroimaging results. For each type of consistency (row), the variable (column) that is held constant (•, bullet) or allowed to vary (D=different) is indicated. In each instance, a bullet (•) indicates the exact same setting; for the variable "Subject" this means the very same acquired data is used, while a double bullet (••) indicates the same subjects are scanned multiple times. Examples of different Experimental Methods include fundamental changes like an event-related design vs. a block design; examples of different Experimental Code/Stimuli, include different sets of pictures used in a visual working memory experiment, or using different paradigm software; an example of different Stimuli Type would be number vs. shape vs. image stimuli in a working memory experiment. An example of different analysis methods for intrasubject (first level) fMRI data would be a confirmatory regression-based modelling vs. an exploratory data-driven method like independent components analysis; examples of different analysis code would be intrasubject fMRI fit with a regression model in two different software packages.

Type of Stability	**Investigator**	**Experimental Method**	**Stimuli Type**	**Experimental Code/Stimuli**	**Subject Population**	**Subjects**	**Scanner**	**Acquistion Day**	**Data Analyst**	**Analysis Method**	**Analysis Code**
Measurement Stability											
ISO Repeatability e.g. 30-minute intra-scanner reliability	•	•	•	•	•	••	•	•	•	•	•
ISO Intermediate Reproducibility e.g. 7-day intra-scanner reliability	•	•	•	•	•	••	•	D	•	•	•
ISO Reproducibility e.g. 7-day inter-scanner reliability	•	•	•	•	•	••	D	D	•	•	•
Analytical Stability											
Analysis Replicability	•	•	•	•	•	•	•	•	•	•	•
Collegial Analysis Replicability	•	•	•	•	•	•	•	•	D	•	•
Peng's Reproducibility	D	•	•	•	•	•	•	•	D	•	•
Generalisabilty over ...											
Subjects (Near Replicability)	•	•	•	•	•	D	•	D	•	•	•
Labs (Intermediate Replicability)	D	•	•	D	•	D	D	D	D	•	D
Experimental & Analytical methods (Far Replicability)	D	D	•	D	•	D	D	D	D	D	D
Subject & Stimulus Populations (Hypothesis Generalisability)	D	D	D	D	D	D	D	D	D	D	D

Appendix C. Short descriptions of fMRI models

While any analysis software consists of myriad modelling decisions, an author must be able to describe the key facets of an analysis in the methods section of their paper. To facilitate this, and to suggest a level of detail that is useful to readers unfamiliar with the software yet not distractingly long, we provide short descriptions for the most commonly used statistical models in widely used software packages.

C1. Task fMRI

Summaries for AFNI[44], Freesurfer[45], FSL[46], & SPM[47] are based on versions AFNI_2011_12_21_1014, FreeSurfer 5.3, FSL 5.0.8 and SPM 12 revision 6470, respectively.

AFNI 1st level – 3dDeconvolve: Linear regression at each voxel, using ordinary least squares, drift fit with polynomial.
AFNI 1st level – 3dREMLfit: Linear regression at each voxel, using generalised least squares with a voxel-wise ARMA(1,1) autocorrelation model, drift fit with polynomial.
AFNI 2nd level – 3dTtest: Linear regression at each voxel, using ordinary least squares.
AFNI 2nd level – 3dMEMA: Linear mixed effects regression at each voxel, using generalized least squares with a local estimate of random effects variance.
AFNI 2nd level – 3dMVM: Multivariate ANOVA or ANCOVA at each voxel.
AFNI 2nd level – 3dLME: General linear mixed-effects modeling at each voxel, with separate specification of fixed and random variables.
Freesurfer 1st Level – selxavg3-sess: Linear regression at each surface element, using generalized least squares with a element-wise AR(1) autocorrelation model, drift fit with polynomial.
Freesurfer 2st Level – mri_glmfit: Linear regression at each surface element, using ordinary least squares.
FSL 1st level: Linear regression at each voxel, using generalized least squares with a voxel-wise, temporally and spatially regularized autocorrelation model, drift fit with Gaussian-weighted running line smoother (100s FWHM).
FSL 2nd level – "OLS": Linear regression at each voxel, using ordinary least squares.
FSL 2nd level – "FLAME1": Linear mixed effects regression at each voxel, using generalized least squares with a local estimate of random effects variance.
SPM 1st level: Linear regression at each voxel, using generalized least squares with a global approximate AR(1) autocorrelation model, drift fit with Discrete Cosine Transform basis (128s cut-off).

[44] http://afni.nimh.nih.gov.
[45] https://surfer.nmr.mgh.harvard.edu.
[46] http://fsl.fmrib.ox.ac.uk.
[47] http://www.fil.ion.ucl.ac.uk/spm.

SPM 2nd level – no repeated measures: Linear regression at each voxel, using ordinary least squares.
SPM 2nd level – repeated measures: Linear regression at each voxel, using generalized least squares with a global repeated measures correlation model.

C2. Single-Modality ICA

Methods for ICA analyses are not as consolidated as mass univariate linear modelling, but we provide short summaries of some typical analyses in GIFT[48] and MELODIC[49] (alphabetical order), based on versions GIFTv3.0a and FSL 5.0.8, respectively. [Optional aspects, depending on particular variants used, indicated in brackets.]

GIFT, single-subject fMRI with ICASSO stability: Spatial ICA estimated with infomax where scaling of original data, spatial components and time courses constrained to unit norm, resulting best-run selected from 10 runs; post-ICA Z statistics produced for maps, between temporal component correlation (Functional Network Correlation), time courses, spectra, tested within a GLM framework.
GIFT, multi-subject PCA-based back-reconstruction with ICASSO stability: Single-subject PCA followed by temporal concatenation, group-level PCA and then spatial ICA with infomax; calculation of single subject maps using PCA-based back-reconstruction, resulting best-run selected from 10 runs; post-ICA Z statistics produced for maps, time courses, spectra, and between temporal component correlation (Functional Network Correlation) tested within a GLM framework. [Time-varying states computed using moving window between temporal components (Dynamic Functional Network Correlation).]
GIFT, spatio-temporal (dual) regression of new data: Using provided component maps calculates per-subject components from new data using regression-based back-reconstruction; produces component maps, time courses and spectra and between temporal component correlation (Functional Network Correlation) tested within a GLM framework.
GIFT, spatial ICA with reference: Spatial ICA using one or more provided seed or component maps. Components found by joint maximization of non-Gaussianity and similarity to spatial maps resulting in subject specific component maps and timecourses for each subject, scaled to Z-scores, following by testing voxelwise (within network connectivity), between temporal component correlation (Functional Network Correlation), spectra, tested within a GLM framework.
GIFT, source based morphometry of gray matter maps: Spatial ICA of multi-subject gray matter segmentation maps (from SPM, FSL, etc) resulting in spatial components and subject-loading parameters tested within a GLM framework.
MELODIC, single-subject ICA: Spatial ICA estimated by maximising non-Gaussian sources, using robust voxel-wise variance-normalisation of data, automatic model-order selection and Gaussian/Gamma mixture-model based inference on component maps.

[48] http://mialab.mrn.org/software/gift.
[49] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC.

MELODIC, group level (concat ICA): Temporally concatenation of fMRI data, followed by spatial ICA estimated by maximising non-Gaussian sources, using using robust voxel-wise variance-normalisation of data, automatic model order selection and Gaussian/Gamma mixture-model based inference on component maps
MELODIC, group-level (tensor-ICA): Higher-dimensional decomposition of all fMRI data sets into spatial, temporal and subject modes; automatic model order selection and Gaussian/Gamma mixture-model based inference on component maps
MELODIC dual regression: Estimation of subject-specific temporal and spatial modes from group-level ICA maps or template maps using spatial followed by temporal regression.

C3. Multi-Modalitiy ICA

Available multi-modality ICA methods include FIT[50] and FSL-FLICA[51] (alphabetical order), based on versions FITv2.0c and flica_2013-01-15, respectively.

FIT, joint ICA, two-group, fMRI + EEG fusion: Joint spatial ICA of GLM contrast maps and temporal ICA of single or multi-electrode event-related potential time course data (can be non-concurrent) with infomax ICA; produces joint component maps (each with an fMRI component map and ERP component timecourse(s)) and subject loading parameters which are then tested for group differences with a GLM framework.
FIT, N-way fusion using multiset CCA+joint ICA: Multiset canonical correlation analysis applied to several spatial maps to extract components, then submitted to spatial ICA with infomax ICA; produces multi-modal component maps and subject-specific loading parameters which are tested within a GLM framework.
FIT, parallel ICA, fusion of gray matter maps and genetic polymorphism array data: Joint spatial ICA of gray matter segmentation maps and genetic ICA of single nucleotide polymorphism data performed through a maximization of independence among gray matter components, genetic components, and subject-wise correlation among one or more gray matter and genetic components. Produces linked and unlinked gray matter and genetic components and subject loading parameters which are then tested within a GLM framework.
FSL-FLICA multi-subject/multi-modality (Linked-ICA): ICA-based estimation of common components across multiple image modalities, linked through a shared subject-courses.

[50] http://mialab.mrn.org/software/fit.
[51] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLICA.

Appendix D. Itemized lists of best practices and reporting items

This section contains checklists for practices and items to report. Each item has been included because it is an essential piece of information needed to understand, evaluate and reproduce an experiment. Authors should strive to include all these items, but items marked as “Mandatory” are particularly crucial, and a published work cannot be considered complete without such information.

The rest of this page intentionally left blank.

Table D.1. Experimental Design Reporting

Aspect	Notes	Mandatory
Number of subjects	*Elaborate each by group if have more than one group.*	
Subjects approached		N
Subjects consented		N
Subjects refused to participate	Provide reasons.	N
Subjects excluded	Subjects excluded after consenting but before data acquisition; provide reasons.	N
Subjects participated and analyzed	Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.	Y
Inclusion criteria and descriptive statistics	*Elaborate each by group if have more than one group.*	
Age	Mean, standard deviation and range.	Y
Sex	Absolute counts or relative frequencies.	Y
Race & ethnicity	Per guidelines of NIH or other relevant agency.	N
Education, SES	Education is essential for studies comparing patient and control groups; complete SES reporting less important for single-group studies, but still useful. Specify measurement instrument used; may be parental SES and education if study has minors.	Y
IQ	Specify measurement instrument used.	N

Handedness	Absolute or relative frequencies; basis of handedness-attribution (self-report, EHI, other tests). (Important for fMRI, may be less important for structural studies.)	Y
Exclusion criteria	Describe any screening criteria, including those applied to “normal” sample such as MRI exclusion criteria.	Y
Clinical criteria	Detail the area of recruitment (in- vs. outpatient setting, community hospital vs. tertiary referral center etc.) as well as whether patients were currently in treatment.	Y
Clinical instruments	Describe the instruments used to obtain the diagnosis and provide tests of intra- or inter-rater reliability. Clarify whether a “clinical diagnosis” or “inventory diagnosis” was used (if applicable). State the diagnostic system (ICD, DSM etc) that was used.	Y
Matching strategy	If applicable.	Y
Population & recruitment strategy	Population from which subjects were drawn, and how and where recruitment took place, e.g., schools, clinics, etc. If possible, note if subjects are research-naive or have participated in other studies before.	Y
Subject scanning order	With multiple groups, information on ordering and or balance over time; especially report relative to scanner changes/upgrades. (Ideally, use randomized or interleaved order to avoid bias due to scanner changes/upgrades.)	Y
Neurocognitive measures	All measures collected on subjects should be described and reported.	Y
Ethical considerations		
Ethical approval	Describe approval given, including the particular institutional review board, medical ethics committee or equivalent that granted the approval. When data is shared, describe the ethics/institutional approvals required from either the author (source) or recipient.	Y
Informed consent	Record whether subjects provided informed consent or, if applicable, informed assent.	Y

Design specifications		
Design type	Task or resting state. Event-related or block design. (See body text for usage of 'block design' terminology.)	Y
Condition & stimuli	Clearly describe each condition and the stimuli used. Be sure to completely describe baseline (e.g. blank white/black screen, presence of fixation cross, or any other text), especially for resting-state studies. When possible provide images or screen snapshots of the stimuli.	Y
Number of blocks, trials or experimental units	Specify per session, and if differing by subject, summary statistics (mean, range and/or standard deviation) of such counts.	Y
Timing and duration	Length of each trial or block (both, if trials are blocked), and interval between trials. Provide the timing structure of the events in the task, whether a random/jittered pattern or a regular arrangement; any jittering of block onsets.	Y
Length of the experiment	Describe the total length of the scanning session, as well as the duration of each run. (Important to assess subject fatigue.)	Y
Design optimization	Whether design was optimized for efficiency, and how.	Y
Presentation software	Name software, version and operating system on which the stimulus presentation was run. When possible, provide code used to drive experiment.	Y
Task specification		
Condition	Enumerate the conditions and fully describe and reference each. Consider using a shorthand name, e.g. AUDSTIM, VISSTIM, to refer to each condition, to clarify the distinction between a specific modeled effect and a psychological construct. Naming should reflect the distinction between instruction periods and actual stimuli, and between single parameters and contrasts of parameters.	Y
Instructions	Specify the instructions given to subjects for each condition (ideally the exact text in supplement or appendix). For resting-state, be sure to indicate eyes-closed,	Y

	eyes-open, any fixation. Describe if the subjects received any rewards during the task, and state if there was a familiarization / training inside or outside the scanner.	
Stimuli	Specifics of stimuli used in each run. For example, the unique number of stimuli used, and whether/how stimuli were repeated over trials or conditions.	Y
Randomization	Describe block or event ordering as deterministic, or report manner of randomization, in terms of order and timing. If pseudo-randomized, i.e. under constraints, describe how and the criteria used to constrain the orders/timings.	Y
Stimulus presentation & response collection.	Specify the presentation hardware (e.g. back projection, in-room display, goggles, etc), and the response systems (e.g. button boxes, eye tracking, physiology). Note how equipment was synched to the scanner (e.g. scanner TTL, or manual sync.)	Y
Run order	Order in which tasks runs are conducted in the scanner.	Y
Power analysis		
Outcome	Specify the type of outcome used as the basis of power computations, e.g. signal in a pre-specified ROI, or whole image voxelwise (or cluster-wise, peak-wise, etc.).	Y
Power parameters	Specify ● Effect size (or effect magnitude and standard deviation separately). ● Source of predicted effect size (previous literature with citation; pilot data with description, etc). ● Significance level (e.g. uncorrected alpha 0.05 for an ROI, or FWE-corrected significance ● Target power (typically 80%). ● Any other parameters set (e.g., for spatial methods a brain volume and smoothness may be needed to be specified).	Y
Behavioral performance		

Variables recorded	State number of type of variables recorded (e.g. correct button press, response time).	Y
Summary statistics	Summaries of behavior sufficient to establish that subjects were performing the task as expected. For example, correct response rates and/or response times, summarized over subjects (e.g. mean, range and/or standard deviation).	Y

Table D.2. Acquisition Reporting

Aspect	**Notes**	**Mandatory**
Subject preparation		
Mock scanning	Use of an MRI simulator to acclimate subjects to scanner environment. Report type of mock scanner and protocol (i.e. duration, types of simulated scans, experiments).	N
Special accommodations	For example, for pediatric scanning, presence of parent/guardian in the room.	Y
Experimenter personnel	Whether a single or multiple experimenters interacted with the subjects.	N
MRI system description		
Scanner	Provide make, model & field strength in tesla (T).	Y
Coil	Receive coil (e.g. “a 12-channel phased array coil”, but more details for a custom coil) and (if nonstandard) transmit coil. Additional information on the gradient system, e.g. gradient strength (if non-standard for the make and model, or switchable).	Y
Significant hardware modifications	For example, special gradient inserts/sets.	N

Software version	Highly recommended when sharing vendor-specific protocols or exam cards, as version may be needed to correctly interpret that information.	N
MRI acquisition		
Pulse sequence type	For example, gradient echo, spin echo, etc.	Y
Imaging type	For example, echo planar imaging (EPI), spiral, 3D. Number of shots (if multi-shot); partial Fourier scheme & reconstruction method (if used);	Y

Essential sequence & imaging parameters.	For all acquisitions:
	● Echo time (TE).
	● Repetition time (TR).
	  o For multi-shot acquisitions, additionally the time per volume.
	● Flip angle (FA).
	● Acquisition time (duration of acquisition).
	Functional MRI:
	● Number of volumes.
	● Sparse sampling delay (delay in TR) if used.
	Inversion recovery sequences:
	● Inversion time (TI).
	B0 field maps:
	● Echo time difference (dTE).
	Diffusion MRI:
	● Number of directions.
	  o Direction optimization, if used and type.
	● b-values.
	● Number of b=0 images.
	● Number of averages (if any).
	● Single shell, multi-shell (specify equal or unequal spacing).
	● Single- or dual-spin-echo, gradient mode (serial or parallel).
	● If cardiac gating used.
	Imaging parameters:
		Y

	● Field of view.
	● In-plane matrix size, slice thickness and interslice gap, for 2D acquisitions.
	● Slice orientation:
	  ○ Axial, sagittal, coronal or oblique.
	  ○ Angulation: If acquistion not aligned with scanner axes, specify angulation to AC-PC line (see Slice position procedure).
	● 3D matrix size, for 3D acquisitions.	

Phase encoding	Specify phase encoding direction (e.g. as A/P, L/R, or S/I). For 3D, specify “partition encode” (aka slice) direction. Phase encoding reversal: Mention if used (aka “blip-up/blip-down”).	Y
Parallel imaging method & parameters	Report: ● Method, e.g. SENSE, GRAPPA or other parallel imaging method, and acceleration factor. ● Matrix coil mode, and coil combining method (if non-standard).	Y
Multiband parameters	Multiband factor and field-of-view shift (only if applicable).	Y
Readout parameters	Receiver bandwidth, readout duration, echo spacing.	N
Fat suppression	For anatomical scans, whether it was used or not.	Y
Shimming	Any specialized shimming procedures.	Y
Slice order & timing	For fMRI acquisitions, interleaved vs. sequential ordering and direction (ascending/descending), location of 1st slice; any specialized slice timing.	Y
Slice position procedure	For example, landmark guided vs. auto-alignment.	N
Brain coverage	Report whether coverage was whole-brain, and whether cerebellum and brainstem were included. If not whole-brain, note the nature of the partial area of coverage. If axial and co-planar with AC-PC line, the volume coverage in terms of Z in mm.	Y

Scanner-side preprocessing	Including: ● Reconstruction matrix size differing from acquisition matrix size. ● Prospective-motion correction (including details of any optical tracking, and how motion parameters are used). ● Signal inhomogeneity correction. ● Distortion-correction.	Y
Scan duration	In seconds	N
Other non-standard procedures	Including: ● Turning off the cold head(s) (e.g. during EEG/fMRI or spectroscopy measurements). ● Reduce sound pressure by limiting the gradient slew rate.	N
T1 stabilization	Number of initial “dummy” scans acquired and then discarded by the scanner.	Y
Diffusion MRI gradient table	Also referred to as the b-matrix (but not to be confused with the 3×3 matrix that describes diffusion weighting for a single diffusion weighted measurement).	N
Perfusion: Arterial Spin Labelling MRI	● ASL Labelling method (e.g. continuous ASL (CASL), pseudo-continuous ASL (PCASL), Pulsed ALS (PASL), velocity selective ASL (VSASL)). ● Use of background suppression pulses and their timing. ● For either PCASL or CASL report: ○ Label Duration. ○ Post-labeling delay (PLD). ○ Location of the labeling plane. ● For PCASL also report: ○ Average labeling gradient. ○ Slice-selective labeling gradient. ○ Flip angle of B1 pulses. ○ Assessment of inversion efficiency; QC used to ensure off-resonance artifacts not problematic, signal obtained over whole brain. ● For CASL also report:	Y

	○ Use of a separate labeling coil. ○ Control scan/pulse used. ○ B1 amplitude. ● For PASL report ○ TI. ○ Labeling slab thickness. ○ Use of QUIPSS pulses and their timing. ● For VSASL ○ TI. ○ Choice of velocity selection cutoff (“VENC”).	
Perfusion: Dynamic Susceptibility Contrast MRI	Specify: ● Number of baseline volumes. ● Type, name and manufacturer of intravenous bolus (e.g. gadobutrol, Gadavist, Bayer). ● Bolus amount and concentration (e.g. 0.1 ml/kg and 0.1 mmol/kg). ● Injection rate (e.g. 5 ml/s). ● Post-injection of saline (e.g. 20 ml). ● Injection method (e.g. power injector).	Y
Preliminary quality control		
Motion monitoring	For functional or diffusion acquisitions, any visual or quantitative checks for severe motion; likewise, for structural images, checks on motion or general image quality.	Y
Incidental findings	Protocol for review of any incidental findings, and how they are handled in particular with respect to possible exclusion of a subject’s data.	N

Table D.3. Preprocessing Reporting

Aspect	Notes	Mandatory
Software	For each software used, be sure to include version and revision number.	Y
Software citation	Include URL and Research Resource Identifier for each software used.	N
T1 stabilization	Number of initial “dummy” scans discarded as part of preprocessing (if not already performed by scanner).	Y
Brain extraction	If performed, report: ● Name of software/method (e.g., BET, recon-all in FreeSurfer, etc). ● Parameter choices (e.g. BET’s fractional intensity threshold). ● Any manual editing applied to the brain masks.	Y
Segmentation	For structural images, method used to extract gray, white, CSF and other tissue classes.	Y
Slice time correction	If performed, report: ● Name of software/method. ● Whether performed after or before motion correction. ● Reference slice. ● Interpolation type and order (e.g., 3rd order spline or sinc).	Y
Motion correction	Report: ● Name of software/method. ● Use of non-rigid registration, and if so the type of transformation. ● Use of motion susceptibility correction (fieldmap-based unwarping), as well as the particular software/method. ● Reference scan (e.g. 1st scan or middle scan). ● Image similarity metric (e.g. normalized correlation, mutual information, etc).	Y

	● Interpolation type (e.g., spline, sinc), and whether image transformations are combined to allow a single interpolation. ● Use of any slice-to-volume registration methods, or integrated with slice time correction.	
Gradient distortion correction	(If not already described as part of motion susceptibility correction.)	Y
Diffusion MRI eddy current correction	Report: ● Name of software/method, and if integrated with motion correction ● Image similarity / cost function. ● Type of transformation (e.g. rigid body, affine) and whether constrained only along the phase encode direction. ● Note if gradient table (b-matrix) is then re-oriented. ● Volumetric change applied for eddy current along the phase-encode axis (by the Jacobian determinant).	Y

Diffusion estimation	For all methods, report
	● Model, parameterisation and number of free parameters.
	● Estimation method.
	● Outlier handling approach.
	● Some evidence of fit quality; e.g sample of slices of diffusion weighted data, or residual maps.
	Items to note for particular approaches:
	● Tensor or Kurtosis.
	  ○ Any parameter constraints, like cylindrical symmetry.
	● Multi-compartmental models.
	  ○ Compartments of the model.
	● Orientation distribution function.
	  ○ Parametric (model) or nonparametric (basis function) model.
	  ○ Whether orientation distribution function or fibre orientation density is reported.
	  ○ For spherical deconvolution, note how the canonical fibre response function is derived (e.g. from the data themselves, or simulated data).
		Y

Diffusion processing	Report: ● Summary measures computed (FA, MD, AD, RD, MK, AK, RK, etc.). ● Whether a track based or voxel-wise method is used. ● Threshold used to define analysis voxels. ● Use of population reference track atlas vs. custom atlas (specify set of subjects used to create atlas). ● Standard deviation map (across subjects).	N
Diffusion tractography	Report: ● Name of software/method. ● Step size, turning angle and stopping criteria. ● For ROI based analysis, definition of ROIs (e.g. specify the images used to draw ROIs; manual, semi-automatic or automatic definition of ROIs). ● For tracking, note step-size, turning angle, any anatomical constraints imposed, and stopping criteria. ● If a measure of path probability / “connectivity” is extracted, clearly define this measure.	Y
Perfusion: Arterial Spin Labeling	Report modelling/post-processing scheme: ● For subtraction, specify whether simple subtraction, running, sinc-subtraction, etc. ● For quantitative model, specify model used, number of free parameters.	Y
Perfusion: Dynamic Susceptibility Contrast MRI	● How concentration time curves are calculated, e.g. use of T1 corrections (if short TR) or corrections for leakage. ● Selection of arterial input function (e.g. manual or automatic with reference to method). ● Deconvolution method (kinetic model) to estimate residue function (e.g. SVD or parametric model). ● Details of parameter calculations (e.g. CBF, CBV, MTT, TTP, Tmax).	Y
Function-structure (intra-subject) coregistration	Report: ● Name of software/method. ● Type of transformation (rigid, nonlinear); if nonlinear, type of transformation	Y

	● Cost function (e.g., correlation ratio, mutual information, boundary-based registration, etc). ● Interpolation method (e.g., spline, linear). Note this step might not be necessary if direct T2* to a functional template registration is used.	
Distortion correction	Use of any distortion correction due to field or gradient nonlinearity.	Y
Intersubject registration	Report: ● Name of software/method (e.g., FSL flirt followed by fnirt, FreeSurfer, Caret, Workbench, etc) ● Whether volume and/or surface based registration is used (if not already clearly implied). ● Image types registered (e.g. T2* or T1). ● Any preprocessing to images; e.g. for T1, bias field correction, or segmentation of gray matter; for T2*, single image (specify image) or mean image. ● Template space (e.g., MNI, Talairach, fsaverage, FS_LR), modality (e.g., T1, T2*), resolution (e.g., 2mm, fsaverage5, 32k_FS_LR), and the specific name of template image used; note the domain of the template if not whole brain, i.e. cortical surface only, cerebellum only, CIFTI 'grayordinates' (cortical surface vertices + subcortical gray matter voxels), etc. ● Additional template transformation for reporting; e.g., if using a template in MNI space, but reporting coordinates in Talairach, clearly note and report method used (e.g., Brett's mni2tal, Lancaster's icbm_spm2tal). ● Choice of warp (rigid, nonlinear); if nonlinear, transformation type (e.g., B-splines, stationary velocity field, momentum, non-parametric displacement field); if a parametric transformation is used, report resolution, e.g., 10x10x10 spline control points. ● Use of regularization, and the parameter(s) used to set degree of regularization.	Y

	● Interpolation type (e.g., spline, linear); if projection from volume to surface space, how were voxels sampled from the volume (e.g., trilinear; nearest neighbor; ribbon-constrained specifying inner and outer surface used). ● Cost function (e.g., correlation ratio, mutual information, SSD). ● Use of cost-function masking.	
Intensity correction	Bias field corrections for structural MRI, but also correction of odd versus even slice intensity differences attributable to interleaved EPI acquisition without gaps.	Y
Intensity normalization	Scan-by-scan or run-wide scaling of image intensities before statistical modelling. E.g. SPM scales each run such that the mean image will have mean intracerebral intensity of 100; FSL scales each run such that the mean image will have an intracerebral mode of 10,000.	N
Artifact and structured noise removal	Use of physiological noise correction method. Report: ● Name of software/method used (e.g. CompCor, ICA-FIX, ICA-AROMA, etc.). ● If using a nuisance regression method, specify regressors used; for each type, include key details, as follows: ○ Motion parameters. ■ Expansion basis and order (e.g. 1st temporal derivatives; Volterra kernel expansion) ○ Tissue signals. ■ Tissue type (e.g., whole brain, gray matter, white matter, ventricles). ■ Tissue definition (e.g., a priori seed, automatic segmentation, spatial regression). ■ Signal definition (e.g., mean of voxels, first singular vector, etc.). ○ Physiological signals ■ e.g., heart rate variability, respiration.	Y

	■ Modeling choices (e.g. RETROICOR, cardiac and/or respiratory response functions) and number of computed regressors.	
Volume censoring	Remediation of problem scans, also known as “scrubbing” or “de-spiking”. Report: ● Name of software/method. ● Criteria (e.g., frame-by-frame displacement threshold, percentage BOLD change). ● Use of censoring or interpolation; if interpolation, method used (e.g., spline, spectral estimation).	Y
Resting state fMRI feature	Creation of summary measure like ALFF, fALFF, ReHo. For ALFF, fALFF report: ● Lower and upper band pass frequencies. For ReHo, report: ● Neighborhood size used to compute local similarity measures (e.g. 6, 18 or 26). ● Similarity measure (e.g. Kendall’s coefficient of concordance).	Y
Spatial smoothing	If this preprocessing step is performed, report: ● Name of software/method. ● Size and type of smoothing kernel. ● Filtering approach, e.g., fixed kernel or iterative smoothing until fixed FWHM. ● Space in which smoothing is performed (i.e. native volume, native surface, MNI volume, template surface).	Y
Quality control reports	Summaries of subject motion (e.g. mean framewise displacement), image variance (e.g. DVARS), and note of any other irregularities found (e.g. motion or poor SNR not sufficiently severe to warrant exclusion). Should be included with any publically shared data.	N

Table D.4. Statistical Modeling & Inference

Aspect	Notes/Ontology	Mandatory
Mass univariate analyses		
Dependent variable: Data submitted to statistical modeling	Report the number of time points, number of subjects; specify exclusions of time points / subjects, if not already specified in experimental design.	Y
Dependent variable: Spatial region modeled	If not “Full brain”, give a specification of an anatomically or functionally defined mask.	Y

Independent variables	For first level fMRI, specify:
	● Event-related design predictors.
	  ○ Modeled duration, if other than zero.
	  ○ Parametric modulation.
	● Block Design predictors.
	  ○ Note whether baseline was explicitly modeled.
	● HRF basis, typically one of:
	  ○ Canonical only.
	  ○ Canonical plus temporal derivative.
	  ○ Canonical plus temporal and dispersion derivative.
	  ○ Smooth basis (e.g. SPM “informed” or Fourier basis; FSL’s FLOBS).
	  ○ Finite Impulse Response model.
	● Drift regressors (e.g. DCT basis in SPM, with specified cut-off).
	● Movement regressors; specify if squares and/or temporal derivative used.
	● Any other nuisance regressors, and whether they were entered as interactions (e.g. with a task effect in 1st level fMRI, or with group effect).
	● Any orthogonalization of regressors, and set of other regressors used to orthogonalize against.
	For second level fMRI or general group model, specify:
	● Group effects (patients vs. controls).
		Y

	● Clearly state whether or not covariates are split by group (i.e. fit as a group-by-covariate interaction). ● Other between subject effects (age, sex; for VBM, total GM or ICV). For group model with repeated measures, specify: ● How condition effects are modeled (e.g. as factors, or as linear trends). ● Whether subject effects are modeled (i.e. as regressors, as opposed to with a covariance structure).	
Model type	Some suggested terms include: ● “Mass Univariate”. ● “Multivariate” (e.g. ICA on whole brain data). ● “Mass Multivariate” (e.g. MANOVA on diffusion or morphometry tensor data). ● “Local Multivariate” (e.g. “searchlight”). ● “Multivariate, intra-subject predictive” (e.g. classify individual trials in event-related fMRI). ● “Multivariate inter-subject predictive” (e.g. classify subjects as patient vs. control). ● “Representational Similarity Analysis”.	Y

Model settings	The essential details of the model. For mass-univariate, first level fMRI, these include:
	● Drift model, if not already specified as a dependent variable (e.g. locally linear detrending of data & regressors, as in FSL).
	● Autocorrelation model (e.g. global approximate AR(1) in SPM; locally regularized autocorrelation function in FSL).
	
	For mass-univariate second level fMRI these include:
	● Fixed effects (all subjects’ data in one model).
	● Random or mixed-effects model, implemented with:
	  ○ Ordinary least squares (OLS, aka unweighted summary statistics approach; SPM default, FSL FEAT’s “Simple OLS”).
	  ○ weighted least squares (i.e. FSL FEAT’s “FLAME 1”), using voxel-wise estimate of between subject variance.
		Y

	○ Global weighted least squares (i.e. SPM's MFX). With any group (multi-subject) model, indicate any specific variance structure, e.g. ● Un-equal variance between groups (and if globally pooled, as in SPM). ● If repeated measures, the specific covariance structure assumed (e.g. compound symmetric, or arbitrary; if globally pooled). For local-multivariate report: ● The number of voxels in the local model. ● Local model used (e.g. Canonical Correlation Analysis) with any constraints (e.g. positive weights only).	
Inference: Contrast/effect	● Specification of the precise effect tested, often as a linear contrast of parameters in a model. When possible, define these in terms of the task or stimulus conditions instead of psychological concepts (See *Task Specification* in *Experimental Design Reporting*). ● Provide tables/figures on main effects (e.g. in supplement), not just differences or interactions. For example, an inference on a difference of two fMRI conditions, A-B, doesn't indicate if both A & B induced positive changes; likewise, to fully interpret an interaction requires knowledge of the main effects. ● Indicate any use of any omnibus ANOVA tests. ● All contrasts explored as part of the research should be fully described in the methods section, whether or not they are considered in the results. ● If performing a two-sided test via two one-sided tests, double the one-sided p-values to convert them into two-sided p-values. For example, if looking at both a contrast [-1 1] and [1 -1] together, each with cluster-forming threshold p=0.001, double the FWE cluster p-values from each contrast to obtain two-sided inferences.	Y
Inference: Search region	● Whole brain or "small volume"; carefully describe any small volume correction used for each contrast. ● If a small-volume correction mask is defined anatomically, provide named anatomical regions from a publicly available ROI atlas.	Y

	● If small-volume correction mask is functionally defined, clearly describe the functional task and identify any risk of circularity. ● All small-volume corrections should be fully described in the methods section, not just mentioned in passing in the results.	

Inference: Statistic type	Typically one of:
	● Voxel-wise (aka peak-wise in SPM).
	● Cluster-wise.
	  ○ Cluster size.
	  ○ Cluster mass.
	  ○ Threshold-free Cluster Enhancement (TFCE).
	For cluster size or mass, report:
	● Cluster-forming threshold.
	For all cluster-wise methods, report:
	● Neighborhood size used to form clusters (e.g. 6, 18 or 26).
	For TFCE, report:
	● Use of non-default TFCE parameters.
		Y

Inference: P-value computation	Report if anything but standard parametric inference used to obtain (uncorrected) P-values. If nonparametric method was used, report method (e.g. permutation or bootstrap) and number of permutations/samples used.	Y

Inference: Multiple testing correction	For mass-univariate, specify the type of correction and how it is obtained, especially if not the typical usage. Usually one of:
	● Familywise Error.
	  ○ Random Field Theory (typical).
	  ○ Permutation.
	  ○ Monte Carlo.
	  ○ Bonferroni.
	● False Discovery Rate.
	  ○ Benjamini & Hochberg FDR (typical).
	  ○ Positive FDR.
	  ○ Local FDR.
	  ○ Cluster-level FDR.
		Y

	● None/Uncorrected. If permutation or Monte Carlo, report the number of permutations/samples. If Monte Carlo, note the brain mask and smoothness used, and how smoothness was estimated.	
Functional connectivity		
Confound adjustment & filtering	Report: ● Method for detecting movement artifacts, movement-related variation, and remediation (e.g. 'scrubbing', 'despiking', etc). ● Use of global signal regression, exact type of global signal used and how it was computed. ● Whether a high- or low-pass temporal filtering is applied to data, and at which point in the analysis pipeline. Note, any temporal regression model using filtered data should have it's regressors likewise filtered.	Y
Multivariate method: Independent Component Analysis	Report: ● Algorithm to estimate components. ● Number of components (if fixed), or algorithm for estimating number of components. ● If used, method to synthesize multiple runs. ● Sorting method of IC's, if any. ● Detailed description of how components were chosen for further analysis.	Y
Dependent variable definition	For seed-based analyses report: ● Definition of the seed region(s). ● Rationale for choosing these regions. For region-based analyses report: ● Number of ROIs. ● How the ROI's are defined (e.g. citable anatomical atlas; auxiliary fMRI experiments); note if ROIs overlap. ● Assignment of signals to regions (i.e. how a time series is obtained from each region, e.g. averaging or first singular vector) ● Note if considering only bilateral (L+R) merged regions.	Y

	● Note if considering only interhemispheric homotopic connectivity.	
Functional connectivity measure/ model	Report: ● Measure of dependence used, e.g. Pearson's (full) correlation, partial correlation, mutual information, etc; also specify: ○ Use of Fisher's Z-transform (Yes/No) and, if standardised, effective N is used to compute standard error (to account for any filtering operations on the data). ○ Estimator used for partial correlation. ○ Estimator used for mutual information. ● Regression model used to remove confounding effects (Pearson or partial correlation).	Y
Effectivity connectivity	Report: ● Model. ● Algorithm used to fit model. ● If per-subject model, method used to generalize inferences to population. ● Itemize models considered, and method used for model comparison.	Y
Graph analysis	Report the 'dependent variable' and 'functional connectivity measure' used (see above). Specify either: ● Weighted graph analysis or, ● Binarized graph analysis is used, clarifying the method used for thresholding (e.g. a 10% density threshold, or a statistically-defined threshold); consider the sensitivity of your findings to the particular choice of threshold used. Itemise the graph summaries used (e.g. clustering coefficient, efficiency, etc), whether these are global or per-node/per-edge summaries. In particular with fMRI or EEG, clarify if measures applied to individual subject networks or group networks.	Y
Multivariate modelling & predictive analysis		

Independent variables	Specify:
	● Variable type (discrete or continuous).
	● Class proportions in classification settings.
	● Variable dimension.
	  o For whole-brain prediction, this is a voxel count.
	  o For searchlight analyses, the exact number of voxels in the search region, not just a radius.
	  o Provide dimension before and after any feature selection and/or dimension reduction.
	If available, report on population stratification:
	● Information on how the target values relate to the population (e.g. male/female frequency or age distribution by group).
	● Specify how this is taken into account in the predictive model.
		Y

Features extraction and dimension reduction	Specify the use of any: ● Feature transformation. ● Feature selection. ● Dimension reduction. When these techniques are data-driven, specify the procedures used to learn the parameters involved.	Y
Model	For traditional multivariate analyses, report: ● Type of model, e.g. MANOVA. ● Assumptions made on the covariance structure, e.g. independence, or a common arbitrary covariance between groups. ● Statistic used to assess significance, e.g. Wilk's lambda, Hotelling-Lawley trace, etc. For predictive models, report: ● Type of model, e.g. Linear discriminant analysis, support vector machines, logistic regression, etc. ● For kernel-based methods (i.e. SVM) report type of kernel used, type and number of parameters needed to be estimated.	Y
Learning method	Report:	Y

	● Figure-of-merit optimised. ● Fitting method. ● Parameter settings, those fixed and those estimated; specify how fixed parameter values were chosen. ● How the convergence of the learning method is monitored.	
Training procedure	Describe: ● Pipeline structure applied uniformly to all cases (e.g. that could be independently applied to a new case). ● Method for hyper-parameter setting. ● Data splitting (cross validation).	Y
Evaluation metrics: Discrete response	Describe the evaluation metrics that are to be computed. Always compute: ● Accuracy. ● If group sizes unequal, balanced (or average) accuracy. When there are only 2 classes, and one can be labeled “positive”: ● Precision (1 − false discovery rate). ● Recall (sensitivity). ● False positive rate (1-specificity). ● F1 (incorporates both precision and recall). ● Receiver operating characteristic (ROC) curves, e.g. summarised by area under the curve (AUC); AUC for only high specificity (e.g. false positive rates no greater than 10%) are also useful. When there are 3 or more classes: ● Report the confusion matrix.	Y
Evaluation metrics: Continuous response	“Prediction R^2”, the percentage of variance explained by prediction, computed as one minus the ratio of prediction sum-of-squares to total sum-of-squares. (Note this *is not* the squared correlation coefficient between true and predicted values).	Y
Evaluation metrics: Representational similarity analysis	Report the Kendall Tau statistic for each candidate model considered.	Y

Evaluation metrics: Significance	When possible use formal test to obtain P-value to assess whether evaluation metric is “significant” or consistent with noise.	Y
Fit interpretation	Procedure used to interpret the fit of the classifier, identifying the relative importance of the features (e.g. the weight vector in linear discriminant).	N

Table D.5. Results Reporting

Aspect	**Notes/Ontology**	**Mandatory**
Mass univariate analysis		
Effects tested	Provide a complete list of tested and omitted effects.	Y
Extracted data	● Define how voxels/elements were selected; if region is based on the same data, clarify how circularity was accounted for. ● For any summary reported, give units. Ideally these are as interpretable as possible (e.g. percent change). ● If reporting R^2 (coefficient of determination) clarify how nuisance variability is considered. For instance, in task fMRI the vast majority of variance is explained by slow temporal drift, and R^2 values for an effect of interest will be vastly different if computed with or without counting drift in the total variance.	Y
Tables of coordinates	Provide one table of coordinates including: ● Contrast / effect to which it refers. ● XYZ coordinate (with coordinate system, MNI, Talairach, noted in caption; also clarify whether peak or center-of-mass location). ● Anatomical region (in caption or body text, describe source of labels, e.g. subjective, atlas, etc).	Y

	● P-value forming basis of inference (e.g. voxel-wise FWE corrected P; or cluster-wise FDR corrected P). ● T/Z/F statistic (with degrees of freedom in table caption) ● In caption, state whether coordinates are from whole brain, or from a specific constrained volume. ● If cluster-wise inference is used, the cluster size. Report in mm^3 or, if in voxels, be explicit about the size of voxels. If a cluster statistic other than size is used (e.g. mass) it should be listed as well. ● In caption or body text, note criterion for peak per cluster reporting; e.g. “one peak per cluster listed”, or “up to 3 per cluster that are at least 8mm apart” (SPM default), etc.	
Thresholded maps	For each effect, provide images of maps of significant regions, ensuring that each caption describes: ● Type of inference and the correction method, as well as form of any sub-volume corrections applied when computing corrected significance. ● Include color bars; when presenting multiple maps in a figure, use a common color bar to ensure the results are comparable.	N
Unthresholded maps	Share, via supplementary material or repository: ● Unthresholded statistic maps. ● Optionally, the thresholded statistic maps. ● Optionally, the effect size map (e.g. % BOLD change, % GM change).	Y
Extracted data	State whether data extracted from an ROI (e.g. to compute an effect size) is defined based on independent data, as otherwise it is susceptible to bias. If ROIs are circularly defined, best not to provide any statistical summary (i.e. P-values, R^2, etc).	Y
Spatial features	Report the ● Size of the analysis volume in voxels, mm. ● Spatial smoothness of noise (e.g. FWHM) and Resel count (if using Random Field Theory).	Y

Functional connectivity		
ICA analyses	Report the total number of components (especially when estimated from the data and not fixed). Report the number of these analyzed and the reason for their selection.	Y
Graph analyses: Null hypothesis tested	For graph-based methods, carefully state what is the null hypothesis of the test and how the statistic distribution under the null is computed.	Y
Multivariate modelling & predictiveanalysis		
Optimised evaluation metrics	Report the values obtained for the evaluation metrics chosen (see Evaluation Metrics, above), as well as any P-values to justify above-chance performance.	Y

Table D.6. Data Sharing

Aspect	Notes	Mandatory
Reporting a data sharing resource		
Material shared	List types of images and non-imaging data provided. Report on the completeness of the data (e.g., number of subjects where all types of imaging, demographic, and behavioral data is available).	Y
URL, access information	Provide: ● Stable URL or DOI. ● Specific instructions on how to gain access. Specifically mention whether application must be vetted for particular intended research use (e.g. to preclude multiple users investigating the same question), or whether a research collaboration must be established.	Y

	• Cost of access.	
Ethics compliance	Confirm that the ethics board of the host institution generating the data approves the sharing of the data made available. Clarify any constraints on uses of shared data, for example, whether users downloading the data also need ethics approval from their own institution.	Y
Documentation	Provide URL to documentation, and specify its scope (e.g. worked examples, white papers, etc).	N
Data format	Report the format of the image data shared, e.g. DICOM, MINC, NIFTI, etc.	Y
Ontologies	Data organization structures, including Data Dictionaries and Schemas. Is the software using an established ontology?	N
Visualization	Availability of in-resource visualization of the imaging or non-imaging data.	N
De-identification	How, if at all, data are de-identified.	N
Provenance and history	Availability of detailed provenance of preprocessing and analysis of shared data.	N
Interoperability	Ability of a repository to work in a multi-database environment, availability of API's and ability to connect to analysis pipelines.	N
Querying	Mechanisms available for constructing queries on the repository (e.g. SQL, SPARQL).	N
Versioning	How users can check version of downloaded data and compare it to the current version at a later time.	N

Table D.7. Reproducibility

Aspect	Notes/Ontology	Mandatory
Documentation		
Tools used	Tool names, versions, and URLs.	Y
Infrastructure	Machine CPU model, operating system version, any use of parallelization.	Y
Workflow	Use of a workflow system, its version and URL.	N
Provenance trace	State whether detailed provenance information is available.	N
Literate program implementing results	Provide a URL linking to the relevant resource; for example, an ipython notebook implementing key analyses.	N
English language version	As the scientific lingua franca, documentation should be provided in English in addition to any other languages.	N
Archiving		
Tools availability	Note if tools are publically available.	N
Virtual appliances	Note if a virtual environment to facilitate a repeated analysis is available.	N
Citation		
Data	Provide permanent identifier if possible.	N
Workflow	Provide permanent identifier if possible.	N