Best Practices in Data Analysis and Sharing in Neuroimaging using MRI

A report by
Committee on Best Practices in Data Analysis and Sharing (COBIDAS)
Organization for Human Brain Mapping (OHBM)[1]
20 October, 2015

A Draft For Consideration by The OHBM Community

0. Introduction

In many areas of science and even the lay community, 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"[2] 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 ultimately (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 inherently 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].

Beyond data analysis, COBIDAS also addresses best practices in data sharing. The sharing of data can enable reuse, saving costs of data acquisition. 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.

[1] Please see Appendix 1 for authorship and acknowledgment information.

[2] http://www.humanbrainmapping.org/files/2014MeetingFiles/6c%20OHBM%20Data%20Integrity%20Statement.pdf

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 make these principles practical and usable, we created explicit lists of items to be shared (Appendix 2).

Working closely with OHBM Council, this document has been prepared by COBIDAS, released to the OHBM community for comment. Members will be given one month to provide comments, those comments will be integrated and the revised document will be presented to the membership for up/down vote, and finally submitted for publication[3]. 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 approval of the OHBM community.

Approach

There are different responses to the perceived crisis of reproducibility, from simply letting the problem `self-correct' as reviewers and readers become more aware of the problem, to dramatic measures like requiring registration of all research hypotheses before data collection. We take the view that the most pragmatic way forward 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 re-use of data.

The reader may be daunted by the sheer scale and detail of recommendations and checklists in this work (Appendix 2). 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.

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

[3] This passage will be updated to account for the actual process that transpires.

task-based and task-free functional MRI (fMRI), analyzed voxel-wise and on the surface, but inevitably 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 3), but in this work we make the distinction between reproducing results *with the same data* versus replicating a result *with different data* and possibly methods. Hence while this 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.

1. Experimental Design Reporting

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 design choices.

General Principles

For experimental design, the goal of open 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.

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 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 1. 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 tensor imaging, functional MRI, spectroscopy, etc.

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

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.

Block. A block (or alternatively "epoch") is a temporally contiguous period when a subject is presented with a particular condition.

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 [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) [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.

Subjects

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

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 30% of the population and this restriction could induce a bias towards a ‘super healthy,’ thus limiting the generalization to the population.

Behavioral Performance

The successful execution of a task is essential for interpreting the cognitive effects of a task. Be sure to report appropriate 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.

2. Acquisition Reporting

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.

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 2). Thus in the remainder of this section we only briefly review the categories of information that should be conveyed.

Device Information

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

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, field of view), parallel imaging parameters, use of field maps, and acquisition orientation are all critical information. Further details are needed for functional acquisitions (e.g. scans per session, discarded dummy scans) and diffusion acquisitions (e.g. number of directions and averages and magnitude and number of b-values).

Format for sharing

While there is some overlap with Section 6. 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). Take care, though, of sensitive protected personal information in the acquisition metadata and use appropriate anonymization procedures before sharing (see Section 6. Data Sharing).

3. Preprocessing Reporting

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.) are the most common.

General Principles

As with other areas of practice, openness here requires authors to clearly detail each manipulation done to the data before a statistical or predictive model is fit. This is also essential for reproducibility, as the exact outcome of preprocessing is dependent on the exact steps, their order and the particular software used.

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). Consider adding a Research Resource Identifier (RRID[4]) [Bandrowski2015] citation for each tool used. RRID's index everything from software to mouse strains, and provide a consistent and searchable reference.

[4] http://www.force11.org/group/resource-identification-initiative

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.

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.

4. Statistical Modeling & Inference

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.

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 example, the list of contrasts and small-volume corrections should be fully described in the methods section, whether or not it is also summarised in the results section.

Software

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

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 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). Often complicated contrasts, linear combinations of independent variables, are needed to interrogate the experimental effect of interest.

While software may make the *model* and *estimation method* seem 'automatic', a short description is needed for a complete scientific report. See Appendix 4 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 distinguish true effects from background noise, and is a crucial final step. In brain imaging, inference usually amounts to a thresholding procedure, though if ROIs are used, it could also include computation of confidence intervals. Be sure to clearly and separately state both the type of inference and the manner of multiple-testing correction. For example, the inference method description "5% cluster wise inference" doesn't specify the cluster-forming threshold nor the multiple-testing correction method (e.g. familywise vs. false discovery rate). Clearly describe the volume, sub-volume, or surface domain for which multiple-testing correction has been performed.

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 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, both to the data before fitting and as a constraint between spatial and temporal components; describe the type of scaling applied to data and extracted components. When considering multiple components, report how the components were sorted and the use of any post-hoc task regression model (with 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.

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 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.

Other Resting-State Analyses

Analysis of resting state data need not 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) [Zhang2004] measures local consistency of signals. These methods produce a map per subject that can be analyzed with a mass univariate model (see above).

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] or search-light mapping [Kriegeskorte2006]. A complete description should include details of the following: Target values, features, predictive model, and training 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).

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. 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.

The *predictive model* is the type of method used to map features to targets. Typical examples include linear discriminant analysis, support vector machines or logistic regression. It 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 method* is perhaps the most important facet of a predictive analysis, and comprises the algorithm used to build a working classifier. 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. 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 (e.g. cross-validation), 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 indeed valid only if *all* hyper-parameters are optimized inside the loop as part of a nested cross-validation procedure or chosen and fixed *a priori*.

5. Results Reporting

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.

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 inflates the significance of results shown and will stymie efforts to replicate a finding.

Mass Univariate Modelling

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 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 should be created with care. Each table or sub-portion of a table should be clearly labeled as to what contrast / effect it refers to, 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. Finally, the space (i.e., Talairach, MNI) 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.

Extracted data from images aids the interpretation of the complex imaging results, and is presented as effect magnitudes (in normalized "effect sizes" or percent change), bar plots, or scatter plots. 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 circularity problem; e.g. "derived from independently-formed ROI", or "values based on voxels in a significant cluster and are susceptible to selection bias".

Functional Connectivity

The critical issues when reporting functional connectivity differ between voxel-wise, seed-based and structured models.

When reporting multivariate decomposition methods like PCA, ICA, MDS or SOM, 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 four IC maps for significance.

Multivariate Modelling & Predictive Analysis

While it may appear that predictive analyses are trivial to report ("Accuracy was X%"), 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 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 predictive accuracy, 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 not contain any predictive signal at all, but may serve to cancel correlated noise, thus improving classifier performance. Mapping procedures that conduct the same analysis at every location, such as multivariate searchlight mapping, can identify regions that are predictive in isolation of activity elsewhere and thus complement whole-brain classification methods.

6. 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.

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 data repositories and how to convey details of retrieval.

General Principles

Data sharing is one of the cornerstones of open 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. There are four elements to a successful data sharing effort: Planning, databases, documentation, & ethics.

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 key to 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 QC. Best practice is to use an established data structure; for example, the recently developed BIDS standard[5] 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[6], 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, while sharing preprocessed images makes it easier for others to immediately start analyzing your data. Finally, sharing of extensively processed data, such as 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.

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. 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[7] or Open Data Commons[8], 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.

[5] http://bids.neuroimaging.io
[6] http://openfmri.org
[7] http://creativecommons.org/licenses/
[8] http://opendatacommons.org/licenses/

One last 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).

In short, no matter what is shared it is essential is that data sharing, as a part of a data management plan, 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.

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.

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 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 necessary 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.

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 [Sherif2014]) 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 to obtain the maximal value of the data collected, now and for years to come.

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[9] 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[10] and MINC[11] formats have forms of provenance tracking, and the NIDM project[12] is developing a framework to save this information in a standard format. Pipeline software like LONI Pipeline[13] or nipype[14] explicitly provide such provenance records.

[9] http://humanconnectome.org/documentation
[10] http://afni.nimh.nih.gov/afni/doc/faq/39
[11] http://www.bic.mni.mcgill.ca/software/minc/
[12] http://nidm.nidash.org
[13] http://pipeline.bmap.ucla.edu
[14] http://nipy.org/nipype/

Ethics
Data sharing can be difficult if ethics and consent documents are not suitably crafted. While in the United States de-identified data is not "protected health information" and should be able to be shared, regulations differ between countries and institutions and are subject to change. Hence be sure to consult your ethics or institutional review board before acquiring data with the intent of sharing, as well as before releasing data. The Open Brain Consent project[15] 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 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.

7. 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 1). 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.

Scope
We focus on analysis-level replication, i.e. the ability to reproduce 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. To further make your analysis as reproducible as possible, ensure it is *documented*, *archived* and *citable*.

Documentation
As detailed in the Preprocessing Reporting section above, be sure to cite the software and computational infrastructure used to obtain your results.

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], and SPM batch [Penny2006] may help structure and describe workflows. myExperiment [DeRoure2009] can be used to share and run workflows online (see for instance this FSL fMRI workflow from the LONI Pipeline environment[16]).

[15] http://open-brain-consent.readthedocs.org
[16] http://www.myexperiment.org/workflows/2048

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[17] 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.

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 center[18], the NITRC Resource Registry[19], or in software suites such as NeuroDebian [Halchenko2012] or Lin4Neuro [Nemoto2011].

The best 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[20] 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[21] (used e.g. in [Ziegler2014]), both NeuroDebian-based VMs, and Niak[22] (available on DockerHub[23]). 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.

Citation

[17] http://en.wikibooks.org/wiki/MINC/Reference/MINC2.0_Users_Guide
[18] http://software.incf.org
[19] http://www.nitrc.org
[20] http://neuro.debian.net (also available on Dockerhub: https://hub.docker.com/_/neurodebian)
[21] http://www.nitrc.org/plugins/mwiki/index.php/nitrc:User_Guide_-_NITRC_Computational_Environment
[22] http://simexp.github.io/niak
[23] http://hub.docker.com

URLs tend to "decay" over time, 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 create DOIs to your data and workflows, such as Zenodo[24], figshare[25] or DataCite[26] (see examples in [Tustison2014] & [Soelter2014]).

8. 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.

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 http://www.humanbrainmapping.org/cobidas[27].

Acknowledgements

Please see Appendix 1.

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

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., 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

Chavas, J. (2014). A Docker image for spiking neural network simulators. *Frontiers in Neuroinformatics*. doi:10.3389/conf.fninf.2014.18.00028

[24] http://zenodo.org
[25] http://figshare.com
[26] http://www.datacite.org
[27] URL to be confirmed.

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.

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

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

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.

De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, *25*(5), 561–567.

Delescluse, M., Franconville, R., Joucla, S., Lieury, T., & Pouzat, C. (2012). Making neurophysiological data analysis reproducible: Why and how? *Journal of Physiology-Paris*, *106*(3), 159–170.

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.

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.

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.

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

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*.

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.

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*.

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

Himberg, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage*, 22(3), 1214–22.

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.

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

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

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

Kriegeskorte, N., Goebel, R., & Bandettini, P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America,* 103(10), 3863–8. doi:10.1073/pnas.0600244103

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

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., … Zhao, J. (2013). Prov-o: The prov ontology. *W3C Recommendation, 30th April*.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat* (pp. 575–580). Springer.

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.

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.

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.

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.

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

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., … Wipat, A. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science*, 349(6251), 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., 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.

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

Sherif, T., Kassis, N., Rousseau, M.-É., Adalat, R., & Evans, A. C. (2014). BrainBrowser: distributed, web-based neurological data visualization. *Frontiers in Neuroinformatics*, 8(January), 89. doi:10.3389/fninf.2014.00089

Soelter, J., Schumacher, J., Spors, H., & Schmuker, M. (2014). Erratum to“ Automatic segmentation of odour maps in the mouse olfactory bulb using regularized non-negative matrix factorization”[NeuroImage 98 (2014) 279-288]. *NeuroImage*, *107*, 116.

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

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.

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.

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.

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.

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.

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.

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 1: 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 covered 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; gender; 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.
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, CA, United States.
Erika Proal, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz & Neuroingenia, Mexico City, Mexico.
Bertrand Thirion, Inria, Paris-Saclay University.
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 reporting checklist 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 and Robert Welsh, Jörg Stadler 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; and Ting Xu for Statistical Modeling & Inference.

Appendix 2. Itemized lists. See separate file.

Appendix 3. Defining Reproducibility.

A number of terms with overlapping meaning are used to refer to the merits of scientific findings, including reproducibility, replicability, 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 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, ISO repeatability could be measured with same-scanner, same-session test-retest variability, and ISO reproducibility to be cross-scanner or cross-site test-retest variability.

Of course, ISO reproducibility is at odds with Roger Peng’s definition, which somehow seems closer to ISO repeatability. However, as brain imaging is highly computational and has close links to computational sciences like machine learning, statistics and engineering, we retained used the “computational” notion reproducibility notion throughout this work.

Appendix 4. Short descriptions of commonly used fMRI statistical 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 many of the most commonly models available in widely used software packages.

Task fMRI. Summaries for AFNI[28], FSL[29], & SPM[30] (alphabetical order) are current as of versions AFNI_2011_12_21_1014, 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.
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).
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.

Independent Component Analysis (ICA). Methods for ICA analyses are not as consolidated as mass univariate linear modelling, but we provide some short summaries of some typical analyses in GIFT[31] and MELODIC[32] (alphabetical order), current as of GIFTv3.0a and FSL

[28] http://afni.nimh.nih.gov
[29] http://fsl.fmrib.ox.ac.uk
[30] http://www.fil.ion.ucl.ac.uk/spm
[31] http://mialab.mrn.org/software/gift

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 component moving window correlation (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.
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.

[32] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC

Available multi-modality ICA methods include FIT[33] and FSL-FLICA[34] (alphabetical order), current as of 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.

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