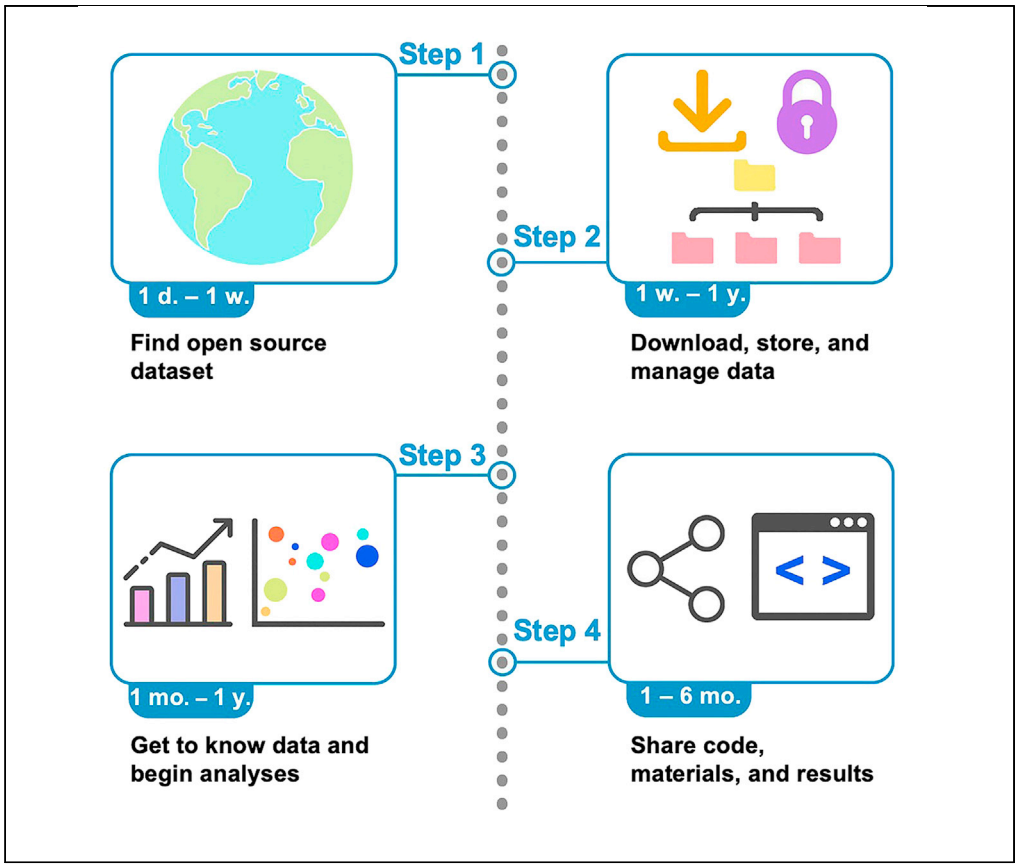
STAR Protocols

CellPress
OPEN ACCESS

Protocol

A protocol for working with open-source neuroimaging datasets



Large, publicly available neuroimaging datasets are becoming increasingly common, but their use presents challenges because of insufficient knowledge of the tool options for data processing and proper data organization. Here, we describe a protocol to lessen these barriers. We describe the steps for the search and download of the open-source dataset. We detail the steps for proper data management and practical guidelines for data analysis. Finally, we give instructions for data and result sharing on public repositories and preprint services.

Corey Horien, Kangjoo Lee, Margaret L. Westwater, ..., Teimur Kayani, R. Todd Constable, Dustin Scheinost

corey.horien@yale.edu (C.H.)
dustin.scheinost@yale.edu (D.S.)

Highlights

A protocol for working with open-source neuroimaging datasets is provided

All stages of a project are covered, from downloading data to writing up results

Instructions for data and result sharing on public repositories and preprint services

Horien et al., STAR Protocols 3, 101077
March 18, 2022
https://doi.org/10.1016/j.xpro.2021.101077

STAR Protocols

CellPress
OPEN ACCESS

Protocol

A protocol for working with open-source neuroimaging datasets

Corey Horien,[1,2,7,8,*] Kangjoo Lee,[3] Margaret L. Westwater,[3] Stephanie Noble,[3] Link Tejavibulya,[1] Teimur Kayani,[3] R. Todd Constable,[1,3,4] and Dustin Scheinost[1,3,5,6,*]

[1]Interdepartmental Neuroscience Program, Yale School of Medicine, New Haven, CT 06510, USA

[2]MD-PhD Program, Yale School of Medicine, New Haven, CT 06510, USA

[3]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT 06510, USA

[4]Department of Neurosurgery, Yale School of Medicine, New Haven, CT 06510, USA

[5]Yale Child Study Center, New Haven, CT 06510, USA

[6]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

[7]Technical contact

[8]Lead contact

*Correspondence: corey.horien@yale.edu (C.H.), dustin.scheinost@yale.edu (D.S.)
https://doi.org/10.1016/j.xpro.2021.101077

SUMMARY

Large, publicly available neuroimaging datasets are becoming increasingly common, but their use presents challenges because of insufficient knowledge of the tool options for data processing and proper data organization. Here, we describe a protocol to lessen these barriers. We describe the steps for the search and download of the open-source dataset. We detail the steps for proper data management and practical guidelines for data analysis. Finally, we give instructions for data and result sharing on public repositories and preprint services.

For complete details on the use and execution of this profile, please refer to Horien et al. (2021).

BEFORE YOU BEGIN

Large, publicly available neuroimaging datasets are becoming increasingly common in neuroscience. From the Adolescent Brain Cognitive Development (ABCD) (Casey et al., 2018) study, which investigates pediatric participants, to the Human Connectome Project Aging (HCP-A) dataset, designed to study older adults (e.g., those aged 36–100+) (Bookheimer et al., 2019), samples of various populations are available to research brain structure and function in health and disease. Despite the growing availability of openly available datasets, using them can be challenging, especially for more junior researchers. For example, software packages to download, store, manage, process, and analyze these datasets are appearing with increasing frequency, and simply navigating the tools to work with these data can feel like learning a new language. Protocols are therefore needed to help first-time users navigate the often-overwhelming world of working with large datasets.

Here, we provide a step-by-step example of issues to consider when working with open datasets. We focus on all stages of the data life cycle, highlighting steps that are often overlooked when working with these samples. Specifically, this framework is intended to help investigators navigate the myriad choices they may face when getting started using open-source neuroimaging datasets. Given the wide number of open datasets, our aim is to provide general guidelines that can be easily adapted depending on the sample, but where appropriate, specific examples are offered (particularly when

discussing how to download a sample). We refrain from offering analysis recommendations; interested readers are directed to (Bzdok et al., 2019; Bzdok and Yeo, 2017; Fan et al., 2014; Smith and Nichols, 2018).

Working with open data has taken on new importance in the ongoing COVID-19 pandemic. With laboratory-based studies largely disrupted, many investigators have had to turn to publicly available datasets or biobanks to continue their research. Our intended reader is one who has little to no prior experience working with open-source datasets. Therefore, the goal of this manuscript is to compile accessible, easy-to-follow recommendations in one place, which can serve as a resource for researchers to return to as they work through data processing over the course of their research. Due to the collaborative nature of working with these datasets, it is possible a single investigator might not need to refer to all aspects of this manuscript. Hence, we have attempted to make the tips as accessible as possible in each section—each section can be referred to as needed to serve as a guide when working with open samples.

Consider your computer processing and storage needs, as well as your timeline

1. Ensure you have considered computational processing needs
   a. For smaller studies (i.e., 20–30 subjects with an anatomical and a resting-state fMRI scan), processing data on a single local computer might be feasible for some analyses.
   b. For larger studies (i.e., hundreds or thousands of subjects) or smaller N studies with more data and/or more computationally-intensive analyses, consider using dedicated high performance (i.e., cluster) or cloud computing resources.
      i. Labs can use pre-existing cluster resources (or can potentially establish their own).
      ii. We encourage researchers to investigate options at their institution to determine if existing resources can be used to fit their needs.
      iii. If a cluster does exist at an institution, researchers may need to apply for access.
      iv. Using Amazon Web Services (AWS) is a popular option for cloud computing (https://aws.amazon.com/getting-started/); other resources are available as well (Microsoft Azure, Google Cloud, IBM Cloud, etc.).
      v. These resources are not free, and cost should be considered prior to use.
      vi. For example, it costs $552 to store 2 TB for a year using AWS; analyzing data using AWS will add to this cost (see 'team up and collaborate to save time and money' below for more about AWS analysis costs).
      vii. For more about using cloud-based resources for neuroimaging, see (https://training.incf.org/cloud-based-computer-matrix).
2. Ensure you have sufficient storage for data
   a. For the dataset used in this paper (Yale Resting State fMRI/Pupillometry: Arousal Study; https://openneuro.org/datasets/ds003673/) (Lee et al., 2021), downloading the initial dataset of 27 subjects requires 27.32 GB of disk space for neuroimaging data, eye-tracking data, and basic demographic information.
      i. Per participant, these data comprise a T1-weighted 3D anatomical image acquired using a magnetization prepared rapid gradient echo (MPRAGE) sequence and two resting-state fMRI functional scans (~7 min per scan).
      ii. All data were acquired on a 3 Tesla magnet.
      iii. More data per participant and/or higher resolution data would require more storage space.
   b. Alternatively, downloading raw imaging data from the full first release of ABCD (~10,000 subjects) would require ~13.5 TB of storage space (Horien et al., 2021).
   c. Over the course of data preprocessing, data intermediates (skull-stripped data, motion-corrected data, etc.) will take up additional storage space.
      i. Intermediate data should be backed up; this will typically double the storage space needed.

ii. Careful planning of what intermediates to back up is essential and can help reduce costs.
iii. Raw data can also be backed up if desired.
iv. However, raw data can always be redownloaded if needed, so choosing not to back up raw data can reduce costs associated with storing data.
v. It might be sufficient to delete skull-stripped images, for example, when backing up data.

3. Consider what data are needed to address the research question and the anticipated timeline for processing.
   a. Data can be available in two forms: raw data and processed data.
      i. Raw data typically comes in two forms: digital imaging and communications in medicine (DICOM) or neuroimaging informatics technology initiative (NIfTI) images.
      ii. Processed data can consist of fully processed data like connectivity matrices, statistical parametric maps, or some other form of intermediate data
   b. There are benefits and drawbacks of both raw vs. processed data.
      i. A first factor to consider in using raw vs. processed data is time.
      ii. Downloading the raw imaging data from ABCD can take days to weeks, depending on computational resources. Converting the imaging data from DICOM to NIfTI format can also take weeks. For a team of 3–4 investigators working to generate connectivity matrices, processing and QC can take upwards of 8–9 months.
      iii. Alternatively, downloading the processed connectivity matrices from ABCD can be completed in approximately one day.
      iv. Using raw vs. preprocessed data also affects storage space and is a second factor to consider.
      v. For example, the processed connectivity matrices from ABCD would require only ~25.6 MB of disk space, approximately 0.0001 percent of the space required to store the NIfTI images and intermediates if starting from raw data (Horien et al., 2021).
      vi. We note that using processed data does not mean processing steps can be ignored.
      vii. Oftentimes, it can be challenging to follow what precisely other teams have done to the data (even with knowledge of the processing steps).
      viii. Hence, using processed data might require even more time/expertise to understand other teams' processing pipelines.
      ix. For a full discussion about the strengths and weaknesses of raw vs. processed data, see (Barron and Fox, 2015).

Hone your coding and computer skills

4. Investigators should have the ability to perform basic file management operations and adapt computer code.
   a. Proficiency in a programming language (bash, MATLAB, Python, R) can help processing and analysis. For reference, bash, Python, and MATLAB are commonly used for manipulating fMRI data and using preprocessing software.
      i. It is worth noting that a researcher does not have to master all of these languages. It might be appropriate to choose a subset of popular programming languages that is sufficient for one's scientific needs.
   b. At a minimum, investigators should be able to adapt existing code to fit their research needs.
5. Resources exist to help gain familiarity with these methods:
   a. MATLAB: https://www.mathworks.com/help/matlab/getting-started-with-matlab.html
   b. Python: https://www.python.org/about/gettingstarted/
   c. R: https://support.rstudio.com/hc/en-us/articles/201141096-Getting-Started-with-R
   d. bash: https://www.computerhope.com/unix/ubash.htm
6. There are many choices of preprocessing software.
   a. The software used will depend on processing/analysis goals, the population under study, the users' familiarity with neuroimaging software, etc.

b. A full discussion is beyond the scope of this paper, but we list a few common examples; we encourage researchers to investigate each through the links included below (see 'key resources table' for the citation of each software tool).
   i. FMRIB Software Library (FSL; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki)
   ii. Statistical Parametric Mapping (SPM; https://www.fil.ion.ucl.ac.uk/spm/)
   iii. Analysis of Functional NeuroImages (AFNI; https://afni.nimh.nih.gov/)
   iv. Advanced Normalization Tools (ANTS; https://stnava.github.io/ANTs/)
   v. fMRIPrep (https://fmriprep.org/en/stable/)

7. There are many choices for analysis software; a full listing is also beyond the scope of this paper, but we point the interested reader to (Soares et al., 2016) for a discussion about the basics of fMRI data analysis as well as tools that can help researchers get started. (See the 'troubleshooting' section below for more about how to get started with processing and analysis tools.)

Team up and collaborate to save time and money

8. Working with open-source samples, especially large samples (e.g., hundreds or thousands of subjects) can be onerous for a single investigator
   a. Multiple lab members can work together throughout various phases of the data lifecycle. For example, one lab member can locate and download the data, another could lead preprocessing efforts, and so on.
   b. In our own experience, teams of 3–4 individuals can work effectively on downloading and processing data from larger samples (i.e., ~5500 participants) (Rapuano et al., 2020), with additional lab members (5–6 others) helping to manually inspect anatomical images for quality control (QC) purposes.
   c. Alternatively, multiple labs can collaborate on processing of the data, or intermediates of the data can be shared with other labs as needed.
9. Working together can save time and money
   a. Computational resources needed to store, process, and analyze large datasets (e.g., cloud-based resources) can be expensive.
   b. For example, the amount of stored data can balloon when using large available datasets, particularly if multiple users copy data or generate additional derivatives.
   c. For reference, it costs $552 to store 2 TB for a year using AWS (2000 GB * $0.023/GB/month * 12 mo). Furthermore, analyzing data on a 16-CPU compute-optimized instance, as was used in (Noble et al., 2020), currently costs about $0.7 per hour (https://aws.amazon.com/ec2/pricing/).
   d. Depending on the pipeline (which can take up to 2 h per subject), one may be able to process 1,000 subjects in about three days (1,000 subjects * 2 h/subject / (16 jobs/instance * 2 instances) = 62.5 h) using two of these instances, costing only about $40 (62.5 h * $0.7/h=$43.75).
   e. However, if anything needs to be corrected (likely during the initial setup of a pipeline) or data needs to be processed in different ways, costs can begin to add up.
   f. One of the biggest ways labs can save time and money is by sharing the same preprocessed data rather than re-processing data themselves. Processed data can also be more readily used and shared with collaborators.
   g. Computing hardware and/or cluster access can also be shared amongst labs.
   h. See the 'troubleshooting' section below for what to do if it is not possible to team up with other neuroimagers at your institution.
10. For more about effective collaborations in research, see (Bennett and Gadlin, 2012).

⚠ CRITICAL: Working with large, open-source samples can be a slow process and quite expensive. Investigators should consider the anticipated timeline for the analyses that will be necessary to address their research question, as well as associated costs for data storage and processing.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Yale Resting State fMRI/Pupillometry: Arousal Study	Lee et al. (2021)	https://openneuro.org/datasets/ds003673/versions/1.0.1
Software and algorithms		
Amazon Web Services (AWS)	Amazon Web Services	https://aws.amazon.com/
Bash	Free Software Foundation, Inc.	https://www.gnu.org/software/bash/
MATLAB (ver R2021b)	MathWorks	https://www.mathworks.com
Python (ver 3.10)	Python Software Foundation	https://www.python.org/
R (ver 3.6.2)	The R Foundation	https://www.r-project.org/
Brain Imaging Data Structure (BIDS)	Gorgolewski et al. (2016)	https://bids.neuroimaging.io/
FMRIB Software Library (FSL; ver 6.0)	Jenkinson et al. (2012)	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
Statistical Parametric Mapping (ver SPM12)	Frackowiak et al., 2004	https://www.fil.ion.ucl.ac.uk/spm/
Analysis of Functional NeuroImages (AFNI; ver AFNI_21.3.05)	Cox (1996)	https://afni.nimh.nih.gov/
Advanced Normalization Tools (ANTS)	Avants et al. (2011)	https://stnava.github.io/ANTs/
fMRIPrep	Esteban et al. (2019)	https://fmriprep.org/en/stable/
Jupyter	Project Jupyter	https://jupyter.org/
GitHub	GitHub, Inc.	https://github.com/

STEP-BY-STEP METHOD DETAILS

Find open-source dataset

Timing: 1 day to 1 week

There are many open-source samples available; it is first necessary to identify a dataset, or datasets, of interest.

1. Determine the research question(s) to be addressed.
   a. Determining the population of interest (i.e., infants, children, young adults, aging adults, individuals with a disorder, etc.) will help hone in on the appropriate sample.
   b. Specify what sorts of data are needed from this sample.
      i. For example, in a study to address the neural correlates of autism spectrum disorder (ASD), are neuroimaging data and clinical labels sufficient (i.e., case/control status)? Alternatively, are continuous symptom scores necessary?
      ii. Note that samples with many contributing sites may not have standardized data across all sites (e.g., ABIDE I/II) (Di Martino et al., 2014, 2017), whereas other studies with many sites may have harmonized measures (i.e., ABCD (Casey et al., 2018) and UK-Biobank (Miller et al., 2016)).

2. Find the dataset(s) of interest.
   a. There are many samples openly available; we list some of the large samples (i.e., those with 700+ participants) in Figure 1.
   b. Samples consist of a variety of data modalities, including imaging, genetics, and phenotypic data.
   c. Most have raw imaging data; some have data that have been minimally or fully processed (see 'troubleshooting' below for what to do if a dataset of interest cannot be accessed).
   d. Samples differ with respect to access.
      i. Some datasets, like those hosted on OpenNeuro (Poldrack et al., 2013; Poldrack and Gorgolewski, 2017), do not require an application; download and use of data are available to anyone.

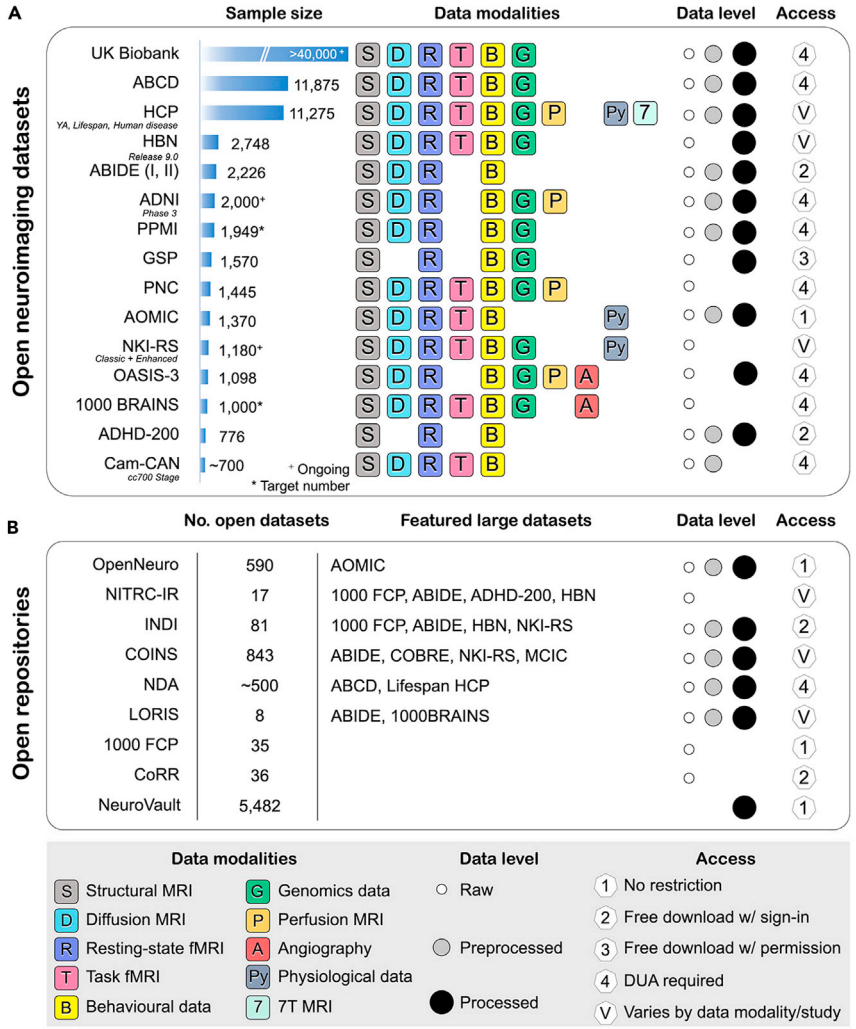


Figure 1. An overview of open-source datasets and open repositories

(A) For each dataset listed in the leftmost column, sample size is indicated, along with the type of data included ('Data modalities'). 'Data level' refers to the level of preprocessing: white circle, raw data; gray circle, some level of preprocessed data; black, processed data (for example, statistical maps, connectivity matrices, etc.).

(B) For each open repository (i.e., a collection of open datasets) listed in the leftmost column, an estimate of the number of open datasets is listed. Datasets of interest are highlighted ('Featured large datasets'). Sample sizes and the number of open datasets are current as of September 2021. Users are encouraged to visit the website associated with each dataset before use, as sample sizes, access conditions, etc. may change. Figure adapted with permission from (Horien et al., 2021).

ii. Other datasets, like ABCD (Casey et al., 2018), require a formal data usage agreement (DUA) to be approved by the organization hosting the dataset.

e. It is often useful to access multiple datasets that contain the variables of interest to assess the replicability/generalizability of any significant findings.

f. Investigators may wish to preregister their study and analysis plan at this stage (see 'trouble-shooting' below for how to preregister a study).

⚠ CRITICAL: DUAs must be approved prior to working with the data. Requirements vary by sample, but most DUAs require that all individuals in contact with the data be on the DUA and approved to work with the sample (including rotating graduate students, visiting

scholars, etc.). Some datasets require a new DUA to be approved annually. DUAs often must be signed by an institutional signing official.

g. Investigators should confer with their institutional review board (IRB) and/or human investigation committee (HIC) before downloading the data, as human research exceptions or data-sharing agreements might be required.

⚠ CRITICAL: Because the data are already collected, obtaining IRB/HIC approval is often (erroneously) neglected. Regulations with respect to using open data vary by institution and by country or region. Due to data anonymization concerns, regulations will continue to evolve. It is therefore essential that investigators consult with their institutional regulatory bodies prior to beginning work with these samples and, if necessary, once they have obtained the data.

h. Datasets differ with respect to participant preferences about data usage.
   i. For instance, participants in the UK-Biobank (Miller et al., 2016) may withdraw their data at any time (or request that research teams delete it).
   ii. Alternatively, participants in the National Institute of Mental Health Data Archive can request to have their data withdrawn for future download but are unable to request that research teams with the data delete it.
   iii. Researchers should be aware of privacy standards with their dataset and ensure that they are compliant with participant requests.
   iv. For the remainder of this protocol, we will focus on data obtained from OpenNeuro, but, where appropriate, we will highlight points of divergence from other open-source samples.

Download, store, and manage data

Timing: 1 week to 1 year

In this section, we will discuss how to download, store, and manage data from an example dataset (Yale Resting State fMRI/Pupillometry: Arousal Study. https://openneuro.org/datasets/ds003673/) (Lee et al., 2021). This dataset includes resting-state fMRI data that were acquired with simultaneous pupillometry from 27 healthy adults (26.5 $\pm$ 4 years old; 16 females). We will focus on imaging data in this example.

3. Downloading data
   a. Check the version of the dataset: Visiting the dataset link (https://openneuro.org/datasets/ds003673/) from your browser will take you to the latest available version of this dataset (Figure 2). At this time, you can see version 1.0.1 published on 2021-08-21, as shown in the left panel. If this is not the version you are looking for, navigate the left panel to select the right one.

   Note: Open-source datasets often have multiple releases, and researchers should check the study website to see which version they want. This is important to check because 1) sample size and/or the number of follow-up time points tends to increase with the version and 2) there might have been processing errors in prior releases that have been corrected.

   b. Check the information of the dataset in the README and CHANGE files to review the history and data information provided by the authors.
   c. You can download the full dataset by clicking the DOWNLOAD button below the title panel. Several download options will appear (e.g., Download with your browser, Node.js, S3, or DataLad).
   d. Or, you can manually download parts of the dataset depending on your interest.
      i. Click the subject data directory (e.g., sub-pa1372) you wish to download to view the list of files.

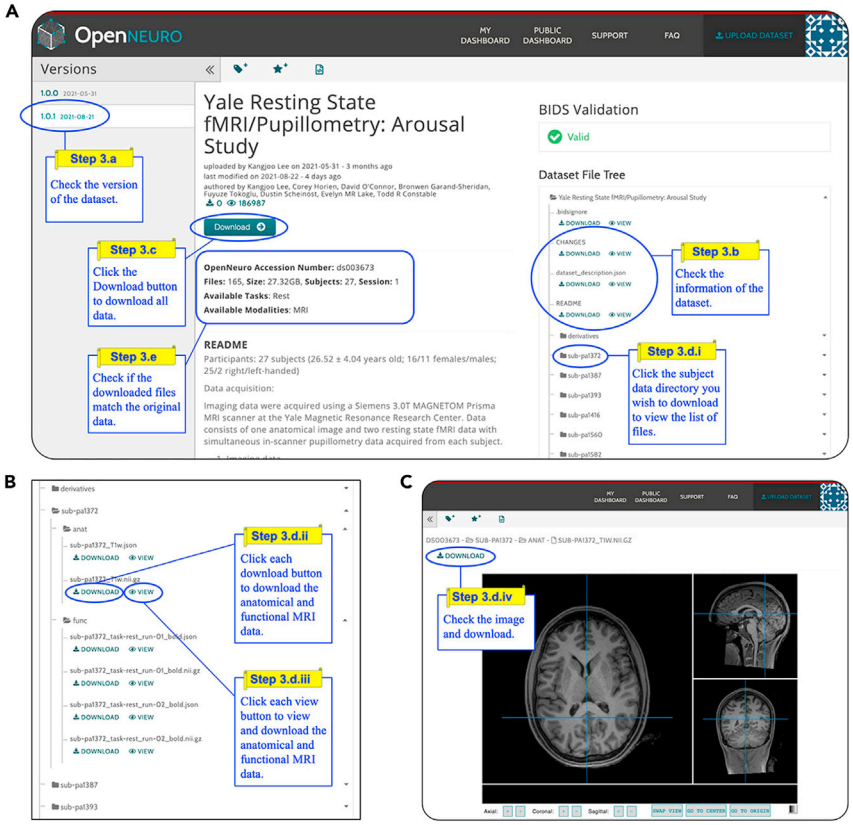


Figure 2. Steps to download data from OpenNeuro

ii. Click each DOWNLOAD button below the name of the file to download the anatomical and functional MRI images.
iii. Or, click each VIEW button below the name of the file, to view the image.
iv. You can download the image after viewing.

e. Check if the downloaded files match with the original data.
f. Make sure you cite the dataset or related publications when you use the dataset. The citation information is usually provided on the title page and README files provided by the authors.

4. Storage
   a. Brain imaging data structure (BIDS) (Gorgolewski et al., 2016) is a common data organizational standard in neuroimaging that helps facilitate use and reuse by investigators.
      i. For help with BIDS, see https://github.com/bids-standard/bids-starter-kit
   b. All datasets from OpenNeuro are organized according to BIDS.
   c. Raw anatomical and functional data are stored in specific folders per subject (Figure 3A).
   d. Data derived from each subject over the course of the study are stored in a subject-specific folder in the derivatives directory.
      i. For example, pupillometry data are shown in the derivatives folder for a representative participant (Figure 3B).
   e. Some legacy, open-source datasets (i.e., early HCP releases) might not be organized according to BIDS.
      i. Investigators can restructure their dataset to match BIDS standards or retain the original data structure.
      ii. The main goal is to have consistent organization across all participants.

Figure 3. BIDS format

(A) Each participant has folders containing raw anatomical and functional data.

(B) Data generated during the course of the study are stored in a derivatives folder.

5. Data management
   a. When working with the data, depending on a team member's experience/role, raw data files can be made read-only (so that they cannot be accidentally modified or deleted).
      i. For example, if a team member has little experience coding and is tasked with performing QC on image registration, read-only privileges are sufficient.
   b. Keep track of what was done to the data
      i. Documentation should enable a knowledgeable researcher within the field to exactly recreate the workflow.
      ii. This includes what was done to the data, why it was done, the code/software used, and who performed each step.
      iii. Resources like Google Docs and Jupyter (https://jupyter.org/) can be used as virtual lab notebooks if desired.
   c. Resources like Slack and Microsoft Teams can be helpful to facilitate communication among team members when managing the data (as well as during all aspects of a project).

⚠ CRITICAL: Maintaining a well-documented lab notebook is critical, particularly for larger datasets that might take up to a year to process and involve many team members. Given that junior personnel are often tasked with managing these large samples, keeping a clear, concise record can help maintain progress if/when lab members move on with their training.

⚠ CRITICAL: An important aspect of data management that is often overlooked is checking for updates to a dataset. Oftentimes, issues are discovered by the team collecting the data that can significantly affect processing and analyses. Most teams hosting the data have a QC page, a wiki, or a point person responsible for fielding issues. These resources should be frequently consulted as soon as downloading the data.

d. After data download, it can be helpful to have a lab member designated to managing data and watching for updates.
   i. For instance, this lab member could be responsible for maintaining documentation, managing which team members have access to the data, and checking to see if data/QC updates are available for the sample.
   ii. In terms of updates, be on the lookout for new data releases, scanner/software upgrades, different behavioral indices being used, and basic QC issues.
e. Social network services, such as Twitter, can also be a useful resource for obtaining advice from colleagues working with the same data or determining if other groups have noticed a QC issue. Neurostars.org can also be a helpful resource for posting questions/issues.
f. Most databases have listservs that email about new data releases and bugs that might have been discovered in previously released data.

⚠ CRITICAL: If issues are discovered, researchers should share this information with the team hosting the data, so fixes can be put into place.

Get to know data, begin analyses

Timing: 1 month to 1 year

Getting to know your dataset is crucial, especially as a user who did not participate in data collection.

6. Investigate how many subjects have the data of interest.
   a. Missing imaging and behavioral data can affect analyses and should be investigated to arrive at a final sample ready for analysis.
      i. If data are missing, determine how this will affect analyses.
      ii. There are multiple ways to handle missing data (i.e., listwise deletion, pairwise deletion, imputation) (Kang, 2013).
      iii. Researchers may want to consult a statistician for assistance, given the complexity of missing data.
      iv. In terms of the imaging data, some participants may have incomplete scans, some may be missing scans, and some may have repeated scans.
   b. A full discussion of imaging QC is outside the scope of this protocol, but we note that automated QC tools exist to aid users (https://mriqc.readthedocs.io/en/stable/) (Esteban et al., 2017).
      i. In addition, if QC info is made available by the data collection team (or other labs), researchers could use this information about which subjects to exclude.
   c. For the behavioral data, the same steps should be completed: determine if all demographic/behavioral/clinical data is available, reasons data might be missing, or if the desired versions of behavioral indices were used.
      i. If raw data are available, it might also be advisable to determine if scores were computed correctly.
7. Determine if aspects of study design will affect analyses
   a. For example, the scans in the young adult HCP (Van Essen et al., 2013) study were collected on back-to-back days.
   b. In the Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2014), all scans were collected on the same day.
   c. In addition, similar tasks (i.e., a working memory task) can differ substantially across datasets.
   d. These differences can impact analyses within samples and can also affect analyses if one is planning to use certain datasets as test samples (i.e., to determine if an effect observed in one sample is also observed in another, independent sample; see 'expected outcomes' for more about using multiple samples to assess generalizability).

8. Also investigate scanner type, software, and acquisition sequences.
9. Determine if these parameters are consistent across participants or if any of these parameters were updated in between data releases.
10. Imaging task-data can contain errors impacting analyses.
    a. If block order in a task is counterbalanced, ensure it is consistent across all participants.
        i. For example, approximately 30 subjects in the S900 release in the young adult HCP sample have a different block order in the working memory task than is reported for most participants (Horien et al., 2021).
    b. Investigate task timing; task regressors can fail to match overall task duration (as is the case for some subjects in the emotion task in the young adult HCP sample; http://protocols.humanconnectome.org/HCP/3T/task-fMRI-protocol-details.html).
    c. Task stimuli may occasionally be missed or presented for different durations (as is the case for the stop-signal task in ABCD) (Bissett et al., 2021).

⚠ CRITICAL: Issues like these may or may not be reported with the data release. Block order and task timing are only a few examples of issues to be aware of; it is crucial to determine if other issues are present in the data.

11. After obtaining the final sample for analysis, perform basic steps to get to know the data.
    a. First investigate basic demographics, like age, sex, and handedness.
    b. Family structure should also be considered.
        i. The young adult HCP sample and the ABCD sample consists of many twins and siblings; these should be accounted for when working with the data (e.g., (Winkler et al., 2015)).

⚠ CRITICAL: These basic steps are essential. Because they are somewhat obvious, however, they are often overlooked. After performing QC and excluding subjects, distributions of basic demographic variables can be skewed, and this can negatively impact analyses.

    c. Some open samples contain many contributing sites (e.g., ABIDE I/II (Di Martino et al., 2017; Di Martino et al., 2014), ABCD (Casey et al., 2018), UK-Biobank (Miller et al., 2016)); determine if sites differ in systematic ways that will affect analyses (see 'troubleshooting' section below for what to do when confounds are present in the data).
    d. If available, also investigate what time of day participants were scanned (Orban et al., 2020; Trefler et al., 2016), what time of year, smoking status, etc. In larger samples, these factors could amplify uninteresting sources of variance in datasets and act as confounds.
    e. The following site contains examples of basic visualizations that can be performed to get to know a dataset, along with R packages and toy data (http://uc-r.github.io/gda).

12. Investigate behavioral measures.
    a. Participant measures beyond basic demographic information should be examined prior to use (i.e., performance on cognitive tests, self-report measures, clinician assessments)
        i. A good place to start is to determine if data were collected in a similar manner across participants. Measures within a dataset might differ, particularly in multi-site data.
        ii. For instance, in ABIDE, sites used different versions of the Autism Diagnostic Observation Schedule (ADOS), and only some sites had research-certified clinicians administer ADOS.
        iii. Issues like these can impact other behavioral metrics.
    b. Behavioral data can also be released as summary scores for a measure, standardized scores, subscale-specific scores, etc.
    c. Ensure you are using the behavioral score you intended to use.
13. Conduct analyses.
    a. Thinking carefully about reproducible inference is key when using open samples.
    b. Particularly with larger samples, small correlations that might bear little practical importance can become statistically significant.

c. It is therefore useful to define what effect size would be meaningful for this particular study before beginning analyses.
d. Reporting multiple lines of converging evidence can increase confidence that a given result is replicable.
    i. Using multiple open-source samples in a given study is one way to test if results are convergent (see 'expected outcomes' for examples of using multiple open samples in a study).
e. Negative results should be reported.
    i. The literature as a whole can be skewed by not doing so (Easterbrook et al., 1991; Rosenthal, 1979).
    ii. Other labs may be planning similar analyses, so the reporting of negative results could ensure duplicate efforts are not being conducted.
f. Investigators might be tempted to cherry-pick positive findings within their sample or search for a sample that gives positive results (both examples of "*p*-hacking"; see 'troubleshooting' for tips on ways to reduce the inclination to *p*-hack).
g. See (Smith and Nichols, 2018) for more on statistical issues that might be encountered when working with large, open-source datasets.

Share code, materials, and results

Timing: 1–6 months

To contribute to the open-science ecosystem, investigators should openly share code, materials, and results.

14. Processing and analysis code should be shared.
    a. Code can be shared while the project is ongoing or when a paper is submitted/published.
    b. GitHub is a popular option for sharing code (https://github.com/).
        i. GitHub Guides is a helpful resource for getting started (https://guides.github.com/).
        ii. Additionally, adopting the standards used by popular open-source projects can be helpful (https://github.com/scikit-learn/scikit-learn).
        iii. Code should be well-documented and well-organized. Ideally, code should also be bug-free and efficient (in terms of time to run and overall structure).
        iv. Including readme files, adding comments to code, and fixing any bugs that are discovered by other users is good practice.
        v. See https://code.tutsplus.com/tutorials/top-15-best-practices-for-writing-super-readable-code–net-8118 for more recommendations for structuring code.

Note: While ideally code will be well-documented and polished, it does not have to be perfect to share.

15. Share materials generated from working with the sample.
    a. There are many repositories to share materials (Table 1).
        i. Preprocessed data (i.e., skull-stripped anatomical images, motion-corrected functional data) can be shared.
        ii. Derived, statistical data (i.e., parametric brain maps, parcellations) can also be shared.
        iii. Data should be shared with a clear license so that other researchers know what usage restrictions accompany reuse of the data (if any).
        iv. See https://creativecommons.org/about/cclicenses/ for more about Creative Commons licenses.

⚠ CRITICAL: Before sharing materials, researchers should check their DUA to determine what can be shared. Some datasets (for example, those obtained through the Consortium for Reliability and Reproducibility; Figure 1) allow materials to be shared openly, whereas

Table 1. A sampling of online data repositories available for sharing different levels of data

	Available repositories								
Data level	Balsa	COINS	INDI	Loris	NDA	NeuroVault	NITRC-IR	Omega	Open neuro
Preproc. data	Y	Y		Y	Y		Y	Y	
Derived, statistical parametric data	Y	Y		Y	Y	Y			
Access	https://balsa.wustl.edu	https://coins.trendscenter.org/	http://fcon_1000.projects.nitrc.org/	https://loris.ca/	https://nda.nih.gov/	https://neurovault.org/	https://www.nitrc.org/	https://www.mcgill.ca/bic/resources/omega	https://openneuro.org/

Table adapted with permission from (Horien et al., 2021).

others are more restrictive and do not permit the sharing of materials. Investigators should also consult with their IRB/HIC to determine if participants consented to data sharing when they took part in the original study.

16. Results can be shared via preprint servers
    a. These services are free and allow the dissemination of results prior to publication.
    b. Different preprint servers exist and can be used depending on the nature of a study
        i. bioRxiv (https://www.biorxiv.org/) can be used to share papers in the life sciences (i.e., neurotypical participants are studied).
        ii. medRxiv (https://www.medrxiv.org/) can be used to share papers in the medical sciences (i.e., patients are studied).
        iii. PsyArXiv (https://psyarxiv.com/) hosts papers in the psychological sciences.
        iv. Other preprint servers (arXiv: https://arxiv.org/ and OSF Preprints: https://osf.io/preprints/) are available for posting preprints as well.
    c. Upon manuscript acceptance, many journals have an open-access option (for a fee).
    d. Funding agencies also may require posting papers to publicly available servers (e.g., PubMed).
    e. When writing up results, consult the Committee on Best Practices in Data Analysis and Sharing (COBIDAS) (Nichols et al., 2017) guidelines about what to include in a manuscript.
        i. COBIDAS includes 'mandatory' and 'not mandatory' recommendations.
        ii. The full list of mandatory recommendations is outside the scope of this paper but can be viewed here: http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf
        iii. Special considerations about what to report for open datasets include participant IDs, the data release used, the date the data were accessed, and the URL from where the data were obtained.
        iv. Standards regarding participant IDs can differ across datasets.
        v. For example, ABIDE participant IDs are the same for all researchers across downloads.
        vi. For UK-Biobank, unique participant IDs are generated for each group working with the data.
        vii. Prior to sharing participant IDs, researchers should determine if this is allowed in their DUA.
        viii. If no official data release is available, researchers should include as much information about the data as possible, as well as where the data were obtained.
        ix. If data are released in a continuous fashion (i.e., the ABCD Fast Track data releases new imaging data monthly), reporting the date the data were downloaded can be helpful.
        x. When working with open samples, some of the details of the dataset being used may have been reported in previous papers, so it might be sufficient to include a reference to these original studies.
        xi. Nevertheless, providing a concise summary of critical details is still advised.

xii. For example, providing a summary of imaging acquisition parameters, the preprocessing pipeline, and behavioral measures should be included, along with a description of how the data were used and analyzed.

EXPECTED OUTCOMES

We have detailed steps for how to work with open-source datasets throughout all phases of the data lifecycle. In our own experience, we have been able to leverage 2–4 open-source datasets in a single study. Such an approach can be used to aid generalizability and ensure that results hold under multiple contexts. For instance, we (Greene et al., 2018) have used open-source data from HCP young adult sample and the PNC to show that when predicting participant traits from functional connectivity data, generating predictive models using task-based data resulted in higher prediction performance compared to predictive models generated using resting-state data. This finding was strengthened by holding for both samples, which are composed of different populations (i.e., HCP comprises healthy young adults, whereas PNC comprises a population-based sample of youths aged 8–21 years). In other work, we (Horien et al., 2019) have used four open-source resting-state fMRI samples to show that the functional connectome is unique and stable across months to years. Similar to above, the results held for different populations (i.e., from adolescent children up to adults aged 70–80 years).

Such studies would be extremely difficult for single labs to conduct in isolation, and they point to the power of using shared data. More generally, the use of open data depends on and reinforces the open-science ecosystem that is rapidly becoming the norm in neuroimaging. Such an ecosystem is necessary, given concerns about a lack of reproducibility in fMRI (Poldrack et al., 2017). Both a lack of reproducibility and a lack of reliability are factors that can be addressed in single studies and by single labs if data are used thoughtfully. (We point out that a lack of reliability has been proposed as one of the reasons (Milham et al., 2021) fMRI has largely failed to have much of an effect on clinical practice.) Hence, using open-source data to assess the reproducibility and replicability of findings, and using these samples to bolster the generalizability of results, is an important step for the field.

In addition, using the steps outlined allows access to large, publicly available samples. Given recent work suggesting that large sample sizes are needed to obtain reliable effect sizes in brain-behavior associations (Marek et al., 2020), investigators using large samples will be well powered to detect their effect of interest. This can also improve the reproducibility of findings, which will aid researchers in their quest to understand the human brain.

LIMITATIONS

The availability of open-source datasets and tools to work with them are both constantly growing, and we have only described a few of the many possible choices here. We have attempted to account for this by making the protocol as general as possible. However, the specific steps one may have to undertake will likely vary depending on where data have been accessed (e.g., the steps for working with a dataset from OpenNeuro might vary slightly compared to working with a dataset from NDAR). In addition, some of the steps and/or tools described here might not be appropriate for all samples (e.g., a dataset comprising infants).

More generally, working with open-source datasets presents several other issues that should be considered. For instance, working with data collected by others ties future researchers to choices made by the data collection team, ranging from data acquisition parameters to certain preprocessing choices. Therefore, it is important that researchers think critically about their research question and the type of data that might be needed to address it before working with large datasets. Moreover, issues with the software/hardware used to acquire the data might be known at a later date (or might never be discovered by the data collection team). When working with data

that others have collected, it can often be impossible to know of such issues. If such issues are discovered, this can necessitate re-downloading the data, reprocessing the data, rerunning analyses, and so on, which are nontrivial tasks. Finally, the dataset needed to address a specific research question or aim might not be publicly available. In this case, collecting one's own data would be required.

TROUBLESHOOTING

Problem 1

I am not sure where to turn to learn more about processing and analysis tools ('before you begin', step 7).

Potential solution

There are many choices when it comes to deciding on processing and analysis tools, and getting started with these packages can be daunting. The following reference (Soares et al., 2016) is a good resource to get acquainted with some common tools used in fMRI processing and analysis and also serves as a good introduction to become acquainted with fMRI basics—from background on study design to tips on reporting/interpreting results.

Problem 2

I am one of the few neuroimaging researchers at my university—how can I team up with other investigators? ('before you begin', step 9).

Potential solution

In this case, there are more formal collaborations that researchers can join to foster new collaborations with researchers who have a shared interest, like the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium (Thompson et al., 2014), Multi-centre Epilepsy Lesion Detection (MELD) Project (https://meldproject.github.io//), to list a few. In addition, neuroimaging hackathons (Gau et al., 2021) are held around the world and are forums for researchers to meet and solve problems (as well as brainstorm research ideas). Attending events like these can help connect individual researchers to others in their community.

Further, collaborations do not need to stay within neuroimaging; teaming up with investigators in other fields can help advance research aims. For instance, if conducting imaging-genetics research, it may be useful to reach out to geneticists, statisticians, or even clinicians, depending on study aims.

Problem 3

The dataset of interest cannot be accessed, or there is another issue discovered after download (step 2 of protocol).

Potential solution

Occasionally, researchers may not be able to access a dataset of interest on a platform, or portions of the data might be missing due to technical errors. Additionally, a researcher may discover that there was an issue with the data download, in that subjects are missing, or data file types might be corrupted. In all of these cases, the organization hosting the dataset should be contacted. For datasets like ABCD, there are dedicated research staff available to help with data access issues. Additionally, datasets like HCP have a website where QC issues can be reported (https://wiki.humanconnectome.org/). This could be consulted to determine if solutions are available to any issues that are encountered. Finally, reaching out to others who have accessed or published on the data is a viable solution. Reaching out to other members of the research community through social media platforms like Twitter is common, and in our experience, can be a helpful means to help solve problems encountered with open-source samples.

Problem 4

I want to preregister my study as a way to reduce *p*-hacking (step 2, step 13 of protocol).

Potential solution

With the number of datasets available, it might be tempting for investigators to run many different analyses on many different datasets until "something works," and then write up this result for publication. Preregistration of the dataset one is intending to work with, as well as the research question and analysis plan, is one solution to help prevent the temptation to engage in "*p*-hacking" or data dredging. See https://www.cos.io/initiatives/prereg for more about how to preregister a study. In addition, we reiterate that publishing null results is important for the field, particularly in large datasets. Making this the norm in the field could help discourage other researchers from *p*-hacking.

Problem 5

Confounds are present in the dataset (step 11 of protocol).

Potential solution

Confounds are a fact of life when working with large neuroimaging datasets. Of particular importance to consider are site effects. Many openly available datasets (e.g., ABCD, ABIDE, UK-Biobank) comprise multiple sites, so care must be taken to account for this potential confound in analyses. Specifically, tools like ComBat can be used to remove between-site variance (Fortin et al., 2017, 2018). Additionally, if using a prediction-based approach, entire sites can be left out as a test sample (i.e., models are trained in N-1 sites and then tested in site N; (Lake et al., 2019; Scheinost et al., 2019; Sripada et al., 2019)). Exact solutions will depend on analysis goals; these are simply two solutions that might be useful. See (Alfaro-Almagro et al., 2020) for a thorough discussion of confounds in the UK Biobank and ways to address them.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Corey Horien (corey.horien@yale.edu).

Materials availability

This study did not generate any new materials.

Data and code availability

This study did not generate any new datasets or code.

ACKNOWLEDGMENTS

This work was supported by NIH grants T32GM007205 (CH), K00MH122372 (SN), and T32DA 022975 (MLW). The authors thank Dr. Francesca Mandino for helpful comments on a draft of this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, CH and DS; methodology, CH, SN, LT, KL, DS; writing, CH, KL, MLW, TK, SN, LT, RTC, DS; supervision, RTC and DS.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Anderson, J.L.R., Bastiani, M., Miller, K.M., Nichols, T.E., and Smith, S.M. (2020). Confound modelling in UK Biobank brain imaging. Neuroimage *224*, 117002.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., and Gee, J.C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage *54*, 2033–2044.

Barron, D.S., and Fox, P.T. (2015). BrainMap database as a resource for computational modeling. Brain Mapp. Encyclopedic Reference *1*, 675–683.

Bennett, L.M., and Gadlin, H. (2012). Collaboration and team science: from theory to practice. J. Investig. Med. *60*, 768–775.

Bissett, P.G., Hagen, M.P., Jones, H.M., and Poldrack, R.A. (2021). Design issues and solutions for stop-signal data from the adolescent brain cognitive development (ABCD) study. Elife *10*, e60185. https://doi.org/10.7554/eLife.60185.

Bookheimer, S.Y., Salat, D.H., Terpstra, M., Ances, B.M., Barch, D.M., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Diaz-Santos, M., Elam, J.S., et al. (2019). The lifespan human connectome project in aging: an overview. Neuroimage *185*, 335–348.

Bzdok, D., Nichols, T.E., and Smith, S.M. (2019). Towards algorithmic analytics for large-scale datasets. Nat. Mach Intell. *1*, 296–306.

Bzdok, D., and Yeo, B.T.T. (2017). Inference in the age of big data: future perspectives on neuroscience. Neuroimage *155*, 549–564.

Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., et al. (2018). The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. *32*, 43–54.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162–173.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci. Data *4*, 170010.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry *19*, 659–667.

Easterbrook, P.J., Berlin, J.A., Gopalan, R., and Matthews, D.R. (1991). Publication bias in clinical research. Lancet *337*, 867–872.

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., and Gorgolewski, K.J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One *12*, e0184661.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. Natl. Sci. Rev. *1*, 293–314.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. Neuroimage *167*, 104–120.

Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. Neuroimage *161*, 149–170.

Frackowiak, R.S.J., Ashburner, J.T., Penny, W.D., Zeki, S., Friston, K.J., Frith, C.D., Dolan, R.J., and Price, C.J. (2004). Human Brain Function (Academic Press).

Gau, R., Noble, S., Heuer, K., Bottenhorn, K.L., Bilgin, I.P., Yang, Y.F., Huntenburg, J.M., Bayer, J.M.M., Bethlehem, R.A.I., Rhoads, S.A., et al. (2021). Brainhack: developing a culture of open, inclusive, community-driven neuroscience. Neuron *109*, 1769–1775.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data *3*, 160044.

Greene, A.S., Gao, S., Scheinost, D., and Constable, R.T. (2018). Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun. *9*, 2807.

Horien, C., Noble, S., Greene, A.S., Lee, K., Barron, D.S., Gao, S., O'Connor, D., Salehi, M., Dadashkarimi, J., Shen, X., et al. (2021). A hitchhiker's guide to working with large, open-source neuroimaging datasets. Nat. Hum. Behav. *5*, 185–193.

Horien, C., Shen, X., Scheinost, D., and Constable, R.T. (2019). The individual functional connectome is unique and stable over months to years. Neuroimage *189*, 676–687.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., and Smith, S.M. (2012). Fsl. Neuroimage *62*, 782–790.

Kang, H. (2013). The prevention and handling of the missing data. Korean J. Anesthesiol *64*, 402–406.

Lake, E.M.R., Finn, E.S., Noble, S.M., Vanderwal, T., Shen, X., Rosenberg, M.D., Spann, M.N., Chun, M.M., Scheinost, D., and Constable, R.T. (2019). The functional brain organization of an individual allows prediction of measures of social abilities transdiagnostically in autism and attention-deficit/hyperactivity disorder. Biol. Psychiatry *86*, 315–326.

Lee, K., Horien, C., O'Connor, D., Garand-Sheridan, B., Tokoglu, F., Scheinost, D., Lake, E.M.R., and Constable, R.T. (2021). Arousal impacts distributed hubs modulating the integration of brain functional connectivity. bioRxiv. https://doi.org/10.1101/2021.07.12.452041.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Feczko, E., et al. (2020). Towards reproducible brain-wide association studies. bioRxiv. https://doi.org/10.1101/2020.08.21.257758.

Milham, M.P., Vogelstein, J., and Xu, T. (2021). Removing the reliability bottleneck in functional magnetic resonance imaging research to achieve clinical utility. JAMA Psychiatry *78*, 587–588.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. *19*, 1523–1536.

Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.B., et al. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. *20*, 299–303.

Noble, S., Scheinost, D., and Constable, R.T. (2020). Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. Neuroimage *209*, 116468.

Orban, C., Kong, R., Li, J., Chee, M.W.L., and Yeo, B.T.T. (2020). Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. Plos Biol. *18*, e3000602.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafo, M.R., Nichols, T.E., Poline, J.B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. *18*, 115–126.

Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., and Milham, M.P. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform *7*, 12.

Poldrack, R.A., and Gorgolewski, K.J. (2017). OpenfMRI: open sharing of task fMRI data. Neuroimage *144*, 259–261.

Rapuano, K.M., Rosenberg, M.D., Maza, M.T., Dennis, N.J., Dorji, M., Greene, A.S., Horien, C., Scheinost, D., Todd Constable, R., and Casey, B.J. (2020). Behavioral and brain signatures of substance use vulnerability in childhood. Dev. Cogn. Neurosci. *46*, 100878.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychol. Bull *86*, 638–641.

Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al. (2014). Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage *86*, 544–553.

Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D.S., et al. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage *193*, 35–45.

Smith, S.M., and Nichols, T.E. (2018). Statistical challenges in "big data" human neuroimaging. Neuron *97*, 263–268.

Soares, J.M., Magalhaes, R., Moreira, P.S., Sousa, A., Ganz, E., Sampaio, A., Alves, V., Marques, P., and Sousa, N. (2016). A Hitchhiker's guide to functional magnetic resonance imaging. Front Neurosci. *10*, 515.

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W.K., Luciana, M., Weigard, A., Hyde, L.H., and Heitzeg, M. (2019). Prediction of neurocognition in youth from resting state fMRI. Mol. Psychiatry *25*, 3413–3421.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., et al. (2014). The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. *8*, 153–182.

Trefler, A., Sadeghi, N., Thomas, A.G., Pierpaoli, C., Baker, C.I., and Thomas, C. (2016). Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. Neuroimage *133*, 41–52.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., and Consortium, W.U.-M.H. (2013). The Wu-Minn human connectome project: an overview. Neuroimage *80*, 62–79.

Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., and Smith, S.M. (2015). Multi-level block permutation. Neuroimage *123*, 253–268.