Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Title

A Hitchhiker's Guide to Working with Large, Open-Source Neuroimaging Datasets

Corey Horien[1,2#], Stephanie Noble[3], Abigail S. Greene[1,2], Kangjoo Lee[3], Daniel S Barron[4], Siyuan Gao[5], David O'Connor[5], Mehraveh Salehi[5,6], Javid Dadashkarimi[7], Xilin Shen[3], Evelyn MR Lake[3], R. Todd Constable[1,3,5,8], Dustin Scheinost[1,3,5,9,10#]

[1]Interdepartmental Neuroscience Program, Yale School of Medicine
[2]MD/PhD program, Yale School of Medicine
[3]Department of Radiology and Biomedical Imaging, Yale School of Medicine
[4]Department of Psychiatry, Yale School of Medicine
[5]Department of Biomedical Engineering, Yale University
[6]Summary Analytics Inc., Seattle, WA
[7]Deparment of Computer Science, Yale University
[8]Department of Neurosurgery, Yale School of Medicine
[9]Deparment of Statistics & Data Science, Yale University
[10]Child Study Center, Yale School of Medicine

[#]Corresponding authors:
Corey Horien, Dustin Scheinost
Magnetic Resonance Research Center
300 Cedar St
PO Box 208043
New Haven, CT 06520-8043
corey.horien@yale.edu
dustin.scheinost@yale.edu

Keywords
Open-science, big data, fMRI, data sharing, data management

Highlights
Practical tips for working with large, open-source neuroimaging datasets
Suggestions regarding data management
Recommendations for confounds to consider
Tips for communicating results
Discussion of emerging issues with large datasets

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Abstract

Large datasets that enable researchers to perform investigations with unprecedented rigor are growing increasingly common in neuroimaging. Due to the simultaneous increasing popularity of open science, these state-of-the-art datasets are more accessible than ever to researchers around the world. While analysis of these samples has pushed the field forward, they pose a new set of challenges that might cause difficulties for novice users. Here, we offer practical tips for working with large datasets from the end-user's perspective. We cover all aspects of the data life cycle: from what to consider when downloading and storing the data, to tips on how to become acquainted with a dataset one did not collect, to what to share when communicating results. This manuscript serves as a practical guide one can use when working with large neuroimaging datasets, thus dissolving barriers to scientific discovery.

Keywords: Open-science; big data; fMRI; data sharing; data management

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Introduction

As a part of the open science movement in neuroimaging, many large-scale datasets, including the Human Connectome Project (HCP)[1], the Adolescent Brain Cognitive Development (ABCD) study[2], and the UK Biobank[3], have been released to investigators around the world (Figure 1). These initiatives have advanced efforts to understand human brain function. Notably, they have been collected in response to—and helped provide support for—the realization that many questions in the field are associated with small effect sizes only detectable with large samples[4, 5]. Since adequately large samples can be difficult for any single lab to collect in isolation, these large datasets unlock a path to investigate previously inscrutable questions.

Nevertheless, use of these large datasets can be daunting. With thousands of subjects and substantial imaging data per subject, simply downloading and storing the data can be difficult. The complex structure of these large datasets (e.g. multiple data releases from HCP, multiple sites contributing to ABCD, etc.) presents considerable challenges and requires adherence to best practices. Even day-to-day concerns, like maintaining a lab notebook, take on new importance when handling such data.

Here, we present tips for those who will be handling these data as end-users. We offer recommendations for the entire life cycle of data use—from downloading and storing data, to becoming acquainted with a dataset one did not collect, to reporting and sharing results (Table 1). Note that we do not provide recommendations for specific analytical approaches using large datasets, as these topics have been discussed elsewhere[4, 6, 7, 8]. Our intention is to bring together in one place accessible and general recommendations, incorporating practical suggestions based on our experience working with numerous large datasets. Our intended reader is one who might

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

be tasked with working with a large dataset for the first time, and we envision this manuscript to serve as an ongoing guide throughout this exciting process.

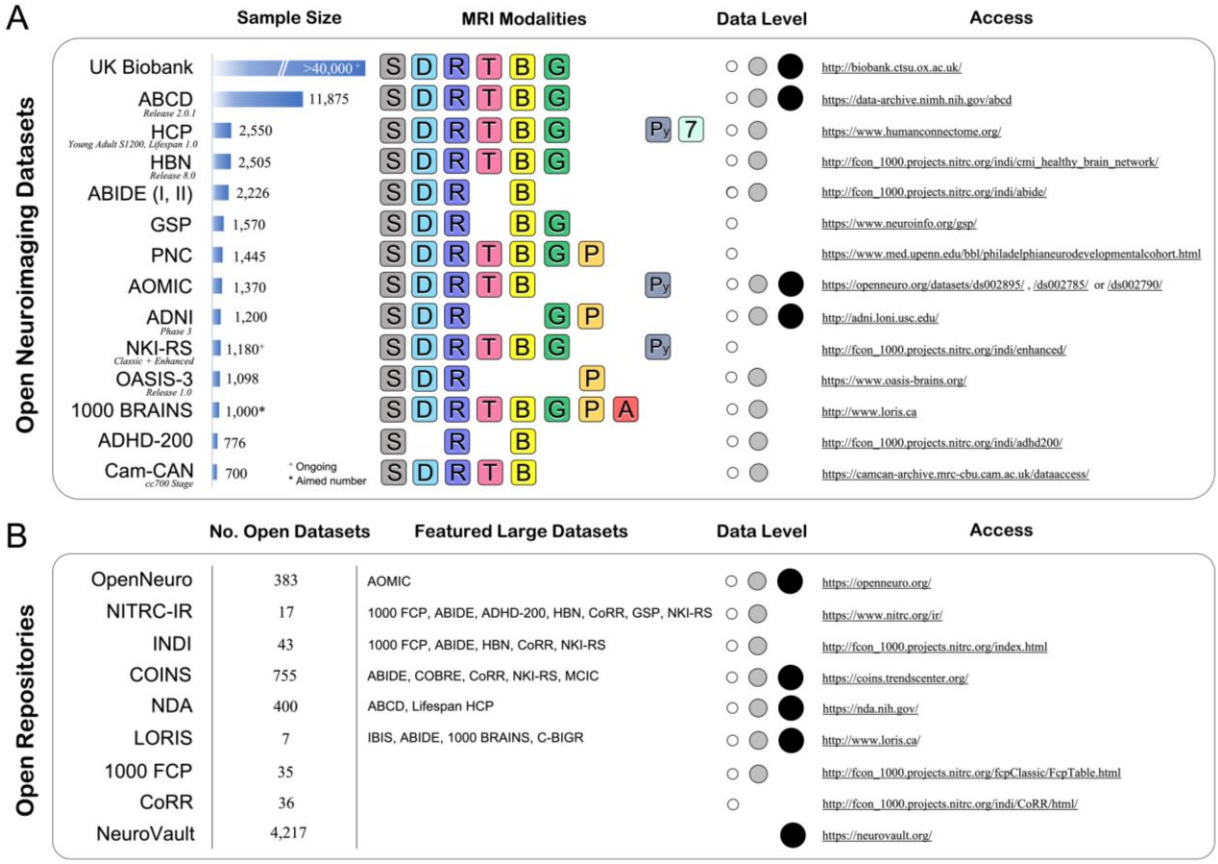

Figure 1. A list of large, open-source datasets and open repositories. A: For each dataset listed in the left-most column, sample size is indicated, along with the type of MRI data included ('MRI Modalities'). 'Data level' refers to the level of preprocessing: white circle = raw data; gray circle = some level of preprocessed data; black = processed data (e.g. statistical maps, connectivity matrices, etc.). A URL is included ('Access') for each sample. B: For each open repository listed in the left-most column, an estimate of the number of open datasets is listed. Datasets of particular interest are highlighted ('Featured Large Datasets'). 'Data level' and 'Access' are provided as in A. Note that all sample sizes and the number of open datasets are current as of May 2020.

Abbreviations for datasets and repositories: '1000 BRAINS' = 1000 brains study[9]; '1000 FCP' = 1000 Functional Connectomes Project[10, 11]; 'ABCD' = Adolescent Brain Cognitive Development study[2]; 'ABIDE' = Autism Brain Imaging Data Exchange initiative[12, 13]; 'ADHD-200' = Attention Deficit Hyperactivity Disorder 200 sample[14]; 'ADNI' = Alzheimer's Disease Neuroimaging Initiative[15]; 'AOMIC' = Amsterdam Open MRI Collection[16]; 'Cam-CAN' = Cambridge Centre for Ageing Neuroscience[17, 18]; 'C-BIGR' = Clinical Biologic Imaging and Genetic Repository[19]; 'COBRE' = Center for Biomedical Research Excellence; 'COINS' = Collaborative Informatics and Neuroimaging Suite[20]; 'CoRR' = Consortium for Reliability and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Reproducibility[21]; 'GSP'= Brain Genomics Superstruct Project[22]; 'HBN' = Healthy Brain Network[23]; 'HCP' = Human Connectome Project[1]; 'IBIS' = Infant Brain Imaging Study; 'INDI' = International Neuroimaging Data-sharing Initiative[11]; 'LORIS' = Longitudinal Online Research and Imaging System[24]; 'MCIC' = MIND Clinical Imaging Consortium; 'NDA' = National Institute of Mental Health Data Archive; 'NITRC-IR' = NeuroImaging Tools & Resources Collaboratory Image Repository[25]; 'NKI-RS' = Nathan Kline Institute Rockland Sample[26]; 'OASIS-3' = Open Access Series of Imaging Studies[27]; 'PNC' = Philadelphia Neurodevelopmental Cohort[28].

Abbreviations for 'MRI Modalities': 'S' = structural MRI; 'D' = diffusion MRI; 'R' = resting-state fMRI; 'T' = task fMRI (including movie data); 'B' = behavioral, cognitive, or psychiatric measures; 'G' = genomics data; 'P' = perfusion MRI; 'A' = MR angiography; 'Py' = physiological data (respiratory, pulse recordings); '7' = 7T MRI.

	References
Part 1: Obtaining and managing data	Barron and Fox, 2015 (ref. [29]): Describes strengths and limitations of raw and processed imaging data Gorgolewski et al., 2016 (ref. [30]): Describes brain imaging data structure (BIDS)
Part 2: Getting to know your data	Alfaro-Almagro et al., 2020 (ref. [31]): Examination of confounds in the UK Biobank, along with recommendations for confound modelling in large datasets http://uc-r.github.io/gda Tips for exploring a new dataset, along with code and toy data
Part 3: Communicating results	Weston et al., 2019 (ref. [32]): Suggestions for analyzing pre-existing datasets Mennes et al., 2013 (ref. [11]); Poldrack and Gorgolewski, 2014 (ref [33]): Discussion of how and why to share data, along with issues and opportunities accompanying data sharing
Further reading	Milham and Klein, 2019 (ref. [34]): Practical suggestions for practicing open science Nowogrodzki, 2020 (ref. [35]): Tips from a variety of fields for working with large datasets

Table 1. Key references and resources for working with large, publically available datasets.

Part 1. Obtaining and managing data

In the first section, we discuss obtaining and managing large datasets. Careful planning can help ensure that preprocessing and analysis goes smoothly, saving time in the future.

Obtaining data: identify research questions

Given that large, open-source datasets consist of many different types of data, the first step is identifying the dataset that can address a study's question of interest. Most large datasets have some combination of imaging, genetic, and behavioral data (Figure 1) that may not be harmonized across different datasets. Some may include specific clinical populations and/or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

related measures. To more robustly address the research question, a researcher may leverage multiple datasets to bolster sample sizes (i.e., for a rare subset of the data or for participants with a rare disease) or demonstrate reproducibility of findings across samples. Whatever the intended use, giving careful thought to the scientific question at hand will help focus the researcher and identify which types of data are needed. Once a question and dataset are identified, researchers should consult with their local Institutional Review Board (IRB) and/or Human Investigation Committee (HIC) before proceeding, as human research exceptions or data sharing agreements may be needed.

Managing data: what to download

Typically, neuroimaging data is released in two formats: primary, raw data in the form of Digital Imaging and Communications in Medicine (DICOM) or Neuroimaging Informatics Technology Initiative (NIfTI) images; or some form of processed data (e.g. connectivity matrices or activation maps). Both types of data possess strengths and limitations[29].

Disk storage. The first difference between raw and processed imaging data is the amount of disk storage required. When a sample comprises thousands of subjects, storage of raw data can become a challenge. For example, in the ABCD[2] dataset, the raw data in NifTI format takes up ~1.35 GB per subject, or ~13.5 TB for the entire first release of ~10,000 subjects. Note that this is simply the NifTI data—this does not include space that will be needed to store intermediate files, processed data, or results. On the other hand, processed data from ABCD, such as preprocessed connectivity matrices[36], would only require ~25.6 MB amount of disk space, approximately 0.0001 percent of the space needed to store the NifTI images and intermediate files needed to generate connectivity matrices starting from the raw data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

One may elect to use local or cloud storage, depending on the funds available, security needs, ability and intent to process data on the cloud, accessibility needs, etc. Finally, these storage estimates do not include the need to back up the data, which will typically double the amount of storage needed. To decrease the backup volume, certain intermediate files (e.g. skull stripped images) may be excluded from the backup. Further, some may choose not to back up already processed data, as these data can simply be re-downloaded.

Time. When choosing what data to download, it is important to consider the amount of time that will need to be invested in obtaining the data, as well as the additional time it will take to process the raw data. For instance, when obtaining raw data from thousands of subjects, it can take weeks to download the DICOM data. Depending on the computational resources at hand, converting the data into NifTI format can similarly take weeks. Coupled with the amount of time it takes to skull-strip subjects' anatomical images, register them into common space, motion correct functional images, and perform quality control (QC), in our experience, it can take 6-9 months for 2-3 researchers to download, process, and prepare the data for analysis. (It should be noted that this is still far less time than it would take for a site to generate such a large dataset on its own.) Alternatively, some databases include already processed data; in principle, these are ready for use immediately. However, we still recommend performing basic QC steps on processed data before analysis begins.

Flexibility. With the amount of time it takes to process raw data, one might ask: why go through all this trouble instead of simply downloading the processed data? The main answer is that by choosing to use processed data, one is tied to preprocessing decisions that were made to generate the processed data, which may not suit a particular study. For instance, in the ABCD dataset, network-based connectivity matrices were released[36] using network definitions from the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Gordon atlas[37]. If a researcher wanted to test the generalizability of their results to the choice of parcellation, it would not be possible with only the processed ABCD data. On the other hand, by having access to the raw data, it is straightforward to generate matrices with different parcellations. Given the impact of analytic flexibility on results[38], this idea holds for other preprocessing steps as well: the impact of different motion artifact removal pipelines could be assessed[39], the effect of region size on behavioral prediction accuracy could be investigated[40], and so on. Many datasets have multiple forms of processed data available (e.g. functional connectivity data with and without global signal regression), so documentation should be investigated to see what is available and if this coincides with analysis goals.

Organizing and keeping track of what was done to the data

Once data are obtained, efficient management of data is key. If using raw data, it is becoming the norm to organize data according to the brain imaging data structure (BIDS[30]; for help with BIDS, see https://github.com/bids-standard/bids-starter-kit). Regardless of how data are stored and managed, documentation is essential. Keeping track of what was done to the data, how it was done (i.e. what code and software were used; see "Part 3: Communicating results" for tips on sharing code), who performed each step, and the motivation for each choice should be documented. As in other areas of research, the aim of documentation is that a knowledgeable researcher within the field should be able to exactly recreate the workflow that is described. Although exact formats may vary by lab and by needs, examples include keeping track of progress in shared Google Docs and Jupyter notebooks. In addition, using a platform like Slack can facilitate communication between project members and might be useful for some teams. Whatever the method, a record needs to be accessible to others who will use the data in the future, and it is helpful to avoid jargon—a point especially pertinent given that junior personnel,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

who are often responsible for obtaining and managing the data, have a high turnover rate as they progress with their training. While these steps take time to implement, careful organization and documentation saves time in the long run when performing analyses and writing up results.

Obtaining and managing data: closing thoughts

Check for updates. Investigators need to regularly check for updates to a dataset—it is not enough to simply download the data and forget about it. Besides checking for new data releases, other important information is released—whether it is scanner updates, different preprocessing pipelines, or QC issues that were noticed and corrected. In addition, it is not unusual for data collection sites to discover errors in acquisition or processing that could significantly impact downstream findings (see "Part 2: Getting to know your data" for issues to be on the lookout for). Each large dataset typically has a QC wiki, a forum where issues can be discussed, or an email list that users can subscribe to. In the case of the UK-Biobank and ABCD, there are research staff members dedicated to help investigators as issues arise. It is important that researchers utilize these resources frequently.

Team up when possible. If multiple labs at the same institution are interested in the same dataset, working together to download, manage, and store the data helps to reduce duplicate efforts, saving time and resources. Team members can work together to handle different aspects of the workload. However, sharing between labs across different institutions can be more difficult, as privacy laws and other regulations can vary by institute, region, or country. A researcher's local IRB and/or HIC should be consulted when sharing curated data across labs. Whatever the solution, the point is to work together and be collaborative whenever possible. At the same time, it is particularly important to be prudent with write permissions (e.g., read-only is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

sufficient for team members performing visual inspection of skull stripping results); while raw data can always be re-downloaded and re-processed, it can be unpleasant, to say the least.

Ask questions. As noted above, each large dataset typically has mechanisms where problems can be explained and potential solutions can be offered (i.e. forums, a contact person dedicated to QC issues, etc.). In addition, social media platforms (e.g., Twitter) are increasingly popular for obtaining advice from colleagues for using large datasets. Whatever the resource, asking questions (and making the solutions known to the community) is an essential part of working with any open resource, including large datasets.

Part 2. Getting to know your data

Once all data are downloaded, the next step is becoming acquainted with the raw data. This is particularly important when using large, open-source datasets—as these data have not been collected by the end-user, it may be easier for the user to overlook subtle issues.

Demographic and participant factors

The first factors that should be considered are sample demographics and other basic participant attributes. Depending on the analyses planned, one should investigate factors like age, sex, race, and family structure within the dataset. In addition, datasets like the Autism Brain Imaging Data Exchange (ABIDE) samples[12, 13] and ABCD[2] are comprised of multiple sites, so this step enables users to understand characteristics of the data collected at various sites in order to plan analyses that account for potential site effects and/or generalizability of results across sites[41, 42, 43]. Given that potentially uninteresting sources of variance can be amplified in large datasets, other possible factors—like smoking status, the time of day a participant was scanned[44], or the time of year a participant was scanned—could be explored to determine if they might act as confounds.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Imaging measures

Imaging basics. After considering sample demographics and other participant characteristics, the next step is getting to know the imaging data. To start, researchers should determine which participants have complete scans that are needed for a given analysis. For example, some participants may have had scans cut short for technical reasons, some may have multiple scans (i.e. if a scan had to repeated to obtain quality data), etc. The scanner type, software, and acquisition parameters that were used should also be considered, as sometimes scanner software is updated during a study[22]. Scanning site has also been shown to introduce systematic bias into measures of functional connectivity, especially for multivariate analyses[45], as has scanner manufacturer[46]. Further, general aspects of study design should be taken into account: it should be noted if all scans for a participant were conducted on the same day (as in the Philadelphia Neurodevelopmental Cohort (PNC)[28]), or if the scans were split into back-to-back days (as in HCP)[1], in addition to if the scan/task order was counterbalanced or fixed across subjects.

Task-specific factors. Most of the datasets mentioned have released task-based functional scans; these data must be thoroughly investigated before use. In our own experience, in the HCP S900 release, we observed that at least 30 subjects had a different block order in the working memory task during the RL run than that reported for a majority of the other subjects. Possible discrepancies in task timing should be examined as well. In the emotion task in HCP, a bug in the E-prime scripts resulted in the last block ending prematurely for some subjects. Nevertheless, the task regressors released do not reflect this incongruity (https://protocols.humanconnectome.org/HCP/3T/task-fMRI-protocol-details.html). In addition, issues with the stop-signal task have also recently been reported in the ABCD sample, including

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

different durations of stimuli across trials and stimuli occasionally not being presented[47]. While none of these discrepancies preclude using the data per se (though analyses might have to be adapted considerably), we use these as examples of possible issues to be on the lookout for.

In addition, there are potential differences in similar tasks across datasets. For instance, many datasets have a working memory task (Table 2). In the HCP, a 2- and 0-back paradigm was used with places, faces, tools, and body parts as stimuli[48], whereas in the PNC, 0-, 1-, and 2-back conditions were used with fractals as stimuli[28, 49]. Along with other differences in task design (duration of blocks, other timing parameters, etc.), these must be kept in mind when planning analyses and when comparing results to those obtained in other samples.

	HCP	ABCD	PNC
Type	0-back, 2-back	Emotional 0-back, 2-back	0-back, 1-back, 2-back
Stimuli	places, faces, tools and body parts	happy, fearful and neutral facial expressions; place stimuli	fractals
Run duration	5 min	5 min	11.6 min
Task cue at start of each block	2.5 s	2.5 s	9 s
# of task blocks/run	8 × 25 s/block (4 for each n-back)	8 × 25 s/block (4 for each n-back)	9 x 60 s/block (3 for each n-back)
# of trials/block	10 × 2.5 s/trial	10 × 2.5 s/trial	20 × 3 s/trial
Target to non-target trials ratio	1:5	1:5	1:3
Each trial	2 s stimulus + 0.5 s ITI	2 s stimulus + 0.5 s ITI	0.5 s stimulus + 2.5s ITI
# of fixation blocks/run	4 × 15 s/block	4 × 15 s/block	3 × 24 s/block
Reference	Barch et al., 2013 (ref. [48])	Casey et al., 2018 (ref. [2])	Satterthwaite et al., 2014 (ref. [28])

Table 2. Differences in working memory task across datasets. Abbreviations of datasets as in Fig. 1.

Behavioral measures

Another important category of data with which researchers should familiarize themselves are measures collected outside of the scanner, which we refer to as "behavioral measures." Specifically, we are referring to participant measures obtained beyond demographic information (i.e. performance on cognitive tests, self-report measures, or clinician assessments).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

Within dataset differences. Measures within a dataset may differ. For instance, different versions of the autism diagnostic observation schedule (ADOS) were released by different sites in the ABIDE samples. In addition, only some sites had the ADOS administered by research certified clinicians—the gold standard for multisite reliability in diagnosing ASD[50, 51]. Both of these factors could introduce unaccounted for variance into the sample. In the same dataset, different instruments, and versions of instruments, to assess full scale IQ were collected at different sites—some used the Wechsler Adult Intelligence Scale (WAIS), some used the Differential Abilities Scale (DAS), while others used different versions of the Wechsler Intelligence Scale for Children (WISC)[12, 13].

Between dataset differences. Measure across different datasets may differ as well. For instance, in the HCP dataset, fluid intelligence was measured using a 24-item version of the Penn Progressive Matrices assessment[48], whereas in the PNC dataset, both a 24- and 18-item version of the Penn Matrix Reason Test were used[28, 49]. As with task design, these differences must be acknowledged when interpreting previous findings or planning future analyses—specifically when trying to use specific datasets as validation samples[52]. In addition, it should be noted that multiple measures can typically be reported for each behavioral scale—a raw score, a standardized score, scores on specific subscales, etc.—so it is important to ensure that one is using the behavioral score that is intended.

Getting to know your data: closing thoughts

We encourage investigators to calculate descriptive statistics, visualize distributions, and explore bivariate—or even multivariate—associations of the variables in a dataset. Additionally, outliers, higher leverage data points (i.e. a data point with an extreme predictor value), and missing data should be identified. (See http://uc-r.github.io/gda for examples of factors to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

investigate, as well as R packages and toy data.) All of these steps can help detect potential issues with the data that might preclude planned analyses. If potential issues are found, steps should be taken to address them. Exact solutions will differ depending on analysis goals[31, 53, 54, 55, 56, 57], and other resources exist to understand confounds in more detail[31]. Nevertheless, the main point of this section is that getting to know all aspects of an open-source dataset and how it was acquired is key, especially as an end-user who did not collect the data.

Part 3. Communicating results

The last phase of working with large datasets is reporting and sharing results. In addition, it might be appropriate for researchers to share processed data at the conclusion of their study.

What to report

Ideally, a manuscript should include all needed details for another researcher in the field to reproduce the work. A good start is the Committee on Best Practices in Data Analysis and Sharing (COBIDAS) guidelines for reporting neuroimaging methods, which includes both "mandatory" and "not mandatory" recommendations[58]. When working with big data, some of this information may have been reported elsewhere. It can be cumbersome to repeat this information in every manuscript, so it may be sufficient to include a reference to the original studies following the guidelines established by the creators of the database. When taking this route, we also advise researchers to include a brief summary of critical details to facilitate comprehension by reviewers and readers.

Data release. To ensure transparency, the data release version should be reported. Similar to software releases, datasets will be updated to include new subjects, new preprocessing pipelines, or fixes for QC issues (see "Part 1: Obtaining and managing data"). Reporting is straightforward when data are released as discrete packages with specific names (i.e. the HCP

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

1200 Subjects Data Release*)*. For data released in a continuous fashion (i.e. the ABCD Fast Track Data releases new subjects monthly), reporting when the data were obtained will allow other researchers to see how results fit into the context of previous findings using the same dataset. If details about the data release are less clear, as much information should be provided as possible, including the date the dataset was downloaded, the number of subjects downloaded, and a URL detailing the location of the release. In addition, when there are multiple releases available (i.e. 900 Subjects HCP release, 1200 Subjects HCP release, etc.), we recommend that the most recent release should be downloaded to ensure the highest quality data are used, as well as the highest number of subjects. However, if older releases are utilized, reasons for doing so should be reported.

Subject IDs. Reporting subject IDs of the participants used, as well as those excluded from final analyses (and reasons for exclusion), can help aid transparency. This information can be included in supplementary material. It should be noted, however, that datasets often have different systems regarding subject IDs. Some datasets have IDs that are consistent across all downloads (e.g. ABIDE) and straightforward to share with others, whereas other datasets have unique IDs generated for each group working with the sample (e.g. UK Biobank). In addition, data usage agreements (DUA) for each dataset often dictate what can and cannot be published in a manuscript. Researchers should check their DUA to determine if publishing subject IDs is allowed.

What to share

There has been an increased push to share resources among the neuroimaging community in recent years, and open-source datasets are a prime example of how sharing has accelerated progress in the field[59]. Hence, users of large datasets should pay it forward by sharing materials

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

related to their study, which will further help progress and allow other researchers to attempt to replicate and extend their findings.

Processed forms of data. Similar to subject IDs, researchers should check their DUA to determine if sharing processed forms of data is allowed (i.e. skull-stripped anatomical images, motion-corrected images, connectivity matrices, etc.). For example, when accessing data through the Consortium for Reliability and Reproducibility (CoRR)[21], once a user has registered, they can share all forms of data with other labs. On the other hand, datasets like ABCD require all users who interact with the data be approved and listed on the DUA. In this case, sharing with others would necessitate that the researchers being given data are approved in advance. Some datasets, like the HCP, stipulate that derivative data be shared only if it is impossible to infer anything about any particular participant from the data. Before sharing data, researchers should consult with their local IRB and/or HIC. When sharing data with the larger community is appropriate, there are many options to do so (Table 3). Specialized tools have been developed to facilitate working with many of these datasets (e.g. DataLad, https://www.datalad.org/datasets.html and OpenNeuro[60, 61]).

Data Level	Available Repositories								
	BALSA	**COINS**	**INDI**	**LORIS**	**NDA**	**Neuro-Vault**	**NITRC-IR**	**OMEGA**	**Open Neuro**
Primary, Raw Data		Y	Y	Y	Y		Y	Y	Y
Preprocessed Data	Y	Y		Y	Y			Y	
Derived, Statistical Parametric Data	Y			Y	Y	Y			

Table 3. A sampling of online data repositories available for sharing different levels of data. Abbreviations for datasets and repositories: 'BALSA' = Brain Analysis Library of Spatial maps and Atlases[62]; 'COINS' = Collaborative Informatics and Neuroimaging Suite[20]; 'INDI' = International Neuroimaging Data-sharing Initiative[11]; 'LORIS' = Longitudinal Online Research

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

and Imaging System[24]; 'NDA' = National Institute of Mental Health Data Archive; 'NITRC-IR' = NeuroImaging Tools & Resources Collaboratory Image Repository[25]; 'OMEGA' = Open MEG Archive[63].

Results. When possible, we also advocate for sharing aspects of results that might not be included in manuscripts. Unthresholded statistical maps, as well as parcellations, can be shared via NeuroVault[64]. If performing a predictive modelling study, there is currently no standard for sharing. However, Python's pickle protocols (https://docs.python.org/3/library/pickle.html), JavaScript Object Notation (JSON) files, and MATLAB's MAT-files are popular options. Once converted to these file formats, models can be shared via GitHub (e.g. https://github.com/canlab/Neuroimaging_Pattern_Masks).

Code. With the availability of online platforms such as GitHub, sharing code has become straightforward. Ideally, all code used for preprocessing and analysis should be shared, and a link to a project repository should be included in each manuscript. It is necessary to keep code well-documented and well-structured. This includes adding proper readme files, adding comments to the code describing what is being done, maintaining a well-structured project repository, and regularly checking and fixing "open issues" (i.e. bugs). Some useful resources can be found in GitHub Guides (https://guides.github.com/) or by following the standards adopted by popular open-source projects, such as scikit-learn (https://github.com/scikit-learn/scikit-learn)[65, 66].

Reproducible inference

Effect sizes and P-values. When writing up the results of a study, it is also important to keep in mind some of the statistical issues associated with common null hypothesis statistical testing using large datasets. (For a deeper discussion, see ref.[4].) While a large number of subjects permits a closer estimate of how sample effect sizes map onto true population effect sizes[67, 68, 69], even small effects with potentially little practical importance can be "statistically significant."

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

For instance, in the UK Biobank sample (n = 14,500), a correlation of $r = 0.017$ would be considered significant at $P < 0.05$. Hence, such findings must be interpreted with caution, particularly when relying on a single *P*-value to determine significance[70, 71]. Reporting multiple lines of converging evidence—whether through the use of effect sizes or Bayesian analyses, in addition to *P*-values—will help determine the practical significance of a given result. See refs. [72, 73, 74, 75] for more on alternatives to *P*-values.

Reporting negative results. Finally, negative results can be particularly informative when derived from large datasets. Much has been written about the importance of publishing null findings and how the literature can be skewed by not doing so[76, 77, 78, 79, 80, 81]. Because of the statistical power associated with large datasets, reporting such negative results can help clear up potentially conflicting effects obtained with smaller samples. Reporting negative findings can also save time and reduce duplicate efforts as other labs may be planning similar analyses.

Communicating results: closing thoughts

When communicating results from large datasets, transparency is essential. Clearly reporting what version of the dataset was downloaded, which subjects were used in analyses, and the practical significance of associations should drive what is included in manuscripts. Sharing materials is a key step as well and should be performed wherever possible.

Emerging issues and final thoughts

We close with arising issues with large datasets to alert first time users to these potential concerns. The first issue is known as data decay, or the fact that having multiple investigators analyze the same dataset inadvertently increases the number of false positives. This problem increases as the number of researchers analyzing the data increases[82]. In essence, the utility of the dataset decays as the number of users increases. A related notion has been advanced before:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

it has been suggested that a lack of generalizability might begin to be seen in the Alzheimer's Disease Neuroimaging Initiative (ADNI)[15] dataset, given that more and more Alzheimer's disease researchers based their conclusions on the same data[4] and results began to become overfit to sample noise. The issue of overfitting is well-known to the machine learning community and is discussed elsewhere[83, 84, 85, 86].

Because of issues like data decay and a potential for decreasing generalizability, continuing to collect new data—that might be of smaller size than the samples highlighted here—is essential. Generating new datasets, with varied characteristics, and sharing them can help ensure conclusions are not based on idiosyncratic quirks of samples[4]. Environments will continue to change and evolve—from the exposures affecting an individual to the way they interact with technology.

Also, conducting a smaller-scale study allows unique training opportunities for younger personnel. Taking part in the data collection process can provide a fuller appreciation of neuroimaging as a whole, from strengths of the technique to potential weaknesses. Finally, smaller samples can also be contributed to larger consortiums and become a part of the big data ecosystem—indeed, efforts like ABIDE, the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium[87], and the International Neuroimaging Data-sharing Initiative (INDI)[11] have taken this approach to much success. Hence, we need to continue to collect datasets, large and small, to ensure results are generalizable and also to ensure that neuroimagers are studying factors relevant to society at large.

Conclusion

The use of large datasets is becoming more and more common in human neuroimaging. While these datasets can be a powerful resource, their use introduces new issues that must be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

considered. We have detailed practical tips that investigators can use as they download and manage their data, potential confounds to be aware of, as well as what to share when communicating results. Careful consideration of the many challenges associated with these datasets and ways to deal with these issues will allow researchers the chance to make new discoveries and push forward our understanding of the human brain.

Acknowledgements

The authors acknowledge funding from the following sources: CH and ASG, T32GM007205; SMN, K00MH122372; KL, R01MH111424 and P50MH115716; DSB, T32 MH019961 and R25 MH071584; and DS, R24 MH114805.

References

1. Van Essen DC*, et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62-79 (2013).

2. Casey BJ*, et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci* **32**, 43-54 (2018).

3. Miller KL*, et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523-1536 (2016).

4. Smith SM, Nichols TE. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron* **97**, 263-268 (2018).

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

6. Bzdok D, Nichols TE, Smith SM. Towards Algorithmic Analytics for Large-scale Datasets. *Nat Mach Intell* **1**, 296-306 (2019).

7. Bzdok D, Yeo BTT. Inference in the age of big data: Future perspectives on neuroscience. *Neuroimage* **155**, 549-564 (2017).

8. Fan J, Han F, Liu H. Challenges of Big Data Analysis. *Natl Sci Rev* **1**, 293-314 (2014).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

9. Caspers S, *et al.* Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Front Aging Neurosci* **6**, 149 (2014).

10. Biswal BB, *et al.* Toward discovery science of human brain function. *Proc Natl Acad Sci U S A* **107**, 4734-4739 (2010).

11. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: the FCP/INDI experience. *Neuroimage* **82**, 683-691 (2013).

12. Di Martino A, *et al.* Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data* **4**, 170010 (2017).

13. Di Martino A, *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* **19**, 659-667 (2014).

14. Consortium HD. The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. *Front Syst Neurosci* **6**, 62 (2012).

15. Mueller SG, *et al.* Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement* **1**, 55-66 (2005).

16. Snoek LvdM, M.M; Beemsterboer, T.; van der Leij, A.; Eigenhuis, A.; Scholte, H.S. The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *bioRxiv*, (2020).

17. Shafto MA, *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol* **14**, 204 (2014).

18. Taylor JR, *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* **144**, 262-269 (2017).

19. Das S, *et al.* Cyberinfrastructure for Open Science at the Montreal Neurological Institute. *Front Neuroinform* **10**, 53 (2016).

20. Scott A, *et al.* COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Front Neuroinform* **5**, 33 (2011).

21. Zuo XN, *et al.* An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci Data* **1**, 140049 (2014).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

22. Holmes AJ, *et al.* Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Sci Data* **2**, 150031 (2015).

23. Alexander LM, *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data* **4**, 170181 (2017).

24. Das S, Zijdenbos AP, Harlap J, Vins D, Evans AC. LORIS: a web-based data management system for multi-center studies. *Front Neuroinform* **5**, 37 (2011).

25. Luo XZ, Kennedy DN, Cohen Z. Neuroimaging informatics tools and resources clearinghouse (NITRC) resource announcement. *Neuroinformatics* **7**, 55-56 (2009).

26. Nooner KB, *et al.* The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Front Neurosci* **6**, 152 (2012).

27. LaMontagne PJB, T.L.S.; Morris, J.C.; Keefe, S.;, Hornbeck RX, C; Grant, E.; Hassenstab, J.; Moulder,, K.; Vlassenko AGR, M.E.; Cruchaga, C.; Marcus, D. . OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, (2019).

28. Satterthwaite TD, *et al.* Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage* **86**, 544-553 (2014).

29. Barron DSF, P.T. BrainMap Database as a Resource for Computational Modeling. *Brain Mapping: An Encyclopedic Reference* **1**, 675-683 (2015).

30. Gorgolewski KJ, *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).

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

32. Weston SJ, Ritchie SJ, Rohrer JM, Przybylski AK. Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. *Adv Methods Pract Psychol Sci* **2**, 214-227 (2019).

33. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. *Nat Neurosci* **17**, 1510-1517 (2014).

34. Milham MP, Klein A. Be the change you seek in science. *BMC Biol* **17**, 27 (2019).

35. Nowogrodzki A. Eleven tips for working with large data sets. *Nature* **577**, 439-440 (2020).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

36. Hagler DJ, Jr.*, et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage* **202**, 116091 (2019).

37. Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb Cortex* **26**, 288-303 (2016).

38. Botvinik-Nezer R*, et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84-88 (2020).

39. Ciric R*, et al.* Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* **154**, 174-187 (2017).

40. Dadi K*, et al.* Benchmarking functional connectome-based predictive models for resting-state fMRI. *Neuroimage* **192**, 115-134 (2019).

41. Lake EMR*, et al.* 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 (2019).

42. Pomponio R*, et al.* Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).

43. Sripada C*, et al.* Prediction of neurocognition in youth from resting state fMRI. *Mol Psychiatry*, (2019).

44. Orban C, Kong R, Li J, Chee MWL, Yeo BTT. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS Biol* **18**, e3000602 (2020).

45. Noble S*, et al.* Multisite reliability of MR-based functional connectivity. *Neuroimage* **146**, 959-970 (2017).

46. Marek S*, et al.* Identifying reproducible individual differences in childhood functional brain networks: An ABCD study. *Dev Cogn Neurosci* **40**, 100706 (2019).

47. Bissett PGH, M.P; Poldrack, R.A. A cautionary note on stop-signal data from the Adolescent Brain Cognitive Development [ABCD] study *bioRxiv*, (2020).

48. Barch DM*, et al.* Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169-189 (2013).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

49. Gur RC, *et al.* Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. *Neuropsychology* **26**, 251-265 (2012).

50. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).

51. Lord C, *et al.* A multisite study of the clinical diagnosis of different autism spectrum disorders. *Arch Gen Psychiatry* **69**, 306-313 (2012).

52. Greene AS, Gao S, Scheinost D, Constable RT. Task-induced brain state manipulation improves prediction of individual traits. *Nat Commun* **9**, 2807 (2018).

53. Duncan NW, Northoff G. Overview of potential procedural and participant-related confounds for neuroimaging of the resting state. *J Psychiatry Neurosci* **38**, 84-96 (2013).

54. Pervaiz U, Vidaurre D, Woolrich MW, Smith SM. Optimising network modelling methods for fMRI. *Neuroimage* **211**, 116604 (2020).

55. Rao A, Monteiro JM, Mourao-Miranda J, Alzheimer's Disease I. Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage* **150**, 23-49 (2017).

56. Snoek L, Miletic S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage* **184**, 741-760 (2019).

57. Chyzhyk D, Varoquaux, G., Thirion, B., Milham, M. Controlling a confound in predictive models with a test set minimizing its effect. *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1-4 (2018).

58. Nichols TE, *et al.* Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci* **20**, 299-303 (2017).

59. Milham MP, *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* **9**, 2818 (2018).

60. Poldrack RA, *et al.* Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform* **7**, 12 (2013).

61. Poldrack RA, Gorgolewski KJ. OpenfMRI: Open sharing of task fMRI data. *Neuroimage* **144**, 259-261 (2017).

62. Van Essen DC, *et al.* The Brain Analysis Library of Spatial maps and Atlases (BALSA) database. *Neuroimage* **144**, 270-274 (2017).

63. Niso G, *et al.* OMEGA: The Open MEG Archive. *Neuroimage* **124**, 1182-1187 (2016).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

64. Gorgolewski KJ*, et al.* NeuroVault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain. *Neuroimage* **124**, 1242-1244 (2016).

65. Fabian Pedregosa GV, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825−2830 (2011).

66. Pedregosa F*, et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

67. Lombardo MV, Lai MC, Baron-Cohen S. Big data approaches to decomposing heterogeneity across the autism spectrum. *Mol Psychiatry* **24**, 1435-1450 (2019).

68. Button KS*, et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365-376 (2013).

69. Szucs D, Ioannidis JP. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol* **15**, e2000797 (2017).

70. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". *Am Stat* **73**, 1-19 (2019).

71. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* **7**, 342-346 (2014).

72. Bzdok D, Ioannidis JPA. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci* **42**, 251-262 (2019).

73. Chen G, Taylor PA, Cox RW. Is the statistic value all we should care about in neuroimaging? *Neuroimage* **147**, 952-959 (2017).

74. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front Hum Neurosci* **11**, 390 (2017).

75. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat* **70**, 129-131 (2016).

76. Earp BD. The need for reporting negative results - a 90 year update. *J Clin Transl Res* **3**, 344-347 (2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 July 2020 doi:10.20944/preprints202007.0153.v1

77. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* **337**, 867-872 (1991).

78. Greenwald AG. Consequences of prejudice against the null hypothesis. *Psychological Bulletin* **82**, 1-20 (1975).

79. Heger M. Editor's inaugural issue foreword: perspectives on translational and clinical research. *J Clin Transl Res* **1**, 1-5 (2015).

80. Pautasso M. Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics* **85**, 193-202 (2010).

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

82. Thompson WH, Wright J, Bissett PG, Poldrack RA. Dataset decay and the problem of sequential analyses on open datasets. *Elife* **9**, (2020).

83. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11**, 2079-2107 (2010).

84. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Computmg Survevs* **27**, (1995).

85. Lawrence S, Giles CL. Overfitting and neural networks: Conjugate gradient and backpropagation. *Ieee Ijcnn*, 114-119 (2000).

86. Reunanen J. Overfitting in Making Comparisons Between Variable Selection Methods. *Journal of Machine Learning Research* **3**, 1371-1382 (2003).

87. Thompson PM*, et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav* **8**, 153-182 (2014).