Published in final edited form as:

Nat Hum Behav. 2021 February ; 5(2): 185–193. doi:10.1038/s41562-020-01005-4.

A hitchhiker's guide to working with large, open-source neuroimaging datasets

Corey Horien[1,2,✉], **Stephanie Noble**[3], **Abigali 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 M. R. Lake**[3], **R. Todd Constable**[1,3,5,8], **Dustin Scheinost**[1,3,5,9,10,✉]

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

[2]MD/PhD program, Yale School of Medicine, New Haven, CT, USA

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

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

[5]Department of Biomedical Engineering, Yale University, New Haven, CT, USA

[6]Summary Analytics Inc., Seattle, WA, USA

[7]Department of Computer Science, Yale University, New Haven, CT, USA

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

[9]Department of Statistics & Data Science, Yale University, New Haven, CT, USA

[10]Child Study Center, Yale School of Medicine, New Haven, CT, USA

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 lifecycle: from what to consider when downloading and storing the data to tips on how to become acquainted with

✉**Correspondence** should be addressed to C.H. or D.S. corey.horien@yale.edu; dustin.scheinost@yale.edu.

Author contributions

C.H. wrote the first draft of the manuscript. C.H., S.N., A.S.G., K.L., D.S.B., S.G., D.O'C., M.S., J.D., X.S., E.M.R.L., R.T.C. and D.S. contributed to the conceptualization, writing and editing of the manuscript. C.H., S.N., A.S.G., K.L., D.S.B., S.G., D.O'C., M.S., J.D., X.S., E.M.R.L., R.T.C. and D.S. read and approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-01005-4.

Peer review information Primary Handling Editor: Marike Schiffer

a dataset one did not collect and 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.

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 (Fig. 1; abbreviations for datasets provided in Supplementary Table 1)[4–28]. These initiatives build upon efforts dating back to the early twentieth century to collate large-scale brain datasets (for example, ref. [29]) and 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[30,31]. 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 participants and substantial imaging data per individual, simply downloading and storing the data can be difficult. The complex structure of these large datasets (for example, 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[30,32–34]. 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 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.

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.

Identifying 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, behavioural and other phenotypic data (Fig. 1) that may not be harmonized across different datasets. Some may include specific clinical populations and/or 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 to 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. At this stage, investigators may also wish to preregister their research question and analysis plans. In addition, it is important to consider the original purpose of the dataset, as it might influence the sort of questions that can be addressed, as well as the feasibility of using it in conjunction with other open-source datasets. Indeed, understanding the original purpose of a dataset can facilitate analyses, including in some cases analyses performed many years after the original data were collected[35].

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. We note that the ethics of data sharing are complex[22,36–38], standards are evolving[37–40], there are many models of what constitutes open data[22,38,41,42] and standards vary by institution and/or country. For example, participants in the UK Biobank can elect to have their data withdrawn and/or deleted at any point (https://www.ukbiobank.ac.uk/withdrawal/). On the other hand, participants can elect to have their data withdrawn from the National Institute of Mental Health Data Archive, but data already distributed to research teams would not be withdrawn (https://nda.nih.gov/about/policy.html). Standards and regulations will continue to evolve (for example, the General Data Protection Regulation recently enacted by the European Union; https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en), and issues associated with anonymization, participant consent and data sharing will continue to be refined; it is imperative researchers remain alert for possible changes in these issues with the large datasets they are using.

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 (for example, connectivity matrices or activation maps). Both types of data possess strengths and limitations (for more, see ref. [41]).

The first difference between raw and processed imaging data is the amount of disk storage required. When a sample comprises thousands of participants, 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 individual or ~13.5 TB for the entire first release of ~10,000 individuals. 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[43], would only require ~25.6 MB 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.

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 (for example, 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.

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 participants, 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 participants' 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 two or three 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 QC steps on processed data before analysis begins (see "Getting to know your data" below for the QC steps we discuss).

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[43] using network definitions from the Gordon atlas[44]. 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, having access to the raw data makes it straightforward to generate matrices with different parcellations. Given the impact of analytic flexibility on results[45], this idea holds for other preprocessing steps as well: the impact of different motion artefact removal pipelines could be assessed[46], the effect of region size on behavioural prediction accuracy could be investigated[47], and so on. Many datasets have multiple forms of processed data available (for example, 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. When using raw data, it is becoming the norm to organize data according to the brain imaging data structure (BIDS[48]; for help with BIDS, see https://github.com/bids-standard/bids-starter-kit). It may also be advisable to make the raw data files read-only, so that they cannot be inadvertently modified or deleted. Regardless of how data are stored and managed, documentation is essential. Keeping track of what was done to the data should also include documenting how it was done (i.e., what code and software were used; see the "Communicating results" section for tips on sharing code), who performed each step and the motivation for each choice. 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, 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.

Closing thoughts.

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: scanner updates, different preprocessing pipelines, QC issues that were noticed and corrected, etc. In addition, it is not unusual for data collection sites to discover errors in acquisition or processing that could significantly impact down-stream findings (see “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, research staff includes members dedicated to help investigators as issues arise. It is important that researchers utilize these resources frequently.

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, whether this involves formal collaborations with clearly delineated roles (and specified in grants, perhaps) or more informal working agreements that are still conducted in accordance with data usage agreements (DUA). For more on working together effectively in science, see ref. [49]. At the same time, it is particularly important to be prudent with write permissions (for example, read-only is sufficient for team members performing visual inspection of skull-stripping results); while raw data can always be re-downloaded and re-processed, this can be unpleasant, to say the least.

As noted above, each large dataset typically has a channel 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 (for example, 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.

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[10,11] and ABCD[2] comprise data collected at multiple sites, so this step enables users to understand characteristics of the data collected at various sites, allowing them to plan analyses that account for potential site effects and/or generalizability of results across sites[50–52]. ComBat is one method of removing between-site and between-sample effects and has been used in both structural and functional MRI analyses[51,53–56]. (See ref. [57] for a recent review of data harmonization of diffusion MRI data.) Given that potentially uninteresting sources of variance (i.e., variance unrelated to the question at hand) can be amplified in large datasets, other possible factors—like smoking status, the time of day a participant was scanned[58] or the time of year a participant was scanned—could be explored to determine whether they might act as confounds. The exact confounds investigated will depend on attributes of the dataset, as well as the reason the data were initially collected.

Imaging measures.

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[13]. Scanning site has also been shown to introduce systematic bias into measures of functional connectivity, especially for multivariate analyses[59], as has scanner manufacturer[60]. (For a full discussion of the quantitative effects of factors like site, software upgrades and changes in hardware in the UK Biobank, see ref. [61]) Further, general aspects of study design should be taken into account: it should be noted whether all scans for a participant were conducted on the same day (as in the Philadelphia Neurodevelopmental Cohort (PNC)[23]) or were split into back-to-back days (as in HCP)[1], in addition to whether the scan or task order was counterbalanced or fixed across participants. We note that tools designed with large datasets in mind are available to aid imaging QC (for example, https://mriqc.readthedocs.io/en/stable/)[62].

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 participants had a different block order in the working memory task during the right–left phase-encoding run than that reported for a majority of the other

participants. 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 participants. Nevertheless, the task regressors released do not reflect this incongruity (http://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 different durations of stimuli across trials and stimuli occasionally not being presented[63]. 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 two-back and zero-back paradigm was used with places, faces, tools and body parts as stimuli[64], whereas in the PNC, zero-, one- and two-back conditions were used with fractals as stimuli[23,65]. 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.

Behavioural measures.

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

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 autism spectrum disorder[66,67]. Both of these factors could introduce unaccounted-for variance into the sample. In the same dataset, different instruments, and versions of instruments, were used to assess full-scale intelligence quotient (IQ) at different sites—some used the Wechsler Adult Intelligence Scale, some used the Differential Abilities Scale, while others used different versions of the Wechsler Intelligence Scale for Children[10,11].

Measures 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[64], whereas in the PNC dataset, both 24- and 18-item versions of the Penn Matrix Reason Test were used[23,65]. 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[68]. In addition, it should be noted that multiple measures can typically be reported for each behavioural scale—a raw score, a standardized score, scores on specific subscales, etc.—so it is important to ensure that one is using the behavioural score that is intended.

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 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[61,69–72], and other resources exist to understand confounds in more detail[61]. Nevertheless, the main point of this section is that getting to know all aspects of an open-source dataset and how it was acquired are key, especially as an end-user who did not collect the data.

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 include both ‘mandatory’ and ‘not mandatory’ recommendations[39]. 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.

To ensure transparency, the data release version should be reported. Similarly to software releases, datasets will be updated to include new participants, new preprocessing pipelines or fixes for QC issues (see “Obtaining and managing data”). Reporting is straightforward when data are released as discrete packages with specific names (i.e., the HCP 1200 Subjects Data Release). For data released in a continuous fashion (i.e., the ABCD Fast Track Data releases data from new participants 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 participants downloaded and a URL detailing the location of the release. In addition, when there are multiple releases available (i.e., HCP 900 Subjects release, 1200 Subjects release, etc.), we recommend that the most recent release should be downloaded to ensure that the highest quality data are used, as well as the greatest number of participants. However, if older releases are used, reasons for doing so should be reported (i.e., when an issue has been discovered in the latest release).

Reporting participant IDs of the individuals 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 participant IDs. Some datasets have IDs that are consistent across all downloads (for example, ABIDE) and straightforward to share with others, whereas other datasets have unique IDs generated for each group working with the sample (for example, UK Biobank). In addition, DUAs for each dataset often dictate what can and cannot be published in a manuscript. Researchers should check their DUA to determine whether publishing participant 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[73]. Hence, users of large datasets should pay it forward by sharing materials related to their study, which will further help progress and allow other researchers to attempt to replicate and extend their findings.

As with participant IDs, researchers should check their DUA to determine whether sharing processed forms of data is allowed (for example, skull-stripped anatomical images, motion-corrected images, connectivity matrices, etc.). For example, when accessing data through the Consortium for Reliability and Reproducibility[28], once a user has registered, they can share all forms of data with other labs. On the other hand, datasets like ABCD require that 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). Shared data should be released with a clear license, so that other investigators know what restrictions are placed on reuse of the data, if any. Specialized tools have been developed to facilitate working with many of these datasets (for example, DataLad, https://www.datalad.org/datasets.html and OpenNeuro[21,22]).

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[12]. 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) and MATLAB's MAT-files are popular options. Platform-independent formats, such as JavaScript object notation (JSON) files and comma-separated value (CSV) files, can also be shared and do not tie investigators to the use of a specific programming language. Once converted to these file formats, models can be shared via GitHub (for example, https://github.com/canlab/Neuroimaging_Pattern_Masks).

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)[74].

Reproducible inference.

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. [30]). While a large number of participants permits a closer estimate of how sample effect sizes map onto true population effect sizes[75–77], even small effects with potentially little practical importance can be 'statistically significant'. 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[78,79]. Reporting multiple lines of converging evidence—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.[80–83] for more on alternatives to P-values.

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[84–89]. 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.

Closing thoughts.

When communicating results from large datasets, transparency is essential. Clearly reporting what version of the dataset was downloaded, which participants 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 remarks

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 analyse the same dataset inadvertently increases the number of false positives. This problem increases as the number of researchers analysing the data increases[90]. In essence, the utility of the dataset decays as the number of users increases. A related notion has been advanced before: it has been suggested that a lack of generalizability might begin to be seen in the Alzheimer's Disease Neuroimaging Initiative[19] dataset, given that more and more Alzheimer's disease researchers have based their conclusions on the same data[30] and results began to become overfit to sample noise. The issue of over-fitting is well-known to the machine learning community and is discussed elsewhere[91–93].

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[30]. 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[94] and the International Neuroimaging Data-sharing Initiative (INDI)[18] have taken this approach to much success. 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. An important consideration in this process will be properly crediting investigators who generated the original dataset. One solution is assigning a specific ID to each dataset, facilitating the citation of the original dataset generators, as well as potentially allowing easier searches for existing datasets[95]. Such a system would thus allow the original data generators to receive recognition for their contribution and encourage others to publically share data.

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 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors acknowledge funding from the following NIH grants: C.H. and A.S.G., T32GM007205; S.N., K00MH122372; K.L., R01MH111424 and P50MH115716; D.S.B., T32 MH019961 and R25 MH071584; and D.S., R24 MH114805. The funders had no role in the conception or writing of this manuscript.

References

1. Van Essen DC et al. The WU-Minn Human Connectome Project: an overview. Neuroimage 80, 62–79 (2013). [PubMed: 23684880]
2. Casey BJ et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54 (2018). [PubMed: 29567376]
3. Miller KL et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536 (2016). [PubMed: 27643430]

4. Alexander LM et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci. Data 4, 170181 (2017). [PubMed: 29257126]
5. Biswal BB et al. Toward discovery science of human brain function. Proc. Natl Acad. Sci. USA 107, 4734–4739 (2010). [PubMed: 20176931]
6. 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). [PubMed: 25071558]
7. HD-200 Consortium. The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Front. Syst. Neurosci. 6, 62 (2012). [PubMed: 22973200]
8. Das S et al. Cyberinfrastructure for open science at the Montreal Neurological Institute. Front. Neuroinform. 10, 53 (2017). [PubMed: 28111547]
9. 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 (2012). [PubMed: 22319489]
10. 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). [PubMed: 28291247]
11. 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). [PubMed: 23774715]
12. 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). Pt B. [PubMed: 25869863]
13. Holmes AJ et al. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Sci. Data 2, 150031 (2015). [PubMed: 26175908]
14. LaMontagne PJ et al. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. Preprint at medRxiv 10.1101/2019.12.13.19014902 (2019).
15. Luo XZ, Kennedy DN & Cohen Z Neuroimaging informatics tools and resources clearinghouse (NITRC) resource announcement. Neuroinformatics 7, 55–56 (2009). [PubMed: 19184562]
16. Marek K et al. The Parkinsons progression markers initiative (PPMI) - establishing a PD biomarker cohort. Ann. Clin. Transl. Neurol. 5, 1460–1477 (2018). [PubMed: 30564614]
17. Marek K et al. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). Prog. Neurobiol. 95, 629–635 (2011). [PubMed: 21930184]
18. Mennes M, Biswal BB, Castellanos FX & Milham MP Making data sharing work: the FCP/INDI experience. Neuroimage 82, 683–691 (2013). [PubMed: 23123682]
19. Mueller SG et al. Ways toward an early diagnosis in Alzheimers disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement. 1, 55–66 (2005). [PubMed: 17476317]
20. Nooner KB et al. The NKI-Rockland Sample: a model for accelerating the pace of discovery science in psychiatry. Front. Neurosci. 6, 152 (2012). [PubMed: 23087608]
21. Poldrack RA et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front. Neuroinform. 7, 12 (2013). [PubMed: 23847528]
22. Poldrack RA & Gorgolewski KJ OpenfMRI: Open sharing of task fMRI data. Neuroimage 144, 259–261 (2017). Pt B. [PubMed: 26048618]
23. Satterthwaite TD et al. Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage 86, 544–553 (2014). [PubMed: 23921101]
24. Scott A et al. COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. Front. Neuroinform. 5, 33 (2011). [PubMed: 22275896]
25. 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). [PubMed: 25412575]
26. Snoek L et al. The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. Preprint at bioRxiv 10.1101/2020.06.16.155317 (2020).
27. 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). Pt B. [PubMed: 26375206]

28. Zuo XN et al. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 1, 140049 (2014). [PubMed: 25977800]
29. Southard EE On the topographical distribution of cortex lesions and anomalies in dementia praecox, with some account of their functional significance. Am. J. Insanity 71, 603–671 (1915).
30. Smith SM & Nichols TE Statistical challenges in “Big Data” human neuroimaging. Neuron 97, 263–268 (2018). [PubMed: 29346749]
31. Noble S, Scheinost D & Constable RT Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. Neuroimage 209, 116468 (2020). [PubMed: 31852625]
32. Bzdok D, Nichols TE & Smith SM Towards algorithmic analytics for large-scale datasets. Nat. Mach. Intell. 1, 296–306 (2019). [PubMed: 31701088]
33. Bzdok D & Yeo BTT Inference in the age of big data: Future perspectives on neuroscience. Neuroimage 155, 549–564 (2017). [PubMed: 28456584]
34. Fan J, Han F & Liu H Challenges of big data analysis. Natl. Sci. Rev. 1, 293–314 (2014). [PubMed: 25419469]
35. Sandu AL, Paillere Martinot ML, Artiges E & Martinot JL 1910s’ brains revisited. Cortical complexity in early 20th century patients with intellectual disability or with dementia praecox. Acta Psychiatr. Scand. 130, 227–237 (2014). [PubMed: 24400850]
36. Brakewood B & Poldrack RA The ethics of secondary data analysis: considering the application of Belmont principles to the sharing of neuroimaging data. Neuroimage 82, 671–676 (2013). [PubMed: 23466937]
37. Meyer MN Practical tips for ethical data sharing. Adv. Methods Pract Psychol Sci. 1, 131–144 (2018).
38. White T, Blok E & Calhoun VD Data sharing and privacy issues in neuroimaging research: opportunities, obstacles, challenges, and monsters under the bed. Hum. Brain Map. 10.1002/hbm.25120 (2020).
39. Nichols TE et al. Best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci. 20, 299–303 (2017). [PubMed: 28230846]
40. Poline JB et al. Data sharing in neuroimaging research. Front. Neuroinform. 6, 9 (2012). [PubMed: 22493576]
41. Barron DS & Fox PT BrainMap Database as a Resource for Computational Modeling, in Brain Mapping: An Encyclopedic Reference (ed. Toga AW) 1, 675–683 (Elsevier, 2015).
42. Poldrack RA & Gorgolewski KJ Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517 (2014). [PubMed: 25349916]
43. Hagler DJ Jr. et al. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. Neuroimage 202, 116091 (2019). [PubMed: 31415884]
44. Gordon EM et al. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb. Cortex 26, 288–303 (2016). [PubMed: 25316338]
45. Botvinik-Nezer R et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88 (2020). [PubMed: 32483374]
46. 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). [PubMed: 28302591]
47. Dadi K et al. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage 192, 115–134 (2019). [PubMed: 30836146]
48. Gorgolewski KJ et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 3, 160044 (2016). [PubMed: 27326542]
49. Bennett LM & Gadlin H Collaboration and team science: from theory to practice. J. Investig. Med. 60, 768–775 (2012).
50. 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). [PubMed: 31010580]
51. Pomponio R et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 208, 116450 (2020). [PubMed: 31821869]

52. Sripada C et al. Prediction of neurocognition in youth from resting state fMRI. Mol. Psychiatry 10.1038/s41380-019-0481-6 (2019).
53. Fortin JP et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120 (2018). [PubMed: 29155184]
54. Fortin JP et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170 (2017). [PubMed: 28826946]
55. Yamashita A et al. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. 17, e3000042 (2019). [PubMed: 30998673]
56. Yu M et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp. 39, 4213–4227 (2018). [PubMed: 29962049]
57. Pinto MS et al. Harmonization of brain diffusion MRI: concepts and methods. Front. Neurosci. 14, 396 (2020). [PubMed: 32435181]
58. 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). [PubMed: 32069275]
59. Noble S et al. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970 (2017). [PubMed: 27746386]
60. Marek S et al. Identifying reproducible individual differences in childhood functional brain networks: an ABCD study. Dev. Cogn. Neurosci. 40, 100706 (2019). [PubMed: 31614255]
61. Alfaro-Almagro F et al. Confound modelling in UK Biobank brain imaging. Neuroimage 224, 117002 (2021). [PubMed: 32502668]
62. Esteban O et al. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One 12, e0184661 (2017).
63. Bissett PG, Hagen MP & Poldrack RA A cautionary note on stop-signal data from the Adolescent Brain Cognitive Development [ABCD] study. Preprint at bioRxiv 10.1101/2020.05.08.084707(2020).
64. Barch DM et al. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80, 169–189 (2013). [PubMed: 23684877]
65. 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). [PubMed: 22251308]
66. Fischbach GD & Lord C The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68, 192–195 (2010). [PubMed: 20955926]
67. Lord C et al. A multisite study of the clinical diagnosis of different autism spectrum disorders. Arch. Gen. Psychiatry 69, 306–313 (2012). [PubMed: 22065253]
68. Greene AS, Gao S, Scheinost D & Constable RT Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun. 9, 2807 (2018). [PubMed: 30022026]
69. 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). [PubMed: 22964258]
70. Pervaiz U, Vidaurre D, Woolrich MW & Smith SM Optimising network modelling methods for fMRI. Neuroimage 211, 116604 (2020). [PubMed: 32062083]
71. Rao A, Monteiro JM & Mourao-Miranda J Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150, 23–49 (2017). [PubMed: 28143776]
72. Snoek L, Miletić S & Scholte HS How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184, 741–760 (2019). [PubMed: 30268846]
73. Milham MP et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat. Commun. 9, 2818 (2018). [PubMed: 30026557]
74. Pedregosa F et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
75. Lombardo MV, Lai MC & Baron-Cohen S Big data approaches to decomposing heterogeneity across the autism spectrum. Mol. Psychiatry 24, 1435–1450 (2019). [PubMed: 30617272]
76. Button KS et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376 (2013). [PubMed: 23571845]

77. 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). [PubMed: 28253258]

78. Wasserstein RL, Schirm AL & Lazar NA Moving to a world beyond "$P < 0.05$". Am. Stat 73 Suppl. 1, 1–19 (2019).

79. 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). [PubMed: 25043853]

80. Bzdok D & Ioannidis JPA Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci. 42, 251–262 (2019). [PubMed: 30808574]

81. Chen G, Taylor PA & Cox RW Is the statistic value all we should care about in neuroimaging? Neuroimage 147, 952–959 (2017). [PubMed: 27729277]

82. Szucs D & Ioannidis JPA When null hypothesis significance testing is unsuitable for research: a reassessment. Front. Hum. Neurosci. 11, 390 (2017). [PubMed: 28824397]

83. Wasserstein RL & Lazar NA The AS As statement on *P*-values: context, process, and purpose. Am. Stat. 70, 129–133 (2016).

84. Earp BD The need for reporting negative results - a 90 year update. J. Clin. Transl. Res. 3, 344–347 (2017). Suppl 2. [PubMed: 30873480]

85. Easterbrook PJ, Berlin JA, Gopalan R & Matthews DR Publication bias in clinical research. Lancet 337, 867–872 (1991). [PubMed: 1672966]

86. Greenwald AG Consequences of prejudice against the null hypothesis. Psychol. Bull. 82, 1–20 (1975).

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

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

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

90. Thompson WH, Wright J, Bissett PG & Poldrack RA Dataset decay and the problem of sequential analyses on open datasets. eLife 9, e53498 (2020). [PubMed: 32425159]

91. Cawley GC & Talbot NLC On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107 (2010).

92. Dietterich T Overfitting and undercomputing in machine learning. ACM Comp. Surv. 27, 326–327 (1995).

93. Reunanen J Overfitting in making comparisons between variable selection methods, J. Mach. Learn. Res. 3, 1371–1382 (2003).

94. Thompson PM et al. Alzheimers Disease Neuroimaging Initiative, EPIGEN Consortium, IMAGEN Consortium, Saguenay Youth Study (SYS) Group. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 8, 153–182 (2014). [PubMed: 24399358]

95. Pierce HH, Dev A, Statham E & Bierer BE Credit data generators for data reuse. Nature 570, 30–32 (2019). [PubMed: 31164773]

96. 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). [PubMed: 32190814]

97. Milham MP & Klein A Be the change you seek in science. BMC Biol. 17, 27 (2019). [PubMed: 30914050]

98. Nowogrodzki A Eleven tips for working with large data sets. Nature 577, 439–440 (2020). [PubMed: 31932750]

99. Van Essen DC et al. The Brain Analysis Library of Spatial Maps and Atlases (BALSA) database. Neuroimage 144, 270–274 (2017). Pt B. [PubMed: 27074495]

100. Niso G et al. OMEGA: the open MEG archive. Neuroimage 124, 1182–1187 (2016). Pt B. [PubMed: 25896932]

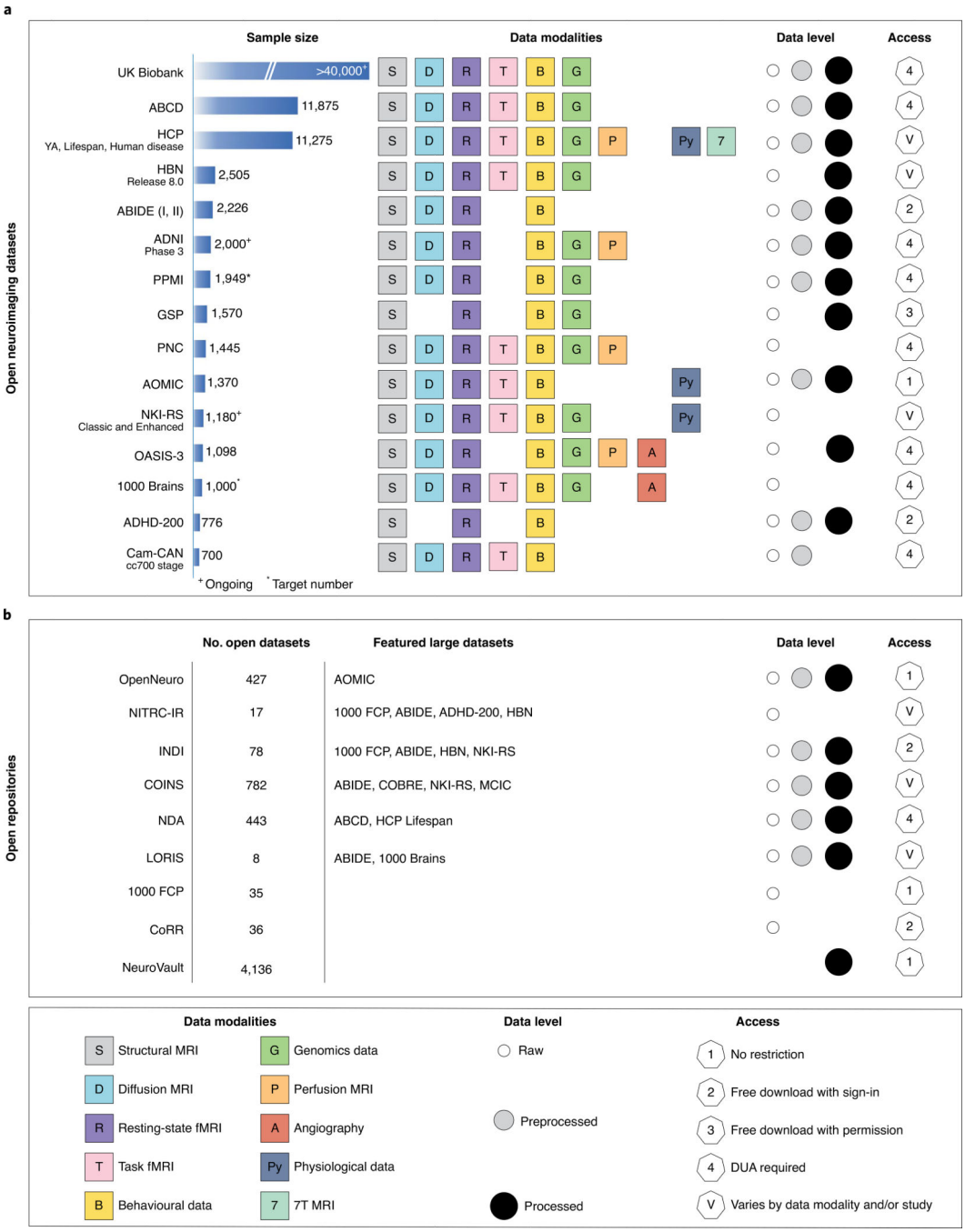

Fig. 1. A list of large, open-source datasets and open repositories.

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; grey 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 particular interest are highlighted ('Featured large datasets'). Sample sizes and the number of open datasets are current as of October 2020.

Users are encouraged to visit the website associated with each dataset before use, as sample sizes, access conditions, etc. may change. YA, HCP Young Adult study; COINS, Collaborative Informatics and Neuroimaging Suite[24]; LORIS, Longitudinal Online Research and Imaging System[9]; NITRC-IR, NeuroImaging Tools & Resources Collaboratory Image Repository[15]; NDA, National Institute of Mental Health Data Archive; ADNI, Alzheimer's Disease Neuroimaging Initiative; HBN, Healthy Brain Network; PPMI, Parkinson's Progression Markers Initiative; GSP, Brain Genomics Superstruct Project; AOMIC, Amsterdam Open MRI Collection; NKI-RS, Nathan Kline Institute Rockland Sample; OASIS-3, Open Access Series of Imaging Studies; ADHD-200, Attention Deficit Hyperactivity Disorder 200 sample; Cam-CAN, Cambridge Centre for Ageing Neuroscience dataset; 1000 FCP, 1000 Functional Connectomes Project; MCIC, MIND Clinical Imaging Consortium.

Table 1 |

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

	References
Obtaining and managing data	Barron and Fox[41]: describes strengths and limitations of raw and processed imaging data Gorgolewski et al.[48]: describes brain imaging data structure
Getting to know your data	Alfaro-Almagro et al.[61]: 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
Communicating results	Weston et al.[96]: suggestions for analysing pre-existing datasets Mennes et al.[18]; Poldrack and Gorgolewski[42]: discussions of how and why to share data, along with issues and opportunities accompanying data sharing
Further reading	Milham and Klein[97]: practical suggestions for practicing open science Nowogrodzki[98]: tips from a variety of fields for working with large datasets

Table 2 |

Differences in working memory task across datasets

	HCP	ABCD	PNC
Type	Zero-back, two-back	Emotional zero-back, two-back	Zero-back, one-back, two-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
No. of task blocks per run	8 × 25 s per block (4 for each *n*-back)	8 × 25 s per block (4 for each *n*-back)	9 × 60 s per block (3 for each *n*-back)
No. of trials per block	10 × 2.5 s per trial	10 × 2.5 s per trial	20 × 3 s per 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 + 2.5 s ITI
No. of fixation blocks per run	4 × 15 s per block	4 × 15 s per block	3 × 24 s per block
Reference	Barch et al.[64]	Casey et al.[2]	Satterthwaite et al.[23]

ITI, intertrial interval.

Table 3 |

A sampling of online data repositories available for sharing different levels of data

Data level	Available repositories								
	BALSA	**COINS**	**INDI**	**LORIS**	**NDA**	**NeuroVault**	**NITRC-IR**	**OMEGA**	**OpenNeuro**
Primary, raw data		Y	Y	Y	Y		Y	Y	Y
Preprocessed 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/
Reference	99	24	18	9	—	12	15	100	21,22

BALSA, Brain Analysis Library of Spatial maps and Atlases[99]; INDI, International Neuroimaging Data-sharing Initiative[18]; OMEGA, Open MEG Archive[100].