bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

OpenNeuro: An open resource for sharing of neuroimaging data

Christopher J. Markiewicz[1], Krzysztof J. Gorgolewski[1], Franklin Feingold[1], Ross Blair[1], Yaroslav O. Halchenko[2], Eric Miller[3], Nell Hardcastle[3], Joe Wexler[1], Oscar Esteban[1, 4], Mathias Goncalves[1], Anita Jwa[1], Russell A. Poldrack[1]

1. Department of Psychology, Stanford University, Stanford, CA, USA
2. Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH, USA
3. Squishymedia, Portland, OR, USA
4. Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

Abstract

The sharing of research data is essential to ensure reproducibility and maximize the impact of public investments in scientific research. Here we describe OpenNeuro, a BRAIN Initiative data archive that provides the ability to openly share data from a broad range of brain imaging data types following the FAIR principles for data sharing. We highlight the importance of the Brain Imaging Data Structure (BIDS) standard for enabling effective curation, sharing, and reuse of data. The archive presently shares more than 500 datasets including data from more than 18,000 participants, comprising multiple species and measurement modalities and a broad range of phenotypes. The impact of the shared data is evident in a growing number of published reuses, currently totalling more than 150 publications. We conclude by describing plans for future development and integration with other ongoing open science efforts.

Introduction

There is growing recognition of the importance of data sharing for scientific progress (National Academies of Sciences, Engineering, and Medicine et al., 2018). However, not all shared data are equally useful. The FAIR principles (Wilkinson et al., 2016) have formalized the notion that in order for shared data to be maximally useful, they need to be Findable, Accessible, Interoperable, and Reusable. An essential necessity for achieving these goals is that the data and associated metadata follow a common standard for organization, so that data users can easily understand and reuse the shared data. Here we describe the OpenNeuro data archive [RRID:SCR_005031], accessible at https://openneuro.org, which enables FAIR-compliant data sharing for a growing range of neuroscience data types (currently including magnetic resonance imaging [MRI], electroencephalography [EEG], magnetoencephalography [MEG], and positron emission tomography [PET]) through the use of a common community standard, the Brain Imaging Data Structure (BIDS) [RRID:SCR_016124] (Gorgolewski et al., 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Starting with early pioneering efforts by Gazzaniga and Van Horn to establish an fMRI Data Center in 1999 (Van Horn and Gazzaniga, 2013), data sharing has become well established in the domain of neuroimaging (Milham et al., 2018; Poldrack and Gorgolewski, 2014; Poline et al., 2012). A major impetus for the growth of data sharing was the International Neuroimaging Data Sharing Initiative (INDI) (Mennes et al., 2013), which published a landmark paper in 2010 (Biswal et al., 2010) demonstrating the scientific utility of a large shared resting fMRI dataset. The most prominent recent examples have been large-scale prospective data sharing projects, including the Human Connectome Project [HCP] (Van Essen et al., 2013), the NKI-Rockland sample (Nooner et al., 2012), Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018), and the UK Biobank (Littlejohns et al., 2020). These datasets have provided immense value to the field, and have strongly demonstrated the utility of shared data. However, their scientific scope is necessarily limited, given that each dataset includes only a limited number of imaging tasks and measurement types. Beyond these large focused data sharing projects, there is a “long tail” of smaller neuroimaging datasets that have been collected in service of specific research questions. Making these available is essential to ensure reproducibility as well as to allow aggregation across many different types of measurements in service of novel scientific questions. The OpenNeuro archive addresses this challenge by providing researchers with the ability to easily share a broad range of neuroimaging data types in a way that adheres to the FAIR principles.

Goals and principles

The OpenNeuro archive evolved from the OpenfMRI archive (Poldrack et al., 2013), which was focused solely on the sharing of task-based human fMRI data. Some of the principles behind OpenNeuro were inherited from OpenfMRI, whereas others grew out of our experiences in that project as well as from new developments in the domain of open science.

Minimal restrictions on sharing

There is a range of restrictiveness across data archives with regard to their data use agreements (Jwa and Poldrack, 2021). At one end of the spectrum are highly restricted databases such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI), which requires researchers to submit their scientific question for review and requires the consortium to be included as a corporate author on any publications. OpenNeuro represents the other pole of restrictiveness, by releasing data (by default) under a Creative Commons Zero (CC0) Public Domain Dedication which places no restrictions on who can use the data or what can be done with them. While not legally required, researchers using the data are expected to abide by community norms and cite the data following the guidelines included within each dataset. The primary motivation for this policy is that it makes the data maximally accessible to the largest possible number of researchers and citizen-scientists.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Standards-focused data sharing

To ensure the utility of shared data for the purposes of efficient discovery, reuse, repeatability, and reproducibility, standards are required for data and metadata organization. These standards make the structure of the data clear to users and thus reduce the need for support by data owners and curation by repository owners, as well as enabling automated QA, preprocessing, and analytics. Unfortunately, most prior data sharing projects have relied upon bespoke organizational schemes, which can lead to misunderstanding and can also require substantial reorganization to adapt to common analysis workflows. The need for a clearly defined standard for neuroimaging data emerged from our experiences in the OpenfMRI project; while the repository had developed a bespoke scheme for data organization and file naming, this scheme was ad hoc and limited in its coverage, and datasets often required substantial manual curation (involving laborious interaction with data owners). In addition, there was no way to directly validate whether a particular dataset met the standard.

For these reasons, we focused at the outset of the OpenNeuro project on developing a more robust data organization standard that could be implemented in an automated validator. We engaged a large group of researchers from the neuroimaging community to establish a standard that ultimately became the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016), which is now a highly successful community standard for a broad and growing range of neuroimaging data types. BIDS defines a set of schemas for file and folder organization and naming, along with a schema for metadata organization. The framework was inspired by the existing data organization frameworks used in many research laboratories, so that transitioning to the standard is relatively easy for most researchers. One of the important features of BIDS is its extensibility; using a scheme inspired by open source software projects, community members can propose extensions to BIDS that encompass new data types. To date, modality extensions include magnetoencephalography (Niso et al., 2018), scalp electroencephalography (Pernet et al., 2019), intracranial EEG (Holdgraf et al., 2019), positron emission tomography (Norgaard et al., 2021), and arterial spin labeling MRI. In addition to standards for raw data, the BIDS community has also developed a standard for the organization of the outputs of processing operations (known as “BIDS Derivatives”), providing a framework for sharing processed as well as raw data.

While BIDS and OpenNeuro are now independent projects, there is a strongly synergistic relationship. All data uploaded to OpenNeuro must first pass a BIDS validation step, such that all data in OpenNeuro are compliant with the BIDS specifications at upload time. Conversely, the OpenNeuro team has made substantial contributions to the BIDS standard and validator. The BIDS standard has been remarkably successful, with tens of thousands of datasets now available in the format, including but not limited to those contained in the OpenNeuro database. As a consequence, this model maximizes compatibility with processing and analysis tools (Gorgolewski et al., 2017), but more importantly, it effectively minimizes the potential for data

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

misinterpretation (e.g., when owner and re-user have slightly different definitions of a critical acquisition parameter). Through the adoption of BIDS, OpenNeuro has moved away from project- or database-specific data structures designed by the owner or the distributor (as used in earlier projects such as OpenfMRI and HCP) and toward a uniform and unambiguous representation model agreed upon by the research community prior to sharing and reuse.

FAIR sharing

The FAIR principles (Wilkinson et al., 2016) have provided an important framework to guide the development and assessment of open data resources. OpenNeuro implements each of these principles.

Findable. Each dataset within OpenNeuro is associated with metadata, both directly from the BIDS dataset along with additional dataset-level metadata provided by the submitter at time of submission. Both data and metadata are assigned a persistent unique identifier (Digital Object Identifier [DOI]). Within the repository, a machine-readable summary of BIDS metadata is collected by the BIDS validator and indexed with an ElasticSearch mapping. In addition, dataset-level metadata are exposed according to the schema.org standard, which allows indexing by external resources such as Google Dataset Search.

Accessible. Data and metadata can be retrieved using a number of access methods (directly from Amazon S3, using the openneuro command-line tool, or using DataLad) via standard protocols (http/https). Metadata are also accessible programmatically via a web API. Metadata remain available even in the case that data must be removed (e.g., in cases of human subjects concerns). No authentication is necessary to access the data.

Interoperable. The data and metadata use the BIDS standard to ensure accessible representation and interoperation with analysis workflows, such as BIDS Apps (Gorgolewski et al., 2017). Ongoing work is extending the metadata representation to use richer formats and to link to relevant FAIR ontologies or vocabularies.

Reusable. The data are released with a clear data use agreement (currently defaulting to a CC0 public domain dedication). Through use of the BIDS standard, the data and metadata are consistent with community standards in the field.

Data versioning and preservation

OpenNeuro keeps track of all changes in stored datasets, and allows researchers to unambiguously report the exact version of the data used for any analysis. OpenNeuro preserves all versions of the data through the creation of “snapshots” that unequivocally point to one specific point in the lifetime of a dataset. Data management and snapshots are supported by

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DataLad (RRID:SCR_003931) (Halchenko et al., 2016), a free and open-source distributed data management system (Hanke et al., 2021).

Protecting privacy and confidentiality of data

There is a direct relationship in data sharing between the openness of the data and their reuse potential; all else being equal, data that are more easily or openly available will be more easily and readily reused. However, all else is not equal, as openness raises concern regarding risks to subject privacy and confidentiality of data in human subjects research. Researchers are ethically bound to both minimize the risks to their research participants (including risks to confidentiality), and to maximize the benefits of their participation (United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978) . Because sharing of data will necessarily increase the potential utility of the data, researchers are ethically bound to share human subject data unless the benefits of sharing are outweighed by risks to the participant (Brakewood and Poldrack, 2013).

In general, risks to data privacy and confidentiality are addressed through deidentification of the data to be shared. For example, under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the US, deidentification can be achieved through the removal of any of 18 personal identifiers, unless the researcher has knowledge that the remaining data could be re-identified (known as the “safe harbor” method). With regard to neuroimaging data, a particularly challenging feature is the facial structure that is present in some forms of imaging data, such as structural MRI images. It is often possible to reconstruct facial structures from these images, and there are proofs of concept that such data could be used to re-identify individuals from photographic databases (Schwarz et al., 2019). It is thus essential to remove any image features that could be used to reconstruct facial structure (Bischoff-Grethe et al., 2007). For this reason, all MRI data shared through OpenNeuro must have facial features removed prior to upload, in addition to the 18 personal identifiers outlined by HIPAA. An exception is provided in cases where an investigator has explicit permission to openly share the data without defacing, usually when the data are collected by the investigator themself. At present, data are examined by a human curator to ensure that this requirement has been met. In the future, we plan to deploy an automated face detection tool (Bansal et al., n.d.) to detect any uploads that inadvertently contain facial features.

Truly informed consent requires that subjects be made aware that their data may be shared publicly, and that confidentiality cannot be absolutely guaranteed in the future. For this reason, we recommend that researchers planning to share their data via OpenNeuro use a consent form based on the Open Brain Consent (Bannier et al., 2021), which includes language that ensures subject awareness of the intent to share and its potential impact on the risk of participating.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Open source

The entirety of the code for OpenNeuro is available under a permissive open source software license (MIT License) at https://github.com/OpenNeuroOrg/openneuro. This enables any researcher who wishes to reuse part or all of the code or to run their own instance of the platform.

Data submission and access

Figure 1 outlines the steps required for sharing a dataset using OpenNeuro. Once shared, data can be accessed by several available mechanisms:

Web download. Each snapshot is associated with a link that provides immediate downloading of the dataset.

DataLad. DataLad (Halchenko et al., 2016) is a decentralized data management system built on top of git and git-annex. Through DataLad, researchers may install a complete copy of a dataset, while deferring the retrieval of file contents until needed, permitting lightweight views of large datasets. OpenNeuro's versioned snapshots are implemented as git tags, which allows specific versions to be easily retrieved or compared. The decentralized protocol also allows mirrors of the datasets to be hosted on GitHub and https://datasets.datalad.org, ensuring access during service interruptions of the OpenNeuro website.

OpenNeuro command line tool. The OpenNeuro command line tool provides access to the latest snapshot of all datasets, and is generally more stable than browser downloads for large datasets.

Amazon S3. The latest snapshot as well as all previous versions of a dataset may be fetched using the Amazon Web Services (AWS) clients or directly via https.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

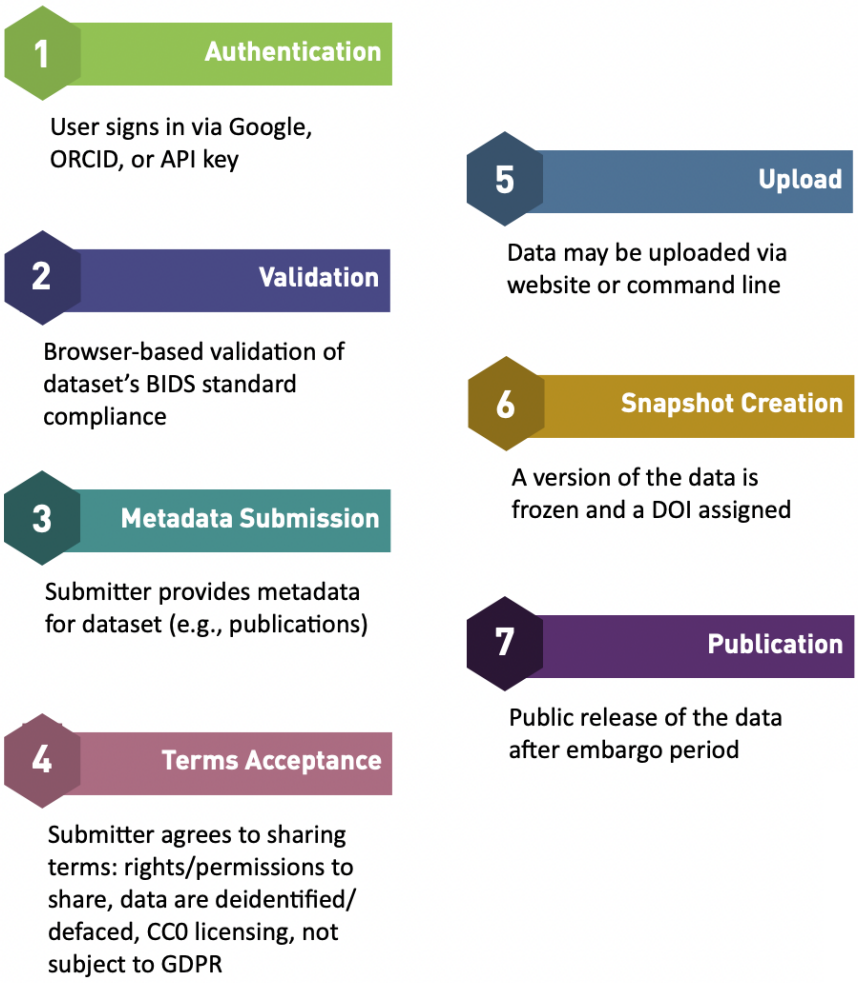


Figure 1. A schematic overview of the data upload process.

User support

Support for individual datasets. Data users sometimes have questions regarding particular datasets. In order to facilitate discussion of these issues and to make those discussions available to the entire community, a discussion forum is provided on each dataset page. The dataset owner is automatically notified by email of any questions that are posted. In addition, users can "follow" a dataset of interest and receive notifications of any comments posted to the dataset.

Site support. Two mechanisms are provided for users of the OpenNeuro site to obtain help with site issues. First, a helpdesk is available directly from the site, through which users can submit specific help questions. Second, users are recommended to post general questions to the

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Neurostars.org question and answer forum, so that the answers will be available to the entire community.

Data processing

Data processing was initially envisioned as an incentive for researchers to share their data, and the OpenNeuro site was launched in 2017 with the ability to perform cloud-based data processing using a limited set of analysis workflows. This feature was disabled in 2018, after an overhaul of the site's initial storage infrastructure. At that time, we determined that it would be preferable to collaborate with an existing platform dedicated to cloud processing rather than rebuilding our own execution platform. At present, OpenNeuro has partnered with the Brainlife.io platform (RRID:SCR_020940), which provides a large set of cloud-based neuroimaging workflows for data analysis and visualization. Data hosted on OpenNeuro can be easily imported into Brainlife for analysis, and more than 400 OpenNeuro datasets are cached for quick access; in the first 6 months of 2021, more than 700 analyses were performed on these datasets. In the future we plan to partner with additional platforms, including the NEMAR platform for EEG/MEG analysis; the availability of the data via DataLad and Amazon S3 also enables any platform to make the data available to their users without requiring any agreement or effort from OpenNeuro.

Results

Usage and impact

The OpenNeuro site was launched in June 2017, and was originally seeded with all of the datasets previously shared through OpenfMRI, after converting them to the BIDS standard. All data presented below are current as of June 18, 2021. The database contains 548 datasets comprising data from 18,758 individual participants. Figure 2 shows cumulative figures for numbers of datasets and subjects since 2018, demonstrating sustained and continual growth in the archive since its inception.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

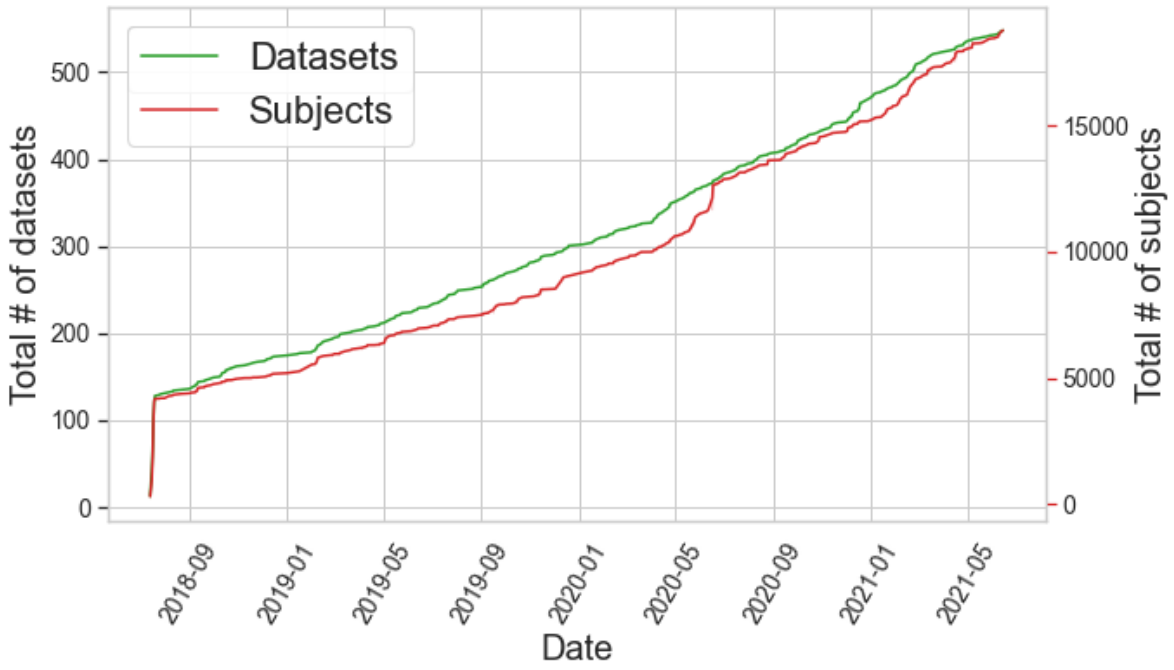


Figure 2: Cumulative growth of the number of datasets (green) and subjects (red) in the OpenNeuro archive from July 2018 through May 2021. The initial offset reflects the transfer of datasets from OpenfMRI to OpenNeuro after BIDS conversion.

The overwhelming majority of datasets are from humans (517 datasets, 95%), with a small but growing number of nonhuman species including mouse (17 datasets), rat (5 datasets), nonhuman primates (2 datasets), and juvenile pigs (1 dataset). Table 1 presents data for the prevalence of different modalities; while the majority of datasets include some form of MRI data, other supported modalities are present including electrophysiological measures and positron emission tomography.

Modality	**Number of datasets**
Anatomical MRI	465
Functional MRI	413
Electroencephalography	63
Diffusion-weighted MRI	52
Magnetoencephalography	21

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Positron emission tomography	8
Intracranial EEG	8
Arterial spin labeling MRI	3

Table 1. Number of datasets by imaging modality; additional modalities present in fewer than 3 datasets are not included here.

OpenNeuro is a recommended data repository for a number of publishers and journals, including: Nature Scientific Data, PLOS, eLife, F1000 Research, Gigascience, BioMed Central, American Heart Association, and Wellcome Open Research. The database contains 376 DOIs for publications associated with datasets (including both primary scientific publications and data descriptors).

Multiple dimensions of “big data”

Discussions of “big data” in neuroimaging (Poldrack and Gorgolewski, 2014; Smith and Nichols, 2018) have largely focused on datasets including large numbers of individuals. While these analyses are essential for robust population inference, it is also important to recognize that large numbers of subjects are only one dimension over which a neuroimaging dataset can be “big”. Here we will define the number of subjects as the “width” of the dataset, the number of different phenotypes measured for each individual as the “breadth” of the dataset, and the number of measurements per individual as the “depth” of the dataset.

The OpenNeuro database is distinguished by sharing datasets that are extensive along each of these dimensions (see Figure 3). With regard to width, the median dataset size is 22 subjects, with 28 studies having sample sizes larger than 100, and a maximum sample size of 928. With regard to breadth, notable datasets include: the BOLD5000 dataset (Chang et al., 2019), which includes data from subjects viewing a total of 5000 natural images; the Individual Brain Charting dataset (Pinho et al., 2020, 2018), which includes data from individuals each completing 24 different tasks, and the Multidomain Task Battery dataset (King et al., 2019), which includes data from individuals each completing 26 tasks. With regard to depth, the database currently includes: the MyConnectome dataset (Poldrack et al., 2015), which includes extensive task, resting, and diffusion MRI data from more than 100 sessions for a single individual; the Midnight Scan Club dataset (Gordon et al., 2017) which includes extensive task and resting fMRI data from ten individuals; and a number of other dense scanning datasets (Gonzalez-Castillo et al., 2015; Newbold et al., 2020; Salehi et al., 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

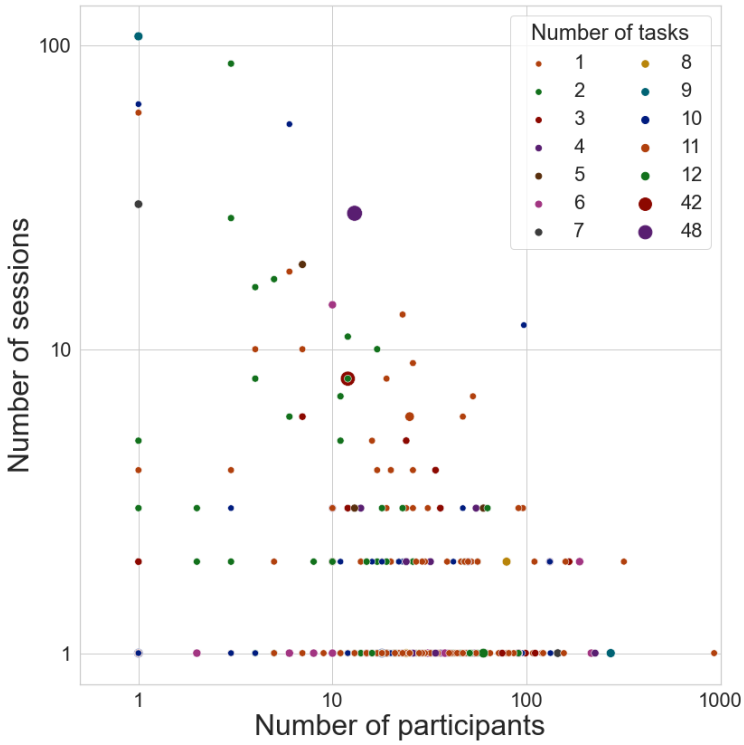
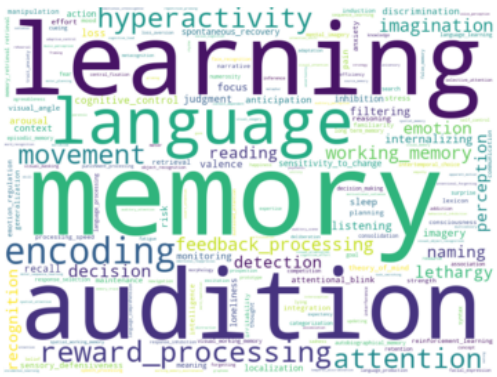


Figure 3. OpenNeuro datasets vary substantially in number of participants (X axis), number of sessions per participant (Y axis), and number of tasks per participant (size/color of datapoints); axes are log-scaled for easier visualization. Results are based on metadata derived directly from the 469 OpenNeuro datasets available via DataLad as of 6/18/2021.

Another unique feature of OpenNeuro is the breadth of phenotypes across datasets. To further characterize this, we searched the text associated with OpenNeuro datasets to identify terms related to psychological concepts and tasks as defined in the Cognitive Atlas ontology (Poldrack et al., 2011). Word clouds showing the top terms identified in this analysis are shown in Figure 4. This analysis shows a broad range of tasks and concepts associated with these datasets, highlighting the substantial conceptual and methodological breadth of the archive.



bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4. Word clouds based on Cognitive Atlas terms for psychological concepts (left) and tasks (right) identified from titles and README files associated with OpenNeuro datasets.

Data reuse

OpenNeuro has distributed a substantial amount of data; from May 2020 through April 2021, a total of 406 terabytes of data were distributed. Because data reuse is not directly measurable, we utilize published reuse of the shared data as a proxy. To identify published reuses of OpenNeuro data, we used Google Scholar and CrossRef to identify potential reuses, and then manually examined them to confirm that they were a legitimate reuse (as opposed to a primary publication of the data or data descriptor); note that this is an underestimate since many papers during this period reported analyses of data downloaded from OpenFMRI, which would not have been identified in our searches. We identified 165 publications that reused OpenNeuro datasets; this showed a sharp increase over time (see Figure 5). Of these publications, 112 were journal or conference papers, 42 were preprints, and 11 were other types of publications (such as theses or project reports). A total of 111 OpenNeuro datasets were reused at least once, with the most popular dataset (Poldrack et al., 2016) appearing in 28 published reuses. A significant number of publications reused multiple datasets; 31 of the 165 papers reused at least two datasets, with a maximum of 40 datasets reused (Esteban et al., 2019). Collecting these data from scratch would have required more than 21,000 individual subject visits; at an estimated scanning cost of $1000/session (based on the conservative cost estimate from (Milham et al., 2018)), this represents a total data reuse value of nearly twenty-one million US dollars. These reuses have a total of 1329 citations (according to Google Scholar as of June 15, 2021); the most highly cited reuse (Esteban et al., 2019) has more than 500 citations.

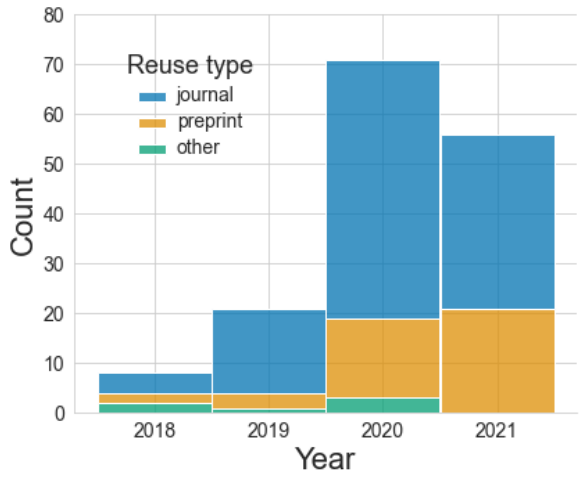


Figure 5. Published reuses of OpenNeuro datasets, split by the type of reuse. Note that the final bar includes only reuses identified through June 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The published reuses of OpenNeuro data span from basic neuroscience to methodological studies and software development. In particular, several studies demonstrate how OpenNeuro data have enabled new insights into brain function. For example, Martins et al. (2021) used structural MRI data from several OpenNeuro datasets along with other shared data to examine different patient groups suffering from physical pain or depression. Their analyses demonstrated a specific pattern of anatomical change common to patients with pain syndromes but distinct from depression. This kind of analysis highlights the way in which OpenNeuro enables researchers to combine smaller datasets in order to test hypotheses using convergent data, which can help overcome the confounds and biases present in any particular study as well as increasing statistical power. Other basic neuroscience studies have used OpenNeuro data to model the role of temporal context in forgetting (Chien and Honey, 2020), characterize the role of edge communities in brain networks (Faskowitz et al., 2020), understand the relationship between functional connectivity and sustained attention (Rosenberg et al., 2020), and to demonstrate that functional parcellation changes as a function of task (Salehi et al., 2020).

Data from OpenNeuro have been particularly useful for the development of new software tools. Esteban et al. (2019) used the breadth and variety of datasets in the archive to assess the robustness of the fMRIPrep preprocessing workflow to many different fMRI datasets, incorporating a total of 40 datasets from OpenNeuro. Importantly, these datasets were used in an iterative manner to improve the robustness of the tool; thus, the breadth of the data were essential both for assessment as well as for improvement of the tool. Without OpenNeuro (and BIDS), amassing such a large and diverse group of datasets would have required immense efforts to reach out to many different research groups, request their data, and then format the data for common usage, whereas with OpenNeuro the entirety of these datasets can be downloaded automatically within a number of hours, immediately ready for analysis. Other software development projects have taken advantage of some of the particular unique datasets in OpenNeuro; for example, Takeda et al. (2019) took advantage of a unique dataset that combines EEG, MEG, and MRI data on the same individuals (Wakeman and Henson, 2015) to demonstrate the broad range of functions of their VBMEG toolbox. Other software publications using OpenNeuro data include FastSurfer (Henschel et al., 2020) for structural MRI analysis, and Brainstorm (Tadel et al., 2019) for MEG/EEG analysis.

The data in OpenNeuro have been particularly useful for methodological researchers. One prominent example was published by Bowring et al. (2019), who examined how the use of different analysis software impacted statistical results from fMRI activation analyses. Their study included an in-depth analysis of the publications associated with each of 55 datasets, in order to identify studies with analysis pipelines and activation results that could be easily compared with their multi-platform results. Based on this process, they selected three datasets and processed each using several different analysis pipelines; their results highlighted substantial similarity in unthresholded maps but substantial discordance in thresholded maps, highlighting the need for better understanding of the impact of software packages on statistical results. Another example that would have been challenging to perform without OpenNeuro was published by Dadi et al. (2020), who developed a set of functional atlases using 27 datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This breadth allowed them to ensure that the specific features of the atlas were not driven by any particular dataset or task. Other examples include studies that used OpenNeuro data to assess the impact of confound regression on fMRI signals and develop new methods for confound modeling (Aquino et al., 2020), and to develop and benchmark new methods for multiple comparison correction (Spisák et al., 2019).

Discussion

The OpenNeuro data archive plays an important role in advancing neuroscience research and ensuring its reproducibility by enabling the sharing of a broad range of neuroscience data types according to the FAIR principles. Its tight integration with the community-driven BIDS standard enhances the ease of sharing, the reusability of the shared data, and the extensibility of the archive in the future. The shared data have enabled a growing number of publications that provide novel neuroscientific insights, as well as supporting novel methodological advances and software development.

Lessons learned

The experiences of our group in developing the OpenNeuro project have provided a number of lessons that may be useful more generally for researchers interested in establishing a culture of data sharing within their scientific subdomain.

Foremost, we have found that the use of a common community-driven format for data organization is essential to effective sharing. In our case, the BIDS standard has enabled data owners to easily share a growing range of data types (through the use of client-side validator), and has enabled researchers to easily reuse the data. Because any dataset that passes the validator can be shared, the community's efforts on extending the standard (which are implemented in the validator) has provided a steady stream of additions to the types of data that OpenNeuro can share. Another important point is that data sharing does not only include sharing with other researchers, but also with one's own research group in the future; thus, the use of a well-structured data standard can help researchers ensure that data collected by current lab members can be effectively utilized by other lab members in future, as well as making it easy to share the data beyond one's own lab. On the flipside, we continue to see that conversion of data into the BIDS standard remains a stumbling block for many researchers; the continued development of conversion tools is necessary to support these researchers.

Second, we have found that "it takes an ecosystem" to make data sharing successful. OpenNeuro is only one of the data sharing projects within the field of neuroimaging, and each of the projects has its own particular features and advantages, but together these projects have increasingly led the field to view data sharing as a net positive for our field. In addition, the availability of these data resources has allowed others to build projects that support new mechanisms for data representation and distribution (such as the Datalad project) and new

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

platforms for analysis (such as Brainlife.io). Together, these tools have provided researchers with additional incentives to share their data via OpenNeuro through its deep integration with those projects.

Finally, we would highlight the importance of domain-specific data repositories that support a particular research community. All of the sharing activities accomplished using OpenNeuro could in principle have been accomplished using more general data sharing repositories (such as Figshare or Dryad). A unique benefit of OpenNeuro has been in making a large number of datasets easily findable by researchers, rather than requiring a trawl through a much larger body of datasets to find ones that are relevant. By developing upload and download systems that are tailored for imaging data, OpenNeuro has also greatly lowered the barrier to sharing and reusing data. These benefits argue for the continued need for domain-specific data sharing projects designed in close consultation with researchers in the area.

Long-term sustainability

A continual challenge for any investigator-initiated data repository is the long-term sustainability of the archive, in order to ensure researchers' trust in the platform (Lin et al., 2020). The ongoing costs of running a repository are substantial, primarily due to the continuing cost of technological upkeep of a web platform with regard to security and stability, as well as the ongoing costs of storage and bandwidth on cloud platforms or hardware maintenance when using on-premise computing systems. Performant web applications require the use of cutting-edge software tools, which can often become deprecated or unstable over time, leading to substantial technical debt that must be continually addressed to maintain stable and secure operation.

One major challenge for repositories that are reliant upon federal grants is the usual three year funding period, in addition to the preference of standard grant mechanisms for funding novel projects rather than ongoing maintenance and operations. One welcome development has been the instigation of longer-term funding for data archives through the US BRAIN Initiative (Koroshetz et al., 2018), which has explicitly dedicated funding to the development and long-term sustainability of data archives for neuroscience data. These renewable five-year grants (of which OpenNeuro is one of the recipients) provide a much-needed longer term funding source for data repositories.

Another resource for longer term sustainability is institutional data repositories, which are increasingly available at many universities. OpenNeuro is working with the Stanford Digital Repository to develop a plan to deposit all raw datasets within the university's archive, which would provide a digital backstop to the archive's cloud storage.

OpenNeuro has also been fortunate to be part of the Amazon Public Datasets project (https://registry.opendata.aws/openneuro/), which has provided free data storage and bandwidth for the openly available datasets in the OpenNeuro archive.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Current limitations and future directions

There are a number of additional features planned for future development. These include:

Enhanced metadata. At present, a limited amount of dataset-level metadata is collected beyond that present within the BIDS metadata. Working with the CEDAR Metadata Center (Musen et al., 2015), we plan to add the ability for researchers to enter additional metadata that is linked to standard ontologies, including those being developed for BIDS data in the context of the Neuroimaging Data Model (Maumet et al., 2016). These annotations will provide the basis for more powerful queries of the archive.

Sharing of derivatives. At present, OpenNeuro only shares raw data. However, the availability of a BIDS standard for the outputs of data processing (i.e. “derivative” data) now provides the ability to include derivative data within a BIDS dataset. We plan to enable researchers to share derivatives, e.g. allowing the sharing of preprocessed MRI data in addition to raw data. This will greatly enhance the reuse of data by researchers who do not have the resources or expertise to preprocess these complex datasets as well as provide a standard baseline for downstream analyses, reducing the potential effects of analytic flexibility (Botvinik-Nezer et al., 2020; Bowring et al., 2019).

Bringing computing to data. The availability of the OpenNeuro data on the Amazon Web Services allows researchers direct access to computing on the data, but doing so requires a substantial degree of cloud computing expertise. To ease the application of computing to the data, we plan to adapt the DANDI Hub infrastructure developed by the Distributed Archives for Neurophysiology Data Integration (DANDI: https://www.dandiarchive.org/), which will allow direct access to the data via a Jupyter notebook.

Beyond MRI data. Driven by the initial seeding of data from OpenfMRI, and reflecting the fact that BIDS was originally MRI-centric, the data currently available from OpenNeuro are heavily skewed towards MRI, and fMRI in particular (Table 1). However, BIDS is quickly expanding to other modalities that can readily be uploaded to OpenNeuro, and there has been a rapid increase in sharing of other modalities; for example, more than 60 EEG datasets have been deposited since the publication of the BIDS-EEG standard in 2019 (Pernet et al., n.d.). . This organic expansion beyond MRI will be supported with the necessary adaptations (e.g., online visualization of new modalities) of OpenNeuro’s user interface.

Conclusion

Data sharing ensures the transparency and reproducibility of scientific research, and allows aggregation across datasets that improves statistical power and enables new research questions. The OpenNeuro repository plays a central role in the data sharing ecosystem by

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

promoting maximally open sharing of data, and by enhancing open availability of data from a wide range of datasets spanning The growth and impact of the repository demonstrate the viability of minimally restrictive sharing, and the importance of common standards such as BIDS for the effective sharing and reuse of data.

Materials and Methods

OpenNeuro Infrastructure

Code for the OpenNeuro platform is available at https://github.com/OpenNeuroOrg/openneuro. The application utilizes a cloud-based containerized architecture and is built in JavaScript and Python with a MongoDB database for application data storage. OpenNeuro is hosted on Amazon Web Services (AWS) using the Kubernetes container orchestration platform. Services are deployed as containers and integrated via a JavaScript GraphQL API gateway and the AWS Application Load Balancer. Several clients access this API, the React website, OpenNeuro command line interface, and an ElasticSearch indexer. Datasets are stored as DataLad repositories and managed by a Python backend service container. Each DataLad repository is assigned to a ZFS pool backed by AWS Elastic Block Store. This allows DataLad versioning and filesystem level access to datasets with existing processing and validation tools. Persistent metadata such as user accounts and permissions are maintained in a MongoDB database. Ephemeral caching is provided by Redis. Search indexes, performance monitoring, and logging are implemented with ElasticSearch. CloudFront is used as a global cache and network to provide global presence.

Content analysis

Data regarding OpenNeuro contents and usage were current as of June 18, 2021. Code and data needed to execute all analyses and generate all figures are available from https://doi.org/10.5281/zenodo.5189758.

Reuse analyses. Potential reuses were identified by first searching Google Scholar for the term “openneuro.”; note that this will exclude any paper that mention “OpenfMRI” instead of OpenNeuro, thus the reported results are underestimates of the true impact of the data, given that many of the datasets in OpenNeuro came from OpenFMRI. Papers matching this search were examined manually to confirm that they had reused data; data descriptor papers were excluded from further analysis. Citation counts were obtained from Google Scholar using the Python package 'scholarly'.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dataset size analyses. Dataset size analyses were performed using DataLad to obtain the full BIDS metadata for the 469 datasets available as of 6/18/2021, and then using pybids (Yarkoni et al., 2019) to load the metadata for each dataset.

Acknowledgements

The work described here has been supported by the National Institute Of Mental Health of the National Institutes of Health under Award Numbers R24MH117179 and R24MH114705. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Development of OpenNeuro and OpenfMRI was also supported by a grant from the Laura and John Arnold Foundation, and the National Science Foundation (OAC-1131441). Sharing of OpenNeuro datasets has been enabled by support from Amazon Web Services. We would like to thank all of the users who have uploaded data to OpenNeuro. Thanks to Franco Pestili for providing usage data on Brainlife.io, and Armin Thomas for helpful comments on an earlier draft.

References

Aquino KM, Fulcher BD, Parkes L, Sabaroedin K, Fornito A. 2020. Identifying and removing widespread signal deflections from fMRI data: Rethinking the global signal regression problem. *Neuroimage* **212**:116614.

Bannier E, Barker G, Borghesani V, Broeckx N, Clement P, Emblem KE, Ghosh S, Glerean E, Gorgolewski KJ, Havu M, Halchenko YO, Herholz P, Hespel A, Heunis S, Hu Y, Hu C-P, Huijser D, de la Iglesia Vayá M, Jancalek R, Katsaros VK, Kieseler M-L, Maumet C, Moreau CA, Mutsaerts H-J, Oostenveld R, Ozturk-Isik E, Pascual Leone Espinosa N, Pellman J, Pernet CR, Pizzini FB, Trbalić AŠ, Toussaint P-J, Visconti di Oleggio Castello M, Wang F, Wang C, Zhu H. 2021. The Open Brain Consent: Informing research participants and obtaining consent to share brain imaging data. *Hum Brain Mapp* **42**:1945–1951.

Bansal S, Kori A, Zulfikar W, Wexler J, Markiewicz C, Feingold F, Poldrack R, Esteban O. n.d. High-sensitivity detection of facial features on MRI brain scans with a convolutional network. doi:10.1101/2021.04.25.441373

Bischoff-Grethe A, Burak Ozyurt I, Busa E, Quinn BT, Fennema-Notestine C, Clark CP, Morris S, Bondi MW, Jernigan TL, Dale AM, Brown GG, Fischl B. 2007. A technique for the deidentification of structural brain MR images. *Human Brain Mapping*. doi:10.1002/hbm.20312

Biswal BB, Mennes M, Zuo X-N, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski A-M, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kötter R, Li S-J, Lin C-P, Lowe MJ, Mackay C, Madden DJ, Madsen KH, Margulies DS, Mayberg HS, McMahon K, Monk CS, Mostofsky SH, Nagel BJ, Pekar JJ, Peltier SJ, Petersen SE, Riedl V, Rombouts SARB, Rypma B, Schlaggar BL, Schmidt S, Seidler RD, Siegle GJ, Sorg C, Teng G-J, Veijola J, Villringer A, Walter M, Wang L, Weng X-C, Whitfield-Gabrieli S, Williamson P, Windischberger C, Zang Y-F, Zhang H-Y, Castellanos FX, Milham MP. 2010. Toward discovery science of human brain function. *Proc*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Natl Acad Sci U S A **107**:4734–4739.

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, Avesani P, Baczkowski BM, Bajracharya A, Bakst L, Ball S, Barilari M, Bault N, Beaton D, Beitner J, Benoit RG, Berkers RMWJ, Bhanji JP, Biswal BB, Bobadilla-Suarez S, Bortolini T, Bottenhorn KL, Bowring A, Braem S, Brooks HR, Brudner EG, Calderon CB, Camilleri JA, Castrellon JJ, Cecchetti L, Cieslik EC, Cole ZJ, Collignon O, Cox RW, Cunningham WA, Czoschke S, Dadi K, Davis CP, Luca AD, Delgado MR, Demetriou L, Dennison JB, Di X, Dickie EW, Dobryakova E, Donnat CL, Dukart J, Duncan NW, Durnez J, Eed A, Eickhoff SB, Erhart A, Fontanesi L, Fricke GM, Fu S, Galván A, Gau R, Genon S, Glatard T, Glerean E, Goeman JJ, Golowin SAE, González-García C, Gorgolewski KJ, Grady CL, Green MA, Guassi Moreira JF, Guest O, Hakimi S, Hamilton JP, Hancock R, Handjaras G, Harry BB, Hawco C, Herholz P, Herman G, Heunis S, Hoffstaedter F, Hogeveen J, Holmes S, Hu C-P, Huettel SA, Hughes ME, Iacovella V, Iordan AD, Isager PM, Isik AI, Jahn A, Johnson MR, Johnstone T, Joseph MJE, Juliano AC, Kable JW, Kassinopoulos M, Koba C, Kong X-Z, Koscik TR, Kucukboyaci NE, Kuhl BA, Kupek S, Laird AR, Lamm C, Langner R, Lauharatanahirun N, Lee H, Lee S, Leemans A, Leo A, Lesage E, Li F, Li MYC, Lim PC, Lintz EN, Liphardt SW, Losecaat Vermeer AB, Love BC, Mack ML, Malpica N, Marins T, Maumet C, McDonald K, McGuire JT, Melero H, Méndez Leal AS, Meyer B, Meyer KN, Mihai G, Mitsis GD, Moll J, Nielson DM, Nilsonne G, Notter MP, Olivetti E, Onicas AI, Papale P, Patil KR, Peelle JE, Pérez A, Pischedda D, Poline J-B, Prystauka Y, Ray S, Reuter-Lorenz PA, Reynolds RC, Ricciardi E, Rieck JR, Rodriguez-Thompson AM, Romyn A, Salo T, Samanez-Larkin GR, Sanz-Morales E, Schlichting ML, Schultz DH, Shen Q, Sheridan MA, Silvers JA, Skagerlund K, Smith A, Smith DV, Sokol-Hessner P, Steinkamp SR, Tashjian SM, Thirion B, Thorp JN, Tinghög G, Tisdall L, Tompson SH, Toro-Serey C, Torre Tresols JJ, Tozzi L, Truong V, Turella L, van 't Veer AE, Verguts T, Vettel JM, Vijayarajah S, Vo K, Wall MB, Weeda WD, Weis S, White DJ, Wisniewski D, Xifra-Porxas A, Yearling EA, Yoon S, Yuan R, Yuen KSL, Zhang L, Zhang X, Zosky JE, Nichols TE, Poldrack RA, Schonberg T. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**:84–88.

Bowring A, Maumet C, Nichols TE. 2019. Exploring the impact of analysis software on task fMRI results. *Hum Brain Mapp* **40**:3362–3384.

Brakewood B, Poldrack RA. 2013. The ethics of secondary data analysis: considering the application of Belmont principles to the sharing of neuroimaging data. *Neuroimage* **82**:671–676.

Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, Orr CA, Wager TD, Banich MT, Speer NK, Sutherland MT, Riedel MC, Dick AS, Bjork JM, Thomas KM, Chaarani B, Mejia MH, Hagler DJ Jr, Daniela Cornejo M, Sicat CS, Harms MP, Dosenbach NUF, Rosenberg M, Earl E, Bartsch H, Watts R, Polimeni JR, Kuperman JM, Fair DA, Dale AM, ABCD Imaging Acquisition Workgroup. 2018. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci* **32**:43–54.

Chang N, Pyles JA, Marcus A, Gupta A, Tarr MJ, Aminoff EM. 2019. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci Data* **6**:49.

Chien H-YS, Honey CJ. 2020. Constructing and Forgetting Temporal Context in the Human Cerebral Cortex. *Neuron* **106**:675–686.e11.

Dadi K, Varoquaux G, Machlouzarides-Shalit A, Gorgolewski KJ, Wassermann D, Thirion B, Mensch A. 2020. Fine-grain atlases of functional modes for fMRI analysis. *Neuroimage* **221**:117126.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ. 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods* **16**:111–116.

Faskowitz J, Esfahlani FZ, Jo Y, Sporns O, Betzel RF. 2020. Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. *Nat Neurosci* **23**:644–1654.

Gonzalez-Castillo J, Hoy CW, Handwerker DA, Roopchansingh V, Inati SJ, Saad ZS, Cox RW, Bandettini PA. 2015. Task Dependence, Tissue Specificity, and Spatial Distribution of Widespread Activations in Large Single-Subject Functional MRI Datasets at 7T. *Cereb Cortex* **25**:4667–4677.

Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, Ortega M, Hoyt-Drazen C, Gratton C, Sun H, Hampton JM, Coalson RS, Nguyen AL, McDermott KB, Shimony JS, Snyder AZ, Schlaggar BL, Petersen SE, Nelson SM, Dosenbach NUF. 2017. Precision Functional Mapping of Individual Human Brains. *Neuron* **95**:791–807.e7.

Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, Churchill NW, Cohen AL, Craddock RC, Devenyi GA, Eklund A, Esteban O, Flandin G, Ghosh SS, Guntupalli JS, Jenkinson M, Keshavan A, Kiar G, Liem F, Raamana PR, Raffelt D, Steele CJ, Quirion P-O, Smith RE, Strother SC, Varoquaux G, Wang Y, Yarkoni T, Poldrack RA. 2017. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput Biol* **13**:e1005209.

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline J-B, Rokem A, Schaefer G, Sochat V, Triplett W, Turner JA, Varoquaux G, Poldrack RA. 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**:160044.

Halchenko YO, Poldrack B, Hanke M. 2016. DataLad--decentralized data distribution for consumption and sharing of scientific datasetsOrganization of Human Brain Mapping Poster. Organization of Human Brain Mapping Annual Meeting, Geneva, Switzerland.

Hanke M, Pestilli F, Wagner AS, Markiewicz CJ, Poline J-B, Halchenko YO. 2021. In defense of decentralized research data management. *Neuroforum* **0**. doi:10.1515/nf-2020-0037

Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. 2020. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* **219**:117012.

Holdgraf C, Appelhoff S, Bickel S, Bouchard K, D'Ambrosio S, David O, Devinsky O, Dichter B, Flinker A, Foster BL, Gorgolewski KJ, Groen I, Groppe D, Gunduz A, Hamilton L, Honey CJ, Jas M, Knight R, Lachaux J-P, Lau JC, Lee-Messer C, Lundstrom BN, Miller KJ, Ojemann JG, Oostenveld R, Petridou N, Piantoni G, Pigorini A, Pouratian N, Ramsey NF, Stolk A, Swann NC, Tadel F, Voytek B, Wandell BA, Winawer J, Whitaker K, Zehl L, Hermes D. 2019. iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Sci Data* **6**:102.

Jwa A, Poldrack R. 2021. The Spectrum of Data sharing Policies in Neuroimaging Data Repositories. *PsyArXiv*. doi:10.31234/osf.io/cnuy7

King M, Hernandez-Castillo CR, Poldrack RA, Ivry RB, Diedrichsen J. 2019. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nat Neurosci* **22**:1371–1378.

Koroshetz W, Gordon J, Adams A, Beckel-Mitchener A, Churchill J, Farber G, Freund M, Gnadt J, Hsu NS, Langhals N, Lisanby S, Liu G, Peng GCY, Ramos K, Steinmetz M, Talley E, White S. 2018. The State of the NIH BRAIN Initiative. *J Neurosci* **38**:6427–6438.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, Westbrook J. 2020. The TRUST Principles for digital repositories. *Scientific Data*. doi:10.1038/s41597-020-0486-7

Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, Bell JD, Boultwood C, Collins R, Conroy MC, Crabtree N, Doherty N, Frangi AF, Harvey NC, Leeson P, Miller KL, Neubauer S, Petersen SE, Sellors J, Sheard S, Smith SM, Sudlow CLM, Matthews PM, Allen NE. 2020. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat Commun* **11**:2624.

Martins D, Dipasquale O, Veronese M, Turkheimer F, Loggia ML, McMahon S, Howard MA, Williams SCR. 2021. Transcriptional and cellular signatures of cortical morphometric similarity remodelling in chronic pain. *bioRxiv*. doi:10.1101/2021.03.24.436777

Maumet C, Auer T, Bowring A, Chen G, Das S, Flandin G, Ghosh S, Glatard T, Gorgolewski KJ, Helmer KG, Jenkinson M, Keator DB, Nichols BN, Poline J-B, Reynolds R, Sochat V, Turner J, Nichols TE. 2016. Sharing brain mapping statistical results with the neuroimaging data model. *Sci Data* **3**:160102.

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

Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, Koo B, Krishnakumar A, Biswal BB, Castellanos FX, Colcombe S, Di Martino A, Zuo X-N, Klein A. 2018. Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* **9**:2818.

Musen MA, Bean CA, Cheung K-H, Dumontier M, Durante KA, Gevaert O, Gonzalez-Beltran A, Khatri P, Kleinstein SH, O'Connor MJ, Pouliot Y, Rocca-Serra P, Sansone S-A, Wiser JA, CEDAR team. 2015. The center for expanded data annotation and retrieval. *J Am Med Inform Assoc* **22**:1148–1152.

National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, Committee on Toward an Open Science Enterprise. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. National Academies Press.

Newbold DJ, Laumann TO, Hoyt CR, Hampton JM, Montez DF, Raut RV, Ortega M, Mitra A, Nielsen AN, Miller DB, Adeyemo B, Nguyen AL, Scheidter KM, Tanenbaum AB, Van AN, Marek S, Schlaggar BL, Carter AR, Greene DJ, Gordon EM, Raichle ME, Petersen SE, Snyder AZ, Dosenbach NUF. 2020. Plasticity and Spontaneous Activity Pulses in Disused Human Brain Circuits. *Neuron* **107**:580–589.e6.

Niso G, Gorgolewski KJ, Bock E, Brooks TL, Flandin G, Gramfort A, Henson RN, Jas M, Litvak V, T Moreau J, Oostenveld R, Schoffelen J-M, Tadel F, Wexler J, Baillet S. 2018. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Sci Data* **5**:180110.

Nooner KB, Colcombe SJ, Tobe RH, Mennes M, Benedict MM, Moreno AL, Panek LJ, Brown S, Zavitz ST, Li Q, Sikka S, Gutman D, Bangaru S, Schlachter RT, Kamiel SM, Anwar AR, Hinz CM, Kaplan MS, Rachlin AB, Adelsberg S, Cheung B, Khanuja R, Yan C, Craddock CC, Calhoun V, Courtney W, King M, Wood D, Cox CL, Kelly AMC, Di Martino A, Petkova E, Reiss PT, Duan N, Thomsen D, Biswal B, Coffey B, Hoptman MJ, Javitt DC, Pomara N, Sidtis JJ, Koplewicz HS, Castellanos FX, Leventhal BL, Milham MP. 2012. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Front Neurosci* **6**:152.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Norgaard M, Matheson GJ, Hansen HD, Thomas A, Searle G, Rizzo G, Veronese M, Giacomel A, Yaqub M, Tonietto M, Funck T, Gillman A, Boniface H, Routier A, Dalenberg JR, Betthauser T, Feingold F, Markiewicz CJ, Gorgolewski KJ, Blair RW, Appelhoff S, Gau R, Salo T, Niso G, Pernet C, Phillips C, Oostenveld R, Gallezot J-D, Carson RE, Knudsen GM, Innis RB, Ganz M. 2021. PET-BIDS, an extension to the brain imaging data structure for positron emission tomography. *bioRxiv*. doi:10.1101/2021.06.16.448390

Pernet CR, Appelhoff S, Flandin G, Phillips C, Delorme A, Oostenveld R. n.d. BIDS-EEG: an extension to the Brain Imaging Data Structure (BIDS) Specification for electroencephalography. doi:10.31234/osf.io/63a4y

Pernet CR, Appelhoff S, Gorgolewski KJ, Flandin G, Phillips C, Delorme A, Oostenveld R. 2019. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data* **6**:103.

Pinho AL, Amadon A, Gauthier B, Clairis N, Knops A, Genon S, Dohmatob E, Torre JJ, Ginisty C, Becuwe-Desmidt S, Roger S, Lecomte Y, Berland V, Laurier L, Joly-Testault V, Médiouni-Cloarec G, Doublé C, Martins B, Salmon E, Piazza M, Melcher D, Pessiglione M, van Wassenhove V, Eger E, Varoquaux G, Dehaene S, Hertz-Pannier L, Thirion B. 2020. Individual Brain Charting dataset extension, second release of high-resolution fMRI data for cognitive mapping. *Sci Data* **7**:353.

Pinho AL, Amadon A, Ruest T, Fabre M, Dohmatob E, Denghien I, Ginisty C, Becuwe-Desmidt S, Roger S, Laurier L, Joly-Testault V, Médiouni-Cloarec G, Doublé C, Martins B, Pinel P, Eger E, Varoquaux G, Pallier C, Dehaene S, Hertz-Pannier L, Thirion B. 2018. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. *Sci Data* **5**:180105.

Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, Cumba C, Koyejo O, Milham MP. 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform* **7**:12.

Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt KH, Mumford JA, Sabb FW, Freimer NB, London ED, Cannon TD, Bilder RM. 2016. A phenome-wide examination of neural and cognitive function. *Sci Data* **3**:160110.

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

Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, Parker DS, Sabb FW, Bilder RM. 2011. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front Neuroinform* **5**:17.

Poldrack RA, Laumann TO, Koyejo O, Gregory B, Hover A, Chen M-Y, Gorgolewski KJ, Luci J, Joo SJ, Boyd RL, Hunicke-Smith S, Simpson ZB, Caven T, Sochat V, Shine JM, Gordon E, Snyder AZ, Adeyemo B, Petersen SE, Glahn DC, Reese Mckay D, Curran JE, Göring HHH, Carless MA, Blangero J, Dougherty R, Leemans A, Handwerker DA, Frick L, Marcotte EM, Mumford JA. 2015. Long-term neural and physiological phenotyping of a single human. *Nat Commun* **6**:8885.

Poline J-B, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, Haselgrove C, Helmer KG, Keator DB, Marcus DS, Others. 2012. Data sharing in neuroimaging research. *Front Neuroinform* **6**.

Rosenberg MD, Scheinost D, Greene AS, Avery EW, Kwon YH, Finn ES, Ramani R, Qiu M, Constable RT, Chun MM. 2020. Functional connectivity predicts changes in attention observed across minutes, days, and months. *Proc Natl Acad Sci U S A* **117**:3797–3807.

Salehi M, Greene AS, Karbasi A, Shen X, Scheinost D, Constable RT. 2020. There is no single functional atlas even for a single individual: Functional parcel definitions change with task. *Neuroimage* **208**:116366.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450168; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, Arani A, Spychalla AJ, Kantarci K, Knopman DS, Petersen RC, Jack CR Jr. 2019. Identification of Anonymous MRI Research Participants with Face-Recognition Software. *N Engl J Med* **381**:1684–1686.

Smith SM, Nichols TE. 2018. Statistical Challenges in “Big Data” Human Neuroimaging. *Neuron* **97**:263–268.

Spisák T, Spisák Z, Zunhammer M, Bingel U, Smith S, Nichols T, Kincses T. 2019. Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. *Neuroimage* **185**:12–26.

Tadel F, Bock E, Niso G, Mosher JC, Cousineau M, Pantazis D, Leahy RM, Baillet S. 2019. MEG/EEG Group Analysis With Brainstorm. *Front Neurosci* **13**:76.

Takeda Y, Suzuki K, Kawato M, Yamashita O. 2019. MEG Source Imaging and Group Analysis Using VBMEG. *Front Neurosci* **13**:241.

United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. The Commission.

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, WU-Minn HCP Consortium. 2013. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**:62–79.

Van Horn JD, Gazzaniga MS. 2013. Why share data? Lessons learned from the fMRIDC. *Neuroimage* **82**:677–682.

Wakeman DG, Henson RN. 2015. A multi-subject, multi-modal human neuroimaging dataset. *Sci Data* **2**:150001.

Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**:160018.

Yarkoni T, Markiewicz CJ, de la Vega A, Gorgolewski KJ, Salo T, Halchenko YO, McNamara Q, DeStasio K, Poline J-B, Petrov D, Hayot-Sasson V, Nielson DM, Carlin J, Kiar G, Whitaker K, DuPre E, Wagner A, Tirrell LS, Jas M, Hanke M, Poldrack RA, Esteban O, Appelhoff S, Holdgraf C, Staden I, Thirion B, Kleinschmidt DF, Lee JA, Visconti di Oleggio Castello M, Notter MP, Blair R. 2019. PyBIDS: Python tools for BIDS datasets. *J Open Source Softw* **4**. doi:10.21105/joss.01294