Published in final edited form as:

Neuroimage. 2016 January 1; 124(0 0): 1089–1092. doi:10.1016/j.neuroimage.2015.04.022.

Neuroimaging Data Sharing on the Neuroinformatics Database Platform

Gregory A Book^{1,*}, Michael Stevens¹, Michal Assaf^{1,2}, David Glahn¹, and Godfrey D Pearlson^{1,2}

¹Olin Neuropsychiatry Research Center, Hartford Hospital, Hartford CT

²Yale University, Department of Psychiatry, New Haven, CT

Abstract

We describe the Neuroinformatics Database (NiDB), an open-source database platform for archiving, analysis, and sharing of neuroimaging data. Data from the multi-site projects Autism Brain Imaging Data Exchange (ABIDE), Bipolar-Schizophrenia Network on Intermediate Phenotypes parts one and two (B-SNIP1, B-SNIP2), and Monetary Incentive Delay task (MID) are available for download from the public instance of NiDB, with more projects sharing data as it becomes available. As demonstrated by making several large datasets available, NiDB is an extensible platform appropriately suited to archive and distribute shared neuroimaging data.

Background

Neuroinformatics Database (NiDB) was created to solve the problem of organizing and analyzing very large neuroimaging datasets and has since grown into a neuroimaging database platform [1]. When development of the platform began in 2005, a publication with a sample size of one hundred subjects was considered very large, while now sample sizes in the thousands are common [2, 3]. There are diminishing returns when using sample sizes larger than 1000 subjects, however the ability to store and analyze data from multiple patient cohorts and longitudinal datasets is extremely valuable, especially when testing reproducibility [4].

Development of NiDB began as a system for searching and downloading of MRI scans collected in the previous 30 days, using flat file storage of meta-data. The system could only search by subject ID, protocol name, and scan date. However, as the amount of stored data grew, the system was re-written to use a SQL database and catalog more meta-data. As data sizes grew further, the system architecture was redesigned to be subjectcentric, following a Subject→Enrollment→Imaging Study→Series hierarchy (figure 1). A subject-centric design allows association of multiple modalities of data with an imaging session, multiple imaging sessions with a subject's enrollment in a project, and enrollment of subjects in multiple

*Corresponding Author: gregory.a.book@gmail.com, gregory.book@hhchealth.org, 860-545-7267.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

projects. This architecture provides a standardized hierarchy into which new imaging modalities are stored in the database and makes the addition of project permissions and security straightforward. NiDB currently stores magnetic resonance (MR), computed tomography (CT), ultrasound (US), positron-emission tomography (PET), electroencephalography (EEG), pre-pulse inhibition (PPI), eye-tracking (ET), and genome data, but can expand with minimal effort to include any modality. NiDB is web-based, using PHP and JavaScript as the front-end, MySQL as the middle layer, and Perl as the backend. A separate uploader for large datasets is written in C++ and QT. Regular users access the system through the web-based GUI or QT based uploader, and administrators perform many maintenance operations through the web-based GUI. A small amount of maintenance is required in the back-end by a developer to backup data, add new modalities, fix bugs, or add enhancements.

Data importing, searching, and exporting features are available, as well as storage of subject demographics, system statistics, and project permissions. NiDB contains several features beyond data storage and searching, including: pipeline analysis, inter-instance sharing, and modular automated quality control (QC). Automated QC is 'modular', meaning a user can create a QC module/script which takes a data path as input, performs specified QC analysis, and inserts the results into the database. NiDB's pipeline system is connected to a compute cluster where analyses are automatically performed and results are imported back into the database to be associated with the original data. Data are analyzed using a normal bash script, with special NiDB variables that are replaced with full paths when the pipeline is run. Each pipeline has a set of data criteria, and all imaging studies that match the criteria are sent through the pipeline, which creates a custom cluster job with the correct paths and IDs for each imaging study. During the cluster job processing, output from the original bash script is logged and the status of an analysis can be viewed, along with summary statistics such as number completed, running, or in error state. Upon completion of each analysis, important results and figures (defined by the user) are automatically imported back into NiDB and are available for searching alongside the raw data. Data processed through a NiDB freesurfer pipeline was included in very large scale study of genetic association with subcortical brain structures [5]. The NiDB pipeline system includes inter-pipeline dependencies to allow efficient processing of data. Examples of tested NiDB pipelines include freesurfer, SPM fMRI processing, FSL DTI, FSL fMRI, and Human Connectome Pipeline (HCP) analyses.

A feature important to future data sharing is NiDB's export to National Database on Autism Research (NDAR) format. NDAR is a large scale data repository hosted by the NIMH, and the underlying database system is now used for the Research Domain Criteria (RDoC) project which seeks to archive data collected under NIMH sponsored projects. NiDB compatibility with NDAR allows for direct, seamless data export.

Design

The current iteration of NiDB was designed as an active study management system for neuroimaging and clinical research data. Data is imported from one of several sources (DICOM receiver, web-based importer, GUI based importer, or inter-instance sharing),

archived and QC'd. The data is then searchable to users. Archived data is associated with existing or new subjects, at which time demographic data may be imported from meta-data (eg DICOM header) or manually entered. Data stored on the public server is static for the main projects listed in the available data section, but may be dynamic for other projects.

Available Data

NiDB is currently hosted in two different instances, each with different data and accessibility. The internal instance of NiDB, only accessible within the Olin Neuropsychiatry Research Center network, contains 12.9TB of raw data from 199,951 imaging series from 22,464 imaging sessions from 11,147 subjects in 158 projects. In total, 366 days of CPU time have been used to compute QC metrics and 16.6TB of data have been requested. While these data are not all available publicly, it attests to the scalability of NiDB. The external (public) instance of NiDB contains the data described in this paper, available at <http://olinnidb.org> (table 1). Five major projects comprise the data currently shared on the public server: ABIDE, B-SNIP1, PARDIP, B-SNIP2, MID.

Autism Brain Imaging Data Exchange (ABIDE) data was aggregated by the International Data-sharing Initiative (INDI) and imported into NiDB [6, 7]. The ABIDE dataset contains resting fMRI, structural MR, and phenotypic data from 16 projects (sites) examining autism spectrum disorder. The original downloads from ABIDE were single blocks of data from each site, but after importing into NiDB, subsets of the data can be searched for and downloaded. Part one of the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP1) study, examines multiple phenotypes in individuals with schizophrenia, psychotic bipolar disorder, and schizoaffective disorder, and their first-degree relatives. Data were collected on five different MR scanners from five sites and will be available for download as a complete dataset by summer of 2015 [8–13]. Psychosis and Related Domains Intermediate Phenotypes (PARDIP) is a continuation of B-SNIP1, but is conducted on non-psychotic bipolar patients and controls, and is only collected at three of the original five B-SNIP sites. Part two of B-SNIP will be similar to B-SNIP part one, with different sites (some overlapping from B-SNIP1) and more standardized imaging across sites. A group of four sites is retrospectively combining data from the monetary incentive delay task (MID, also known as the Hommer task) collected from substance abusers [14]. In this ongoing project, MR data from the MID task, T1 images, and some resting state scans are archived in NiDB, as well as some assessment data from several reward tasks.

Subject age-at-scan, sex, and group (control, patient group, etc) are stored for all subjects in all projects. ABIDE and MID are currently the only projects to store phenotypic and assessment data in addition to imaging. All projects store at least one T1 MR image, and one or more functional MR images for each subject. B-SNIP1, BSNIP2 and PARDIP also store EEG and eye-tracking data, though only the BSNIP projects store PPI data.

Data is stored in the original (raw) format in which it was collected or contributed. MR imaging data from the B-SNIP and PARDIP studies are stored in DICOM or Philips .par/.rec format, while ABIDE data was contributed in Nifti 4D format. NiDB performs data format conversion at the time of request if the original data was in DICOM

or .par/.rec format, otherwise the original format is made available for download. Data in DICOM format can be converted into Analyze 3D/4D, Nifti 3D/4D, and anonymized DICOM. Raw data in .par/.rec format can only be converted to Analyze or Nifti format, or downloaded as raw data, but cannot be anonymized. Analyzed or processed data is not available as part of any downloads, nor are any citable digital-object-identifiers(doi's)/uniform resource identifier (uri's) available for downloads.

Quality Control

Multiple quality control methods are available on NiDB, which are either manual or automated, and most are currently only available for MR. After MR images are archived, they begin an automated QC process, which calculates SNR and motion estimation for 3D and 4D images. SNR on 3D data is calculated by comparing the mean signal from a brain extracted (BET) volume to the corners of the volume. SNR on 4D data is calculated by comparing the mean signal over time extracted from the BET volume compared to the corners of the volume over time. Motion estimation, consisting of X,Y,Z translation, P,R,Y rotation, and derivative of translation, are calculated using FSL's mcflirt [15]. Motion estimation of 3D images is done by taking the fast-Fourier transform (FFT) of each slice, performing radial averages of the resulting plots, and finding the mean slope of the line. More 3D motion corresponds to less high frequency signal and therefore a steeper line. In practice, the SNR and 3D motion values are only comparable within an imaging study, not across imaging studies or across subjects.

Motion estimation from 4D datasets is comparable across subjects and imaging studies. All QC metrics are color coded from green (good) to red (bad), depending on the QC method's scale, for easy identification by a user of data quality. NiDB also provides area for user entered notes and ratings (scale of 0–5, 0=good, 5=bad) of data. Since different users may have different QC criteria, a combination of subjective and objective QC observations is useful.

Data Access

Users of the external (public) NiDB must create an account on the website to access shared data, providing their email address so that information about changes, additions, or deletions to data can be emailed to users. Datasets are available in two ways 1) public downloads, which are single .zip files containing all data for a particular project or subset of a project 2) searching and downloading subsets of data individually. Once registered, users can download any of the public downloads, but must request access to particular datasets and projects if they wish to explore them individually and view QC and other information. Access is requested by clicking a "request access" link within NiDB.

Data are anonymized by removing subject name, subject ID, and month and day of birth. Birth year is retained so that age-at-scan can be calculated. No data usage agreements are necessary to download the data. Internally, the NiDB instance uses subject names, birthdates, and IDs to uniquely identify subjects; however the public server uses an encrypted form of the subject ID to identify a subject. Because of the presence of

identifiable information on the internal server, access is restricted by project to only IRB approved personnel.

Data is downloaded via the HTTP protocol. Large files of several hundred gigabytes are also downloaded via HTTP, with a resume option. Data is zipped using the Linux zip command. Data available through the public downloads is available only as 4D Nifti format (for images which were originally DICOM) or their original format if non-DICOM (e.g. EEG data). Data available through the search and download method can be converted prior to download into several formats including Analyze, Nifti 3D or 4D, or anonymized DICOM.

Contribution

Researchers wishing to contribute new, unshared data or data shared in another repository to the NiDB repository are welcome to do so, with the understanding that their data must be completely anonymous, including removal of dates of service, and must be freely available to the public when uploaded to NiDB. Contributed data must be neuroimaging in nature or related to neuroimaging. The contributor must be the original owner/collector of the data. Any data format or modality is acceptable as long as it can be anonymized, including dates of service, patient information, or other identifiers. Collaboration is welcome to contribute to the development of the NiDB platform itself, including pipeline and QC development. Utilization of NiDB as a primary or backup database for actively maintained studies is encouraged. The pipeline system is part of the NiDB code, but no compute cluster is available on the public NiDB server.

Maintenance

NiDB is maintained by the Olin Neuropsychiatry Research Center (ONRC), and supported both internally and by data sharing grants. The long term goals for development of NiDB are increasing performance, reliability, and efficiency of the system and increasing the user-base. More significant long term goals include porting the pipeline system to run on Amazon Web Services (AWS), a cloud computing platform [16]. Using StarCluster, we intend to create a dynamic compute cluster attached to NiDB within AWS that would stand in place of the computer cluster at the ONRC [17].

References

1. Book GA, et al. Neuroinformatics Database (NiDB)--a modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics*. 2013; 11(4):495–505. [PubMed: 23912507]
2. Kiehl KA, et al. An adaptive reflexive processing model of neurocognitive function: supporting evidence from a large scale (n = 100) fMRI study of an auditory oddball task. *Neuroimage*. 2005; 25(3):899–915. [PubMed: 15808990]
3. Meda SA, et al. Multivariate analysis reveals genetic associations of the resting default mode network in psychotic bipolar disorder and schizophrenia. *Proc Natl Acad Sci U S A*. 2014; 111(19):E2066–E2075. [PubMed: 24778245]
4. Kennedy DN. Data persistence insurance. *Neuroinformatics*. 2014; 12(3):361–363. [PubMed: 24985144]
5. Hibar DP, et al. Common genetic variants influence human subcortical brain structures. *Nature*. 2015

6. 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*. 2014; 19(6):659–667. [PubMed: 23774715]
7. Mennes M, et al. Making data sharing work: the FCP/INDI experience. *Neuroimage*. 2013; 82:683–691. [PubMed: 23123682]
8. Hill SK, et al. Neuropsychological impairments in schizophrenia and psychotic bipolar disorder: findings from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) study. *Am J Psychiatry*. 2013; 170(11):1275–1284. [PubMed: 23771174]
9. Hill SK, et al. Regressing to Prior Response Preference After Set Switching Implicates Striatal Dysfunction Across Psychotic Disorders: Findings From the B-SNIP Study. *Schizophr Bull*. 2014
10. Ivleva EI, et al. Gray matter volume as an intermediate phenotype for psychosis: Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP). *Am J Psychiatry*. 2013; 170(11):1285–1296. [PubMed: 24185241]
11. Mathew I, et al. Medial temporal lobe structures and hippocampal subfields in psychotic disorders: findings from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) study. *JAMA Psychiatry*. 2014; 71(7):769–777. [PubMed: 24828364]
12. Ruocco AC, et al. Emotion recognition deficits in schizophrenia-spectrum disorders and psychotic bipolar disorder: Findings from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) study. *Schizophr Res*. 2014; 158(1–3):105–112. [PubMed: 25052782]
13. Tamminga CA, et al. Clinical phenotypes of psychosis in the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP). *Am J Psychiatry*. 2013; 170(11):1263–1274. [PubMed: 23846857]
14. Knutson B, et al. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci*. 2001; 21(16):RC159. [PubMed: 11459880]
15. Jenkinson M, et al. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. 2002; 17(2):825–841. [PubMed: 12377157]
16. Services AW. Amazon Web Services. 2015 Available from: <http://aws.amazon.com>.
17. MIT. StarCluster. 2015 Available from: <http://star.mit.edu/cluster/>.

Highlights

- We describe the available datasets on the Neuroinformatics Database public website
- Datasets include ABIDE (n=1011) and BSNIP1 (n=2600)
- To be released data include PARDIP, BSNIP2, MID
- We describe the NiDB pipeline and QC systems

Figure 1.

Neuroinformatics database hierarchy. This is a near universal format into which any modality of imaging can be stored.

Analysis

Pipelines > HCPfMRI-APPSad > Analysis List >

63 items Page 1 of 1 (1000/page)

« ‹ › »

☐ Study	Pipeline version	Study date	# series	Status	Complete?	Logs	Files	Notes	Message	Size bytes	Hostname	Setup time completed date	Cluster time completed date	Delete ☐
☐ S7224HRW3	11	Oct 6, 2014 08:13	4	processing					processing step 39 of 53 2015-01-12 15:11:32	0	hhlxncas22	00:01:13 2015-01-12 15:10	1969-12-31 19:00	☐
☐ S7224HRW1	11	Jul 8, 2013 08:21	4	processing					processing step 39 of 53 2015-01-13 03:22:44	0	compute20	00:01:49 2015-01-13 03:22	1969-12-31 19:00	☐
☐ S0053MBH3	11	Jan 6, 2015 08:09	4	complete					Cluster processing complete 2015-01-12 18:13:08	18,972,753,599	compute20	00:01:14 2015-01-12 14:39	03:33:45 2015-01-12 18:13	☐
☐ S1808RTC3	9	Dec 31, 2014 09:00	4	complete					Cluster processing complete 2015-01-05 19:56:47	8,864,832,482	hhlxncas22	00:00:39 2015-01-05 13:59	05:56:47 2015-01-05 19:56	☐
☐ S6024ZVW3	9	Dec 26, 2014 10:37	4	complete					Cluster processing complete 2015-01-05 20:02:50	8,922,779,478	compute21	00:01:26 2015-01-05 14:01	06:01:04 2015-01-05 20:02	☐
☐ S0861URH2	9	Dec 22, 2014 08:01	4	complete					Cluster processing complete 2015-01-05 20:02:43	8,990,348,486	compute21	00:01:38 2015-01-05 14:09	05:53:27 2015-01-05 20:02	☐
☐ S8560ORS3	8	Dec 16, 2014 08:10	4	complete					Default status message 2015-01-02 12:59:12	7,981,269,739	compute23	00:01:38 2014-12-18 13:01	04:04:30 2014-12-18 17:06	☐
☐ S6425HVM3	8	Dec 13, 2014 07:46	4	complete					Default status message 2015-01-02 13:00:12	8,504,564,438	compute23	00:01:24 2014-12-18 13:03	04:20:33 2014-12-18 17:23	☐
☐ S1115002	8	Dec 12, 2014 10:36	4	complete					Default status message	7,656,943,518	compute23	00:01:37	04:09:26	☐

Figure 2.

NiDB pipeline system, analysis list. Imaging studies that meet the data criteria are processed through the pipeline and their status is displayed.

Table 1

Available data in summer 2015.

Dataset	Number of Subjects	Number of Imaging Sessions	Data	Population	Number of Sites	Access	Date Available to Public
ABIDE	1011	1011	MR (rest, T1), 73 assessment summary scores	539 autism spectrum disorders; 573 controls	16	Public	2013
B-SNIP1	2447*	9788*	MR (rest, T1, DTI), EEG, PPI, eye-tracking, age, sex, dx group	933 schizophrenia, schizoaffective, psychotic bipolar I; 1055 first degree relatives; 459 controls	5	Public	2015
PARADIP	405*	1215*	MR (rest, T1, DTI), EEG, eye-tracking, age, sex, dx group	Collection ongoing	3	Open only to consortium members	2016
B-SNIP2	3000*	12,000*	MR (rest, T1, DTI), EEG, PPI, eye-tracking, age, sex, dx group	Collection ongoing	5	Open only to consortium members	2019
MID	1800*	2400*	MR (T1, MID, rest, DTI), age, sex, summary scores from selected assessments	Collection ongoing	4	Open only to consortium members	2016*

Items with an asterisk (*) are projections and may change based on QC results, subject exclusion, and enrollment. NiDB account creation is required to access all datasets.