Tools for Importing and Evaluating BIDS-EEG Formatted Data

Arnaud Delorme, Dung Truong, Ramon Martinez-Cancino, Cyril Pernet, Subha Sivagnanam, Kenneth Yoshimoto, Russ Poldrack, Amit Majumdar, Scott Makeig

***Abstract*— This article outlines a set of plug-in tools running on MATLAB to automatically import, preprocess, and evaluate the quality of electro-encephalography (EEG) data stored using the Brain Imaging Data Structure BIDS-EEG standard. As a proof of concept, we apply several possible data quality metrics to 30 EEG studies (sets of compatible datasets) currently available in BIDS-EEG format on the *OpenNeuro.org* platform. The *bids-matlab-tools* plug-in for EEGLAB (*sccn.ucsd.edu/eeglab*) checks for the presence of information needed for analysis, then applies preprocessing pipelines that compute data quality metrics and retain dataset portions (e.g., those channels and time points) that appear suitable for analysis. These and other data measures and visualizations will be made available to the EEG community through a new *NEMAR.org* web portal now under development to support the use of the *OpenNeuro.org* data archive by the human electrophysiology research community.**

I. Introduction

The Brain Imaging Data Structure (BIDS) standards first introduced in 2016 for storing human neuroimaging data are gaining increasing breadth, acceptance and use [1]. The advantages of using the BIDS standards to store data for analysis use and reuse are several. First, they do not rely on a complex database schema; they simply store data and relevant metadata in text and binary files within an ordinary file folder structure. Second, BIDS text files are both human and machine-readable, making it easier for new users to examine, understand, and contribute to advancing the BIDS standards through their now manifold community development efforts. Third, a growing number of data repositories already have tools to process and visualize BIDS data [2].

BIDS was first designed [1] to define a standard for storing MRI and fMRI data. However, its purpose has rapidly evolved, and BIDS now has published extension standards for storing EEG [3], magneto-encephalography (MEG) [4], and intracranial EEG [5, 6] data, with many volunteer development groups working on adding to the number of included data types. Modality-specific BIDS extensions (for EEG, MEG, etc.) set minimal standards for data encoding and metadata requirements appropriate to the data modality, while inheriting the basic structure and definitions of the top-level BIDS standard.

Thus BIDS-EEG, the BIDS extension for EEG [3] recommends two, while accepting four, of the many EEG data formats now in use by EEG laboratories and system manufacturers (EEGLAB; EDF and BDF; Brain Vision Data Exchange). The BIDS-EEG standard also specifies how to document recording parameters, electrode locations and EEG channel types as well as event records, all parameters needed for useful analysis and interpretation of stored EEG data.

Although standard BIDS validator functions check the consistency of BIDS studies with the BIDS standards; the quality of BIDS-archived data and metadata may vary widely. Therefore, it is important to define and supply data quality metrics to allow data users to assess the potential for reuse of available BIDS-EEG formatted studies. Here we present automated pipelines to provide information to users about the level metadata available in BIDS-EEG data sets, and to estimate channel, time course, and independent source quality. We apply these metrics to the collection of BIDS EEG studies currently available on the OpenNeuro platform.

II. Methods

○ *A. Importing BIDS data*

We have released a *bids-matlab-tools* plug-in (version 5.2) for the popular EEGLAB [7] software environment (*sccn.ucsd.edu/eeglab*) running on MATLAB (The Mathworks, Inc.) to import and export BIDS studies. Installing the plug-in using the EEGLAB plug-in manager adds menu items to the EEGLAB *File* menu to import and export BIDS-EEG formatted data. Upon calling the BIDS import menu item, a graphic interface window pops up (Fig. 1), asking for the type of information to be imported. The plug-in also allows importing BIDS-EEG formatted data from the MATLAB command line.

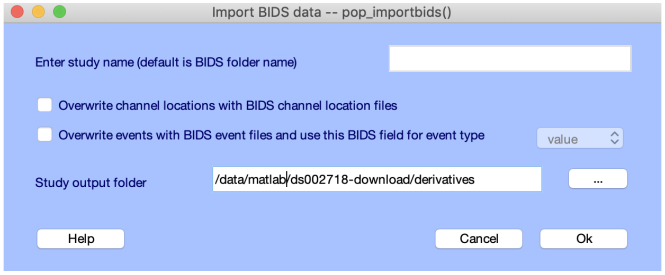

Figure 1. *The import interface of the bids-matlab-tools plug-in for EEGLAB.*

EEGLAB uses the term STUDY for a set of 'dataset' files collected from individual participants and/or recording sessions (though BIDS refers to such data collections as 'data sets', here we use 'study'). Raw EEG data files typically contain metadata about experimental, behavioral, and other types of events whose times of occurrence (onset) are noted for use in data analysis. BIDS allows defining events in dedicated text files that may contain detailed event information in the Hierarchical Event Descriptor (HED) system, which is currently being upgraded (HED 3rd-gen.) [8]. If the BIDS event files do contain additional event information, not in the raw EEG data file, users may choose to overwrite raw EEG data events with the event information contained in the BIDS event files by checking the first checkbox of the *bids-matlab-tool* EEGLAB plug-in import interface (see Fig. 1).

Raw EEG data files also typically include labels preserving channel location information; the BIDS-EEG standard also allows dedicated text files (**_channels.tsv* and **_electrodes.tsv*) giving the same or more information about the data channels and sensor locations. By selecting the

second checkbox (Fig. 1), users may choose to use the channel label and electrode location information contained in the BIDS-EEG files.

A BIDS-EEG study is imported into EEGLAB as an EEGLAB STUDY, allowing the many study-level and individual dataset-level processing tools in the EEGLAB environment to be readily applied to the data. Our tools import channel information from the BIDS-EEG text files, when available, and check its consistency with metadata in the raw EEG data files (see Table 1).

Known current limitations of the *bids-matlab-tools* plug-in are the number of channel coordinate systems (only one is currently supported), and the way file are saved, unnecessarily saving multiple copies of the same metadata files.

o *B. BIDS-EEG data used for testing*

To test our tools, we used all the EEG studies currently available (October, 2020) on OpenNeuro. A search for data described using the term 'EEG' on *OpenNeuro.org* returned 40 studies. Some of these contained no EEG data (N=3: ds000248; ds001408; ds003082), or contained EEG data that was unusable or not formatted to the BIDS format (N=5; ds000116; ds002000; ds002181; ds002734; ds002739). These used the *.bidsignore* settings file to include non BIDS-compliant data in OpenNeuro. One other BIDS-EEG study was a duplicate (N=1; ds002087); yet another contained processed instead of raw data (N=1; ds003004). This left a total of 30 usable BIDS-EEG formatted studies containing raw data comprising approximately 1.5 TB. To determine channel locations, we either used the electrode and channel location information specified in the study metadata or, when absent, added template electrode locations for the montage named in the BIDS-EEG metadata.

o *C. Advanced checking of BIDS-EEG study contents*

The publicly available BIDS-EEG validator [9] checks for basic data consistency. However, the type, quantity, and quality of the data may vary widely across archived studies and datasets. We have designed a simple approach to testing the overall quality of BIDS-EEG meta-data for an archived study. This answers 12 questions:

1. Is there a README file?
2. Is there a task description of more than 400 characters? A shorter description would likely not capture the complexity of the experiment.
3. Are the instructions to participants included?
4. Is there an event description file? Studies with inadequate event information make interpretation of events difficult.
5. Is the EEG reference electrode site specified?
6. Is the power line frequency specified?
7. Is channel type information included?
8. Are electrode locations specified?
9. Are participant age and gender specified?
10. Are non-brain artifacts in each dataset described?
11. Are the BIDS-EEG and raw data file channel information consistent (i.e., are the channel numbers the same?).
12. Is the BIDS event information consistent with the dataset-specified events (i.e., are the numbers of events in the BIDS and raw dataset files the same?) (Note: there may be instances where inconsistency does not reflect a problem with the data, as additional events may have been added to the BIDS metadata).

The *pop_importbids.m* of the *bids-matlab-tool* EEGLAB plug-in reports the percent of positive answers to the 12 questions. To aggregate this information across study datasets, if one of the EEG files contains information that is presumably valid for all datasets, the study is considered to fulfill that criterion. However, event and channel location information must be consistent for every dataset to fulfill criteria 11 and 12. Table 1 shows the results of this analysis on the collection of current OpenNeuro BIDS-EEG studies.

o *D. Data quality metrics and pre-processing pipelines*

We used 3 quality metrics to assess the quality of the EEG data within the BIDS-EEG studies. All computations were performed using EEGLAB v2020.0.

- *How many of the channels are 'good' for analysis?* We calculated the percentage of 'good' channels for each dataset of a study and assessed the 95% confidence interval for the mean percentage value (using percentile bootstrap function *bootci*). We calculated the number of unanalyzable ('bad') channels using the EEGLAB *clean_rawdata* (v2.2) plug-in with parameters: *FlatlineCriterion* 5, *ChannelCriterion* 0.8 (minimum correlation between channels, taking into account channel locations), and *LineNoiseCriterion* 4. For further information about this channel rejection procedure, see [10-12].
- *How much of the data is 'good' for analysis?* Here, *clean_rawdata* calculated the percentage of 'good' data for each dataset in the study and assessed the 95% confidence interval for the mean percentage value. We calculated the percentage of not-rejected ('good') data using Artifact Subspace Reconstruction (ASR) [10-12] with parameters: *BurstCriterion* 20, *WindowCriterion* 0.25, and Euclidean *Distance*. First, 'bad' data points were removed (rather than corrected). Then, 'bad' data windows (of default length 1 second with 66% overlap) were removed (*WindowCriterionTolerances* with range [-Inf 7]).

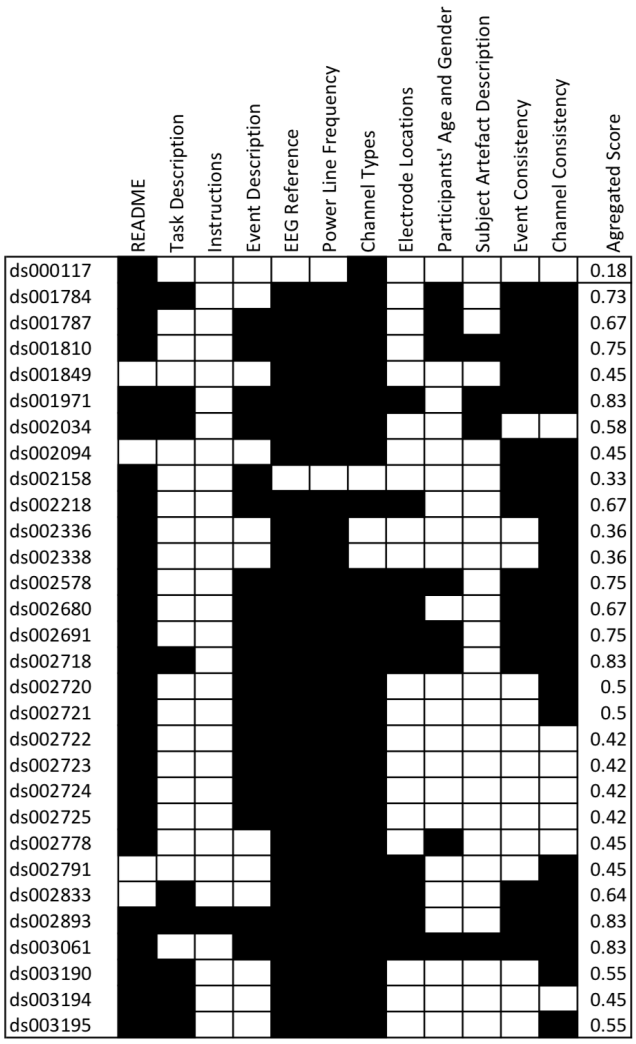

Table 1. *Advanced checking of BIDS study content using 12 questions (white cells = yes; black cells = no), and percent 'yes' (right column). (see Methods). Only 4 studies were 10/12 positive.*

- *How many independent components are 'brain-based'?* The percentage of component processes identified by Independent Component Analysis (ICA) decomposition that are compatible with activity originating within a limited brain area can indicate the quality of the data; data with fewer of these typically include more (spatially non-stereotyped) noise [13]. The plug-in thus computed the percentage of brain-based independent components for each dataset of a study, and we then calculated the 95% confidence interval of the mean percentage. Before decomposition, we applied the data cleaning steps outlined above, high-pass filtered the data (FIR, transition band 0.25 to 0.75 Hz) and converted the data to average reference. ICA decomposition used *Picard* (Infomax ICA with Newton descent available as EEGLAB plugin *Picard* v1.0) [14]), after reducing the data dimension by 1 by PCA. *Picard* is a fast decomposition approach comparable with the *runica* and *AMICA* algorithms standard previously used in EEGLAB [13]. We then used the EEGLAB plug-in *ICLabel* (v1.2.6) [15] to classify the type of each component, using a normed log-likelihood (probability) threshold of 60% for labeling an independent component process as 'brain' generated (rather than non-brain or 'other').

Table 2 shows the results of data quality assessment on the 30 usable BIDS-EEG studies now in OpenNeuro. For each study, we also indicate the numbers of EEG files and channels and the number of empty ('flat') EEG files. These computations required about 100k core hours on the Comet and Expanse supercomputers of the San Diego Supercomputer Center.

Dataset	n	Flat	Chan.	Good chan.	Good data	Brain ICs
ds000117	84	-	74	89 - 90	68 - 78	28 - 32
ds001784	30	-	60	85 - 92	63 - 77	5 - 11
ds001787	40	-	64	93 - 95	85 - 90	24 - 30
ds001810	263	-	64	78 - 80	71 - 74	35 - 37
ds001849	120	-	30	86 - 88	66 - 70	41 - 45
ds001971	273	-	112*	91 - 95	81 - 84	8 - 10
ds002034	167	21	62	95 - 98	47 - 52	29 - 33
ds002094	43	-	30	88 - 93	66 - 77	37 - 45
ds002158	8	-	63	80 - 86	19 - 66	11 - 23
ds002218	18	-	32	94 - 97	66 - 78	35 - 46
ds002336	54	1	63	81 - 83	74 - 89	33 - 39
ds002338	85	-	63	74 - 77	87 - 95	24 - 28
ds002578	2	-	18	78 - 89	96 - 97	46 - 60
ds002680	350	-	19	92 - 93	81 - 84	53 - 56
ds002691	20	-	32	93 - 97	82 - 87	30 - 40
ds002718	18	-	74	95 - 96	58 - 76	20 - 28
ds002720	165	-	19	93 - 95	68 - 72	52 - 56
ds002721	185	-	19	95 - 96	67 - 72	45 - 49
ds002722	94	-	32	92 - 95	63 - 70	36 - 39
ds002723	44	-	32	95 - 97	66 - 76	36 - 40
ds002724	96	-	32	94 - 96	72 - 77	35 - 39
ds002725	105	-	31	90 - 94	80 - 85	53 - 58
ds002778	46	-	32	90 - 95	68 - 78	41 - 48
ds002791	92	-	256	90 - 92	62 - 70	6 - 7
ds002833	80	-	256	91 - 94	82 - 88	6 - 8
ds002893	55	12	36	89 - 93	76 - 83	29 - 37
ds003061	39	-	64	84 - 89	86 - 92	22 - 28
ds003190	384	2	8	81 - 83	89 - 91	74 - 78
ds003194	29	-	19	93 - 97	90 - 95	72 - 79
ds003195	20	-	19	90 - 97	89 - 94	65 - 76

* Some datasets had 108 EEG channels

Table 2. *Quality statistics for all BIDS-EEG formatted studies on OpenNeuro. n: the number of EEG datasets (possibly more than one per participant). Flat: the number of empty datasets. Chan: the number of EEG channels. Good chan., Good data, Brain ICs: 95% confidence intervals (across datasets) for the percent 'good' channels, 'good' data points, and ICA components of brain origin, respectively (see Methods).*

Figure 2 shows histograms of 3 quality metrics across the 30 BIDS-EEG studies. Most of the studies have 90% to 100% 'good' channels and 60% to 90% 'good' data points.

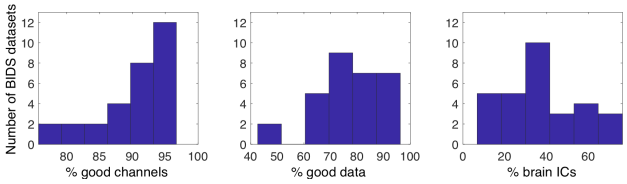

Figure 2. *Histograms showing numbers of BIDS studies for three quality measures: percent 'good' channels, percent 'good' data, and*

percent estimated 'brain- based' independent components (see Methods).

III. Discussion

Here we report EEGLAB-based tools for assessing the quality of metadata and data made available in BIDS-EEG studies and reported their application to the 30 analyzable studies made available to date on the *OpenNeuro.org* human neuroimaging data archive. The development of these tools is timely, as the National Institute of Health (NIH) recently updated their policy [16] to push researchers to publish their data collected using public funding.

Surprisingly, although the BIDS-EEG structure itself enforces the presence of some metadata in archived studies, we found that data quality across the submitted studies was nonetheless sub-optimal. 25% of the archived BIDS-EEG studies on OpenNeuro unusable. Data quality varied widely: 4 of the 30 analyzable BIDS-EEG studies contained some empty EEG files. Metadata essential for EEG analysis was often omitted for other BIDS studies.

For the 30 studies we could analyze, the fraction of analyzable channels and data points were within the range we would expect for EEG data. We noticed that a larger percentage of 'brain' based component processes were identified in studies using fewer scalp channels (ds003190; ds003194; ds003195) – a possible bias of the *ICLabel* tool, which processes component scalp maps after interpolation. Scalp maps for independent components of datasets with low channel count tend to be smoother and thus may be more likely to be classified as 'brain' based; they may also need to include more types of activity than independent components of high-channel number datasets.

IV. Conclusion

The metrics presented in this article will be made available on *NEMAR.org* (NEuroelectroMagnetic data Archive and tools Resource), a portal to *OpenNeuro.org* for the EEG research community. Metadata quality metrics will also be computed when users export BIDS data from EEGLAB using the EEGLAB *bids-matlab-tools* plug-in (v5.2 and higher). We hope the data quality metrics presented in this paper will help researchers contribute high-quality BIDS-EEG studies for both personal and public use for advanced and large-scale analysis.

Acknowledgment

This work was supported by NIH grant 5R24-MH120037-02. Supercomputer time was provided via XSEDE allocations.

References

1. Gorgolewski, K.J., et al., *The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments.* Scientific Data, 2016. **3**: p. 160044.
2. Gorgolewski, K., et al., *OpenNeuro—a free online platform for sharing and analysis of neuroimaging data.* Organization for human brain mapping. Vancouver, Canada, 2017. **1677**(2).
3. Pernet, C.R., et al., *EEG-BIDS, an extension to the brain imaging data structure for electroencephalography.* Scientific Data, 2019. **6**(1): p. 1-5.
4. Niso, G., et al., *MEG-BIDS, the brain imaging data structure extended to magnetoencephalography.* Scientific data, 2018. **5**.
5. Holdgraf, C., et al., *BIDS-iEEG: an extension to the brain imaging data structure (BIDS) specification for human intracranial electrophysiology.* 2018.
6. Holdgraf, C., et al., *iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology.* Scientific Data, 2019. **6**.
7. Delorme, A. and S. Makeig, *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.* Journal of neuroscience methods, 2004. **134**(1): p. 9-21.
8. Robbins, K., et al., *Building FAIR functionality: Annotating event-related imaging data using Hierarchical Event Descriptors (HED).* 2020.
9. *Bids-Standard/Bids-Validator*. 2020 [cited 2020 11/20/2020]; Git repository for Bids-Validator tools]. Available from: https://github.com/bids-standard/bids-validator.
10. Kothe, C.A.E. and T.-P. Jung, *Artifact removal techniques with signal reconstruction*. 2016, Google Patents.
11. Chang, C.-Y., et al., *Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings.* IEEE Transactions on Biomedical Engineering, 2019. **67**(4): p. 1114-1121.
12. Chang, C.-Y., et al. *Evaluation of artifact subspace reconstruction for automatic EEG artifact removal*. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018. IEEE.
13. Delorme, A., et al., *Independent EEG sources are dipolar.* PloS one, 2012. **7**(2): p. e30135.
14. Ablin, P., J.-F. Cardoso, and A. Gramfort, *Faster independent component analysis by preconditioning with Hessian approximations.* IEEE Transactions on Signal Processing, 2018. **66**(15): p. 4040-4049.
15. Pion-Tonachini, L., K. Kreutz-Delgado, and S. Makeig, *ICLabel: An automated electroencephalographic independent component classifier, dataset, and website.* NeuroImage, 2019. **198**: p. 181-197.
16. *NIH Data Sharing Policy and Implementation Guidance*. 2020 [cited 2020 11/20/2020]; NIH guidelines for data sharing policy]. Available from: https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.