NeuroImage 144 (2017) 275–286

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The Neuro Bureau ADHD-200 Preprocessed repository

Pierre Bellec[a, b, c, *], Carlton Chu[a, d], François Chouinard-Decorte[a, b, e], Yassine Benhajali[a, b, f], Daniel S. Margulies[a, g], R. Cameron Craddock[a, h, i, *]

[a] The Neuro Bureau, Germany
[b] Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Canada
[c] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada
[d] Google DeepMind, London, UK
[e] Integrated Program in Neuroscience, McGill University, Montreal, Canada
[f] Département d'Anthropologie, Université de Montréal, Montréal, Canada
[g] Max Planck Research Group for Neuroanatomy & Connectivity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[h] Computational Neuroimaging Laboratory, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA
[i] Center for the Developing Brain, Child Mind Institute, New York, NY, USA

ARTICLE INFO

Article history:
Received 25 December 2015
Received in revised form 28 May 2016
Accepted 17 June 2016
Available online 15 July 2016

Keywords:
Preprocessed fMRI
Data sharing
Neuroimaging competition

ABSTRACT

In 2011, the "ADHD-200 Global Competition" was held with the aim of identifying biomarkers of attention-deficit/hyperactivity disorder from resting-state functional magnetic resonance imaging (rs-fMRI) and structural MRI (s-MRI) data collected on 973 individuals. Statisticians and computer scientists were potentially the most qualified for the machine learning aspect of the competition, but generally lacked the specialized skills to implement the necessary steps of data preparation for rs-fMRI. Realizing this barrier to entry, the Neuro Bureau prospectively collaborated with all competitors by preprocessing the data and sharing these results at the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) (http://www.nitrc.org/frs/?group_id=383). This "ADHD-200 Preprocessed" release included multiple analytical pipelines to cater to different philosophies of data analysis. The processed derivatives included denoised and registered 4D fMRI volumes, regional time series extracted from brain parcellations, maps of 10 intrinsic connectivity networks, fractional amplitude of low frequency fluctuation, and regional homogeneity, along with grey matter density maps. The data was used by several teams who competed in the ADHD-200 Global Competition, including the winning entry by a group of biostaticians. To the best of our knowledge, the ADHD-200 Preprocessed release was the first large public resource of preprocessed resting-state fMRI and structural MRI data, and remains to this day the only resource featuring a battery of alternative processing paths.

Introduction

In 2011, the "ADHD-200 Global Competition" was held with the aim of engaging researchers from a variety of analytical backgrounds to identify biomarkers of attention-deficit/hyperactivity disorder (ADHD) from resting-state functional magnetic resonance imaging (rs-fMRI) and structural MRI (s-MRI) data (Milham et al., 2012). The competition made use of the "ADHD-200 Sample" data collection that was aggregated from eight independent sites and shared through the Intenational Neuroimaging Datasharing Initiative (INDI) (Mennes et al., 2013). The data includes rs-fMRI, structural MRI (s-MRI), and basic phenotypic information for 973 individuals: some typically-developing controls (TDC) and patients diagnosed with ADHD (Milham et al., 2012). Competitors were given five and a half months to optimize a classification algorithm on training data (776 individuals) and submit their predicted clinical labels on test data for which diagnostic information was withheld. The competition data was distributed in a raw form and, before any analysis could begin, the images had to be preprocessed to make them comparable across individuals and reduce noise. These preprocessing steps present a significant hurdle for would-be competitors who do not have the specialist knowledge of neuroimaging methods, or access to high

* Corresponding authors.
E-mail addresses: pierre.bellec@criugm.qc.ca (P. Bellec), carltonchu1@gmail.com (C. Chu), francois.chouinard@gmail.com (F. Chouinard-Decorte), yanamarji@gmail.com (Y. Benhajali), margulies@cbs.mpg.de (D. Margulies), ccraddock@nki.rfmh.org (R. Craddock).
URLs: http://www.bellec.simexp-lab.org (P. Bellec), http://www.computational-neuroimaging-lab.org (R. Craddock).

http://dx.doi.org/10.1016/j.neuroimage.2016.06.034
1053-8119/

performance computing resources. Realizing this barrier to entry, the Neuro Bureau, a non-profit organization aimed at facilitating open science grassroots initiatives,[1] prospectively collaborated with all competitors by preprocessing the data and sharing these results.

The "ADHD-200 Preprocessed" is a repository of preprocessed rs-fMRI and s-MRI data along with statistical derivatives from the ADHD-200 Sample. Rather than favoring a specific processing strategy, we followed a pluralistic approach by preprocessing the data using multiple pipelines (called "Athena", "Burner", and "NIAK") that differed in the toolsets used, the philosophy motivating choices of algorithms and parameters, and the statistical derivatives calculated. The Athena pipeline processed rs-fMRI and s-MRI images using a combination of AFNI (Cox, 1996) and FSL (Smith et al., 2004) neuroimaging toolkits. The Burner pipeline used SPM8 (Ashburner et al., 2012) to process s-MRI data for voxel-based morphometry. The NIAK pipeline processed rs-fMRI and s-MRI using the NeuroImaging Analysis Kit (Bellec et al., 2011).

Organization and access to the repository

The ADHD-200 Preprocessed data was released in 2011 and can be downloaded from NITRC.[2] No data usage agreement is required to access or download the data, the only requirement is registering for a free NITRC account. This registration enables downloads to be tracked for usage statistics and users to be contacted in the event that errors are found in the dataset. The ADHD-200 Sample allows unrestricted data usage for non-commercial research purposes provided that the specific datasets included in an analysis be cited appropriately and that their funding sources be acknowledged.[3] There are no more restrictions placed on the preprocessed data or derivatives other than the request that the ADHD-200 Preprocessed Initiative is cited appropriately and that the specific pipeline is acknowledged in publications using the data. A forum is available on the Neuro Bureau's NITRC project page for users to ask questions or report problem.[4] Questions regarding data acquisition or phenotypic variables should be directed to INDI's support forum.[5]

Contents of the repository

The ADHD-200 Preprocessed repository contains preprocessed outputs and derivatives for data from the ADHD-200 Sample, which includes 973 individuals (352 F) between the ages of 7 and 27 aggregated from 17 different studies conducted across 8 different sites (for a breakdown of age and sex by diagnosis, see Table 1). For each individual, phenotypic data includes sex, age, handedness, ADHD diagnosis (585 TDC, 362 ADHD, 26 with diagnosis unavailable), ADHD subtype (ADHD-combined, ADHD-inattentive, ADHD-hyperactive/impulsive), one of three different measures of ADHD severity, one of five measures of intelligence, co-morbid diagnoses, and whether or not they have used medication to treat their symptoms (Milham et al., 2012). Imaging data for each individual includes one or more T1-weighted high-resolution s-MRI scan (s-MRI) and one or more rs-fMRI scan. The majority of data was acquired during a single imaging session, although a second session is available for 15 individuals from the Washington University at Saint Louis (WUSTL) site. There is a substantial amount of variation in data acquisition procedures across sites including the type of MRI system and scanning parameters, the length of the rs-fMRI scans, and the instructions given to participants prior to the scan (see Tables 2 and 3).

Nearly all of the imaging data from the ADHD-200 Sample was included in the preprocessing effort, though some individuals were excluded for poor quality or missing data.[6] The results of the preprocessing are made available as a collection of compressed tar files that are organized by pipeline, sites of data collection, training and test samples, as well as by derivatives. A group-level file containing the phenotypic data is available in comma-separated-values format (.csv).

Shared preprocessed data and extracted features include:

- 3D grey matter density maps suitable for voxel-based morphometry – Athena and Burner (see Fig. 1),
- 4D preprocessed resting-state fMRI data including limited intermediaries and quality assessment – Athena and NIAK,
- Average time series for brain regions from structurally defined parcellations – Athena (see Figs. 2 and 3),
- Average time series for brain regions for regions defined by functional parcellation – Athena and NIAK (see Figs. 2 and 3),
- Spatial maps for 10 intrinsic connectivity networks (ICNs), fractional amplitude of low frequency fluctuations (fALFF), and regional homogeneity (ReHo) – Athena (see Fig. 4).

Athena pipeline

The Athena pipeline[7] processed rs-fMRI and s-MRI images using a custom BASH script that combined AFNI (Cox, 1996) and FSL (Smith et al., 2004) neuroimaging toolkits and was run on the Athena computer cluster at Virginia Tech's Advanced Research Computing center.[8] The processing scripts for each site are distributed in the repository, as well as on Github,[9] along with output log files for each processed dataset.

Structural processing

Athena's s-MRI pipeline began with skull-stripping to remove non-brain tissue and background from the images (Smith, 2002) and segmenting the results into white matter (WM), cerebro-spinal fluid (CSF), and grey matter (GM) probability maps (Zhang et al., 2001). A non-linear warp was calculated between the skull-off image and MNI space as represented by the NIHPD 4.5–18.5y age-specific assymmetric template (Fonov et al., 2011) using a two step procedure that calculates a linear transform (Jenkinson et al., 2002), that is subsequently refined by a non-linear registration procedure (Andersson et al., 2007). *Shared s-MRI outputs include:* skull-stripped whole-brain images and smoothed (by a 6 mm FWHM Gaussian) and unsmoothed GM density maps in MNI space at $1\times1\times1$ mm^3 resolution, along with the FSL fNIRT non-linear warp, as compressed NIfTI files (.nii.gz).

Functional processing

Preprocessing. Athena's rs-fMRI pipeline involved removing the first four volumes to allow for magnetization to reach equilibrium, site-specific slice timing correction to the middle slice, re-aligning each volume to the first volume to correct for motion (Cox and Jesmanowicz, 1999), and calculating a linear transform between the mean functional volume and the corresponding s-MRI (Jenkinson et al., 2002). The rs-fMRI to s-MRI transform was then combined with the s-MRI to MNI non-linear warp to write the functional data into

[1] See http://www.neurobureau.org/mission-statement/ for the full mission statement. The Neurobureau is a non-profit organization registered in Germany.
[2] http://www.nitrc.org/frs/?group_id=383.
[3] http://fcon_1000.projects.nitrc.org/indi/ADHD-200/.
[4] http://www.nitrc.org/forum/forum.php?forum_id=2046.
[5] http://www.nitrc.org/forum/forum.php?forum_id=1735.
[6] Further information regarding excluded data can be found at the respective pipeline wiki page: Athena: http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline#Excluded_Data; Burner: http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:BurnerPipeline; NIAK: http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:NIAKPipeline.
[7] http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline.
[8] http://www.arc.vt.edu/.
[9] https://github.com/preprocessed-connectomes-project/adhd200_athena_scripts.