Data in Brief 22 (2019) 570–573

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Identifying neuropsychiatric disorders using unsupervised clustering methods: Data and code

Xinyu Zhao [a], D. Rangaprakash [a,b], Thomas S. Denney Jr. [a,c,d,h], Jeffrey S. Katz [a,c,d,h], Michael N. Dretsch [e,f], Gopikrishna Deshpande [a,c,d,g,h,*]

[a] AU MRI Research Center, Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA
[b] Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, USA
[c] Department of Psychology, Auburn University, Auburn, AL, USA
[d] Alabama Advanced Imaging Consortium, Birmingham, USA
[e] Human Dimension Division, HQ TRADOC, Fort Eustis, VA, USA
[f] US Army Aeromedical Research Laboratory, Fort Rucker, AL, USA
[g] Center for Health Ecology and Equity Research, Auburn University, USA
[h] Center for Neuroscience, Auburn University, USA

ARTICLE INFO

Article history:
Received 19 August 2017
Received in revised form
10 January 2018
Accepted 29 January 2018
Available online 2 February 2018

Keywords:
Functional magnetic resonance imaging
Functional connectivity
Effective connectivity
Unsupervised learning
Clustering
Genetic algorithm
Psychiatric disorders

ABSTRACT

This article provides data for five different neuropsychiatric disorders—Attention Deficit Hyperactivity Disorder, Alzheimer's Disease, Autism Spectrum Disorder, Post-Traumatic Stress Disorder, and Post-Concussion Syndrome–along with healthy controls. The data includes clinical diagnostic labels, phenotypic variables, and resting-state functional magnetic resonance imaging connectivity features obtained from individuals. In addition, it provides the source MATLAB codes used for data analyses. Three existing clustering methods have been incorporated into the provided code, which do not require *a priori* specification of the number of clusters. A genetic algorithm based feature selection method has also been included to find the relevant subset of features and clustering the subset of data simultaneously. Findings from this data set and further detailed interpretations are available in our recent research study (Zhao et al., 2017) [1]. This contribution is a valuable asset for performing unsupervised machine learning on fMRI data to

* Correspondence to: Auburn University, 560 Devall Dr, Suite 266D, Auburn, AL 36849, USA. Fax: +1 334 844 0214.
E-mail address: gopi@auburn.edu (G. Deshpande).

https://doi.org/10.1016/j.dib.2018.01.080

investigate the correspondence of clinical diagnostic grouping with the underlying neurobiological/phenotypic clusters.

Specifications Table

Subject area	Brain imaging
More specific subject area	Unsupervised machine learning applications in functional magnetic resonance imaging of neuropsychiatric disorders
Type of data	Functional MRI connectivity features, Tables, MATLAB code
How data was acquired	Data was acquired using 3T MRI scanners. Details for each data set can be found here: ADHD: http://fcon_1000.projects.nitrc.org/indi/adhd200/ AD: htpp://adni/loni.ucla.edu ASD: http://fcon_1000.projects.nitrc.org/indi/abide/index.html PTSD/PCS: Siemens MAGNETOM Verio 3T MRI scanner at Auburn University MRI research center, Auburn, AL, USA
Data format	Processed data in MATLAB (*.mat) format
Experimental factors	Resting state fMRI connectivity features from ADHD, AD, ASD, PTSD and PCS subjects, with matched healthy controls
Experimental features	Resting-state: participants were requested to have their eyes open and fixated on a white cross, displayed on a dark background on the display. They were asked to not think of anything specific.
Data source location	AU MRI research center, Auburn University, AL, USA (GPS coordinates: 32.586, − 85.494)
Data accessibility	Data is available with this article and also in a public repository: https://github.com/xinyuzhao/identification-of-brain-based-disorders.git

Value of the data

- We describe an analysis pipeline that can identify different neuropsychiatric disorders in an unsupervised fashion and measure their correspondence with clinically defined labels as well as phenotypic clusters.
- This approach can be used to develop imaging biomarkers of neuropsychiatric disorders, which could potentially be used as an aid by the clinician, in addition to currently available subjective markers, to improve diagnostic precision.
- The data and code provided in this article can be used for reproducing the results in the research article entitled “Investigating the correspondence of clinical diagnostic grouping with underlying neurobiological and phenotypic clusters using unsupervised learning” [1]. They can also be applied to study other neuropsychiatric disorders.
- Cluster labels identified using the proposed analysis pipeline can be used by other researchers to improve upon our clustering results.

1. Data

Four different datasets are presented in this article. Each dataset includes:

a) Clinical diagnostic labels.
b) Phenotypic variables.
c) Top significant connectivity features ($p < 0.01$) and corresponding feature names.
d) Regions-of-interest (ROIs) table, which includes region names and coordinates.

The codes are composed in MATLAB, which contains implementation of three existing unsupervised clustering methods, i.e., hierarchical clustering [2], ordering points to identify the clustering structure (OPTICS) [3], and density peak clustering (DPC) [4], along with genetic algorithm (GA)-based feature selection [5].

2. Experimental design, materials and methods

In this work, a general pipeline has been developed that can cluster subjects into different neuropsychiatric disorders in an unsupervised way, along with further characterizing the similarity of such neurobiologically-informed clusters with phenotypic/behavioral clusters and clinically determined clusters. Four different datasets were used. Three of them – ADNI (Alzheimer's Disease Neuroimaging Initiative), ADHD-200, and ABIDE (Autism Brain Imaging Data Exchange) – are publicly available. The fourth one – posttraumatic stress disorder (PTSD) and post-concussion syndrome (PCS) – was acquired in-house. The proposed pipeline was verified using these datasets which contains the following different disorders: Attention Deficit Hyperactivity Disorder (ADHD) from ADHD-200, Alzheimer's Disease (AD) from ADNI, Autism Spectrum Disorder (ASD) from ABIDE, PTSD and PCS from data acquired in-house. The following four different connectivity matrices obtained from functional magnetic resonance imaging (fMRI) data were used to define the neurobiologically informed feature space: statistic functional connectivity (SFC), variance of dynamic functional connectivity (vDFC) [6], statistic effective connectivity (SEC) [7], and variance of dynamic effective connectivity [8]. Hierarchical clustering, DPC and OPTICS were separately used as the clustering algorithms. GA was used to select the most optimal set of features, which maximizes the similarity between clusters obtained from connectivity, phenotype/behavior and clinical labels.

Further details about the data acquisition and analysis pipeline are presented in the research paper associated with this data release [1]. The documentation of data and code are presented within this article, as well as in GitHub repository: https://github.com/xinyuzhao/identification-of-brain-based-disorders.git.

Disclosures

The authors report no competing interests.

Transparency document. Supplementary material

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.01.080.

References

[1] X. Zhao, D. Rangaprakash., B. Yuan, T. S. Denney Jr, J. S. Katz, M. N. Dretsch, G. Deshpande, Investigating the Correspondence of Clinical Diagnostic Grouping With Underlying Neurobiological and Phenotypic Clusters Using Unsupervised Machine Learning, Front. App. Math. Stat. 4 (2018), 25 https://www.frontiersin.org/article/10.3389/fams.2018.00025. (ISSN=2297-4687).
[2] S. Dasgupta, P.M. Long, Performance guarantees for hierarchical clustering, J. Comput. Syst. Sci. 70 (2005) 555–569.
[3] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, ACM Sigmod Rec. (1999) 49–60.

[4] A. Rodriguez, A. Laio, Machine learning. Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[5] J.G. Dy, C.E. Brodley, Feature selection for unsupervised learning, J. Mach. Learn. Res. 5 (2004) 845–889.
[6] H. Jia, X. Hu, G. Deshpande, Behavioral relevance of the dynamics of the functional brain connectome, Brain Connect. 4 (9) (2014) 741–759.
[7] G. Deshpande, P. Santhanam, X. Hu, Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data, NeuroImage 54 (2) (2011) 1043–1052.
[8] M.M. Grant, K. Wood, K. Sreenivasan, M. Wheelock, D. White, J. Thomas, D.C. Knight, G. Deshpande, Influence of early life stress on intra- and extra-amygdaloid causal connectivity, Neuropsychopharmacology 40 (7) (2015) 1–12.