NeuroImage 147 (2017) 736–745

Contents lists available at ScienceDirect

NeuroImage

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

Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example

Alexandre Abraham[a,b,*], Michael P. Milham[e,f], Adriana Di Martino[g], R. Cameron Craddock[e,f], Dimitris Samaras[c,d], Bertrand Thirion[a,b], Gael Varoquaux[a,b]

[a] Parietal Team, Saclay-INRIA le-de-France,Saclay,France
[b] CEA, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[c] Stony Brook University, NY 11794, USA
[d] Ecole Centrale, 92290 Châtenay Malabry, France
[e] Center for the Developing Brain, Child Mind Institute, New York, USA
[f] Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA
[g] The Child Study Center at NYU Langone Medical Center, New York, USA

ARTICLE INFO

Keywords:
Data heterogeneity
Resting-state fMRI
Data pipelines
Biomarkers
Connectome
Autism spectrum disorders

ABSTRACT

Resting-state functional Magnetic Resonance Imaging (R-fMRI) holds the promise to reveal functional biomarkers of neuropsychiatric disorders. However, extracting such biomarkers is challenging for complex multi-faceted neuropathologies, such as autism spectrum disorders. Large multi-site datasets increase sample sizes to compensate for this complexity, at the cost of uncontrolled heterogeneity. This heterogeneity raises new challenges, akin to those face in realistic diagnostic applications. Here, we demonstrate the feasibility of inter-site classification of neuropsychiatric status, with an application to the Autism Brain Imaging Data Exchange (ABIDE) database, a large (N=871) multi-site autism dataset. For this purpose, we investigate pipelines that extract the most predictive biomarkers from the data. These R-fMRI pipelines build participant-specific connectomes from functionally-defined brain areas. Connectomes are then compared across participants to learn patterns of connectivity that differentiate typical controls from individuals with autism. We predict this neuropsychiatric status for participants from the same acquisition sites or different, unseen, ones. Good choices of methods for the various steps of the pipeline lead to 67% prediction accuracy on the full ABIDE data, which is significantly better than previously reported results. We perform extensive validation on multiple subsets of the data defined by different inclusion criteria. These enables detailed analysis of the factors contributing to successful connectome-based prediction. First, prediction accuracy improves as we include more subjects, up to the maximum amount of subjects available. Second, the definition of functional brain areas is of paramount importance for biomarker discovery: brain areas extracted from large R-fMRI datasets outperform reference atlases in the classification tasks.

1. Introduction

In psychiatry, as in other fields of medicine, both the standardized observation of signs, as well as the symptom profile are critical for diagnosis. However, compared to other fields of medicine, psychiatry lacks accompanying objective markers that could lead to more refined diagnoses and targeted treatment (Kapur et al., 2012). Advances in non-invasive brain imaging techniques and analyses (e.g. Craddock et al. (2013); Van Essen and Ugurbil (2012)) are showing great promise for uncovering patterns of brain structure and function that can be used as objective measures of mental illness. Such *neurophenotypes* are important for clinical applications such as disease staging, determination of risk prognosis, prediction and monitoring of treatment response, and aid towards diagnosis (e.g. Castellanos et al. (2013)).

Among the many imaging techniques available, resting-state fMRI (R-fMRI) is a promising candidate to define functional neurophenotypes (Kelly et al., 2008; Van Essen and Ugurbil, 2012). In particular, it is non-invasive and, unlike conventional task-based fMRI, it does not require a constrained experimental setup nor the active and focused participation of the subject. It has been proven to capture interactions between brain regions that may lead to neuropathology diagnostic biomarkers (Greicius, 2008). Numerous studies have linked variations

* Corresponding author at: Parietal Team, INRIA Saclay-Île-de-France, Saclay, France.
E-mail address: abraham.alexandre@gmail.com (A. Abraham).

http://dx.doi.org/10.1016/j.neuroimage.2016.10.045
Received 20 July 2016; Accepted 21 October 2016
Available online 16 November 2016

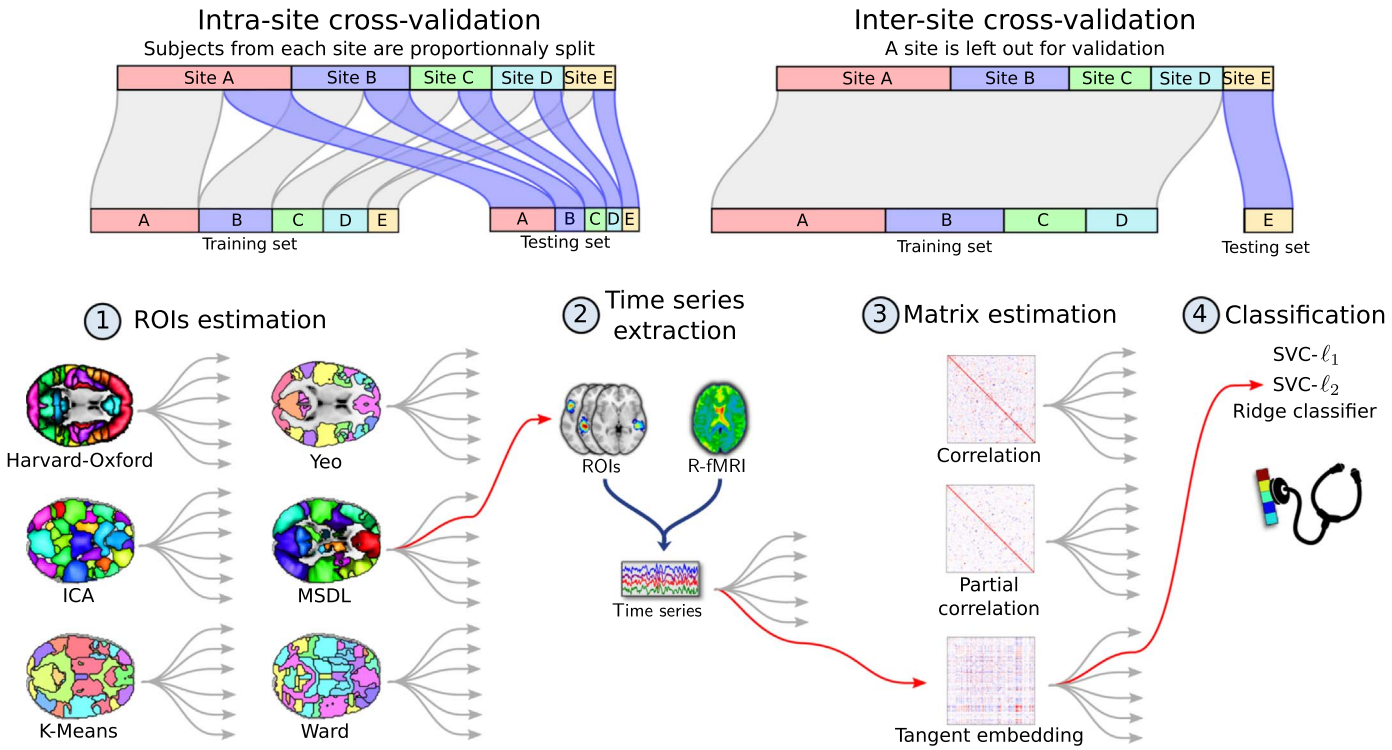

Fig. 1. *Functional MRI analysis pipeline.* Cross-validation schemes used to validate the pipeline are presented above. *Intra-site* cross-validation consists of randomly splitting the participants into training and testing sets while preserving the ratio of samples for each site and condition. *Inter-site* cross-validation consists of leaving out participants from an entire site as testing set. In the first step of the pipeline, regions of interest are estimated from the training set. The second step consists of extracting signals of interest from all the participants, which are turned into connectivity features via covariance estimation at the third step. These features are used in the fourth step to perform a supervised learning task and yield an accuracy score. An example of pipeline is highlighted in red. This pipeline is the one that gives best results for inter-site prediction. Each model is decribed in the section relative to material and methods.

in brain functional architecture measured from R-fMRI to behavioral traits and mental health conditions such as Alzheimer disease (e.g. Greicius et al. (2004); Chen et al. (2011), Schizophrenia (e.g. Garrity et al. ; Zhou et al. (2007); Jafri et al. (2008); Calhoun et al. (2011)), ADHD, autism (e.g. Plitt et al., 2015 and others (e.g. Anderson et al. 2011). Extending these findings, predictive modeling approaches have revealed patterns of brain functional connectivity that could serve as biomarkers for classifying depression (e.g. Craddock et al.), ADHD (e.g. Consortium et al.), autism (e.g. Anderson et al. (2011)), and even age (Dosenbach et al., 2010). This growing number of studies has shown the feasibility of using R-fMRI to identify biomarkers. However questions about the readiness of R-fMRI to detect clinically useful biomarkers remain (Plitt et al., 2015). In particular, the reproducibility and generalizability of these approaches in research or clinical settings are debatable. Given the modest sample size of most R-fMRI studies, the effect of cross-study differences in data acquisition, image processing, and sampling strategies (Desmond and Glover, 2002; Murphy and Garavan, 2004; Thirion et al., 2007) has not been quantified.

Using larger datasets is commonly cited as a solution to challenges in reproducibility and statistical power (Button et al., 2013). They are considered a prerequisite to R-fMRI-based classifiers for the detection of psychiatric illness. Recent efforts have accelerated the generation of large databases through sharing and aggregating independent data samples (Fair et al., Mennes et al., 2013; Di Martino et al., 2014). However, a number of concerns must be addressed before accepting the utility of this approach. Most notably, the many potential sources of uncontrolled variation that can exist across studies and sites, which range from MRI acquisition protocols (e.g. scanner type, imaging sequence, see Friedman et al. (2008)), to participant instructions (e.g. eyes open vs. closed, see Yan et al. (2013)), to recruitment strategies (age-group, IQ-range, level of impairment, treatment history and acceptable comorbidities). Such variation in aggregate samples is often viewed as dissuasive, as its effect on diagnosis and biomarker extraction is unknown. It commonly motivates researchers to limit the number of sites included in their analyses at the cost of sample size.

Cross-validated results obtained from predictive models are more robust to inhomogeneities: they measure model generalizability by applying it to unseen data, i. e. , data not used to train the model. In particular, leave-out cross-validation strategies, which remove single individuals (or random subsets), are common in biomarkers studies. However, these strategies do not measure the effect of potential site-specific confounds. In the present study we leverage aggregated R-fMRI samples to address this problem. Instead of leaving out random subsamples as test sets, we left out entire sites to measure performance in the presence of uncontrolled variability.

Beyond challenges due to inter-site data heterogeneity, choices in the functional-connectivity data-processing pipeline further add to the variability of results (Carp, 2012; Yan et al., 2013; Shirer et al.,). While preprocessing procedures are now standard, the different steps of the prediction pipeline vary from one study to another. These entail specifying regions of interest, extracting regional time courses, computing connectivity between regions, and identifying connections that relate to subject's phenotypes (Craddock et al.,, Richiardi et al., ; Shirer et al., 2012; Eickhoff et al., 2015).

Lack of ground truth for brain functional architecture undermines the validation of R-fMRI data-processing pipelines. The use of functional connectivity for individual prediction suggests a natural figure of merit: prediction accuracy. We contribute quantitative evaluations, to help settling down on a parameter-free pipeline for R-fMRI. Using efficient implementations, we were able to evaluate many pipeline options and select the best method to estimate atlases, extract connectivity matrices, and predict phenotypes.

To demonstrate that pipelines to extract R-fMRI neuro-phenotypes can reliably learn inter-site biomarkers of psychiatric status on