NeuroImage 149 (2017) 220–232

Contents lists available at ScienceDirect

NeuroImage

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

Statistical power and prediction accuracy in multisite resting-state fMRI connectivity

Christian Dansereau[a,b], Yassine Benhajali[a,c], Celine Risterucci[d], Emilio Merlo Pich[d,1], Pierre Orban[a], Douglas Arnold[e], Pierre Bellec[a,b,*]

[a] Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Canada
[b] Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada
[c] Département d'anthropologie, Université de Montréal, Montréal, Canada
[d] Clinical Imaging, pRED, F.Hoffman-La Roche, Basel, Switzerland
[e] NeuroRx inc., Montréal, Canada

ARTICLE INFO

Keywords:
Multisite
Statistical power
Prediction accuracy
Monte-Carlo simulation
Sample size
Resting-state
FMRI connectivity
SVM

ABSTRACT

Connectivity studies using resting-state functional magnetic resonance imaging are increasingly pooling data acquired at multiple sites. While this may allow investigators to speed up recruitment or increase sample size, multisite studies also potentially introduce systematic biases in connectivity measures across sites. In this work, we measure the inter-site effect in connectivity and its impact on our ability to detect individual and group differences. Our study was based on real, as opposed to simulated, multisite fMRI datasets collected in N=345 young, healthy subjects across 8 scanning sites with 3 T scanners and heterogeneous scanning protocols, drawn from the 1000 functional connectome project. We first empirically show that typical functional networks were reliably found at the group level in all sites, and that the amplitude of the inter-site effects was small to moderate, with a Cohen's effect size below 0.5 on average across brain connections. We then implemented a series of Monte-Carlo simulations, based on real data, to evaluate the impact of the multisite effects on detection power in statistical tests comparing two groups (with and without the effect) using a general linear model, as well as on the prediction of group labels with a support-vector machine. As a reference, we also implemented the same simulations with fMRI data collected at a single site using an identical sample size. Simulations revealed that using data from heterogeneous sites only slightly decreased our ability to detect changes compared to a monosite study with the GLM, and had a greater impact on prediction accuracy. However, the deleterious effect of multisite data pooling tended to decrease as the total sample size increased, to a point where differences between monosite and multisite simulations were small with N=120 subjects. Taken together, our results support the feasibility of multisite studies in rs-fMRI provided the sample size is large enough.

1. Introduction

Main objective. Multisite studies are becoming increasingly common in resting-state functional magnetic resonance imaging (rs-fMRI). In particular, some consortia have retrospectively pooled rs-fMRI data from multiple independent studies comparing clinical cohorts with control groups, e.g. normal controls in the 1000 functional connectome project (FCP) (Biswal et al., 2010), children and adolescents suffering from attention deficit hyperactivity disorder from the ADHD200 (Milham et al., 2012; Fair et al., 2012), individuals diagnosed with autism spectrum disorder in ABIDE (Nielsen et al., 2013), individuals suffering from schizophrenia (Cheng et al., 2015), or elderly subjects suffering from mild cognitive impairment (Tam et al., 2015). The rationale behind such initiatives is to dramatically increase the sample size at the cost of decreased sample homogeneity. The systematic variations of connectivity measures derived using different scanners, called site effects, may decrease the statistical power of group comparisons, and somewhat mitigate the benefits of having a large sample size (Brown et al., 2011; Jovicich et al., 2016). In this work, our main objective was to quantitatively assess the impact of site effects on group comparisons in rs-fMRI connectivity.

Group comparison in rs-fMRI connectivity. In this work, we focused on the most common measure of individual functional connectivity, which is the Pearson's correlation coefficient between

* Corresponding author at: Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Canada.
E-mail address: pierre.bellec@criugm.qc.ca (P. Bellec).
[1] Present adress: CNS TAU, Takeda Development Centre Europe, London, UK.

http://dx.doi.org/10.1016/j.neuroimage.2017.01.072
Received 9 July 2016; Accepted 30 January 2017
Available online 02 February 2017

the average rs-fMRI time series of two brain regions. To compare two groups, a general linear model (GLM) is typically used to establish the statistical significance of the difference in average connectivity between the groups. Finally a p-value is generated for each connection to quantify the probability that the difference in average connectivity is significantly different from zero (Worsley and Friston, 1995; Yan et al., 2013). If the estimated p-value is smaller than a prescribed tolerable level of false-positive findings (see for more detail Table 1), generally adjusted for the number of tests performed across connections, say $\alpha = 0.001$, then the difference in connectivity is deemed significant.

Statistical power in group comparisons at multiple sites. The statistical power of a group comparison study is the probability of finding a significant difference, when there is indeed a true difference. A careful study design involves the selection of a sample size that is large enough to reach a set level of statistical power, e.g. 80%. In the GLM, the statistical power actually depends on a series of parameters (Desmond and Glover, 2002; Durnez et al., 2014): (1) the sample size (the larger the better); (2) the absolute size of the group difference (the larger the better), and, (3) the intrinsic variability of measurements (the smaller the better) (4) the rejection threshold α for the null hypothesis.

Sources of variability: factors inherent to the scanning protocol. In a multisite (or multi-protocol) setting, differences in imaging or study parameters may add variance to rs-fMRI measures, e.g. the scanner make and model (Friedman and Glover, 2006; Friedman et al., 2008), repetition time, flip angle, voxel resolution or acquisition volume (Friedman and Glover, 2006), experimental design such as eyes-open/eyes-closed (Yan et al., 2009), experiment duration (Van Dijk et al., 2010), and scanning environment such as sound attenuation measures (Elliott et al., 1999), or head-motion restraint techniques (Edward et al., 2000; Van Dijk et al., 2012), amongst others. These parameters can be harmonized to some extent, but differences are unavoidable in large multisite studies. The recent work of Yan et al. (2013) has indeed demonstrated the presence of significant site effects in rs-fMRI measures in the 1000 FCP. Site effects will increase the variability of measures, and thus decrease statistical power. To the best of our knowledge, it is not yet known how important this decrease in statistical power may be.

Sources of variability: within-subject. The relative importance of site effects in rs-fMRI connectivity depends on the amplitude of the many other sources of variance. First, rs-fMRI connectivity only has moderate-to-good test-retest reliability using standard 10-minute imaging protocols (Shehzad et al., 2009), even when using a single scanner and imaging session. Differences in functional connectivity across subjects are also known to correlate with a myriad of behavioural and demographic subject characteristics (Anand et al., 2007; Sheline et al., 2010; Kilpatrick et al., 2006). Taken together, these sources of variance reflect a fundamental volatility of human physiological signals.

Sources of variability: factors inherent to the site. In addition to physiology, some imaging artefacts will vary systematically from session to session, even at a single site. For example, intensity non-uniformities across the brain depend on the positioning of subjects (Caramanos et al., 2010). Room temperature has also been shown to impact MRI measures (Vanhoutte et al., 2006). Given the good consistency of key findings in resting-state connectivity across sites, such as the organization of distributed brain networks (Biswal et al., 2010), it is reasonable to hypothesize that site effects will be small compared to the combination of physiological and within-site imaging variance.

Multivariate analysis. Another important consideration regarding the impact of site effects on group comparison in rs-fMRI connectivity is the type of method used to identify differences. The concept of statistical power is very well established in the GLM framework, which tests one brain connection at a time (mass univariate testing). However, multivariate methods that combine several or all connectivity values in a single prediction are also widely used and likely affected by the site effects. A popular multivariate technique in rs-fMRI is support-vector machine (SVM) (Cortes and Vapnik, 1995). In this approach, the group sample is split into a training set and a test set. The SVM is trained to predict group labels on the training set, and the accuracy of the prediction is evaluated independently on the test set. The accuracy level of the SVM captures the quality of the prediction of clinical labels from resting-state connectivity, but does not explicitly tell which brain connection is critical for the prediction. The accuracy score can thus be seen as a separability index between the individuals of two groups in high dimensional space. Altogether, the objectives and measures of statistical risk for SVM and GLM are quite different. Because SVM has the ability to combine measures across connections, unlike univariate GLM tests, we hypothesized that the GLM and SVM will be impacted differently by site effects. Even though the accuracy is expected to be lower for the multisite than the monosite configuration, it as been shown that the generalizability of a predictive model to unseen sites is greater for models trained on multisite than monosite datasets as shown by Abraham et al. (2016).

Specific objectives. Our first objective was to characterize, using real data, the amplitude of systematic site effects in rs-fMRI connectivity measures across sites, as a function of within-site variance. We based our evaluation on images generated from independent groups at 8 sites equipped with 3 T scanners, in a subset (N=345) of the 1000 FCP. Our second objective was to evaluate the impact of site effects on the detection power of group differences in rs-fMRI connectivity. To answer this question directly, one would need to scan two different cohorts of participants at least twice, once in a multisite setting and once in a monosite setting. Such an experiment may be too costly to implement for addressing a purely technical objective. As a more feasible alternative, we implemented a series of Monte Carlo simulations, adding synthetic "pathological" effects in the 1000 FCP sample. One interesting feature of the "1000 FCP" dataset is the presence of one large site of ~200 subjects and 7 small sites of ~20 subjects per site. We were therefore able to implement realistic scenarios following either a monosite or a multisite design (with 7 sites), with the same total sample size. Our simulations gave us full control on critical aspects for the detection of group differences, such as the amplitude of the group difference, sample size, and the balancing of groups across sites. We evaluated the ability of detecting group differences both in terms of sensitivity for a GLM and in terms of accuracy for a SVM model.

Table 1
Confusion matrix.

		Detected value: patho	Detected value: no patho
Actual value	patho	True Positive	False Negative
	no patho	False Positive	True Negative

2. Method

2.1. Imaging sample characteristics

The full 1000 FCP sample includes 1082 subjects, with images acquired over 33 sites spread across North America, Europe, Australia and China. As the 1000 FCP is a retrospective study, no effort was made to harmonize population characteristics or imaging acquisition