Statistical power and measurement bias in multi-centric resting-state fMRI connectivity

Christian Dansereau[a,b], Celine Risterucci[c], Emilio Merlo Pich[c], Douglas Arnold[d], Pierre Bellec[a,b,*]

[a] *Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, CA*
[b] *Département d'Informatique et de recherche opérationnelle, Université de Montréal, Montréal,CA*
[c] *F. Hoffmann-La Roche Ldt., Basel, Switzerland*
[d] *NeuroRx, Montreal, Quebec, Canada*

Abstract

Keywords: multisite, multiprotocol, bias, statistical power, sample size, resting-state, fMRI, connectivity

Highlights

- etc

1. Introduction

Main objective. Studies collecting brain images at multiple sites are becoming increasingly common in resting-state functional magnetic resonance imaging (rs-fMRI). In particular, some consortia have retrospectively shared rs-fMRI data from multiple independent studies of comparable populations, with the objective of dramatically increasing the sample size at the cost of decreased sample homogeneity, 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), or individual diagnosed with autism spectrum disorder in ABIDE (Nielsen et al., 2013). Recruitment of patient populations in a limited time frame may also require acquisitions at multiple sites, e.g. the Alzheimers disease neuroimaging initiative (ADNI) (Mueller et al., 2005) (REF2) fBIRN (Friedman et al., 2006; Friedman and Glover, 2006), which is a common practice in pharmaceutical clinical trials at phase II and III [1]. An important benefit of multisite acquisitions is to offer improved generalization compared to single site

*Corresponding author
Email address: pierre.bellec@criugm.qc.ca (Pierre Bellec)
[1] http://www.roche-trials.com/trialDetailsGet.action?studyNumber=BP28248

studies, due to more diversity in scanners and populations. These additional sources of variance may however decrease the statistical power, and somewhat mitigate the benefits of having a large sample size. In this work, our main objective was to quantitatively assess the impact of inter-site variability on statistical power, for rs-fMRI group comparison.

Statistical power. A simple and popular measure of individual resting-state connectivity is the Pearsons correlation coefficient between the average temporal rs-fMRI fluctuations of two brain parcels. To compare two groups, a general linear model (GLM) is typically used to establish statistical difference in average connectivity between the groups, while accounting for possible confounding variables such as age, sex or the amount of head motion during the scan. Based on the parameters estimated by the GLM, a p-value is generated for each connection to quantify the probability that the difference in average connectivity is significantly distant from zero (Worsley and Friston, 1995). If the estimated p-value is smaller than a prescribed false-positive rate, say $\alpha = 0.05$, then the difference in connectivity is deemed significant. Although in classical statistical testing the statistical decision is made purely based on the p-value, or type I errors, the statistical power is as critical as it controls for type II errors, i.e. failing to detect true differences. The statistical power is defined as the probability of finding a significant difference, when there is indeed a true difference. In the GLM, the statistical power in addition to sample size (Desmond and Glover, 2002): (1) the sample size; (2) the absolute size of the effect, i.e. the difference in mean connectivity between groups; and, (3) the variability of measures. The careful design of a study typically involves selecting the sample size in order to achieve a reasonable statistical power, say larger than 80%. But the choice of a multisite vs monosite study design may also impact statistical power by increasing the variability of measures. The recent work of **?** have shown the presence of a significant bias in rs-fMRI measures between site, yet, to the best of our knowledge, the amplitude of this bias and its implications for statistical power calculation are not currently documented in the literature.

Sources of variance in rs-fMRI. The variability of rs-fMRI connectivity measures has both physiological and instrumental origins. Some sources of variability are shared by monosite and multisite studies, while others are specific of multisite studies. We can first note that rs-fMRI connectivity only has moderate-to-good test-rest reliability, when the measure is repeated for the same subjects in the same scanner (Shehzad et al., 2009). Many physiological factors likely contribute to these variations, such as the cognitive state of the subject, the level of alertness/drowsiness, circadian rhythm, hunger, medical regimen, potential neurostimulants, amongst others. There is also an instrumental variability even within subject and within scanner. The thermal noise in rs-fMRI only has a small amplitude (REF), but there are non-uniformity artefacts which have a strong impact on the signal, and will vary from session to session with the positioning of subjects as well as the adjustment of shimming (REF). Another source of within-site variations is the difference in connectivity across subjects.

These variations are substantial and have been associated to a myriad of variables and clinical conditions (REF review). Taken together, the combination of within- and between-subjects variance is expected to have a large amplitude even at a single site. Multisite studies will add additional sources of physiological variations, as populations at different sites may substantially differ in terms of ethnicity, language, diet, socioeconomic status, exposure to pollutants, typical medication, quality of health services, etc. Some of these factors will be present even if stringent and harmonized inclusion/exclusion criteria are applied, e.g. diet or language. In terms of instrumentation, the fMRI measurements across sites can be affected by the scanner make and model (Friedman et al., 2006), sequence parameters such as repetition time, flip angle, or acquisition volume (Friedman and Glover, 2006), experimental design such as eyes-open/eyes-closed (Yan et al., 2009) or experiment duration (Van Dijk et al., 2010), and scanning environment such as sound attenuation measures (Elliott et al., 1999), room temperature (Vanhoutte et al., 2006), or head-motion restraint techniques (Edward et al., 2000). Many of these parameters can be harmonized to some extent, but some differences may always remain, e.g. even identical scanners may have different software versions or upgrades.

Specific objectives. To establish the impact of multisite acquisitions on statistical power in rs-fMRI, we first specifically aimed at characterizing the amplitude of the inter-site bias in real rs-fMRI measures, relative to intra-site variance. We based our evaluation on N=345 young healthy participants from the 1000 Functional Connectomes Project (FCP), including rs-fMRI samples independently collected at 8 imaging sites with 3T scanners in Germany, the United Kingdom, Australia and the United States of America. Datasets in this study were shared retrospectively and every documented parameters of image acquisition varied across studies. This data sample thus represents a worst-case-scenario in terms of 3T instrumental inter-site variations. Our second specific aim was to evaluate the impact of such inter-site bias on the detection power of rs-fMRI group comparison, in relation with sample size, group balancing and interaction between sites and group differences. We implemented for this purpose a series of simulation, mixing synthetic data with real data from the 1000 FCP. One of the particularity of the 1000 FCP 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 the same total sample size.

2. Method

2.1. Data samples

Participants. The paper studies 345 cognitively normal young adults (CNY) from a subset of subjects from the 1000 functional connectome project[2]. Only 3

[2] `http://fcon_1000.projects.nitrc.org/`

Site	Magnet	Scanner make	Channels	N	N final	Sex	Age	TR	# Slices	# Frames
Baltimore, USA	3T	N/A	N/A	23	21	8M/15F	20-40	2.5	47	123
Berlin, Germany	3T	Siemens Tim Trio	12	26	26	13M/13F	23-44	2.3	34	195
Cambridge, USA	3T	Siemens Tim Trio	12	198	195	75M/123F	18-30	3	47	119
Newark, USA	3T	N/A	N/A	19	17	9M/10F	21-39	2	32	135
NewYork_b, USA	3T	Siemens	N/A	20	18	8M/12F	18-46	2	33	175
Oxford, UK	3T	Siemens Tim Trio	12	22	20	12M/10F	20-35	2	34	175
Queensland, Australia	4T	Bruker	1	19	17	11M/8F	20-34	2.1	36	190
SaintLouis, USA	3T	Siemens Tim Trio	12	31	31	14M/17F	21-29	2.5	32	127

Table 1: Sites selected from the 1000 Functional Connectome Project.

tesla scanners from the following geographical locations: Germany, the United Kingdom, Australia and the United States of America were used, with a minimum of 15 subjects / site, including male and female representation. The sample was restricted to sites with full brain coverage in term of their field of view and a minimum of 50 functional volumes / run after high-motion censoring. From the 345 subjects (150 males, age range = 18-46 years) the total time of acquisition ranged between 6 and 7.5 min / run. The average age of the pooled dataset is 23.8 $\pm$5.14 years old with a male ratio of 41% across all sites, see Table 1 for more details on each site selected in the study. The experimental protocols for all datasets were approved by their respective ethic boards.

2.2. Computational environment

All experiments were performed using the NeuroImaging Analysis Kit (NIAK[3]) (Bellec et al., 2011) version 0.12.18, under CentOS version 6.3 with Octave[4] version 3.8.1 and the Minc toolkit[5] version 0.3.18. Analyses were executed in parallel on the Mammouth supercomputer[6] , using the pipeline system for Octave and Matlab (Bellec et al., 2012), version 1.0.2. The scripts used for processing can be found on Github[7]. For visualization Python 2.7.9 from the Anaconda 2.2.0[8] distribution were used along with Matplotlib[9] (Hunter, 2007), Seaborn[10] and Nilearn[11] for brain map visualizations.

2.3. Preprocessing

Each fMRI dataset was corrected for slice timing; a rigid-body motion was then estimated for each time frame, both within and between runs, as well as between one fMRI run and the T1 scan for each subject (Collins et al., 1994). The T1 scan was itself non-linearly co-registered to the Montreal Neurological

[3] http://simexp.github.io/niak/

[4] http://gnu.octave.org/

[5] http://www.bic.mni.mcgill.ca/ServicesSoftware/ServicesSoftwareMincToolKit

[6] http://www.calculquebec.ca/index.php/en/resources/compute-servers/mammouth-serie-ii

[7] http://www.calculquebec.ca/index.php/en/resources/compute-servers/mammouth-serie-ii

[8] http://docs.continuum.io/anaconda/index

[9] http://matplotlib.org/

[10] http://stanford.edu/~mwaskom/software/seaborn/index.html

[11] http://nilearn.github.io/

Institute (MNI) ICBM152 stereotaxic symmetric template (Fonov et al., 2011), using the CIVET pipeline (Ad-Dab'bagh et al., 2006). The rigid-body, fMRI-to-T1 and T1-to-stereotaxic transformations were all combined to resample the fMRI in MNI space at a 3 mm isotropic resolution. To minimize artifacts due to excessive motion, all time frames showing a displacement greater than 0.5 mm were removed (Power et al., 2012). A minimum of 50 unscrubbed volumes per run was required for further analysis (13 subjects were rejected). The following nuisance covariates were regressed out from fMRI time series: slow time drifts (basis of discrete cosines with a 0.01 Hz highpass cut-off), average signals in conservative masks of the white matter and the lateral ventricles as well as the first principal components (accounting for 95% variance) of the six rigid-body motion parameters and their squares (Giove et al., 2009; Lund et al., 2006). The fMRI volumes were finally spatially smoothed with a 6 mm isotropic Gaussian blurring kernel. A more detailed description of the pipeline can be found on the NIAK website[12] and Github[13].

2.4. Inter-site bias in resting-state connectivity

We first compared the functional connectivity measures derived from different sites of the 1000 FCP. The connectivity measures were obtained from the Pearsons correlation coefficient between the average temporal rs-fMRI fluctuations of a given pair of brain parcels. $R = 100$ parcels were obtained from a brain parcellation described in **?**, and generated using a bootstrap analysis of stable clusters (Bellec et al., 2010) from the Cambridge cohort of the 1000 FCP. For each subject, a 100×100 functional connectome matrix was generated including all parcel-to-parcel connections.

To quantify the inter-site bias we proposed to calculate the difference between the intra-site variability and the inter-site variability, we first obtained the intra-site variance for each site by computing the standard deviation (STD) distribution across-connections. In a second step we obtained the inter-site variance distribution by computing the STD across-site for each average connection of the connectivity matrix.

In order to quantify if the functional connectivity differences were significative between sites, we used a general linear model (GLM). The age, sex and frame displacement (FD) confounding variables were included in the model. The variables were corrected to have a zero mean across subjects, and an intercept (i.e. a column filled with 1) was added to capture the inter-site global average. A p-value was generated for each connection to quantify the probability that the difference in average connectivity was significantly distant from zero (Worsley and Friston, 1995). The number of false discovery was also controlled ($\alpha = 0.05$) using a BenjaminiHochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995).

[12] `http://niak.simexp-lab.org/pipe_preprocessing.html`
[13] `https://github.com/SIMEXP/multisite`

2.5. Simulations

Data generation process. We implemented Monte-Carlo simulations based on the 1000 FCP sample, with 8 sites for a total of 345 subjects. The multisite simulations were based on 7 sites, totaling 150 subjects, and the monosite simulations were based on the Cambridge site, with 195 subjects. The following data generation model was used to simulate tests of group difference. A subsample of subjects of fixed size was first selected randomly at each site. For each site, a ratio W of the selected subjects were randomly assigned to a 'treatment' group. For the subjects in the treatment group, a fixed value was added to achieve a given relative effect size (Cohen's d, see Section below) to 11 connections, previously described in **?**. **?** first selected from a literature review the pairs of regions believed to be affected in Alzheimer disease progression due to their difference in connectivity between cognitively normal subjects and patient with dementia of the Alzheimer type. Second, from that selection a subset of candidate pairs was retained based on their reproducibility score at a test-retest experiment conducted using the NYU TRT dataset (Zuo et al., 2010) (a dataset where each subject is scan 3 times, across 3 sessions).

In summary, the simulations had the following parameters:

- W: The allocation ratio of participants,
- d effect size (Cohens d),
- $di = 0.5$ interaction site-pathology effect,
- $B = 10^3$ number of simulations,
- s average standard deviation from the reference dataset,
- $\alpha = 0.001$ threshold of false-positive rate,

Effect size (Cohen's d). The normalized Cohen's d was used to estimate the effect size and it is defined as the difference between two means $\bar{x_1}, \bar{x_2}$ divided by a standard deviation from the data s.

For each site an effect is added to the connectivity of W of the subjects, selected randomly ("pathological" group):

$$y_{i,j} = y_{i,j} + \mu. \tag{1}$$

The parameter μ is chosen to obtain a particular effect size (expressed as the d of Cohen)

$$d = \frac{\mu}{s_{i,j}}, \tag{2}$$

where $s_{i,j}$ is the standard deviation between region i and j for the reference population (mono-site).

In order to introduce the same effect-size across the monosite and multisite datasets we are taking the standard deviation from the previously mentioned Cambridge cohort as the reference. The connection $y_{i,j}$ of the randomly affected subjects ("treatment" group) are therefore calculated $y_{i,j} = y_{i,j} + d \times s_{i,j}$.

GLM model. In order to detect changes on each connection pair between the artificially created groups, a general linear model (GLM) was applied and the following confounding variables were modelled in the analysis: age, sex and frame displacement (FD). To account for site-specific bias $S-1$ dummy-variables (binary vectors $1 \times S$) were added to the model with S being the total number of sites used in the study. The variables are corrected to have a zero mean across subjects, and an intercept (i.e. a column filled with 1) is added to $\mathbf{X}$ to capture the global average. The GLM relies on the following stochastic model Equation3.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{V}\gamma + \mathbf{E}, \tag{3}$$

- $\mathbf{Y}$: $N \times 1$, connectivity value for the pair (i, j),
- $\mathbf{X}$: $N \times K$, explainable variables,
- β: $1 \times K$, regression values for each explainable variable,
- $\mathbf{V}$: $N \times S$, each column code for a site (0/1),
- γ: $1 \times S$, site average connectivity,
- $\mathbf{E}$: $N \times 1$, residual values from the regression,

with N the number of subjects, K the number of explainable variables and S the number of sites. Where β is an unknown $1 \times K$ vector of linear regression coefficients, γ is a $1 \times S$ vector of linear regression coefficients representing the contribution of each site and $\mathbf{E}$ is a $N \times 1$ random (noise) multivariate Gaussian variable. As data generated from different subjects are statistically independent, and under an homoscedasticity assumption, the regression coefficients β can be estimated with ordinary least squares.

Statistical detection and sensitivity. We relied on the following parametric assumptions on the noise $\mathbf{E}$ (1) that its rows are independent; (2) that each element follows a normal distribution with zero mean, and (3) that the variance of all elements are constant within a column, also called the homoscedasticity assumptions. As the data generated from different subjects are statistically independent the first assumption is reasonable. We tested the normality and homoscedasticity assumptions on real datasets. Under these parametric assumptions, the regression coefficients β and γ can be estimated with ordinary least squares and, for a given contrast (difference between the control and 'treatment' group) the significance of the contrast is assessed by a Student t-test.

The sensitivity of the test was evaluated by the average detection performance of all the samples (see Equation 4). For each sample b, we have a p-value p_b^* and the detection sensitivity is estimated by the probability of p_b^* being inferior or equal to $\alpha = 0.001$.

$$\frac{1}{B}\sum_{b=1}^{B}\left(p_b^* \leq 0.001\right). \tag{4}$$

Simulation scenarios. For each experiment using real data, all effect size in the range 0 to 2 Cohens d with a step of 0.01 were considered. For each experiment we performed $B = 10^3$ simulations using 80% subsampling and the reported allocation ratio of participants W and its inverse $1 - W$ was attributed randomly to the treatment population for each site. We implemented the following experiments:

- We first checked how the total number of subject impact sensitivity using 3 different sample size (40, 80, 120) for an allocation ratio of $W = 0.5$,
- We then checked how the allocation ratio of participants W impact sensitivity using 3 different ratio ($W = 0.5, 0.3, 0.15$) for a total sample size of 120,
- Including an interaction site-pathology effect of $d = 0.5$, we checked how the allocation ratio of participants W impact sensitivity using 3 different ratio ($W = 0.5, 0.3, 0.15$) for a total sample size of 120,
- We checked how the allocation ratio of participants $W = 0.3$ impact sensitivity using only 2 sites one large (80 subjects) and one small ($\sim$ 20 subjects) instead of the 7 sites previously used,

In order to obtain more control on each of the parameter of the simulation and simulate some configuration that were not possible to do with the real data (like the simulation of 2 medium size sites of of 40 and 60 subjects per site, and the size of the site effect) we have therefore used a synthetic model using only synthetic data and the average standard deviation of the Cambridge site connectivity.

For each experiment using synthetic data, all effect size in the range 0 to 1.5 with a step of 0.01 were considered. The simulation were performed using randomly generated values. All experiments show four scenarios,

For each experiment, all combinations of effect size in the grid (without site-effect and with site-effect) and the grid (without interaction site-pathology and with interaction site-pathology) were considered. We implemented a series of experiments:

- We first checked how 2 sites with respectively 50 subjects in each site with an allocation ratio of $W = 0.5$ in both sites are impacted in the 4 scenarios in term of sensitivity,
- We then checked how 2 sites with respectively 50 subjects in each site with an allocation ratio of $W = 0.7$ for the first site and $W = 0.3$ for the second site are impacted in the 4 scenarios in term of sensitivity,
- We checked how 2 sites (one small of 20 subjects and one large of 80 subjects) with an allocation ratio of $W = 0.5$ in both sites are impacted in the 4 scenarios in term of sensitivity,

- We checked how 2 sites (one small of 20 subjects and one large of 80 subjects) with an allocation ratio of $W = 0.7$ for the first site and $W = 0.3$ for the second site are impacted in the 4 scenarios in term of sensitivity,
- We checked how 2 sites (one small of 20 subjects and one large of 80 subjects) with an allocation ratio of $W = 0.7$ for the first site and $W = 0.3$ for the second site are impacted in the 4 scenarios in term of sensitivity,
- We checked how 50 sites (with a number of subject per site randomly assigned between 2 and 15) with an allocation ratio W randomly assigned between 0.1 and 0.9 for each sites are impacted in term of sensitivity for 2 scenarios (without and with interaction site pathology),

3. Results

3.1. Inter-site bias

The first assessment perform on the dataset was to verify the distribution of the intra-site and inter-site functional connectivity variance to assess the magnitude of the bias. This analysis of Figure 1 shows the distribution of the standard deviation in connectivity across subjects (the distribution is over the full brain connectome, with several 1000s connections) for the 8 sites against the inter-sites standard deviation of connectomes averaged at each site. The first observation is the consistency across site in term of there STD. Secondly the inter-site (between-site) variability is smaller than the intra-site (between subjects) variability, in fact the amplitude of inter-site bias (~ 0.06) is about 3-fold smaller than the within-site standard deviation (~ 0.18).

In order to verify how spatial structure vary across sites the average standard deviation and the average connectivity map of the DMN were extracted for each site and reported in Figure 2. At the intersection between two sites the difference in average connectivity is illustrated (red set of brain cuts). First the mean DMN at each site is consistent with the expected spatial distribution reported in other studies (Damoiseaux et al., 2006; Dansereau et al., 2014; Yan et al., 2013b). The most salient changes between-sites are located in the mesio-frontal region associated with the anterior part of the DMN. In order to verify if significant changes are not only found in the DMN we have computed the entire connectome to obtain the other connectivity patterns and the findings can be generalized to the full connectome as illustraded in Figure 3.

3.2. Simulation on real data

In order to evaluate the impact of a multisite setup on our ability to detect changes in rs-functional connectivity studies we performed various simulation on real fMRI data. For each site and each sub-sample, W of the subjects were randomly assigned to a 'treatment' group. For the subjects in this group, a value was added to achieve a given relative effect size expressed in Cohen's d. In order to detect changes on each connection pair between the artificially created

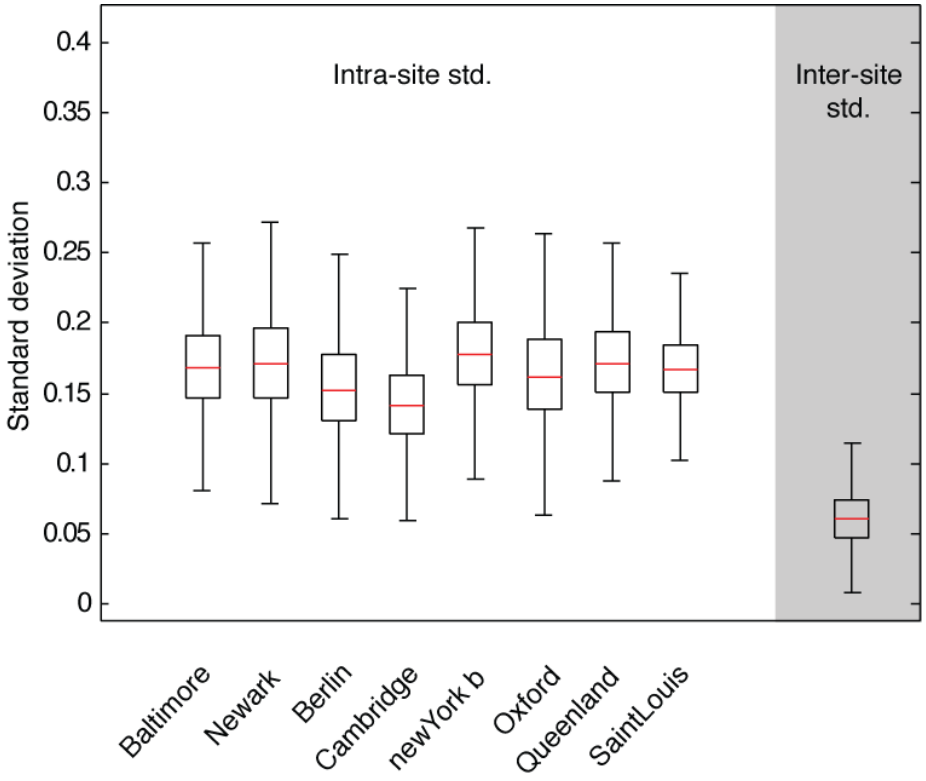


Figure 1: Distribution of intra-site (between-subject) standard deviation vs. inter-site (between-site) standard deviation, based on the standard deviation of the connectivity matrices from a subset of 8 sites from the 1000 functional connectome dataset.

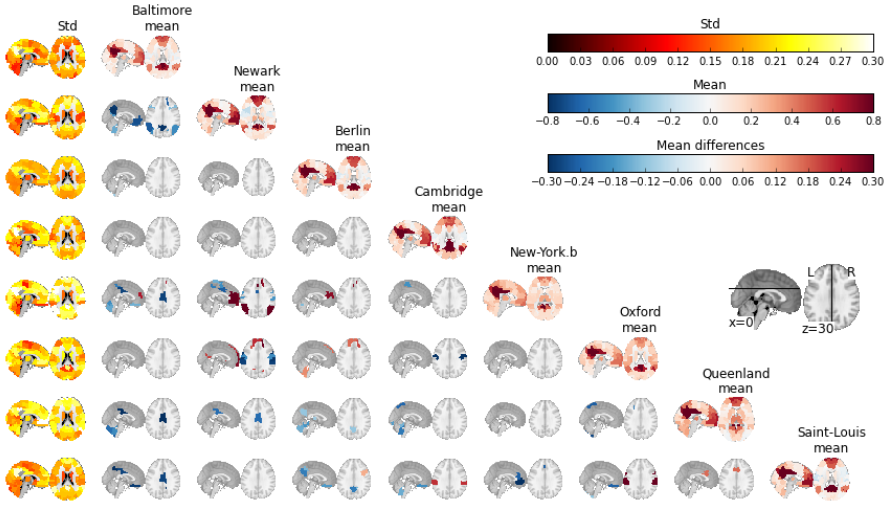


Figure 2: Average functional connectivity maps of the default-mode network at on 8 sites (Baltimore, Newark, Berlin, Cambridge, New-Yorkb, Oxford, Queenland and SaintLouis). The average connectivity map are shown on the diagonal. The standard deviation across subjects and within site is shown on the first column. Each off-diagonal block represent the significant differences between the average functional connectivity maps of two sites (called the inter-site bias). Statistical differences are obtained using a GLM procedure include age, sex and FD for each subject with an FDR correction procedure ($\alpha = 0.05$).

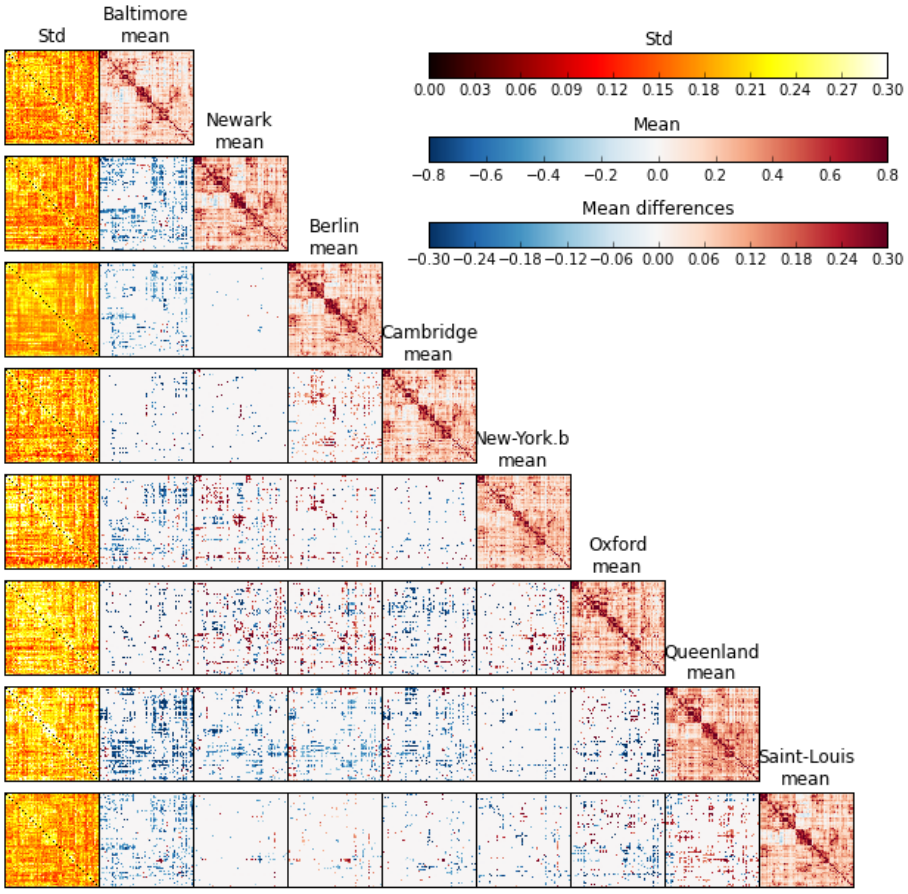


Figure 3: Average functional connectome of 8 sites (Baltimore, Newark, Berlin, Cambridge, New-Yorkb, Oxford, Queenland and SaintLouis) are shown on the diagonal. The standard deviation across subjects and within site is shown on the first column. Each off-diagonal block represent the significant differences between the average functional connectome of two sites (called the inter-site bias). Statistical differences are obtained using a GLM procedure include age, sex and FD for each subject with an FDR correction procedure ($\alpha = 0.05$).

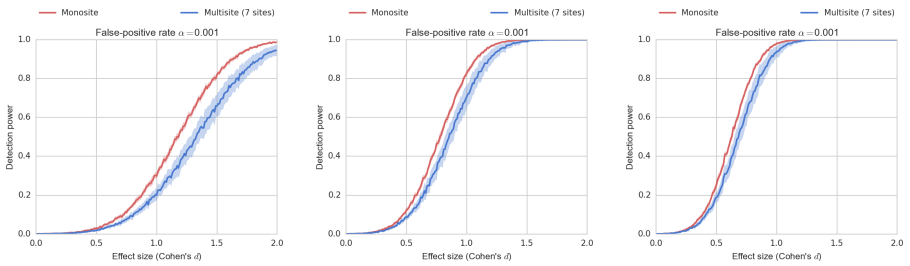


Figure 4: Simulation on real data, detection power of two groups with an allocation ratio W of 50% between the 7 sites. Every plot two scenarios: 1) monosite and 2) multisite 7 sites with correction for multisite differences using dummy variables. Each plot show the detection power in function of the effect-size for 3 different sample size 40, 80 and 120 subjects in total.

groups, a general linear model (GLM) was applied and the following confounding variables were modelled in the analysis: age, sex and frame displacement (FD).

In Figure 4 we show the effect of the sample size on the detection power. As expected we are able to detect smaller and smaller effect size as we increase the sample size. For an effect size of 1 which is considered a large effect we are able to detect significant changes in only 20% of the cases at 40 subjects, 80% at 80 subjects and almost 95% at 120% subjects.

In Figure 5 show the sample-size in function of the effect-size for a detection power of 80% using an allocation ratio W of 50%. A threshold on the statistical power of each test is applied at three alpha values: 0.001, (see alpha of 0.01 and 0.05 in supplementary material S1,S2). We are reporting the parametric curve and the points obtained using our simulations for the monosite and the multisite. As we can see the monosite is concordant with the parametric estimation and the multisite is offset of approximately 20 subjects for the same effect-size. The variance in detection power across connections in multisite also diminish with larger sample-size and the difference between the parametric estimation and the monosite and multisite tend to vanish as we increase the sample size.

Figure 6 show the effect of debalancing the two groups at various allocation ratio W (50% , 30% and 15%) for a total sample-size of 120 subjects. Has the debalancing increase our ability to detect effect is diminished. As an example for an effect size of 1 we would detect the effect in 95% of the cases in a 50 balanced scenario, this would go down to 90% in a 30%-70% scenario and to 60% in a 15%-85% scenario.

Figure 7 show the effect of debalancing the two groups at various allocation ratio W (50% , 30% and 15%) with an interaction site-pathology (an effect of 0.5 Cohen's d added to the model of half the sites) for a total sample size of 120 subjects. Has the allocation ratio is more debalanced our ability to detect effect is diminished. As an example for an effect size of 1 we would detect the effect in 98% of the cases using a W of 50% scenario, this would go down to 95% in

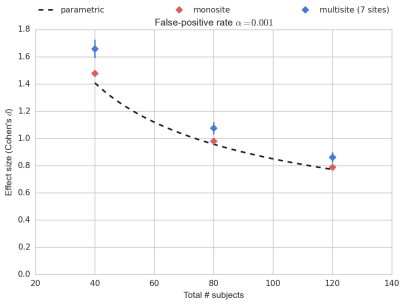
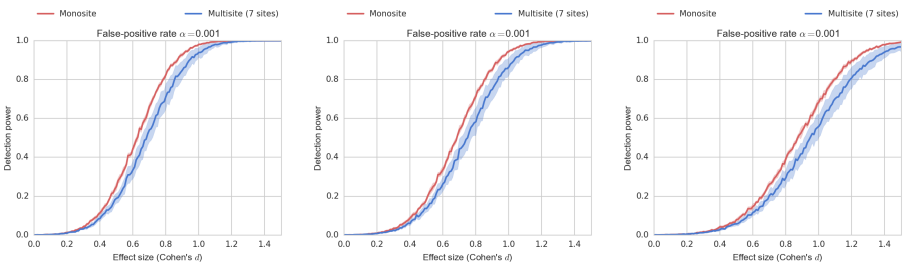
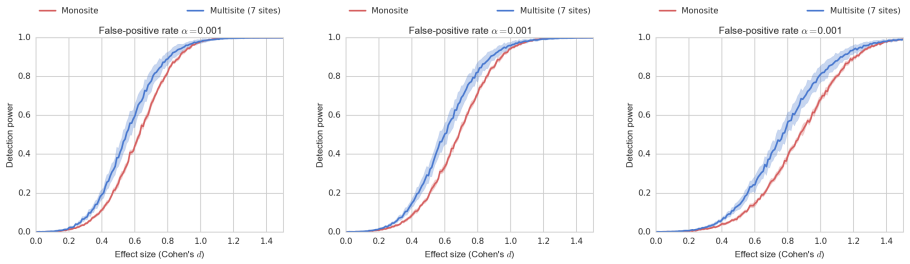


Figure 5: Sample-size in function of the effect-size for a detection power of 80% using a threshold on the probability of having false-positive rate ($\alpha = 0.001$) on a balanced dataset using an allocation ratio W of 50%.



Debalancing 50% Debalancing 30% Debalancing 15%

Figure 6: Simulation on real data, detection power of two groups for a total of 120 subjects between 7 sites. Every plot show two scenarios, 1) monosite and 2) multisite 7 sites with correction for multisite differences using dummy variables. Each plot show the detection power in function of the effect-size for 3 different allocation ratio W of 50%, 30% and 15% for each simulation.



Debalancing 50% Debalancing 30% Debalancing 15%

Figure 7: Simulation on real data, detection power of two groups for a total of 120 subjects between 7 sites with a site-pathology interaction of 0.5 Cohens d. Every plot show two scenarios, 1) monosite and 2) multisite 7 sites with correction for multisite differences using dummy variables. Each plot show the detection power in function of the effect-size for 3 different allocation ratio W of 50%, 30% and 15% for each simulation.

a W =30% scenario and to 75% in a W =15% scenario. The particularity of this experiment is the fact that the multisite configuration perform better than the single site meaning that it is better to have interaction of various amplitude across small sites than an average interaction on one large site.

Figure 8 show a scenario of two sites one large (80 subjects) and one small ($\sim$ 20 subjects) unbalanced at an allocation ratio W of 30% and the inverse (W =70%) with an interaction site-pathology (an effect of 0.5 Cohen's d added to the model of half the sites). As we can see the multisite configuration is as good as the monosite and the variance between connection is of the same order for the monosite and multisite.

Using these Monte-Carlo simulations we have shown that the power of detecting an effect is marginally affected by the site acquisition configuration (monosite or multisite) when the sites are balanced in term of the amount of subject.

3.3. Simulation on synthetic data

In order to obtain more control on each of the parameter of the simulation and simulate some configuration that were not possible to do with the real data (like the simulation of 2 medium size sites of of 40 and 60 subjects per site, and the size of the site effect) we have therefore used a synthetic model using only synthetic data and the average standard deviation of the Cambridge site connectivity. All plots show four scenarios: The top left plot represent the two configurations one monosite and two sites with correction for multisite differences using dummy variables without site effect. Has expected there is no difference between using 1 large site than combining two site of half the size. The plot on the upper right corner represent the detection power when we apply a site effect on balanced sites, here again not much differences an additive effect is fully compensated by the dummy variables corrective method. the lower left

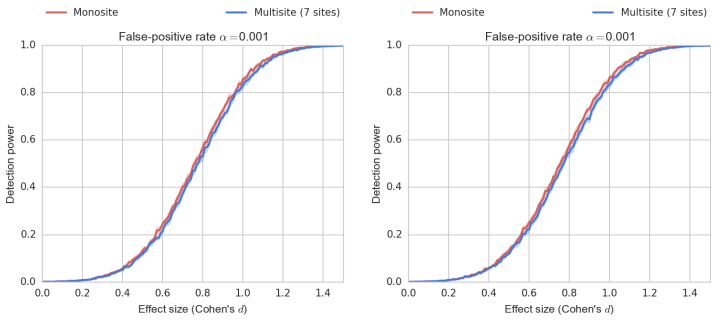


Figure 8: Simulation on real data, detection power of two groups for a total of 100 subjects between 2 sites with a site-pathology interaction of 0.5 Cohens d, one small site of 20 subjects and one large of 80 subjects. Every plot show two scenarios, 1) monosite and 2) multisite 2 sites with correction for multisite differences using dummy variables. Each plot show the detection power in function of the effect-size for 2 different allocation ratio W of 30% and 70% between 2 sites.

plot represent no site effect but an interaction between site and pathology. An the last plot on the lower right corner show the detection power with a site effect of 0.5 and a interation site pathology.

Figure 13 show the same configuration but the sample size are inverted. Has we can see we have the same pattern as in the real data with...

Figure 14 show a more realistic case in clinical trials where the number of scanning sites is very large and the number of subjects per site is small. There is usually no control on the exact balancing of those sites therefore we have randomly assign debalancing for each site (between 10% and 90%) and have randomly assigned a number of subject to each site (between 2 and 15 subjects).

4. Discussion

This work confirm the existence of an inter-site connectivity bias and compared it to the intra-site bias.

Connectivity bias that can impact interpretation.

Type of scanner most of them are Siemens we may have more variability if combining various brand

Talk about the impact in small acquisition 40 subjects total and the importance of pooling data among PI.

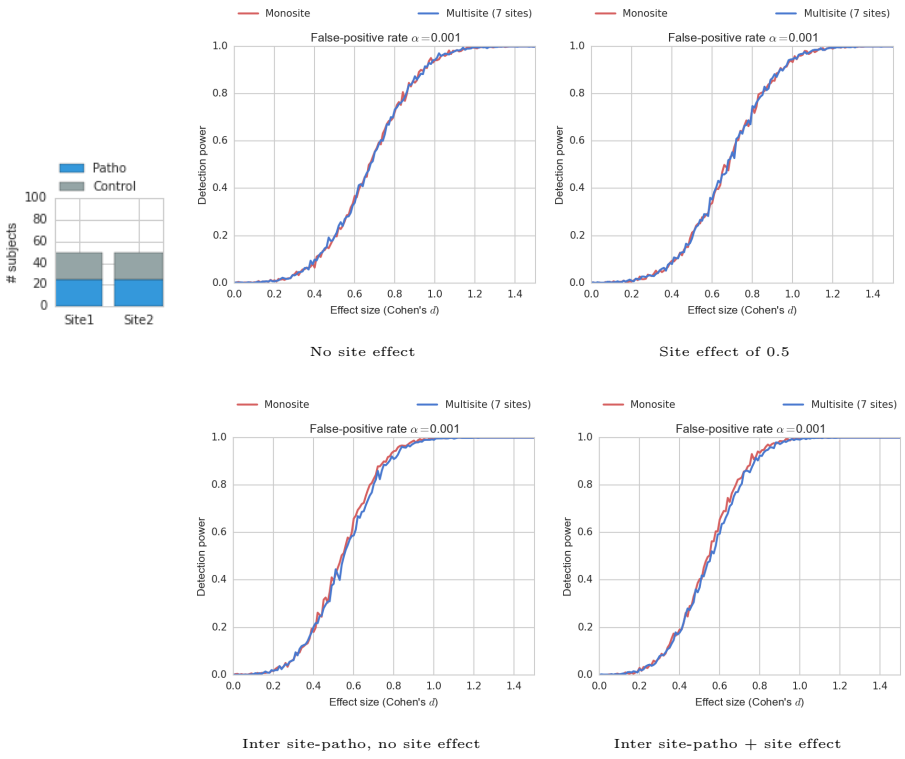


Figure 9: Simulation of the detection power of two groups balanced 50%-50% between two sites. All plots show four scenarios in two configuration one monosite and two sites with correction for multisite differences using dummy variables. The first column represent scenarios without site effect and the second column show with a site effect the first row show simulation with out interaction site-pathology and the second row show with interaction.

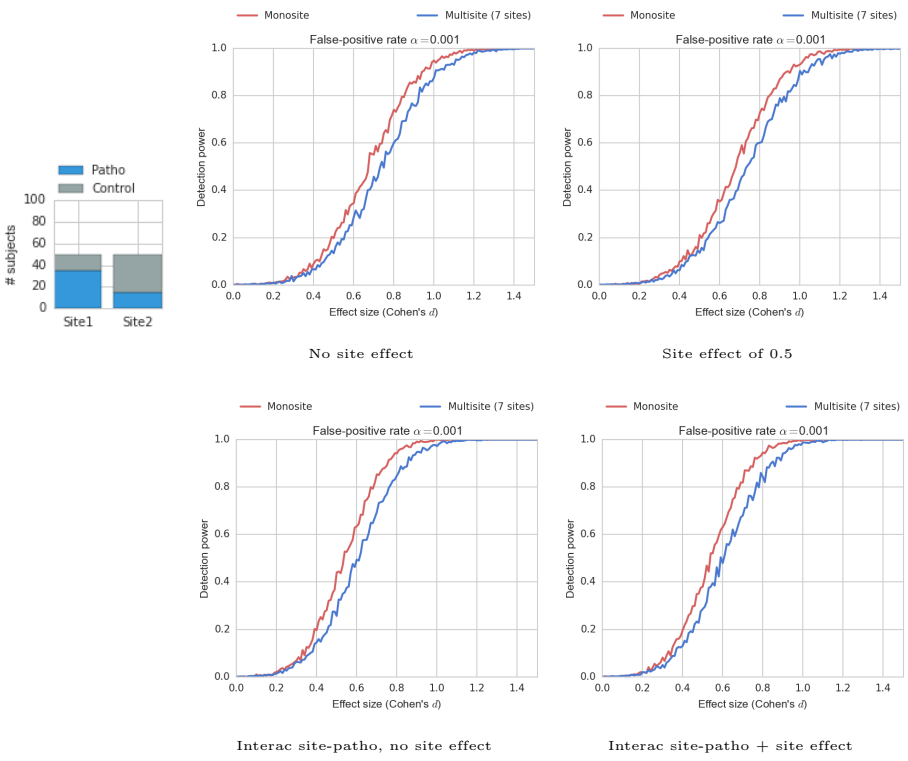


Figure 10: Simulation of the detection power of two groups unbalanced 70%-30% between two sites. All plots show four scenarios in two configuration one monosite and two sites with correction for multisite differences using dummy variables. The first column represent scenarios without site effect and the second column show with a site effect the first row show simulation with out interaction site-pathology and the second row show with interaction.

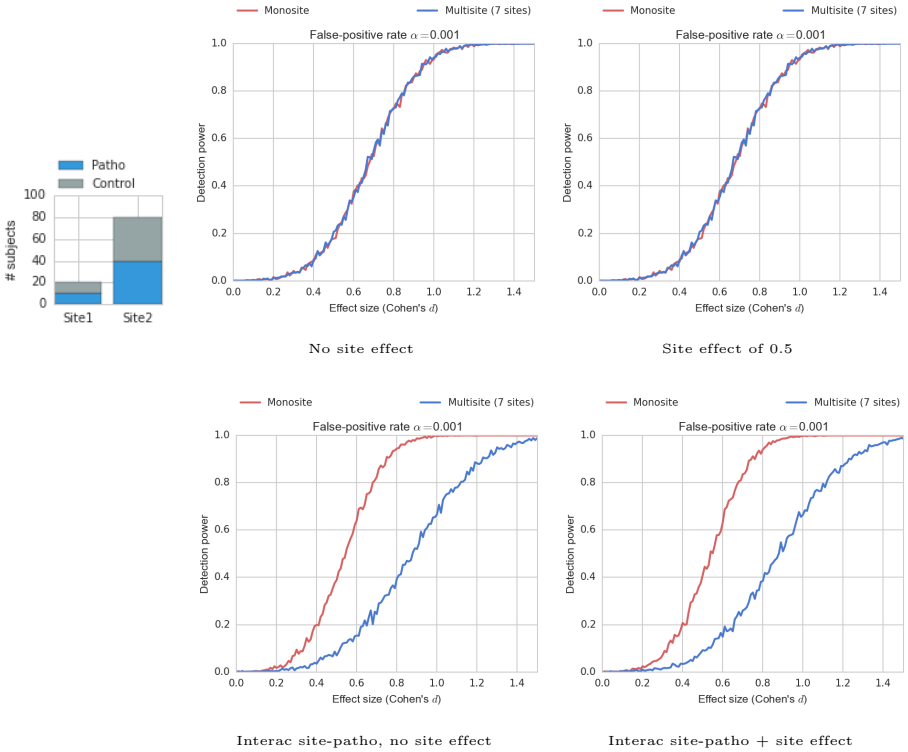


Figure 11: Simulation of the detection power of two groups unbalanced 50%-50% between two sites, one small site of 20 subjects and one large of 80 subjects. All plots show four scenarios in two configuration one monosite and two sites with correction for multisite differences using dummy variables. The first column represent scenarios without site effect and the second column show with a site effect the first row show simulation with out interaction site-pathology and the second row show with interaction.

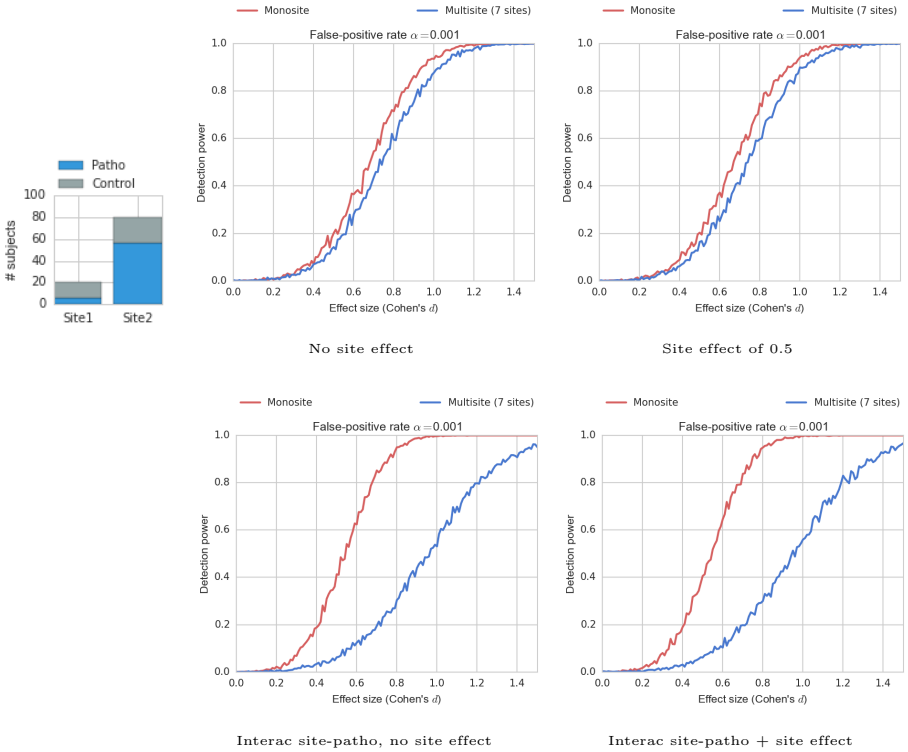


Figure 12: Simulation of the detection power of two groups unbalanced 70%-30% between two sites, one small site of 20 subjects and one large of 80 subjects. All plots show four scenarios in two configuration one monosite and two sites with correction for multisite differences using dummy variables. The first column represent scenarios without site effect and the second column show with a site effect the first row show simulation with out interaction site-pathology and the second row show with interaction.

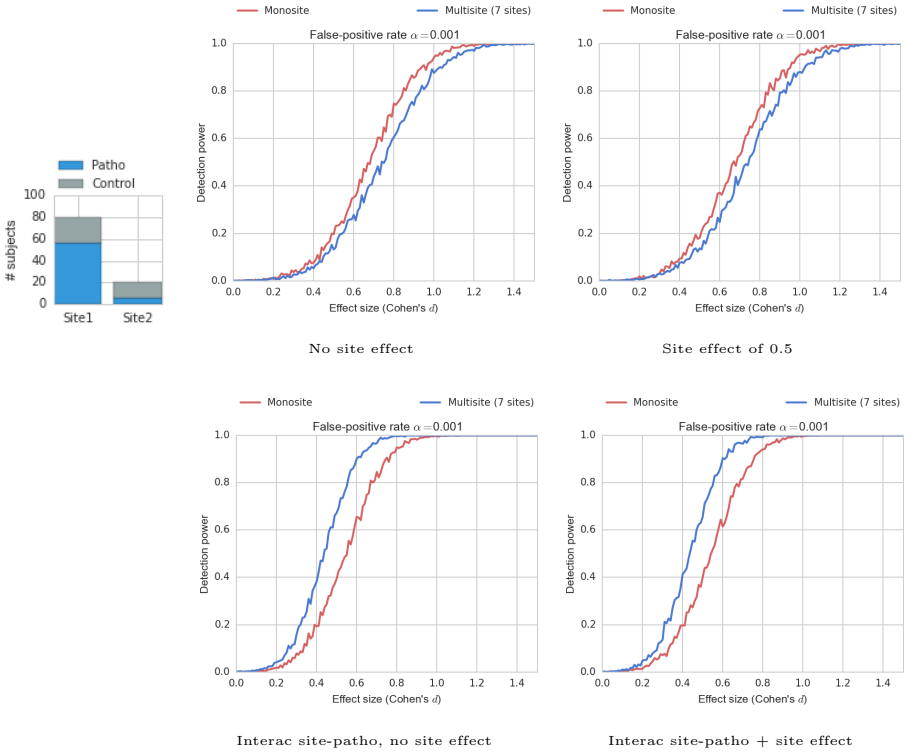


Figure 13: Simulation of the detection power of two groups unbalanced 70%-30% between two sites, one small site of 20 subjects and one large of 80 subjects. All plots show four scenarios in two configuration one monosite and two sites with correction for multisite differences using dummy variables. The first column represent scenarios without site effect and the second column show with a site effect the first row show simulation with out interaction site-pathology and the second row show with interaction.

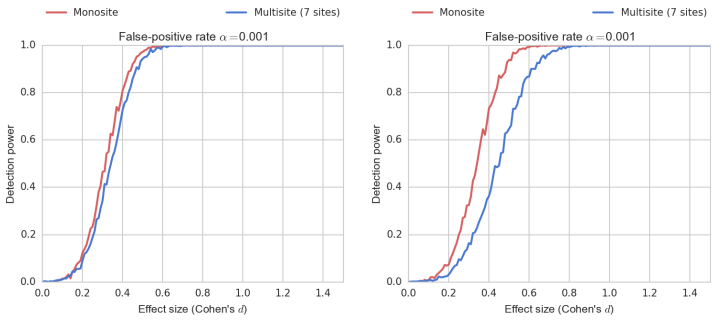


Figure 14: Simulation of the detection power of two groups balanced randomly between 10%-90% between 50 sites, the number of subject per site is randomly assigned between 2 and 15. All plot show two scenarios, one monosite and two sites with correction for multisite differences using dummy variables. The plots show the detection power when the variability is greater in one side then the other for the pathology group (twice the reference variability in one site and half the reference variability in the other one).

Choosing the right sample size for a given detection power to obtain reproducible results.

Talk about the impact in clinical trial and best practice.
most of the time we cannot correct for site with one population only or with too few subjects per site. Although these kind of multisite studies generally sum-up to large dataset of more than 200 subjects. has we demonstrated in the synthetic simulation such configuration are comparable to single site analysis for this range of sample size.

It is better to have interaction of various amplitude across small sites than an average interaction on one large site

The number of subject / site will impact the bias between multisite and monosite (like in the example of 2 sites the diff is marginal). Resulting in the following conclusion combining large site (e.g greater than 50 subjects) may results in low bias compared to very small sites.

talk about the reduction of the bias of the simulation compared to the fully parametric estimate as we increase the sample size. The estimation is more realistic over 80 subjects and overestimate the detection power at low sample-size.

Despite justifiable scepticism, feasibility analyses demonstrated that meaningful explorations of the aggregate dataset, composed of 24 imaging sites for a grand total of 1093 subjects, could be performed (Biswal et al., 2010). Although no explicit correction for multi-site variability was used, they only use global signal correction (GSC) to normalize subjects which may introduce anti-correlation in the data (Fox et al., 2009; Murphy et al., 2009; Saad et al., 2012;

Carbonell et al., 2014; Power et al., 2014). After accounting for site-related differences, the analysis showed brain-behaviour relationships with phenotypic variables such as age, gender, and diagnostic label, and confirmed a variety of prior hypotheses (Biswal et al., 2010; Fair et al., 2012; Tomasi and Volkow, 2010; Zuo et al., 2012). While encouraging, many uncontrolled and unknown factors in the 1000 FCP remain a source of concern, as they spread beyond simple site effects and can limit the datasets utility as highlighted by Yan et al. (2013a). Another compelling proof of multisite bias is the study reported by Nielsen et al. (2013) where they did an analysis on a single site dataset and a multi-site dataset of subject with autism and concluded that the multi-site autism study classification accuracy significantly outperformed chance but was much lower for multi-site prediction than for previous single site results (Nielsen et al., 2013). We therefore need to keep in mind that the site effect must be taken in account in the analysis or we may reduce our detection power.

5. Conclusion

6. Acknowledgments

Parts of this work were presented at the 2013 annual meetings of the organization for human brain mapping (**?**), as well as the Alzheimer's Association International Conference (AAIC) (2013) (Boston) (**?**). The authors are grateful to the members of the 1000 functional connectome consortium for publicly releasing there dataset. The computational resources used to perform the data analysis were provided by ComputeCanada[14] and CLUMEQ[15], which is funded in part by NSERC (MRS), FQRNT, and McGill University. This project was funded by NSERC grant number RN000028, a salary award from "Fonds de recherche du Québec – Santé" to PB as well as a salary award by the Canadian Institute of Health Research to CD.

References

Ad-Dab'bagh, Y., Einarson, D., Lyttelton, O., Muehlboeck, J. S., Mok, K., Ivanov, O., Vincent, R. D., Lepage, C., Lerch, J., Fombonne, E., Evans, A. C., 2006. The CIVET Image-Processing environment: A fully automated comprehensive pipeline for anatomical neuroimaging research. In: Corbetta, M. (Ed.), Proceedings of the 12th Annual Meeting of the Human Brain Mapping Organization. Neuroimage, Florence, Italy.

Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P., Evans, A. C., 2012. The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows.

[14] `https://computecanada.org/`
[15] `http://www.clumeq.mcgill.ca/`

Frontiers in neuroinformatics 6.
URL http://dx.doi.org/10.3389/fninf.2012.00007

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., Evans, A. C., Jul. 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51 (3), 1126–1139.
URL http://dx.doi.org/10.1016/j.neuroimage.2010.02.082

Benjamini, Y., Hochberg, Y., 1995. Controlling the false-discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57, 289–300.

Biswal, B. B., Mennes, M., Zuo, X.-N. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A.-M. M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S., Kiviniemi, V. J., Kötter, R., Li, S.-J. J., Lin, C.-P. P., Lowe, M. J., Mackay, C., Madden, D. J., Madsen, K. H., Margulies, D. S., Mayberg, H. S., McMahon, K., Monk, C. S., Mostofsky, S. H., Nagel, B. J., Pekar, J. J., Peltier, S. J., Petersen, S. E., Riedl, V., Rombouts, S. A., Rypma, B., Schlaggar, B. L., Schmidt, S., Seidler, R. D., Siegle, G. J., Sorg, C., Teng, G.-J. J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.-C. C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.-F. F., Zhang, H.-Y. Y., Castellanos, F. X., Milham, M. P., Mar. 2010. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences of the United States of America 107 (10), 4734–4739.
URL http://dx.doi.org/10.1073/pnas.0911855107

Carbonell, F., Bellec, P., Shmuel, A., Feb 2014. Quantification of the impact of a confounding variable on functional connectivity confirms anti-correlated networks in the resting-state. Neuroimage 86, 343–353.
URL http://dx.doi.org/10.1016/j.neuroimage.2013.10.013

Collins, D. L., Neelin, P., Peters, T. M., Evans, A. C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. Journal of computer assisted tomography 18 (2), 192–205.
URL http://view.ncbi.nlm.nih.gov/pubmed/8126267

Damoiseaux, J. S., Rombouts, S. A. R. B., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., Beckmann, C. F., Sep. 2006. Consistent resting-state networks across healthy subjects. Proceedings of the National Academy of Sciences 103 (37), 13848–13853.
URL http://dx.doi.org/10.1073/pnas.0601417103

Dansereau, C., Bellec, P., Lee, K., Pittau, F., Gotman, J., Grova, C., 2014. Detection of abnormal resting-state networks in individual patients suffering from focal epilepsy: An initial step toward individual connectivity assessment. Frontiers in Neuroscience 8 (419).
URL http://www.frontiersin.org/brain_imaging_methods/10.3389/fnins.2014.00419/abstract

Desmond, J., Glover, G., Aug. 2002. Estimating sample size in functional mri (fmri) neuroimaging studies: Statistical power analyses. Journal of Neuroscience Methods 118 (2), 115–128.
URL http://dx.doi.org/10.1016/s0165-0270(02)00121-8

Edward, V., Windischberger, C., Cunnington, R., Erdler, M., Lanzenberger, R., Mayer, D., Endl, W., Beisteiner, R., Nov 2000. Quantification of fmri artifact reduction by a novel plaster cast head holder. Hum Brain Mapp 11 (3), 207–213.

Elliott, M. R., Bowtell, R. W., Morris, P. G., Jun 1999. The effect of scanner sound in visual, motor, and auditory functional mri. Magn Reson Med 41 (6), 1230–1235.

Fair, D. A., Nigg, J. T., Iyer, S., Bathula, D., Mills, K. L., Dosenbach, N. U. F., Schlaggar, B. L., Mennes, M., Gutman, D., Bangaru, S., Buitelaar, J. K., Dickstein, D. P., Martino, A. D., Kennedy, D. N., Kelly, C., Luna, B., Schweitzer, J. B., Velanova, K., Wang, Y.-F., Mostofsky, S., Castellanos, F. X., Milham, M. P., 2012. Distinct neural signatures detected for adhd subtypes after controlling for micro-movements in resting state functional connectivity mri data. Front Syst Neurosci 6, 80.
URL http://dx.doi.org/10.3389/fnsys.2012.00080

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., Brain Development Cooperative Group, Jan. 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54 (1), 313–327.
URL http://dx.doi.org/10.1016/j.neuroimage.2010.07.033

Fox, M. D., Zhang, D., Snyder, A. Z., Raichle, M. E., Jun 2009. The global signal and observed anticorrelated resting state brain networks. J Neurophysiol 101 (6), 3270–3283.
URL http://dx.doi.org/10.1152/jn.90777.2008

Friedman, L., Glover, G., Jun. 2006. Report on a multicenter fmri quality assurance protocol. Journal of magnetic resonance imaging : JMRI 23 (6), 827–839.
URL http://dx.doi.org/10.1002/jmri.20583

Friedman, L., Glover, G., Consortium, T. F., Nov. 2006. Reducing interscanner variability of activation in a multicenter fmri study: Controlling for signal-to-fluctuation-noise-ratio (sfnr) differences. NeuroImage 33 (2), 471–481.
URL http://dx.doi.org/10.1016/j.neuroimage.2006.07.012

Giove, F., Gili, T., Iacovella, V., Macaluso, E., Maraviglia, B., Oct. 2009. Images-based suppression of unwanted global signals in resting-state functional connectivity studies. Magnetic resonance imaging 27 (8), 1058–1064.
URL http://dx.doi.org/10.1016/j.mri.2009.06.004

Hunter, J. D., 2007. Matplotlib: A 2d graphics environment. Computing In Science & Engineering 9 (3), 90–95.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., Nichols, T. E., Jan. 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29 (1), 54–66.
URL http://dx.doi.org/10.1016/j.neuroimage.2005.07.005

Milham, M. P., Fair, D., Mennes, M., Mostofsky, S. H., 2012. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in Systems Neuroscience 6 (62).
URL http://www.frontiersin.org/systems_neuroscience/10.3389/fnsys.2012.00062/full

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., Beckett, L., Nov 2005. The alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am 15 (4), 869–77, xi–xii.
URL http://dx.doi.org/10.1016/j.nic.2005.09.008

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., Bandettini, P. A., Feb. 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? NeuroImage 44 (3), 893–905.
URL http://dx.doi.org/10.1016/j.neuroimage.2008.09.036

Nielsen, J., Zielinski, B., Fletcher, T., Alexander, A., Lange, N., Bigler, E., Lainhart, J., Anderson, J., 2013. Multisite functional connectivity mri classification of autism: Abide results. Frontiers in human neuroscience 7, –.
URL http://view.ncbi.nlm.nih.gov/pubmed/24093016

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Feb. 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage 59 (3), 2142–2154.
URL http://dx.doi.org/10.1016/j.neuroimage.2011.10.018

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Jan. 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage 84, 320–341.
URL http://dx.doi.org/10.1016/j.neuroimage.2013.08.048

Saad, Z. S., Gotts, S. J., Murphy, K., Chen, G., Jo, H. J. J., Martin, A., Cox, R. W., 2012. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. Brain connectivity 2 (1), 25–32.
URL http://dx.doi.org/10.1089/brain.2012.0080

Shehzad, Z., Kelly, C. M., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., Lee, S. H. H., Margulies, D. S., Roy, A. K. K., Biswal, B. B., Petkova, E., Castellanos, F. X., Milham, M. P., Oct. 2009. The resting brain: unconstrained yet reliable. Cerebral cortex (New York, N.Y. : 1991) 19 (10), 2209–2229.
URL http://dx.doi.org/10.1093/cercor/bhn256

Tomasi, D., Volkow, N. D., May 2010. Functional connectivity density mapping. Proceedings of the National Academy of Sciences 107 (21), 9885–9890.
URL http://dx.doi.org/10.1073/pnas.1001414107

Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., Buckner, R. L., Jan. 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. Journal of neurophysiology 103 (1), 297–321.
URL http://dx.doi.org/10.1152/jn.00783.2009

Vanhoutte, G., Verhoye, M., der Linden, A. V., May 2006. Changing body temperature affects the t2* signal in the rat brain and reveals hypothalamic activity. Magn Reson Med 55 (5), 1006–1012.
URL http://dx.doi.org/10.1002/mrm.20861

Worsley, K. J., Friston, K. J., Sep. 1995. Analysis of fMRI Time-Series Revisited—Again. NeuroImage 2 (3), 173–181.
URL http://dx.doi.org/10.1006/nimg.1995.1023

Yan, C., Liu, D., He, Y., Zou, Q., Zhu, C., Zuo, X., Long, X., Zang, Y., May 2009. Spontaneous Brain Activity in the Default Mode Network Is Sensitive to Different Resting-State Conditions with Limited Cognitive Load. PLoS ONE 4 (5), e5743+.
URL http://dx.doi.org/10.1371/journal.pone.0005743

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., Li, Q., Zuo, X.-N., Castellanos, F. X., Milham, M. P., Aug. 2013a. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. NeuroImage 76, 183–201.
URL http://dx.doi.org/10.1016/j.neuroimage.2013.03.004

Yan, C.-G. G., Craddock, C. C., Zuo, X.-N. N., Zang, Y.-F. F., Milham, M. P., Oct. 2013b. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. NeuroImage 80, 246–262.
URL http://view.ncbi.nlm.nih.gov/pubmed/23631983

Zuo, X.-N., Kelly, C., Adelstein, J. S., Klein, D. F., Castellanos, F. X., Milham, M. P., Feb. 2010. Reliable intrinsic connectivity networks: Test–retest evaluation using ICA and dual regression approach. NeuroImage 49 (3), 2163–2177.
URL http://dx.doi.org/10.1016/j.neuroimage.2009.10.080

Zuo, X.-N., Xu, T., Jiang, L., Yang, Z., Cao, X.-Y., He, Y., Zang, Y.-F., Castellanos, F. X., Milham, M. P., Oct. 2012. Toward reliable characterization of functional homogeneity in the human brain: Preprocessing, scan duration, imaging resolution and computational space. NeuroImage.
URL http://dx.doi.org/10.1016/j.neuroimage.2012.10.017

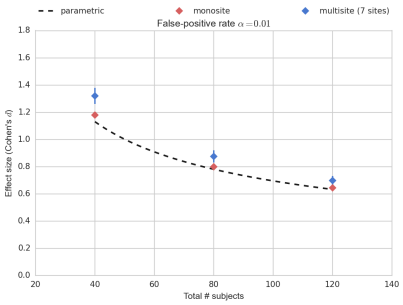


Figure S1: Sample-size in function of the effect-size for a detection power of 80% using a threshold on the probability of having false-positive rate ($\alpha = 0.01$) on a balanced dataset using an allocation ratio W of 50%.

Supplementary Material – Feasibility of multi-centric fMRI connectivity studies of Alzheimer's disease

Submitted to Neuroimage.

C. Dansereau[1,2], C. Risterucci[3], E. Merlo Pich[3], D. Arnold[4], P. Bellec[1,2]

[1]Functional Neuroimaging Unit, Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal
[2]Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada
[3]F. Hoffmann-La Roche Ldt., Basel, Switzerland
[4]NeuroRx, Montreal, Quebec, Canada

For all questions regarding the paper, please address correspondence to Pierre Bellec, CRIUGM, 4545 Queen Mary, Montreal, QC, H3W 1W5, Canada. Email: pierre.bellec (at) criugm.qc.ca.

Supplementary Material

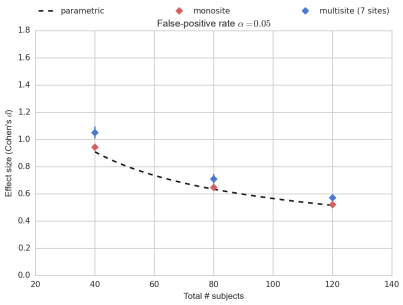


Figure S2: Sample-size in function of the effect-size for a detection power of 80% using a threshold on the probability of having false-positive rate ($\alpha = 0.05$) on a balanced dataset using an allocation ratio W of 50%.