

# 1 Statistical power and measurement bias in multisite 2 resting-state fMRI connectivity

3 Christian Dansereau<sup>a,b</sup>, Yassine Benhajali<sup>a,c</sup>, Celine Risterucci<sup>d</sup>, Emilio Merlo  
4 Pich<sup>d,\*\*</sup>, Pierre Orban<sup>a</sup>, Douglas Arnold<sup>e</sup>, Pierre Bellec<sup>a,b,\*</sup>

5 <sup>a</sup>*Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, CA*

6 <sup>b</sup>*Department of Computer Science and Operations Research, University of Montreal,  
7 Montreal, CA*

8 <sup>c</sup>*Département d'anthropologie, Université Montréal, Montréal, CA*

9 <sup>d</sup>*Clinical Imaging, pRED, F.Hoffman-La Roche, Basel, CH*

10 <sup>e</sup>*NeuroRx inc., Montréal, CA*

---

## 11 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 bias in connectivity and its impact on our ability to detect individual and group differences. Our study was based on real multisite fMRI datasets collected in  $N = 345$  young, healthy subjects across 8 scanning sites with 3T 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 bias 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 bias 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 more serious 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 sample size is large enough.

12 **Keywords:** multisite, statistical power, prediction accuracy, Monte-Carlo  
13 simulation, sample size, resting-state, fMRI connectivity, SVM

---

\*Corresponding author: pierre.bellec@criugm.qc.ca  
Preprint submitted to *NeuroImage* on July 8, 2016  
Present address: CNRS, IMAG, Fakeda Development Centre Europe, London, UK  
Email address: pierre.bellec@criugm.qc.ca (Pierre Bellec)

## 14 Highlights

- 15 • Small to moderate systematic site bias in fMRI connectivity.
- 16 • Small impact of site bias on the detection of group differences for sample  
17 size  $> 100$ .
- 18 • Linear regression of the sites prior to multivariate prediction do not im-  
19 prove prediction accuracy.

## 20 1. Introduction

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

38 *Group comparison in rs-fMRI connectivity.* We focused in this work on the most  
39 common measure of individual functional connectivity, which is the Pearson’s  
40 correlation coefficient between the average rs-fMRI time series of two brain re-  
41 gions. To compare two groups, a general linear model (GLM) is typically used  
42 to establish the statistical significance of the difference in average connectiv-  
43 ity between the groups. Finally a  $p$ -value is generated for each connection to  
44 quantify the probability that the difference in average connectivity is signifi-  
45 cantly different from zero (Worsley and Friston, 1995). If the estimated  $p$ -value  
46 is smaller than a prescribed tolerable level of false-positive findings, generally  
47 adjusted for the number of tests performed across connections, say  $\alpha = 0.001$ ,  
48 then the difference in connectivity is deemed significant.

49 *Statistical power in group comparisons at multiple sites.* The statistical power  
50 of a group comparison study is the probability of finding a significant difference,  
51 when there is indeed a true difference. A careful study design involves to select  
52 a sample-size large enough to reach a given level of statistical power, e.g. 80%.  
53 In the GLM, the statistical power actually depends on a series of parameters

(Desmond and Glover, 2002): (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). 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 et al., 2006), 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. (2013a) has indeed demonstrated the presence of a significant bias in rs-fMRI measures between sites 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 variance in rs-fMRI.* 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. 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. Because SVM has the ability to combine mea-

100 sures across connections, unlike univariate GLM tests, we hypothesized that the  
101 GLM and SVM will be impacted differently by site effects.

102 *Specific objectives.* Our first objective was to characterize, using real data, the  
103 amplitude of systematic biases in rs-fMRI connectivity measures across sites, as  
104 a function of within-site variance. We based our evaluation on images generated  
105 from independent groups at 8 sites equipped with 3T scanners, in a harmonized  
106 subset ( $N = 345$ ) of the 1000 FCP. Our second objective was to evaluate the  
107 impact of site effects on the detection power of group differences in rs-fMRI  
108 connectivity, as a function of the amplitude of the group difference, sample size,  
109 as well as the balancing of groups across sites. We implemented for this purpose  
110 a series of Monte Carlo simulations, mixing synthetic data with real data in the  
111 1000 FCP sample. One of the particularity of the 1000 FCP is the presence of  
112 one large site of  $\sim 200$  subjects and 7 small sites of  $\sim 20$  subjects per site. We  
113 were therefore able to implement realistic scenarios following either a monosite  
114 or a multisite design (with 7 sites), with the same total sample size. Finally,  
we repeated the Monte-Carlo using a SVM instead of a GLM, and assessed the  
impact of site effects on prediction accuracy rather than statistical power.

## 115 2. Method

### 116 2.1. Imaging sample characteristics

117 The full 1000 FCP sample includes 1082 subjects, with images acquired over  
118 33 sites spread across North America, Europe, Australia and China. As the  
119 1000 FCP is a retrospective study, no effort was made to harmonize population  
120 characteristics or imaging acquisition parameters. Many sites thus featured  
121 some outliers characteristics within the sample, such as images acquired at 1.5T  
122 or 4T field strengths (5 sites), a population composed mainly of older (4 sites)  
123 or Asian (6 sites) participants, samples composed almost exclusively of male or  
124 females (8 sites), or partial brain coverage in rs-fMRI. To avoid possible biases  
125 in rs-fMRI measures related to such outliers characteristics, a subset of sites  
126 was selected based on the following harmonization criteria: (1) 3T scanner field  
127 strength, (2) full brain coverage for the rs-fMRI scan, and, (3) a minimum of 15  
128 young or middle aged adult participants, with a mixture of males and females  
129 (4) samples drawn from a population with a predominant Caucasian ethnicity.  
130 In addition, only young and middle aged participants (18-46 years old) were  
131 included in the study, and we further excluded subjects with excessive motion  
132 (see next Section). The final sample for our study thus included 345 cognitively  
133 normal young adults (150 males, age range: 18-46 years, mean $\pm$ std:  $23.8 \pm 5.14$ )  
134 with images acquired across 8 sites located in Germany, the United Kingdom,  
135 Australia and the United States of America. The total time of available rs-  
136 fMRI data for these subjects ranged between 6 and 7.5 min and only one run  
137 was available. See Table 1 for more details on the demographics and imaging  
138 parameters at each site selected in the study. The experimental protocols for all  
139 datasets as well as data sharing in the 1000 FCP were approved by the respective

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 Allegra	1	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	3T	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.

ethics committee of each site. This secondary analysis of the 1000 FCP sample was approved by the local ethics committee at CRIUGM, University of Montreal, QC, Canada.

## 2.2. Computational environment

All experiments were performed using the NeuroImaging Analysis Kit, NIAK<sup>1</sup> (Bellec et al., 2011) version 0.12.18, under CentOS version 6.3 with Octave<sup>2</sup> version 3.8.1 and the Minc toolkit<sup>3</sup> version 0.3.18. Analyses were executed in parallel on the Mammouth supercomputer<sup>4</sup>, using the pipeline system for Octave and Matlab, PSOM (Bellec et al., 2012) version 1.0.2. The scripts used for processing can be found on Github<sup>5</sup>. Prediction was performed using the LibSVM library (Chang and Lin, 2011). Visualization was implemented using Python 2.7.9 from the Anaconda 2.2.0<sup>6</sup> distribution, along with Matplotlib<sup>7</sup> (Hunter, 2007), Seaborn<sup>8</sup> and Nilearn<sup>9</sup> 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 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 re-sample the fMRI in MNI space at a 3 mm isotropic resolution. To minimize artifacts due to excessive motion, all time frames showing a frame displacement, as defined in Power et al. (2012), greater than 0.5 mm were removed and a residual motion estimated after scrubbing. A minimum of 50 unscrubbed volumes per run was

<sup>1</sup><http://simexp.github.io/niak/>

<sup>2</sup><http://gnu.octave.org/>

<sup>3</sup><http://www.bic.mni.mcgill.ca/ServicesSoftware/ServicesSoftwareMincToolKit>

<sup>4</sup><http://www.calculquebec.ca/index.php/en/resources/compute-servers/mammouth-serie-ii>

<sup>5</sup><https://github.com/SIMEXP/Projects/tree/master/multisite>

<sup>6</sup><http://docs.continuum.io/anaconda/index>

<sup>7</sup><http://matplotlib.org/>

<sup>8</sup><http://stanford.edu/~mwaskom/software/seaborn/index.html>

<sup>9</sup><http://nilearn.github.io/>

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<sup>10</sup> and Github<sup>11</sup>.

#### 2.4. Inter-site bias in resting-state connectivity

*Functional connectomes.* We compared the functional connectivity measures derived from different sites of the 1000 FCP. A functional brain parcellation with 100 regions was first generated using a bootstrap analysis of stable clusters (Bellec et al., 2010b), on the Cambridge cohort of the 1000 FCP ( $N = 195$ ), as described in Orban et al. (2015). For a given pair of regions, the connectivity measure was defined by the Fisher transformation of the Pearson’s correlation coefficient between the average temporal rs-fMRI fluctuations of the two regions. For each subject, a  $100 \times 100$  functional connectome matrix was thus generated, featuring the connections for every possible pair of brain regions.

*Inter-site bias.* The inter-site bias at a particular connection was defined as the absolute difference in average connectivity between two sites. In order to formally test the significance of the inter-site bias, we used a GLM including age, sex and residual motion as covariates (corrected to have a zero mean across subjects), as well as dummy variables coding for the average connectivity at each site. For each site, a “contrast” vector was coded to measure the difference in average connectivity between this site and the grand average of functional connectivity combining all other sites. A  $p$ -value was generated for each connection to quantify the probability that the observed effect using this contrast was significantly different from zero (Worsley and Friston, 1995). The number of false discovery was also controlled ( $q = 0.05$ ) using a BenjaminiHochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995). To quantify the severity of inter-site bias, we derived Cohen’s  $d$  effect size measure for each connection:  $|\beta_c|/\hat{\sigma}$ , with  $\beta_c$  being the weight associated with the contrast. The standard deviation from the noise  $\hat{\sigma}$  was calculated as  $\hat{\sigma} = \sqrt{\sum e^2/(N - K)}$ ,  $e$  being the residuals from the GLM,  $N$  the sample size and  $K$  the number of covariates in the model. As secondary analyses,  $t$ -tests were also implemented in the GLM to validate that age, sex as well as residual motion made significant contributions to the model.

<sup>10</sup>[http://niak.simexp-lab.org/pipe\\_preprocessing.html](http://niak.simexp-lab.org/pipe_preprocessing.html)

<sup>11</sup><https://github.com/SIMEXP/>

## 2.5. Simulations

*Data generation process.* We implemented Monte-Carlo simulations to assess the detection sensitivity of group differences in rs-fMRI connectivity. The simulations were based on the 1000 FCP sample, with 8 sites totaling 345 subjects. The multisite simulations were sampled from 148 subjects, available across  $S = 7$  sites. The monosite simulations were sampled from 195 subjects available at  $S = 1$  site (Cambridge). For each simulation, a subset of subjects of a given size  $N$  was selected randomly and stratified by site. For each site, a ratio  $W$  of the selected subjects were randomly assigned to a so-called “patient” group. We focus our analysis on connections showing a fair-to-good test-retest reliability based on a previous study reporting 11 connections likely impacted by Alzheimer’s disease, see Orban et al. (2015) for details. For each connection, a “pathology” effect was added to the connectivity measures of the subjects belonging to the “patient” group. This additive shift in connectivity for “patients” was selected as to achieve a specified effect size, defined below.

*Effect size (Cohen’s  $d$ ).* The Cohen’s  $d$  was used to quantify the effect size. For a group comparison, Cohen’s  $d$  is defined as the difference  $\mu$  between the means of the two groups, divided by the standard deviation of the measures within each group, here assumed to be equal. For a given connection between brain regions  $i$  and  $j$ , let  $y_{i,j}$  be the functional connectivity measure for a particular subject of the 1000 FCP sample. If the subject was assigned to the “patient” group in a particular simulation, an effect was added to generate a simulated connectivity measure  $y_{i,j}^*$  equals to  $y_{i,j} + \mu$ . For a specified effect size  $d$ , the parameter  $\mu$  was set to  $d \times s_{i,j}$ , where  $s_{i,j}$  is the standard deviation of connectivity between region  $i$  and  $j$ . The parameter  $s_{i,j}$  was estimated as the standard deviation of connectivity measures across subjects in the mono-site sample (Cambridge), without any “pathological” effect simulated.

*GLM tests.* In order to detect changes between the simulated groups at each pair of connection, a GLM was estimated from the simulated data, using age, sex and residual motion as confounds (corrected to have a zero mean across subjects). To account for site-specific biases,  $S - 1$  dummy variables (binary vectors coding for each site) were added to the model, with  $S$  being the total number of sites used in the study, in addition to an intercept accounting for the global average. Finally, one dummy variable coded for the “patient” group. The regression coefficients of the linear model were estimated with ordinary least squares, and a  $t$ -test, with associated  $p$ -value, was calculated for the coefficient of the “patient” variable. A significant pathology effect was detected if the  $p$  value was smaller than a prescribed  $\alpha$  level. The  $\alpha$  level needs to be adjusted for multiple comparisons (in our case 11 connections, but this would depend on the number of connections selected in a particular study), which can be done in an adaptive manner using FDR. We tested different typical values for  $\alpha$  in  $\{0.001, 0.01, 0.05\}$ . For each simulation sample  $b$  and each connection, we derived a  $p$ -value  $p^{(b)}$ , and the effect was deemed detected if  $p^{(b)}$  was lesser than

246  $\alpha$ . The sensitivity of the test for a particular connection was evaluated by the  
 247 frequency of positive detections over all simulation samples.

248 *Prediction accuracy.* In addition to mass univariate GLM tests, we also investi-  
 249 gated a linear SVM (Cortes and Vapnik, 1995) using a Monte Carlo simulation  
 250 framework similar to the one described above. For SVM simulations, all possible  
 251 connections between the 100 brain regions were used simultaneously to predict  
 252 the presence of the simulated pathology in a given subject. For a participant  
 253 assigned to the “patient” group, a “pathology” effect was only simulated in a  
 254 set percentage of connections, which were randomly selected. The proportion  
 255 of connections with a non-null effect was denoted as  $\pi_1$ . For a given simulation  
 256 at sample size  $N$ , the SVM model was trained on  $N$  subjects selected randomly  
 257 and stratified by site. The accuracy of the model was evaluated on a separate  
 258 simulation implemented with the remaining subjects, unused during training.  
 259 For example, for a multisite simulation with  $N = 80$  subjects for SVM training,  
 260 the model accuracy was estimated on 68 subjects: 148 (available subjects) minus  
 261 80 (subjects in the training set). During training, a 10-fold cross-validation was  
 262 used to optimize the hyper-parameters of the SVM independently for each sim-  
 263 ulation. The mean and standard deviation of accuracy scores across all samples  
 264 were derived for each simulation scenario.

265 *Simulation experiments.* All the simulation parameters have been summarized  
 266 below:

- 267 • Sample size  $N$ .
- 268 • Patient allocation ratio  $W$ .
- 269 • Number of sites  $S$ .
- 270 • The type of detection method, either GLM or SVM.
- 271 • For GLM tests, the false-positive rate  $\alpha$ .
- 272 • For SVM tests, the proportion of “pathological” connections  $\pi_1$ .
- 273 • The effect size  $d$ .

274 For a given set of simulation parameters, we generated  $B = 10^3$  Monte-Carlo  
 275 samples to estimate either the sensitivity (for GLM test) or the accuracy (for  
 276 SVM prediction) of the method. For all experiments, we investigated effect sizes  
 277  $d \in \{0, 2\}$  with a step of 0.01 and  $\alpha \in \{0.001, 0.01, 0.05\}$ . The number of site(s)  
 278 was  $S = 1$  for the monosite analysis and  $S = 7$  for the multisite analysis. We  
 279 implemented the following experiments:

- 280 • ( $\mathcal{E}_1$ ) Test the impact of the sample size on GLM  $N \in \{40, 80, 120\}$ , with  
 281 a fixed allocation ratio  $W = 0.5$ .
- 282 • ( $\mathcal{E}_2$ ) Test the impact of the allocation ratio on GLM  $W \in \{0.5, 0.3, 0.15\}$   
 283 for a fixed sample size  $N = 120$ .



- ( $\mathcal{E}_3$ ) Test the impact of multisite correction and affected connection volume ( $\pi_1$ ) on the prediction accuracy. For the prediction scenario, we used a range of  $\pi_1 \in \{0.1, 1, 5\%\}$ , and two sample sizes  $N \in \{80, 120\}$  subjects for training, with model accuracy estimated on  $N = 68$  and  $N = 28$ , respectively.

### 3. Results

#### 3.1. Inter-site bias in fMRI connectivity

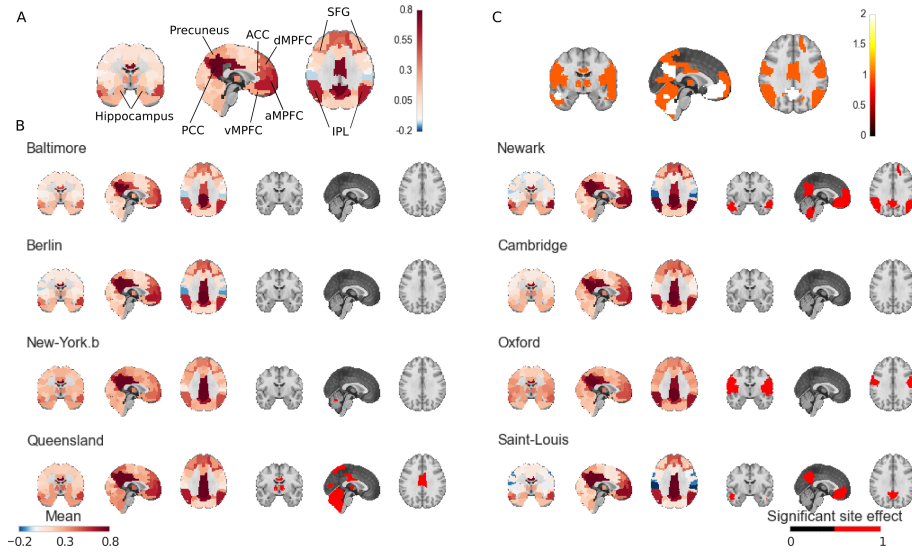


Figure 1: Panel A: map of the DMN obtained using a seed in the posterior cingulate cortex, averaging all subjects and sites together. Panels B: The first column shows the average functional connectivity maps of the DMN at 8 sites. The second column shows the significant differences between the average functional connectivity maps of one site versus all the others. Panel C shows the number of sites with a significant inter-site difference for each brain region.

*Site bias in the default-mode network.* We first focused on the connections associated with a seed region located in the posterior cingulate cortex, a key node of the default-mode network (DMN), which is one of the most widely studied resting-state network (Greicius et al., 2004). The connections were based on the Cambridge 100 parcellation, and were represented as a connectivity map, (Figure 1). Figure 1A shows the posterior cingulate cortex connectivity map, averaged across all subjects and all sites. The key regions of the DMN are easily identifiable, and include the posterior cingulate cortex, precuneus, inferior parietal lobule, anterior cingulate cortex, medial pre-frontal cortex (dorsal, anterior and ventral), superior frontal gyri and the medial temporal lobe (Damoiseaux et al., 2006; Dansereau et al., 2014; Yan et al., 2013b). Using a GLM, the

302 average connectivity map of the DMN was then extracted for each site, Figure  
 303 1B. Qualitatively, the DMN maps were consistent across sites, as expected  
 304 based on the literature. We then tested for the significance of the site bias, i.e.  
 305 the difference in average connectivity at a given site and the average connec-  
 306 tivity at all remaining sites. The statistical maps were corrected for multiple  
 307 comparisons across the brain with FDR at  $q \leq 0.05$  (Benjamini and Hochberg,  
 308 1995). A significant bias for at least one connection could be identified for ev-  
 309 ery site, without exception, Figure 1B. Figure 1C shows how reproducible were  
 310 the significant biases in connectivity across the brain and sites. The identified  
 311 significant connections locations were quite variable across sites, most of them  
 312 being identified at less than three sites.

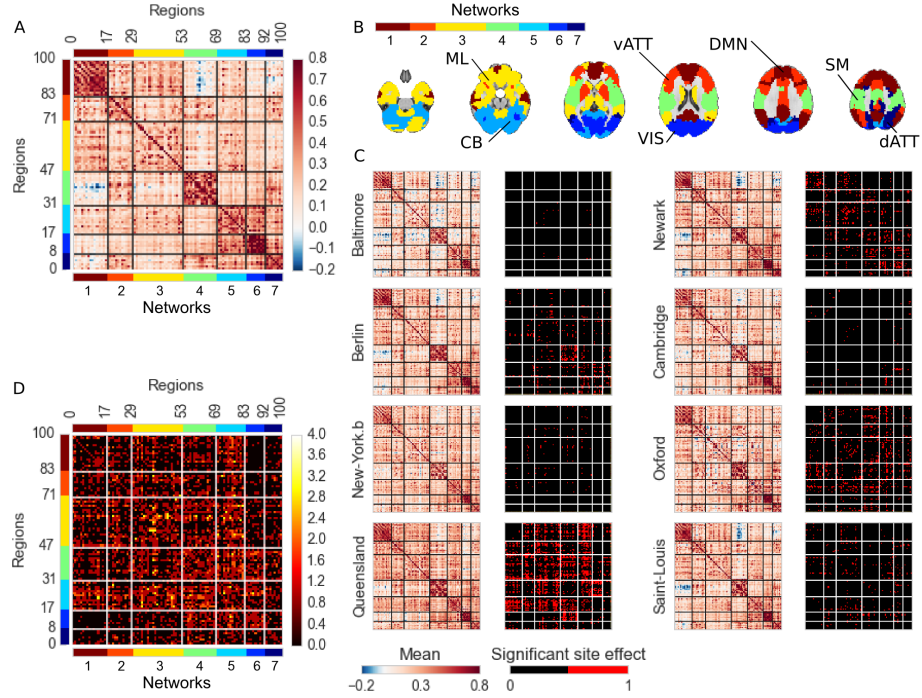


Figure 2: Panel A shows the average functional connectomes for 8 sites of the 1000 FCP. Colors next to the  $x$  and  $y$  axis correspond to different networks in a 7-cluster solution of the matrix, obtained from a hierarchical clustering (Ward criterion). Panel B presents the corresponding 7 brain networks, along with labels. Panel C shows average connectomes for individual sites, as well as connections with a significant site bias. Panel D shows the number of sites at which a given connection was detected as significantly biased. ML: mesolimbic, CB: cerebellar, VIS: visual, vATT: ventral attentional, dATT: dorsal attentional, DMN: default mode network, SM: sensorimotor.

313 *Site bias across the connectome.* In order to extend these observations outside  
 314 of the DMN, we derived the entire connectome using the Cambridge 100 par-

cellation. Figure 2A shows the average connectome, pooling all subjects and sites together. The regions have been re-ordered based on a hierarchical clustering with Ward criterion. A network structure is clearly visible as squares of high connectivity on the diagonal of the connectome (as outlined by black lines). Each diagonal square corresponds to the intra-network connectivity for a partition into 7 networks (Figure 2A). These 7 networks were consistent with the major resting-state networks reported using a cluster analysis in previous works (e.g. van den Heuvel et al., 2008; Bellec et al., 2010a; Yeo et al., 2011; Power et al., 2011): the DMN, visual, sensorimotor, dorsal and ventral attentional networks, mesolimbic and cerebellar networks were identified (Figure 2B). Figure 2C shows how this large-scale connectome organization varied from site to site. The average connectivity per site as well as significant differences with the average of the remaining sites ( $q \leq 0.05$ ) is shown in Figure 2C. Visually, consistent with our previous observations in the DMN, the organization of the average connectome into large-scale resting-state networks was preserved across all sites. Some significant site effects were still detected in the connectivity both within each network, as well as between networks. By counting the number of sites showing a significant effect for each pair of region, it was apparent that significant site effects were quite variable in their localization and spread across the full connectome (Figure 2D). Concerning the association with the other confounding variables in the model (sex, age and motion) many connections were found to be significantly associated with motion, see Supplementary Material Figure S3, although very few connections were found to be significantly associated with the sex and age, see Supplementary Material Figure S4 and S5. This finding supports the inclusions of age, sex and FD in a GLM in order to remove their confounding effects in addition to site effects.

*Site bias vs. within-site variations across subjects.* We measured the amplitude of inter-site bias, represented as violin plots across connections using either the absolute difference in average connectivity (Figure 3A,C) or Cohen’s  $d$  effect size measures (Figure 3B,D). The violin plots include either every pairs of connections from the BASC Cambridge parcellation (Figure 3A,B), or only the 11 connections selected for Monte-Carlo simulations (Figure 3C,D). For absolute differences, the distributions were mostly consistent across sites, with a median around 0.06, 5% percentile near 0 and 95% percentiles in the 0.08- 0.1 range. For Cohen’s  $d$ , the distributions were also consistent across sites, with a median around 0.33, 5% percentile near 0 and 95% percentiles in the 0.4- 0.6 range. This effect size would be deemed small-to-moderate, which suggests that the impact of additive inter-site bias on statistical tests will be limited. Similar findings were observed across all possible connections, or across the 11 pairs of connections selected in the simulation study. The region-to-region maps of within-site, across-subjects standard deviations are presented for the DMN in Supplementary Figure S1, and for the entire connectome in the Supplementary Material S2.

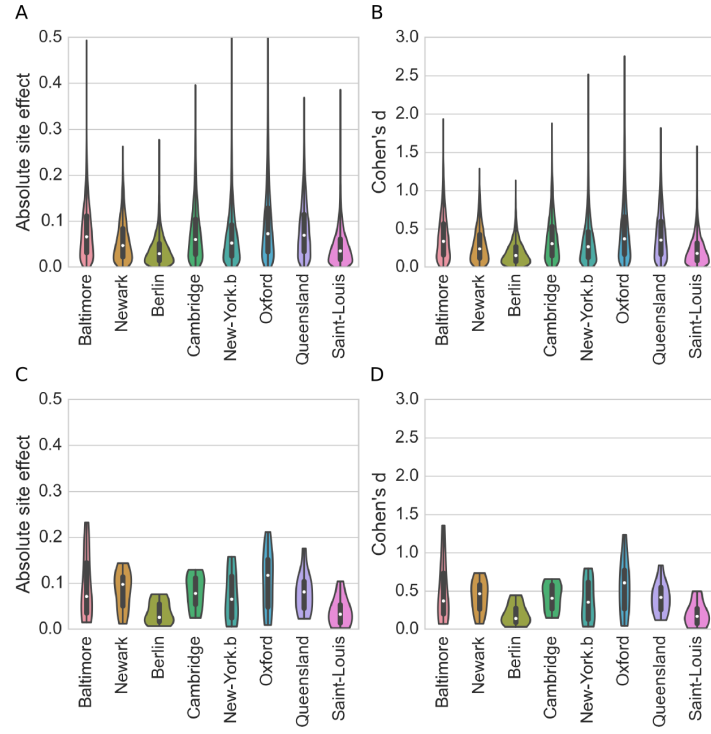


Figure 3: Effect size of the inter-site bias from a subset of 8 sites from the 1000 FCP. Panels A,C show the distribution of absolute differences in functional connectivity, while panels B,D show Cohen's  $d$  measures of inter-site bias. Panels A,B show violin plots across every connections in the BASC Cambridge 100 parcellation, while Panels C,D focus on the selected 11 functional connections used in simulations, only.

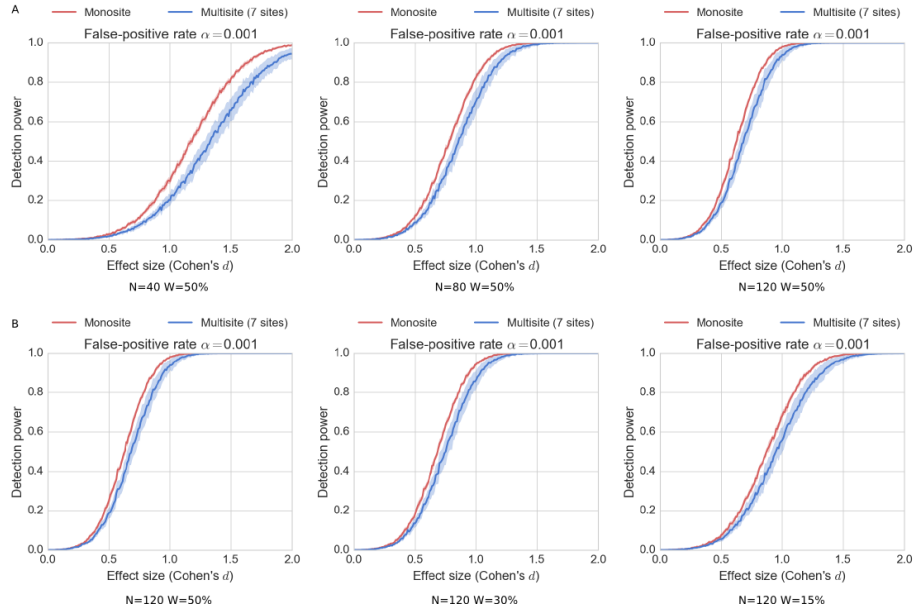


Figure 4: Monte-Carlo simulation of detection power as a function of the effect size  $d \in [0, 2]$ , either for a monosite ( $S = 1$ , in red) or a multisite ( $S = 7$ , in blue) sample, when testing differences between two groups with a GLM and a false-positive rate  $\alpha = 0.001$ . The plain curves are the average statistical power across 11 connections, and the shaded area represent  $\pm 1$  standard deviation across connections. In panel A, the patient allocation ratio is fixed ( $W = 50\%$ ) and three different sample sizes have been tested,  $N \in \{40, 80, 120\}$  (Experiment ( $\mathcal{E}_1$ )). In panel B, the sample size is fixed ( $N = 120$ ) and three different patient allocation ratios have been tested  $W \in \{15\%, 30\%, 50\%\}$  (Experiment ( $\mathcal{E}_2$ )).

### 3.2. Multisite Monte-Carlo simulations

*Statistical power and effect size.* Figure 4A shows the relationship between effect size and a GLM detection power in experiment ( $\mathcal{E}_1$ ), i.e. for a fixed allocation ratio ( $W = 50\%$ ) and three different sample sizes,  $N \in \{40, 80, 120\}$ . The average and std of detection power was plotted across the 11 selected connections. The variations of statistical power across connections were very small for monosite simulations, as the effect size was adjusted based on the standard deviation of each connection within that sample. As expected, the sensitivity increased with sample size, quite markedly. In multisite simulations ( $S = 7$ ), for a large effect size ( $d = 1$ ), the detection power was 20% with 40 subjects, 80% with 80 subjects and 95% with 120 subjects. The sensitivity was larger with a single site than multisite sample, yet the difference between the two decreased as sample size increased. With  $N = 40$  and  $d = 1$ , the detection power was close to 30% for a single site sample, compared to 20% for the multisite sample. With  $N = 120$  and  $d = 1$ , the difference in sensitivity was only of a few percent. The same trend was apparent for all tested effect sizes as well as for  $\alpha \in \{0.01, 0.05\}$  (not shown).

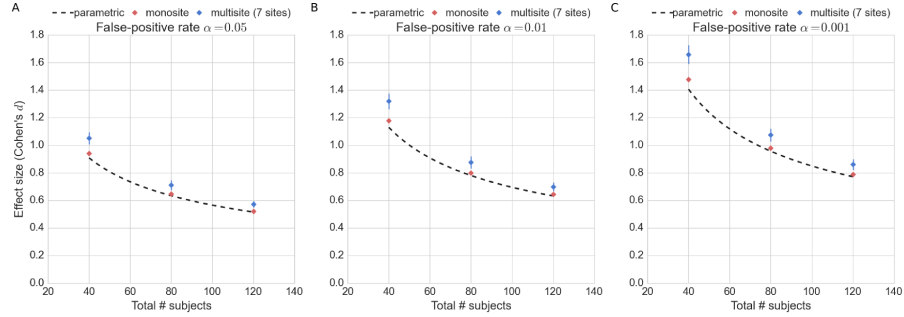


Figure 5: Effect size detectable at 80% sensitivity as a function of sample size, for different false-positive rate  $\alpha \in \{0.05, 0.01, 0.001\}$  (experiment ( $\mathcal{E}_1$ )). All simulations used a balanced patient allocation ratio  $W = 50\%$ . The monosite performance is shown in red and the multisite in blue. The dotted black line shows the detectable effect size for a classical parametric  $t$ -test.

*Statistical power and group allocation ratio.* Figure 4B shows the relationship between effect size and a GLM detection power in experiment ( $\mathcal{E}_2$ ), i.e. for a fixed sample size ( $N = 120$ ) and three different patient allocation ratio,  $W \in \{15\%, 30\%, 50\%\}$ . Overall, we found that the detection power increased with  $W$ . For example, with  $d = 1$ , the detection power was 65% for  $W = 15\%$ , and increased to 90% with  $W = 30\%$ , and finally 95% for  $W = 50\%$ . The impact of  $W$  was observed in both monosite and multisite samples, with an optimal allocation ratio of  $W = 50\%$  for both. This observation was also made for  $\alpha \in \{0.01, 0.05\}$  (not shown).

*Detectable effect size, as a function of sample size.* An alternative summary of experiment ( $\mathcal{E}_1$ ) is to represent the effect size that can be detected with 80%

386 sensitivity, as a function of sample size for monosite and multisite configurations,  
 387 see Figure 5. As a reference, we computed the same curve for parametric  $t$ -test  
 388 comparisons, under assumptions of normality. As expected, the detectable effect  
 389 size for parametric  $t$ -tests closely followed the monosite estimation. For a small  
 390 sample size ( $N = 40$ ), the detectable effect size was notably larger in multisite  
 391 configurations than in a monosite configuration (difference of about 0.25 in  
 392 Cohen's  $d$  for  $\alpha = 0.001$ ). However, the difference decreased for large sample  
 393 sizes to become smaller than 0.1 with  $N = 120$  and  $\alpha = 0.001$ . The lowest  
 394 detectable effect size for a sensitivity of 80% at  $\alpha = 0.05$  was about  $d = 0.8$ ,  
 395 achieved in a monosite configuration with  $N = 120$ . At this sample size, the  
 396 difference between single and multisite configurations was marginal, with only  
 397 a few percent's of difference in detectable effect sizes.

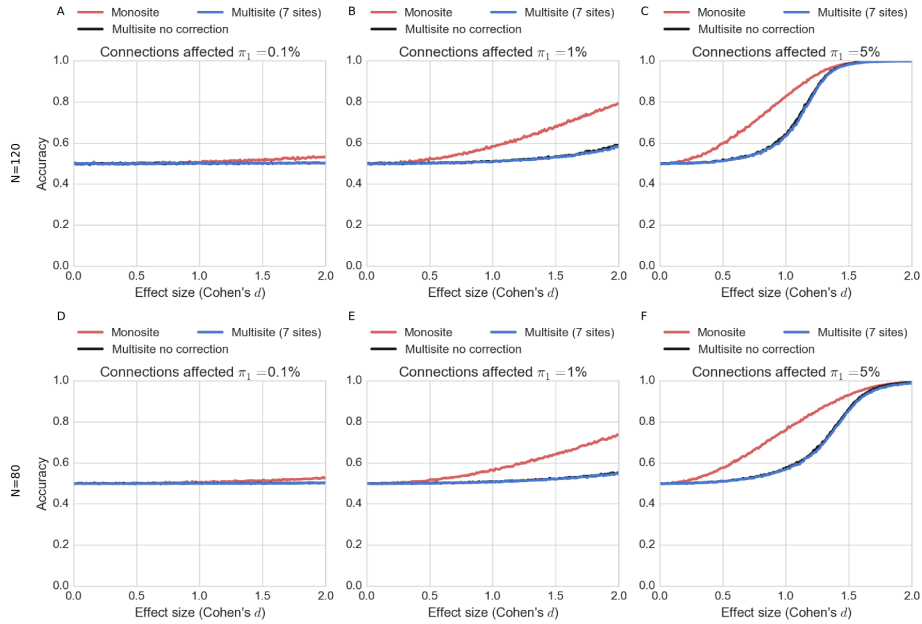


Figure 6: Prediction accuracy of patient vs. controls as a function of effect size. Three simulation settings are presented on each plot: monosite (red curve), multisite with regression of site effects ( $S = 7$ , blue curve), and multisite without regression of site effects ( $S = 7$ , black curve). Accuracy was estimated over  $B = 10^3$  simulation samples with a patient allocation ratio  $W = 50\%$  and 3 volumes of affected connections  $\pi_1 = 0.1\%$  (left column),  $\pi_1 = 1\%$  (middle column) and  $\pi_1 = 5\%$  (right column). Two sample sizes were tested:  $N = 120$  randomly selected subjects for training, with the remaining  $N = 28$  to estimate accuracy (first row), and  $N = 80$  randomly selected subjects for training, with the remaining  $N = 68$  to estimate accuracy (second row).

398 *Prediction accuracy.* In experiment ( $\mathcal{E}_3$ ), we examined the impact of effect size  
 399 and the volume of affected connections on prediction accuracy in a SVM, see  
 400 Figure 6. The volume of changes  $\pi_1$  had a major impact on prediction accu-

401 racy. At  $\pi_1 = 0.1\%$  (around 5 connections) the accuracy level was at chance  
 402 level across all tested effect sizes, (Figure 6A). With  $\pi_1 = 1\%$ , accuracy slightly  
 403 increased, but effect sizes larger than  $d = 2$  were still required to reach over 80%  
 404 accuracy (Figure 6B). With  $\pi_1 = 5\%$ , 95% accuracy was achieved at the same  
 405 effect size (about  $d = 1.5$ ) for monosite and multisite simulations, although the  
 406 accuracy in multisite simulations was notably lower than for monosite simula-  
 407 tions across most effect sizes (Figure 6C). The relationship between effect size  
 408 and accuracy followed a sigmoidal curve in both settings, yet a sharper, and  
 409 latter transition between very low and very high accuracy was observed in mul-  
 410 tisite simulations. Interestingly, correcting for site effects by regressing out the  
 411 dummy variable before running the SVM classifier had no impact on accuracy  
 412 levels. The sample size ( $N = 80$  vs  $N = 120$  for training) did have a moderate  
 413 effect on prediction accuracy: for  $\pi_1 = 5\%$  and  $d = 1$  and monosite simulations,  
 414 accuracy was about 85% with  $N = 120$  (Figure 6C) and 75% with  $N = 80$   
 415 (Figure 6F).

#### 416 4. Discussion and conclusions

417 *Inter-site bias in rs-fMRI connectivity.* Typical resting-state networks, such as  
 418 the DMN, the attentional, visual and sensorimotor networks, were reliably found  
 419 across sites. This was strongly expected given the relative consistency of their  
 420 distribution across individuals, studies, preprocessing approaches or even meth-  
 421 ods used to extract networks (e.g. Damoiseaux et al., 2006; van den Heuvel  
 422 et al., 2008; Bellec et al., 2010b; Yeo et al., 2011; Power et al., 2011). We how-  
 423 ever found that significant differences in average connectivity existed between  
 424 sites (i.e. site biases), as previously reported by Yan et al. (2013b). This connec-  
 425 tivity bias may undermine the generalization of the results derived at a single  
 426 site. The inter-subject (intra-site) standard deviation of the connections was  
 427 found to be more than twice as large as the inter-site absolute bias, on aver-  
 428 age across brain connections. This effect size measured in Cohen’s  $d$  would be  
 429 deemed small-to-moderate, which suggests that the impact of additive inter-site  
 430 bias on statistical tests will be limited. This is a reassuring finding supporting  
 431 the feasibility of statistical tests pooling fMRI data across multiple sites. Pre-  
 432 vious studies (Sutton et al., 2008; Brown et al., 2011) had reported inter-site  
 433 variance up to 10 times smaller than inter-subject variability, but these studies  
 434 had much more homogeneous scanning environments than ours and also used  
 435 different fMRI outcome measures. In our case, we still investigated only 3T  
 436 scanners, mostly Siemens, and inter-site bias may be larger when considering  
 437 other manufacturers or field strengths.

439 *Statistical power and multisite rs-fMRI.* After accounting for site-related addi-  
 440 tive bias in a GLM, the multisite simulation pooling 7 sites together showed  
 441 detection power close to that of a monosite simulation with equivalent sample  
 442 size. The difference was noticeable for small sample size (total  $N = 40$ ), and  
 443 became very small for a sample size  $N = 120$ . Another observation was that,



for a given detection power, the lowest effect size that we were able to detect was more variable across connections for a low sample size. Taken together, these observations suggest to use sample size larger than 100 subjects for GLM multisite studies. A limitation of this conclusion is that it is likely dependent on the number of sites pooled in the study and the actual number of subjects in each of those sites.

*Statistical power and sample size.* For a medium effect size, e.g.  $d = 0.5$ , the sensitivity was low (below 20%), even for monosite simulations with  $N = 120$  subjects. This sobering result supports the current trend in the literature to pool multiple data samples to increase sample size, at the cost of decreased homogeneity. We also found that resting-state studies based on 40 subjects or less, even at a single site, are seriously underpowered, except for extremely large effect sizes (Cohen’s  $d$  greater than 1.5). Finally, unbalanced patient allocation ratio in site samples greatly reduces sensitivity, even in monosite studies. Balanced datasets, i.e. with equal numbers of patients and controls at each site, should therefore be favored.

*Prediction.* Comparing the monosite and the multisite accuracy curves reveals a substantial drop in accuracy from monosite to multisite across a broad range of effect sizes. However, it should be noted that classifiers trained across multiple data sources will likely generalize better to new observations, which is likely a critical feature in most applications and reflects the true potential clinical utility of this type of technique. Our conclusions are consistent with the work of Nielsen et al. (2013), which compares the prediction of a clinical diagnosis of autism in monosite vs. multisite settings. The authors concluded that the prediction accuracy for the multisite sample was significantly smaller than for the monosite sample. A somewhat surprising observation in our analysis was that linear correction for site-specific biases did not improve accuracy of prediction using SVM. The SVM model seems to learn features that are invariant across sites, maybe focusing on connections with the smallest site bias, or looking at differences between connections similarly impacted by a site bias. Finally, an important conclusion of our simulations was that the volume of brain connections affected by a disease impacts as much accuracy as the effect size per connection. This suggests that feature reduction and/or selection is a very important step to improve sensitivity to small effect sizes.

*Beyond additive bias.* An important limitation to our study is that we only investigated the impact of additive bias in brain connectivity across sites. Areas of future works include interactions between site effects and pathology, possibly in the form of polynomial and non-linear interactions. We hope that, in the future, fMRI data acquired on clinical cohorts at tens of sites will become available, which will enable researchers to test empirically the presence of such interaction effects.

485 *Other types of multisite data.* Another limitation of our study is that we only  
 486 investigated multisite data featuring roughly equal sample sizes with fairly bal-  
 487 anced patient allocation ratios at each site. Multisite studies including a very  
 488 large number of sites with sometimes only a few subjects per site are however  
 489 quite common, e.g. the Alzheimer’s disease neuroimaging initiative (ADNI)  
 490 (Mueller et al., 2005) and many pharmaceutical clinical trials at phase II and  
 491 III<sup>12</sup>. In this type of design, the multisite effect may play a much more pro-  
 492 nounced role than in our simulations as it cannot be modeled in the GLM, and  
 493 will become an intrinsic added source of inter-subject variance. Unfortunately,  
 494 this type of design could no be tested with the current dataset due to the limited  
 495 number of sites available. This represents an important avenue of future work.

496 *Underlying causes of the site bias.* Not all sites seemed to be equally biased,  
 497 with sites like Berlin or Saint-Louis showing a small number of connections sig-  
 498 nificantly different then the grand average connectivity matrix, while sites like  
 499 Baltimore, Queensland and Oxford showed much more biased connectivity mea-  
 500 sures. These differences may not be statistically significant, or they may reflect  
 501 real differences due to protocol or scanner characteristics at these sites. Multiple  
 502 causes may be interacting together to produce the site bias, as reported by Yan  
 503 et al. (2013b), although some of these sources of variance could be better con-  
 504 trolled like the scanner parameters, paired with the use of a phantom to promote  
 505 more homogeneous configurations across sites (Friedman et al., 2006; Friedman  
 506 and Glover, 2006; Glover et al., 2012). Even in standardized experiments, it  
 507 should be noted that differences in scanner protocols remain (Brown et al.,  
 508 2011). A much larger multisite sample with systematically varying parameters  
 509 could enable a data-driven identification of the critical parameters impacting  
 510 site bias. The various releases made by the INDI initiative may fill that gap in  
 511 the literature in the future, as the scanner protocols are much better described  
 512 in recent releases, such as CoRR (Zuo et al., 2014), than they were in the initial  
 513 FCP release.

## 514 5. Acknowledgments

515 Parts of this work were presented at the 2013 annual meeting of the Orga-  
 516 nization for Human Brain Mapping, as well as the 2013 Alzheimer’s Associa-  
 517 tion International Conference (AAIC) (Dansereau et al., 2013). The authors  
 518 are grateful to the members of the 1000 functional connectome consortium for  
 519 publicly releasing their datasets. The computational resources used to per-  
 520 form the data analysis were provided by ComputeCanada<sup>13</sup> and CLUMEQ<sup>14</sup>,  
 521 which is funded in part by NSERC (MRS), FQRNT, and McGill University.  
 522 This project was funded by NSERC grant number RN000028 and the Canadian

<sup>12</sup><http://www.rochetrials.com/trialDetailsGet.action?studyNumber=BP28248>

<sup>13</sup><https://computeCanada.org/>

<sup>14</sup><http://www.clumeq.mcgill.ca/>

523 Consortium on Neurodegeneration in Aging (CCNA), through a grant from the  
 524 Canadian Institute of Health Research and funding from several partners in-  
 525 cluding SANOFI-ADVENTIS R&D. PB is supported by a salary award from  
 526 “Fonds de recherche du Québec – Santé” and the Courtois Foundation.

## 527 References

- 528 Ad-Dab’bagh, Y., Einarson, D., Lyttelton, O., Muehlboeck, J. S., Mok, K.,  
 529 Ivanov, O., Vincent, R. D., Lepage, C., Lerch, J., Fombonne, E., Evans,  
 530 A. C., 2006. The CIVET Image-Processing environment: A fully automated  
 531 comprehensive pipeline for anatomical neuroimaging research. In: Corbetta,  
 532 M. (Ed.), Proceedings of the 12th Annual Meeting of the Human Brain Map-  
 533 ping Organization. Neuroimage, Florence, Italy.
- 534 Anand, A., Li, Y., Wang, Y., Gardner, K., Lowe, M. J., 2007. Reciprocal ef-  
 535 fects of antidepressant treatment on activity and connectivity of the mood  
 536 regulating circuit: an fmri study. The Journal of neuropsychiatry and clinical  
 537 neurosciences.
- 538 Bellec, P., Carbonell, F., Perlberg, V., Evans, A. C., 2010a. A neuroimaging  
 539 analysis kit for Octave and Matlab.  
 540 URL <http://code.google.com/p/niak/>
- 541 Bellec, P., Carbonell, F. M., Perlberg, V., Lepage, C., Lyttelton, O., Fonov,  
 542 V., Janke, A., Tohka, J., Evans, A. C., 2011. A neuroimaging analysis kit for  
 543 Matlab and Octave. In: Proceedings of the 17th International Conference on  
 544 Functional Mapping of the Human Brain. pp. In Press+.
- 545 Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P.,  
 546 Evans, A. C., 2012. The pipeline system for Octave and Matlab (PSOM): a  
 547 lightweight scripting framework and execution engine for scientific workflows.  
 548 Frontiers in neuroinformatics 6.  
 549 URL <http://dx.doi.org/10.3389/fninf.2012.00007>
- 550 Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., Evans, A. C., Jul. 2010b.  
 551 Multi-level bootstrap analysis of stable clusters in resting-state fMRI. Neu-  
 552 roImage 51 (3), 1126–1139.  
 553 URL <http://dx.doi.org/10.1016/j.neuroimage.2010.02.082>
- 554 Benjamini, Y., Hochberg, Y., 1995. Controlling the false-discovery rate: a prac-  
 555 tical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B  
 556 57, 289–300.
- 557 Biswal, B. B., Mennes, M., Zuo, X.-N. N., Gohel, S., Kelly, C., Smith, S. M.,  
 558 Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski,  
 559 A.-M. M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S.,  
 560 Kiviniemi, V. J., Kötter, R., Li, S.-J. J., Lin, C.-P. P., Lowe, M. J., Mackay,

561 C., Madden, D. J., Madsen, K. H., Margulies, D. S., Mayberg, H. S., McMa-  
562 hon, K., Monk, C. S., Mostofsky, S. H., Nagel, B. J., Pekar, J. J., Peltier,  
563 S. J., Petersen, S. E., Riedl, V., Rombouts, S. A., Rypma, B., Schlaggar, B. L.,  
564 Schmidt, S., Seidler, R. D., Siegle, G. J., Sorg, C., Teng, G.-J. J., Veijola, J.,  
565 Villringer, A., Walter, M., Wang, L., Weng, X.-C. C., Whitfield-Gabrieli, S.,  
566 Williamson, P., Windischberger, C., Zang, Y.-F. F., Zhang, H.-Y. Y., Castel-  
567 lanos, F. X., Milham, M. P., Mar. 2010. Toward discovery science of human  
568 brain function. *Proceedings of the National Academy of Sciences of the United*  
569 *States of America* 107 (10), 4734–4739.  
570 URL <http://dx.doi.org/10.1073/pnas.0911855107>

571 Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N.,  
572 McCarthy, G., Voyvodic, J., Glover, G., Diaz, M., et al., 2011. Multisite  
573 reliability of cognitive bold data. *Neuroimage* 54 (3), 2163–2175.

574 Caramanos, Z., Fonov, V. S., Francis, S. J., Narayanan, S., Pike, G. B., Collins,  
575 D. L., Arnold, D. L., 2010. Gradient distortions in mri: Characterizing  
576 and correcting for their effects on siena-generated measures of brain volume  
577 change. *NeuroImage* 49 (2), 1601–1611.

578 Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines.  
579 *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27, soft-  
580 ware available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

581 Cheng, W., Palaniyappan, L., Li, M., Kendrick, K. M., Zhang, J., Luo, Q., Liu,  
582 Z., Yu, R., Deng, W., Wang, Q., Ma, X., Guo, W., Francis, S., Liddle, P.,  
583 Mayer, A. R., Schumann, G., Li, T., Feng, J., May 2015. Voxel-based, brain-  
584 wide association study of aberrant functional connectivity in schizophrenia  
585 implicates thalamocortical circuitry. *Npj Schizophrenia* 1, –.  
586 URL <http://dx.doi.org/10.1038/npjschz.2015.16>

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

591 Cortes, C., Vapnik, V., Sep. 1995. Support-vector networks. *Machine Learning*  
592 20 (3), 273–297.  
593 URL <http://dx.doi.org/10.1007/BF00994018>

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

599 Dansereau, C., Bellec, P., Lee, K., Pittau, F., Gotman, J., Grova, C., 2014.  
600 Detection of abnormal resting-state networks in individual patients suffering

601 from focal epilepsy: An initial step toward individual connectivity assess-  
602 ment. *Frontiers in Neuroscience* 8 (419).  
603 URL [http://www.frontiersin.org/brain\\_imaging\\_methods/10.3389/fnins.2014.00419/abstract](http://www.frontiersin.org/brain_imaging_methods/10.3389/fnins.2014.00419/abstract)  
604

605 Dansereau, C., Risterucci, C., Pich, E. M., Arnold, D., Bellec, P., 2013. A  
606 power analysis for multisite studies in resting-state functional connectivity,  
607 with an application to clinical trials in alzheimer’s disease. Vol. 9. pp. P248  
608 – P249, alzheimer’s Association International Conference 2013 Alzheimer’s  
609 Association International Conference 2013.  
610 URL <http://www.sciencedirect.com/science/article/pii/S1552526013011461>  
611

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

616 Edward, V., Windischberger, C., Cunnington, R., Erdler, M., Lanzenberger, R.,  
617 Mayer, D., Endl, W., Beisteiner, R., Nov 2000. Quantification of fmri artifact  
618 reduction by a novel plaster cast head holder. *Hum Brain Mapp* 11 (3), 207–  
619 213.

620 Elliott, M. R., Bowtell, R. W., Morris, P. G., Jun 1999. The effect of scanner  
621 sound in visual, motor, and auditory functional mri. *Magn Reson Med* 41 (6),  
622 1230–1235.

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

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

635 Friedman, L., Glover, G., Jun. 2006. Report on a multicenter fmri quality assur-  
636 ance protocol. *Journal of magnetic resonance imaging : JMRI* 23 (6), 827–839.  
637 URL <http://dx.doi.org/10.1002/jmri.20583>

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

- 642 Giove, F., Gili, T., Iacovella, V., Macaluso, E., Maraviglia, B., Oct. 2009.  
 643 Images-based suppression of unwanted global signals in resting-state func-  
 644 tional connectivity studies. *Magnetic resonance imaging* 27 (8), 1058–1064.  
 645 URL <http://dx.doi.org/10.1016/j.mri.2009.06.004>
- 646 Glover, G. H., Mueller, B. A., Turner, J. A., van Erp, T. G., Liu, T. T., Greve,  
 647 D. N., Voyvodic, J. T., Rasmussen, J., Brown, G. G., Keator, D. B., et al.,  
 648 2012. Function biomedical informatics research network recommendations for  
 649 prospective multicenter functional mri studies. *Journal of Magnetic Reso-*  
 650 *nance Imaging* 36 (1), 39–54.
- 651 Greicius, M. D., Srivastava, G., Reiss, A. L., Menon, V., Mar. 2004. Default-  
 652 mode network activity distinguishes Alzheimer’s disease from healthy aging:  
 653 Evidence from functional MRI. *Proceedings of the National Academy of Sci-*  
 654 *ences of the United States of America* 101 (13), 4637–4642.  
 655 URL <http://dx.doi.org/10.1073/pnas.0308627101>
- 656 Hunter, J. D., 2007. Matplotlib: A 2d graphics environment. *Computing In*  
 657 *Science & Engineering* 9 (3), 90–95.
- 658 Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartrés-  
 659 Faz, D., Arnold, J., Benninghoff, J., Fiedler, U., Roccatagliata, L., et al.,  
 660 2016. Longitudinal reproducibility of default-mode network connectivity in  
 661 healthy elderly participants: A multicentric resting-state fmri study. *Neu-*  
 662 *roImage* 124, 442–454.
- 663 Kilpatrick, L., Zald, D., Pardo, J., Cahill, L., 2006. Sex-related differences  
 664 in amygdala functional connectivity during resting conditions. *Neuroimage*  
 665 30 (2), 452–461.
- 666 Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., Nichols, T. E., Jan. 2006.  
 667 Non-white noise in fMRI: does modelling have an impact? *NeuroImage* 29 (1),  
 668 54–66.  
 669 URL <http://dx.doi.org/10.1016/j.neuroimage.2005.07.005>
- 670 Milham, M. P., Fair, D., Mennes, M., Mostofsky, S. H., 2012. The adhd-200  
 671 consortium: a model to advance the translational potential of neuroimaging  
 672 in clinical neuroscience. *Frontiers in Systems Neuroscience* 6 (62).  
 673 URL [http://www.frontiersin.org/systems\\_neuroscience/10.3389/](http://www.frontiersin.org/systems_neuroscience/10.3389/fnsys.2012.00062/full)  
 674 [fnsys.2012.00062/full](http://www.frontiersin.org/systems_neuroscience/10.3389/fnsys.2012.00062/full)
- 675 Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust,  
 676 W., Trojanowski, J. Q., Toga, A. W., Beckett, L., Nov 2005. The alzheimer’s  
 677 disease neuroimaging initiative. *Neuroimaging Clin N Am* 15 (4), 869–77, xi–  
 678 xii.  
 679 URL <http://dx.doi.org/10.1016/j.nic.2005.09.008>

- 680 Nielsen, J., Zielinski, B., Fletcher, T., Alexander, A., Lange, N., Bigler, E.,  
681 Lainhart, J., Anderson, J., 2013. Multisite functional connectivity mri classi-  
682 fication of autism: Abide results. *Frontiers in human neuroscience* 7, -.  
683 URL <http://view.ncbi.nlm.nih.gov/pubmed/24093016>
- 684 Orban, P., Madjar, C., Savard, M., Dansereau, C., Tam, A., Das, S., Evans,  
685 A. C., Rosa-Neto, P., Breitner, J. C., Bellec, P., oct 2015. Test-retest resting-  
686 state fMRI in healthy elderly persons with a family history of alzheimer's  
687 disease. *Scientific Data* 2, 150043.  
688 URL <http://dx.doi.org/10.1038/sdata.2015.43>
- 689 Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E.,  
690 Feb. 2012. Spurious but systematic correlations in functional connectivity  
691 MRI networks arise from subject motion. *NeuroImage* 59 (3), 2142–2154.  
692 URL <http://dx.doi.org/10.1016/j.neuroimage.2011.10.018>
- 693 Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church,  
694 J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., Pe-  
695 tersen, S. E., Nov. 2011. Functional Network Organization of the Human  
696 Brain. *Neuron* 72 (4), 665–678.  
697 URL <http://dx.doi.org/10.1016/j.neuron.2011.09.006>
- 698 Shehzad, Z., Kelly, C. M., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q.,  
699 Lee, S. H. H., Margulies, D. S., Roy, A. K. K., Biswal, B. B., Petkova, E.,  
700 Castellanos, F. X., Milham, M. P., Oct. 2009. The resting brain: uncon-  
701 strained yet reliable. *Cerebral cortex (New York, N.Y. : 1991)* 19 (10), 2209–  
702 2229.  
703 URL <http://dx.doi.org/10.1093/cercor/bhn256>
- 704 Sheline, Y. I., Price, J. L., Yan, Z., Mintun, M. A., 2010. Resting-state func-  
705 tional mri in depression unmasks increased connectivity between networks via  
706 the dorsal nexus. *Proceedings of the National Academy of Sciences* 107 (24),  
707 11020–11025.
- 708 Sutton, B. P., Goh, J., Hebrank, A., Welsh, R. C., Chee, M. W., Park, D. C.,  
709 2008. Investigation and validation of intersite fmri studies using the same  
710 imaging hardware. *Journal of Magnetic Resonance Imaging* 28 (1), 21–28.
- 711 Tam, A., Dansereau, C., Badhwar, A., Orban, P., Belleville, S., Chertkow, H.,  
712 Dagher, A., Hanganu, A., Monchi, O., Rosa-Neto, P., Shmuel, A., Wang,  
713 S., Breitner, J., Bellec, P., 2015. Common effects of amnesic mild cognitive  
714 impairment on resting-state connectivity across four independent studies.  
715 *Frontiers in Aging Neuroscience* 7 (242).  
716 URL [http://www.frontiersin.org/aging\\_neuroscience/10.3389/fnagi.2015.00242/abstract](http://www.frontiersin.org/aging_neuroscience/10.3389/fnagi.2015.00242/abstract)
- 718 van den Heuvel, M., Mandl, R., Hulshoff Pol, H., Apr. 2008. Normalized Cut  
719 Group Clustering of Resting-State fMRI Data. *PLoS ONE* 3 (4), e2001+.  
720 URL <http://dx.doi.org/10.1371/journal.pone.0002001>

- 721 Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W.,  
722 Buckner, R. L., Jan. 2010. Intrinsic functional connectivity as a tool for human  
723 connectomics: theory, properties, and optimization. *Journal of neurophysiol-*  
724 *ogy* 103 (1), 297–321.  
725 URL <http://dx.doi.org/10.1152/jn.00783.2009>
- 726 Van Dijk, K. R., Sabuncu, M. R., Buckner, R. L., 2012. The influence of head  
727 motion on intrinsic functional connectivity mri. *Neuroimage* 59 (1), 431–438.
- 728 Vanhoutte, G., Verhoye, M., der Linden, A. V., May 2006. Changing body  
729 temperature affects the t2\* signal in the rat brain and reveals hypothalamic  
730 activity. *Magn Reson Med* 55 (5), 1006–1012.  
731 URL <http://dx.doi.org/10.1002/nrm.20861>
- 732 Worsley, K. J., Friston, K. J., Sep. 1995. Analysis of fMRI Time-Series  
733 Revisited—Again. *NeuroImage* 2 (3), 173–181.  
734 URL <http://dx.doi.org/10.1006/ning.1995.1023>
- 735 Yan, C., Liu, D., He, Y., Zou, Q., Zhu, C., Zuo, X., Long, X., Zang, Y., May  
736 2009. Spontaneous Brain Activity in the Default Mode Network Is Sensitive to  
737 Different Resting-State Conditions with Limited Cognitive Load. *PLoS ONE*  
738 4 (5), e5743+.  
739 URL <http://dx.doi.org/10.1371/journal.pone.0005743>
- 740 Yan, C.-G., Craddock, C., Zuo, X.-N., Zang, Y.-F., Milham, M., 2013a. Stan-  
741 dardizing the intrinsic brain: towards robust measurement of inter-individual  
742 variation in 1000 functional connectomes. *NeuroImage* 80, 246–262.  
743 URL <http://view.ncbi.nlm.nih.gov/pubmed/23631983>
- 744 Yan, C.-G. G., Craddock, C. C., Zuo, X.-N. N., Zang, Y.-F. F., Milham, M. P.,  
745 Oct. 2013b. Standardizing the intrinsic brain: towards robust measurement  
746 of inter-individual variation in 1000 functional connectomes. *NeuroImage* 80,  
747 246–262.  
748 URL <http://view.ncbi.nlm.nih.gov/pubmed/23631983>
- 749 Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D.,  
750 Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R.,  
751 Fischl, B., Liu, H., Buckner, R. L., Sep. 2011. The organization of the human  
752 cerebral cortex estimated by intrinsic functional connectivity. *Journal of neu-*  
753 *rophysiology* 106 (3), 1125–1165.  
754 URL <http://dx.doi.org/10.1152/jn.00338.2011>
- 755 Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J.,  
756 Breitner, J. C. S., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., Chen,  
757 A., Chen, B., Chen, J., Chen, X., Colcombe, S. J., Courtney, W., Craddock,  
758 R. C., Di Martino, A., Dong, H.-M., Fu, X., Gong, Q., Gorgolewski, K. J.,  
759 Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.-H., Huckins, J., Jiang,  
760 T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S. M., Lainhart, J. E.,  
761 Lei, X., Li, H.-J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu,



762 G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D. S., Mayer,  
763 A. R., Meindl, T., Meyerand, M. E., Nan, W., Nielsen, J. A., O'Connor,  
764 D., Paulsen, D., Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z.,  
765 Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.-X., Weng,  
766 X.-C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.-F., Zhang, L., Zhang,  
767 Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.-T., Milham,  
768 M. P., 9 Dec. 2014. An open science resource for establishing reliability and  
769 reproducibility in functional connectomics. *Sci Data* 1, 140049.

772 Submitted to Neuroimage.

773 C. Dansereau<sup>1,2</sup>, Y. Benhajali<sup>1,3</sup>, C. Risterucci<sup>4</sup>, E. Merlo Pich<sup>4</sup>, P. Orban<sup>1</sup>, D.  
774 Arnold<sup>5</sup>, P. Bellec<sup>1,2</sup>

775 <sup>1</sup>Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal,  
776 Montréal, CA

777 <sup>2</sup>Department of Computer Science and Operations Research, University of Mon-  
778 treal, Montreal, CA

779 <sup>3</sup>Département d'anthropologie, Université Montréal, Montréal, CA

780 <sup>4</sup>Clinical Imaging, pRED, F.Hoffman-La Roche, Basel, CH

781 <sup>5</sup>NeuroRx inc., Montréal, CA

782  
783 For all questions regarding the paper, please address correspondence to  
784 Pierre Bellec, CRIUGM, 4545 Queen Mary, Montreal, QC, H3W 1W5, Canada.  
785 Email: pierre.bellec (at) criugm.qc.ca.  
786

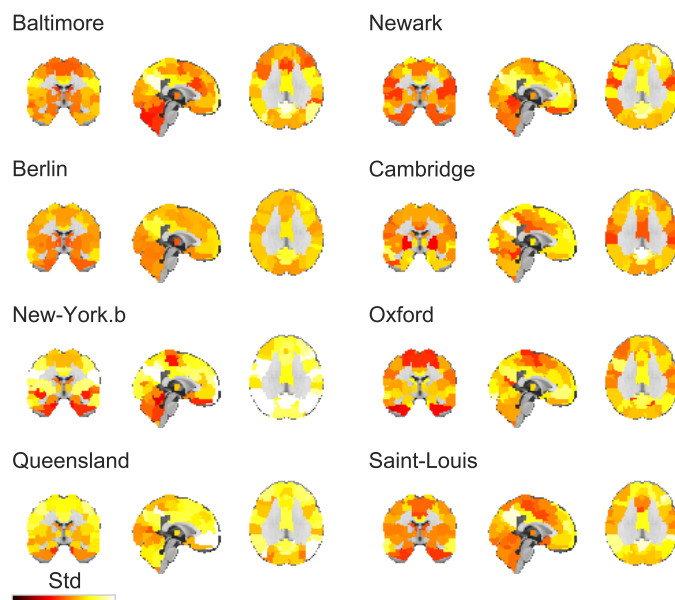


Figure S1: Overlay of the standard deviation of the DMN for each site on the MNI152 template.

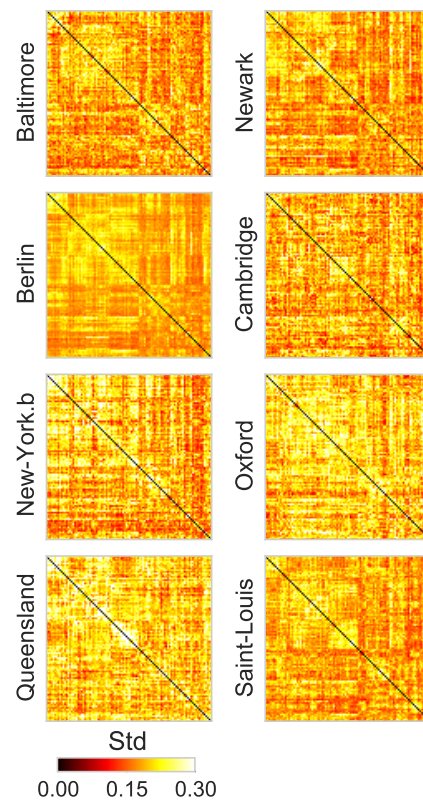


Figure S2: The standard deviation of the connectome for each site.

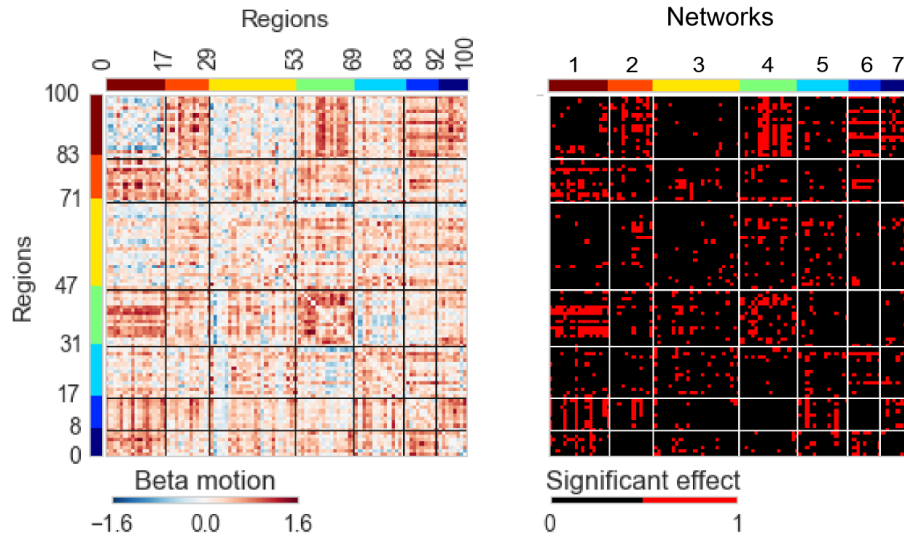


Figure S3: The figure shows average connectomes for individual sites, as well as connections with a significant motion bias.

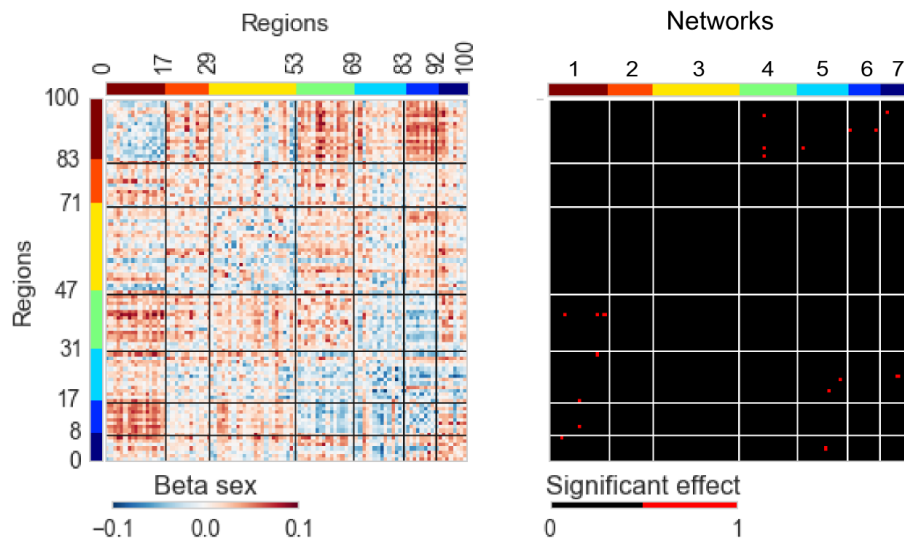


Figure S4: The figure shows average connectomes for individual sites, as well as connections with a significant sex bias.

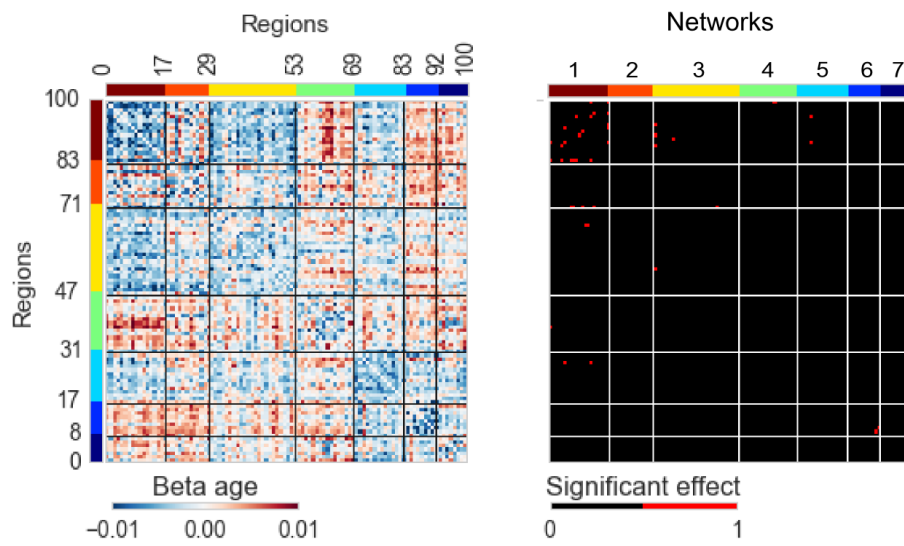


Figure S5: The figure shows average connectomes for individual sites, as well as connections with a significant age bias.