# A Comprehensive Assessment of Reproducibility of R-fMRI Metrics on the Impact of Different Strategies for Multiple Comparison Correction and Small Sample Size

Xiao Chen[1,2], Bin Lu[1,2], Chao-Gan Yan[1,2,3,4,*]

[1]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China; [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China; [3]Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; [4]Department of Child and Adolescent Psychiatry, NYU Langone Medical Center School of Medicine, New York, NY, USA

***Corresponding author:**

Chao-Gan Yan, Ph.D.

CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China

16 Lincui Road, Chaoyang District, Beijing 100101, China

Tel: +86-10-64101582

Fax: +86-10-64101582

E-mail: ycg.yan@gmail.com

**ABSTRACT**

Reproducibility is one of the key defining features of science and plays a central role in knowledge accumulation. In the field of resting-state functional magnetic resonance imaging (R-fMRI), concerns regarding the reproducibility of findings have been raised. In response, we comprehensively assessed the reproducibility of widely used R-fMRI metrics and systematically investigated the impact of different strategies correcting for multiple comparisons and for small sample sizes. We found that multiple comparison correction strategies with liberal thresholds yield higher reproducibility but can dramatically increase the family wise error rate (FWER) to unacceptable levels. We noted permutation test with Threshold-Free Cluster Enhancement (TFCE), a strict multiple comparison correction strategy, reached the best balance between FWER (under 5%) and reproducibility (e.g., 0.68 for within-subject reproducibility of amplitude of low-frequency fluctuations). Although the sex differences in R-fMRI metrics can be moderately reproduced from a scan to another scan within subjects, they are poorly reproduced in another different dataset (between-subject reproducibility < 0.3). Among the brain regions showing the most reproducible sex differences, posterior cingulate cortex demonstrated consistent lower spontaneous activity in males than in females. Defining the most reproducible brain regions in two large sample datasets as “gold standard”, we found that small sample size not only minimized power (sensitivity < 5%), but also decreased the likelihood that significant results reflect true effects. For the liberal multiple comparison correction, results were unlikely to reflect true effects (positive predictive value = 10%). Fortunately, voxels determined to be significant using permutation test with TFCE have a 71% probability of reflecting true effects. Our findings have implications for how to select

multiple comparison correction strategies and highlight the need for sufficiently large sample sizes in future R-fMRI studies.

**KEYWORDS:** reproducibility; resting-state fMRI; multiple comparison correction strategies; sample size; sensitivity; positive predictive value

## 1. INTRODUCTION

The ability to replicate an entire experiment is essential to the scientific method (Open Science Collaboration, 2015). Much of the scientific enterprise, such as providing detailed descriptions of methods and peer-reviewing manuscripts before publication, are intended to optimize agreement of results when performed by different researchers. These efforts are crucial because science cannot progress if results cannot be reproduced (Blackford, 2017). However, concerns regarding the reproducibility of biomedical and psychological research are increasingly being expressed (Open Science Collaboration, 2015; Ioannidis, 2005; Prinz et al., 2011). This is particularly relevant to the field of resting-state functional magnetic resonance imaging (R-fMRI) (Carp, 2012a; Poldrack et al., 2017), which has appeared to be a fruitful approach for basic, translational and clinical neuroscience (Biswal et al., 1995; Fox and Raichle, 2007; Fox et al., 2005). Beyond its reported sensitivity to developmental, aging and pathological processes (Hjelmervik et al., 2014; Luo et al., 2011; Tomasi and Volkow, 2012), R-fMRI is being increasingly adopted due to its relative ease of data collection and amenability to aggregation across studies and sites (Zuo et al., 2014). These advantages are balanced against high dimensionality of data, relatively small sample size of most studies and the great amount of flexibility in data analysis, all of which threaten reproducibility.

Despite the importance of reproducibility, the way to quantitatively examine to what extent the findings of one study can be reproduced by another study is still limited (Raemaekers et al., 2007). Some used intra-class correlation (ICC) (Caceres et al., 2009; Shrout and Fleiss, 1979)

to assess reproducibility, and found moderate to high ICC for most R-fMRI metrics (Cao et al., 2014; Shehzad et al., 2009; Zuo and Xing, 2014; Zuo et al., 2013; Zuo et al., 2010a). However, ICC is a general measure defined as the ratio of the between-subject variance to the total variance, it is basically a scaling of reliability instead of reproducibility (Caceres et al., 2009). ICC is less informative because most investigators interpret fMRI results on conventional *P*- or *Z*-thresholded statistical maps (Kristo et al., 2014). In the current study, we sought to propose a quantitative method to calculate reproducibility of R-fMRI metrics, which has a diversity of potential applications, such as evaluating the reproducibility of new neuroimaging metrics. We did so by comparing differences of common R-fMRI metrics between males and females and examining how the significant clusters can be reproduced in further retest or in totally different datasets (studies). Sex differences were choosing because it is a relatively objective category and can be readily investigated across a large scale of datasets. Besides, the differences between men and women's brain function have been well documented in the R-fMRI literature (Allen et al., 2011; Beltz et al., 2015; Bluhm et al., 2008; Filippi et al., 2013; Hjelmervik et al., 2014; Kilpatrick et al., 2015; Scheinost et al., 2015; Tomasi and Volkow, 2012; Xu et al., 2015).

By examining the likelihood that significant results can be reproduced in a retest of the same group of subjects or in a dataset with completely different group of subjects, we can scale both between-subject and within-subject reproducibility. However, in such a way, reproducibility is highly sensitive to the statistical threshold used to define significance. The reported reproducibility decreases as the significance threshold is enhanced (Duncan et al., 2009). However, introducing a liberal statistical threshold can dramatically increase the family wise

error rate (FWER), as shown in a recent study which systematically evaluated the FWERs of widely-used statistical methods (Eklund et al., 2016). The trade-off between reproducibility and FWER requires a comprehensive investigation into different statistical approaches for multiple comparison correction to try to reach a balance. Accordingly, the impact of statistical method, especially multiple comparison correction strategies, on reproducibility is the second focus of the present study.

Another major concern is the low statistical power of small samples, which are prevalent in the field of neuroscience. Carp reviewed over 200 fMRI studies published since 2007, and found the median sample size was 15 for one group studies and 14.75 for two group studies, resulting in unacceptable statistical power for most studies (Carp, 2012b). Another recent analysis (Poldrack et al., 2017), reviewed 1131 sample sizes in neuroimaging studies over more than 20 years. Despite the steady increase in sample size (with median sample size up to 28.5 for single-group studies and 19 per group in multi-group studies), the median study in 2015 was only sufficiently powered to detect effects greater than 0.75 SD units. Button and colleagues calculated the statistical power of neuroscience studies with data extracted from meta-analyses. They found that the median statistical power of studies in the field of neuroscience was optimistically estimated to be between ~8% and ~31% (Button et al., 2013). Moreover, the statistical findings of low power studies are unlikely reflecting a true effect (i.e., with low positive predictive value, PPV) (Button et al., 2013; Ioannidis, 2005). Although these concerns exist for a long time, the empirical reproducibility of small sample size studies, as well as their power and PPV are unknown. The attempt to establish the sensitivity and PPV is

hampered by the problem that how to define the truly positive results. Using findings that are reproducible in many datasets as the “gold standard”, it’s possible to examine the empirical quantifications of sensitivity and PPV in small sample size studies.

Here, to address the above issues, we systematically analyzed three independent datasets to quantify both the within-subject and between-subject reproducibility of R-fMRI data and investigate how multiple comparison correction strategies impact reproducibility. We also considered how statistically significant findings emerging from small sample size studies might be reproducible and reflect a true effect. Five common R-fMRI metrics, namely, the amplitude of low frequency fluctuation (ALFF) and its fractional version (fALFF), regional homogeneity (ReHo), degree centrality (DC) and voxel-mirrored homotopic connectivity (VMHC) were employed to encompass possible sex differences. We conclude by providing a recommended guideline based on this quantitative analysis to address the reproducibility challenge in the field of R-fMRI research.

## 2. MATERIALS AND METHODS

### 2.1. Participants and Imaging Protocols

We performed our analyses on publicly available imaging data from the Consortium for Reliability and Reproducibility (CORR) (Zuo et al., 2014) and the 1000 Functional Connectomes Project (FCP) (Biswal et al., 2010), which are open science resources that share data via the International Neuroimaging Data-sharing Initiative (INDI, all data are available at http://fcon_1000.projects.nitrc.org). Two independent dataset were analyzed to

evaluate within-subject reproducibility and between-subject reproducibility. A third dataset was employed to explore the implication in studies with small sample size. Although datasets differ in participant demographics and scanning parameters, the general data acquisition procedures were similar across different sites. Participants were instructed to simply rest while awake in a 3T scanner (data from 3 sites within the “FCP” dataset were scanned on 1.5T scanners), and R-fMRI data were acquired using an echo-planer imaging (EPI) sequence. A high-resolution T1-weighted anatomical image was also obtained for each participant for spatial normalization and localization. The corresponding institutional review boards of each collection center approved or provided waivers for the submission of anonymized data to INDI, which were obtained with written informed consent from each participant.

The first dataset originally included 549 subjects who underwent 2 scanning sessions (mean time range = 205 ± 161 days) available at CORR. Four hundred and twenty subjects (age 21.45±2.67, 208 females, referred as “2-session dataset” from this point on) were selected after quality control with the following exclusion criteria. To avoid the confounds of development or aging, only young adults (age between 18 and 32) were included. Subjects were excluded from analysis if their functional scans showed extreme motion, measured by mean frame-wise displacement (FD), see (Jenkinson et al., 2002)), that is, mean FD exceeding 0.2mm. Furthermore, participants with poor T1 or functional images, low quality normalization or inadequate brain coverage were also excluded. The second dataset consisted of 716 young healthy subjects (age 22.34 ± 2.92, 420 females, referred as “1-session dataset” from this point on) selected from FCP with the same inclusion criteria as

the 2-session dataset. The third dataset consisted of 30 subjects (age 24.37±2.41, 15 females, referred to as the "10-session dataset" from this point on) who underwent 10 scanning sessions (every 3 days in 1 month) from CORR. No subjects were excluded from this dataset. For further information on these three datasets including scanning protocols please refer to the CORR (http://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html) and FCP (http://fcon_1000.projects.nitrc.org/index.html) websites.

### 2.2. Preprocessing

Unless otherwise stated, all preprocessing was performed using the Data Processing Assistant for Resting-State fMRI (DPARSF, Yan and Zang, 2010, http://rfmri.org/DPARSF), which is based on Statistical Parametric Mapping (SPM, http://www.fil.ion.ucl.ac.uk/spm) and the toolbox for Data Processing & Analysis of Brain Imaging (DPABI, Yan et al., 2016, http://rfmri.org/DPABI). First, the initial 10 volumes were discarded, and slice-timing correction was performed with all volume slices were corrected for different signal acquisition time by shifting the signal measured in each slice relative to the acquisition of the slice at the mid-point of each repetition time (TR). Then, the time series of images for each subject were realigned using a six-parameter (rigid body) linear transformation with a two-pass procedure (registered to the first image and then registered to the mean of the images after the first re-alignment). After realignment, individual T1-weighted MPRAGE were co-registered to the mean functional image using a 6 degree-of-freedom linear transformation without re-sampling and then segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) (Ashburner and Friston, 2005). Finally, the transformations from individual native space to MNI

space were computed with the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool (Ashburner, 2007).

### 2.3. Nuisance Regression

To control for head motion confounds, we utilized the Friston 24-parameter model (Friston et al., 1996) to regress out head motion effects. The Friston 24-parameter model (i.e., 6 head motion parameters, 6 head motion parameters one time point before, and the 12 corresponding squared items) was chosen based on prior work that higher-order models remove head motion effects better (Satterthwaite et al., 2013; Yan et al., 2013a). Additionally, mean FD was used to address the residual effects of motion in group analyses. Mean FD is derived from Jenkinson's relative root mean square (RMS) algorithm (Jenkinson et al., 2002). As global signal regression (GSR) is still a controversial practice in the R-fMRI field, and given the recent advice that analyses with and without GSR are complementary (Murphy and Fox, 2016), we evaluated results both with and without GSR. Other sources of spurious variance (WM and CSF signals) were also removed from the data through linear regression to reduce respiratory and cardiac effects. Additionally, linear trends were included as a regressor to account for drifts in the blood oxygen level dependent (BOLD) signal. We performed temporal filtering (0.01-0.1Hz) on all time series except for ALFF and fALFF analyses.

### 2.4. A Broad Array of R-fMRI Metrics

Amplitude of Low Frequency Fluctuations (ALFF) (Zang et al., 2007) and fractional ALFF (fALFF) (Zou et al., 2008): ALFF is the sum of amplitudes within a specific frequency domain

(here, 0.01-0.1Hz) from a fast Fourier transform of a voxel's time course. fALFF is a normalized version of ALFF and represents the relative contribution of specific oscillations to the whole detectable frequency range.

Regional Homogeneity (ReHo) (Zang et al., 2004): ReHo is a rank-based Kendall's coefficient of concordance (KCC) that assesses the synchronization among a given voxel and its nearest neighbors' (here, 26 voxels) time courses.

Degree Centrality (DC) (Buckner et al., 2009; Zuo et al., 2012): DC is the number or sum of weights of significant connections for a voxel. Here, we calculated the weighted sum of positive correlations by requiring each connection's statistical significance to exceed a threshold of $r > 0.25$ (Buckner et al., 2009).

Voxel-mirrored homotopic connectivity (VMHC, Anderson et al., 2011; Zuo et al., 2010b): VMHC corresponds to the functional connectivity between any pair of symmetric inter-hemispheric voxels - that is, the Pearson's correlation coefficient between the time series of each voxel and that of its counterpart voxel at the same location in the opposite hemisphere. The resultant VMHC values were Fisher-Z transformed. For better correspondence between symmetric voxels, VMHC requires that individual functional data are further registered to a symmetric template and smoothed (4 mm FWHM). The group averaged symmetric template was created by first computing a mean normalized T1 image across participants, and then this image was averaged with its left–right mirrored version (Zuo et al., 2010b).

Before entering into further analyses, all of the metric maps were Z-standardized (subtracting the mean value for the entire brain from each voxel, and dividing by the corresponding

standard deviation) and then smoothed (4 mm FWHM), except for VMHC (which were smoothed and Fisher-Z transformed beforehand). Of note, all the R-fMRI metrics of the 3 datasets have been openly shared through the R-fMRI Maps Project (http://rfmri.org/maps), thus readers can easily replicate the current results based on these shared maps.

### 2.5. Strategies to Correct Multiple Comparisons

We first evaluated the FWER of 31 different kinds of statistical strategies (see Tables 1 and 2). Statistical maps were thresholded using eight versions of the one tailed Gaussian random field theory (GRF) (Friston et al., 1994; Nichols and Hayasaka, 2003) correction procedure, as implemented in DPABI (Yan et al., 2016). These eight thresholding approaches used uncorrected single-voxel thresholds of $P < 0.01$ ($Z > 2.33$), $P < 0.005$ ($Z > 2.58$), $P < 0.001$ ($Z > 3.09$), or $P < 0.0005$ ($Z > 3.29$), and cluster size thresholds of $P < 0.05$, or $P < 0.025$. Given that GRF correction is only performed on one tailed tests, we set $P < 0.025$ to perform two one-tailed tests, which is equivalent to two-tailed $P < 0.05$ after Bonferroni correction. Furthermore, we evaluated FWER of two versions of Monte Carlo simulation (simulated by 1000 times) based corrections (Ledberg et al., 1998), which is implemented in AFNI (AFNI 3dClusterSim, https://afni.nimh.nih.gov/afni/doc/manual/3dclust.pdf) and DPABI (DPABI AlphaSim), separately. We note the bug reported in Eklund et al. (2016) had been fixed in the software versions used in the current study. Statistical maps were also thresholded using seven versions of permutation tests (PT), as implemented in PALM (Winkler et al., 2016) and integrated into DPABI. For PALM approaches, two tailed $P < 0.05$ (compared to 1000 permutations in FWER evaluation, and 5000 permutations for the remaining analyses) was set

as the final threshold. For cluster-extend PT, voxel thresholds of two-tailed $P < 0.02$ ($Z > 2.33$), $P < 0.01$ ($Z > 2.58$), $p < 0.002$ ($Z > 3.09$) and $p < 0.001$ ($Z > 3.29$) were set. The threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009) and voxel-wise correction (VOX) with PT were also tested at two-tailed $P < 0.05$. Finally, false discovery rate (FDR) (Genovese et al., 2002) correction was also examined. Of note, after FWER evaluation, two sets of strategies (AFNI 3dClusterSim and DPABI AlphaSim) were excluded from further analyses due to their computational demands and higher FWER than GRF correction (see Results below).

### 2.6. Evaluating FWER of Different Strategies to Correct Multiple Comparisons

To calculate the FWERs of different approaches for multiple comparisons corrections, we performed permutation tests (1000 permutations in this study). In this permutation test, we first selected 106 female young subjects from the Beijing site within the 1-session dataset to make the sample homogenous. Then, 40 subjects were randomly picked from the entire set of 106 subjects and randomly assigned to two equal groups (20 for each group). Because this assignment was fully random, no significant results were supposed to emerge when these two groups' R-fMRI metrics were compared. If a significant difference was detected after multiple comparison correction, a family wise error had occurred. Thus, FWER was calculated as the proportion of false positives in all comparisons within the permutation test.

### 2.7. Assessing Reproducibility of Different Datasets

For each dataset, we employed a general linear model to examine the sex differences in

R-fMRI measures while taking the confounding effects of age, head motion (mean FD) and site (except for the 10-session dataset) into account. Sex effect was estimated by the *t* value of the regressor corresponding to sex. Then the group difference map was corrected using different multiple comparison correction approaches described above to obtain statistically significant clusters.

The Dice coefficient was used to evaluate the within-subject reproducibility. It is calculated by the following equation:

$$Dice = \frac{2 \times V_{overlap}}{V_1 + V_2}$$

Where $V_1$ and $V_2$ represents the number of supra-threshold voxels in test 1 and test 2 of the 2-session dataset, and $V_{overlap}$ stands for the number of supra-threshold voxels in both tests.

To calculate between-subject reproducibility, we selected the voxels which are significant in both sessions in the 2-session dataset, and then calculate how they overlap with the significant voxels in the 1-session dataset. We use the similar dice formula, while $V_1$ represents the number of voxels significant in both sessions in the 2-session dataset, $V_2$ represents the number of voxels significant in the 1-session dataset. And $V_{overlap}$ stands for the number of voxels that are significant in both sessions in the 2-session dataset as well as significant in the 1-session dataset.

For each multiple comparison correction strategy, we calculated within-subject reproducibility

and between-subject reproducibility. To figure out which multiple comparison correction strategy yields the best reproducibility, a non-parametric one-way repeated measures ANOVA (Friedman's test) on 5 metrics by 2 operations (with and without GSR) was conducted, and followed by post-hoc analyses corrected by Tukey's honest significant difference criterion.

Finally, we used the voxels that are significant in both sessions in the 2-session dataset as well as significant in the 1-session dataset as the "gold standard" for further evaluation (see section below). We believe these significant voxels may be able to reflect the true differences between males and females for their high reproducibility in two large sample size datasets.

### 2.8. Sensitivity and Positive Predictive Value in Small Sample Size Studies

A simple voxel counting method (Rutten et al., 2002) was applied to determine reproducibility in the 10-session dataset. We counted how many voxels were significant for a given frequency (range from 1 to 10) in all 10 sessions. Then we evaluated the sensitivity and positive predictive value (PPV) based on the "gold standard" defined above. The sensitivity of a study measures the proportion of positives that are correctly identified as such (Altman and Bland, 1994), while PPV is the probability that a positive finding reflects a true effect (Ioannidis, 2005). A recent analysis (Button et al., 2013) demonstrated that studies with small sample size not only reduce the chance of detecting a true effect, but also reduce the probability that significant findings reflect a true effect. To determine this effect of small samples, the sensitivity and PPV of the 10-session dataset were calculated:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

where TP is equal to the number of statistical significant “true positive voxels”, which are statistical significant and also reflect the true effect. As the true effect is difficult to define, we used the voxels that are significant in both sessions in the 2-session dataset as well as significant in the 1-session dataset (the “gold standard” defined above). FN represents the number of “false negative voxels”, that are statistically insignificant but reflect a true effect. And FP stands for the number of the false positive voxels that are statistically significant but do not reflect the true effect.

## 3. RESULTS

### 3.1. FWER of Different Multiple Comparison Correction Strategies

To evaluate the reproducibility of R-fMRI metrics, an appropriate statistical threshold and multiple comparison correction strategy must be defined in advance. The appropriate multiple comparison correction strategy must control the false positive rate at an acceptable level. Here, we evaluated 31 different multiple comparison correction strategies with 5 different R-fMRI metrics by 2 different operations (with and without GSR) in 106 female young adults (selected from the Beijing site of the 1-session dataset). Based on the group differences of two randomly assigned groups (20 subjects per group, permuted 1000 times), we calculated FWER for each multiple comparison correction strategy. Table 1 presents FWERs and cluster sizes of GRF and Monte Carlo Simulation based correction strategies on ALFF. Other metric’s FWERs can be found in supplementary materials (Table S1-S4). For FWERs under GRF and Monte Carlo Simulation based corrections, the liberal voxel *P* thresholds

(cluster-defining threshold) ($P < 0.01$ ($Z > 2.33$), $P < 0.005$ ($Z > 2.58$)) far exceeded nominal 5% level (Table 1, Figure 1 & Tables S1-S4). Furthermore, as most researchers are interested in two tailed effects (e.g., both patients > controls and patients < controls), if they perform one-tailed thresholding twice (i.e., each tail $P < 0.05$), then the final FWER is higher than the nominal 5% level. Only if the researcher corrects the two tests of each tail (e.g., Bonferroni correction, each tail controls at $P < 0.025$), the FWER can reach the nominal 5% level. For example, GRF was almost valid (FWER = 5.4%) under the strictest threshold (voxel wise $P < 0.0005$ and cluster wise $P < 0.025$, each tail), while FWERs of Monte Carlo Simulation based correction exceeds their nominal 5% level (Table 1 & Table S1-S4), especially in metrics with higher smoothness. For example, for ReHo maps which have relatively higher smoothness (9.4×8.7×8.4mm), FWER can reach 15.7% for AFNI 3dClusterSim (or 9.7% for DPABI AlphaSim), which is much worse than GRF (5.4% FWER) (Table 2). Given their high computational demands and their higher FWER than GRF correction, two versions of Monte Carlo Simulation based correction (AFNI 3dClusterSim and DPABI AlphaSim) were excluded from further analyses. Almost all the remaining versions of PT and FDR correction controlled the FWER at the nominal 5% level (Table 2 and Figure 1).

### 3.2. Within-Subject Reproducibility of R-fMRI Metrics under Different Multiple Comparison Correction Strategies

Some argue liberal statistical threshold can achieve better reproducibility at the cost of higher FWER. After evaluating the FWER, we systematically evaluated the within-subject reproducibility of 5 R-fMRI metrics under 15 different multiple comparison correction

strategies on the 2-session dataset (Table 3). On average, within-subject reproducibility reached 0.49 (SD: 0.14, Range: 0.11 ~ 0.75) among different R-fMRI metrics. ALFF, fALFF and ReHo have relatively high within-subject reproducibility: ALFF: 0.65 ± 0.01, fALFF: 0.60 ± 0.11, ReHo: 0.53 ± 0.03. In contrast, DC and VMHC have lower within-subject reproducibility: DC: 0.39 ± 0.05, VMHC: 0.40 ± 0.07. Interestingly, we found GSR slightly decreased within-subject reproducibility of R-fMRI metrics. For example, DC's within-subject reproducibility decreased from 0.37 to 0.11 under correction of PT with VOX.

Similar to the findings of Duncan et al. (Duncan et al., 2009) , we also found reproducibility under GRF correction with stricter cluster defining threshold were lower than those with looser threshold. To fully investigate reproducibility under different thresholds, we further performed a Friedman test on 5 metrics by 2 operations (with and without GSR) to identify the best multiple comparison correction strategy that can balance reproducibility and FWER (see Figure 2). Results showed significant differences among 15 different multiple comparison correction strategies (Friedman's chi-square = 74.45, df = 14, N = 10, $P < 0.001$). Further post-hoc analysis revealed that GRF with liberal thresholds (voxel-wise $P < 0.01$ and cluster-wise $P < 0.05$ or 0.025, each tail) and PT with TFCE achieved better reproducibility. For example, PT with TFCE has significantly higher within-subject reliability than GRF (voxel-wise threshold of $P < 0.0005$ ($Z > 3.29$) & cluster-wise thresholds of $P < 0.05$, or $P < 0.025$, each tail), PT (voxel-wise threshold of $P < 0.002$ ($Z > 3.09$) and $P < 0.001$ ($Z > 3.29$) & cluster-wise thresholds of $P < 0.05$ (two tailed)) and PT with VOX in the post-hoc analysis ($P < 0.05$, multiple comparison corrected by Tukey's honest significant difference criterion)

(Figure 2A). However, even at the cost of high FWER, GRF with loose thresholds (voxel-wise $P < 0.01$ and cluster-wise $P < 0.05$, each tail) did not show significantly higher within-subject reproducibility than PT with TFCE. Thus we conclude PT with TFCE could best balance FWER and reproducibility.

### 3.3. Between-Subject Reproducibility of R-fMRI Metrics under Different Multiple Comparison Correction Strategies

To calculate between-subject reproducibility, we selected the voxels that are significant in both sessions in the 2-session dataset, and then calculate how they overlap with the significant voxels in the 1-session dataset (Table 3). Generally, between-subject reproducibility was lower than within-subject reproducibility, achieving a mean of 0.10 (SD: 0.07, Range: 0.00 ~ 0.25). Under the multiple comparison correction of PT with TFCE, ALFF can reach a between-subject reproducibility of 0.25. None of the measures reached between-subject reproducibility higher than 0.3. This means that even voxels that could be reliably detected in 2 different sessions in the same subjects, are difficult to be reproduced in a totally different dataset. This might be due to many different factors between the two different datasets, for example, variation in ethnicity, sequence type, coil type, scanning parameters, participant instructions, head-motion restraint techniques, etc.

A Friedman's test was conducted to compare between-subject reproducibility under different multiple comparison correction strategies. Results showed significant differences among 15 different multiple comparison correction strategies (Friedman's chi-square = 86.11, df = 14, N

= 10, $P < 0.001$). PT with TFCE had significantly higher between-subject reliability than PT with VOX in the post-hoc analysis ($P < 0.05$, multiple comparison corrected by Tukey's honest significant difference criterion) (Figure 2B). Again, we found that, even at the cost of high FWER, GRF with liberal thresholds (voxel-wise $P < 0.01$ and cluster-wise $P < 0.05$, each tail) did not show significantly higher between-subject reproducibility than PT with TFCE.

### 3.4. Core Brain Regions with Reproducible Sex Differences

Sections 3.1 ~ 3.3 showed that PT with TFCE yielded moderate reproducibility while maintaining FWER under the nominal 5% level, thus outperforming the alternative multiple comparison correction strategies. This allowed us to determine the core brain regions with sex differences in R-fMRI metrics by identifying voxels that were reproduced across both sessions of the 2-session dataset and the 1-session dataset when applying PT with TFCE correction. As shown in Figure 3, significant differences between males and females were reproducibly observed for all R-fMRI metrics. Brain regions with sex differences varied across R-fMRI metrics, although they converged at the posterior cingulate cortex (PCC). PCC demonstrated lower spontaneous activity in males compared with females in all the metrics except for DC (i.e., ALFF, fALFF, ReHo and VMHC). The voxels with reproducible sex differences were considered the “gold standard” in the following analyses to calculate sensitivity and PPV using the 10-session dataset.

### 3.5. Reproducibility, Power and PPV in Small Sample Size Studies

Figure 4 shows number of significant voxels on ALFF under correction of different strategies

of multiple comparison correction in the 10-session dataset. Voxel number indicates how many voxels that are significant for a given frequency (ranged from 1 to 10) in all the 10 sessions. Results of the other metrics are provided in supplementary materials (Figures S1-S4). As shown in Figure 4, many voxels (286 voxels) could be classified as significant in at least 1 test among 10 sessions even in small sample studies (N = 30, 15 per group) when applying a liberal statistical threshold (GRF with voxel wise $P < 0.01$ and cluster wise $P < 0.05$, each tail). This increased possibility to yield significant results explains why liberal statistical thresholds are prevalent in many studies. However, if the criterion for significance requires that the findings be reproduced in many retests, the number of overlapping voxels is dramatically reduced (e.g., 49 voxels were significant in 7 of 10 retests, but no voxels were significant in all 10 retests even with the liberal statistical threshold (GRF with voxel wise $P < 0.01$ and cluster wise $P < 0.05$, each tail)). On the other hand, if more stringent statistical thresholds were used, the number of significant voxels was minimal. For example, with multiple comparison correction of PT with TFCE, only 13 voxels were significant in at least 1 test out of 10 sessions, and none were reproduced in more than 1 test. To investigate the impact of small sample design on the power (sensitivity) and PPV, we used the “gold standard” results obtained in Section 3.4 to calculate the sensitivity and PPV of all sessions in the 10-session dataset. As shown in Figure 5 (other metrics’ results were displayed in supplementary Figures S5-S12), sensitivity was below 5% in a small sample design, even at a liberal threshold. As indicated by the PPV, for the liberal threshold, although a large number of voxels were found to be significant in at least 1 of 10 sessions, they do not seem to reflect a true effect (PPV = 0.10). The PPV increased if the voxels could be repeatedly detected in

multiple sessions. For example, for GRF with voxel wise $P < 0.01$ and cluster wise $P < 0.05$, if the voxels were significant in 7 of 10 sessions, the PPV reached up to 0.45. Of note, voxels found to be significant in any of the 10 sessions under stringent multiple comparison correction (PT with TFCE) had a 71% probability of reflecting a true effect (PPV = 0.71, see Figure 5B).

## 4. DISCUSSION

A recent analysis observed that the conclusions drawn from many neuroimaging studies are probably irreproducible (Poldrack et al., 2017). Lack of reproducibility may partly due to (a) the abuse of liberal multiple comparison correction strategies and (b) the high prevalence of small sample size studies. Here, we provided a comprehensive examination of the impact of different multiple comparison correction strategies and small sample size on reproducibility across widely used R-fMRI metrics. We found that multiple comparison correction strategies with liberal thresholds could yield higher reproducibility but would dramatically increase the family wise error rate (FWER) to unacceptable levels. We noted that permutation test with TFCE, a strict multiple comparison correction strategy, reached the best balance between FWER (under 5%) and reproducibility (e.g., 0.68 within-subject reproducibility of sex differences in ALFF). Although the sex differences in R-fMRI metrics can be moderately reproduced from a scan to another scan within subjects, they are poorly reproduced in another different dataset (between-subject reproducibility < 0.3). Among the brain regions showing the most reproducible sex differences, PCC demonstrated consistent lower spontaneous activity in males compared with females. Defining the most reproducible brain

regions in 2 large sample datasets as a “gold standard”, we found that small sample size not only minimized power (sensitivity < 5%), but also decreased the likelihood that the significant results reflect a true effect. For the liberal multiple comparison correction, results are very unlikely to reflect a true effect (PPV = 10%). Fortunately, voxels found to be significant in a permutation test with TFCE had a 71% probability of reflecting a true effect. Here we discuss the implications of our findings on decision-making regarding the choice of multiple comparison correction strategies and approach towards addressing the challenge of reproducibility.

### 4.1. Selecting a Multiple Comparison Correction Strategy with Respect to FWER

Appropriate multiple comparison correction strategies must control the false positive rate at an acceptable level. Our results replicated the findings of prior work (Eklund et al., 2016), which analyzed R-fMRI data with a putative task design to compute FWER in task fMRI studies. They also performed between-group comparisons on simulated null task activation maps and calculated the FWER. They found an unacceptably high FWER for most widely used multiple comparison correction strategies. Our results provide additional evidence from group comparisons with a range of R-fMRI metrics. Our results confirmed that multiple comparison correction strategies with a liberal threshold (e.g., with voxel wise $P < 0.01$ and cluster wise $P < 0.05$) led to an unacceptably high FWER, while PT can maintain the FWER at the nominal 0.05 levels.

Beyond replicating Eklund et al.’s conclusions regarding FWER, two additional points should

be noted. First, researchers should pay close attention to whether the test is one-tailed or two-tailed. As most researchers are interested in two-tailed effects (e.g., both patients > controls and patients < controls), if they perform one-tailed thresholding twice (i.e., each tail $P < 0.05$), then the final FWER will be higher than 10% even if the voxel-level p is set to 0.0005 ($Z > 3.29$). Such researchers have to correct for the two tests at each tail, that is, researchers could perform one-tailed correction twice, with each tail voxel wise $P < 0.0005$ and cluster wise $P < 0.025$. With such a setting in GRF correction, the FWER almost reaches the nominal level of 5%. Second, we recommend against using Monte Carlo Simulation based corrections, given their high computational demands and higher FWER than GRF correction. At the strict level ($P < 0.0005$ and cluster wise $P < 0.025$), GRF is almost valid, while Monte Carlo Simulation based corrections inflated FWER, especially in metrics with higher smoothness (e.g., ReHo).

In sum, in considering FWER, 8 different multiple comparison correction strategies can be used: 1) GRF correction with strict p values (voxel wise $P < 0.0005$ and cluster wise $P < 0.025$ for each tail); 2) 4 kinds of PT with extent thresholding; 3) PT with TFCE; 4) PT with voxel-wise correction; and 5) FDR correction.

### 4.2. Selecting a Multiple Comparison Correction Strategy with Regard to Reproducibility

FWER is not the only criterion in selecting a multiple comparison correction strategy; reproducibility is even more crucial. An appropriate strategy should best balance FWER and

reproducibility. For example, GRF with liberal threshold (e.g., with voxel wise $P < 0.01$ and cluster wise $P < 0.05$) has relatively high reproducibility, but it is not usable because of its unacceptably high FWER. On the other hand, PT with voxel-wise correction can control FWER at a low level (< 5%), but results in the lowest reproducibility (either within-subject or between-subject), thus it is not an appropriate strategy to correct multiple comparisons. Fortunately, PT with TFCE provides a good balance between FWER and reproducibility. PT with TFCE can maintain the FWER under 5%, while yielding moderate reproducibility, e.g., 0.68 within-subject reproducibility for ALFF. Of note, the reproducibility of PT with TFCE is not significantly lower than for the liberal GRF threshold (e.g., with voxel wise $P < 0.01$ and cluster wise $P < 0.05$), whether for within-subject or between-subject reproducibility.

In considering both FWER and reproducibility, we recommend using PT with TFCE. As an approach for defining a cluster-like voxel-wise statistic, TFCE avoids the limitation of defining the initial cluster-forming threshold as do other common cluster-based strategy thresholding strategies (Smith and Nichols, 2009). TFCE uses the height parameter (H) and the extent parameter (E) to enhance cluster-like features in a statistical image. Although tweaking of these two parameters is possible, we found the default parameters (H = 2, E = 0.5) already perform well. Of note, PT with TFCE can be easily performed for many different kinds of statistical tests in DPABI, which integrated functions from PALM (Winkler et al., 2016).

### 4.3. Can R-fMRI Findings Be Reproducible?

Concerns regarding the reproducibility of R-fMRI findings are increasing (Poldrack et al.,

2017). Assessing reproducibility is highly sensitive to the statistical threshold used to define significance (Rombouts et al., 1998). After identifying the appropriate statistical approach (PT with TFCE), we could evaluate the reproducibility of common R-fMRI metrics. We found most R-fMRI metrics demonstrated moderate within-subject reproducibility (Table 3). Without GSR, fALFF reached the highest within-subject reproducibility (0.75), followed by ALFF (0.68) and ReHo (0.54). DC (0.48) and VMHC (0.44) had the lowest within-subject reproducibilities. Using a within-subject design, prior studies reported within-subject reproducibility of R-fMRI networks localized by either seed based analysis (Kristo et al., 2014) or independent component analysis (Meindl et al., 2010; Pinter et al., 2016), showing moderate to high reproducibility (between 0.29 and 0.76 in most regions). Our study confirmed the moderate within-subject reproducibility, while extending the within-subject design (e.g., the pattern of default mode component) to a between-subject design (the sex differences between females and males as in the current study), as the latter is more common and informative in clinical studies (Kristo et al., 2014). Interestingly, we found GSR, a controversial practice (Murphy and Fox, 2016) in the R-fMRI field, reduced reproducibility. Our results suggest the perspective of reproducibility should be taken into consideration in future studies investigating the mechanism of GSR.

Beyond within-subject reproducibility, a unique contribution of our study is the investigation of between-subject reproducibility. That is, to what extent can a finding in one dataset (usually one study) be reproduced in another dataset (another study)? We found between-subject reproducibility was much lower than within-subject reproducibility: between-subject

reproducibility of all the R-fMRI metrics was below 0.3. ALFF had the best balance between within-subject reproducibility (0.68) and between-subject reproducibility (0.25), outperforming the other R-fMRI metrics. Although fALFF reached a high within-subject reproducibility, between-subject reproducibility was poor (0.06), possibly because it is sensitive to variations in repetition time (TR) used in different datasets. It is not surprising to see such a low between-subject reproducibility, given the large differences between two different datasets, e.g., variation in ethnicity, sequence type, coil type, scanning parameters, participant instructions and head-motion restraint techniques. However, the present results question the generalizability of between-group differences reported in R-fMRI studies, and support the suggestion that future studies incorporate advanced data standardization techniques (Yan et al., 2013b) to improve between-subject reproducibility.

It is noteworthy that we found convergent sex differences in PCC across all metrics and all datasets, despite low between-subject reproducibility. Greater activity in females versus males was found in PCC, which is similar to previous studies (Allen et al., 2011; Biswal et al., 2010). As this phenomenon replicated in two sessions of the same dataset, and was reproduced in two different datasets, we believe this reflects a true sex difference that should be reproducible in future studies. PCC has been shown to be more active in females than in males in several fMRI activation experiments based on working and episodic memory (Filippi et al., 2013). It has been suggested that the PCC is associated with self-referential thoughts, emotions relating to others, remembering the past and thinking about the future (Fransson and Marrelec, 2008; Leech and Sharp, 2014; Maddock et al., 2001), thus our results are

consistent with more inward thinking and empathy in women compared to men.

### 4.4. What Can Be Done for Small Sample Size R-fMRI Studies?

A recent theoretical study (Button et al., 2013) highlighted the detrimental effect of low statistical power induced by small sample size on reproducibility. Our findings indicate that the reproducibility of small sample size (15 vs. 15) results was very low. For example, under PT with TFCE correction, significant voxels could never be replicated in any other of 10 sessions. According to the mathematical model of bias in scientific research (Button et al., 2013), studies with a small sample size not only have a reduced chance to detect true effects, but they also reduce the likelihood that a statistically significant result reflects a true effect. The current study used empirical data (R-fMRI metrics) to confirm that the power (sensitivity) of small sample size comparisons is extremely low (<5%), which is consistent with prior finding that median statistical power across 461 neuroimaging studies was 8% (Button et al., 2013). Despite the generally low PPV of small sample size studies, fortunately, if a voxel was determined to be significant in PT with TFCE, the probability that it reflects a true effect was as high as 71%. Further, using small, under powered samples is more likely to provide a positive result through selective analysis and outcome reporting, which are prevalent in R-fMRI studies across a broad range of experimental design and data analytic strategies (Carp, 2012a; Poldrack et al., 2017). Our results further indicated that, even though large amount of significant results could be found in a study with a small sample size under correction of liberal statistical thresholds, most of them were unlikely to reflect true effects, unless they could be reproduced in many retests. In sum, we recommend using a strict

multiple comparison correction strategy (e.g., PT with TFCE) in small sample size studies whenever possible.

Many suggestions have been proposed to address the challenges of reproducibility, e.g., establishing large-scale consortia to accumulate big data, sharing custom analysis code, following accepted standards for reporting methods, and encouraging replication studies (Button et al., 2013; Poldrack et al., 2017). Recently, data-sharing initiatives (e.g., grassroots such as FCP/INDI, openfMRI, fMRIDC and coordinated efforts such as ADNI, HCP, PING and UKBiobank) enable big data research models to address the reproducibility challenge. However, raw data sharing requires intensive coordinating efforts, huge manpower demand and large-amount data storing/management facilities. Furthermore, sharing raw data is mired with privacy concerns arising from the possibility of being able to identify participants from high dimensional raw data. These concerns, together with the demands of data organization and the limit of large data uploading, prevents the wider imaging community from sharing valuable brain imaging datasets to the public. The R-fMRI Maps project (http://rfmri.org/maps) was proposed to address the above concerns by only sharing the final maps of various R-fMRI indices, which only need light data storing/uploading requirements and remove the privacy concerns regarding raw data sharing. All of the R-fMRI metric maps of the current study have been made available through the R-fMRI Maps project, thus readers can easily confirm/reanalyze this data. Through the R-fMRI Maps project, we hope to build an unprecedented big data repository of brain imaging analyses across a wide variety of individuals: including different neurological and psychiatric diseases and disorders, as well as

healthy people with different traits. We hope the availability of such a big data repository will help to address the challenge of reproducibility.

## CONCLUSIONS

To our knowledge, this was the first effort to comprehensively evaluate the impact of different strategies to correct multiple comparisons as well as small sample size on the reproducibility of group differences in R-fMRI metrics. Our results revealed that PT with TFCE, a strict multiple comparison correction strategy, reached the best balance between FWER and reproducibility. We found moderate within-subject reproducibility of the R-fMRI metrics we assessed. By contrast, between-subject reproducibility was low, thus questioning the generalizability of between-group differences reported in R-fMRI studies. Finally, the present research demonstrated that findings from R-fMRI studies with small sample sizes are poorly reproducible, as well as yielding low sensitivity and PPV, which reinforces the calls for increasing sample size in future R-fMRI studies.

## ACKNOWLEDGEMENTS

The authors appreciate the editorial assistance and support of Dr. F. Xavier Castellanos. This work was supported by the National Natural Science Foundation of China (81671774 and 81630031), the Hundred Talents Program of the Chinese Academy of Sciences, and Beijing Municipal Science & Technology Commission (Z161100000216152).

## CONFLICT OF INTEREST

The authors declare no competing financial interests.

## REFERENCES

Allen, E.A., Erhardt, E.B., Damaraju, E., Gruner, W., Segall, J.M., Silva, R.F., Havlicek, M., Rachakonda, S., Fries, J., Kalyanam, R., Michael, A.M., Caprihan, A., Turner, J.A., Eichele, T., Adelsheim, S., Bryan, A.D., Bustillo, J., Clark, V.P., Feldstein Ewing, S.W., Filbey, F., Ford, C.C., Hutchison, K., Jung, R.E., Kiehl, K.A., Kodituwakku, P., Komesu, Y.M., Mayer, A.R., Pearlson, G.D., Phillips, J.P., Sadek, J.R., Stevens, M., Teuscher, U., Thoma, R.J., Calhoun, V.D., 2011. A baseline for the multivariate comparison of resting-state networks. Front Syst Neurosci 5, 2.

Altman, D.G., Bland, J.M., 1994. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ 308, 1552.

Anderson, J.S., Druzgal, T.J., Froehlich, A., DuBray, M.B., Lange, N., Alexander, A.L., Abildskov, T., Nielsen, J.A., Cariello, A.N., Cooperrider, J.R., Bigler, E.D., Lainhart, J.E., 2011. Decreased interhemispheric functional connectivity in autism. Cerebral cortex 21, 1134-1146.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95-113.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839-851.

Beltz, A.M., Berenbaum, S.A., Wilson, S.J., 2015. Sex differences in resting state brain function of cigarette smokers and links to nicotine dependence. Exp Clin Psychopharmacol 23, 247-254.

Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34, 537-541.

Biswal, B.B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.-M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.-J., Lin, C.-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.R.B., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.-J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.-C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.-F., Zhang, H.-Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc Natl Acad Sci U S A 107, 4734-4739.

Blackford, J.U., 2017. Leveraging Statistical Methods to Improve Validity and Reproducibility of Research Findings. JAMA Psychiatry 74, 119-120.

Bluhm, R.L., Osuch, E.A., Lanius, R.A., Boksman, K., Neufeld, R.W., Theberge, J., Williamson, P., 2008. Default mode network connectivity: effects of age, sex, and analytic approach. Neuroreport 19, 887-891.

Buckner, R.L., Sepulcre, J., Talukdar, T., Krienen, F.M., Liu, H., Hedden, T., Andrews-Hanna, J.R., Sperling, R.A., Johnson, K.A., 2009. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. J Neurosci 29, 1860-1873.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafo, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14, 365-376.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45, 758-768.

Cao, H., Plichta, M.M., Schafer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A., Tost, H., 2014. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage 84, 888-900.

Carp, J., 2012a. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Front Neurosci 6, 149.

Carp, J., 2012b. The secret lives of experiments: Methods reporting in the fMRI literature. Neuroimage 63, 289-300.

Collaboration, O.S., 2015. Estimating the reproducibility of psychological science. Science 349.

Duncan, K.J., Pattamadilok, C., Knierim, I., Devlin, J.T., 2009. Consistency and variability in functional localisers. Neuroimage 46, 1018-1026.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113, 7900-7905.

Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G., Rocca, M.A., 2013. The organization of intrinsic brain activity differs between genders: a resting-state fMRI study in a large cohort of young healthy subjects. Hum Brain Mapp 34, 1330-1343.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8, 700-711.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci U S A 102, 9673-9678.

Fransson, P., Marrelec, G., 2008. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. Neuroimage 42, 1178-1184.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J., Turner, R., 1996. Movement-Related effects in fMRI time-series. Magn Reson Med 35, 346-355.

Friston, K.J., Worsley, K.J., Frackowiak, R.S., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1, 210-220.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870-878.

Hjelmervik, H., Hausmann, M., Osnes, B., Westerhausen, R., Specht, K., 2014. Resting states are resting traits--an FMRI study of sex differences and menstrual cycle effects in resting state cognitive control networks. PLoS One 9, e103492.

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLOS Med 2, e124.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images.

Neuroimage 17, 825-841.

Kilpatrick, L.A., Istrin, J.J., Gupta, A., Naliboff, B.D., Tillisch, K., Labus, J.S., Mayer, E.A., 2015. Sex commonalities and differences in the relationship between resilient personality and the intrinsic connectivity of the salience and default mode networks. Biol Psychol 112, 107-115.

Kristo, G., Rutten, G.-J., Raemaekers, M., de Gelder, B., Rombouts, S.A.R.B., Ramsey, N.F., 2014. Task and task-free FMRI reproducibility comparison for motor network identification. Hum Brain Mapp 35, 340-352.

Ledberg, A., Akerman, S., Roland, P.E., 1998. Estimation of the Probabilities of 3D Clusters in Functional Brain Images. Neuroimage 8, 113-128.

Leech, R., Sharp, D.J., 2014. The role of the posterior cingulate cortex in cognition and disease. Brain 137, 12-32.

Luo, C., Li, Q., Lai, Y., Xia, Y., Qin, Y., Liao, W., Li, S., Zhou, D., Yao, D., Gong, Q., 2011. Altered functional connectivity in default mode network in absence epilepsy: a resting-state fMRI study. Hum Brain Mapp 32, 438-449.

Maddock, R.J., Garrett, A.S., Buonocore, M.H., 2001. Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. Neuroscience 104, 667-676.

Meindl, T., Teipel, S., Elmouden, R., Mueller, S., Koch, W., Dietrich, O., Coates, U., Reiser, M., Glaser, C., 2010. Test-retest reproducibility of the default-mode network in healthy individuals. Hum Brain Mapp 31, 237-246.

Murphy, K., Fox, M.D., 2016. Towards a Consensus Regarding Global Signal Regression for

Resting State Functional Connectivity MRI. Neuroimage.

Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res 12, 419-446.

Pinter, D., Beckmann, C., Koini, M., Pirker, E., Filippini, N., Pichler, A., Fuchs, S., Fazekas, F., Enzinger, C., 2016. Reproducibility of Resting State Connectivity in Patients with Stable Multiple Sclerosis. PLoS One 11, e0152158.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafo, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18, 115-126.

Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10, 712-712.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532-542.

Rutten, G.J., Ramsey, N.F., van Rijen, P.C., van Veelen, C.W., 2002. Reproducibility of fMRI-determined language lateralization in individual subjects. Brain Lang 80, 421-437.

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64, 240-256.

Scheinost, D., Finn, E.S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M., Constable,

R.T., 2015. Sex differences in normal age trajectories of functional brain networks. Hum Brain Mapp 36, 1524-1535.

Shehzad, Z., Kelly, A.M., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Lee, S.H., Margulies, D.S., Roy, A.K., Biswal, B.B., Petkova, E., Castellanos, F.X., Milham, M.P., 2009. The resting brain: unconstrained yet reliable. Cereb Cortex 19, 2209-2229.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86, 420-428.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83-98.

Tomasi, D., Volkow, N.D., 2012. Gender differences in brain functional connectivity density. Hum Brain Mapp 33, 849-860.

Winkler, A.M., Ridgway, G.R., Douaud, G., Nichols, T.E., Smith, S.M., 2016. Faster permutation inference in brain imaging. Neuroimage 141, 502-516.

Xu, C., Li, C., Wu, H., Wu, Y., Hu, S., Zhu, Y., Zhang, W., Wang, L., Zhu, S., Liu, J., Zhang, Q., Yang, J., Zhang, X., 2015. Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state FMRI study. Biomed Res Int 2015, 183074.

Yan, C., Zang, Y., 2010. DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI. Front Syst Neurosci 4, 13.

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., 2013a. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics.

Neuroimage 76, 183-201.

Yan, C.G., Craddock, R.C., Zuo, X.N., Zang, Y.F., Milham, M.P., 2013b. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. Neuroimage 80, 246-262.

Yan, C.G., Wang, X.D., Zuo, X.N., Zang, Y.F., 2016. DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. Neuroinformatics 14, 339-351.

Zang, Y., Jiang, T., Lu, Y., He, Y., Tian, L., 2004. Regional homogeneity approach to fMRI data analysis. Neuroimage 22, 394-400.

Zang, Y.F., He, Y., Zhu, C.Z., Cao, Q.J., Sui, M.Q., Liang, M., Tian, L.X., Jiang, T.Z., Wang, Y.F., 2007. Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. Brain Dev 29, 83-91.

Zou, Q.-H., Zhu, C.-Z., Yang, Y., Zuo, X.-N., Long, X.-Y., Cao, Q.-J., Wang, Y.-F., Zang, Y.-F., 2008. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. J Neurosci Methods 172, 137-141.

Zuo, X.-N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C.S., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S.J., Courtney, W., Craddock, R.C., Di Martino, A., Dong, H.-M., Fu, X., Gong, Q., Gorgolewski, K.J., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.-H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S.M., Lainhart, J.E., Lei, X., Li, H.-J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D.S., Mayer, A.R., Meindl, T., Meyerand, M.E., Nan, W., Nielsen, J.A., O'Connor, D., Paulsen, D., Prabhakaran, V.,

Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.-X., Weng, X.-C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.-F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.-T., Milham, M.P., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Scientific Data 1, 140049.

Zuo, X.-N., Xing, X.-X., 2014. Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. Neurosci Biobehav Rev 45, 100-118.

Zuo, X.-N., Xu, T., Jiang, L., Yang, Z., Cao, X.-Y., He, Y., Zang, Y.-F., Castellanos, F.X., Milham, M.P., 2013. Toward reliable characterization of functional homogeneity in the human brain: Preprocessing, scan duration, imaging resolution and computational space. Neuroimage 65, 374-386.

Zuo, X.N., Di Martino, A., Kelly, C., Shehzad, Z.E., Gee, D.G., Klein, D.F., Castellanos, F.X., Biswal, B.B., Milham, M.P., 2010a. The oscillating brain: Complex and reliable. Neuroimage 49, 1432-1445.

Zuo, X.N., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F.X., Sporns, O., Milham, M.P., 2012. Network Centrality in the Human Functional Connectome. Cereb Cortex 22, 1862-1875.

Zuo, X.N., Kelly, C., Di Martino, A., Mennes, M., Margulies, D.S., Bangaru, S., Grzadzinski, R., Evans, A.C., Zang, Y.F., Castellanos, F.X., Milham, M.P., 2010b. Growing together and growing apart: regional and sex differences in the lifespan developmental trajectories of functional homotopy. J Neurosci 30, 15034-15043.

## TABLES

Table 1. Family wise error rate and cluster size of ALFF (smoothness: 7.9×7.3×6.9) under corrections of Gaussian Random Field Theory, AFNI 3dClusterSim and DPABI AlphaSim.

| Voxel Threshold (One Tailed Twice) | Cluster Threshold | AFNI 3dClusterSim | | DPABI AlphaSim | | Gaussian Random Field | |
|---|---|---|---|---|---|---|---|
| | | Family Wise Error Rate | Cluster Size | Family Wise Error Rate | Cluster Size | Family Wise Error Rate | Cluster Size |
| $P < 0.01$ ($Z > 2.33$) | $P < 0.05$ | 41.2% | 65.2±1.3 | 50.2% | 60.2±1.7 | 38.0% | 69.3±1.1 |
| $P < 0.005$ ($Z > 2.58$) | $P < 0.05$ | 30.7% | 42.9±0.9 | 35.3% | 39.5±1.1 | 26.1% | 46.7±0.8 |
| $P < 0.001$ ($Z > 3.09$) | $P < 0.05$ | 10.8% | 19.9±0.4 | 15.3% | 18.4±0.6 | 9.9% | 21.3±0.5 |
| $P < 0.0005$ ($Z > 3.29$) | $P < 0.05$ | 12.3% | 14.2±0.4 | 13.1% | 13.9±0.5 | 10.3% | 15.8±0.4 |
| $P < 0.01$ ($Z > 2.33$) | $P < 0.025$ | 28.7% | 73.8±1.9 | 37.0% | 67.7±2.4 | 25.4% | 79.0±1.2 |
| $P < 0.005$ ($Z > 2.58$) | $P < 0.025$ | 24.9% | 47.1±1.0 | 26.8% | 44.5±1.6 | 18.2% | 53.5±0.8 |
| $P < 0.001$ ($Z > 3.09$) | $P < 0.025$ | 8.9% | 22.4±0.4 | 11.2% | 21.0±0.9 | 7.9% | 24.9±0.4 |
| $P < 0.0005$ ($Z > 3.29$) | $P < 0.025$ | 6.9% | 16.7±0.3 | 7.1% | 16.0±0.7 | 5.4% | 18.5±0.5 |

Table 2. Family wise error rate under correction of 3 versions of cluster based correction, 6 versions of Permutation Test (PT) based correction as well as False Discovery Rate (FDR) correction.

| | Voxel Threshold | Cluster Threshold | Family Wise Error Rate | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ALFF | fALFF | ReHo | DC | VMHC | ALFF with GSR | fALFF with GSR | ReHo with GSR | DC with GSR | VMHC with GSR |
| Smoothness (mm, x×y×z) | | | 7.9×7.3×6.9 | 7.3×7.4×7.2 | 9.4×8.7×8.4 | 7.9×8.0×7.8 | 6.3×6.9×6.6 | 8.0×7.3×6.8 | 7.3×7.4×7.2 | 9.2×8.6×8.2 | 8.1×8.2×8.1 | 6.1×6.6×6.4 |
| Gaussian Random Field (One Tailed) | | | 5.4% | 5.9% | 5.4% | 6.3% | 7.1% | 4.7% | 6.9% | 5.9% | 5.6% | 7.5% |
| AFNI 3dClusterSim (One Tailed) | $P < 0.0005$ (Z > 3.29) | $P < 0.025$ | 6.9% | 6.7% | 15.7% | 10.2% | 4.1% | 7.7% | 7.8% | 16.2% | 10.6% | 4.8% |
| DPABI AlphaSim (One Tailed) | | | 7.1% | 8.5% | 9.7% | 10.5% | 9.7% | 6.9% | 8.1% | 9.2% | 9.6% | 9.3% |
| PT Cluster Extent Correction (Two Tailed) | $P < 0.02$ (Z > 2.33) | $P < 0.05$ | 5.8% | 3.6% | 5.8% | 4.6% | 5.2% | 4.8% | 3.9% | 5.5% | 5.2% | 4.3% |
| | $P < 0.01$ (Z > 2.58) | $P < 0.05$ | 5.4% | 4.0% | 5.7% | 4.6% | 5.5% | 5.3% | 3.8% | 5.3% | 5.0% | 4.5% |
| | $P < 0.002$ (Z > 3.09) | $P < 0.05$ | 4.5% | 4.1% | 5.3% | 4.8% | 4.2% | 4.5% | 5.0% | 5.1% | 4.7% | 4.3% |
| | $P < 0.001$ (Z > 3.29) | $P < 0.05$ | 4.8% | 4.5% | 4.5% | 4.9% | 3.4% | 4.3% | 4.8% | 5.4% | 4.2% | 3.9% |
| PT Threshold-Free Cluster Enhancement (TFCE) | | | 4.6% | 3.9% | 5.7% | 5.0% | 4.3% | 5.3% | 4.2% | 5.5% | 4.7% | 4.8% |
| PT Voxel-Wise Correction (VOX) | | | 4.9% | 4.9% | 5.7% | 3.9% | 4.7% | 6.0% | 4.5% | 5.6% | 4.0% | 4.6% |
| FDR Correction | | | 5.0% | 5.0% | 6.0% | 4.0% | 4.9% | 6.1% | 4.6% | 5.7% | 4.0% | 4.8% |

Table 3. Within-subject reproducibility of all R-fMRI metrics with and without Global Signal Regression (GSR) under correction of Guassian Random Field (GRF), Permutation Test (PT) and False Discovery Rate (FDR) correction, calculated between the first and second sessions in the 2-session dataset.

| | | | Reproducibility (Dice Coefficient) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Voxel Threshold | Cluster Threshold | ALFF | fALFF | ReHo | DC | VMHC | ALFF with GSR | fALFF with GSR | ReHo with GSR | DC with GSR | VMHC with GSR |
| GRF (One Tailed) | $P < 0.01$ (Z > 2.33) | $P < 0.05$ | 0.67 | 0.71 | 0.58 | 0.45 | 0.49 | 0.65 | 0.67 | 0.50 | 0.40 | 0.44 |
| | $P < 0.005$ (Z > 2.58) | $P < 0.05$ | 0.67 | 0.67 | 0.58 | 0.42 | 0.47 | 0.63 | 0.63 | 0.50 | 0.35 | 0.45 |
| | $P < 0.001$ (Z > 3.09) | $P < 0.05$ | 0.64 | 0.56 | 0.49 | 0.34 | 0.42 | 0.63 | 0.53 | 0.47 | 0.27 | 0.32 |
| | $P < 0.0005$ (Z > 3.29) | $P < 0.05$ | 0.65 | 0.51 | 0.48 | 0.34 | 0.40 | 0.64 | 0.48 | 0.44 | 0.28 | 0.27 |
| | $P < 0.01$ (Z > 2.33) | $P < 0.025$ | 0.66 | 0.71 | 0.56 | 0.45 | 0.45 | 0.64 | 0.67 | 0.47 | 0.40 | 0.44 |
| | $P < 0.005$ (Z > 2.58) | $P < 0.025$ | 0.66 | 0.66 | 0.55 | 0.41 | 0.49 | 0.63 | 0.63 | 0.47 | 0.35 | 0.42 |
| | $P < 0.001$ (Z > 3.09) | $P < 0.025$ | 0.64 | 0.56 | 0.53 | 0.34 | 0.41 | 0.64 | 0.53 | 0.47 | 0.28 | 0.33 |
| | $P < 0.0005$ (Z > 3.29) | $P < 0.025$ | 0.64 | 0.51 | 0.50 | 0.35 | 0.39 | 0.65 | 0.48 | 0.43 | 0.28 | 0.24 |
| PT Cluster Extent Correction (Two Tailed) | $P < 0.02$ (Z > 2.33) | $P < 0.05$ | 0.65 | 0.70 | 0.56 | 0.45 | 0.40 | 0.62 | 0.68 | 0.45 | 0.30 | 0.40 |
| | $P < 0.01$ (Z > 2.58) | $P < 0.05$ | 0.67 | 0.66 | 0.52 | 0.32 | 0.33 | 0.60 | 0.63 | 0.46 | 0.27 | 0.32 |
| | $P < 0.002$ (Z > 3.09) | $P < 0.05$ | 0.63 | 0.55 | 0.51 | 0.36 | 0.38 | 0.63 | 0.52 | 0.47 | 0.23 | 0.32 |
| | $P < 0.001$ (Z > 3.29) | $P < 0.05$ | 0.64 | 0.51 | 0.48 | 0.37 | 0.38 | 0.64 | 0.48 | 0.44 | 0.28 | 0.26 |
| PT Threshold-Free Cluster Enhancement (TFCE) | | | 0.68 | 0.75 | 0.54 | 0.48 | 0.44 | 0.66 | 0.74 | 0.44 | 0.31 | 0.42 |
| PT Voxel-Wise Correction (VOX) | | | 0.66 | 0.34 | 0.48 | 0.37 | 0.22 | 0.65 | 0.31 | 0.38 | 0.11 | 0.14 |
| FDR Correction | | | 0.64 | 0.67 | 0.54 | 0.39 | 0.37 | 0.63 | 0.64 | 0.47 | 0.23 | 0.29 |

Table 4. Between-subject reproducibility of all R-fMRI metrics with and without Global Signal Regression (GSR) under correction of Guassian Random Field (GRF), Permutation Test (PT) and False Discovery Rate (FDR) correction, calculated using significant results in both sessions in the 2-session dataset and those significant in the 1-session dataset.

| | | | Reproducibility (Dice Coefficient) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Voxel Threshold | Cluster Threshold | ALFF | fALFF | ReHo | DC | VMHC | ALFF with GSR | fALFF with GSR | ReHo with GSR | DC with GSR | VMHC with GSR |
| GRF (One Tailed) | $P < 0.01$ (Z > 2.33) | $P < 0.05$ | 0.21 | 0.13 | 0.17 | 0.20 | 0.07 | 0.20 | 0.10 | 0.11 | 0.26 | 0.09 |
| | $P < 0.005$ (Z > 2.58) | $P < 0.05$ | 0.19 | 0.11 | 0.11 | 0.17 | 0.05 | 0.17 | 0.09 | 0.11 | 0.24 | 0.05 |
| | $P < 0.001$ (Z > 3.09) | $P < 0.05$ | 0.14 | 0.10 | 0.08 | 0.10 | 0.02 | 0.12 | 0.10 | 0.04 | 0.10 | 0.03 |
| | $P < 0.0005$ (Z > 3.29) | $P < 0.05$ | 0.12 | 0.09 | 0.07 | 0.07 | 0.02 | 0.10 | 0.11 | 0.02 | 0.08 | 0.02 |
| | $P < 0.01$ (Z > 2.33) | $P < 0.025$ | 0.21 | 0.13 | 0.15 | 0.20 | 0.07 | 0.19 | 0.10 | 0.11 | 0.28 | 0.09 |
| | $P < 0.005$ (Z > 2.58) | $P < 0.025$ | 0.19 | 0.11 | 0.11 | 0.17 | 0.05 | 0.16 | 0.09 | 0.10 | 0.24 | 0.05 |
| | $P < 0.001$ (Z > 3.09) | $P < 0.025$ | 0.14 | 0.10 | 0.08 | 0.10 | 0.02 | 0.12 | 0.10 | 0.04 | 0.11 | 0.03 |
| | $P < 0.0005$ (Z > 3.29) | $P < 0.025$ | 0.13 | 0.10 | 0.07 | 0.07 | 0.01 | 0.10 | 0.11 | 0.02 | 0.08 | 0.02 |
| PT Cluster Extent Correction (Two Tailed) | $P < 0.02$ (Z > 2.33) | $P < 0.05$ | 0.21 | 0.13 | 0.14 | 0.17 | 0.05 | 0.21 | 0.06 | 0.12 | 0.22 | 0.10 |
| | $P < 0.01$ (Z > 2.58) | $P < 0.05$ | 0.19 | 0.11 | 0.11 | 0.16 | 0.02 | 0.17 | 0.09 | 0.08 | 0.24 | 0.08 |
| | $P < 0.002$ (Z > 3.09) | $P < 0.05$ | 0.14 | 0.10 | 0.08 | 0.11 | 0.02 | 0.12 | 0.10 | 0.03 | 0.05 | 0.03 |
| | $P < 0.001$ (Z > 3.29) | $P < 0.05$ | 0.12 | 0.10 | 0.07 | 0.07 | 0.01 | 0.10 | 0.11 | 0.02 | 0.08 | 0.02 |
| PT Threshold-Free Cluster Enhancement (TFCE) | | | 0.25 | 0.06 | 0.13 | 0.20 | 0.01 | 0.25 | 0.03 | 0.09 | 0.26 | 0.02 |
| PT Voxel-Wise Correction (VOX) | | | 0.02 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.05 | 0.00 | 0.00 | 0.00 |
| FDR Correction | | | 0.15 | 0.06 | 0.11 | 0.09 | 0.02 | 0.13 | 0.04 | 0.05 | 0.08 | 0.00 |

## FIGURE LEGENDS

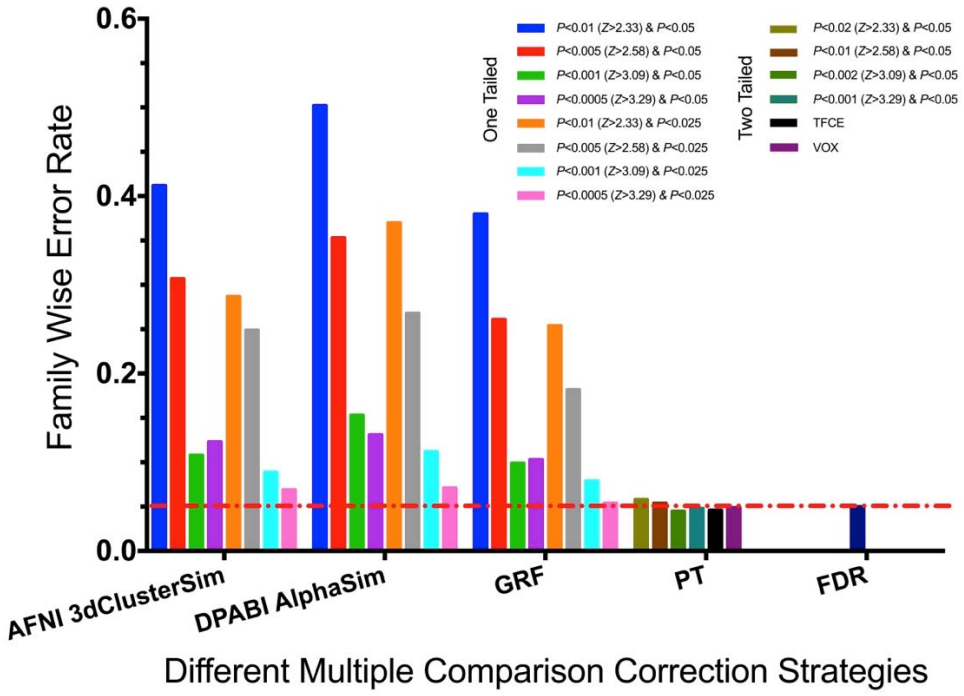


Figure 1**.** Family wise error rates of ALFF under 31 kinds of different multiple comparison correction strategies. AFNI 3dClusterSim and DPABI AlphaSim are two versions of Monte Carlo simulation based correction implemented in AFNI and DPABI, separately. GRF, PT and FDR are Guassian Random Field correction, Permutation Test and False Discovery Rate correction implemented in DPABI, separately. TFCE stands for Threshold-Free Cluster Enhancement and VOX stands for Voxel-Wise Correction which are both correction approaches accompanied with PT.

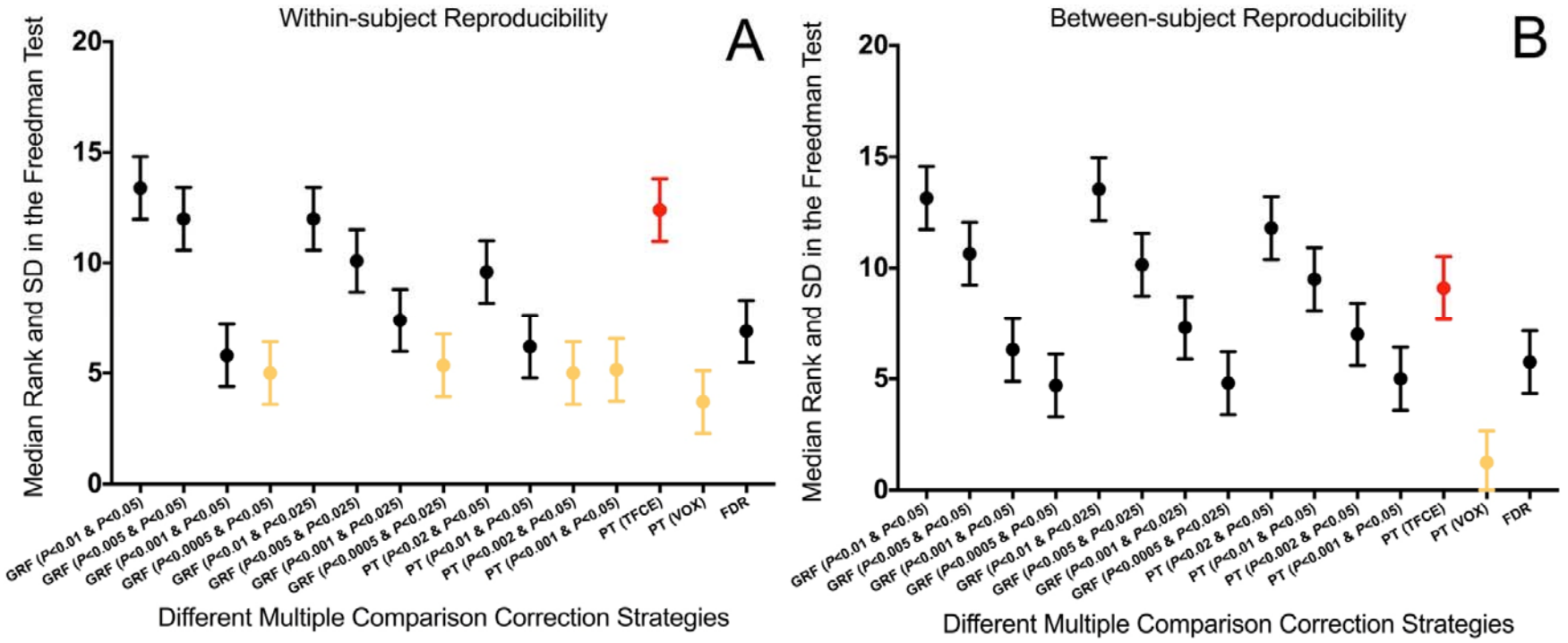


Figure 2. Results of the Freedman Test of reproducibility on 5 metrics by 2 operations (with and without GSR) among all multiple comparison correction strategies (A: within-subject reproducibility B: between-subject reproducibility). Larger median rank numbers represent the better reproducibility compared with other statistical threshold approaches. PT with TFCE is outlined with red, and those are significantly different from PT with TFCE in reproducibility are outlined with yellow (multiple comparison corrected by Tukey's honest significant difference criterion). GRF, PT and FDR stand for Guassian Random Field correction, Permutation Test and False Discovery Rate correction, separately. All versions of GRF correction are one-tailed *P* values while all versions of PT are two tailed *P* values.

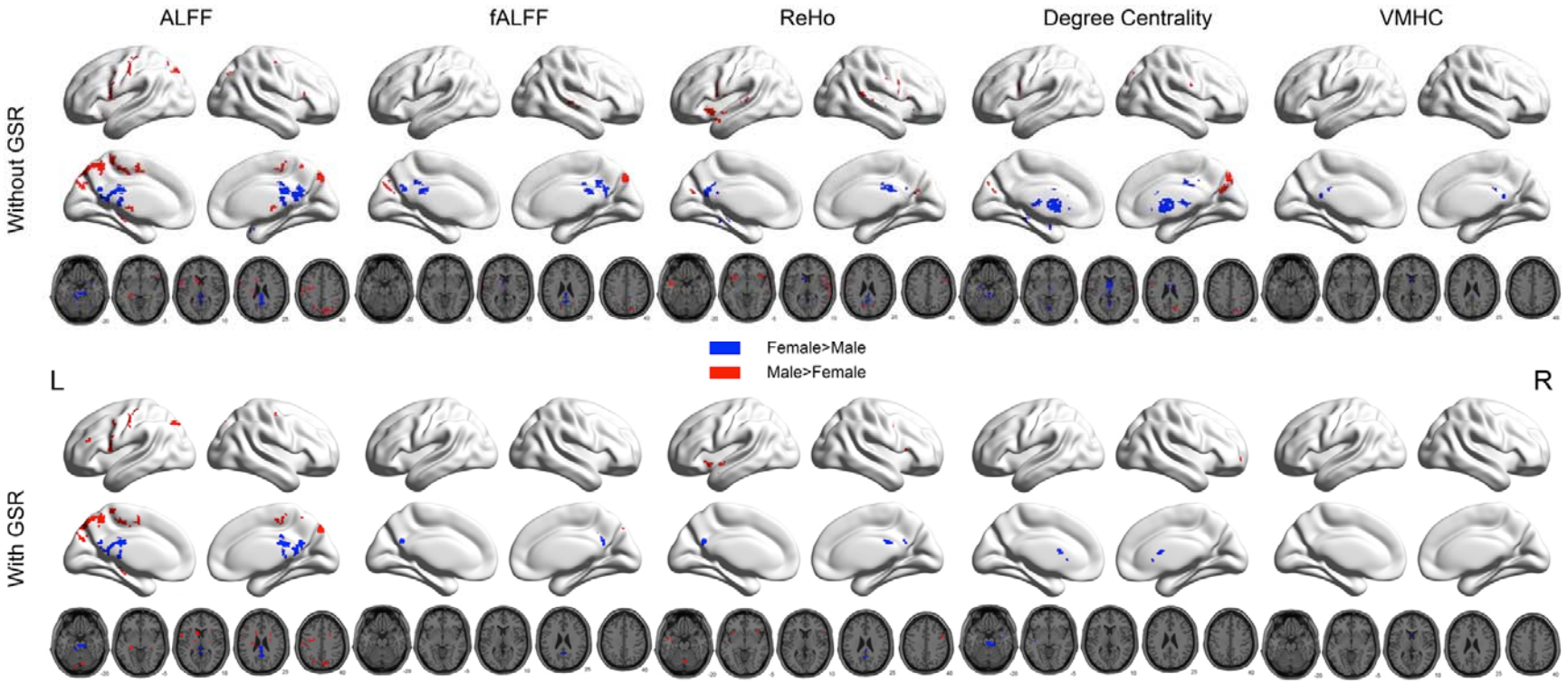


Figure 3. Sex differences those are significant in both sessions in the 2-session dataset as well as significant in the 1-session dataset ("gold standard"), under the correction of Permutation Test (PT) with Threshold-Free Cluster Enhancement (TFCE).

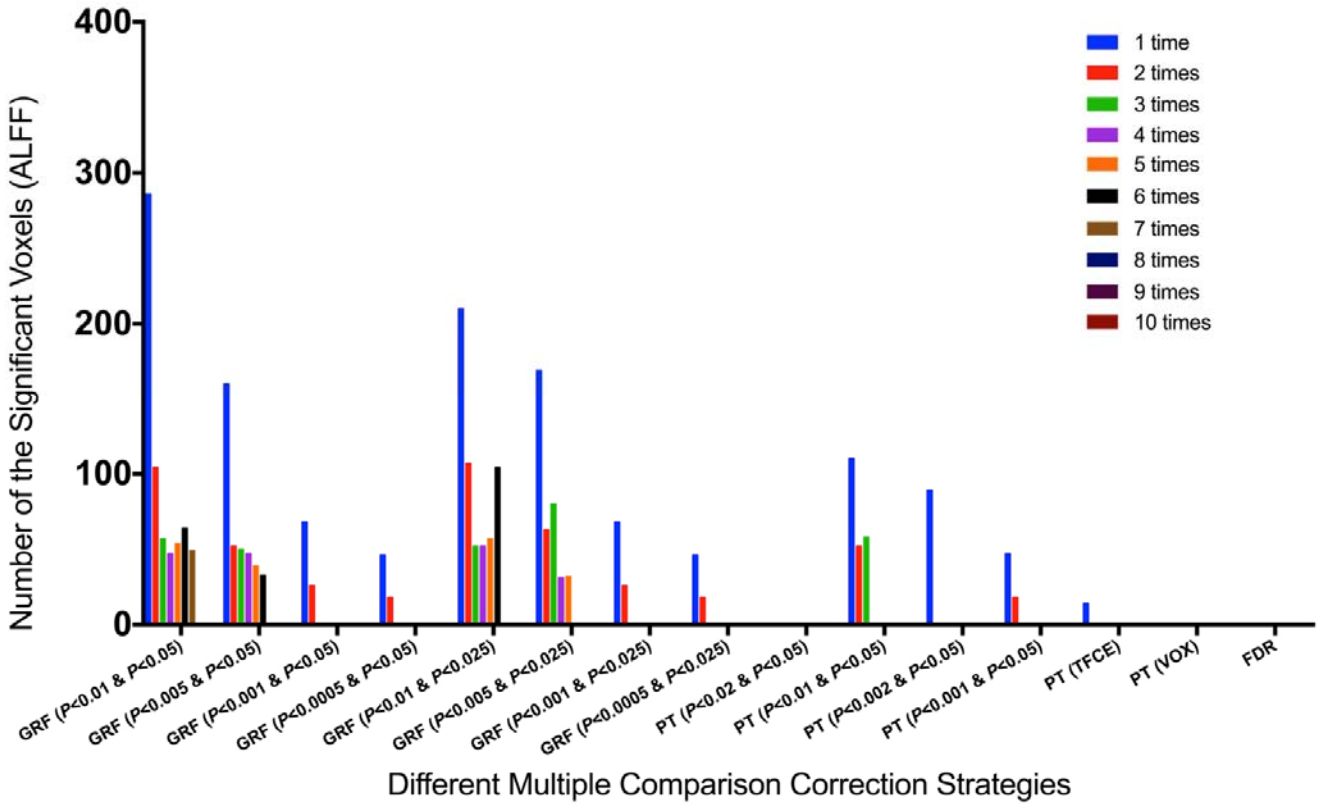


Figure 4. Number of significant voxels (on ALFF) under correction of different strategies of multiple comparison correction in the 10-session dataset. Voxel number indicates how many voxels that are significant for a given frequency (ranged from 1 to 10, indicated by different color) in all the 10 sessions. GRF, PT and FDR stand for Guassian Random Field correction, Permutation Test and False Discovery Rate correction, separately. All versions of GRF correction are one-tailed *P* values while all versions of PT are two tailed *P* values.

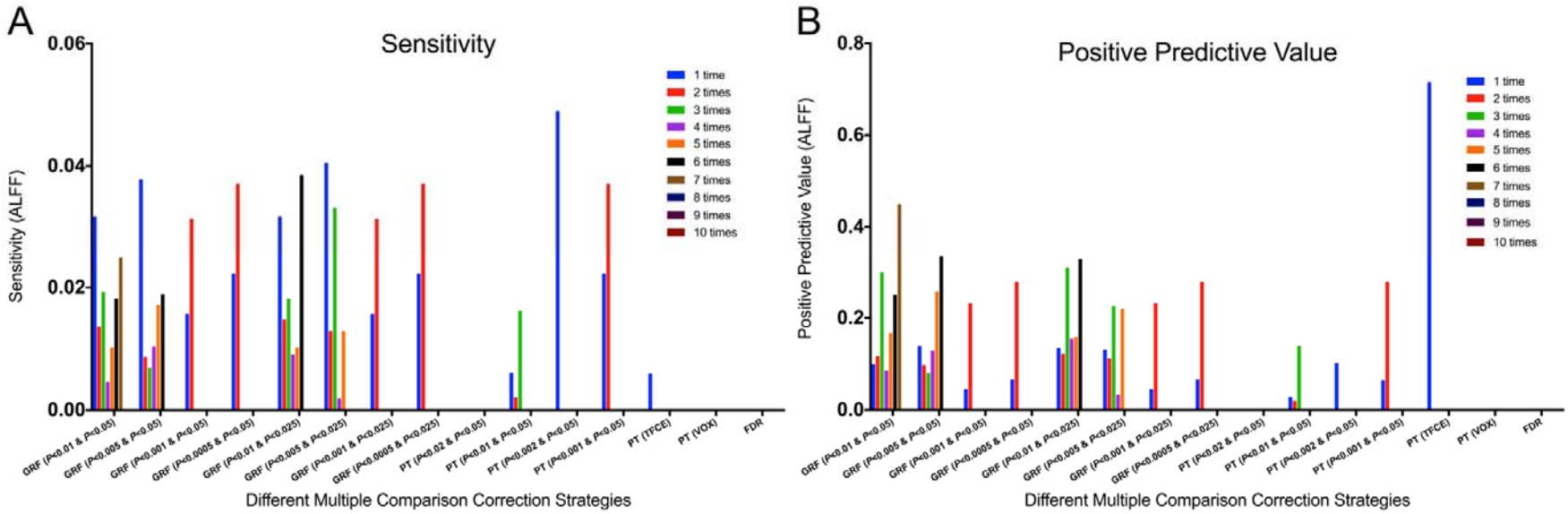


Figure 5. Sensitivity (A) and Positive Predictive Value (PPV, B) of ALFF under different multiple comparison correction strategies within the 10-session dataset (different color indicates the voxels are significant for a given frequency of sessions, ranged from 1 to 10). GRF, PT and FDR stand for Guassian Random Field correction, Permutation Test and False Discovery Rate correction, separately. All versions of GRF correction are one-tailed *P* values while all versions of PT are two tailed *P* values.