



Verification of Reproducibility of R-fMRI Metrics and Reproducible Network Underpinnings of Rumination

Xiao Chen
陈晓

chenxiao@psych.ac.cn
The R-fMRI Lab, Institute of Psychology, Chinese Academy of Sciences

1

Outline

- Verification of Reproducibility of R-fMRI Metrics
- Reproducible Network Underpinnings of Rumination

Introduction

“Reproducibility Crisis”

RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration¹

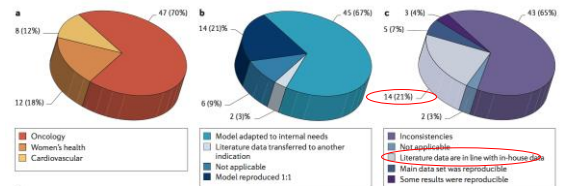
Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 502 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size. 29% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Open Science Collaboration, 2015. Science

3

Introduction

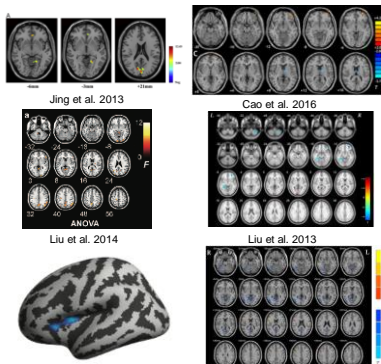
False findings may be the majority majority of published research claims



Analysis of the reproducibility of published data in 67 in-house projects

Prinz et al., 2011. Nat Rev Drug Discov 4

Introduction



5

Introduction

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Button et al., 2013. Nat Rev Neurosci

ANALYSIS

Scanning the horizon: towards transparent and reproducible neuroimaging research

Poldrack, et al., 2017. Nat Rev Neurosci

6

Reproducibility and Multiple Comparison Correction

Multiple Comparisons

Bonferroni correction

The Bonferroni correction rejects the null hypothesis for each $p_{5\alpha/m}$, thereby controlling the FWER at α .

$$\text{FWER} = P\left\{\bigcup_{i=1}^m \left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^m P\left\{p_i \leq \frac{\alpha}{m}\right\} = m \cdot \frac{\alpha}{m} = \alpha.$$



Carlo Emilio Bonferroni

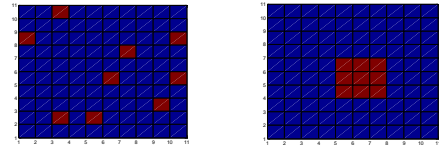
13

Reproducibility and Multiple Comparison Correction

Multiple Comparisons

Gaussian Random Field Theory Correction

Monte Carlo simulations (AlphaSim)

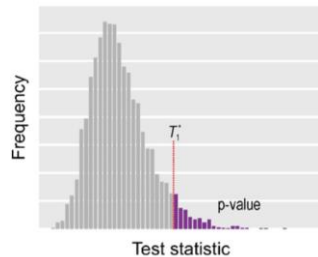


14

Reproducibility and Multiple Comparison Correction

Permutation Test

Permutations



Ronald Aylmer Fisher

Winkler et al., 2016. Neuroimage

15

Reproducibility and Multiple Comparison Correction

Threshold-Free Cluster Enhancement (TFCE)

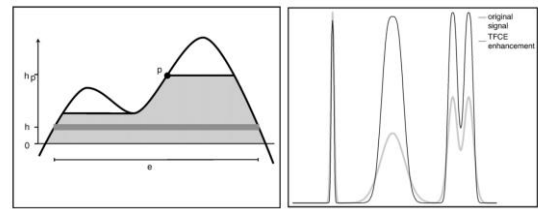


Fig. 1. Illustration of the TFCE approach. Left: the TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height h and extent s . Right: example input image and TFCE-enhanced output. The input contains a focal, high signal, a much more spatially extended, lower, signal and a pair of overlapping signals of intermediate extent and height. The TFCE output has the same maximal values for all three cases, and preserves the distinct local maxima in the third case.

Smith et al., 2009. Neuroimage

16

Multiple Comparison Correction

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund^{1,2,3,4}, Thomas E. Nichols^{4,5}, and Hans Knutsson^{4,6}

¹Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, 581 85 Linköping, Sweden; ²Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden; ³Center for Medical Image Science and Visualization, Linköping University, 581 83 Linköping, Sweden; ⁴Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and ⁵WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

Technology

15 years of brain research has been invalidated by a software bug, say Swedish scientists

Up to 70% of fMRI analyses produce at least one false positive, challenging the validity of over 40,000 studies.

Eklund et al., 2016. PNAS

17

Reproducibility and Multiple Comparison Correction

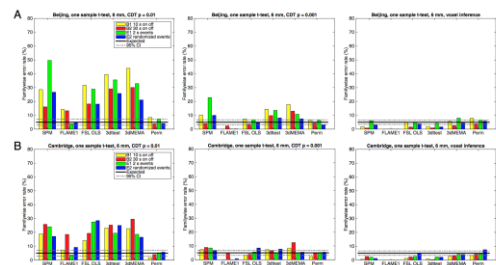


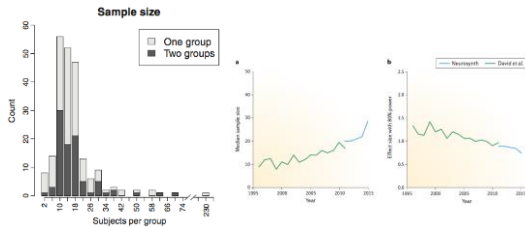
Fig. 1. Results for one-sample t test, showing estimated FWE rates for (A) Beijing and (B) Cambridge data analyzed with 5 mm of smoothing and four different activity parameters (S1, S2, L1, and L2). For SPM, FSL, AFNI, and a permutation test. These results are for a group size of 20. The estimated FWE rates are simply the number of analyses with any significant group activation divided by the number of analyses (1,000). From Left to Right: Cluster inference using a cluster-defining threshold (COT) of $P = 0.01$ and a FWE-corrected threshold of $P = 0.05$, cluster inference using a COT of $P = 0.001$ and a FWE-corrected threshold of $P = 0.05$, and voxel inference using a FWE-corrected threshold of $P = 0.05$. Note that the default COT is $P = 0.001$ in SPM and $P = 0.01$ in FSL (AFNI does not have default settings).

Eklund et al., 2016. PNAS

18

Introduction

Small samples in neuroscience



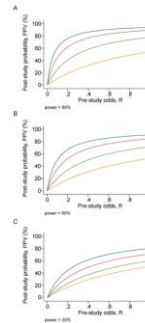
Median sample size: 15 for one group studies and 14.75 per group for two group studies (Carp, 2012)

Poldrack et al., 2017

19

Introduction

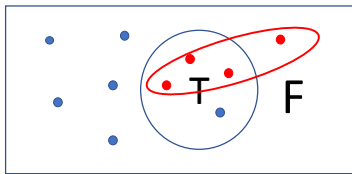
Low power studies are unlikely reflecting a true effect



John P. A. Ioannidis



Introduction



Positive Predictive Value, PPV
After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true

$$PPV = (1 - \beta)R / (R - \beta R + \alpha)$$

Research Finding	True Relationship	No	Total
Yes	$\alpha = 1 - \beta R$	αR	$\alpha R + \beta R$
No	βR	βR	$\beta R + \alpha R$
Total	$\alpha R + \beta R$	$\alpha R + \beta R$	$\alpha R + \beta R$

DOI: 10.1371/journal.pmed.1001001

Introduction

Summary

- The impact of **multiple comparison correction strategy** (considering FWER) on reproducibility (**test-retest reliability** and **replicability**)
- The impact of **sample size** on reproducibility (test-retest reliability)

Introduction

Defining reproducibility

We sought to propose a quantitative method to calculate reproducibility of R-fMRI metrics

Sex differences



Eyes open eyes closed (EOEC) differences



23

Materials and Methods

Participants and Imaging Protocols



Consortium for Reliability and Reproducibility (CORR)

Sample	Task	Modality	Sample Size	Sample Type
1	Rest	fMRI	100	Rest
2	Rest	fMRI	100	Rest
3	Rest	fMRI	100	Rest
4	Rest	fMRI	100	Rest
5	Rest	fMRI	100	Rest
6	Rest	fMRI	100	Rest
7	Rest	fMRI	100	Rest
8	Rest	fMRI	100	Rest
9	Rest	fMRI	100	Rest
10	Rest	fMRI	100	Rest

1000 Functional Connectomes Project (FCP)

24

Materials and Methods

CORR dataset

Sample size: 420 (212 M vs. 208 F)
Scanned 2 times
Inclusion criteria (from 549):
Age between 18 and 32
No extreme head motion
No poor T1 or functional images, low quality normalization or inadequate brain coverage

Beijing EOEC1 dataset

Sample size: 48
Eyes-open vs. eyes-closed
Same Inclusion criteria

1000 Functional Connectomes Project (FCP) dataset

Sample size: 716 (296 M vs. 420 F)
Same inclusion criteria

Beijing EOEC2 dataset

Sample size: 20
Eyes-open vs. eyes-closed
Same inclusion criteria

Chen, Lu, Yan*, 2018. Human Brain Mapping

Materials and Methods

Preprocessing

- 1. The first 10 volumes were discarded
- 2. Slice-timing correction
shifted to the slice at the mid-point of each TR
- 3. Realignment
six-parameter (rigid body) linear transformation
two-pass procedure
- 4. Co-registration and segment
six degree-of-freedom linear transformation without re-sampling
- 5. Transformation from native space to MNI space
Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra tool (DARTEL)



26

Materials and Methods

Nuisance Regression

A General Linear Regression Model including:
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- 1. Head motion
Friston 24-parameter model and mean FD
- 2. Global Signal Regression (GSR)
Results both with and without GSR were evaluated
- 3. Other sources of spurious variance
WM and CSF signals
- 4. Linear trends
Temporal bandpass filtering (0.01–0.1 Hz)
All time series except for ALFF and fALFF analyses

27

Materials and Methods

A Broad Array of R-fMRI Metrics

ALFF:
The mean of amplitudes within a specific frequency domain (here, 0.01–0.1Hz) from a fast Fourier transform of a voxel's time course

fALFF:
A normalized version of ALFF and represents the relative contribution of specific oscillations to the whole detectable frequency range

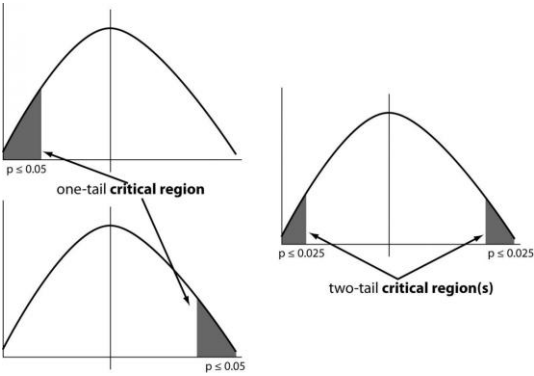
ReHo:
A rank-based Kendall's coefficient of concordance that assesses the synchronization among a given voxel and its nearest neighbors' (here, 26 voxels) time courses

Degree Centrality:
The number or sum of weights of significant connections for a voxel. The weighted sum of positive correlations with a threshold of $r > 0.25$

VMHC:
The functional connectivity between any pair of symmetric inter-hemispheric voxels

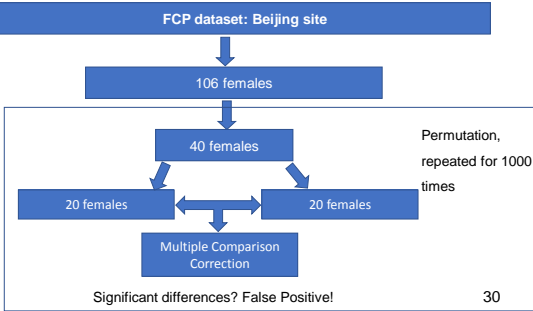
28

Materials and Methods



Materials and Methods

Evaluating FWER of Different Strategies to Correct for Multiple Comparisons



30

Results

Test-retest reliability of between-subject sex difference

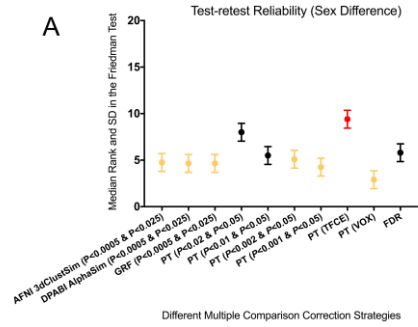
	Voxel threshold	Cluster threshold	Test-retest reliability (dice coefficient)							
			ALFF	fALFF	ReHo	DC	VMHC	ALFF with GSR	fALFF with GSR	ReHo with GSR
AFNI 3dClustSim (one-tailed) ($Z > 3.29$)	$P < 0.0005$	$P < 0.025$	0.65	0.51	0.50	0.34	0.39	0.64	0.48	0.44
DPABI AlphaSim (one-tailed)			0.65	0.51	0.49	0.34	0.39	0.64	0.48	0.45
GRF (one-tailed)			0.64	0.51	0.50	0.35	0.39	0.65	0.48	0.43
PT cluster extent correction (one-tailed) ($Z > 3.39$)	$P < 0.002$	$P < 0.05$	0.65	0.70	0.56	0.45	0.40	0.62	0.68	0.45
PT cluster extent correction (two-tailed) ($Z > 2.58$)	$P < 0.01$	$P < 0.05$	0.67	0.66	0.52	0.32	0.33	0.60	0.63	0.46
PT TCCE ($Z > 3.09$)	$P < 0.002$	$P < 0.05$	0.63	0.55	0.51	0.36	0.38	0.63	0.52	0.47
PT VOX ($Z > 3.29$)	$P < 0.001$	$P < 0.05$	0.64	0.51	0.48	0.37	0.38	0.64	0.48	0.44
PT TCCE			0.68	0.75	0.54	0.48	0.44	0.66	0.74	0.44
PT VOX			0.66	0.34	0.48	0.37	0.22	0.65	0.31	0.38
FDR correction			0.64	0.67	0.54	0.39	0.37	0.63	0.64	0.47

◆ Moderate test-retest reliability

◆ ALFF, fALFF, ReHo are better than DC and VMHC

37

Test-retest Reliability



Chen, Lu, Yan*, 2018. Human Brain Mapping

212 M vs. 208 F × 2 times

38

Results

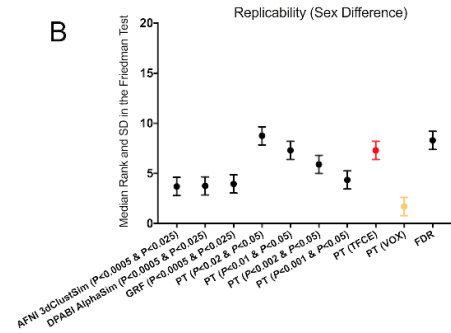
Replicability of between-subject sex difference

	Voxel threshold	Cluster threshold	Replicability (dice coefficient)							
			ALFF	fALFF	ReHo	DC	VMHC	ALFF with GSR	fALFF with GSR	ReHo with GSR
AFNI 3dClustSim (one-tailed) ($Z > 3.29$)	$P < 0.0005$	$P < 0.025$	0.12	0.10	0.07	0.07	0.01	0.10	0.11	0.02
DPABI AlphaSim (one-tailed)			0.13	0.09	0.07	0.07	0.02	0.10	0.11	0.02
GRF (one-tailed)			0.13	0.10	0.07	0.07	0.01	0.10	0.11	0.02
PT cluster extent correction (one-tailed) ($Z > 3.39$)	$P < 0.002$	$P < 0.05$	0.21	0.13	0.14	0.17	0.05	0.21	0.06	0.12
PT cluster extent correction (two-tailed) ($Z > 2.58$)	$P < 0.01$	$P < 0.05$	0.19	0.11	0.11	0.16	0.02	0.17	0.09	0.08
PT TCCE ($Z > 3.09$)	$P < 0.002$	$P < 0.05$	0.14	0.10	0.08	0.11	0.02	0.12	0.10	0.03
PT VOX ($Z > 3.29$)	$P < 0.001$	$P < 0.05$	0.12	0.10	0.07	0.07	0.01	0.10	0.11	0.02
PT TCCE			0.25	0.06	0.13	0.20	0.01	0.25	0.03	0.09
PT VOX			0.02	0.00	0.01	0.00	0.00	0.01	0.05	0.00
FDR correction			0.15	0.06	0.11	0.09	0.02	0.13	0.04	0.05

◆ Poor replicability

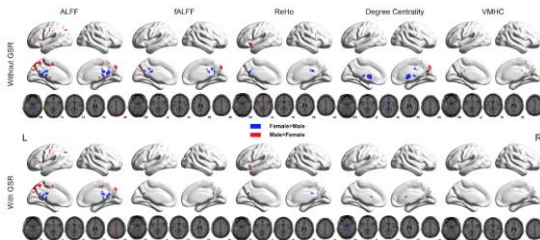
39

Results



40

Results



◆ Female's PCC demonstrate more spontaneous activity than male

41

Results

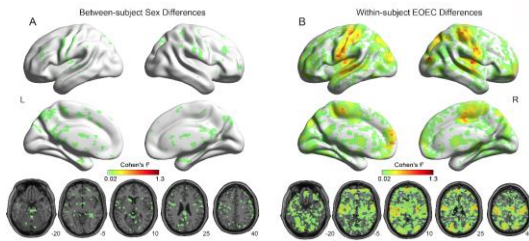
Replicability of within-subject EOEC difference

	Voxel threshold	Cluster threshold	Replicability (dice coefficient)							
			ALFF	fALFF	ReHo	DC	VMHC	ALFF with GSR	fALFF with GSR	ReHo with GSR
AFNI 3dClustSim (one-tailed) ($Z > 3.29$)	$P < 0.0005$	$P < 0.025$	0.15	0.11	0.26	0.03	0.10	0.14	0.11	0.31
DPABI AlphaSim (one-tailed)			0.15	0.11	0.26	0.03	0.10	0.14	0.11	0.31
GRF (one-tailed)			0.15	0.11	0.27	0.04	0.10	0.14	0.11	0.30
PT cluster extent correction (one-tailed) ($Z > 3.39$)	$P < 0.002$	$P < 0.05$	0.46	0.27	0.44	0.24	0.21	0.41	0.30	0.49
PT cluster extent correction (two-tailed) ($Z > 2.58$)	$P < 0.01$	$P < 0.05$	0.39	0.24	0.40	0.20	0.16	0.35	0.21	0.48
PT TCCE ($Z > 3.09$)	$P < 0.002$	$P < 0.05$	0.22	0.16	0.32	0.06	0.14	0.19	0.16	0.35
PT VOX ($Z > 3.29$)	$P < 0.001$	$P < 0.05$	0.15	0.11	0.27	0.04	0.10	0.14	0.11	0.30
PT TCCE			0.49	0.31	0.45	0.29	0.20	0.46	0.32	0.47
PT VOX			0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
FDR Correction			0.09	0.00	0.29	0.03	0.08	0.12	0.00	0.34

◆ Higher than between-subject sex difference but still not moderate

42

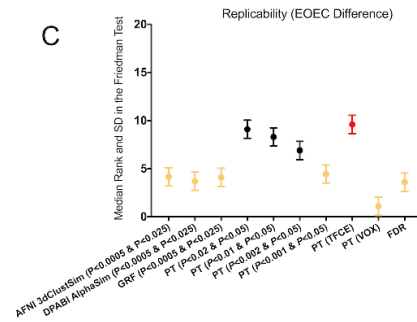
Results



Within-subject design has larger effect size

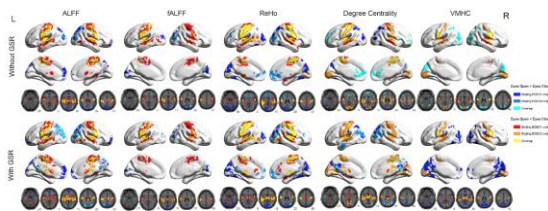
43

Results



44

Results

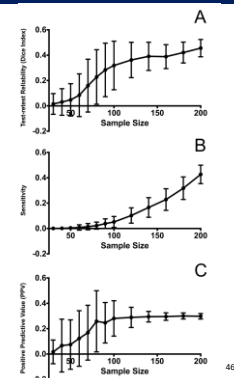


- ◆ Eyes open > Eyes closed in bilateral occipital cortices
- ◆ Eyes open < Eyes closed in bilateral pre- and post-central gyrus

45

Sample Size Matters

Randomly draw k subjects from the "SWU 4" site in the CORR dataset, which has two sessions of 116 males and 105 females



Chen, Lu, Yan*, 2018. Human Brain Mapping

46

Discussion

Main findings:

- ◆ Liberal correction strategies yield unacceptable high FWERs
- ◆ PT with TFCE reach the best balance between FWER and reproducibility
- ◆ Between-subject design has moderate test-retest reliability but poor replicability
- ◆ Within-subject design has better replicability but still not moderate
- ◆ Larger sample size increases reproducibility, sensitivity as well as PPV

47

Discussion

What correction strategy can be used?

According to FWER...

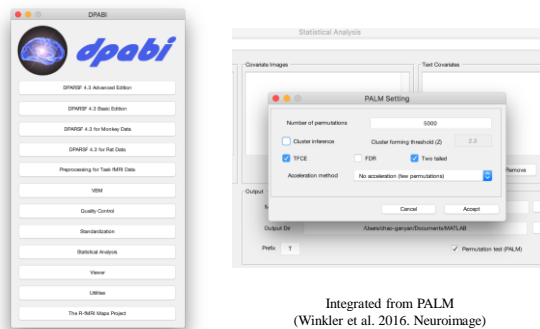
- ◆ GRF correction with strict p values (voxel wise $P < 0.0005$ and cluster wise $P < 0.025$ for each tail)
- ◆ Four kinds of PT with extent thresholding
- ◆ PT with TFCE
- ◆ PT with VOX
- ◆ FDR correction

According to reproducibility...

Strict strategies cannot achieve moderate reproducibility, except PT with TFCE

48

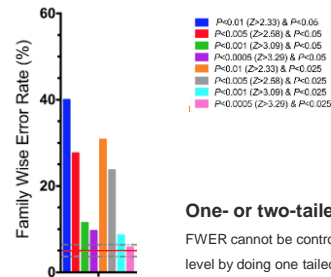
Permutation Test with TFCE



Yan* et al., 2016. Neuroinformatics
ESI Top 1% highly cited (>60 times)

49

Discussion



One- or two-tailed?

FWER cannot be controlled to the nominal level by doing one tailed correction twice

50

Discussion

Sample size (k)	Test-retest reliability (dice index)	Sensitivity	PPV
30	0.02 ± 0.08	0.001 ± 0.004	0.02 ± 0.09
40	0.03 ± 0.11	0.001 ± 0.01	0.07 ± 0.21
50	0.05 ± 0.13	0.004 ± 0.01	0.07 ± 0.19
60	0.08 ± 0.17	0.01 ± 0.02	0.12 ± 0.22
70	0.16 ± 0.21	0.01 ± 0.02	0.17 ± 0.22
80	0.23 ± 0.22	0.02 ± 0.03	0.26 ± 0.26
90	0.28 ± 0.21	0.04 ± 0.04	0.25 ± 0.16
100	0.32 ± 0.19	0.05 ± 0.04	0.28 ± 0.14
120	0.36 ± 0.14	0.10 ± 0.06	0.29 ± 0.08
140	0.39 ± 0.11	0.17 ± 0.08	0.29 ± 0.04
160	0.39 ± 0.09	0.23 ± 0.09	0.30 ± 0.03
180	0.42 ± 0.08	0.32 ± 0.09	0.30 ± 0.02
200	0.46 ± 0.07	0.43 ± 0.07	0.30 ± 0.02

Results from a sample size <80 (40 per group) should be considered preliminary, given their low reliability (< 0.23), sensitivity (< 0.02) and PPV (< 0.26)

51

Discussion



All statistical maps have been shared through the R-fMRI Maps project (<http://rfmri.org/maps>)

Key source code have been shared through (https://github.com/Chaogan-Yan/PaperScripts/tree/master/Chen_2017_HBM)

Thus our findings could be easily reproduced by any researchers

52

Outline

- Verification of Reproducibility of R-fMRI Metrics
- Reproducible Network
- Underpinnings of Rumination

Rumination

Rumination

Repetitive thinking about negative personal concerns and/ or about the implications, causes, and meanings of a negative mood

Example:

What do I do to deserve this?

Why these happen to me?

Features

- Self perpetuate
- Recycled
- Long-lasting



Susan Nolen-Hoeksema (1959 – 2013)

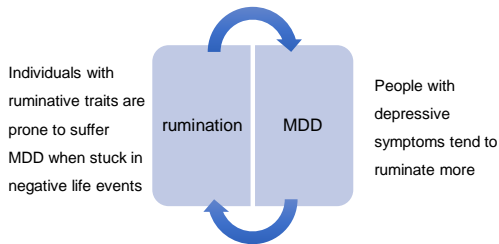


Nolen-Hoeksema et al., 2008. Perspect Psychol Sci

54

Rumination

Rumination and MDD



- Rumination is not only a defining feature, but also a risk factor for MDD

Koster et al., 2011. Clinical Psychol Rev

55

Self-Generated Thoughts



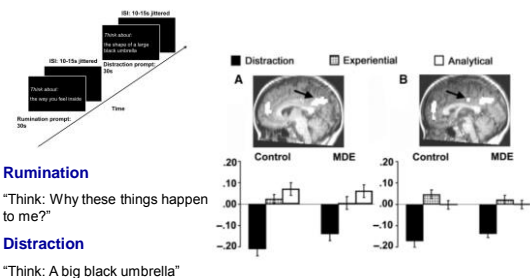
- Resting-state is a complex state
- Focusing on a specific mental state?

Andrews-Hanna et al., 2014. N.Y.Acad.Sci.

56

Literature Review

Rumination and Task-fMRI: DMN

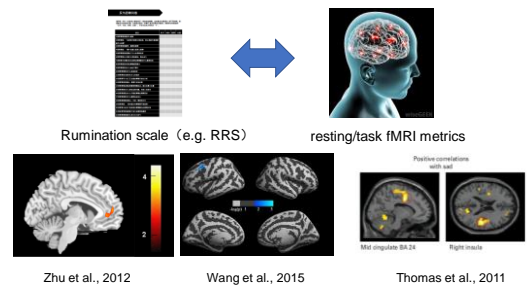


Johnson et al., 2009. SCAN

57

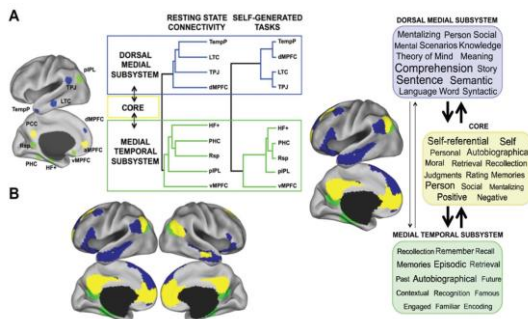
Literature Review

Correlation studies on trait rumination: DMN/CEN/SN



58

Subsystems of DMN

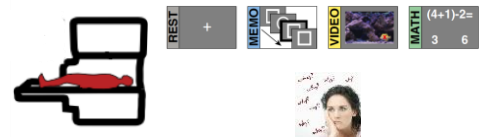


Andrews-Hanna, et al., 2014. Annals of the New York Academy of Sciences

59

Rumination State

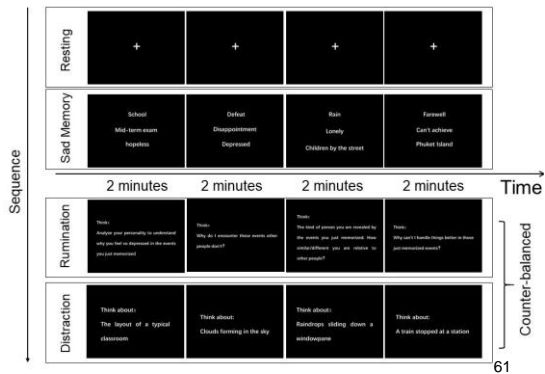
"Rumination State"



- A subject-driven, relatively long period of mental state
- Continuous and dynamic thinking style following the instructions

60

Rumination State Task

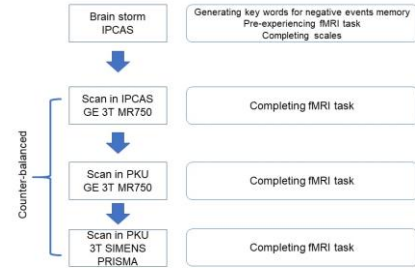


61

Rumination State Task

Materials and Methods

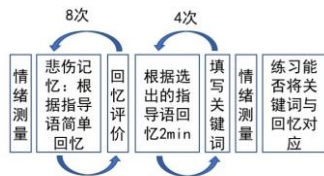
Subjects: Healthy adults (N = 41)



62

Rumination State Task

Brain Storm



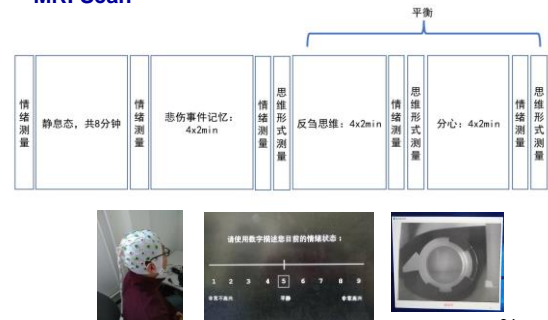
Scale



63

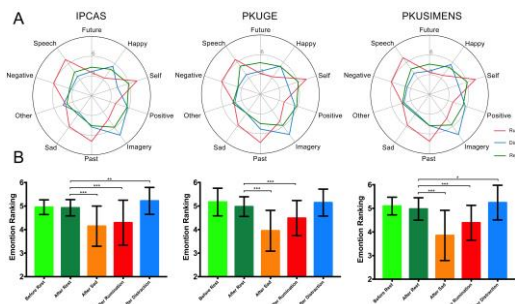
Rumination State Task

MRI Scan



64

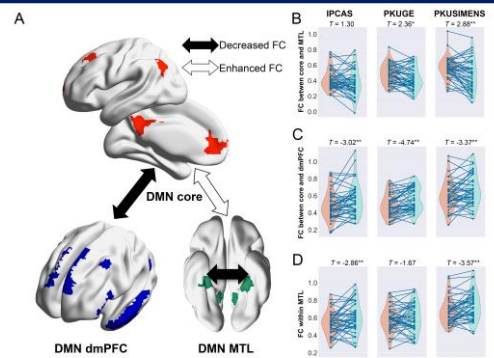
Rumination State



Chen et al., In prep.

65

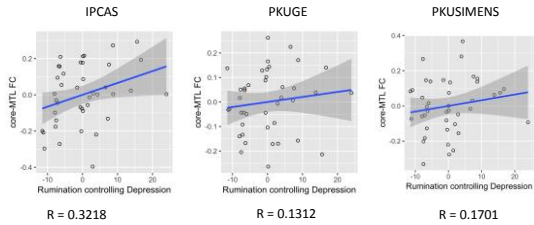
Rumination State



Chen et al., In prep.

66

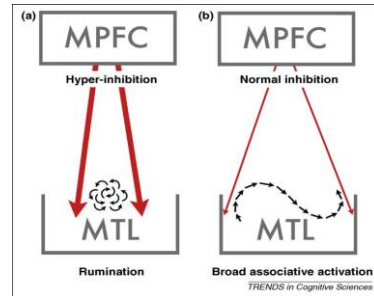
Rumination State



Rum-dis的core和MTL之间的功能连接差 和 rumination得分正相关

67

Discussion

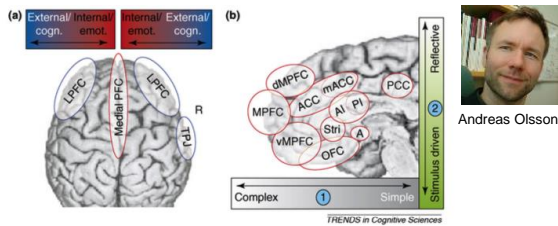


Moshe Bar

Bar, 2009. Trends in cognitive sciences

68

Discussion



- Ventral MPFC: Emotional "hot" psychological process
- Dorsal MPFC: Cognitive "cold" psychological process

Olsson and Ochsner, 2007. Trends in cognitive sciences

69

Future Work

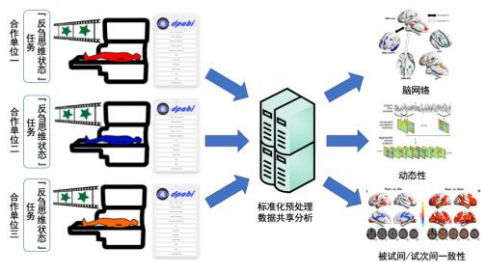
REST-meta-MDD Project



70

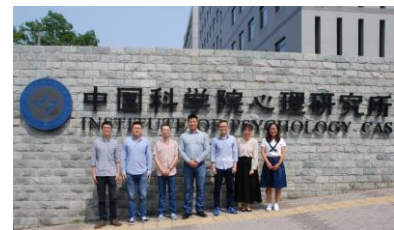
Future Work

Multi-sited rumination state research based on REST-meta-MDD



71

Acknowledgements



Funding

- National Natural Science Foundation of China
- National Key R&D Program of China
- Chinese Academy of Sciences

NYU Child Study Center
F. Xavier Castellanos

72



Thank you for your attention!