Replicability of structure-phenotype associations and its influencing factors

Shahrzad Kharabian Masouleh[1], Marcel Falkiewicz[1],Felix Hoffstaedter[1], Simon B. Eickhoff[1,2*], Sarah Genon[1*]

[1]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany
[2]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
* equally contributed
s.kharabian@fz-juelich.de@fz-juelich.de

Introduction

- Many lines of evidence suggest that inter-individual differences in behavior could be predicted using structure of the brain as assessed for example by voxel-based morphometry (VBM)[1].
- Within "reproducibility crisis" in biomedical research, studies questioned replicability of several previously reported structure-brain-behavior associations (SBB)[2].

Here we aim to define:

- Empirical rate of SBB replicability, over broad range of behavioral measures, among healthy individuals.
- Influence of sample size on spatial variability of significant findings.
- SBB-replicability in a clinical cohort

Methods

Participants:

eNKI[3]: 466 healthy adults (67% female, 18-85years).
ADNI: 371 participants with subjective memory complaints or diagnosed MCI or AD (46% female, 55-91years). See *www.adni-info.org*.

Behavioral data:

eNKI: Age and BMI for validation, standard neuropsychological tests (attention, executive functioning, verbal memory, Intelligence test), anxiety and personality questionnaires.
ADNI: Immediate-recall score of verbal learning task.

Structural data: T1-weighted scans analyzed using VBM (CAT12; Normalized with Dartel algorithm), modulated (only non-linearly), smoothed (8 mm FWHM).

Statistical analysis:

Replicability of whole brain exploratory SBB:
Association between each behavioral score and grey matter volume (GMV) is assessed by fitting a linear model at each voxel, controlling for confounding effects of age, sex, education (+ site and diagnosis, for ADNI) using general linear model in randomise[4] with 1000 permutations. Inference is made at cluster-level, using TFCE ($p < 0.05$, and extent threshold of 100 voxels).

- This procedure is applied on 100 randomly generated subsamples, of same size (e.g. 50% of the original cohort) and binary maps of significant clusters are aggregated to identify rate of spatial overlap of significant findings for each behavioral score.

Replicability of ROI-based SBB:
For every subsample, an independent matched-sample is generated from the main cohort. Partial-correlation of behavioral score and average GMV in the significant clusters are compared between the original and matched-subsamples. Replicated effects are defined based on three criteria:

- Same **direction** in the original and replication sample.
- Same **direction** + **Significant** ($p<0.05$)
- **Bayes factor**[2] (BF_{10}) ≥ 3

(H0: absence of SBB; H1: presence of SBB in the same direction as original effect.)

Results

Exploratory whole-brain SBB in healthy adults

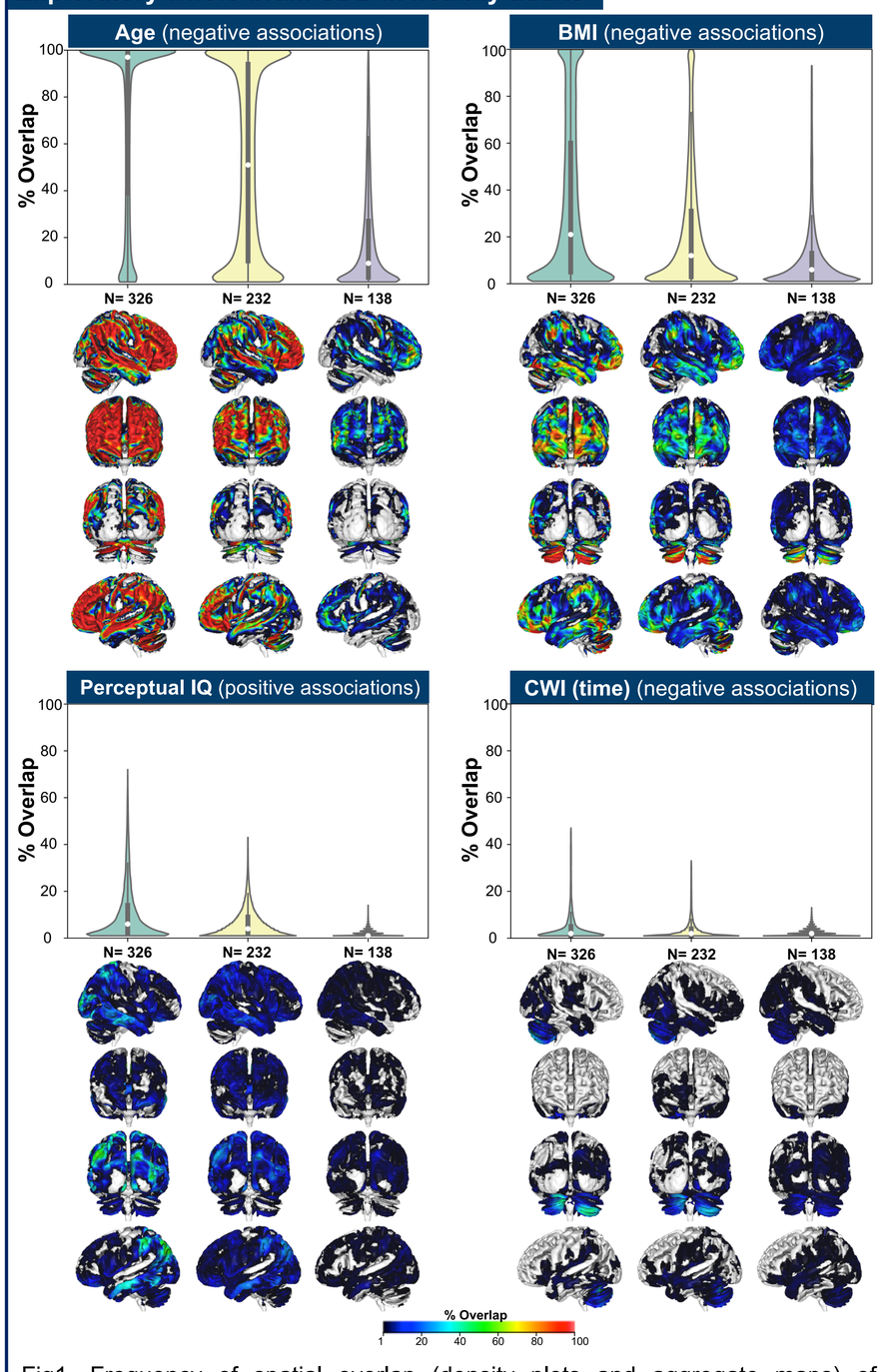

Fig1. Frequency of spatial overlap (density plots and aggregate maps) of significant findings from exploratory analysis over 100 random subsamples, calculated for three different sample sizes (X-axis). Here perceptual IQ and interference time in color-word-interference (CWI) task are shown as the top two behavioral scores with the highest frequency of overlapping results for all tested sample sizes.

ROI-based SBB in healthy adults

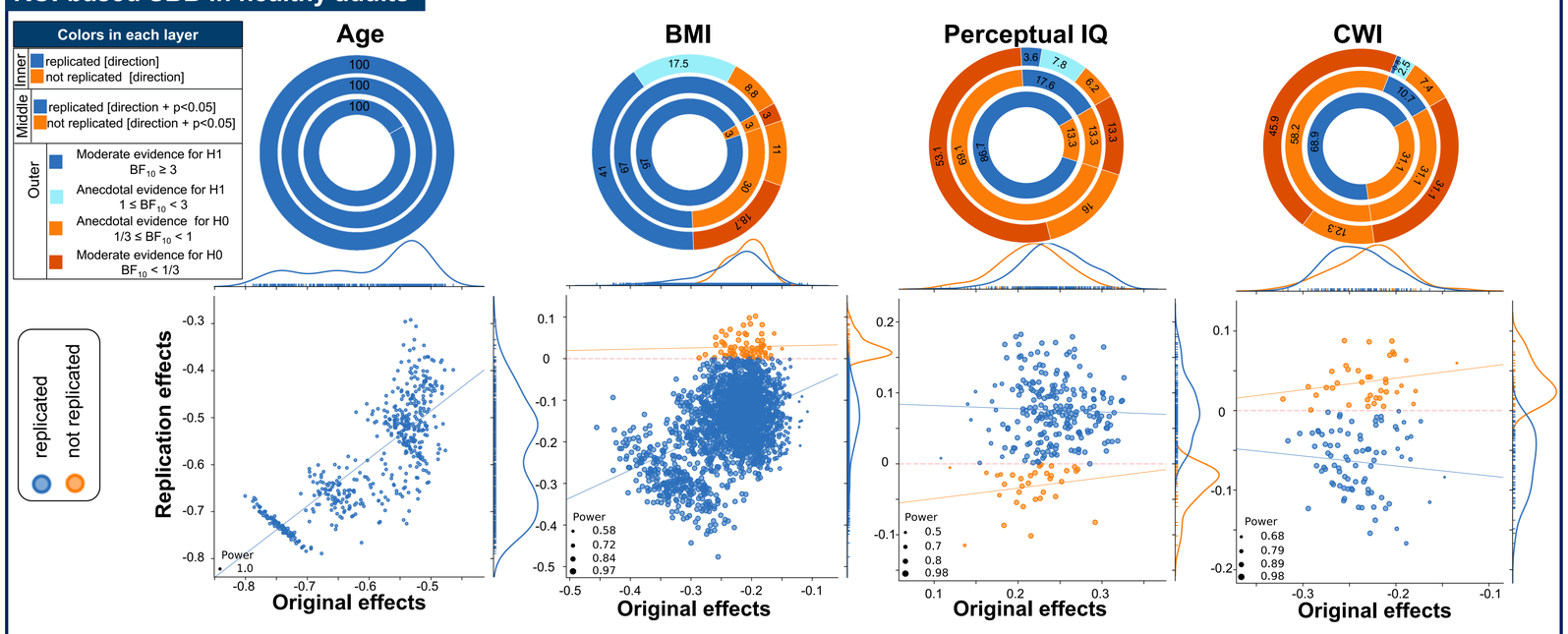

Fig2. Upper row: Summary of ROI-based SBB-replication (% of ROIs) *using three different criteria: Inner layer: "**direction**". Middle layer: **direction** + **Significant**. Outer layer:* BF_{10}. Lower row: Original versus replication effects sizes for all ROIs from 100 splits*; (replication defined using "**direction**" only) and size of each point is proportional to the power of replication. Original and replication samples have equal size(n = 232).*

Healthy adults:

- significant associations between behavior and GMV are highly unlikely.
- Exploratory SBB in similar samples often do **not** overlap.
- Sample size influences rate of spatial overlap.
- <20% of SBB-effects are "significantly replicated" in matched samples.

ADNI patients:

- High spatial overlap of associations of "recall" and GMV over 100 samples, specifically in hippocampus and medial temporal lobes.
- >70% of effects are "significantly replicated" in independent samples.
- Effect sizes of replicated-ROIs are positively correlated in the original and replication sample.

Structural correlates of "immediate-recall" among ADNI patients

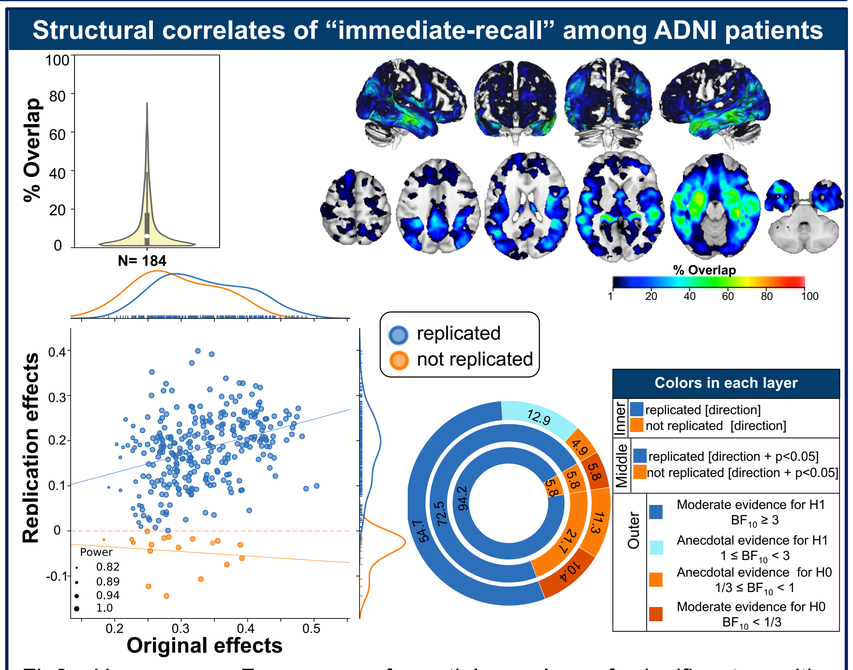

Fig3. Upper row: Frequency of spatial overlap of significant positive association between immediate-recall and GMV from exploratory analysis over 100 random subsamples. Lower row: Original versus replication effects sizes for all ROIs from 100 splits; Original and replication samples have equal size(n = 184) and are matched for age, sex and site.

Conclusions

- For most of the tested behavioral measures, we did not find any significant association in more than 90% of exploratory analyses. These results are alarming, considering the **publication bias**[5].
- Correlations of GMV with noisy behavioral measurements in small samples are frequently reported in the literature. Here we show that due to high variability of spatial location of significant exploratory findings, **harking**[6] (hypothesizing after the results are known) in such context can result in serious misleading conclusions about the true neurobiological associations[7].
- ROI-based analysis: Low rate of "significant" replication; Lack of clear association between original and replication effect sizes; Over-estimation of effects size derived from ***exploratory analysis in small samples***. Use of such effect sizes to design future studies will result in underpowered replications, more variable findings and waste of resources[8].
- Finally we observe higher replicability of structure-phenotype association for age and BMI in healthy cohort, as well as immediate-recall performance among ADNI patients. This suggests with more stably assessed measurements, where within subject variance is lower than between subject variation, structure-phenotype correlations in large samples are more reliable.

References: [1] Kanai R., Rees G., 2011. [2] Boekel, et, al., 2015. [3] Nooner et al., 2012. [4] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Randomise [5] Dwan et. al., 2008. [6] Kerr et. al., 1998. [7] Cremers et. al., 2017. [8] Button et. al., 2013.

Acknowledgments: This study was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/1-1, EI 816/11-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 785907 (HBP SGA2).