NeuroImage 212 (2020) 116601

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The empirical replicability of task-based fMRI as a function of sample size

Han Bossier [a,*], Sanne P. Roels [a], Ruth Seurinck [a], Tobias Banaschewski [b], Gareth J. Barker [c], Arun L.W. Bokde [d], Erin Burke Quinlan [e], Sylvane Desrivières [e], Herta Flor [f,g], Antoine Grigis [h], Hugh Garavan [i], Penny Gowland [j], Andreas Heinz [k], Bernd Ittermann [l], Jean-Luc Martinot [m], Eric Artiges [n], Frauke Nees [b,f], Dimitri Papadopoulos Orfanos [h], Luise Poustka [o], Juliane H. Fröhner Dipl-Psych [p], Michael N. Smolka [p], Henrik Walter [k], Robert Whelan [q], Gunter Schumann [e], Beatrijs Moerkerke [a], IMAGEN Consortium

[a] *Department of Data Analysis, Ghent University, Ghent, Belgium*
[b] *Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, 68159, Mannheim, Germany*
[c] *Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom*
[d] *Discipline of Psychiatry, School of Medicine and Trinity College Institute of Neuroscience, Trinity College Dublin, Ireland*
[e] *Medical Research Council - Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom*
[f] *Department of Cognitive and Clinical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, Mannheim, Germany*
[g] *Department of Psychology, School of Social Sciences, University of Mannheim, 68131, Mannheim, Germany*
[h] *NeuroSpin, CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France*
[i] *Departments of Psychiatry and Psychology, University of Vermont, 05405, Burlington, VT, USA*
[j] *Sir Peter Mansfield Imaging Centre School of Physics and Astronomy, University of Nottingham, University Park, Nottingham, United Kingdom*
[k] *Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charitéplatz 1, Berlin, Germany*
[l] *Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany*
[m] *Institut National de la Santé et de la Recherche Médicale, INSERM Unit 1000 "Neuroimaging & Psychiatry", University Paris Sud, University Paris Descartes, Sorbonne Paris Cité; and Maison de Solenn, Paris, France*
[n] *Institut National de la Santé et de la Recherche Médicale, INSERM Unit 1000 "Neuroimaging & Psychiatry", University Paris Sud, University Paris Descartes, Sorbonne Paris Cité; and Psychiatry Department 91G16, Orsay Hospital, France*
[o] *Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Centre Göttingen, von-Siebold-Str. 5, 37075, Göttingen, Germany*
[p] *Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany*
[q] *School of Psychology and Global Brain Health Institute, Trinity College Dublin, Ireland*

ARTICLE INFO

Keywords:
Task-based fMRI
Replicability
Reproducibility
Reliability
Stability
Coherence

ABSTRACT

Replicating results (i.e. obtaining consistent results using a new independent dataset) is an essential part of good science. As replicability has consequences for theories derived from empirical studies, it is of utmost importance to better understand the underlying mechanisms influencing it. A popular tool for non-invasive neuroimaging studies is functional magnetic resonance imaging (fMRI). While the effect of underpowered studies is well documented, the empirical assessment of the interplay between sample size and replicability of results for task-based fMRI studies remains limited. In this work, we extend existing work on this assessment in two ways. Firstly, we use a large database of 1400 subjects performing four types of tasks from the IMAGEN project to subsample a series of independent samples of increasing size. Secondly, replicability is evaluated using a multi-dimensional framework consisting of 3 different measures: (un)conditional test-retest reliability, coherence and stability. We demonstrate not only a positive effect of sample size, but also a trade-off between spatial resolution and replicability. When replicability is assessed voxelwise or when observing small areas of activation, a larger sample size than typically used in fMRI is required to replicate results. On the other hand, when focussing on clusters of voxels, we observe a higher replicability. In addition, we observe variability in the size of clusters of activation between experimental paradigms or contrasts of parameter estimates within these.

* Corresponding author.
E-mail address: Han.Bossier@Ugent.be (H. Bossier).

https://doi.org/10.1016/j.neuroimage.2020.116601
Received 26 November 2018; Received in revised form 25 January 2020; Accepted 1 February 2020
Available online 7 February 2020
1053-8119/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, challenges related to replicability and reproducibility of scientific findings have been heavily debated in the literature (see for example Baker, 2016; Begley and Ioannidis, 2015; Collins and Tabak, 2014; Munafò et al., 2017). In this paper, we focus on replicability of results from an fMRI data-analysis. While several different definitions exist (Plesser, 2018), we use the definition of replicability given by Patil et al. (2016). A successful replication of an experiment entails obtaining consistent results when repeating the exact same experimental procedure and data analysis plan, using a new (independent) dataset obtained by a different experimenter. The topic of replicating scientific results has been addressed in several fields such as psychology (Open Science Collaboration, 2015), biomedical sciences (Baker, 2015) or neurosciences (Poldrack et al., 2017). In general, there are many factors which can lead to scientific results that cannot be replicated (Munafò et al., 2017). Examples include poor study designs, low power due to small sample sizes (Ioannidis, 2005), questionable research practices (John et al., 2012) and, in the neurosciences, a low signal-to-noise ratio (Button et al., 2013; Poldrack et al., 2017) combined with substantial variability in the analysis pipelines and the reporting of results (Carp, 2012). The inability to replicate experimental results has major consequences for theories derived from these results.

Increased awareness of these problems has led to new initiatives that have been put forward in neuroscience literature. These include pre-registration (see e.g. Munafò et al., 2017) and incentives to share raw data instead of results summarized by (thresholded) statistical maps or coordinates of peaks of activation (Pernet and Poline, 2015; Poline et al., 2012). In case of pre-registration, authors are requested to register the study design, proposed methods and data analysis before data collection takes place. The manuscript is then reviewed and either rejected or accepted in principle (for example see Cortex[1] which accepts registered reports). This reduces analytical flexibility and prohibits questionable ad hoc research practices in the phase of data-analysis. Sharing data not only promotes open science but it also creates the opportunity for other researchers to use previous data as pilot data in an a priori power analysis (Durnez et al., 2016; Mumford and Nichols, 2008) or in a meta-analysis (Costafreda, 2009; Wager et al., 2007).

Although the importance of replicability is widely recognized, insight into the effect of (low) statistical power on the replicability of fMRI results is limited (Mumford, 2012; Thirion et al., 2007; Turner et al., 2018). The relation between underpowered studies and the prevalence of false positive effects is theoretically well documented (see for example Ioannidis, 2005). However, an empirical assessment of the relation between sample size and replicability in neuroimaging studies remains challenging. We identify three main causes. First, an extensive number of individual subjects are needed to effectively study the effect of sample size. Previous studies relied on subsampling from databases with a relatively limited number of subjects. For instance, in Thirion et al. (2007) and Pajula and Tohka (2016) the sample size (N) was equal to 81 and 130 subjects respectively. More recently, large collaborations have been set up and have gathered large amount of data. Examples include the Human Connectome Project (Van Essen et al., 2013) with $N \approx 1200$, the UK Biobank (Sudlow et al., 2015) with $N \approx 500.000$and the IMAGEN consortium (Schumann et al., 2010) with $N \approx 2000$. Second, an empirical assessment of the relation between replicability and sample size based on a single database is not conclusive. Therefore, independent empirical assessments using different databases and different investigators are necessary. Finally, lack of agreement on an exact definition for replicability (Peng, 2011; Pernet and Poline, 2015; Plesser, 2018) complicates conceptualization on how to measure replicability of fMRI data.

In this paper, we aim to extend the empirical assessment of task-based fMRI replicability. To this end, we aim not only to replicate results obtained by Thirion et al. (2007) and Turner et al. (2018) who studied replicability of fMRI results and its relation to sample size (amongst other predictors) but also extend the sample sizes under consideration using a subsampling method. This approach consists of sampling at random independent sets of subjects from the entire database and then incrementally increase the sample size. These groups are subsequently compared to measure replicability of fMRI results. Thirion et al. (2007) collected data on 81 subjects and observed reasonable replicability for a simple left-right button press contrast. In their work, replicability was conceptualized twofold. The first measurement used was voxelwise coherence/concordance (Genovese et al., 1997; Liou et al., 2003, 2006) between several replications. Coherence is defined as the agreement over independent replications in classifying voxels as either active or inactive. Their second measure was the average distance between statistically significant clusters of voxels. Turner et al. (2018) on the other hand used approximately 500 subjects from the Human Connectome Project utilizing various contrasts. They studied test-retest reliability in a replication context. A study is reliable if the deviance between the outcome of the study and a replication is small. The outcome can be a test-statistic such as a t-value (i.e. unconditional reliability) or the result of thresholding for statistical significance (i.e. conditional reliability). Among other measurements, Turner et al. (2018) used the Pearson correlation coefficient between the test-statistics of independent replications to measure unconditional reliability and a Jaccard index to measure conditional reliability. The latter is a measurement of spatial overlap between thresholded images. Interestingly, the authors observed only a low to modest degree of reliability. In this paper, we study how generalizable the set of findings regarding coherence and (un)conditional test-retest reliability are. To do so, we use a different data set from the IMAGEN project (with $N = 1400$). As this database contains more subjects than those previously used, we can assess replicability in larger sample sizes.

Second, we further extend the empirical assessment by studying an additional measure of replicability (i.e. stability) for task-based fMRI. This is needed as results of an fMRI data-analysis can be summarized by features such as peaks or clusters of activation. Furthermore, Roels et al. (2015) demonstrate the need to extend methodological research with measurements regarding the variability of the main characteristics of these features (see also Qiu et al., 2006 for a similar argumentation in gene selection). Main characteristics could be the size or the number of the selected clusters of voxels. Methods where the results show high variability regarding these characteristics are less stable indicating a lower replicability. For example, a specific cluster could be selected in different replications though its size may vary substantial. To assess stability, we focus on clusters of voxels as the main feature of interest.

To summarize, we conceptualize replicability by measuring (un)conditional test-retest reliability, coherence and stability between independent replications created through a subsampling approach, using a large database.

2. Methods

To study the interplay between sample size and replicability, we repeatedly subsample subjects from a large database to create sets of independent groups with a given sample size. Results of the analysis of these groups are then compared to assess replicability.

In this section, we first describe the database and the various single-subject pre-processing steps that are performed to obtain the data from which to subsample. We use six contrasts based on the following experimental paradigms: perform a cognitive task, watch angry faces, participate in a monetary incentive delay task and in a Stop signal task. We then focus on the subsampling scheme which is used to create independent pairs of group analyses. Next, we discuss these analyses and finally describe our measurements to quantify the replicability of fMRI results. In order of discussion, these are *test-retest reliability*, *coherence* and *stability*. Within these, we distinguish between two categories of

[1] https://www.elsevier.com/editors-update/story/peer-review/cortexs-registered-reports.

measurements. The first one is not conditional on any particular threshold chosen for statistical significance. We use the term "unconditional" here to stress that our measurement does not depend on a chosen criterion for thresholding. The second is a conditional assessment, during which thresholded images are used to measure replicability. In this case, results potentially depend on the criteria chosen to threshold the images. Within the latter, we also focus on both voxels and clusters.

The code of the sampling procedure, analysis and figures of this paper together with the processed results (as R objects) to reproduce figures are available at: https://github.com/NeuroStat/replicability_fmri. Note that the database used in this paper is not publicly available.

2.1. Data and pre-processing

In this section, we first discuss the description of the tasks and then the pre-processing. The pre-processed data (i.e. contrast of parameter estimates and their variances) are provided by the IMAGEN consortium (Schumann et al., 2010). This is a European multi-center project to investigate the association between reinforcement-related behaviour in adolescents and the development of frequent psychiatric disorders. For this paper, we used fMRI data of adolescents aged between 13 and 15 years, acquired across several research centres on 3 T scanners from multiple manufactures. The data are stored centrally at the Neurospin[2] center (Paris).

Pre-processing and single subject statistical analyses were performed by the consortium using SPM8[3] (Statistical Parametric Mapping: Wellcome Department of Cognitive Neurology, London, UK) for the cognitive task and SPM12[4] for the other tasks.

2.1.1. Cognitive task

For the first contrast, we use the same data and hence identical scanning protocol and pre-processing steps as those described in Bossier et al. (2018). For each participant, a total of 160 volumes were acquired. The scanning session involved a global cognitive assessment. In this assessment, participants had to perform a series of cognitive tasks. The total series contained up to 10 type of tasks. We restrict our analysis to two tasks: reading sentences in silence (LANGUAGE) and solving math subtractions (MATH). In the latter, single digits (0–9) were presented and had to be subtracted from a double digit between 11 and 20. The design of the experiment was a fast-event related design where each of these two type of trials were presented for 10 times with a probabilistic inter-stimulus interval of on average 3 s (see Pinel et al., 2007). Our contrast of interest is MATH > LANGUAGE.

Due to scanning errors or artefacts, sections of the brain images may be missing in some participants. As the total sample size per study will be up to 700 subjects (more details below), we assured the data quality via a qualitative visual check and a quantitative check on the number of voxels with a measured response value. Subjects who had no data in more than 4% of the median number of voxels over all subjects were excluded from further analysis. This resulted in 87 subjects being removed from the database for this contrast (the total number of available subjects is 1400).

2.1.2. Faces task

In this task, participants were instructed to watch 18-s blocs of either a face or control stimulus. The faces could be angry or neutral (greyscale clips of male or female faces). The control stimulus corresponded to a greyscale video of either an expanding or contracting circle. The fMRI sequence of 160 volumes contained 10 faces and 9 control stimuli. The inter-stimulus interval was 2.2 s. The chosen contrast of interest is ANGRY FACE > CONTROL. After quality control and removing subjects where data was missing, we retain 1400 of the 1890 participants doing this task in the database to sample from.

2.1.3. Monetary incentive delay (MID) task

In this task, participants were presented with sequences of cues, targets and a feedback phase. The cue indicated a possible amount of gain the subjects could win. This could be a small, big or no win. They were instructed to respond when the target was presented (approx. 4 sec. after the cue) after which they received feedback (approx. 1.5 s later) about the win or loss of the trial. The fMRI sequence of 300 volumes contained 144 trials with an inter-stimulus interval of 2.2 s. We have two contrasts of interest using this task. The first is the average effect of participants receiving feedback of a success without a gain (Hit No Gain). The second is Large Win > Small Win. After quality control and removing subjects where data was missing, we retain 1608 of the 1955 participants included in the database to sample from.

2.1.4. Stop signal task

In this task, participants were presented with either a stop/go stimulus where the motor response speed was measured after the go (i.e. signal) stimulus. An event was considered successful if either a motor response was recorded (go) or inhibited (stop). Furthermore, the stop signal delay increased on the next stop by 50 ms if a successful stop was recorded and decreased by 50 ms if participant failed to inhibit on the previous stop trial. The entire fMRI sequence of 444 volumes contained 300 go trials and 80 stop trials. The average inter-stimulus interval was 2.2 s. The chosen contrasts of interest are Stop Failure > Stop Success and its reverse: Stop Success > Stop Failure. After quality control and removing subjects where data was missing, we retain 1450 of the 2026 participants included in the database to sample from.

Note that we do not assess the Stop process as is usually done by including a contrast with a successful Go (e.g. Stop Failure > Go Success). As this corresponds to a completely different cognitive process with a highly delineated set of brain regions associated with, we would expect results to be different when including such a contrast. Written otherwise, we do not consider the results for the chosen contrasts as representative for research done in the Stop signal domain as a whole.

2.1.5. Pre-processing

BOLD time series for each participant were recorded using echoplanar imaging with an isotropic voxel size of 3.4 mm and a repetition time of 2.2 s. For each participant, a number of volumes (depending on the task) were obtained, together with a T1-weighted structural image used for registration. The parameters of the latter were based on the ADNI-protocols.[5]

The pre-processing carried out at the Neurospin center included slice-timing correction, movement correction, coregistration of the functional images to the segmented T1-weighted structural images, non-linear warping of the images into MNI space using a custom EPI template and spatial smoothing of the signal with a 5 mm Gaussian kernel to improve the signal-to-noise ratio, reduce the effects of residual misalignment and meet the conditions for the use of parametric statistics such as those in SPM (Imagen fMRI data analysis methods, revision2, July 2010).

Single subject statistical analyses were performed at each voxel using univariate general linear models (GLM) with all experimental tasks (depending on the four paradigms). For the cognitive task, 18 estimated movement parameters were included as covariates in the design matrix (these correspond to 3 rotations, 3 regular, 3 quadratic and 3 cubic translations, 3 translations shifted 1 TR before, and 3 translations shifted 1 TR later). For the other three tasks, 21 additional columns (corresponding to short and long term movement effects) were added to account for estimated movements. A standard autoregressive [AR(1)] noise model was estimated to account for temporal correlation in the time series.

[2] http://joliot.cea.fr/drf/joliot/en/Pages/research_entities/NeuroSpin.aspx.
[3] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/.
[4] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.
[5] http://adni.loni.usc.edu/methods/documents/mri-protocols/.

2.2. Sampling procedure and group analyses

We use two general sampling procedures to assess replicability: the first to measure test-retest reliability and stability and the second to measure coherence.

The first is obtained through pairwise comparisons of the results of group analyses. Note that results for group analyses are based on statistical parametric maps (SPMs) containing the test statistic (e.g. *t*-values) in each voxel. Thresholded SPMs are binary images where 1 corresponds to a voxel being statistically declared significant and 0 otherwise.

To obtain the independent replications, we start by sampling $N = 10$ subjects at random from the entire database into set *A*. Next, we sample $N = 10$ from the remaining $(1400 - 10)$ subjects at random into set *B*. We fit the group level models and perform statistical inference as described below. The resulting non-thresholded SPMs, together with the thresholded SPMs, are saved and this sampling process is repeated for a total of 50 times. In a next step, we repeat the sampling process, but increase the number of participants in each set by 10. This process continues until we reach a maximum of $N = 700$ subjects per set. In total, we thus generate 3500 comparisons between two independent replications (*A & B*).

To measure coherence, results for more than 2 group analyses are compared. Therefore, we obtain as many independent subsets of subjects as possible from the total database within each sample size. For instance, we create 140 disjoint sets each of 10 subjects for $N = 10$, 70 sets of 20 subjects for $N = 20$, etc. We continue in steps of $N = 10$ until we obtain 3 disjoint subsets of $N = 460$ subjects. This is the maximum sample size for this measurement as we need at least 3 subsets. We iterate the sampling procedure at each sample size 50 times.

The second level group analysis for each sampled group of subjects consists of fitting a linear mixed model in each voxel. This allows us to model within- and between-subject variability separately, and was performed using FLAME1 (Woolrich et al., 2009) from the FSL (RRID:SCR_002823) package (Smith et al., 2004). After estimating the parameters of interest in the group model, we obtain the SPMs. These *t*-images are used for assessing unconditional replicability.

To measure conditional replicability, we threshold the image while apply a correction for multiple testing. The choice depends on our feature of selection. When focussing on voxels, we apply a voxelwise correction where the false discovery rate is controlled at level 0.05 using the Benjamini and Hochberg approach (Benjamini and Hochberg, 1995). When features of interest are clusters, we first define a cluster forming threshold at $Z = 2.3$, then use a 26-voxel neighbourhood search algorithm (default values) and finally control the family wise error rate at 0.05 using the Gaussian Random Field theory (Friston et al., 1994). Note however that in prior work, Woo et al. (2014) argue to use more conservative cluster forming thresholds to ensure valid family-wise error rate control.

2.3. Measures for assessing replicability

2.3.1. (Un)conditional test-retest reliability

To assess unconditional test-retest reliability, we measure the similarity between two non-thresholded SPMs using the Pearson product-moment coefficient (ρ) for the correlation between the *t*-images of study *A* and study *B*.

Conditional test-retest reliability is measured using the percent overlap of activation (Maitra, 2010). This measure is an adaptation of the Dice similarity index (Dice, 1945) or Sørensen similarity coefficient (Sørensen, 1948). Let $V_{A,B}$ represent the number of statistically significant voxels from the intersection between image *A* and *B*, V_A the amount of statistically significant voxels in image *A* and V_B the amount of statistically significant voxels in image *B*. The proportion of overlap $\omega_{A,B}$ in two images is then defined as:

$$\omega_{A,B} = \frac{V_{A,B}}{V_A + V_B - V_{A,B}}.$$

This measure focusses on the spatial pattern of activation in thresholded SPMs and ranges from 0 (no overlap of activation) to 1 (perfect overlap).

Note that we require at least one voxel declared significant in either image *A* or *B*. Furthermore, $\omega_{A,B}$ may be confounded by the proportion of voxels declared active in both images. In one extreme, if all voxels from both image *A* and *B* are declared statistically significant, then the overlap will be perfect. Hence, the effect of the sample size on $\omega_{A,B}$ may be different depending the threshold for significance. A demonstration of this effect is given in a numerical simulation in the appendix. To deal with this possible confounding effect, we also apply different criteria to determine statistical significance of the voxels, using a FDR control at 0.001, 0.01, 0.1 and 0.2 for the cognitive contrast. Furthermore, we stress that results obtained by this measure are conditional on the chosen threshold for significance.

In addition, one can also expect the effect of sample size to be different depending on the amount of true activation. Furthermore, as *N* increases, so does the proportion of voxels classified as statistically significant. To circumvent this issue, we run a second procedure for the cognitive contrast (as an additional check) to calculate $\omega_{A,B}$ in which we condition on the proportion of voxels being classified as active as opposed to the chosen threshold for significance. This is referred to as adaptive thresholding. In this procedure, the total number of voxels to be classified as statistically significant is set *a-priori*, in this case at 20% of the total number of masked voxels. We then adapt the significance thresholding level for each group level analysis so that the percentage of significant voxels matches this target percentage. The variable of interest remains the overlap between images ($\omega_{A,B}$) while increasing the sample size. By adapting the level of threshold for significance within each group analysis, we can check whether we obtain similar response curves as the regular analysis. Although this is not a valid inference procedure (the type I error rate will be higher than 5%), this procedure allows us to remove the effect of the proportion of significant voxels (which increases with sample size). In other words, this strategy allows us to assess the effect of the sample size on the delineation of spatial activation in replication contexts, irrespective of the varying proportion of voxels classified as significant.

2.3.2. Coherence

The second measure of interest is coherence (or concordance) of the thresholded SPMs (Genovese et al., 1997; Liou et al., 2006, 2003; Thirion et al., 2007). Coherence is estimated based on Cohen's kappa (κ). It is a measure that considers agreement over replications based on the proportion of voxels that are correctly labelled as active or inactive, corrected for the probability of being correctly classified by chance alone.

Starting from a binary label per voxel (1/0; active/non-active respectively), we calculate coherence from $R \geq 3$ independent replications. Adapting the notation from Thirion et al. (2007), we obtain over the *R* replications per voxel $v = 1, \ldots, V$ an *R*-dimensional binary vector: $[g_1(v), \ldots, g_R(v)]$. The sum in each voxel over all replications is denoted as $G(v) = \sum_{r=1}^{R} g_r(v)$. Over all voxels, we assume G(v) to be a mixture distribution of two binomials - corresponding to either the inactive or active state of voxels.

We define π_A^1 as the probability of a true active voxel correctly being declared statistically significant. Conversely, we define π_I^1 as the probability of a true inactive voxel incorrectly being classified as active. The complements are defined as $\pi_A^0 = 1 - \pi_A^1$ and $\pi_I^0 = 1 - \pi_I^1$. Finally, let λ denote the proportion of truly active voxels. An overview of these parameters is given in Table 1.

Let $p_0 = \lambda\pi_A^1 + (1-\lambda)\pi_I^0$ represent the proportion of correctly classified voxels. To adjust p_0 for correct classification by chance, we first define the probability of a voxel being declared active: $\pi^1 = \lambda\pi_A^1 + (1 - \lambda)\pi_I^1$. Then we have $p_c = \lambda\pi^1 + (1 - \lambda)(1 - \pi^1)$. The values p_0 and p_c naturally fit in the formula of Cohen's κ:

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \quad (1)$$

with kappa lying between 0 (no coherence) and 1 (perfect coherence).

To estimate the parameters λ, π_A^1 and π_I^1, we assume G to be independent and identically distributed with the following density function

$$f\left(G, R, \lambda, \pi_A^1, \pi_I^1\right) = \lambda P_1\left(G; R, \pi_A^1\right) + (1-\lambda) P_2\left(G; R, \pi_I^1\right),$$

where P_1 and P_2 are characterised by the probability mass function of a binomial distribution.

The E(xpectation) M(aximisation) algorithm is used to estimate λ, π_A^1 and π_I^1. Full details are given in the appendix; while the *NeuRRoStat* package[6] provides an implementation in R. After running the EM-algorithm, we plug the obtained estimates into equation (1) to calculate κ. An illustration of coherence is given in Fig. 1.

2.3.3. Stability

Our third measure of interest probes the stability of fMRI results (Roels et al., 2015). In this paper, we focus on clusters of voxels as the main feature of interest to describe the relation between sample size and stability. The latter is expressed in terms of variability of outcome metrics, where an increased variability indicates less stable results and is therefore an indication for lower replicability.

We describe and quantify stability using several metrics. First, we measure the size (in absolute number of voxels) of the clusters that are classified as significant. We do the same for the largest cluster in each study expressed as the proportion of total masked voxels. These metrics are obtained from the same sampling framework as described above (section 2.2).

Next, we measure the count of total, unique and overlapping clusters when comparing two independent replications. We use two definitions of overlapping cluster: a lenient and a conservative one. In the first, a cluster is replicated as soon as one voxel from a given cluster overlaps with a voxel from a cluster in the corresponding replication. In the second, a cluster is replicated if at least 50% of the voxels in a given cluster overlaps with a cluster in the corresponding replication. To see how much overlap there is between clusters, we also calculate the proportion of overlapping *clustered* voxels in both replications. Finally, we measure the variability of the number of clusters that are declared significant as well as the variability of the cluster size (in number of connected voxels). Results are more stable when the variability is low.

An overview of all measurements, their category, and the corresponding feature selection for fMRI results, is given in Table 2.

3. Results

To begin with, we stress that the relationship between sample size and our measurements of replicability depends on the experimental paradigm or chosen contrast therein. We observe better replicability for tasks such as solving math equations, watching angry faces and the monetary incentive delay task (at least for one contrast). A lower replicability is associated with the Stop signal contrasts. For the latter, a higher statistical power (i.e. sample size) is needed. It is therefore impossible to provide an absolute description of the replicability of fMRI results. The results presented here will inevitable depend on the experimental paradigm and contrast. Our results do demonstrate that replicability for any experimental paradigm/contrast generally increases with sample size.

In Fig. 2, we present the unconditional test-retest reliability of voxelwise inference of independent replications of fMRI results against increasing sample size for all paradigms. We observe a median (over all resampling iterations) ρ between $[0.304; 0.80]$ over all tasks when $N = 50$. Furthermore, the maximum median value of ρ for the cognitive, faces, MID (Hit No Win), MID (Large Win > Small Win) contrast equals 0.976 ($N = 690$), 0.982 ($N = 700$), 0.974 ($N = 700$) & 0.888 ($N = 700$) respectively suggesting ρ asymptotically approaching 1 for these tasks. However, for both contrasts of the Stop signal task, we observe a maximum median value of the correlation of 0.86 ($N = 700$), suggesting a higher sample size is needed.

In Fig. 3 (all contrasts) and 4 (only for the MATH > LANGUAGE contrast), we plot the test-retest reliability conditional on the chosen threshold for significance against increasing sample size. First, we fix the false discovery rate at 0.05 and retain that rate over all sample sizes (Fig. 3). As expected, the overlap increases with the sample size. However, we need at least $N = 200$ to observe a range of $[0.098; 0.744]$ corresponding to the median (over all resampling iterations) overlap between two replications over all tasks/contrasts. Furthermore, we only reach a maximum percentage of overlap between $[0.493; 0.893]$ with $N \in [430; 700]$. Note that the variability of the overlap estimates decreases as the sample size increases. Furthermore, since the overlap is conditional on the chosen threshold for significance, we also investigated the same measurement when controlling the FDR at 0.001, 0.01, 0.1 and 0.2 using the MATH > LANGUAGE contrast. We observe a main effect where liberal thresholds correspond to slightly higher values for the overlap over the entire range of N. Figure A.7 in the appendix shows the effect of varying FDR thresholds, with the overall trends being very similar to those of the cognitive task in Fig. 3.

Fig. 4 shows results for the adaptive thresholding strategy (focussing on the cognitive task), where the number of significant voxels in each test is restricted to 20% of the total number of masked voxels. As expected, the overlap at the lowest sample sizes is higher compared to the previous setting. For instance, the median omega equals 0.303 for $N = 20$ (with a median uncorrected threshold of $P \leq 0.156$). Again, we observe a gradually increasing overlap with increasing sample size and a maximum median of $\omega = 0.794$ at $N = 700$ (median threshold $P \leq 0.00001$). These results suggest that a high number of subjects is needed to replicate the voxelwise spatial shape of activation.

The final measurement to assess replicability of voxelwise inference is the coherence (κ) of fMRI results (Fig. 5). We compare results of different images that are all thresholded such that the FDR remains constant at 0.05. As with the previous measurements, κ increases with increasing sample size. We observe a median coherence ≥ 0.80 when $N = 180$ for both the cognitive and faces task. The maximum median κ for the faces task, the Stop signal task (Stop Failure > Stop Success), Stop signal task (Stop Success > Stop Failure), the MID task (Hit No Win) and MID task (Large Win > Small Win) equals 0.865 ($N = 460$), 0.774 ($N = 460$), 0.664 ($N = 460$), 0.864 ($N = 460$) & 0.774 ($N = 460$) respectively. The highest median κ (0.894) is observed at $N = 460$ for the cognitive task. Note that we observe a considerable gain in κ when N is low. Indeed, it is possible to fit a continuous piecewise linear regression model where break-points can be estimated using an iterative optimization technique (Muggeo, 2003). Fitting this model on the observed values for κ for the cognitive task yields a break-point at $N \approx 60$. Under 60, an increase of 10 subjects in the group analysis is associated with an average improvement of 0.129κ, while above 60 this becomes 0.005κ per 10

Table 1
Parameters of a mixture distribution where voxels correspond either to the distribution characterised by the inactive or active state. A high coherence of independent replications is indicated by a separation of the mixture distribution into the diagonal parameters compared to the off diagonal parameters.

		DECLARED		
		Active	Inactive	
TRULY	Active	π_A^1	π_A^0	λ
	Inactive	π_I^1	π_I^0	$1-\lambda$

[6] See: https://github.com/NeuroStat/NeuRRoStat.

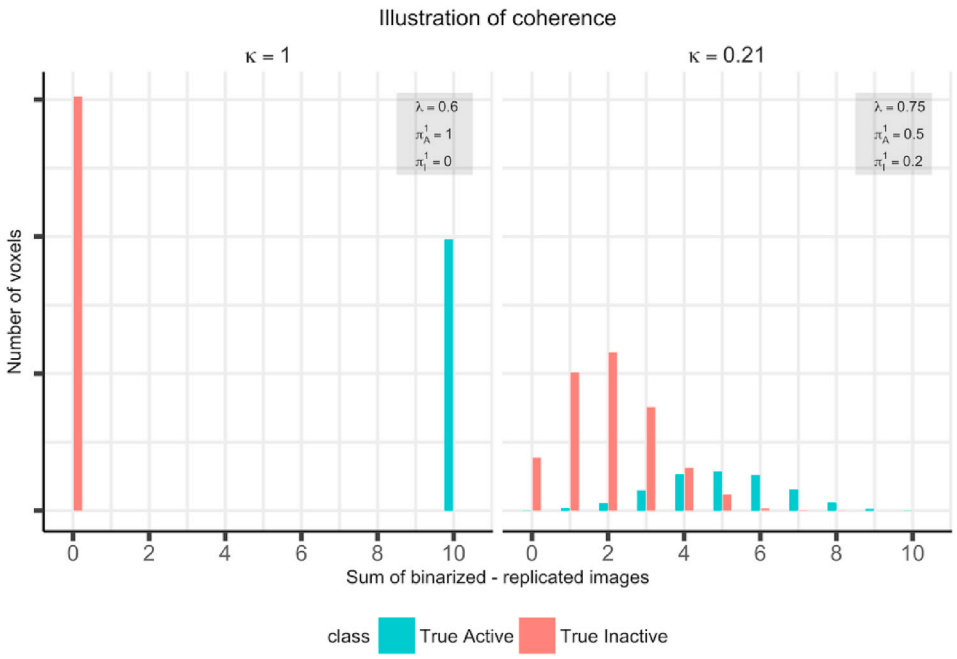

Fig. 1. Illustration of R = 10 thresholded SPMs with on the X-axis the result of summing the binary images (range 0–10) and on the Y-axis the number of voxels. The illustration is shown for brain images with any dimension and therefore no values are plotted on the Y-axis. The left panel shows a perfect agreement over the labels of each state of the voxel (active/inactive, $\kappa = 1$). A more likely scenario is shown in the right panel where a mixture two binomial distributions can be fitted to the observed data ($\kappa = 0.21$). Annotated parameter values correspond to the likelihood function.

Table 2
Overview of the measurements for test-retest reliability, stability and the coherence of results of an fMRI data-analysis. It is also indicated whether the measurement depends on a threshold for statistical significance and whether it is used for voxel- or clusterwise inference.

Measurement	Calculation	Category	Feature
Test-retest reliability	Pearson's Product-Moment Correlation Coefficient (ρ)	Unconditional	Voxel
Test-retest reliability	$\omega_{A,B} = \frac{V_{A,B}}{V_A + V_B - V_{A,B}}$	Conditional	Voxel
Coherence	$\kappa = \frac{p_0 - p_c}{1 - p_c}$	Conditional	Voxel
Stability	(Absolute/relative) average cluster size	Conditional	Cluster
	Number of (unique) clusters	Conditional	Cluster
	Proportion of overlapping voxels in a cluster	Conditional	Cluster
	Variability of cluster count	Conditional	Cluster
	Variability of cluster size	Conditional	Cluster

subjects. Note that for the cognitive, faces and MID (Hit No Win) task, the effect of sample size seems to reach a plateau below 1. Also, we observe a low variability between the different estimates of κ from all the resampling runs for each sample size.

Moving from the voxelwise dimension to the cluster dimension, Fig. 6 shows the stability of fMRI results for the MATH > LANGUAGE contrast. To condense the amount of results, we refer to the appendix for figures containing the results of the other contrasts. Generally, the sample size to response curves are similar. Differences are mainly observed in the absolute values.

First, we observe a convergence in the size and number of clusters as the sample size increases. This is shown in panel A and E of Fig. 6. The average size (in number of voxels) of the significant clusters in a given fMRI data-analysis stabilizes or increases only marginally with sample size. Furthermore, we observe on average 1.02 significant clusters at $N = 700$ (in only two analyses out of 100, we found an additional small significant cluster). Moreover, this single cluster corresponds on average to 31% of the masked brain (panel B from Fig. 6). Interestingly, in some analyses at $N = 100$ it is already possible to observe these large clusters. This is reflected in the gap in panel A and B of Fig. 6. Second, the variability on the number of clusters and their corresponding size at first

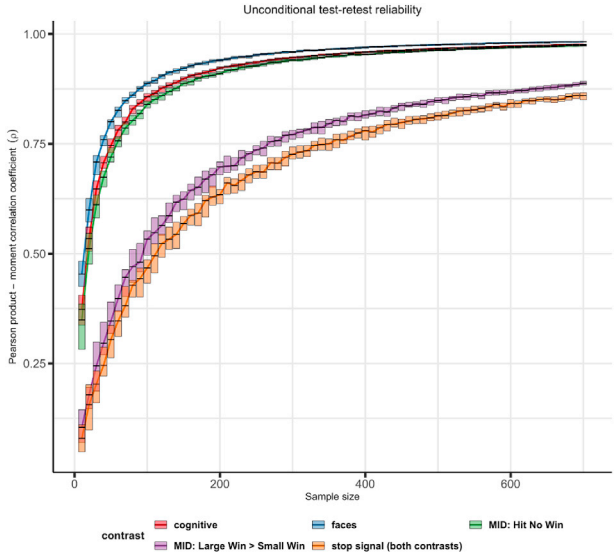

Fig. 2. Pearson product-moment correlation coefficient between two independent replications of non-thresholded fMRI statistical parametric maps under increasing sample size. Solid line represents the median value over all resampling iterations. The height of the boxes represents the distance between the third and first quantile over all iterations. Results indicate a median correlation higher than 0.8 between replications at sample sizes starting from N = 80 for the cognitive, faces and MID (Hit No Win) task. We observe a maximum value for the Stop signal paradigm of 0.860 (N = 690) for both contrasts (the correlation is identical between reverse contrasts as the t-values are identical, only the sign differs and is therefore irrelevant in this case).

increases with the sample size (panel C and D of Fig. 6). However, for $N \geq 100$ in case of the number of clusters and $N \geq 200$in case of the size of clusters, the variability then decreases again, which corresponds to a higher stability. The same trend is observed in the other contrasts. Third, as sample size increases, the average number of clusters that are observed in both the replications increases at first. It then converges to the average total number of significant clusters as the latter decreases to 1 significant cluster per analysis. This is true for both definitions of replicated clusters.

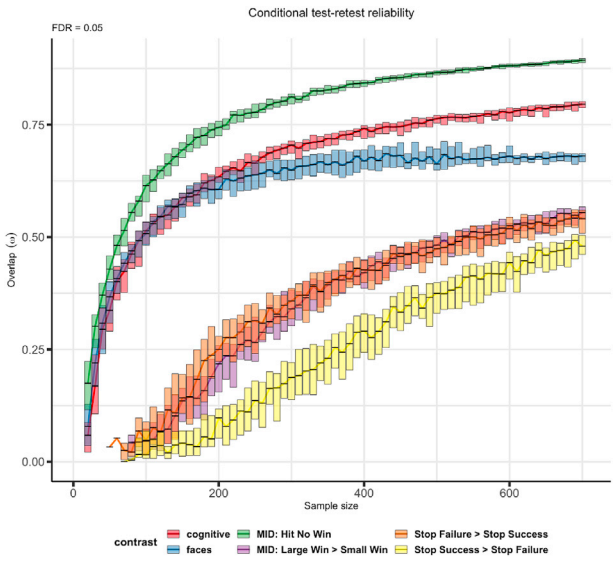

Fig. 3. Overlap under increasing sample size between two independent replications of thresholded fMRI studies. Control for multiple testing is done at the voxelwise False Discovery Rate of 0.05. Solid line represents the median value over all resampling iterations. The height of the boxes represents the distance between the third and first quantile over all iterations. Results indicate variability between and within experimental paradigms.

When comparing these definitions, we observe that a higher sample size is needed to replicate clusters if at least 50% of a cluster has to be observed at the same location in a given replication. Furthermore, only for $N \geq 50$ do we observe a higher number of replicated versus non-replicated clusters. This is not the case when using a lenient definition of replicated clusters. In contrast to the individual voxel results, we observe a better reliability when reporting clusters of activation when using the lenient definition. For instance, once $N \geq 180$, it is possible to replicate almost every cluster (there are no unique clusters). When using a conservative definition, we need $N \geq 570$ to replicate almost every cluster. Note, if we calculate the proportion of clustered voxels that overlap in both the replications (panel F in Fig. 6), we observe values similar to the voxelwise percent overlap of activation. The median proportion overlapping clustered voxels equals 0.645 when $N = 200$ and the maximum median proportion equals 0.799 at $N = 700$. With respect to the other tasks, we observe higher values for stability when the experimental paradigm results in larger areas of activation. For instance, the MID (Hit No Win) task results in clusters with a size over 60% of the masked brain at the highest sample size. Moreover, there are close to zero non-overlapping clusters over $N = 100$. Both the Stop signal contrasts on the other hand result in a high number of smaller clusters (the largest clusters entails approximately 10% & 18% of the masked brain at $N = 700$) which results in lower values for the stability.

Finally, we provide the range of the values over all contrasts for each measurement at $N = 30$. This corresponds roughly to the median sample size of single-group fMRI studies in 2015 (Poldrack et al., 2017). For $N = 30$, we observe a median (over all sampling iterations) correlation $\rho \in [0.203; 0.709]$, a median percent overlap of activation $\omega \in [0; 0.302]$ when the FDR is controlled at 0.05 and a median coherence $\kappa \in [0; 0.504]$. Averaged between both independent replications in each task, we observe a median total number between $[0.75; 45.8]$ significant clusters. The lower bound is observed in the Stop signal task (Stop Success > Stop Failure), while the upper bound is observed in the faces task. When using a lenient definition of overlapping clusters (i.e. only voxel found in both clusters), then we observe a median between $[0; 11.5]$ overlapping clusters and a median between $[0; 34.8]$ non-overlapping clusters. When using a conservative definition (i.e. at least 50% of the voxels within a cluster overlap with the corresponding replication), then we observe a median between $[0; 6.5]$ overlapping clusters and a median between $[0.75; 38.8]$ non-overlapping clusters. The observed median proportions of overlapping voxels in a cluster between both replications are between $[0; 0.33]$.

4. Discussion

The primary aim of this paper was to replicate and extend previous empirical assessments of replicability of fMRI results. We used data from 1400 subjects from the IMAGEN project in a subsampling approach to investigate the role of sample size (i.e. essential to obtain sufficient statistical power). Participants were involved in four types of experimental paradigms (i.e. a cognitive task, watching angry faces, a monetary

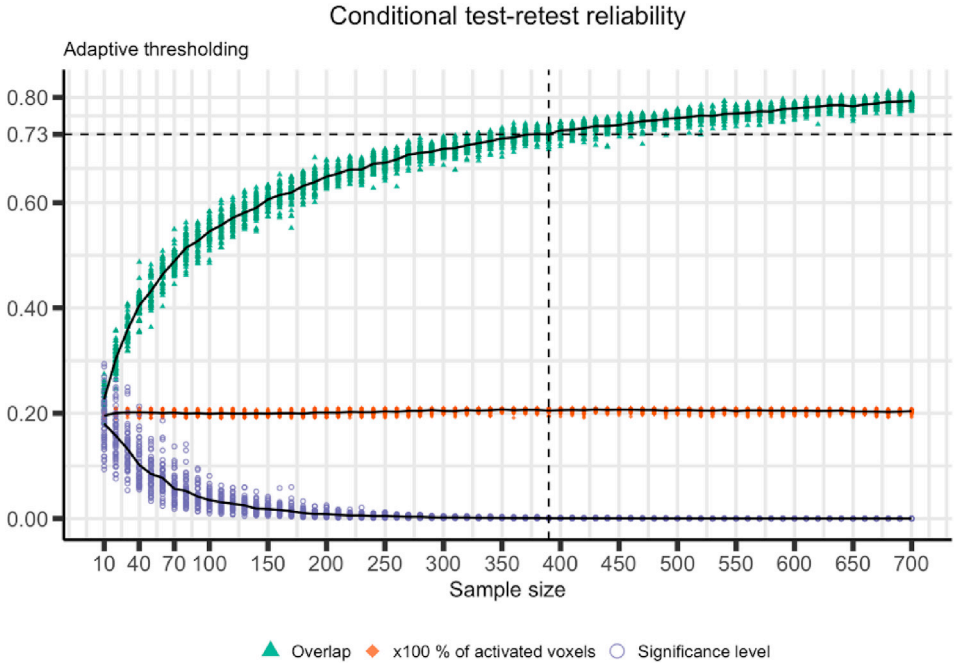

Fig. 4. Overlap (in triangles) under increasing sample size while keeping the number of significant voxels fixed at 20%. The overall pattern is similar to applying the same significance thresholding level over all sample sizes (MATH > LANGUAGE). The dashed lines represent the median overlap when the average significance level equals $P \leq 0.001$, uncorrected for multiple testing. Thus, we observe in this database a median $\omega = 0.73$when we have $N = 390$ and apply a (non-recommended) conventional significance thresholding level of $P \leq 0.001$.

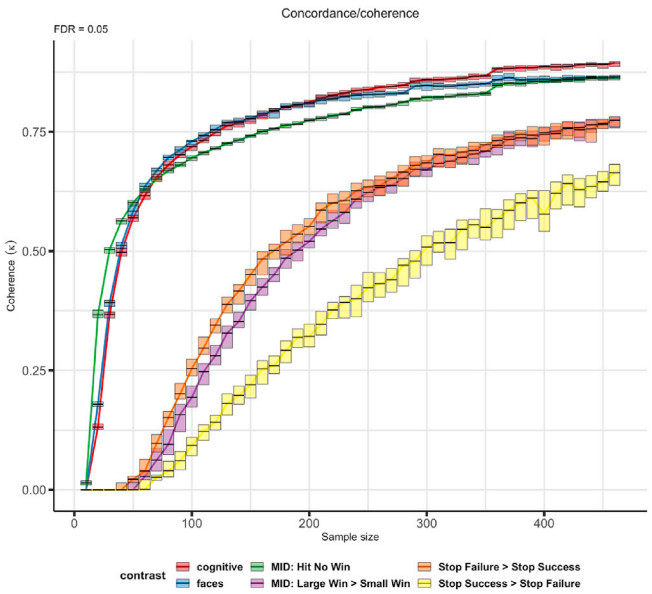

Fig. 5. Coherence in categorizing voxels into true active or true inactive states over R independent replication studies while increasing the sample size. The number of independent studies ranges from 140 (N = 10) to 3 (N = 460). *Solid line represents the median value over all resampling iterations. The height of the boxes represents the distance between the third and first quantile over all iterations.* We observe a median value for κequal to 0.80 when N = 180 for the cognitive and faces task and N = 250 for the MID (Hit No Win) task. The maximum value for the Stop signal task equals 0.774 (N = 460), Stop Failure > Stop Success.

incentive delay task and a Stop signal paradigm). We used independent replications to assess replicability in a multidimensional framework. Furthermore, we quantified replicability using three evaluation measurements: (un)conditional test-retest reliability, coherence and stability of fMRI results.

First we observe variability in terms of the degree of replicability depending on the experimental paradigms and chosen contrast. An fMRI task resulting in larger areas of activation is associated with higher values on our measurements of replicability. We hypothesize that more subjects are needed to obtain replicable results for experimental paradigms associated with small effect sizes and/or small areas of activation. Interestingly, previous research on activation reliability in the MID paradigm showed higher values for test-retest reliability in a large win condition (Wu et al., 2014). In our study however, we observe higher values for replicability in the condition where participants receive no gain. In addition for the Stop signal task, we observe higher values for replicability in the Stop Failure > Stop Success condition compared to its reverse contrast. As pointed out by a reviewer, this is a surprising observation as successful stopping is associated with a relatively circumscribed neural circuit, while failing to stop is probably not. Since participants could fail to stop for many reasons (e.g. subject wanted to stop but responded too late or did not pay attention), we would expect more diverse activation patterns.

With respect to our various measurements of replicability, we note the following key observations. First, we show that the correlation between two images containing the test-statistics from a test and independent replication can be relatively high even at small sample sizes (e.g. lower than 70 subjects). This is consistent with earlier results of Turner et al. (2018) who observed values in the same range. Note that Sochat et al. (2015) also demonstrate relatively high correlations between test-statistics of thresholded SPMs. In this work, we thus extend this observation to complete images.

A high test-retest reliability between full (i.e. non-thresholded) images is important for several applications in neuroimaging. In decoding studies using for instance multivariate pattern classifiers (Haxby et al., 2014), the goal is not only to achieve a high prediction accuracy but also to evaluate the spatial patterns underlying the classification accuracy (Conroy et al., 2013; Poldrack et al., 2009). Our results suggest that replicability of these spatial patterns is likely to be acceptable.

Interestingly, the replicability of thresholded images, using only binary values (i.e. activated or not) can be poor at common fMRI sample sizes (N ≈ 30). This is indicated by the low overlap of activation between replications for the contrasts considered in our study. Furthermore, when enforcing the same proportion of active voxels across analyses by means of adaptive thresholding, we demonstrate the same poor relation between sample size and the delineation of the approximated true spatial pattern of activation. The same observation regarding the overlap values is found in Turner et al. (2018) who used subjects from the Human Connectome Project. In the latter database however, subjects can be genetically related to each other which potentially influences replicability. Multiple factors could explain this pattern including the effect size of interest, characteristics of the studied population (13–15 years old adolescents), design of the experiments (e.g. some tasks are based on a limited number of trials), difficulties with quality control when working with a large database, the signal-to-noise ratio of fMRI data etc. More details regarding these factors are given in the paragraph discussion limitations of this study below.

Next, even though the overlap between thresholded images can be poor, we observe a better performance in terms of coherence between independent replications. This is expected as coherence is measured using more than 2 replications (i.e. from 3 to 140), while overlap is measured using only pairs of images. Furthermore, the coherence is measured using both active and inactive voxels. Note that we observe lower values for the coherence than Thirion et al. (2007). For instance, at $N \approx 30$the median value $\kappa = 0.80$ in the latter, while we observe only a median $\kappa = 0.504$ at best at the same sample size. This difference may be due to the difference in type of cognitive task, experimental paradigms or statistical analysis. For instance, Thirion et al. (2007) show how the coherence declines with more conservative thresholding. Indeed, while Thirion et al. (2007) used an uncorrected $P \leq 0.001$ level for statistical significance, we corrected for multiple comparison at a false discovery rate of 0.05. For larger sample sizes, this is substantially more conservative than thresholding at an uncorrected $P \leq 0.001$.

Finally, while replicability of individual voxels can be low at typical sample sizes, we observe a better performance when looking at clusters of voxels. Although initially, the variability of the number and size of clusters increases with increasing sample sizes, we have shown that it stabilizes for higher sample sizes. As can be expected, the number of *non-replicable* clusters is low if one uses a lenient definition of non-replicable clusters. That is when at least one voxel from the cluster overlaps with a cluster in a given replication. If one considers a cluster only replicated if at least half of it is observed in a given replication, then more subjects are needed. This is true for any experimental paradigm. Moreover, these results potentially depend on the number and size of selected clusters. Small clusters of activation (or highly variable cluster sizes) are more prone to be non-replicable. Furthermore, it has been shown that parametric statistical methods relying on cluster-wise inference are associated with an inflated false positive rate (Eklund et al., 2016). For this reason, Eklund et al. (2016) suggest using non-parametric methods such as permutation testing (Brammer et al., 1997; Bullmore et al., 1999; Nichols and Holmes, 2002; Winkler et al., 2014). We have not investigated the effect of using these methods.

In this paper, we focused on the relation between replicability and statistical power and demonstrate potentially low voxelwise replicability due to low statistical power. However, a lack of replicability is not the only problem associated with low statistical power. There is (1) an increased probability of a positive research claim being false (Ioannidis, 2005), (2) an overestimation of the reported effect sizes compared to the true effect size (Cremers et al., 2017) and (3) an increased risk of missed effects (i.e. false negatives) which potentially induces publication bias (Acar et al., 2018; Rosenthal, 1979; Sterling, 1959). Note that the most

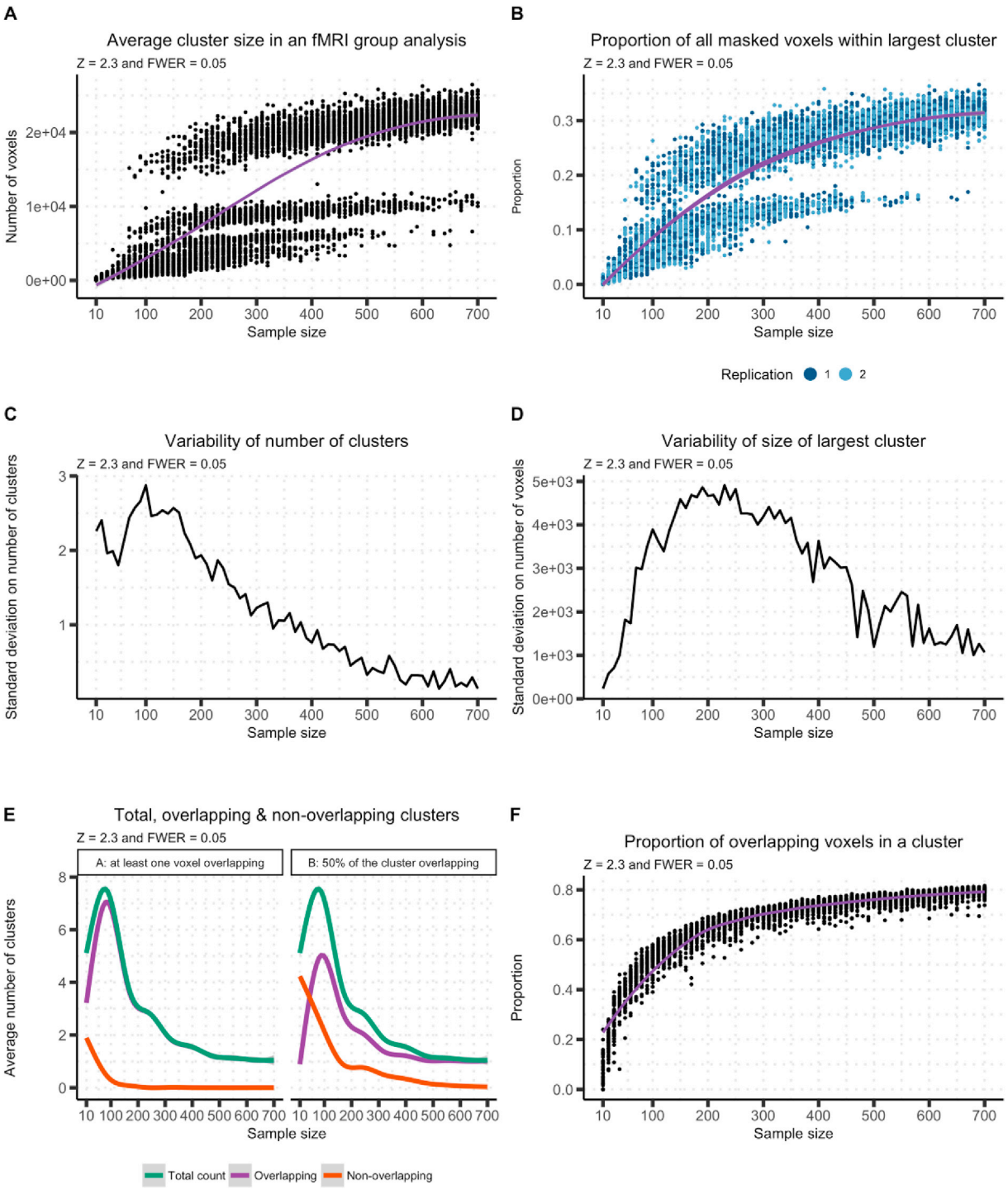

Fig. 6. Results on stability of fMRI data analysis inference on cluster-level (MATH > LANGUAGE). Panel A shows the average cluster size for a given fMRI group analysis at each sample size. Panel B shows the proportion out of the total number of masked voxels of the largest cluster (separated for both replications for visualization purpose). Panel C and D display the standard deviation respectively on the number of clusters and on the size of the largest cluster. The top 4 panels show how the number and size of significant clusters stabilize as the sample size increases. Note that panel A to D do not explicitly compare independent replications of group analyses. In panel E and F, replications are compared pairwise. For panel E, we calculate the total number of clusters and then split between the number of overlapping versus non-overlapping clusters. We separate between two definitions of overlapping clusters: at least one voxel is overlapping between both clusters (A) or at least 50% of both clusters are overlapping (B). Counts are averaged over the two replications. The curves are obtained by fitting a generalized additive model with cubic splines. In panel F, we look at the proportion of overlapping voxels in a cluster. Clusters of voxels converge to the same spatial location and shape as the sample size increases.

optimistic rate of missing contrasts in the fMRI literature is estimated to be 6/100 (Samartsidis et al., 2017). However, this estimate is (1) based on one database and (2) corresponds to missing studies where no single effect in the entire masked brain is observed. Another consequence of underpowered studies is the detection of some but not all true positive voxels within a single fMRI study. In other words, even with low power one is likely to find at *least one* significant voxel (Cremers et al., 2017) which is not sufficient for fMRI studies to be replicable.

Several solutions to increase the sample size have been suggested (Cremers et al., 2017; Poldrack et al., 2017). To begin with, it is advised to run power analyses at the design phase of fMRI studies. Some tools exist to calculate either the power to detect a true effect in a region of interest (Mumford and Nichols, 2008) or when designing a whole brain analysis (Durnez et al., 2016). Second, data sharing initiatives such as NeuroVault[7] (Gorgolewski et al., 2015) create the opportunity to pool data into meta-analyses (see Salimi-Khorshidi et al., 2009; Wager et al., 2007) or mega-analyses. By sharing data, it might be possible to increase the power of subsequent studies. Finally, if increasing the sample size is infeasible, it is possible to complement statistical analyses with sensitivity analyses (Wilke, 2012). The idea here is to iterate the statistical analysis but systematically remove and replace r subjects. The sensitivity can then be calculated using the percent overlap of activation. Another approach presented by Liou et al. (2003) and Liou et al. (2006) is to create so called "reproducibility maps" using multiple sessions or runs of the same participants in an experiment.

Note that our definition of replicability concerns the generalizability of scientific claims as different data are used to answer the same research question. A vast number of studies (e.g. Gorgolewski et al., 2013; Lee et al., 2010; Machielsen et al., 2000) have investigated the reliability of fMRI results over repeated measurements using a small number of subjects ($N \in [1, \ldots, 18]$). The research question in this case corresponds to the ratio of within- versus between-subject variability where one is interested in partitioning and quantifying the total observed variability into both components. We have not addressed this issue in the current paper. Another related research question is to investigate whether individual differences in behaviour can be used as a predictor for replicability in task-based fMRI results.

To end the discussion, we mention some important limitations of this study. First only a limited number of trials are used in the design of the experiments contained in the IMAGEN database. The inter-subject variability of such a fast paced experiment was investigated in (Pinel et al., 2007). Although these researchers could reliably detect most peaks of activation when compared to longer scanning sequences, improved measurements of replicability can be expected with longer scanning sequences. This has been demonstrated in recent work by Nee (2018) who reported an increase of replicability of fMRI results with the duration (i.e. number of time points) of the scanning sequence. An important research question is therefore the interaction between number of time points in the scanning sequence and sample size on replicability of fMRI results. Furthermore, the replicability of fMRI results will also depend on other factors such as the studied population, type of (cognitive) task, etc. It is not unlikely to observe even more different response curves between the sample size and our measurements of replicability. For instance, there may be a higher replicability when using college-aged participants rather than adolescents. Potentially, adolescents are associated with a greater variability in terms of brain development. This in turn could influence replicability between sets of participants from the same age. In addition, as we only use the pre-processed data, we are not able to investigate the effect of design choices or various pre-processing steps on fMRI replicability. For these reasons, we are unable to provide an absolute number as a required sample size for task-based fMRI replicability.

Second, as data are anonymized before the start of our analysis we are unable to investigate the proportion of between-site variability. In the literature, there are mixed results regarding the influence of the scanner/site. In some studies, substantial between-site variability is observed (Rath et al., 2016) while in others only marginal variability is observed (Costafreda et al., 2007). Furthermore, it has been shown that between-site reliability increases as the number of sessions increases (Friedman et al., 2008). It should be noted that the influence of inter-site variability will be averaged out in our study as we sample subjects over all scanning sites. Therefore, our measurements of replicability could also change depending on the effect of between-site variability on the ability to replicate results.

Third, our results are based on only four experimental paradigms for which we did not explore all possible contrasts representing these paradigms (Thirion et al., 2007; Turner et al., 2018). For instance, in the Stop signal paradigm, we were not able to include a contrast with a Go condition. Given the vast literature on this contrast, we hypothesize results will be better when isolating a Stop process. Furthermore, Thirion et al. (2007) observe variability in their reliability analysis related to the statistical model used in the group analysis. However, we compared current results to those obtained using ordinary least squares and did not observe substantial differences (results not shown).

In conclusion, we observe over experimental paradigms and contrasts an effect of sample size and differences between voxel and cluster level analyses in the ability to replicate fMRI results. To our knowledge, the latter has not been demonstrated before. We hope these results further foster designing fMRI studies with appropriate sample sizes and ultimately lead to more replicable scientific findings.

Funding

RS and BM would like to acknowledge the Research Foundation Flanders (FWO) for financial support (Grant G.0149.14N). This work received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT-2007–037,286), the Horizon 2020 funded ERC Advanced Grant 'STRATIFY' (Brain network based stratification of reinforcement-related disorders) (695313), ERANID (Understanding the Interplay between Cultural, Biological and Subjective Factors in Drug Use Pathways) (PR-ST-0416-10004), BRIDGET (JPND: BRain Imaging, cognition Dementia and next generation GEnomics) (MR/N027558/1), the FP7 projects IMAGEMEND(602450; IMAging GEnetics for MENtal Disorders) and MATRICS (603016), the Innovative Medicine Initiative Project EU-AIMS (115300-2), the Medical Research Council Grant 'c-VEDA' (Consortium on Vulnerability to Externalizing Disorders and Addictions) (MR/N000390/1), the Swedish Research Council FORMAS, the Medical Research Council, the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, the Bundesministeriumfür Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; eMED SysAlc01ZX1311A; Forschungsnetz AERIAL 01EE1406A, 01EE1406B), the Deutsche Forschungsgemeinschaft (DFG grants SM 80/7-2, SFB 940/2), the Medical Research Foundation and Medical research council (grant MR/R00465X/1), the Human Brain Project (HBP SGA 2). Further support was provided by grants from: ANR (project AF12-NEUR0008-01 - WM2NA, and ANR-12-SAMA-0004), the Fondation de France, the Fondation pour la Recherche Médicale, the Mission Interministérielle de Lutte-contre-les-Drogues-et-les-Conduites-Addictives (MILDECA), the Assistance-Publique-Hôpitaux-de-Paris and INSERM (interface grant), Paris Sud University IDEX 2012; the National Institutes of Health, Science Foundation Ireland (16/ERCD/3797), U.S.A. (Axon, Testosterone and Mental Health during Adolescence; RO1 MH085772-01A1), and by NIH Consortium grant U54 EB020403, supported by a cross-NIH alliance that funds Big Data to Knowledge Centres of Excellence.

Author contribution

HB, SR, RS, and BM contributed to the conception and design of the manuscript. Data collection and single subject analyses were carried out by the IMAGEN consortium represented by TB, GB, AB, EQ, SD, HF, AG, HG, PG, AH, BI, J-L M, EQ, FN DO, LP, JF, MS, HW, RW and GS. Data analysis and interpretation for this study was performed by HB, SR, RS, and BM. Initial draft of the manuscript was developed by HB. Finally, all authors approve the version to be published.

[7] www.neurovault.org.

Declaration of competing interest

Dr. Banaschewski has served as an advisor or consultant to Bristol-Myers Squibb, Desitin Arzneimittel, Eli Lilly, Medice, Novartis, Pfizer, Shire, UCB, and Vifor Pharma; he has received conference attendance support, conference support, or speaking fees from Eli Lilly, Janssen McNeil, Medice, Novartis, Shire, and UCB; and he is involved in clinical trials conducted by Eli Lilly, Novartis, and Shire; the present work is unrelated to these relationships. Dr. Barker has received honoraria from General Electric Healthcare for teaching on scanner programming courses and acts as a consultant for IXICO. The other authors report no biomedical financial interests or potential conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116601.

References

Aarts, A.A., Anderson, J.E., Anderson, C.J., Attridge, P.R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F.A., Bowman, S.D., Brandt, M.J., Braswell, E., Brohmer, H., Brown, B.T., Brown, K., Brüning, J., Calhoun-Sauls, A., Callahan, S.P., Chagnon, E., Chandler, J., Chartier, C.R., Cheung, F., Christopherson, C.D., Cillessen, L., Clay, R., Cleary, H., Cloud, M.D., Conn, M., Cohoon, J., Columbus, S., Cordes, A., Costantini, G., Alvarez, L.D.C., Cremata, E., Crusius, J., DeCoster, J., DeGaetano, M.A., Penna, N.D., Den Bezemer, B., Deserno, M.K., Devitt, O., Dewitte, L., Dobolyi, D.G., Dodson, G.T., Donnellan, M.B., Donohue, R., Dore, R.A., Dorrough, A., Dreber, A., Dugas, M., Dunn, E.W., Easey, K., Eboigbe, S., Eggleston, C., Embley, J., Epskamp, S., Errington, T.M., Estel, V., Farach, F.J., Feather, J., Fedor, A., Fernández-Castilla, B., Fiedler, S., Field, J.G., Fitneva, S.A., Flagan, T., Forest, A.L., Forsell, E., Foster, J.D., Frank, M.C., Frazier, R.S., Fuchs, H., Gable, P., Galak, J., Galliani, E.M., Gampa, A., Garcia, S., Gazarian, D., Gilbert, E., Giner-Sorolla, R., Glöckner, A., Goellner, L., Goh, J.X., Goldberg, R., Goodbourn, P.T., Gordon-McKeon, S., Gorges, B., Gorges, J., Goss, J., Graham, J., Grange, J.A., Gray, J., Hartgerink, C., Hartshorne, J., Hasselman, F., Hayes, T., Heikensten, E., Henninger, F., Hodsoll, J., Holubar, T., Hoogendoorn, G., Humphries, D.J., Hung, C.O.Y., Immelman, N., Irsik, V.C., Jahn, G., Jäkel, F., Jekel, M., Johannesson, M., Johnson, L.G., Johnson, D.J., Johnson, K.M., Johnston, W.J., Jonas, K., Joy-Gaba, J.A., Kappes, H.B., Kelso, K., Kidwell, M.C., Kim, S.K., Kirkhart, M., Kleinberg, B., Knežević, G., Kolorz, F.M., Kossakowski, J.J., Krause, R.W., Krijnen, J., Kuhlmann, T., Kunkels, Y.K., Kyc, M.M., Lai, C.K., Laique, A., Lakens, D., Lane, K.A., Lassetter, B., Lazarević, L.B., Le Bel, E.P., Lee, K.J., Lee, M., Lemm, K., Levitan, C.A., Lewis, M., Lin, L., Lin, S., Lippold, M., Loureiro, D., Luteijn, I., MacKinnon, S., Mainard, H.N., Marigold, D.C., Martin, D.P., Martinez, T., Masicampo, E.J., Matacotta, J., Mathur, M., May, M., Mechin, N., Mehta, P., Meixner, J., Melinger, A., Miller, J.K., Miller, M., Moore, K., Möschl, M., Motyl, M., Müller, S.M., Munafo, M., Neijenhuijs, K.I., Nervi, T., Nicolas, G., Nilsonne, G., Nosek, B.A., Nuijten, M.B., Olsson, C., Osborne, C., Ostkamp, L., Pavel, M., Penton-Voak, I.S., Perna, O., Pernet, C., Perugini, M., Pipitone, R.N., Pitts, M., Plessow, F., Prenoveau, J.M., Rahal, R.M., Ratliff, K.A., Reinhard, D., Renkewitz, F., Ricker, A.A., Rigney, A., Rivers, A.M., Roebke, M., Rutchick, A.M., Ryan, R.S., Sahin, O., Saide, A., Sandstrom, G.M., Santos, D., Saxe, R., Schlegelmilch, R., Schmidt, K., Scholz, S., Seibel, L., Selterman, D.F., Shaki, S., Simpson, W.B., Sinclair, H.C., Skorinko, J.L.M., Slowik, A., Snyder, J.S., Soderberg, C., Sonnleitner, C., Spencer, N., Spies, J.R., Steegen, S., Stieger, S., Strohminger, N., Sullivan, G.B., Talhelm, T., Tapia, M., Te Dorsthorst, A., Thomae, M., Thomas, S.L., Tio, P., Traets, F., Tsang, S., Tuerlinckx, F., Turchan, P., Valášek, M., Van't Veer, A.E., Van Aert, R., Van Assen, M., Van Bork, R., Van De Ven, M., Van Den Bergh, D., Van Der Hulst, M., Van Dooren, R., Van Doorn, J., Van Renswoude, D.R., Van Rijn, H., Vanpaemel, W., Echeverría, A.V., Vazquez, M., Velez, N., Vermue, M., Verschoor, M., Vianello, M., Voracek, M., Vuu, G., Wagenmakers, E.J., Weerdmeester, J., Welsh, A., Westgate, E.C., Wissink, J., Wood, M., Woods, A., Wright, E., Wu, S., Zeelenberg, M., Zuni, K., 2015. Estimating the reproducibility of psychological science. Science 349 (80). https://doi.org/10.1126/science.aac4716 aac4716–aac4716.

Acar, F., Seurinck, R., Eickhoff, S.B., Moerkerke, B., 2018. Assessing robustness against potential publication bias in Activation Likelihood Estimation (ALE) meta-analyses for fMRI. PloS One 13, e0208177. https://doi.org/10.1371/journal.pone.0208177.

Baker, M., 2015. Reproducibility crisis: blame it on the antibodies. Nature 521, 274–276. https://doi.org/10.1038/521274a.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. https://doi.org/10.1038/533452a.

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in science. Circ. Res. 116, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bossier, H., Seurinck, R., Kühn, S., Banaschewski, T., Barker, G.J., Bokde, A.L.W.W., Martinot, J.-L., Lemaitre, H., Paus, T., Millenet, S., Moerkerke, B., 2018. The influence of study-level inference models and study set size on coordinate-based fMRI meta-analyses. Front. Neurosci. 11, 1–22. https://doi.org/10.3389/fnins.2017.00745.

Brammer, M.J., Bullmore, E.T., Simmons, A., Williams, S.C.R., Grasby, P.M., Howard, R.J., Woodruff, P.W.R., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn. Reson. Imaging 15, 763–770. https://doi.org/10.1016/S0730-725X(97)00135-5.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M.J., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imag. 18, 32–42. https://doi.org/10.1109/42.750253.

Button, K.S., Ioannidis, J.P. a, Mokrysz, C., Nosek, B. a, Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. https://doi.org/10.1038/nrn3475.

Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. Neuroimage 63, 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004.

Collins, F.S., Tabak, L.A., 2014. NIH plans to enhance REproducibility in clinical trials. Nature 505, 612–613. https://doi.org/10.1007/978-1-4471-3719-1_6.

Conroy, B.R., Walz, J.M., Sajda, P., 2013. Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fMRI decoding models. PloS One 8, e79271. https://doi.org/10.1371/journal.pone.0079271.

Costafreda, S., 2009. Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. Front. Neuroinf. 3, 33. https://doi.org/10.3389/neuro.11.033.2009.

Costafreda, S.G., Brammer, M.J., Vêncio, R.Z.N., Mourão, M.L., Portela, L.A.P., de Castro, C.C., Giampietro, V.P., Amaro, E., 2007. Multisite fMRI reproducibility of a motor task using identical MR systems. J. Magn. Reson. Imag. 26, 1122–1126. https://doi.org/10.1002/jmri.21118.

Cremers, H.R., Wager, T.D., Yarkoni, T., 2017. The relation between statistical power and inference in fMRI. PloS One 12, e0184923. https://doi.org/10.1371/journal.pone.0184923.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302. https://doi.org/10.2307/1932409.

Durnez, J., Degryse, J., Moerkerke, B., Seurinck, R., Sochat, V., Poldrack, R., Nichols, T., 2016. Power and sample size calculations for fMRI studies based on the prevalence of active peaks. bioRxiv 049429. https://doi.org/10.1101/049429.

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. Unit. States Am. 113, 7900–7905. https://doi.org/10.1073/pnas.1602413113.

Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., Lauriello, J., Lim, K.O., Cannon, T., Greve, D.N., Bockholt, H.J., Belger, A., Mueller, B., Doty, M.J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S.G., 2008. Test-retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29, 958–972. https://doi.org/10.1002/hbm.20440.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220. https://doi.org/10.1002/hbm.460010306.

Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test-retest reliability in functional MR imaging. I: statistical methodology. Magn. Reson. Med. 38, 497–507.

Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013. Single subject fMRI test–retest reliability metrics and confounding factors. Neuroimage 69, 231–243. https://doi.org/10.1016/j.neuroimage.2012.10.085.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S.S., Maumet, C., Sochat, V.V., Nichols, T.E., Poldrack, R.A., Poline, J.-B., Yarkoni, T., Margulies, D.S., 2015. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front. Neuroinf. 9, 1–9. https://doi.org/10.3389/fninf.2015.00008.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2, e124. https://doi.org/10.1371/journal.pmed.0020124.

John, L.K., Loewenstein, G., Prelec, D., 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol. Sci. 23, 524–532. https://doi.org/10.1177/0956797611430953.

Lee, J.N., Hsu, E.W., Rashkin, E., Thatcher, J.W., Kreitschitz, S., Gale, P., Healy, L., Marchand, W.R., 2010. Reliability of fMRI motor tasks in structures of the corticostriatal circuitry: implications for future studies and circuit function. Neuroimage 49, 1282–1288. https://doi.org/10.1016/j.neuroimage.2009.09.072.

Liou, M., Su, H.-R., Lee, J.-D., Cheng, P.E., Huang, C.-C., Tsai, C.-H., 2003. Bridging functional MR images and scientific inference: reproducibility maps. J. Cognit. Neurosci. 15, 935–945. https://doi.org/10.1162/089892903770007326.

Liou, M., Su, H.-R., Lee, J.-D., Aston, J.A.D., Tsai, A.C., Cheng, P.E., 2006. A method for generating reproducible evidence in fMRI studies. Neuroimage 29, 383–395. https://doi.org/10.1016/j.neuroimage.2005.08.015.

Machielsen, W.C.M., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Witter, M.P., 2000. fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. 9, 156–164. https://doi.org/10.1002/(SICI)1097-0193(200003)9:3<156::AID-HBM4>3.0.CO;2-Q.

Maitra, R., 2010. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. Neuroimage 50, 124–135. https://doi.org/10.1016/j.neuroimage.2009.11.070.

Muggeo, V.M.R., 2003. Estimating regression models with unknown break-points. Stat. Med. 22, 3055–3071. https://doi.org/10.1002/sim.1545.

Mumford, J.A., 2012. A power calculation guide for FMRI studies. Soc. Cognit. Affect Neurosci. 7, 738–742. https://doi.org/10.1093/scan/nss059.
Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261–268. https://doi.org/10.1016/j.neuroimage.2007.07.061.
Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.P.A., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1, 0021 https://doi.org/10.1038/s41562-016-0021.
Nee, D.E., 2018. fMRI replicability depends upon sufficient individual-level data. bioRxiv 2, 352633. https://doi.org/10.1101/352633.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25. https://doi.org/10.1002/hbm.1058.
Pajula, J., Tohka, J., 2016. How many is enough? Effect of sample size in inter-subject correlation analysis of fMRI. Comput. Intell. Neurosci. 2016, 1–10. https://doi.org/10.1155/2016/2094601.
Patil, P., Peng, R.D., Leek, J., 2016. A statistical definition for reproducibility and replicability. https://doi.org/10.1101/066803, 53, 1689-1699.
Peng, R.D., 2011. Reproducible research in computational science. Science 334 (80), 1226–1227. https://doi.org/10.1126/science.1213847.
Pernet, C., Poline, J.-B., 2015. Improving functional magnetic resonance imaging reproducibility. GigaScience 4, 15. https://doi.org/10.1186/s13742-015-0055-8.
Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., Dehaene, S., 2007. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neurosci. 8, 91. https://doi.org/10.1186/1471-2202-8-91.
Plesser, H.E., 2018. Reproducibility vs. Replicability: a brief history of a confused terminology. Front. Neuroinf. 11, 1–4. https://doi.org/10.3389/fninf.2017.00076.
Poldrack, R.A., Halchenko, Y.O., Hanson, S.J., 2009. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol. Sci. 20, 1364–1372. https://doi.org/10.1111/j.1467-9280.2009.02460.x.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, 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. https://doi.org/10.1038/nrn.2016.167.
Poline, J.-B., Breeze, J.L., Ghosh, S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Haselgrove, C., Helmer, K.G., Keator, D.B., Marcus, D.S., Poldrack, R.A., Schwartz, Y., Ashburner, J., Kennedy, D.N., 2012. Data sharing in neuroimaging research. Front. Neuroinf. 6, 9. https://doi.org/10.3389/fninf.2012.00009.
Qiu, X., Xiao, Y., Gordon, A., Yakovlev, A., 2006. Assessing stability of gene selection in microarray data analysis. BMC Bioinf. 7, 50. https://doi.org/10.1186/1471-2105-7-50.
Rath, J., Wurnig, M., Fischmeister, F., Klinger, N., Höllinger, I., Geißler, A., Aichhorn, M., Foki, T., Kronbichler, M., Nickel, J., Siedentopf, C., Staffen, W., Verius, M., Golaszewski, S., Koppelstaetter, F., Auff, E., Felber, S., Seitz, R.J., Beisteiner, R., 2016. Between- and within-site variability of fMRI localizations. Hum. Brain Mapp. 37, 2151–2160. https://doi.org/10.1002/hbm.23162.
Roels, S.P., Bossier, H., Loeys, T., Moerkerke, B., 2015. Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis. J. Neurosci. Methods 240, 37–47. https://doi.org/10.1016/j.jneumeth.2014.10.024.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638–641. https://doi.org/10.1037/0033-2909.86.3.638.
Salimi-Khorshidi, G., Smith, S.M., Keltner, J.R., Wager, T.D., Nichols, T.E., 2009. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. Neuroimage 45, 810–823. https://doi.org/10.1016/j.neuroimage.2008.12.039.
Samartsidis, P., Montagna, S., Laird, A.R., Fox, T., Johnson, T.D., Nichols, T.E., 2017. Estimating the number of missing experiments in a neuroimaging meta-analysis. https://doi.org/10.1101/225425.
Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.-L., Paus, T., Poline, J.-B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Ströhle, A., Struve, M., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatr. 15, 1128–1139. https://doi.org/10.1038/mp.2010.4.
Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.
Sochat, V.V., Gorgolewski, K.J., Koyejo, O., Durnez, J., Poldrack, R.A., 2015. Effects of thresholding on correlation-based image similarity metrics. Front. Neurosci. 9, 1–8. https://doi.org/10.3389/fnins.2015.00418.
Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34.
Sterling, T.D., 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. J. Am. Stat. Assoc. 54, 30–34. https://doi.org/10.1080/01621459.1959.10501497.
Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 https://doi.org/10.1371/journal.pmed.1001779.
Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.-B., 2007. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. Neuroimage 35, 105–120. https://doi.org/10.1016/j.neuroimage.2006.11.054.
Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1, 62. https://doi.org/10.1038/s42003-018-0073-z.
Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The Wu-minn human connectome project: an overview. Neuroimage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
Wager, T.D., Lindquist, M., Kaplan, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. Soc. Cognit. Affect Neurosci. 2, 150–158. https://doi.org/10.1093/scan/nsm015.
Wilke, M., 2012. An iterative jackknife approach for assessing reliability and power of fMRI group Analyses. PloS One 7, e35578. https://doi.org/10.1371/journal.pone.0035578.
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060.
Woo, C.-W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058.
Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S.M., 2009. Bayesian analysis of neuroimaging data in FSL. Neuroimage 45, S173–S186. https://doi.org/10.1016/j.neuroimage.2008.10.055.
Wu, C.C., Samanez-Larkin, G.R., Katovich, K., Knutson, B., 2014. Affective traits link to reliable neural markers of incentive anticipation. Neuroimage 84, 279–289. https://doi.org/10.1016/j.neuroimage.2013.08.055.