Contents lists available at ScienceDirect

NeuroImage

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

A model-based approach to assess reproducibility for large-scale high-throughput MRI-based studies[☆]

Zeyu Jiao[a,b,c,d,e,1], Yinglei Lai[f,1], Jujiao Kang[a,b,c,d,e], Weikang Gong[j], Liang Ma[k], Tianye Jia[b,c,d,e,l], Chao Xie[b,c,d,e], Shitong Xiang[b,c,d,e], Wei Cheng[b,c,d,e], Andreas Heinz[m], Sylvane Desrivières[l], Gunter Schumann[b,l,n], IMAGEN Consortium, Fengzhu Sun[g], Jianfeng Feng[a,b,c,d,e,h,i,*]

[a] *Shanghai Center for Mathematical Sciences, Fudan University, 220 Handan Road, Shanghai, China*
[b] *Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China*
[c] *Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, China*
[d] *MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China*
[e] *Zhangjiang Fudan International Innovation Center, China*
[f] *School of Mathematical Sciences, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui 230026, China*
[g] *Quantitative and Computational Biology Department, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, United States*
[h] *Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom*
[i] *School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai, China*
[j] *Center for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Welcome Center for Integrative Neuroimaging, University of Oxford, Oxford, United Kingdom*
[k] *Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*
[l] *Center for Population Neuroscience and Precision Medicine (PONS), Institute of Psychiatry, Psychology and Neuroscience, SGDP Center, King's College London, United Kingdom*
[m] *Department of Psychiatry and Psychotherapy CCM, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*
[n] *PONS Research Group, Department of Psychiatry and Psychotherapy, Campus Charite Mitte, Humboldt University, Berlin, Germany*

ARTICLE INFO

Keywords:
Reproducibility
Association studies
MRI (magnetic resonance imaging)
Sample size
Heterogeneity

ABSTRACT

Magnetic Resonance Imaging (MRI) technology has been increasingly used in neuroscience studies. Reproducibility of statistically significant findings generated by MRI-based studies, especially association studies (phenotype vs. MRI metric) and task-induced brain activation, has been recently heavily debated. However, most currently available reproducibility measures depend on thresholds for the test statistics and cannot be use to evaluate overall study reproducibility. It is also crucial to elucidate the relationship between overall study reproducibility and sample size in an experimental design. In this study, we proposed a model-based reproducibility index to quantify reproducibility which could be used in large-scale high-throughput MRI-based studies including both association studies and task-induced brain activation. We performed the model-based reproducibility assessments for a few association studies and task-induced brain activation by using several recent large sMRI/fMRI databases. For large sample size association studies between brain structure/function features and some basic physiological phenotypes (i.e. Sex, BMI), we demonstrated that the model-based reproducibility of these studies is more than 0.99. For MID task activation, similar results could be observed. Furthermore, we proposed a model-based analytical tool to evaluate minimal sample size for the purpose of achieving a desirable model-based reproducibility. Additionally, we evaluated the model-based reproducibility of gray matter volume (GMV) changes for UK Biobank (UKB) vs. Parkinson Progression Marker Initiative (PPMI) and UK Biobank (UKB) vs. Human Connectome Project (HCP). We demonstrated that both sample size and study-specific experimental factors play important roles in the model-based reproducibility assessments for different experiments. In summary, a systematic assessment of reproducibility is fundamental and important in the current large-scale high-throughput MRI-based studies.

[☆] A complete list of The IMAGEN Consortium can be found in Supplementary Material.

[*] Corresponding author at: Shanghai Center for Mathematical Sciences, Fudan University, 220 Handan Road, Shanghai, China.
E-mail address: jianfeng64@gmail.com (J. Feng).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2022.119166.
Received 11 May 2021; Received in revised form 26 March 2022; Accepted 30 March 2022
Available online 6 April 2022.

1. Introduction

Magnetic Resonance Imaging (MRI) technology has been widely used in neuroscience (Poldrack and Gorgolewski, 2014). It enables us to conduct experiments on gray matter volume (GMV) changes (structure MRI), task-free scanning (resting state fMRI) and task-based studies (task fMRI) (Ashburner and Friston, 2000; Logothetis, 2008; Snyder and Raichle, 2012). Reproducibility, or results reproducibility (Goodman et al., 2016) for MRI-based studies especially association studies (phenotype vs. MRI metric) and task-induced brain activation has recently received a significant attention. Criticisms have been raised to the phenomena that some MRI-based findings are only modestly reproducible and that some results could be interpreted as inflated or spurious (Nature Neuroscience Editorial, 2017; Bennett and Miller, 2010; Botvinik-Nezer et al., 2020; Eklund et al., 2016). These debates were mostly on the reproducibility of novel discoveries (i.e. findings with statistical significance). The reproducibility of the multi-model MRI data and the reproducibility of novel discoveries were reported, including the reproducibility of functional MRI (Bosnell et al., 2008; Chen et al., 2017; Conti et al., 2019; Tegeler et al., 1999; Zou et al., 2005). However, to our acknowledge, there is a lack of effective statistical model to assess overall study reproducibility in MRI-based studies. A model-based reproducibility assessment can provide us with an adequate confidence in MRI-based research outcomes.

Model-based reproducibility assessment approach aims to evaluate the level of directional concordance among different analysis results (i.e. z-score). Accordingly, a mixture model based approach has been proposed to conduct hypothesis testing on the reproducibility of mass spectrometry studies (Lai et al., 2007). Recently, a Bayesian modeling approach has also been proposed to address the irreproducibility of genome-wide association studies or transcriptome-wide association studies (Zhao et al., 2020). For MRI-based studies, there is still a lack of effective approach to systematic evaluate overall study reproducibility.

In this study, we proposed a model-based reproducibility index to quantify reproducibility for MRI-based association studies (phenotype vs. MRI metric) and task-induced brain activation. To evaluate the performance of our proposed model-based reproducibility index in MRI-based studies, we first present a comprehensive simulation study which is designed based on UK Biobank structure MRI data (Alfaro-Almagro et al., 2018; Sudlow et al., 2015). Then, we present model-based reproducibility assessments of several GMV-related human phenotypes, brain task state activation and connectivity-phenotype studies by using recent large sMRI/fMRI databases. For each database, we focused on a few phenotypes or task types. Furthermore, with a desirable model-based reproducibility requirement, we present the related sample size calculations for several MRI-based study scenarios. These can be achieved with UK Biobank structure and resting-state functional MRI data and IMAGEN task functional MRI data (Bossier et al., 2020; Schumann et al., 2010b). Our model-based analytical tool to evaluate minimal sample size could provide a useful guidance in the related future study planning. Moreover, we demonstrate that the reproducibility between two independent MRI databases can also be evaluated by using our model-based reproducibility approach.

2. Materials and methods

2.1. Study participants

2.1.1. UK Biobank

UK Biobank (Alfaro-Almagro et al., 2018; Sudlow et al., 2015) is a prospective epidemiological resource gathering extensive questionnaires, physical and cognitive measures and biological samples (including genotype), in a cohort of 500,000 participants (Sudlow et al., 2015). Participants which years of age between 40 and 69 at baseline recruitment consent to access to their full health records from the UK National Health Service, enabling researchers to relate phenotypic measures to long-term health outcomes. They also provided blood, urine and saliva samples, which were stored in such a way as to allow many different types of assay to be performed (for example, genetic, proteomic and metabolomics analyzes). In 2014, UK Biobank began the process of inviting back 100,000 of the original volunteers for brain, heart and body imaging. The UK Biobank project received ethical approval from the Research Ethics Committee within the terms of an Ethics and Governance Framework. The initial release of 10,000 UK Biobank imaging and behavioral measures data was used in our manuscript and more details are available online (Alfaro-Almagro et al., 2018).

2.1.2. IMAGEN

One thousand five hundred and six adolescents (mean age = 14.44 y old; SD = 0.42; range = 12.88–16.44 y old) from the baseline assessment of the IMAGEN (Bossier et al., 2020; Schumann et al., 2010b) sample with complete data in fMRI and behavioral measurements were included in the analyzes. Written informed consent and assent had been given by both parents and participants. The study had been approved by the local ethic committees. Detailed descriptions of this study have previously been published (Schumann et al., 2010a).

2.1.3. PPMI

The Parkinson Progression Marker Initiative (PPMI) is a comprehensive observational, international, multi-center study designed to identify PD progression biomarkers both to improve understanding of disease etiology and course and to provide crucial tools to enhance the likelihood of success of PD modifying therapeutic trials. The PPMI cohort will comprise 400 recently diagnosed PD and 200 healthy subjects followed longitudinally for clinical, imaging and biospecimen biomarker assessment using standardized data acquisition protocols at twenty-one clinical sites. All procedures were performed with prior approval from ethical standards committees and with informed consent from all study participants. The PPMI dataset that we could available include T1 structural MRI scans for 543 participants ($n = 374$ patients with PD, $n = 169$ healthy controls). To facilitate the comparison of overall reproducibility between different applications, we only used healthy controls in this cohort ($n = 136$, age range 45–79 to match the UKB dataset age range). For more details of this dataset, see the previous related publication (Marek et al., 2011).

2.1.4. HCP

The WU-Minn HCP consortium (Van Essen et al., 2012) aims to characterize human brain in a population of 1200 healthy adults and to enable detailed comparisons between brain circuits, behavior, and genetics at the level of individual subjects. The HCP was reviewed and approved by the Institutional Ethics Committee of Washington University in St. Louis, Missouri. All participants signed written informed consent. Here, we use the T1 data form this project and more details they were initially reported (Van Essen et al., 2013).

2.2. MRI acquisition

2.2.1. UK Biobank

Details of the image acquisition in UK Biobank are also available online (Alfaro-Almagro et al., 2018). Magnetic resonance imaging (MRI) was performed using a Siemens Skyra 3T running VD13ASP4 (Siemens Healthcare, Erlangen, Germany) with a Siemens 32-channel RF receive head coil. The T1 structural protocol is acquired at 1 mm isotropic resolution using a three-dimensional (3D) MPRAGE acquisition, with inversion and repetition times optimized for maximal contrast. The superior-inferior field-of-view is large (256 mm), at little cost, in order to include reasonable amounts of neck/mouth, as those areas will be of interest to some researchers (for example, in the study of sleep apnea).

Resting-state fMRI use the same acquisition parameters, with 2.4-mm spatial resolution and TR = 0.735 s, with multiband acceleration factor. A 'single band' reference image (without the multiband excitation, exciting each slice independently) is acquired that has higher tissue-type image contrast; this is used as the target for motion correction and alignment. For both databases, the raw data are corrected for motion and distortion and high-pass filtered to remove temporal drift.

2.2.2. IMAGEN

Structural MRI and fMRI data were acquired at eight IMAGEN assessment sites with 3-T MRI scanners of different manufacturers (Siemens, Philips, General Electric, and Bruker). The scanning variables were specially chosen to be compatible with all scanners. The same scanning protocol was used in all cites. In brief, high-resolution T1-weighted 3D structural images were acquired for anatomical localization and coregistration with the functional time series. BOLD functional images were acquired with a gradient echo, echo planar imaging sequence. 300 vol were acquired for each participant, and each volume consisted of 40 slices aligned to the anterior commission/posterior commission line (2.4 mm slice thickness and 1 mm gap). The echo time was optimized (echo time = 30 ms; repetition time = 2200 ms) to provide reliable imaging of subcortical areas. (More details for different task see Supplementary Material)

2.2.3. PPMI

In this research, we use the MRI data acquired by the PPMI study, in which a T1-weighted, 3D sequence (i.e., MPRAGE) is acquired for each subject using 3T SIEMENS MAGNETOM TrioTim syngo scanners. This gives us 374 PD and 169 NC scans. The T1-weighted images were acquired for 176 sagittal slices, with the following parameters: repetition time (TR) = 2300 ms, echo time (TE) = 2.98 ms, flip angle = 9°, and voxel size = $1 \times 1 \times 1$ mm^3.

2.2.4. HCP

The Human Connectome Project (HCP) provides a unique, open source, large-scale collection of about 1200 human head T1 image datasets and we employed 413 healthy subjects which have no clear family related (age-range: 22–36 years) in our study. All HCP imaging data were acquired on a Siemens Skyra 3T scanner with a customized SC72 gradient insert. T1w 3D MPRAGE were acquired with TR = 2400 ms, TE = 2.14 ms, TI = 1000 ms, flip angle = 8 deg, FOV = 224 × 224, 0.7 mm isotropic voxel, bandwidth = 210 Hz/px, iPAT = 2, Acquisition time = 7:40 (min:sec).

2.3. MRI preprocessing

The rs-fMRI data are preprocessed using standard volume-based fMRI pipeline. For each subject, the preprocessing steps include: motion correction (FSL mcflirt), despiking motion artifacts using Brain Wavelet Toolbox (Patel et al., 2014), registering to $3 \times 3 \times 3$ mm^2 standard space by first aligning the functional image to the individual T1 structure image using boundary based registration (Greve and Fischl, 2009) and then to standard space using FSL's linear and non-linear registration tool (FSL flirt and fnirt), regressing out nuisance covariates including Friston-24 parameters, white matter signal, cerebrospinal fluid signal, band-pass filtering (0.01–0.1 Hz) using AFNI (3dTproject) and spatial smoothing by a 3D Gaussian kernel (FWHM = 6 mm). All the images are manually checked to ensure successful preprocessing and insure the mean FD Power not greater than 0.5. After above preprocessing, a large sample size imaging and behavioral measures data which contain 8273 subjects have been used in this study.

T1 data were preprocessed with the voxel-based morphometry (VBM) by using the VBM8 toolbox based on the Statistical Parametric Mapping package (SPM). Firstly, all structural MRI data were manually corrected and divided into gray matters, white matters and cerebrospinal fluid. Secondly, the gray matter images were aligned to a nonlinear deformation field and normalized to Montreal Neurological Institute (MNI) space by using the templates which were created by DARTEL tool. Finally, the normalized images were all smoothed with a full-width at half-maximum (FWHM) 6-mm Gaussian kernel for further analysis. After above procedures, the gray matter images (voxel size: $3 \times 3 \times 3$ mm^2) were obtained for 9850 subjects.

Task-fMRI data were analyzed with SPM. Spatial preprocessing included slice time correction to adjust for time differences caused by multi-slice imaging acquisition, realignment to the first volume in line, nonlinearly warping to the Montreal Neurological Institute space [based on a custom echo planar imaging template ($53 \times 63 \times 46$ mm^3 voxels) created out of an average of the mean images of 400 adolescents], resampling at a resolution of $3 \times 3 \times 3$ mm^3, and smoothing with an isotropic Gaussian kernel of 5-mm FWHM.

2.4. Statistical analyzes

2.4.1. Functional connectivity association study

Based on the automated anatomical labeling (AAL2) atlas, there are 120 brain regions. Each resting-state functional magnetic resonance image (rs-fMRI) included 54,885 voxels (Rolls et al., 2015). For each pair of brain regions, the time series were extracted, and the Pearson correlation was calculated for each subject to provide the measure of functional connectivity (FC), followed by Fisher's z-transformation. The general linear model was used to test the association between the region-wise FC links and a human phenotype or behavior. The effects of age, sex and head motion (mean frame-wise displacement) were regressed out.

2.4.2. Voxel-wise association study

We used the general linear model to define the association between a specific human phenotype or behavior and each intracerebral voxel's gray matter volume, which was included in the automated anatomical labeling (AAL2) atlas (total 54,885 voxels). The effects of age, sex and total intracerebral volume (TIV) were regressed out.

2.4.3. Task fMRI activation

At the first level of analysis, changes in the BOLD response for each subject were assessed by linear combinations at the individual subject level for each experimental condition (e.g. reward anticipation high gain of Monetary Incentive Delay (MID) task), and each trial was convolved with the hemodynamic response function to form regressors that account for potential noise variance (e.g., head movement) associated with the processing of a specific task. Estimated movement parameters were added to the design matrix in the form of 18 additional columns (three translations, three rotations, three quadratic and three cubic translations, and three translations each with a shift of ±1 repetition time). To identify brain activation specific to the task, we contrasted the brain activation patterns between the task status and the control status.

For the MID anticipation phase we contrasted brain activation during 'anticipation of high win [here signaled by a circle] vs anticipation of no-win [here signaled by a triangle]'; For the emotional faces task (EFT) we contrasted brain activation during 'viewing Angry Face vs viewing Control [circles]'; For the stop signal task (SST) we contrasted brain activation during 'successful stop vs successful go'.

2.4.4. Normal distribution quantile-based transformation

z-scores from a normal distribution quantile transformation were used for the analysis (Lai et al., 2007). First, based on an appropriate analysis (functional connectivity association study, voxel-wise association study or task fMRI activation), we acquired a list of one-sided *P*-values. For each *P*-value *P*, the corresponding *z*-score *z* can be calculated as follows:

$$z = \phi^{-1}(1 - P)$$

where $\phi^{-1}(\cdot)$ is the inverse function of the standard normal cumulative distribution function.

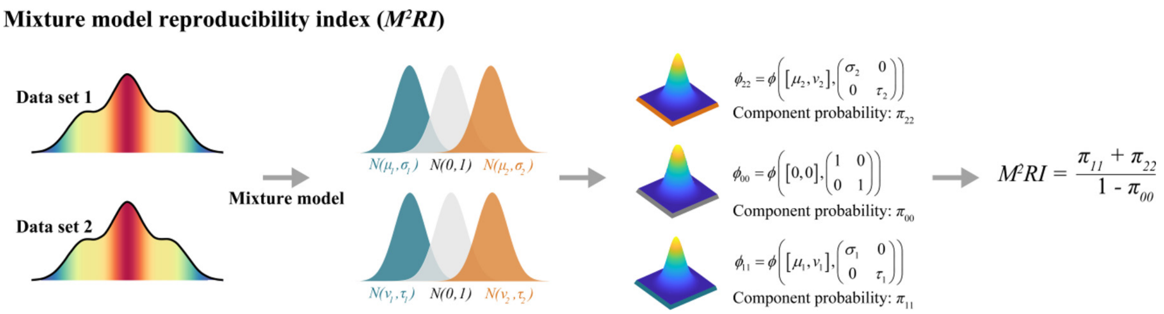
Mixture model reproducibility index (M^2RI)
Data set 1
Data set 2
Mixture model
$N(\mu_1,\sigma_1)$ $N(0,1)$ $N(\mu_2,\sigma_2)$
$N(\nu_1,\tau_1)$ $N(0,1)$ $N(\nu_2,\tau_2)$
$\phi_{22}=\phi\left([\mu_2,\nu_2],\begin{pmatrix}\sigma_2 & 0\\ 0 & \tau_2\end{pmatrix}\right)$
Component probability: π_{22}
$\phi_{00}=\phi\left([0,0],\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$
Component probability: π_{00}
$\phi_{11}=\phi\left([\mu_1,\nu_1],\begin{pmatrix}\sigma_1 & 0\\ 0 & \tau_1\end{pmatrix}\right)$
Component probability: π_{11}
$M^2RI=\frac{\pi_{11}+\pi_{22}}{1-\pi_{00}}$

Fig. 1. An illustration of ***M*²*RI***. $\phi_{i,j}$ is the normal probability distribution function and π_{ij} is the proportion of features consistent with component i in the first analysis and component j in the second analysis.

2.4.5. Definition of model-based reproducibility index

We firstly consider a nine-component normal-mixture model for the joint distribution of paired z-scores $[z^{(1)},\ z^{(2)}]$ (see above for z-score calculation).

$$f\left[z^{(1)},z^{(2)}\right]=\sum\nolimits_{i=0}^{2}\sum\nolimits_{j=0}^{2}\pi_{ij}\phi_{\mu_i,\sigma_i^2}\left[z^{(1)}\right]\phi_{\nu_j,\tau_j^2}\left[z^{(2)}\right]$$

where ϕ_{μ,σ^2} is the normal probability distribution function with mean μ and variance σ^2. We use the first component (index 0) to represent the null (no change/correlation) feature component. Then, $\mu_0=\nu_0=0$ and $\sigma_0^2=\tau_0^2=1$. The second and third components (indices 1 and 2) are used to represent negative changes/correlations and positive changes/correlations. Their corresponding parameters (means and variances) will be estimated from the paired z-scores with the following constraints: $\mu_1,\nu_1\le 0$ and $\mu_2,\nu_2\ge 0$. π_{ij} is the proportion for component i in the first study and component j in the second study, and $\sum_{ij}\pi_{ij}=1$.

This model was termed partial concordance/discordance (PCD) model and more details for PCD model could be available in Supplementary Material (Lai et al., 2007, 2009, 2014, 2017). Then, we define a model-based reproducibility index based on model parameters, mixture model reproducibility index (*M*²*RI*) (The illustration of *M*²*RI* see Fig. 1):

$$M^2RI=\frac{\pi_{11}+\pi_{22}}{1-\pi_{00}}$$

In a recent study (Zhao et al., 2020), two Bayesian models: curved exponential family normal prior model (CEFN) and meta-analysis prior model (META), have been proposed for a similar purpose. In these two models, π_{null} and π_R were the proportions of null and reproducible signals, respectively. Accordingly, we may define the related model-based reproducibility indices:

$$CEFNRI/METARI=\frac{\pi_R}{1-\pi_{null}}.$$

2.4.6. Confidence intervals of M²RI

The confidence intervals (CIs) of *M*²*RI* can be obtained by bootstrapping paired z-scores (Efron and Tibshirani, 1997; Mclachlan, 1987). For our newly developed model-based reproducibility index *M*²*RI*, a theoretical confidence interval will also be highly useful in practice. Therefore, we have derived the asymptotic theoretical CIs for *M*²*RI* based on our proposed mixture model (see Supplementary Material for details).

2.5. Simulations

We conducted a comprehensive simulation study to show the performance of our newly proposed model-based reproducibility index. Our simulations were designed based on the gray matter volume (GMV) data in the UK Biobank. Two-sample comparison is a general analysis scenario in practice, and the reproducibility of a large-scale two-sample study is important. Therefore, we partitioned the whole data randomly into four subsets which have the same sample size (referred to as Data 1A, Data 1B, Data 2A and Data 2B, each subset includes 1/4 sample of the whole data). In our simulations, 1 and 2 represent virtually the case group and the control group, respectively. A and B represent virtually different experiments. Before the analysis, as a widely considered practical approach, we firstly checked that the covariates such as sex, age, total intracerebral volume (TIV) were statistically similar (two-sample *t*-test for age and TIV, chi-square test for sex, $P>0.05$) between Data 1A vs. 2A as well as Data 1B vs. 2B. Then we also checked that total GMV (i.e. dependent variable) was statistically similar (two-sample *t*-test, $P>0.05$) between Data 1A vs. 2A as well as Data 1B vs. 2B. Otherwise, we repeated the random data partition until one passed these similarity requirements. For each feature, there was statistically no differences in distribution between Data 1A vs. 2A nor Data 1B vs. 2B. Then, to generate upward or downward changes, a specified proportion of voxels in a cluster were randomly chosen and 0.0285–0.0855 standard deviations of brain-wise GMV (corresponding to approximately 1–3 effect sizes in z-scores) were randomly added to (or subtracted from) the chosen voxels of each subject in Data 1A and Data 1B. This procedure was repeated 1000 times. For each repetition, we obtained two lists of z-scores: one by voxel-wisely comparing Data 1A vs. Data 2A and the other Data 1B vs. Data 2B. z-scores were calculated based on the traditional two-sample *t*-test. A pair of z-scores were obtained for each voxel. The reproducibility between two lists of z-scores was assessed by our proposed model-based reproducibility index. The following three simulation scenarios were considered.

(a) *Complete reproducible with a moderate proportion of changes*. According to our random data partition, there were statistically no differences between Data 1A vs. 2A nor Data 1B vs. 2B. We modified the 100% of null (no change) to 80% null, 10% upward changes and 10% downward changes as follows. We randomly selected two clusters of voxels, each with 10% of the total voxels. To simulate 10% upward changes, for each voxel in the first cluster of voxels, we randomly added to each subject's GMV a value equivalent to 1–3 effect sizes in z-scores in Data 1A and repeated this in Data 1B so that there were 10% reproducible upward changes. For each voxel in the second cluster of voxels, we randomly subtracted from each subject's GMV a value equivalent to 1–3 effect sizes in z-score in Data 1A and repeated this in Data 1B so that there were 10% reproducible downward changes.

(b) *Partial reproducible*. We randomly selected four clusters of voxels. There were 15% of the total voxels in each of the first two clusters, and the upward changes and downward changes were simulated according to the description in (a). There were 5% of the total voxels in each of the next two clusters. For each voxel in the third cluster, we randomly added to each subject's GMV a value equivalent to 1–3 effect sizes in z-scores in Data 1A (but not in Data 1B). Then, we had 5% discordant changes (up vs. null). For each voxel in the fourth cluster, we similarly subtracted from each subject's GMV in Data 1A (but not in Data 1B) so that we had 5% discordant changes (down vs. null).

(c) *Complete reproducible with a high proportion of changes*. Considering that the number of consistent significant results from different studies

Table 1
The performance of model-based reproducibility index in three simulation analysis scenarios, For each simulation analysis scenario, the true reproducibility is shown in the table. The simulation and evaluation were repeated 1000 times to obtain the median, the lower and upper-quartiles (Q1-Q3) for assessed reproducibility. (For more details, please see section *Model-based Reproducibility Index Recovers the True Reproducibility Accurately in the Simulation Study*).

Model-based Reproducibility Index	*Assessed Model-based Reproducibility*
Simulation (a) ***100% reproducibility***	*Median (Q1-Q3)*
M^2RI	0.9150 (0.7378–0.9950)
CEFNRI	0.9951 (0.9932–0.9961)
METARI	0.9879 (0.9506–0.9933)
Simulation (b) ***75% reproducibility***	*Median (Q1-Q3)*
M^2RI	0.7688 (0.6850–0.8488)
CEFNRI	0.9875 (0.9549–0.9948)
METARI	0.8255 (0.6911–0.9092)
*Simulation (c)****100% reproducibility***	*Median (Q1-Q3)*
M^2RI	0.9597 (0.8731–0.9986)
CEFNRI	0.9980 (0.9973–0.9983)
METARI	0.9939 (0.9734–0.9966)

Table 2
Model-based reproducibility assessment of four analysis scenarios, For each MRI-based analysis scenario, the assessed model-based reproducibility is shown in the table. We obtained 95% confidence intervals (CIs) based on bootstrapping the paired *z*-scores for *M^2RI*. The asymptotic theoretical 95% CIs for *M^2RI* was also presented. (For more details, please see section *Model-based Reproducibility Assessments for Large-scale MRI-based Studies*).

Model-based Reproducibility Index	*Assessed Model-based Reproducibility*
Sex as phenotype vs. RSFC	
M^2RI (95% CIs)	0.9999996 (0.9999993–0.9999998)
M^2RI (asymptotic theoretical 95% CIs)	0.9999996 (0.9807–1)
CEFNRI	0.9993
METARI	0.9992
BMI as phenotype vs. RSFC	
M^2RI (95% CIs)	0.9999934 (0.9999904–0.9999955)
M^2RI (asymptotic theoretical 95% CIs)	0.9999934 (0.9782–1)
CEFNRI	0.9986
METARI	0.9984
Sex as phenotype vs. GMV	
M^2RI (95% CIs)	0.99999995 (0.99999993–0.99999997)
M^2RI (asymptotic theoretical 95% CIs)	0.99999995 (0.9955–1)
CEFNRI	0.9996
METARI	0.9995
Activation in MID task	
M^2RI (95% CIs)	0.99999991 (0.99999989–0.99999994)
M^2RI (asymptotic theoretical 95% CIs)	0.99999991 (0.9968–1)
CEFNRI	0.9997
METARI	0.9997

can vary, we randomly selected two clusters of voxels, each with 20% of the total voxels. The reproducible upward changes (the first cluster) and downward changes (the second cluster) were simulated similarly according to the description in (a).

3. Results

3.1. Model-based reproducibility index recovers the true reproducibility accurately in the simulation study

The simulation results are summarized in Table 1. Based on the scenario (a) as complete reproducible with a moderate proportion of changes, the median *M^2RI, CEFNRI* or *METARI* was 0.915, 0.995 or 0.988, respectively. Furthermore, the related lower- and upper-quartiles (Q1-Q3) was 0.738–0.995, 0.993–0.996 or 0.951–0.993, respectively. It was reasonable to conclude that the assessed model-based reproducibility could be up to the true reproducibility which is 100%. Based on the scenario (b) as a partial reproducibility (75%), the median *M^2RI, CEFNRI* or *METARI* was 0.769, 0.988 or 0.826 when the related lower- and upper-quartiles (Q1-Q3) was 0.685–0.849, 0.955–0.995 or 0.691–0.909, respectively. Based on the scenario (c) as complete reproducible with a high proportion of changes, the median *M^2RI, CEFNRI* or *METARI* reached 0.960, 0.998 or 0.994 with the lower- and upper-quartiles (Q1-Q3) 0.873–0.999, 0.997–0.998 or 0.973–0.997, respectively. It was also reasonable to conclude that the assessed model-based reproducibility could be up to the true reproducibility which is 100%.

3.2. Model-based reproducibility assessments for large-scale MRI-based studies

To investigate the reproducibility of large-scale MRI-based analysis in the data collected for studying human phenotypes/behaviors and task state activations, as well as the brain structure and function, we split each study cohort into two subsets (referred to as Group 1 and Group 2 based on the order of subject number) with (approximately) the same sample sizes. As the sample limitations and missing observations, different sample sizes were used to test different aspects of model-based reproducibility.

For the resting-state functional connectivity (RSFC) data, the sample sizes of the two subsets were 4136 and 4137 for analyzing sex as phenotype vs. RSFC; the sample sizes of the two subsets were 4131 and 4131 for analyzing body mass index (BMI) as phenotype vs. RSFC (as there were missing BMI observations). A general linear model was constructed with sex phenotype as the response in each subset, with age and mean FD adjusted as covariates (hereafter referred to as Sex as phenotype vs. RSFC and BMI as phenotype vs. RSFC; see Fig. 2c,d for the paired *z*-scores). For the GMV data, the sample sizes of the two subsets were 4925 and 4925, respectively. A general linear model was also constructed with sex phenotype as the response in each subset, with age and TIV adjusted as covariates (hereafter referred to as Sex as phenotype vs. GMV; see Fig. 2a for the paired *z*-scores). For the task-related activation data, the sample sizes of the two subsets were 772 and 772, respectively. Student's *t*-test was used to evaluate the activation of the monetary incentive delay (MID) task, one of the most common tasks in fMRI studies (this activity is hereafter referred to as Activation in the MID task; see Fig. 2b for the paired *z*-scores). For each paired *z*-scores, an overall diagonal pattern can be clearly observed. Different paired *z*-scores variation patterns can also be observed for different analysis scenarios, which implies different mixtures of no-change related (null) *z*-scores and upward/downward-change related (non-null) *z*-scores. Evaluation of Gaussian mixture model assumption is available in Supplementary Material.

Model-based reproducibility index was used to evaluate the reproducibility based on the paired *z*-scores in Fig. 2. The results are shown in Table 2. We bootstrapped the paired *z*-score to construct the related 95% confidence intervals (CIs) and we also calculated the asymptotic theoretical 95% CIs for *M^2RI*. For Sex as phenotype vs. RSFC, the *CEFNRI* or *METARI* was 0.9993 or 0.9992, respectively. Furthermore, *M^2RI* was close to one, which also suggested an ideal reproducibility. Its asymptotic theoretical 95% CI was above 0.98. For BMI as phenotype vs. RSFC, *CEFNRI* or *METARI* was 0.9986 or 0.9984, respectively. *M^2RI* was still close to one, and its asymptotic theoretical 95% CIs were above 0.97. For Sex as phenotype vs. GMV, *CEFNRI* or *METARI* was 0.9996 or 0.9995, respectively. *M^2RI* was again nearly one and both 95% CIs were nearly ideal. For activation in the MID task, *CEFNRI* or *METARI* was 0.9997 or 0.9997, respectively. *M^2RI* was still nearly one and both 95% CIs were again nearly ideal.

3.3. Sample size needed to achieve a desirable model-based reproducibility

Sample size calculation is crucial in experimental designs. When designing a large-scale analysis, one may ask what sample size is required to achieve a desirable model-based reproducibility requirement. For a

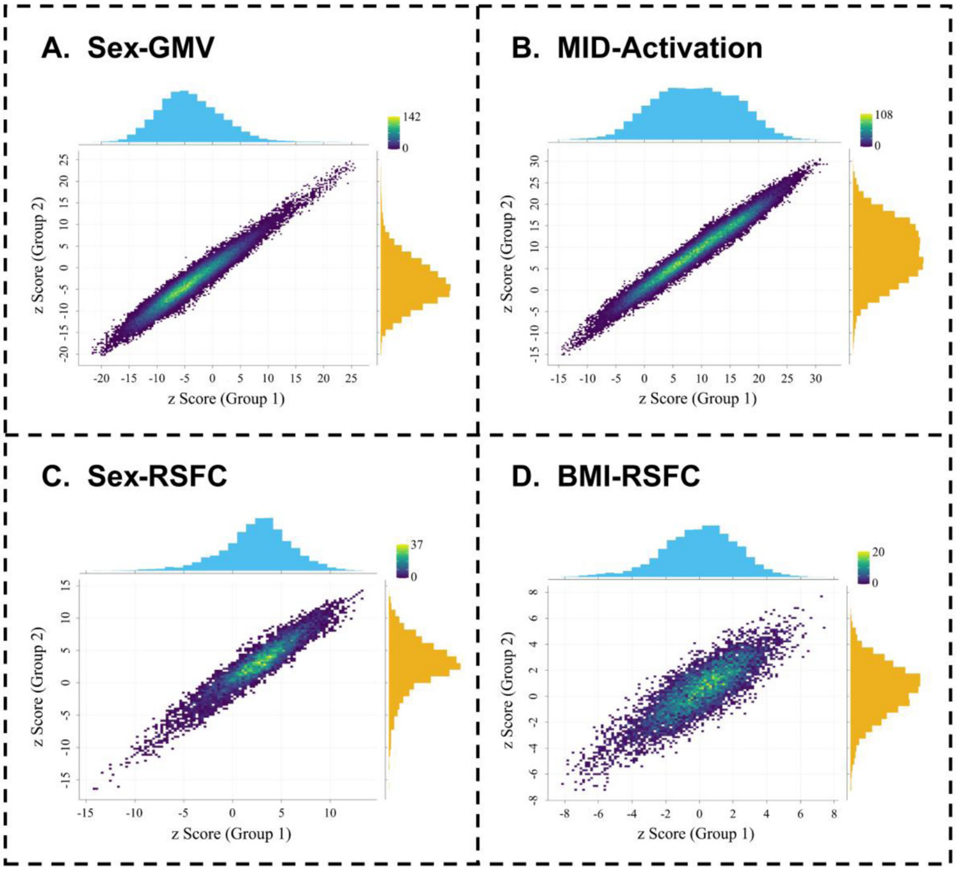

Fig. 2. Paired *z*-scores from four analysis scenarios. (A) Sex as phenotype vs. GMV in UK Biobank data. (B) MID task activation in IMAGEN data. (C) Sex as phenotype vs. RSFC in UK Biobank data. (D) BMI as phenotype vs. RSFC in UK Biobank.

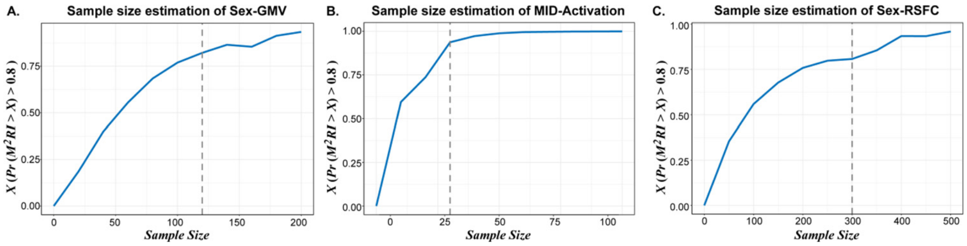

Fig. 3. Sample size calculations for three analysis scenarios in the UK Biobank or IMAGEN data. In each plot, "*X* ($Pr(M^2RI > X) > 0.8$)" means the minimal M^2RI in top 80% resampled repetitions (800 resampled repetitions in total 1000 resampled repetitions). The vertical dashed line indicates the minimum sample size for Pr ($M^2RI > 0.8$) > 0.8. (A) Sex as phenotype vs. GMV in the UK Biobank data. (B) MID task activation in the IMAGEN data. (C) Sex as phenotype vs. RSFC in the UK Biobank data.

comprehensive understanding of sample size requirements in a specific analysis scenario, we proposed a model-based analytical tool based on a large resampling-based simulation study. For a study cohort presented in Table 2, we selected a phenotype available in the study as response. Then, we randomly selected subjects from the cohort to construct two subsets with a given sample size for each subset. This procedure is appropriate for M^2RI. For each given sample size, we repeated the resampling and M^2RI calculation 1000 times.

We evaluated Pr($M^2RI > 0.8$) empirically for each given sample size. Then, we could obtain the minimum sample size to achieve Pr($M^2RI > 0.8$) > 0.8 in each analysis scenario. (In addition to 0.8, other values could be certainly considered, and it is not necessary to always set both values to 0.8.) The results for different analysis scenarios are summarized in Fig. 3 and Table 3. We assessed the minimum sample size for M^2RI. For different response phenotypes in the task-related functional MRI data, the minimum sample size was only approximately 20 to 30. For the GMV data, a sample size of approximately 120 was required when the response was the sex phenotype; a sample size of 70 was required when the response was the age phenotype and a sample size of 300 was required when the response was the BMI phenotype. However, for different response phenotypes in the RSFC data, the results were clearly different. Approximately 200 or 300 were required when the response was the age or sex phenotype, respectively. When the response was BMI, the minimum sample size increased to a very large value (More than 800. When the estimated minimum sample size n close to the sample size of the whole dataset N (e.g. n is more than $\frac{N}{10}$), the minimum sample size estimation could be inaccurate as the sample duplication problem in resampling-based simulation study).

3.4. Application: model-based reproducibility assessments of GMV change for UKB vs. PPMI and UKB vs. HCP

As an application of model-based reproducibility index, we considered two MRI databases: PPMI and UK Biobank cohorts. For the PPMI database, there were 136 normal subjects of age from 45 to 79. As the UK Biobank cohort is much larger, we performed a sample matching based on age and sex for this analysis. Seven age groups of 45–49, 50–54, etc. (5-year intervals) were considered. For each age group, from the UK Biobank cohort, we randomly selected the same number of female/male subjects as that in the PPMI cohort. A total of 136 subjects

Table 3
M^2RI based sample size calculations in different MRI-based analysis scenarios, The minimum sample size to achieve Pr($M^2RI > 0.8$) > 0.8 is presented for each large-scale analysis scenario. For the RSFC data, sex, age or BMI was considered as phenotype. For the GMV data, sex, age or BMI was considered as phenotype. For the fMRI data in task activation, MID, SST or EFT task was considered. (For more details, see section *Sample Size Needed to Achieve a Desirable Model-based Reproducibility*).

MRI Study	*Minimum Sample Size*	*MRI Study*	*Minimum Sample Size*	*MRI Study*	*Minimum Sample Size*
RSFC		*GMV*		*Task fMRI*	
Sex	***300***	*Sex*	***120***	*MID*	***30***
Age	***200***	*Age*	***70***	*SST*	***20***
BMI	***More than 800***	*BMI*	***300***	*EFT*	***30***

Table 4
Model-based reproducibility index for comparing two databases, Model-based reproducibility index applications in comparing the large-scale association analysis results from two closely related databases. Median with the range of interquartile (Q1-Q3) are shown in the table. For the MRI databases PPMI vs. UK Biobank, the analysis was based on age as phenotype vs. GMV. For the MRI databases HCP vs. UK Biobank, the analysis was based on sex as phenotype vs. GMV. (For more details, please see section *Application: Model-based Reproducibility Assessments of GMV Change for UKB* vs. *PPMI and UKB* vs. *HCP*).

Application	*Assessed Model-based reproducibility*
PPMI vs. UKB (Sample size = 136)	*Median (Q1–Q3)*
M^2RI	0.99993 (0.99964–0.99998)
CEFNRI	0.9974 (0.9967–0.9979)
METARI	0.8642 (0.7404–0.9625)
HCP vs. UKB (Sample size = 413)	*Median (Q1-Q3)*
M^2RI	0.6378 (0.5747–0.7032)
CEFNRI	0.4377 (0.3922–0.4838)
METARI	0.0420 (0.0263–0.0656)

were randomly selected from the UK Biobank cohort. Then, for both databases, we calculated the z-scores for the age phenotype as response vs. GMV based on a general linear model with the adjustments for sex and TIV. This was repeated 1000 times and we obtained 1000 lists of paired z-scores.

As another application of model-based reproducibility index, we considered the HCP and UK Biobank cohorts. For the HCP data set, there were 413 subjects of age from 22 to 36. Then, it was not feasible to match the corresponding age ranges in the UK Biobank data because this age range was not available in the UK Biobank. We still performed a sample matching based on sex. From the UK Biobank cohort, we randomly selected the same number of female/male subjects as that in the HCP cohort. A total of 413 subjects were randomly selected from the UK Biobank cohort. Then, for both databases, we calculated the z-scores for the sex phenotype as response vs. GMV based on a general linear model with the adjustments of age and TIV. This was repeated 1000 times and we obtained 1000 lists of paired z-scores. Applications of model-based reproducibility assessments of FCs changes for UKB vs. HCP and pairwise comparing four rs-fMRI scans from HCP database are available in **Supplementary Material**.

For each list of paired z-scores, we applied model-based reproducibility index to assess the related reproducibility (see Table 4 for results). For the PPMI and UK Biobank databases, the median M^2RI, *CEFNRI* or *METARI* was 0.99993,0.9974 or 0.8642 with the lower- and upper-quartiles (Q1-Q3) 0.99964–0.99998, 0.9967–0.9979 or 0.7404–0.9625, respectively. It was reasonable to conclude that both databases were ideally reproducible in term of large-scale association analysis with age as phenotype. For the HCP and UK Biobank databases, the median M^2RI, *CEFNRI* or *METARI* was only 0.6378,0.4377 or 0.0420, with the lower- and upper-quartile (Q1-Q3) 0.5747–0.7032, 0.3922–0.4838 or 0.0263–0.0656, respectively. As the age ranges for both databases were clearly different, it was also reasonable to observe a relatively low reproducibility for this analysis.

4. Discussion

Reproducibility of discoveries based on large-scale data has received a significant attention in recent years. However, there are still a lack of effective statistical model to assess overall study reproducibility in neuroimaging studies (Nature Neuroscience Editorial, 2017; Botvinik-Nezer et al., 2020; Eklund et al., 2016; Poldrack, 2019). To address this need, we proposed a mixture model based reproducibility index, and we discussed its relationship to a recently proposed irreproducibility quantity (Zhao et al., 2020). Through a comprehensive simulation study, we demonstrated the advantages of model-based reproducibility index for an accurate reproducibility assessment in large-scale MRI-based studies. We also demonstrated mode-based reproducibility >0.99 was achieved from several UK Biobank or IMAGEN based large sample size analyzes. Then, we evaluated the sample size necessary for achieving a desirable model-based reproducibility, which is important in a design of experiment. Moreover, we evaluated the reproducibility of GMV changes for UKB vs. PPMI and UKB vs. HCP. We still observed model-based reproducibility >0.99 based on UKB vs. PPMI because of their highly similar study-specific experimental factors (i.e. male to female ratio and age range between these two datasets). However, we observed model-based reproducibility 0.64 based on UKB vs. HCP because of the age range in these two datasets were highly different. Additionally, the functional connectivity networks for rs-fMRI based studies can be affected by some parameters such as brain parcellation and connectivity estimation measures. We conducted some investigation on these and the results were summarized in Supplementary Material. It is clear from our results that the choice of connectivity estimation measures has a significant impact on the model-based reproducibility assessment. For the choice of brain parcellation, when we used the Power 264 atlas as a contrast, similar results were observed. Therefore, the impact from study-specific experimental factors (e.g. the impact of age on the reproducibility for GMV-related studies) plays an important role in the model-based reproducibility assessments for different experiments.

Our newly proposed model-based reproducibility index was derived from the combination of parameters in a mixture model. This index is close related to the irreproducibility quantity ρ_{IR} that was proposed in a previous study (Zhao et al., 2020). Although the model-based reproducibility index can be interpreted as $1 - \rho_{IR}$, their related statistical models are different. The focus of our model-based reproducibility index is on the consistent directional changes between two large-scale high-throughput analysis results. The changes can be positive, negative or no change. However, they cannot be observed in practice (we can only observe z-scores). For analyzing data with unobserved information, mixture models are usually considered in practice. Therefore, our model-based reproducibility index is based on a bivariate Gaussian mixture model.

We conducted minimal sample size calculation for a few phenotypes and task types by using several recent large sMRI/fMRI databases. According to our results in Table 3, to achieve 80% model-based reproducibility, there is minimal sample size requirement. For example, we need at least 30 for MID task in IMAGEN (Fig. 3 and Table 3). Additionally, the sample size requirement is closely related to the strength of associations (or differences), which is largely limited by the signal-to-noise ratio to each measurement. The test-retest reliability is a concept to represent the ability to detect meaningful signals from meaningless background noise (Bennett and Miller, 2010). Therefore, the test-retest reliability of different phenotypes (e.g., Sex, BMI), the test-retest reliability of different predictor data types (e.g., GMV, RSFC and task activation), and technology platforms should all be considered in the study of model-based reproducibility assessment. These conclusions are well illustrated in our results. To achieve the desirable model-based reproducibility, the required sample size for a task fMRI study and a GMV study is clearly lower than that for a RSFC study. For a GMV study, the required sample size for the age phenotype as response is clearly lower than that for the sex phenotype as response. For an RSFC study, the required sample size for the BMI phenotype as response is much larger. These results are consistent with our expectations. A widely used method to quantify test-retest reliability is intraclass correlation coefficient (ICC). A test-retest reliability review for 15 task fMRI studies in healthy individuals showed an average ICC of 0.5 (Bennett and Miller, 2010). The ICC for absolute GMV estimation across sessions more than 0.7 (Jovicich et al., 2013). However, the average ICC equals 0.22 for link-wise FCs, which is commonly considered rather low (Pannunzi et al., 2017). In our analyzes, the phenotype test-retest reliability of BMI is clearly smaller than that of sex or age phenotype. Therefore, our results are highly illustrative and informative for planning the sample size for these large-sale high-throughput MRI-based studies.

The correlation or reproducibility for different association analyzes between neuroimaging features and phenotypes have been widely discussed in recent work (Marek et al., 2022; Masouleh et al., 2019). Moreover, many investigations have indicated that the correlation or reproducibility is influenced by some fundamental experiment factors such as sample size and effect size (Schönbrodt and Perugini, 2013). Our model-based reproducibility index can be interpreted as the level of directional concordance between two large-scale analysis results (i.e. z-score). Therefore, our index (mixture model reproducibility index) can evaluate the overall consistency between two studies in term of the direction of association. It is different from traditional correlation. Instead of focusing individual discoveries, it is a global evaluation based on large-scale high-throughput analyzes. Neuroimaging results include unthresholded and thresholded statistical maps (Botvinik-Nezer et al., 2020). Many existing reproducibility studies were mostly on thresholded statistical maps, which are highly important in literature. The model-based reproducibility assessment approach is based on unthresholded statistical maps. It would be highly convenient if unthresholded statistical maps could be shared, which would facilitate the evaluation of model-based reproducibility index. As data pooling or meta-analysis is frequently considered in practice, the evaluation of model-based reproducibility index between two closely related studies is crucial. Reasonable use of our proposed index allows investigators to understand the heterogeneity among different studies. It also allows investigators to address the generalizability of a large-scale analysis. The relationship between the reproducibility and the heterogeneity among different experiments was also discussed (Zhao et al., 2020). Furthermore, the fundamental experiment factors, such as sample size, effect size, etc., also have significant impact on the model-based reproducibility assessment.

Our model-based reproducibility assessment approach was based on z-scores from univariate association analyzes. As multivariate neuroimaging analysis has received increasing attention, it is highly interesting to investigate the analysis methods and applications based on multivariate neuroimaging associations. Besides, we have demonstrated that our proposed model-based reproducibility index is useful for the reproducibility assessment of large databases. It is still necessary to further develop novel and useful tools for data with relatively small sample sizes. Statistically, when the sample size is relatively small, it is difficult to fit the z-scores with a simple model. As our future research endeavor, we will investigate other approaches so that the model-based reproducibility assessment approach can be achieved for data with a relatively small sample size. We believe that these efforts will also help improve the current approach for data with a relatively large sample size. Moreover, model-based reproducibility index can also be applied to brain-wide association study (BWAS) (Cheng et al., 2015a, 2015b; Gong et al., 2018) and other types of large-scale high-throughput association study. We would point out that the model-based reproducibility assessment approach of BWAS can be computationally time consuming. As the number of features (voxel-wise FCs) is significantly large, we will direct our future research endeavors toward how to conduct such an analysis more effectively and efficiently.

5. Conclusion

The reproducibility of large-scale high-throughput MRI-based research discoveries has been recently debated. There is still a lack of effective and efficient model-based reproducibility assessment approach in neuroimaging experiments. We have developed a model-based reproducibility index for assessing the reproducibility in large-scale MRI-based studies. For several large-scale high-throughput studies, we have evaluated the sample size for a few phenotypes or task types to achieve a desirable model-based reproducibility requirement. Our study provides a scientific contribution to the measurement of reproducibility that is fundamental and important in the current large-scale high-throughput MRI-based research.

Data and code availability statement

Data used in this work were obtained from the publicly available databases: UK Biobank (http://www.ukbiobank.ac.uk/) database, WU-Minn Human Connectome Project (HCP) (http://www.humanconnectomeproject.org/data/hcp-project/) database and Parkinson Progression Marker Initiative (PPMI) (https://www.ppmi-info.org/) database. Details of the data collection can be found at the database websites. Furthermore, IMAGEN database was also used in this work and all data could be accessed based on request to the corresponding committee. Code used in this work could be obtained from the corresponding author on reasonable request.

Declaration of Competing Interest

All authors declare no competing interests.

Credit authorship contribution statement

Zeyu Jiao: Visualization, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Yinglei Lai:** Visualization, Formal analysis, Writing – original draft, Writing – review & editing. **Jujiao Kang:** Formal analysis, Writing – original draft. **Weikang Gong:** Data curation, Formal analysis, Writing – original draft. **Liang Ma:** Formal analysis, Writing – original draft. **Tianye Jia:** Writing – review & editing. **Chao Xie:** Data curation. **Shitong Xiang:** Data curation. **Wei Cheng:** Writing – review & editing. **Andreas Heinz:** Data curation. **Sylvane Desrivières:** Data curation. **Gunter Schumann:** Data curation. **Fengzhu Sun:** Visualization, Writing – original draft, Writing – review & editing. **Jianfeng Feng:** Visualization, Writing – original draft, Writing – review & editing.

Acknowledgements

JFF is supported by National Key R&D Program of China (Nos. 2018YFC1312904 and 2019YFA0709502), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), ZJ Lab, Shanghai Center for Brain Science and Brain-Inspired Technology, and the 111 Project (No. B18015).

This work received other support from the following sources: National Key R&D Program of China (No 2018YFC1312900), National Key

R&D Program of China (No. 2019YFA0709501), National Natural Science Foundation of China (NSFC 91630314, 81801773), National Natural Science Foundation of China (No. 11971459, 81673833), the European Union-funded Integrated Project IMAGEN (Reinforcement-related behavior in normal brain function and psychopathology) (LSHM-CT-2007–037286), the Horizon 2020 funded 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), Human Brain Project (HBP SGA 2, 785907), the FP7 project MATRICS (603016), the Medical Research Council Grant 'c-VEDA' (Consortium on Vulnerability to Externalizing Disorders and Addictions) (MR/N000390/1), the National Institute for Health Research (NIHR) Biomedical Research center at South London and Maudsley NHS Foundation Trust and King's College London, the Bundesministerium für Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; Forschungsnetz AERIAL 01EE1406A, 01EE1406B), the Deutsche Forschungsgemeinschaft (DFG grants SM 80/7–2, SFB 940, TRR 265, NE 1383/14–1), the Medical Research Foundation and Medical Research Council (grants MR/R00465X/1 and MR/S020306/1), the National Institutes of Health (NIH) funded ENIGMA (grants 5U54EB020403–05 and 1R56AG058854–01), 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 centers of Excellence. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119166.

References

Nature Neuroscience Editorial, 2017. Fostering reproducible fMRI research. Nat. Neurosci. 20 298-298.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry–the methods. Neuroimage 11, 805–821.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging. Ann. N. Y. Acad. Sci. 1191, 133–155.

Bosnell, R., Wegner, C., Kincses, Z.T., Korteweg, T., Agosta, F., Ciccarelli, O., De Stefano, N., Gass, A., Hirsch, J., Johansen-Berg, H., Kappos, L., Barkhof, F., Mancini, L., Manfredonia, F., Marino, S., Miller, D.H., Montalban, X., Palace, J., Rocca, M., Enzinger, C., Ropele, S., Rovira, A., Smith, S., Thompson, A., Thornton, J., Yousry, T., Whitcher, B., Filippi, M., Matthews, P.M., 2008. Reproducibility of fMRI in the clinical setting: implications for trial designs. Neuroimage 42, 603–610.

Bossier, H., Roels, S., Seurinck, R., Banaschewski, T., Barker, G.J., Bokde, A.L.W., Quinlan, E.B., Desrivieres, S., Flor, H., Grigis, A., 2020. The empirical replicability of task-based fMRI as a function of sample size. Neuroimage 212, 116601.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., Avesani, P., Baczkowski, B.M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R.G., Berkers, R.M.W.J., Bhanji, J.P., Biswal, B.B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K.L., Bowring, A., Braem, S., Brooks, H.R., Brudner, E.G., Calderon, C.B., Camilleri, J.A., Castrellon, J.J., Cecchetti, L., Cieslik, E.C., Cole, Z.J., Collignon, O., Cox, R.W., Cunningham, W.A., Czoschke, S., Dadi, K., Davis, C.P., Luca, A.D., Delgado, M.R., Demetriou, L., Dennison, J.B., Di, X., Dickie, E.W., Dobryakova, E., Donnat, C.L., Dukart, J., Duncan, N.W., Durnez, J., Eed, A., Eickhoff, S.B., Erhart, A., Fontanesi, L., Fricke, G.M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J.J., Golowin, S.A.E., González-García, C., Gorgolewski, K.J., Grady, C.L., Green, M.A., Guassi Moreira, J.F., Guest, O., Hakimi, S., Hamilton, J.P., Hancock, R., Handjaras, G., Harry, B.B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.P., Huettel, S.A., Hughes, M.E., Iacovella, V., Iordan, A.D., Isager, P.M., Isik, A.I., Jahn, A., Johnson, M.R., Johnstone, T., Joseph, M.J.E., Juliano, A.C., Kable, J.W., Kassinopoulos, M., Koba, C., Kong, X.Z., Koscik, T.R., Kucukboyaci, N.E., Kuhl, B.A., Kupek, S., Laird, A.R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M.Y.C., Lim, P.C., Lintz, E.N., Liphardt, S.W., Losecaat Vermeer, A.B., Love, B.C., Mack, M.L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J.T., Melero, H., Méndez Leal, A.S., Meyer, B., Meyer, K.N., Mihai, G., Mitsis, G.D., Moll, J., Nielson, D.M., Nilsonne, G., Notter, M.P., Olivetti, E., Onicas, A.I., Papale, P., Patil, K.R., Peelle, J.E., Pérez, A., Pischedda, D., Poline, J.B., Prystauka, Y., Ray, S., Reuter-Lorenz, P.A., Reynolds, R.C., Ricciardi, E., Rieck, J.R., Rodriguez-Thompson, A.M., Romyn, A., Salo, T., Samanez-Larkin, G.R., Sanz-Morales, E., Schlichting, M.L., Schultz, D.H., Shen, Q., Sheridan, M.A., Silvers, J.A., Skagerlund, K., Smith, A., Smith, D.V., Sokol-Hessner, P., Steinkamp, S.R., Tashjian, S.M., Thirion, B., Thorp, J.N., Tinghög, G., Tisdall, L., Tompson, S.H., Toro-Serey, C., Torre Tresols, J.J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A.E., Verguts, T., Vettel, J.M., Vijayarajah, S., Vo, K., Wall, M.B., Weeda, W.D., Weis, S., White, D.J., Wisniewski, D., Xifra-Porxas, A., Yearling, E.A., Yoon, S., Yuan, R., Yuen, K.S.L., Zhang, L., Zhang, X., Zosky, J.E., Nichols, T.E., Poldrack, R.A., Schonberg, T., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88.

Chen, X., Lu, B., Yan, C.G., 2017. Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. Hum. Brain Mapp. 39, 300–318.

Cheng, W., Palaniyappan, L., Li, M., Kendrick, K.M., Zhang, J., Luo, Q., Liu, Z., Yu, R., Deng, W., Wang, Q., Ma, X., Guo, W., Francis, S., Liddle, P., Mayer, A.R., Schumann, G., Li, T., Feng, J., 2015a. Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. NPJ Schizophr. 1, 15016.

Cheng, W., Rolls, E.T., Gu, H., Zhang, J., Feng, J., 2015b. Autism : reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. Brain 138, 1382–1393.

Conti, A., Duggento, A., Guerrisi, M., Passamonti, L., Toschi, N., 2019. Variability and reproducibility of directed and undirected functional MRI connectomes in the human brain. Entropy 21 (7), 661.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92, 548–560.

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

Gong, W., Wan, L., Lu, W., Ma, L., Cheng, F., Cheng, W., Grunewald, S., Feng, J., 2018. Statistical testing and power analysis for brain-wide association study. Med. Image Anal. 47, 15–30.

Goodman, S., Fanelli, D., Ioannidis, J., 2016. What does research reproducibility mean? Sci. Transl. Med. 8 341ps312–341ps312.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63–72.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Tränkner, A., Schönknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J.P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalló, N., Blin, O., Frisoni, G.B., 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. Neuroimage 83, 472–484.

Lai, Y., Adam, B., Podolsky, R.H., She, J., 2007. A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups. Bioinformatics 23, 1243–1250.

Lai, Y., Eckenrode, S., She, J., 2009. A statistical framework for integrating two microarray data sets in differential expression analysis. BMC Bioinf. 10, 1–11.

Lai, Y., Zhang, F., Nayak, T.K., Modarres, R., Lee, N.H., Mccaffrey, T.A., 2014. Concordant integrative gene set enrichment analysis of multiple large-scale two-sample expression data sets. BMC Genom. 15, 1–12.

Lai, Y., Zhang, F., Nayak, T.K., Modarres, R., Lee, N.H., Mccaffrey, T.A., 2017. An efficient concordant integrative analysis of multiple large-scale two-sample expression data sets. Bioinformatics 33, 3852–3860.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453, 869–878.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C.M., Simuni, T., Coffey, C.S., Kieburtz, K., Flagg, E., Chowdhury, S., 2011. The Parkinson progression marker initiative (PPMI). Prog. Neurobiol. 95, 629–635.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N.A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660.

Masouleh, S.K., Eickhoff, S.B., Hoffstaedter, F., Genon, S., 2019. Empirical examination of the replicability of associations between brain structure and psychological variables. eLife Sci. 8, e43464.

Mclachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Stat. 36, 318–324.

Pannunzi, M., Hindriks, R., Bettinardi, R.G., Wenger, E., Lisofsky, N., Martensson, J., Butler, O., Filevich, E., Becker, M., Lochstet, M., Kühn, S., Deco, G., 2017. Resting-state fMRI correlations: from link-wise unreliability to whole brain stability. Neuroimage 157, 250–262.

Patel, A.X., Kundu, P., Rubinov, M., Jones, P.S., Vertes, P.E., Ersche, K.D., Suckling, J., Bullmore, E.T., 2014. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. Neuroimage 95, 287–304.

Poldrack, R.A., 2019. The costs of reproducibility. Neuron 101, 11–14.

Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517.

Rolls, E.T., Joliot, M., Tzouriomazoyer, N., 2015. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. Neuroimage 122, 1–5.

Schönbrodt, F.D., Perugini, M., 2013. At what sample size do correlations stabilize? J. Res. Personal. 47, 609–612.

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., 2010a. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G.J., Buchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., 2010b. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Snyder, A.Z., Raichle, M.E., 2012. A brief history of the resting state: the Washington university perspective. Neuroimage 62, 902–910.

Sudlow, C., Gallacher, J., Allen, N.E., Beral, V., Burton, P.R., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M.J., 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.

Tegeler, C., Strother, S.C., And, J.R.A., Kim, S.G., 1999. Reproducibility of BOLD-based functional MRI obtained at 4 T. Hum. Brain Mapp. 7, 267–283.

Van Essen, D.C., Smith, S.M., Deanna, M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Van Essen, D.C., Ugurbil, K., Auerbach, E.J., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Penna, S.D., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62, 2222–2231.

Zhao, Y., Sampson, M.G., Wen, X., 2020. Quantify and control reproducibility in high--throughput experiments. Nat. Methods 17, 1207–1213.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., Wells, W.M., Group, F.B.R., 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by biomedical informatics research network. Radiology 237, 781–789.