Assessing Study Reproducibility through *M²RI*: A Novel Approach for Large-scale High-throughput Association Studies

Zeyu Jiao[1,2,3,#]; Yinglei Lai[4,#]; Jujiao Kang[1,2,3]; Weikang Gong[8]; Liang Ma[10]; Tianye Jia[2,3,9]; Chao Xie[2,3]; Wei Cheng[2,3]; Andreas Heinz[9], Sylvane Desrivières[11], Gunter Schumann[2,9,12], IMAGEN Consortium, Fengzhu Sun[5] and Jianfeng Feng[1,2,3,6,7,*]

1. Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China

2. Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, China

3. Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, China

4. Department of Statistics, The George Washington University, 801 22nd St. NW, Washington, DC 20052, USA

5. Quantitative and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA, 90089, USA

6. Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

7. School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai, China

8. Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Welcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK.

9. Centre for Population Neuroscience and Precision Medicine (PONS), Institute of Psychiatry, Psychology & Neuroscience, SGDP Centre, King's College London, UK.

10. Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

11. 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

12. PONS Research Group, Dept of Psychiatry and Psychotherapy, Campus Charite Mitte, Humboldt University, Berlin, Germany

Keywords: Study reproducibility; Association studies; Sample size; MRI (magnetic resonance imaging); RNA sequencing expression

[#] These authors contributed equally to this work.

* Corresponding author: Professor Jianfeng Feng
Institute of Science and Technology for Brain-inspired Intelligence, School of Life Science and the Collaborative Innovation Center for Brain Science, Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China, and Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
Tel. 86-21-65643621
Email: jianfeng64@gmail.com

Abstract

High-throughput technologies, such as magnetic resonance imaging (MRI) and DNA/RNA sequencing (DNA-seq/RNA-seq), have been increasingly used in large-scale association studies. With these technologies, important biomedical research findings have been generated. The reproducibility of these findings, especially from structural MRI (sMRI) and functional MRI (fMRI) association studies, has recently been questioned. There is an urgent demand for a reliable overall reproducibility assessment for large-scale high-throughput association studies. It is also desirable to understand the relationship between study reproducibility and sample size in an experimental design. In this study, we developed a novel approach: the mixture model reproducibility index (M^2RI) for assessing study reproducibility of large-scale association studies. With M^2RI, we performed study reproducibility analysis for several recent large sMRI/fMRI data sets. The advantages of our approach were clearly demonstrated, and the sample size requirements for different phenotypes were also clearly demonstrated, especially when compared to the Dice coefficient (DC). We applied M^2RI to compare two MRI or RNA sequencing data sets. The reproducibility assessment results were consistent with our expectations. In summary, M^2RI is a novel and useful approach for assessing study reproducibility, calculating sample sizes and evaluating the similarity between two closely related studies.

Introduction

With the rapid development of technologies in biological research, investigators often face staggering quantities of spatiotemporal data. Association studies have been widely conducted in many biomedical research fields, examples include the well-known genome-wide association study (GWAS) approach in genetics and the brain-wide association study (BWAS) approach in neuroimaging (1-4). However, due to the intrinsic complexity of the problems studied, the data are generally noisy, and the correlations among variables are usually weak. As a result, many, if not the majority, of the reported findings in the literature are inevitably false, which has provoked intense debates in the literature. For example, criticisms have been raised to the phenomena that some fMRI findings are only modestly reproducible and that some results could be interpreted as inflated or spurious (5-7). Similarly, the gene lists obtained for the same clinical types of patients by different research groups differed widely and shared very few genes in common (8). Despite all the aforementioned observations, it seems that one of the long-standing key issues is the lack of efficient methodologies for accessing the reproducibility of a given experiment, which should help establish a standard to use when the reproducibility is under debate.

The Dice coefficient (*DC*) is one of the widely used measurements of reproducibility in string matching (9), image segmentation (10), and other areas. It has been recently applied to measure the reproducibility of some large-scale fMRI association studies, including resting-state and task-based designs (11, 12). *DC* calculates the proportion of consistent features: thresholds are first selected to find significant results, and then the proportion of consistently significant results is calculated (see **Figure 1A**).

In this study, we have demonstrated that *DC* is not an efficient reproducibility measure. It requires a threshold selection that is not flexible for measurement errors or statistical summaries. This limitation is especially clear when we focus on the overall reproducibility assessment of a study (termed study reproducibility). To overcome this limitation, we have developed a mixture model based approach, along with a similar idea that has been proposed to test the reproducibility of surface-enhanced laser desorption/ionization–mass spectrometry (SELDI-MS), microarray data analysis and gene set enrichment analysis (13-16). In this model, each mixture component possesses a clear interpretation in practice (see **Figure 1B**).

We have developed a new reproducibility assessment index based on our proposed mixture model. This index is termed the mixture model reproducibility index (M^2RI). To evaluate M^2RI, we present the analysis results from simulations based on sMRI data in the UK Biobank (17, 18). Then, we present a comprehensive reproducibility assessment analysis of gray-matter-related human behaviors, brain task state activation and connectivity-behavior analysis in which the reproducibility rates are known. With a desirable reproducibility rate, we then present the related minimal sample size calculation in various large-scale association analysis scenarios based on the structure and resting-state MRI data from the UK Biobank project (17, 18) and the task activation data from the IMAGEN project (19, 20). In summary, M^2RI has demonstrated a clear robustness and efficiency in the assessment of study reproducibility. Unlike *DC*, it does not require a threshold selection for declaring significant findings. The minimum sample size related to a desirable reproducibility rate should provide a gold standard to use in the planning of a large-scale association study.

Materials and Methods

Data Summary

We used MRI images and phenotypic data from four data sources: the UK Biobank (17, 18), Human Connectome Project (HCP) (21), IMAGEN (the first multicenter genetic-neuroimaging study) (19, 20) and Parkinson's Progression Markers Initiative (PPMI) cohorts (22). All the image data were preprocessed by a standard pipeline. Moreover, we also analyzed RNA sequencing (RNA-seq) data collected by The Cancer Genome Atlas (TCGA) project (23) (see **Supplementary 1** for details).

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 (24). 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 behaviors. The effects of age, sex and head motion (mean frame-wise displacement, or FD) were regressed out.

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.

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, 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.

The Cancer Genome Atlas (TCGA) Gene Expression Data

We downloaded RNA sequencing (RNA-seq) expression data from TCGA data portal (23). Many different cancer types have been studied in TCGA project. For a cancer study, RNA-seq expression data for normal and tumor subjects were collected for 20,531 genes. For each gene, we used the Wilcoxon rank-sum test to compare the RNA-seq expression between the normal and tumor subjects.

Normal Distribution Quantile-based Transformation

z-scores from a normal distribution quantile transformation were used for the analysis (13). First, based on an appropriate association analysis (functional connectivity association study, voxel-wise association study, task fMRI activation or differential expression), 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.

Dice Coefficient

The Dice coefficient (*DC*) is calculated as:

$$DC = \frac{2\times(V_{overlap}^{+} + V_{overlap}^{-})}{V_1 + V_2}$$

where V_1 and V_2 represent the numbers of supra-threshold significant results from large-scale test 1 and 2, respectively, and $V_{overlap}^{+}$ and $V_{overlap}^{-}$ represent the numbers of supra-threshold positive or negative results in both tests.

Mixture Model Reproducibility Index *(M^2RI)*

Motivated by the Dice coefficient that divides *z*-scores into three parts, we consider three categories for the underlying/unobserved association analysis related true status: positive change/correlation, no-change/correlation and negative change/correlation. Instead of thresholding, we consider a three-

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

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

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: $\mu_1, \nu_1 \leq 0$ and $\mu_2, \nu_2 \geq 0$. π_{ij} is the proportion for component i in the first association study and component j in the second association study, and $\sum_{ij} \pi_{ij} = 1$.

This model was termed partial concordance/discordance (PCD) model (13). Then, we define the mixture model reproducibility index (M^2RI) as (see **Supplementary 1** for details):

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

Confidence Intervals

The confidence intervals (CIs) of M^2RI and *DC* can be obtained by bootstrapping paired z-scores (25, 26). For our newly proposed reproducibility index M^2RI, a theoretical confidence interval will also be useful in practice. Therefore, we have derived the asymptotic theoretical CIs for M^2RI based on our proposed mixture model (see **Supplementary 2** for details).

Results

M^2RI Recovers the True Reproducibility Index More Accurately than the Dice coefficient in the Simulation Study

We conducted a comprehensive simulation study to compare the performance of our newly proposed M^2RI vs. the widely used Dice coefficient (*DC*). Our simulations were designed based on the gray matter volume (GMV) data in the UK Biobank. Two-sample comparison is a general association analysis scenario in practice, and the reproducibility of a large-scale two-sample study is important. Therefore, we partitioned the data randomly into four subsets (referred to as Data 1A, Data

1B, Data 2A and Data 2B). Before the analysis, as a widely considered practical approach, we checked that sex, age, total intracerebral volume (TIV) and total GMV were statistically similar between Data 1A vs. 2A as well as Data 1B vs. 2B. Otherwise, we repeated the random data partition until one passed this similarity requirement. The reproducibility was 100% from the random data partition. For each feature, there was statistically no change in distribution between Data 1A vs. 2A nor Data 1B vs. 2B. Then, to generate upward or downward changes, a specified proportion of 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 GMV voxels of each subject in Data 1A and Data 1B. This procedure was repeated 1,000 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 overall reproducibility between two lists of *z*-scores was assessed by M^2RI vs. *DC*. The following three simulation scenarios were considered.

(a) This setting represents complete reproducibility 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) This setting represents partial reproducibility. 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) This setting represents complete reproducibility with a high proportion of changes. 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).

The comparison results are summarized in **Table S1** and **Figure 2**. Based on the scenario (a) as complete reproducibility with a moderate proportion of changes, the median *DC* was only 0.44, 0.38 or 0.13 when the threshold value was 1, 2 or 3 for *z*-score respectively. Furthermore, all the related ranges of interquartile (Q1-Q3) were low. However, the median M^2RI was 0.915 with a range of interquartile (Q1-Q3) 0.738-0.995. (It was reasonable to conclude that the assessed reproducibility could be up to 100%.) Based on the scenario (b) as a partial reproducibility (75%), the median *DC* was 0.52, 0.38 or 0.12 when the threshold value was 1, 2 or 3 for *z*-scores, respectively, and all the related ranges of interquartile (Q1-Q3) were low. The median M^2RI was 0.77, with a true value of 75% in the middle of the range of interquartile (Q1-Q3) 0.6850-0.8488. Based on the scenario (c) as complete reproducibility with a high proportion of changes, the median *DC* was still not satisfactory (0.60, 0.44 and 0.14 when the threshold value was 1, 2 or 3 for *z*-scores, respectively). However, the median M^2RI reached 0.96 with a range of interquartile (Q1-Q3) 0.8731-0.9986. (It was again reasonable to conclude that the assessed reproducibility could be up to 100%.). Overall, M^2RI is a clearly preferred choice for evaluating the reproducibility of large-scale association analysis.

Reproducibility of Large-scale MRI Association Studies

To investigate the reproducibility of large-scale MRI association 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 Data 1 and Data 2 based on the order of subject number) with (approximately) the same sample sizes. For the resting-state functional connectivity (RSFC) data, the sample sizes of the two subsets were 4,136 and 4,137 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 **Figure S1c** and **Figure S1d** for the paired *z*-scores). For the GMV data, the sample sizes of the two subsets were 4,925 and 4,925. 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 **Figure S1a** for the paired *z*-scores). For the task-related activation data, the sample sizes of the two subsets were 772 and 772. 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 **Figure S1b** 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.

Both M^2RI and the *DC* were used to evaluate the reproducibility based on the paired *z*-scores in **Figure S1**. The results are shown in **Table S1**. 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. Clear differences can be observed from the comparisons between the M^2RI and *DC*. For Sex as phenotype vs. RSFC, based on three α-levels for declaring significance 0.05, 0.01 or 0.001 for *P*-value, the *DC* was 0.89, 0.87 or 0.85, respectively, and all the related 95% CIs were below 0.90. However, M^2RI was nearly one, which suggested an ideal reproducibility. Its asymptotic theoretical 95% CI was above 0.98. For BMI as phenotype vs. RSFC. *DC* was only 0.66, 0.63 or 0.59 at α-level 0.05, 0.01 or 0.001, respectively, and all the related 95% CIs were below 0.70. However, M^2RI was still nearly one, and its asymptotic theoretical 95% CI was above 0.97. For Sex as phenotype vs. GMV, *DC* was 0.93, 0.92 or 0.92 at α-level 0.05, 0.01 or 0.001, respectively, and all the related 95% CIs were below 0.94. However, M^2RI was again nearly one and both 95% CIs were nearly ideal. For activation in the MID task, *DC* was 0.97, 0.96 or 0.96 at α-level 0.05, 0.01 or 0.001, respectively, and all the related 95% CIs were below 0.97. However, M^2RI was still nearly one and both 95% CIs were again nearly ideal. The above comparison results clearly suggest that it is important to use an appropriate measure/metric for evaluating the reproducibility of large-scale MRI association analysis.

Sample Size Calculation based on a Desirable M^2RI

Sample size calculation is crucial in experimental designs. When designing a large-scale association analysis, one may ask what minimum sample size is required to achieve a desirable reproducibility rate. For a comprehensive understanding of sample size requirements in different large-scale association analysis scenarios, we conducted a large resampling-based simulation study. For a study cohort presented in **Table S1**, 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. M^2RI (DC) was calculated accordingly. For each given sample size, we repeated the resampling and M^2RI (DC) calculation 1,000 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 **Figure 3** and **Table 1**. We firstly 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 sex phenotype; a sample size of 70 was required when the response was age phenotype and a sample size of 300 was required when the response was 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 age or sex phenotype, respectively. When the response was BMI, the minimum sample size increased to a very large value (2,300).

As a comparison, we also calculated the minimum sample size for DC to achieve Pr($DC > 0.8$). The results are included in **Table 1** as well. As a threshold setting is required for the calculation of DC and the z-scores are transformed from the related P-values. We considered different significance levels (α=0.05, 0.01, 0.001) for the related P-values. At each threshold setting and for each analysis scenario, the minimum sample size for DC was always clearly larger, even for task fMRI studies. When the response was BMI in the RSFC data, the original cohort was not large enough for our sample size calculation. (At most one half of the original cohort sample size could be available for this resampling based sample size calculation.) **Table 1** can be a useful guideline for a future

experimental design of a large-scale association analysis. These results are essential and helpful for us to further understand the reproducibility evaluation in different analysis scenarios.

Application 1: Reproducibility Evaluation of GMV Change for Two Independent Data Sets

As an application of M^2RI, we considered two MRI data sets: PPMI and UK Biobank cohorts. For the PPMI data set, there were 136 normal subjects with 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 were randomly selected from the UK Biobank cohort. Then, for both data sets, we calculated the *z*-scores for age phenotype as response vs. GMV based on a general linear model with the adjustments for sex and TIV. This was repeated 1,000 times and we obtained 1,000 lists of paired *z*-scores.

As another application of M^2RI, we considered HCP and UK Biobank cohorts. For the HCP data set, there were 413 subjects with 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 of 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 data sets, we calculated the *z*-scores for sex phenotype as response vs. GMV based on a general linear model with the adjustments of age and TIV. This was repeated 1,000 times and we obtained 1,000 lists of paired *z*-scores.

For each list of paired z-scores, we applied M^2RI to assess the related reproducibility (see **Table S2** for results). For PPMI and UK Biobank data sets, the median reproducibility was 0.99993 with the range of interquartile (Q1-Q3) 0.99964-0.99998. It was reasonable to conclude that both data sets were ideally reproducible in term of large-scale association analysis with age as phenotype. For HCP and UK Biobank data sets, the median reproducibility was only 0.6378, with the range of interquartile (Q1-Q3) 0.5747-0.7032. As the age ranges for both data sets were clearly different, it was also reasonable to observe a relatively low reproducibility for this analysis. As a comparison, we also

calculated the related *DC*. We considered different significance levels (α=0.05, 0.01, 0.001) for the *P*-values related to *z*-scores for the *DC* threshold setting. At each threshold setting and for each analysis scenario, the calculated *DC* was clearly lower.

Application 2: Reproducibility Evaluation of Differential Expression based on RNA-seq Data

For further applications of M^2RI, we also considered RNA sequencing gene expression data collected by TCGA. We selected the data for studying colon adenocarcinoma (COAD) and stomach adenocarcinoma (STAD). COAD and STAD are both gastrointestinal (GI) carcinoid tumors. We expect a relatively high reproducibility when comparing these two data sets. As a contrast, we also selected the data for studying head and neck squamous cell carcinoma (HNSC) and liver hepatocellular carcinoma (LIHC).

At the time of our study, for COAD cohort, the numbers of normal and tumor samples were 41 and 287, respectively; for STAD cohort, the numbers of normal and tumor samples were 35 and 415, respectively. We randomly selected 287 tumor subjects from COAD cohort and 35 normal subjects from STAD cohort to match the number of normal/tumor subjects in both data sets. Similarly, for HNSC cohort, the numbers of normal and tumor samples were 44 and 522, respectively; for LIHC cohort, the numbers of normal and tumor samples were 50 and 373, respectively. We randomly selected 373 tumor subjects from LIHC cohort and 44 normal subjects from HNSC cohort to match the number of normal/tumor subjects in both data sets. The random subject selection for matching was repeated 1,000 times. For each gene, we used the Wilcoxon rank-sum test to compare the expression difference between the normal and tumor subjects. Then, we obtained 1,000 lists of paired *z*-scores for COAD vs. STAD and 1,000 lists of paired *z*-scores for HNSC vs. LIHC.

For each list of paired z-scores, we applied M^2RI to assess the related reproducibility (see **Table S2** for results). For COAD and STAD, the median reproducibility was 0.9439 with the range of interquartile (Q1-Q3) 0.9403-0.9468. For HNSC and LIHC, the median reproducibility was only 0.8036 with the range of interquartile (Q1-Q3) 0.7970-0.8100. As COAD and STAD are both gastrointestinal (GI) carcinoid tumors, but the organisms related to HNSC and LIHC are clearly separated, these results are consistent with our expectations. As a comparison, we also calculated the related *DC*. We considered different significance levels (α=0.05, 0.01, 0.001) for the *P*-values related

to *z*-scores for the *DC* threshold setting. At each threshold setting and for each analysis scenario, the calculated *DC* was clearly lower.

Discussion

Reproducibility in the study of large data sets has received significant attention in recent years, but there are still few practical approaches to tackle it (5, 7, 27, 28). To address the demand for a reliable overall reproducibility assessment for large-scale high-throughput association studies, we developed a novel approach termed M^2RI. Through a comprehensive simulation study, we demonstrated the advantages of M^2RI, especially when it was compared to the widely used Dice coefficient (*DC*). The model robustness of M^2RI was considered in our simulation study. We confirmed that M^2RI was an informative approach for the study reproducibility assessment of large-scale association analysis with a relatively large sample size. Through the applications of M^2RI to several large MRI/fMRI data sets, we further demonstrated the advantages of M^2RI (also compared with *DC*). M^2RI can also be useful in evaluating the overall similarity between two large-scale association studies. We used it to compare two MRI data sets as well as two TCGA RNA-seq data sets. As data pooling or meta-analysis is frequently considered in practice, the evaluation of overall similarity between two closely related studies is crucial. Such an analysis allows us to understand the biological differences and similarities between the two studies. It also allows us to address the generalizability of a large-scale association analysis. If there are more than two data sets, we can use M^2RI to evaluate the overall similarity for each pair of data sets.

M^2RI was developed for assessing the overall reproducibility of a large-scale high-throughput association study (or the overall similarity between two large-scale studies). This is different from the reproducibility assessment of discoveries from a large-scale association study (or the evaluation of consistency of discoveries from two large-scale studies), which is actually the analytical purpose of *DC*. Furthermore, we have developed a bootstrap-based procedure and a theoretical formula for the calculation of CIs for M^2RI. These also allow us to achieve the sample size calculation, which is always essential in a design of experiment.

We conducted a comprehensive sample size calculation for several recent large sMRI/fMRI data sets. According to our results, an adequate sample size is necessary to report a reliable reproducibility assessment. Additionally, the sample size requirement is closely related to the strength of associations,

which depends largely on the signal-to-noise ratio of response outcome (e.g., phenotypes) and predictors (e.g., MRI signal or gene expression). Therefore, the impact of different phenotypes, predictor data types, and technology platforms should all be considered in the study reproducibility assessment. These results are well illustrated in our results. To achieve the desirable reproducibility, the required sample size for a task fMRI study is clearly lower than that for a GMV study, which is clearly lower than that for an RSFC study. For a GMV study, the required sample size for age phenotype as response is clearly lower than that for sex phenotype as response. For an RSFC study, the required sample size for BMI phenotype as response is much larger. These results are consistent with our expectations. The data signal-to-noise ratios from a task fMRI study are usually clearly large, and the data signal-to-noise ratios from a GMV study are usually comparably larger than those from an RSFC study. The phenotype signal-to-noise ratio 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 a large-sale high-throughput association study.

We have demonstrated that M^2RI is useful for the reproducibility assessment of large data sets. 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 study reproducibility assessment 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. We have illustrated the applications of M^2RI to MRI/fMRI and RNA-seq gene expression data. M^2RI can also be applied to GWAS and other types of large-scale high-throughput association study data (e.g., BWAS). We point out that the study reproducibility assessment of a GWAS data set can be computationally time consuming. The number of features (SNPs) is significantly large. As the dependence among GWAS data is usually strong, we will direct our future research endeavors toward how to conduct such an analysis more effectively and efficiently.

Contributors

Zeyu Jiao, Yinglei Lai, Fengzhu Sun and Jianfeng Feng contributed to the design of the study. Zeyu Jiao, Weikang Gong, Chao Xie, Andreas Heinz, Sylvane Desrivières and Gunter Schumann contributed to preprocess the MRI data. Zeyu Jiao, Yinglei Lai, Jujiao Kang, Liang Ma and Weikang Gong contributed to the analysis of the data and the preparation of the manuscript. Zeyu Jiao, Yinglei Lai, Fengzhu Sun, Wei Cheng, Tianye Jia and Gunter Schumann participated in writing the paper. All collaborators had an opportunity to contribute to the interpretation of the results and to the drafting of the manuscript.

Declaration of Interests

All authors declare no competing interests.

Acknowledgements

This work received support from the following sources: National Key R&D Program of China (2019YFA0709502), the 111 Project (NO.B18015), the key project of Shanghai Science &Technology (No.16JC1420402), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01) and ZJLab, National Key R&D Program of China (No 2018YFC1312900), National Natural Science Foundation of China (NSFC 91630314), the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT- 2007-037286), 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), 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 Centre 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 Centres of Excellence. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

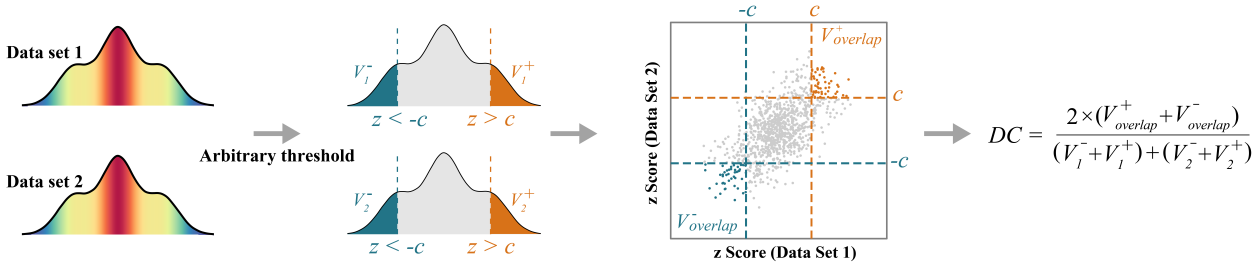
1. W. Cheng, E. T. Rolls, H. Gu, J. Zhang, J. Feng, Autism : reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. *Brain* **138**, 1382-1393 (2015).
2. W. Gong *et al.*, Statistical testing and power analysis for brain-wide association study. *Medical Image Analysis* **47**, 15-30 (2018).
3. J. Marchini, L. R. Cardon, M. S. Phillips, P. Donnelly, The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512-517 (2004).
4. E. Zeggini *et al.*, Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**, 638-645 (2008).
5. Anonymous, Fostering reproducible fMRI research. *Nature Neuroscience* **20**, 298-298 (2017).
6. C. M. Bennett, M. B. Miller, How reliable are the results from functional magnetic resonance imaging. *Annals of the New York Academy of Sciences* **1191**, 133-155 (2010).
7. A. Eklund, T. E. Nichols, H. Knutsson, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 7900-7905 (2016).
8. L. Eindor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5923-5928 (2006).
9. R. J. Bayardo, Y. Ma, R. Srikant (2007) Scaling up all pairs similarity search. in *the web conference*, pp 131-140.
10. B. Norman, V. Pedoia, S. Majumdar, Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. *Radiology* **288**, 177-185 (2018).
11. X. Chen, B. Lu, C. Yan, Reproducibility of R‐fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping* **39**, 300-318 (2018).
12. J. Lu *et al.*, Detectability and reproducibility of the olfactory fMRI signal under the influence of magnetic susceptibility artifacts in the primary olfactory cortex. *NeuroImage* **178**, 613-621 (2018).
13. Y. Lai, B. Adam, R. H. Podolsky, J. She, A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups. *Bioinformatics* **23**, 1243-1250 (2007).
14. Y. Lai, S. Eckenrode, J. She, A statistical framework for integrating two microarray data sets in differential expression analysis. *BMC Bioinformatics* **10**, 1-11 (2009).
15. Y. Lai *et al.*, Concordant integrative gene set enrichment analysis of multiple large-scale two-sample expression data sets. *BMC Genomics* **15**, 1-12 (2014).
16. Y. Lai *et al.*, An efficient concordant integrative analysis of multiple large-scale two-sample expression data sets. *Bioinformatics* **33**, 3852-3860 (2017).
17. F. Alfaro-Almagro *et al.*, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400-424 (2018).
18. C. Sudlow *et al.*, UK biobank: an open access resource for identifying the causes of a wide

range of complex diseases of middle and old age. *PLOS Medicine* **12** (2015).

19. H. Bossier *et al.*, The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage* **212**, 116601 (2020).
20. G. Schumann *et al.*, The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Molecular Psychiatry* **15**, 1128-1139 (2010).
21. D. C. Van Essen *et al.*, The Human Connectome Project: a data acquisition perspective. *NeuroImage* **62**, 2222-2231 (2012).
22. K. Marek *et al.*, The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology* **95**, 629-635 (2011).
23. R. E. Mclendon *et al.*, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
24. E. T. Rolls, M. Joliot, N. Tzouriomazoyer, Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage* **122**, 1-5 (2015).
25. B. Efron, R. Tibshirani, Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association* **92**, 548-560 (1997).
26. G. J. Mclachlan, On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied statistics* **36**, 318-324 (1987).
27. R. Botvinik-Nezer *et al.*, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84-88 (2020).
28. R. A. Poldrack, The Costs of Reproducibility. *Neuron* **101**, 11-14 (2019).

Figures and Tables

A. Dice coefficient (*DC*)



B. Mixture model reproducibility index (*M²RI*)

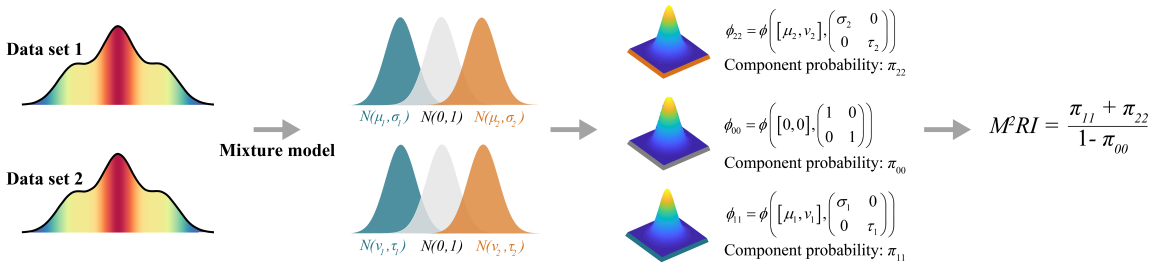


Figure 1: Dice coefficient vs. *M²RI* (A) An illustration of Dice coefficient. V_1^-, V_1^+, V_2^- and V_2^+ represent the numbers of supra-threshold significant results in data set 1 and data set 2, respectively, and $V^+_{overlap}$ and $V^-_{overlap}$ represent the numbers of supra-threshold (threshold c) positive or negative results in both data set. (B) An illustration of *M²RI*. $\phi_{i,j}$ is the normal probability distribution function and π_{ij} is the proportion of features consistent with component i in the first association analysis and component j in the second association analysis.

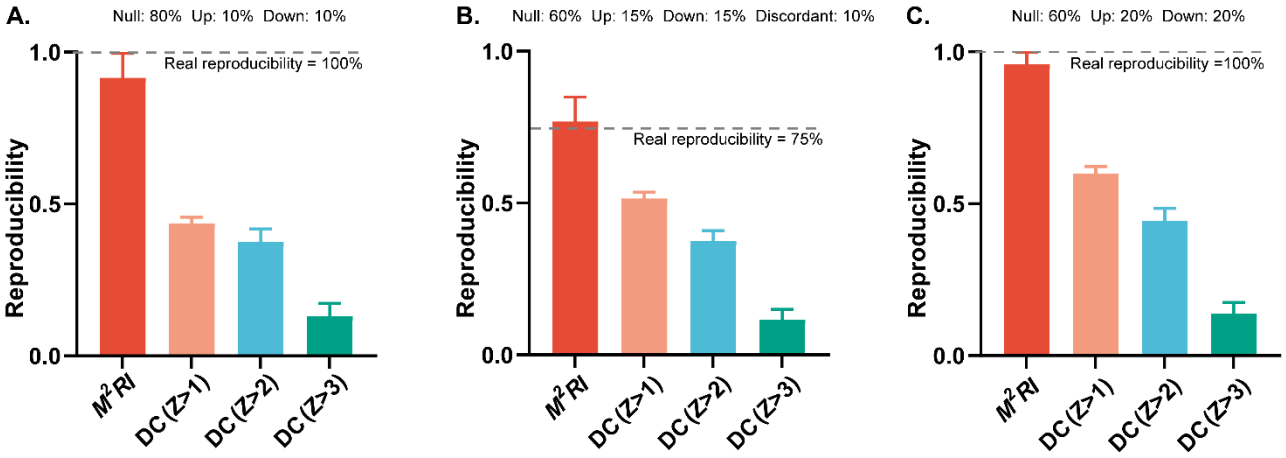


Figure 2: Reproducibility assessed by *M²RI* and Dice coefficient (*DC*) in three simulation scenarios. A bar chat represents the median of 1,000 simulation repetitions with the upper quartile as the error bar. (A) Simulation results based on 80% null, 10% upward change and 10% downward change (100% reproducibility). (B) Simulation results based on 60% null, 15% upward change, 15% downward change and 10% discordant (75% reproducibility). (C) Simulation results based on 60% null, 20% upward change and 20% downward change (100% reproducibility).

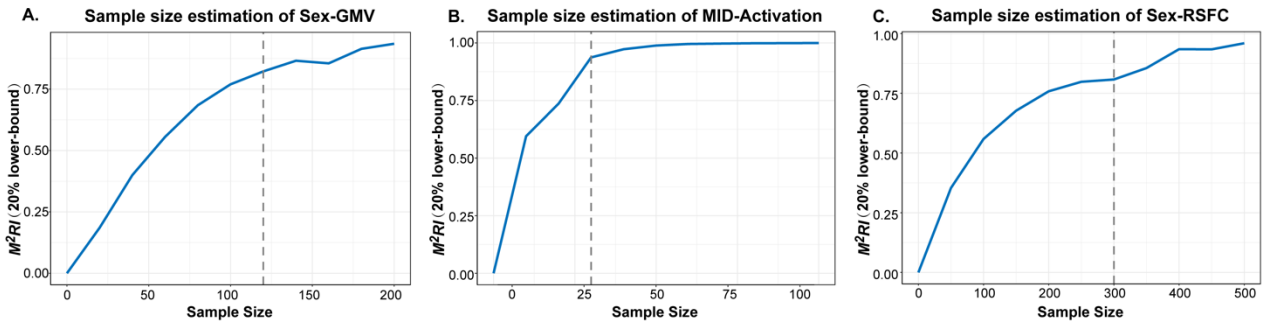


Figure 3: Sample size calculations for three association analysis scenarios in UK Biobank or IMAGEN data. In each plot, "20% lower-bound" means (100-20)-percentile of calculated M^2RI values (1,000 resampled repetitions). The vertical dashed line indicates the minimum sample size for Pr ($M^2RI > 0.8$) > 0.8. (A) Sex as phenotype vs. GMV in UK Biobank data. (B) MID task activation in IMAGEN data. (C) Sex as phenotype vs. RSFC in UK Biobank data.

Table 1: *M^2RI* based sample size calculations in different MRI association analysis scenarios. The minimum sample size to achieve Pr(M^2RI > 0.8) > 0.8 is presented for each large-scale association 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 a comparison purpose, the minimum sample size to achieve Pr(*DC* > 0.8) > 0.8 is also presented at different threshold settings. *DC*: Dice coefficient. (For more details, please see section *M^2RI Discovers the Relationships of Reproducibility and Sample Sizes*.)

MRI Study	*Minimum Sample Size*			
RSFC	*M^2RI*	*DC (P<0.05)*	*DC (P<0.01)*	*DC (P<0.001)*
Sex	*300*	*2600*	*3200*	*3900*
Age	*200*	*1600*	*2100*	*2800*
BMI	*2300*	*>4131*	*>4131*	*>4131*
GMV	*M^2RI*	*DC (P<0.05)*	*DC (P<0.01)*	*DC (P<0.001)*
Sex	*120*	*1300*	*1500*	*1900*
Age	*70*	*500*	*600*	*700*
BMI	*300*	*1800*	*2200*	*2700*
Task fMRI	*M^2RI*	*DC (P<0.05)*	*DC (P<0.01)*	*DC (P<0.001)*
MID	*30*	*60*	*80*	*90*
SST	*20*	*70*	*80*	*90*
EFT	*30*	*210*	*230*	*250*