Cross-validation for the estimation of effect size generalizability in mass-univariate brain-wide association studies

Janik Goltermann[1], Nils R. Winter[1,2], Marius Gruber[1], Lukas Fisch[1], Maike Richter[1,3], Dominik Grotegerd[1], Katharina Dohm[1], Susanne Meinert[1,4], Elisabeth J. Leehr[1], Joscha Böhnlein[1], Anna Kraus[1], Katharina Thiel[1], Alexandra Winter[1], Kira Flinkenflügel[1], Ramona Leenings[1], Carlotta Barkhau[1], Jan Ernsting[1,5,6], Klaus Berger[7], Heike Minnerup[7], Benjamin Straube[8], Nina Alexander[8], Hamidreza Jamalabadi[8], Frederike Stein[8], Katharina Brosch[8], Adrian Wroblewski[8], Florian Thomas-Odenthal[8], Paula Usemann[8], Lea Teutenberg[8], Julia Pfarr[8], Andreas Jansen[8], Igor Nenadić[8], Tilo Kircher[8], Christian Gaser[3,9], Nils Opel[1,3,10,11], Tim Hahn[1], Udo Dannlowski[1]

[1]Institute for Translational Psychiatry, University of Münster, Germany

[2]Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Germany

[3]Department of Psychiatry and Psychotherapy, Jena University Hospital, Germany

[4]Institute for Translational Neuroscience, University of Münster, Germany

[5]Institute for Geoinformatics, University of Münster, Germany

[6]Faculty of Mathematics and Computer Science, University of Münster, Germany

[7]Institute of Epidemiology and Social Medicine, University of Münster, Germany

[8]Department of Psychiatry and Psychotherapy, University of Marburg, Germany

[9]Department of Neurology, Jena University Hospital, Germany

[10]Center for Intervention and Research on adaptive and maladaptive brain Circuits underlying mental health (C-I-R-C), Jena-Magdeburg-Halle, Germany

[11]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany

Abstract

Introduction: Statistical effect sizes are systematically overestimated in small samples, leading to poor generalizability and replicability of findings in all areas of research. Due to the large number of variables, this is particularly problematic in neuroimaging research. While cross-validation is frequently used in multivariate machine learning approaches to assess model generalizability and replicability, the benefits for mass-univariate brain analysis are yet unclear. We investigated the impact of cross-validation on effect size estimation in univariate voxel-based brain-wide associations, using body mass index (BMI) as an exemplary predictor.

Methods: A total of n=3401 adults were pooled from three independent cohorts. Brain-wide associations between BMI and gray matter structure were tested using a standard linear mass-

univariate voxel-based approach. First, a traditional non-cross-validated analysis was conducted to identify brain-wide effect sizes in the total sample (as an estimate of a realistic reference effect size). The impact of sample size (bootstrapped samples ranging from n=25 to n=3401) and cross-validation on effect size estimates was investigated across selected voxels with differing underlying effect sizes (including the brain-wide lowest effect size). Linear effects were estimated within training sets and then applied to unseen test set data, using 5-fold cross-validation. Resulting effect sizes (explained variance) were investigated.

Results: Analysis in the total sample (n=3401) without cross-validation yielded mainly negative correlations between BMI and gray matter density with a maximum effect size of R^2_p=.036 (peak voxel in the cerebellum). Effects were overestimated exponentially with decreasing sample size, with effect sizes up to R^2_p=.535 in samples of n=25 for the voxel with the brain-wide largest effect and up to R^2_p=.429 for the voxel with the brain-wide smallest effect. When applying cross-validation, linear effects estimated in small samples did not generalize to an independent test set. For the largest brain-wide effect a minimum sample size of n=100 was required to start generalizing (explained variance >0 in unseen data), while n=400 were needed for smaller effects of R^2_p=.005 to generalize. For a voxel with an underlying null effect, linear effects found in non-cross-validated samples did not generalize to test sets even with the maximum sample size of n=3401. Effect size estimates obtained with and without cross-validation approached convergence in large samples.

Discussion: Cross-validation is a useful method to counteract the overestimation of effect size particularly in small samples and to assess the generalizability of effects. Train and test set effect sizes converge in large samples which likely reflects a good generalizability for models in such samples. While linear effects start generalizing to unseen data in samples of n>100 for large effect sizes, the generalization of smaller effects requires larger samples (n>400). Cross-validation should be applied in voxel-based mass-univariate analysis to foster accurate effect size estimation and improve replicability of neuroimaging findings. We provide open-source python code for this purpose (https://osf.io/cy7fp/?view_only=a10fd0ee7b914f50820b5265f65f0cdb).

Introduction

Neuroimaging methods such as magnetic resonance imaging (MRI) have been used for several decades now to gain insights into the neurobiological underpinnings of psychological phenotypes. A plethora of scientific publications describes imaging-derived biomarkers of mental health disorders and perpetuates the hope to hope of utilizing such methods to aid clinical decision-making (Nour et al., 2022). However, recently the validity of findings involving relationships between neuroimaging and psychological phenotypes have been questioned due to accumulating reports of low replicability (Boekel et al., 2015; Genon et al., 2022; Marek et al., 2022). While such a replication crisis is not specific to neuroimaging research (e.g., for the field of psychology see Open Science Collaboration, 2015), its contributing factors may be particularly potent in this research domain due to high analytic flexibility in combination with a large number of statistical tests, which is likely to amplify publication bias (Botvinik-Nezer et al., 2020; Ioannidis, 2005; Jennings & Van Horn, 2012). Relatedly, replicability is undermined by small samples (Turner et al., 2018) which lead to unreliable and overestimated effect sizes (Button et al., 2013; Lane & Dunlap, 1978; Maxwell et al., 2008; Schönbrodt & Perugini, 2013). As studies particularly in neuroimaging research frequently use small samples (Elliott et al., 2020; Szucs & Ioannidis, 2020), this likely further contributes to low replicability. Recently, Marek et al. (2022) demonstrated that even without publication bias, associations between psychological and brain phenotypes are not replicable unless thousands of participants are included in the analysis. This seminal study has led to a vibrant discussion regarding replicability, sample size and effect size in the neuroimaging domain (Bandettini et al., 2022; Genon et al., 2022; Nour et al., 2022; Rosenberg & Finn, 2022; Spisak et al., 2023; Tervo-Clemmens et al., 2023). Notably, the demonstrated overestimation of effect sizes in small samples also occurs in statistically significant effects and, paradoxically, using more stringent significance thresholds even aggravates the problem (Lane & Dunlap, 1978; Marek et al., 2022).

In summary, the outlined evidence emphasizes an urgent need for solutions to increase the replicability of psychiatric neuroimaging findings. One possible solution that has been suggested

repeatedly is to validate brain effects in independent data via cross-validation (Klapwijk et al., 2021; Kriegeskorte et al., 2010; Rosenberg & Finn, 2022). While cross-validation methods are standardly used in multivariate brain analyses to identify and counteract overfitting and assure generalizability of models (e.g., Redlich et al., 2014, 2016; Repple et al., 2023; Schaffer, 1993), they are rarely ever used on univariate brain effects, likely also because common neuroimaging analysis software packages do not offer options to conduct cross-validation. However, specifically in the domain of neuroimaging research, cross-validation methods may be useful even for univariate models as the outlined overestimation of effects can be seen as an overfitting of models in the face of high analytic flexibility in combination with a high number of tests in this domain. Thus, we investigate the utility of cross-validation for accurate estimation of effect size in univariate voxel-based brain analysis. Body mass index (BMI) is used as an exemplary predictor due to previous reports of good replicability of BMI with brain structure, as well as its pivotal relevance for various mental disorders (Bond et al., 2014; McWhinney et al., 2022; Opel et al., 2015, 2021). Due to reports of relatively large effect size estimates of the association between BMI and brain structure even in large samples (maximum effect size corresponding to approximately 2.7% explained variance; Opel et al., 2021), this predictor is suitable to investigate the impact of cross-validation for a broader range of effect size, as compared to other predictors where effect size estimates across brain modalities have been shown to be considerably smaller (Marek et al., 2022; Winter et al., 2022). In order to evaluate the relevance of cross-validation for generalizability of effects sizes as a function of the sample size, we conducted our analyses across different sample sizes.

In light of the current lack of brain analysis software implementations regarding this matter, we provide open-source Python code to enable other researchers to apply cross-validation for voxel-based brain analysis.

Method

Participants

A total of n=3401 participants were included from three independent German cohorts: the Marburg-Münster Affective Disorders Cohort Study (MACS; n=1655), the Münster Neuroimaging Cohort (MNC; n=722) and the BiDirect cohort (n=1024). All three cohorts include individuals with major depressive disorder (MDD) and healthy controls (HC) free from any lifetime mental disorder diagnoses. General study methods, exclusion criteria and cohort characteristics were comprehensively described elsewhere (MACS: Kircher et al., 2019 and Vogelbacher et al., 2018; MNC: Dannlowski et al., 2016 and Opel et al., 2019; BiDirect: Teismann et al., 2014). See supplements for an additional description of data exclusion steps and sample characteristics specific to the current analysis.

Measures and procedure

BMI was calculated based on self-reported height and weight of participants. T1-weighted high-resolution anatomical brain images were acquired using 3T MRI scanner in all three studies. For the MACS sample two different MRI scanners were used at the recruitment sites in Marburg (Tim Trio, Siemens, Erlangen, Germany; combined with a 12-channel head matrix Rx-coil) and Münster (Prisma, Siemens, Erlangen, Germany; combined with a 20-channel head matrix Rx-coil). Data from the MNC and BiDirect samples were acquired using the same Gyroscan Intera scanner (later with Achieva update; Philips Medical Systems, Best, The Netherlands). Image preprocessing was conducted using the CAT12-toolbox (Gaser et al., 2022; https://neuro-jena.github.io/cat/) using default parameters for all samples. Briefly, images were bias-corrected, tissue classified, and normalized to MNI-space using linear (12-parameter affine) and non-linear transformations, within a unified model including high-dimensional geodesic shooting normalization (Ashburner & Friston, 2011). The modulated gray matter images were smoothed with a Gaussian kernel of 8 mm FWHM. Absolute threshold masking with a threshold value of 0.2 was used for all second level analyses as recommended for VBM

analyses (https://neuro-jena.github.io/cat12-help/). Image quality was assessed by visual inspection as well as by using the check for homogeneity function implemented in the CAT12 toolbox.

Statistical analysis

We investigated the association between BMI and gray matter brain structure using an established mass-univariate voxel-based morphometry (VBM) approach. To that end, BMI was used as a predictor in a general linear model (GLM) to predict voxel-wise gray matter density. The following nuisance parameters were included in the model: age, sex, total intracranial volume (TIV) and four dummy-coded scanner variables to control for scanner hardware differences. Two-sided whole brain effects of BMI were tested and partial R^2 (R^2_p) was used as a measure of effect size. To assess the impact of sample size and cross-validation on the estimation of voxel-based effect sizes the following steps were undertaken (for a schematic overview see Figure 1):

1. Classical non-cross-validated analysis was conducted in the maximum available sample without bootstrapping (n=3401). The resulting voxel-wise explained variance of BMI was used as an estimate for the realistic underlying effect size (in the following referred to as the *reference effect size*). Subsequently, voxels were selected covering a range of different representative effect sizes, including the voxel with the largest brain-wide reference effect size. In addition, the voxel with the smallest brain-wide reference effect size was selected to investigate effect size estimation based on an underlying null effect. An uncorrected significance threshold of p<.001 and extent threshold of k>200 was used for visualization purposes of brain-wide associations.
2. Based on the findings of Marek et al. (2022) we subsequently investigated the influence of sample size on the non-cross-validated effect sizes. Sample size was manipulated using samples of 18 different sizes: n=25, 35, 50, 70, 100, 150, 200, 300, 400, 600, 800, 1000, 1300, 1600, 2000, 2400, 2900, 3401. For each preselected voxel a bootstrapped resampling

distribution of effect size was obtained containing k=500 bootstrap runs per sample size. For each sample size the mean effect size was calculated across all bootstrap runs, as well as a 95% confidence interval (CI).

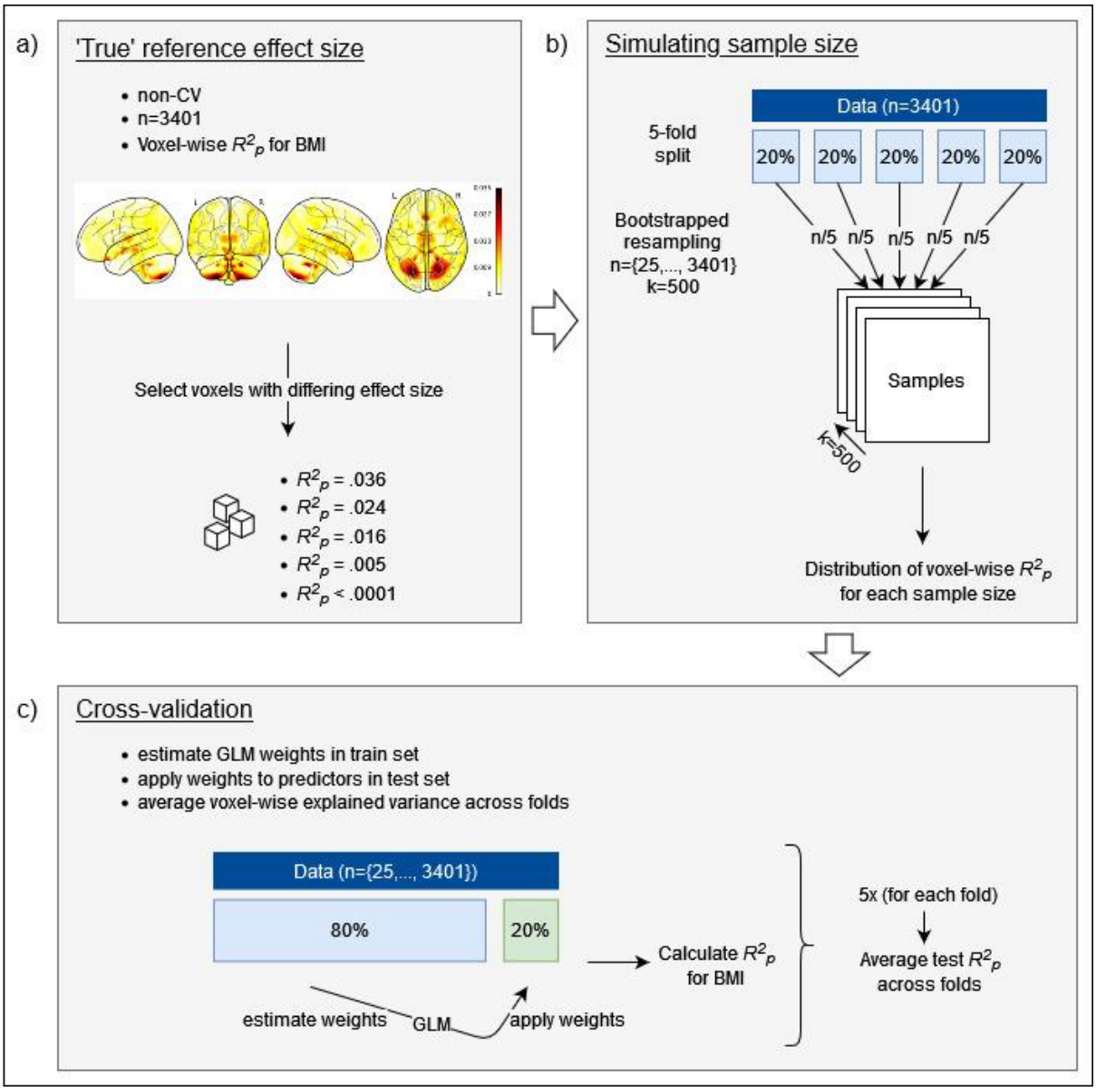
3. Finally, the impact of cross-validation on effect size estimates in the selected voxels was investigated across samples created with the resampling procedure described above. To this end, 5-fold cross-validation was applied by estimating linear effects using the GLM described above within respective train sets and then applying the resulting linear coefficients to the respective test sets. To this end the resulting beta coefficients yielded by model estimation within the train set was applied to the design matrix (individual predictor values) of the test set (i.e., data new to the trained linear model). The predictive value of the BMI predictor was evaluated by calculating the explained variance (based on residuals) with and without BMI as a predictor in the model. Then the mean R^2_p for BMI was calculated across the five test sets. This was used to quantify the extent to which the linear voxel-wise coefficients for BMI obtained from the train sets explain variance within unknown test data (i.e., generalization of effects to unseen data). Note that while R^2_p usually has a range from 0-1, it can become negative (and even exceed -1) in this case due to the application of linear coefficients to unseen data which can result in an effect size lower than prediction by the mean within the same sample (normally the baseline reference corresponding to an effect size R^2_p=0). A *point of initial generalizability* was defined as the minimum sample size necessary to achieve positive effect sizes in the test sets (i.e., the lower bound of the 95% CI above R^2_p=0). This point of initial generalizability can be interpreted as the minimum sample size needed for linear effects of a given effect size to generalize to unknown samples in a 5-fold cross-validation framework. The 5-fold cross-validation was chosen over a higher number of folds to be able to simulate very small (but commonly used) samples sizes. In order investigate the impact of this cross-validation on effect size estimation across different sample sizes we applied the same bootstrap approach described above also to cross-validation-based effect size estimates, resulting in resampling distributions of average test set effect sizes for each

sample size (95% CI were based on these bootstrap runs). To allow this combination of cross-validation and bootstrapping, we first conducted the 5-fold split of the total sample and then performed the bootstrapped resampling for each sample size by drawing with replacement from each split separately – thus only allowing replacement within one split but not across splits. This is necessary to avoid data leakage between train and test sets.

All analyses were conducted in Python (version 3.9.12) using the nilearn package (version 0.9.1) for voxel-based GLMs and k-Fold from the sklearn package (version 1.1.1) for cross-validation. The complete analysis code is provided online (https://osf.io/cy7fp/?view_only=a10fd0ee7b914f50820b5265f65f0cdb).

Figure 1

Analysis steps to obtain voxel-wise effect size estimates across sample sizes with and without cross-validation



Note. CV, cross-validation; BMI, Body Mass Index; GLM, general linear model.

Results

Associations between BMI and gray matter – identification of reference effect sizes in n=3401

The association between BMI and gray matter density in a non-cross-validated standard analysis is shown in Figure 2a and supplementary Figure S1. Liberal thresholding at uncorrected p<.001, k>200

resulted in widespread clusters across the brain. Effect size in significant voxels ranged up to R^2_p=.036. Peak voxel coordinates of significant clusters covering a wide range of effect sizes, as well as the voxel with the brain-wide smallest effect size were selected for further analysis (see Figure 2a). The selected voxels from largest to smallest effects were located in 1) the cerebellum (R^2_p=.036, p<.0001, x-24, y-72, z-60), 2) the posterior medial orbitofrontal cortex (mOFC; R^2_p=.024, p<.0001, x-4, y25, z-30), 3) the thalamus (R^2_p=.016, p<.0001, x6, y-12, z10), 4) the anterior mOFC (R^2_p=.005, p<.0001, x4, y69, z-6) and 5) the calcarine (R^2_p<.0001, p=.999, x-3, y-80, z12).

The impact of sample size on non-cross-validated effect size estimation

Effect size was systematically overestimated in small samples in the non-cross-validated analysis. Averaged across all bootstrapped samples with size n=25, effect size was inflated approximately .06 R^2_p units for all five voxels, resulting in a 2.5-fold inflation for the voxel with the largest reference effect size (R^2_p=.091 instead of R^2_p=.036) and 11.9-fold inflation for the voxel with a reference effect size of R^2_p=.005. Maximum effect size estimates went as high as R^2_p=.734 in samples of n=25. Even the voxel with the brain-wide lowest reference effect size (null-effect) reached a maximum effect size of up to R^2_p=.429 (average R^2_p=.056) in these samples with n=25. Average estimates of effect size (as well as maximum estimates) decreased exponentially with increasing sample size.

Inspection of partial correlation coefficients showed that a broad range of associations was found across bootstrapped samples, particularly with small sample sizes. For the voxel with the largest brain-wide reference effect size, associations ranging from a negative correlation r=-.73 to a positive correlation r=.60 were found in samples with n=25 (see supplementary Figure S2).

Detailed summary statistics for effect size estimates across sample sizes and voxels are presented in supplementary Tables S2-6. While the average effect size estimate across bootstrapped samples may be informative, it should be noted that the distribution of estimates was highly skewed with some

extreme outliers. This distribution across single runs from the resampling procedure is further visualized in supplementary Figure S3.

The impact of cross-validation on effect size estimation

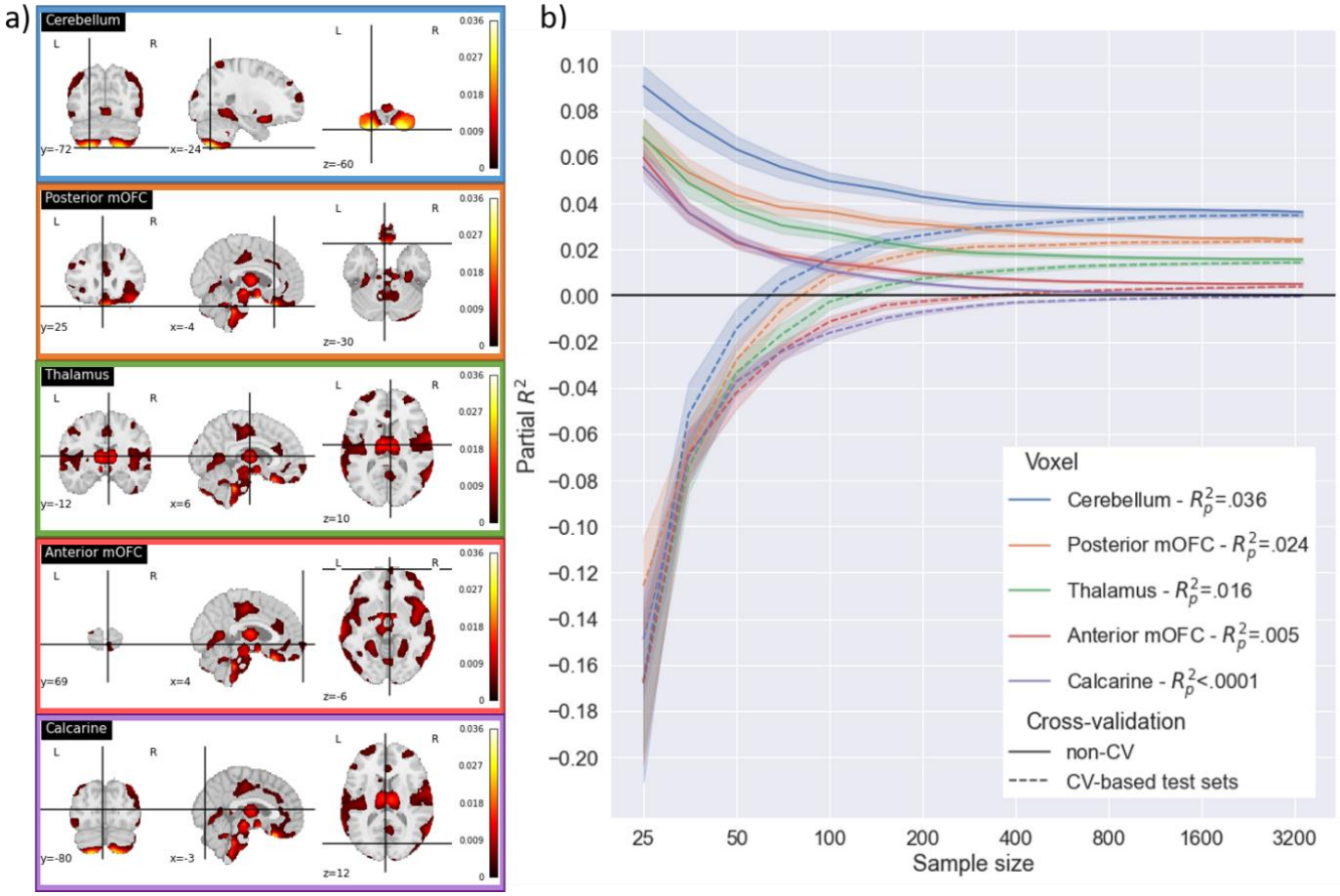
On average test set effect size estimates were descriptively lower compared to non-cross-validated effect sizes across all sample sizes and all voxels (this was true for 93.83% of bootstrapped samples across all sample sizes). This disparity was particularly strong in small samples where non-cross-validated effect sizes were largest while test set effect sizes were mostly negative (indicating no generalization of linear effects to unknown samples).

A *point of initial generalizability* was reached after n=100 for the largest reference effect size, while larger samples were needed to reach this point for voxels with smaller effect size: n=100 for R^2_p=.024, n=150 for R^2_p=.016, and n=400 for R^2_p=.005. For the voxel with the minimum brain-wide effect size this *point of initial generalizability* was never reached, as expectable for an underlying null-effect. In fact, for this voxel even the mean and upper bound of the 95% CI were always below an effect size of zero at any sample size, when applying cross-validation, indicating effective protection against false-positives (see supplementary Tables S2-6).

Estimates derived from non-cross-validated analysis and from test sets using cross-validation approached convergence in larger samples. In samples with n=3401, non- cross-validated and test set effect sizes differed only marginally (average difference across voxels: R^2_p=.001). However, full convergence was never reached between non-cross-validated and test set effect size estimates, meaning that confidence intervals did not overlap, even at largest sample sizes. Detailed descriptive statistics for effect size estimates across sample sizes, voxels and cross-validation are given in supplementary Tables S2-6.

Figure 2

The association between Body Mass Index and gray matter density - selected voxel locations and their effect size estimates across sample sizes, with and without and cross-validation



Note. a) Shows non-cross-validated two-sided effect of BMI on gray matter density at a liberal uncorrected threshold of p<.001 in the maximum sample of n=3401. The location of selected peak voxel coordinates in the cerebellum (largest brain-wide effect), posterior medial prefrontal cortex (mOFC), thalamus, anterior mOFC and calcarine (smallest brain-wide effect) are shown. Color bars represent R^2_p *values. b) Shows effect size estimates for the five different voxels across sample sizes (resampling with k=500 per sample size). Non-cross-validated (non-CV) effect sizes are presented with solid lines and test set effect sizes with dotted lines. Error bands represent the 95% confidence interval. Respective effect sizes of selected voxels in non-cross-validated analysis of full sample are presented in the legend (ranging from* η^2_p*=.036 to* η^2_p*<.0001).*

Discussion

Using BMI as an exemplary predictor variable, we investigated the utility of cross-validation for the accurate estimation of effect sizes of univariate brain-wide associations. Replicating previous findings (Marek et al., 2022), we find that effect size estimates are highly overestimated and unreliable in small samples if no cross-validation is applied. Furthermore, we demonstrate that cross-validation can be used to reveal poor generalizability to new data of these overestimated effects in small samples (indicated by negative effect size estimates). In larger samples, cross-validation-derived test set effect sizes start becoming positive, suggesting a *point of initial generalizability* that can be interpreted as the minimum sample size necessary to achieve generalizable linear effect estimates. Importantly, sample sizes of several hundreds of participants may be sufficient to accurately estimate brain-wide associations, given sufficiently large underlying effect sizes. The cross-validation approach could facilitate a differentiation between artificially inflated large effects and robust 'true' large effects. The standardly implementation of cross-validation to assess the generalizability of brain-wide effects should be considered in addition to conventional reporting of significance and traditional effect size estimates. The utilization of cross-validation to facilitate generalizability of neuroimaging findings has been repeatedly suggested (Klapwijk et al., 2021; Kriegeskorte et al., 2010; Rosenberg & Finn, 2022). However, to the best of our knowledge, we are the first to systematically investigate the utility of cross-validation for effect size estimation in brain-wide univariate analysis.

The presented overestimation of effect sizes in small samples falls in line with previous findings in the literature (Button et al., 2013; Lane & Dunlap, 1978; Maxwell et al., 2008; Schönbrodt & Perugini, 2013). Marek et al. (2022) demonstrated that even the largest brain-wide associations of compound cognitive and psychopathological variables with brain structure and function corresponded to less than 0.5% explained variance, which was inflated to up to approximately 40% explained variance in small samples. The authors conclude that thousands of individuals are needed for robust and replicable estimates in this research domain. While these findings are striking and important, the authors' conclusion has been questioned by various scholars. The main critique is that true effect

sizes may be considerably higher under ideal conditions leading to smaller sample sizes necessary for accurate estimation of effects (DeYoung et al., 2022). In a study of over 1800 adults, Winter et al. (2022) investigated univariate brain-wide, cross-modal brain differences between HC and lifetime MDD individuals and conclude that largest effect sizes go up to 1.7% explained variance. However, they further report that comparisons of HC with chronic or acute MDD individuals yield higher effects sizes of up to 2.7% explained variance (in n>900). Further, studies suggest that effect sizes are larger in other diagnostic groups such as psychotic disorders (Hettwer et al., 2022). Similarly, Repple and Gruber et al. (2022) present transdiagnostic structural connectome alterations, with largest effect sizes found in patients with schizophrenia, possibly rendering smaller samples sufficient to detect effects.

Our findings expand the results by Marek et al. (2022) to a voxel-based analysis framework and to a broader range of effect sizes. Importantly, we provide evidence that for larger effect sizes – which can occur under specific conditions as outlined above – hundreds of participants could suffice to accurately estimate linear brain-wide effects. When smaller effect sizes are assumed, our findings are highly comparable to the findings by Marek et al. as we similarly find that a 'true' effect size of approximately ~0.5% explained variance requires very large samples to obtain robust estimates of effects, although cross-validation may enable the robust detection of accurate estimates already in somewhat smaller samples (n>400). The ongoing debate surrounding effect size and sample size in the neuroimaging domain stresses the importance of consistent reporting of effect sizes in publications, as well as interpreting effect sizes in the context of sample size.

While the above considerations could guide sample size planning of neuroimaging studies, a problem arises when the 'true' effect size is unknown (and effect size inflations in small samples make it particularly difficult to estimate this solely from existing literature). How can researchers know if an obtained effect size is accurate or inflated? Using a cross-validation framework, we demonstrate that the application of identified linear effects to unseen data can be utilized to identify inflated, non-generalizable effect sizes. Notably, this procedure can be applied in smaller studies to verify if a

putative large effect size is robust and may warrant the use of a smaller sample. Importantly, our findings are in line with low test set performance merely reflecting an underpowered sample and not necessarily meaning that an effect does not exist (Helwegen et al., 2023). In other words, even substantial effects of 3.6% explained variance can barely be differentiated from null effects in small samples based on cross-validation-derived effect sizes alone. However, the relative congruence between non-cross-validated and cross-validation-derived effect sizes can provide a strong argument for robust and replicable linear associations with realistic effect size estimates. Thus, calculating non-cross-validated and cross-validation-derived effect sizes combined with a thorough inspection of the similarity between the resulting estimates could be informative. Notably, while cross-validation decreases the probability of effect size inflation, it does not eliminate it. Even unrealistically high effect size estimates can in rare cases generalize to unseen test set data, particularly in small samples, as shown by some extreme outliers in our results.

While it is difficult to deduct a clear recommendation for sample sizes based on our cross-validation analyses, we propose that a *point of initial generalizability* can be defined as the sample size where linear effects start explaining variance in unseen data. This aspect could expand the traditional power analysis framework by the question of generalizability in addition to significance. In other words, while traditional power analysis answers the question of how large a sample needs to be for a given effect to become *significant*, our results open up the possibility for defining sample sizes necessary for linear effects to *generalize* to unseen data.

In large samples effect sizes became highly stable and barely differed whether cross-validation was applied or not. This finding implicates that if several thousands of individuals are available (e.g., due to consortia data), within sample k-fold cross-validation may barely alter the result, regardless of the underlying true effect size.

The current study entails several important limitations. Firstly, it is unclear whether our findings can be generalized to other cross-validation methods. While 5-fold cross-validation was chosen for the current analysis so that small samples of n=25 could be included, other splits (e.g., 10-fold) could in

principle also be suitable to be applied in voxel-based univariate brain analysis. While a systematic comparison of different cross-validation methods for the application to univariate voxel-based analysis is beyond the scope of this work, this may be a fruitful task for future studies. Further, it is open for discussion to what extent our findings are generalizable to other MRI modalities, such as functional MRI and parcellation-based brain analysis, as well as to predictors of other domains. We believe that generalizability to other voxel-based imaging modalities and other predictors is given, as the general pattern of results should not be specific to VBM effects of BMI. Parcellation-based brain analysis approaches could also profit from cross-validation. However, cross-validation could possibly be particularly beneficial in settings that are at higher risk for overfitting statistical models. Such risk is likely to be higher in more complex models, and in settings with higher analytical flexibility and a higher number of statistical tests (the latter being higher in voxel-based as compared to parcellation-based analysis). Interestingly, cross-validation does not seem to ultimately protect from an inflation of effect size in all settings. It has been shown that in complex multivariate modelling of voxel-based associations with psychopathology, even cross-validation-based performance measures are inflated in underpowered studies (Flint et al., 2021). In the same vein, we demonstrated that even cross-validation-derived effect sizes can be inflated in small samples although the probability of an overestimation of effects is lower as compared to non-cross-validated estimates (as outlined above).

In summary the utilization of cross-validation can contribute two major benefits: 1) identify if a study is underpowered and corresponding effect size estimates are likely to be highly inflated and 2) corroborate the accuracy of an effect size estimate of a sufficiently powered study (also for underlying null-effects). Thus, we propose that cross-validation procedures should be applied to foster replicability of neuroimaging research and facilitate accurate estimation of effect sizes. Further, we provide concrete steps for an application of cross-validation to mass-univariate voxel-based analysis, as well as corresponding open-source Python code.

References

Ashburner, J., & Friston, K. J. (2011). Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage*, *55*(3), 954–967. https://doi.org/10.1016/j.neuroimage.2010.12.049

Bandettini, P. A., Gonzalez-Castillo, J., Handwerker, D., Taylor, P., Chen, G., & Thomas, A. (2022). The challenge of BWAS: Unknown unknowns in feature space and variance. *Med*, *3*(8), 526–531.

Boekel, W., Wagenmakers, E. J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Bond, D. J., Ha, T. H., Lang, D. J., Su, W., Torres, I. J., Honer, W. G., Lam, R. W., & Yatham, L. N. (2014). Body mass index-related regional gray and white matter volume reductions in first-episode mania patients. *Biological Psychiatry*, *76*(2), 138–145. https://doi.org/10.1016/j.biopsych.2013.08.030

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. De, 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*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Dannlowski, U., Kugel, H., Grotegerd, D., Redlich, R., Opel, N., Dohm, K., Zaremba, D., Grögler, A., Schwieren, J., Suslow, T., Ohrmann, P., Bauer, J., Krug, A., Kircher, T., Jansen, A., Domschke, K.,

Hohoff, C., Zwitserlood, P., Heinrichs, M., Arolt, V., Heindel, W., & Baune, B. T. (2016). Disadvantage of Social Sensitivity: Interaction of Oxytocin Receptor Genotype and Child Maltreatment on Brain Structure. *Biological Psychiatry*, *80*(5), 398–405. https://doi.org/10.1016/j.biopsych.2015.12.010

DeYoung, C. G., Sassenberg, T. A., Abend, R., Allen, T. A., Beaty, R. E., Bellgrove, M. A., Blain, S. D., Bzdok, D., Chavez, R. S., Engel, S. A., FFeilong, M., Fornito, A., Genc, E., Goghari, V., Grazioplene, R. G., Hanson, J. L., Haxb, J. V, Hilger, K., Homan, P., Joyyner, K., Kaczkurkin, A. N., Latzman, R. D., Martin, E. A., Passamonti, L., Pickering, A. D., Safron, A., Servaas, M. N., Smillie, L. D., Spreng, R. N., Tiego, J., Viding, E., & Wacker, J. (2022). Reproducible between-person brain-behavior associations do not always require thousands of individuals. *PsyArXiv*, *5*(1), 47–55.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D. M. A., Emden, D., Winter, N. R., Leenings, R., Eickhoff, S. B., Kircher, T., Krug, A., Nenadic, I., Arolt, V., Clark, S., Baune, B. T., Jiang, X., Dannlowski, U., & Hahn, T. (2021). Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, *46*(8), 1510–1517. https://doi.org/10.1038/s41386-021-01020-7

Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & Alzheimer's Disease Neuroimaging Initiative. (2022). CAT – A Computational Anatomy Toolbox for the Analysis of Structural MRI Data. *BioRxiv*.

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, *23*(5), 307–318. https://doi.org/10.1038/s41583-022-00584-7

Helwegen, K., Libedinsky, I., & Heuvel, M. P. Van Den. (2023). Statistical power in network neuroscience. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.12.011

Hettwer, M. D., Larivière, S., Park, B. Y., van den Heuvel, O. A., Schmaal, L., Andreassen, O. A., Ching, C. R. K., Hoogman, M., Buitelaar, J., Veltman, D. J., Stein, D. J., Franke, B., van Erp, T. G. M., Jahanshad, N., Thompson, P. M., Thomopoulos, S. I., Bethlehem, R. A. I., Bernhardt, B. C., Eickhoff, S. B., & Valk, S. L. (2022). Coordinated Cortical Thickness Alterations across Psychiatric Conditions: A Transdiagnostic ENIGMA Study. *MedRxiv*, 2022.02.03.22270326. https://doi.org/10.1101/2022.02.03.22270326

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jennings, R. G., & Van Horn, J. D. (2012). Publication bias in neuroimaging research: Implications for meta-analyses. *Neuroinformatics*, *10*(1), 67–80. https://doi.org/10.1007/s12021-011-9125-y

Kircher, T., Wöhr, M., Nenadic, I., Schwarting, R., Schratt, G., Alferink, J., Culmsee, C., Garn, H., Hahn, T., Müller-Myhsok, B., Dempfle, A., Hahmann, M., Jansen, A., Pfefferle, P., Renz, H., Rietschel, M., Witt, S. H., Nöthen, M. M., Krug, A., Dannlowski, U., Nenadić, I., Schwarting, R., Schratt, G., Alferink, J., Culmsee, C., Garn, H., Hahn, T., Müller-Myhsok, B., Dempfle, A., Hahmann, M., Jansen, A., Pfefferle, P., Ranz, H., Rietschel, M., Witt, S. H., Nöthen, M. M., Krug, A., & Dannlowski, U. (2019). Neurobiology of the major psychoses. A translational perspective on brain structure and function: the FOR2107 consortium. *European Archives of Psychiatry and Clinical Neurosciences*, *269*, 949–962. https://doi.org/10.1007/s00406-018-0943-x

Klapwijk, E. T., Bos, W. Van Den, Tamnes, C. K., Raschle, N. M., & Mills, K. L. (2021). Opportunities for

increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience*, *47*, 100902. https://doi.org/10.1016/j.dcn.2020.100902

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow and Metabolism*, *30*(9), 1551–1557. https://doi.org/10.1038/jcbfm.2010.86

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

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. https://doi.org/10.1038/s41586-022-04492-9

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McWhinney, S. R., Brosch, K., Calhoun, V. D., Crespo-Facorro, B., Crossley, N. A., Dannlowski, U., Dickie, E., Dietze, L. M. F., Donohoe, G., Du Plessis, S., Ehrlich, S., Emsley, R., Furstova, P., Glahn, D. C., Gonzalez- Valderrama, A., Grotegerd, D., Holleran, L., Kircher, T. T. J., Knytl, P., Kolenic, M., Lencer, R., Nenadić, I., Opel, N., Pfarr, J.-K., Rodrigue, A. L., Rootes-Murdy, K., Ross, A. J., Sim, K., Škoch, A., Spaniel, F., Stein, F., Švancer, P., Tordesillas-Gutiérrez, D., Undurraga, J., Váquez-Bourgon, J., Voineskos, A., Walton, E., Weickert, T. W., Weickert, C. S., Thompson, P. M., van Erp, T. G. M., Turner, J. A., & Hajek, T. (2022). Obesity and brain structure in schizophrenia – ENIGMA study in 3021 individuals. *Molecular Psychiatry*, *27*(9), 3731–3737. https://doi.org/10.1038/s41380-022-01616-5

Nour, M. M., Liu, Y., & Dolan, R. J. (2022). Functional neuroimaging in psychiatry and the case for failing better. *Neuron*, *110*(16), 2524–2544. https://doi.org/10.1016/j.neuron.2022.07.005

Opel, N., Redlich, R., Dohm, K., Zaremba, D., Goltermann, J., Repple, J., Kaehler, C., Grotegerd, D., Leehr, E. J., Böhnlein, J., Förster, K., Meinert, S., Enneking, V., Sindermann, L., Dzvonyar, F., Emden, D., Leenings, R., Winter, N., Hahn, T., Kugel, H., Heindel, W., Buhlmann, U., Baune, B. T., Arolt, V., & Dannlowski, U. (2019). Mediation of the influence of childhood maltreatment on depression relapse by cortical structure: a 2-year longitudinal observational study. *Lancet Psychiatry*, *6*(4), 318–326. https://doi.org/10.1016/S2215-0366(19)30044-6

Opel, N., Redlich, R., Grotegerd, D., Dohm, K., Heindel, W., Kugel, H., Arolt, V., & Dannlowski, U. (2015). Obesity and major depression: Body-mass index (BMI) is associated with a severe course of disease and specific neurostructural alterations. *Psychoneuroendocrinology*, *51*, 219–226. https://doi.org/10.1016/j.psyneuen.2014.10.001

Opel, N., Thalamuthu, A., Milaneschi, Y., Grotegerd, D., Flint, C., Leenings, R., Goltermann, J., Richter, M., Hahn, T., Woditsch, G., Berger, K., Hermesdorf, M., McIntosh, A. M., Whalley, H. C., Harris, M. A., MacMaster, F. P., Walter, H., Veer, I. M., Frodl, T., Carballedo, A., Krug, A., Nenadic, I., Kircher, T., Aleman, A., Groenewold, N. A., Stein, D. J., Soares, J. C., Zunta-Soares, G. B., Mwangi, B., Wu, M.-J., Walter, M., Li, M., Harrison, B. J., Davey, C. G., Cullen, K. R., Klimes-Dougan, B., Mueller, B. A., Sämann, P. G., Penninx, B., Nawijn, L., Veltman, D. J., Aftanas, L. I., Brak, I. V.,

Filimonova, E. A., Osipov, E. A., Reneman, L., Schrantee, A., Grabe, H. J., Van der Auwera, S., Wittfeld, K., Hosten, N., Völzke, H., Sim, K., Gotlib, I. H., Sacchet, M. D., Lagopoulos, J., Hatton, S. N., Hickie, I. B., Pozzi, E., Thompson, P. M., Jahanshad, N., Schmaal, L., Baune, B. T., & Dannlowski, U. (2021). Brain structural abnormalities in obesity: relation to age, genetic risk, and common psychiatric disorders. *Molecular Psychiatry*, *26*(9), 4839–4852. https://doi.org/10.1038/s41380-020-0774-9

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Redlich, R., Almeida, J. R., Grotegerd, D., Opel, N., Kugel, H., Heindel, W., Arolt, V., Phillips, M. L., & Dannlowski, U. (2014). Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry-pattern classification approach. *JAMA Psychiatry*, *71*(11), 1222–1230. https://doi.org/10.1001/jamapsychiatry.2014.1100

Redlich, R., Opel, N., Grotegerd, D., Dohm, K., Zaremba, D., Burger, C., Munker, S., Muhlmann, L., Wahl, P., Heindel, W., Arolt, V., Alferink, J., Zwanzger, P., Zavorotnyy, M., Kugel, H., & Dannlowski, U. (2016). Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA Psychiatry*, *73*(6), 557–564. https://doi.org/10.1001/jamapsychiatry.2016.0316

Repple, J., Gruber, M., Mauritz, M., de Lange, S. C., Winter, N. R., Opel, N., Goltermann, J., Meinert, S., Grotegerd, D., Leehr, E. J., Enneking, V., Borgers, T., Klug, M., Lemke, H., Waltemate, L., Thiel, K., Winter, A., Breuer, F., Grumbach, P., Hofmann, H., Stein, F., Brosch, K., Ringwald, K. G., Pfarr, J., Thomas-Odenthal, F., Meller, T., Jansen, A., Nenadic, I., Redlich, R., Bauer, J., Kircher, T., Hahn, T., van den Heuvel, M., & Dannlowski, U. (2022). Shared and specific patterns of structural brain connectivity across affective and psychotic disorders. *Biological Psychiatry*. https://doi.org/10.1016/J.BIOPSYCH.2022.05.031

Repple, J., Gruber, M., Mauritz, M., de Lange, S. C., Winter, N. R., Opel, N., Goltermann, J., Meinert, S., Grotegerd, D., Leehr, E. J., Enneking, V., Borgers, T., Klug, M., Lemke, H., Waltemate, L., Thiel, K., Winter, A., Breuer, F., Grumbach, P., Hofmann, H., Stein, F., Brosch, K., Ringwald, K. G., Pfarr, J., Thomas-Odenthal, F., Meller, T., Jansen, A., Nenadic, I., Redlich, R., Bauer, J., Kircher, T., Hahn, T., van den Heuvel, M., & Dannlowski, U. (2023). Shared and Specific Patterns of Structural Brain Connectivity Across Affective and Psychotic Disorders. *Biological Psychiatry*, *93*(2), 178–186. https://doi.org/10.1016/j.biopsych.2022.05.031

Rosenberg, M. D., & Finn, E. S. (2022). How to establish robust brain–behavior relationships without thousands of individuals. *Nature Neuroscience*, *25*(7), 835–837. https://doi.org/10.1038/s41593-022-01110-9

Schaffer, C. (1993). Selecting a Classification Method by Cross-Validation. *Machine Learning*, *13*(1), 135–143. https://doi.org/10.1023/A:1022639714137

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Spisak, T., Bingel, U., & Wager, T. (2023). Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*, E4–E7. https://doi.org/https://doi.org/10.1101/2022.06.22.497072

Szucs, D., & Ioannidis, J. P. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage*, *221*(October 2019), 117164. https://doi.org/10.1016/j.neuroimage.2020.117164

Teismann, H., Wersching, H., Nagel, M., Arolt, V., Heindel, W., Baune, B. T., Wellmann, J., Hense, H.-W., & Berger, K. (2014). Establishing the bidirectional relationship between depression and subclinical arteriosclerosis - rationale, design, and characteristics of the BiDirect Study. *BMC*

Psychiatry, *14*(174). https://doi.org/10.1186/1471-244X-14-174

Tervo-Clemmens, B., Marek, S., Chauvin, R. J., Van, A. N., Kay, B. P., Laumann, T. O., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., & Dosenbach, N. U. F. (2023). Reply to: Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*, E8–E12. https://doi.org/https://doi.org/10.1101/2022.06.22.497072

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(62). https://doi.org/10.1038/s42003-018-0073-z

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*(October 2016), 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Vogelbacher, C., Möbius, T. W. D., Sommer, J., Schuster, V., Dannlowski, U., Kircher, T., Dempfle, A., Jansen, A., & Bopp, M. H. A. (2018). The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. *NeuroImage*, *172*, 450–460. https://doi.org/10.1016/j.neuroimage.2018.01.079

Winter, N. R., Leenings, R., Ernsting, J., Sarink, K., Fisch, L., Emden, D., Blanke, J., Goltermann, J., Opel, N., Barkhau, C., Meinert, S., Dohm, K., Repple, J., Mauritz, M., Gruber, M., Leehr, E. J., Grotegerd, D., Redlich, R., Jansen, A., Nenadic, I., Nöthen, M. M., Forstner, A., Rietschel, M., Groß, J., Bauer, J., Heindel, W., Andlauer, T., Eickhoff, S. B., Kircher, T., Dannlowski, U., & Hahn, T. (2022). Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2022.1780

Funding and Disclosures

The MACS and MNC studies are funded by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to UD; SFB-TRR58, Projects C09 and Z02 to UD), the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Munster (grant Dan3/012/17 to UD), IMF Munster RE111604 to RR und RE111722 to RR, IMF Munster RE 22 17 07 to Jonathan Repple and the Deanery of the Medical Faculty of the University of Munster. TH was supported by the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, HA7070/4). MP was supported by an ERC Consolidator grant (ERC-COG 101001062) and a NWO VIDI grant of the Dutch Research Council (Netherlands Organisation for Scientific Research Grant VIDI-452-16-015). The BiDirect study is funded by German Federal Ministry of Education and Research Grant Nos. 01ER0816, 01ER1506, and 01ER1205. Biomedical financial interests or potential conflicts of interest: TK received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, neuraxpharm. This cooperation has no relevance to the work that is covered in the manuscript.

Acknowledgements

This work is part of the German multicenter consortium “Neurobiology of Affective Disorders. A translational perspective on brain structure and function“, funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; Forschungsgruppe/Research Unit FOR2107).

Principal investigators (PIs) with respective areas of responsibility in the FOR2107 consortium are: Work Package WP1, FOR2107/MACS cohort and brainimaging: Tilo Kircher (speaker FOR2107; DFG grant numbers KI 588/14-1, KI 588/14-2), Udo Dannlowski (co-speaker FOR2107; DA 1151/5-1, DA 1151/5-2), Axel Krug (KR 3822/5-1, KR 3822/7-2), Igor Nenadic (NE 2254/1-2), Carsten Konrad (KO 4291/3-1). CP1, biobank: Petra Pfefferle (PF 784/1-1, PF 784/1-2), Harald Renz (RE 737/20-1, 737/20-2). CP2, administration. Tilo Kircher (KI 588/15-1, KI 588/17-1), Udo Dannlowski (DA 1151/6-1),

Data access and responsibility: All PIs take responsibility for the integrity of the respective study data and their components. All authors and coauthors had full access to all study data.

Acknowledgements and members by Work Package (WP): WP1: Henrike Bröhl, Katharina Brosch, Bruno Dietsche, Rozbeh Elahi, Jennifer Engelen, Sabine Fischer, Jessica Heinen, Svenja Klingel, Felicitas Meier, Tina Meller, Torsten Sauder, Simon Schmitt, Frederike Stein, Annette Tittmar, Dilara Yüksel (Dept. of Psychiatry, Marburg University). Mechthild Wallnig, Rita Werner (Core-Facility Brainimaging, Marburg University). Carmen Schade-Brittinger, Maik Hahmann (Coordinating Centre for Clinical Trials, Marburg). Michael Putzke (Psychiatric Hospital, Friedberg). Rolf Speier, Lutz Lenhard (Psychiatric Hospital, Haina). Birgit Köhnlein (Psychiatric Practice, Marburg). Peter Wulf, Jürgen Kleebach, Achim Becker (Psychiatric Hospital Hephata, Schwalmstadt-Treysa). Ruth Bär (Care facility Bischoff, Neunkirchen). Matthias Müller, Michael Franz, Siegfried Scharmann, Anja Haag, Kristina Spenner, Ulrich Ohlenschläger (Psychiatric Hospital Vitos, Marburg). Matthias Müller, Michael Franz, Bernd Kundermann (Psychiatric Hospital Vitos, Gießen). Christian Bürger, Katharina Dohm, Fanni Dzvonyar, Verena Enneking, Stella Fingas, Katharina Förster, Janik Goltermann, Dominik Grotegerd, Hannah Lemke, Susanne Meinert, Nils Opel, Ronny Redlich, Jonathan Repple, Kordula Vorspohl, Bettina Walden, Dario Zaremba (Dept. of Psychiatry, University of Münster). Harald Kugel, Jochen Bauer, Walter Heindel, Birgit Vahrenkamp (Dept. of Clinical Radiology, University of Münster). Gereon Heuft, Gudrun Schneider (Dept. of Psychosomatics and Psychotherapy, University of Münster). Thomas Reker (LWL-Hospital Münster). Gisela Bartling (IPP Münster). Ulrike Buhlmann (Dept. of Clinical Psychology, University of Münster).

CP1: Julian Glandorf, Fabian Kormann, Arif Alkan, Fatana Wedi, Lea Henning, Alena Renker, Karina Schneider, Elisabeth Folwarczny, Dana Stenzel, Kai Wenk, Felix Picard, Alexandra Fischer, Sandra Blumenau, Beate Kleb, Doris Finholdt, Elisabeth Kinder, Tamara Wüst, Elvira Przypadlo, Corinna Brehm (Comprehensive Biomaterial Bank Marburg, Marburg University). Supplementary information is available at MP's website.