Gray matter correlates of childhood maltreatment in the context of major depression: searching for replicability in a multi-cohort brain-wide association study of 3225 adults

Janik Goltermann[1§], Nils Winter[1], Susanne Meinert[1,2], Dominik Grotegerd[1], Anna Kraus[1], Kira Flinkenflügel[1], Luisa Altegoer[1], Judith Krieger[1], Elisabeth J. Leehr[1], Joscha Böhnlein[1], Linda M. Bonnekoh[1,3], Maike Richter[1,4], Tim Hahn[1], Lukas Fisch[1], Marius Gruber[1,5], Marco Hermesdorf[6], Klaus Berger[6], Volker Arolt[1], Katharina Brosch[7,8], Frederike Stein[7], Florian Thomas-Odenthal[7], Paula Usemann[7], Lea Teutenberg[7], Vincent Hammes[7], Hamidreza Jamalabadi[7], Nina Alexander[7], Benjamin Straube[7], Andreas Jansen[7], Igor Nenadić[7], Tilo Kircher[7], Nils Opel[4,9,10*], Udo Dannlowski[1*]

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

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

[3]Clinic for Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Münster, Germany

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

[5]Department of Psychiatry, Psychosomatic Medicine, and Psychotherapy, Goethe University Frankfurt, University Hospital, Frankfurt, Germany

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

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

[8]Institute of Behavioral Science, Feinstein Institutes for Medical Research, Manhasset, NY, USA

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

[10]German Center for Mental Health (DZPG), Germany

§Corresponding author

*Authors contributed equally to the present work

Corresponding author contact details:

Albert-Schweitzer-Campus 1, Building A9a, 48149 Münster, Germany

Email: jgolt@uni-muenster.de

Phone: +49 251 83 51880

Abstract

Background: Childhood maltreatment has been associated with gray matter alterations, particularly within limbic and prefrontal regions. However, findings are heterogeneous, potentially due to differing methodologies and sample characteristics. Here, we investigate the cross-cohort replicability of gray matter correlates of childhood maltreatment across large clinical and non-clinical adult samples using harmonized assessment, preprocessing and analysis pipelines.

Methods: Three independent adult cohorts comprising a total of N=3225 individuals (healthy control [HC]: n=1898 and participants with major depressive disorder [MDD]: n=1327) underwent structural MRI and maltreatment assessment via the Childhood Trauma Questionnaire (CTQ). Associations between childhood maltreatment and voxel-based gray matter volume (GMV) were tested on a wholebrain level in two steps: 1) pooling all three cohorts together to harvest maximum statistical power (applying a voxel-wise FWE-corrected threshold of $p_{FWE}<.05$) and 2) investigating the replicability of effects by assessing cross-cohort spatial overlap of significant voxels at two liberal uncorrected thresholds ($p_{unc}<.001$ and $p_{unc}<.01$). Twelve statistical models were tested, that varied in maltreatment operationalizations, subsamples and covariates.

Results: Pooling cohorts yielded no significant maltreatment-GMV associations when controlling for lifetime MDD diagnosis. Dropping MDD diagnosis as a covariate yielded significant negative effects of maltreatment within widespread clusters across temporal regions, a fusiform-lingual-parahippocampal complex, the thalamus and the orbitofrontal cortex (k=4970, $p_{FWE}<.05$). Including only HC subsamples, small clusters emerged either when using the CTQ sum score (k=99, $p_{FWE}<.05$, orbitofrontal) or when investigating severe forms of maltreatment in HCs (k=132, $p_{FWE}<.05$, cerebellum). The largest effect size when pooling all three cohorts was partial $R^2=.022$. Replicability analyses using a liberal uncorrected thresholding at $p_{unc}<.001$ yielded maltreatment-GMV associations within all single cohorts and across all statistical models. However, these associations were effectively non-replicable across cohorts, which was largely consistent across statistical models. Even extending the significance threshold to a liberal threshold of $p_{unc}<.01$ yielded only marginal replicability across cohorts.

Conclusions: Gray matter correlates of childhood maltreatment, measured with the CTQ, are non-replicable across large cohorts when adequately controlling for depression diagnosis, even when employing harmonized study protocols, lenient statistical thresholds and exploring various maltreatment operationalizations and subgroups. Previous findings may have been inflated by inadequate control for confounding diagnosis effects or due to publication bias. Our findings underscore the importance of a paradigm shift towards investigating the replicability of neuroimaging findings.

Introduction

Childhood maltreatment (CM) has been identified to be one of the most important risk factors for the development of affective disorders (Lippard & Nemeroff, 2020; Struck et al., 2020) and is associated with chronic disease trajectories and poorer treatment outcomes in major depressive disorder (MDD) (Lippard & Nemeroff, 2020; Nelson et al., 2017). Within the past two decades a plethora of neuroimaging studies has repeatedly suggested that experiences of abuse and neglect during childhood are associated with neurobiological alterations in adults (Dannlowski et al., 2012; Goltermann et al., 2022; McLaughlin et al., 2019; Opel, Redlich, Dohm et al., 2019; Teicher et al., 2012, 2016). Brain regions where these effects have been localized overlap with neural correlates of MDD, giving rise to the notion that neurobiological alterations may mediate the unfavorable effects of CM on clinical trajectories (Meinert et al., 2019; Opel et al., 2014). Thus, studying the neurobiological correlates of CM could give insights into the mechanistic processes of its clinical consequences, potentially informing the optimization of treatments or preventative measures for this population (Teicher & Samson, 2013).

In adults, CM effects on gray matter structure have been observed in an array of regions, with most frequent findings implying the hippocampus, amygdala, dorsolateral prefrontal cortex, insula and anterior cingulate cortex (Lim et al., 2014; McLaughlin et al., 2019; Paquola et al., 2016; Pollok et al., 2022; Teicher et al., 2016). However, the investigation of CM-associated gray matter alterations has yielded considerable heterogeneity in findings regarding localization of effects. Importantly, large-scale consortium studies and meta-analyses do not find these aforementioned regions, but rather report a multitude of other areas to be associated with CM, including, among others the postcentral gyrus and occipital regions (Lim et al., 2014), the median cingulate gyri and supplementary motor area (Yang et al., 2023), the cerebellum and striatum (Gheorghe et al., 2021), as well as the precuneus (Tozzi et al., 2020).

This heterogeneity could result from the diversity of measurement instruments (e.g., different retrospective self-report scales vs. prospective ratings) and operationalizations of CM (e.g., continuous vs. categorical), as well as different subtypes of maltreatment being studied separately (e.g., Ringwald et al., 2022; Sheffield et al., 2013; Van Harmelen et al., 2010) Another potential source of heterogeneity could stem from varying sample characteristics, differing in diagnoses, the degree of psychopathology and the severity of CM exposure (e.g., McLaughlin et al., 2014). Furthermore, different statistical approaches have been used. One statistical challenge is a strong phenomenological co-occurrence with mental health problems. Often, psychiatric diagnosis is statistically controlled for which leads to reduced power to detect maltreatment effects because both constructs strongly covary (Lippard & Nemeroff, 2020) and both explain shared variance in neurobiological alterations (Opel et al., 2014). On the other side, if not controlling for diagnosis, neurobiological effects due to maltreatment or due to depression are impossible to disentangle.

The recent debate around questionable replicability in the neuroimaging domain due to underpowered samples and publication bias suggests the possibility of substantial false-positive findings within the previous body of evidence (Marek et al., 2022; Poldrack et al., 2017). This notion is supported by evidence for considerable publication bias in meta-analyzed findings of gray matter correlates of CM (Paquola et al., 2016). In fact, large-scale neuroimaging consortia, such as the ENIGMA consortium (Frodl et al., 2017; n=3036) or the UK-Biobank (Gheorghe et al., 2021; n=6751), have yielded much smaller effect sizes compared to studies with smaller samples, and have failed to replicate frequently reported associations of CM with the hippocampus or amygdala. However, these consortia still rely exclusively on segmented volumetric brain measures, thus losing spatial resolution, which may account for lower sensitivity to find gray matter alterations, posing a limitation to these findings.

In summary, inconclusive previous findings may result from variability in CM operationalizations, investigated clinical and non-clinical subgroups, varying statistical approaches, insufficient spatial resolution or simply because of false-positive results originating from underpowered studies. In order to shed light on this heterogeneity and re-evaluate our knowledge about the neurobiological underpinnings of adverse childhood experiences, we investigated the cross-cohort replicability of gray matter correlates of CM. We therefore utilized three large-scale, deeply phenotyped clinical cohort data sets, with a broad range of self-reported maltreatment experiences, in combination with high-resolution voxel-based morphometry (VBM). These rich datasets were assessed and processed in a standardized pipelines harmonized across cohorts. We conducted subgroup analyses and probed different operationalizations and subtypes of maltreatment. In order to investigate replicability, we assessed the spatial overlap of significant findings between our three independent well-powered cohorts, in addition to analyzing all cohorts together in a pooled model. Across all models we tested the hypothesis that CM is associated with lower gray matter volume (GMV).

Methods

Participants

Samples from three large-scale independent cohorts were included in the present analysis: the Marburg Münster Affective Disorders Cohort Study (MACS), the Münster Neuroimaging cohort (MNC) and the BiDirect cohort. All three cohorts include adults (age 18-65 years) with and without mental disorders. For the current analyses we included healthy control (HC) individuals, as well as individuals with a lifetime MDD diagnosis, with available information on CM, as well as T1-weighted structural brain images. In total, a sum of n=3225 participants were included (HC: n=1898; MDD: 1327). For details on the methods and inclusion criteria of the study samples we refer to previous publications (MACS: Kircher et al., 2019; Vogelbacher et al., 2018; MNC: Dannlowski et al., 2016; Opel, Redlich, Dohm, et al., 2019; BiDirect: Teismann et al., 2014) and to the supplements. The final samples included in the current analyses comprised n=1752 participants from MACS (HC: n=930; MDD: n=822), n=916 participants from MNC (HC: n=647; MDD: n=269), and n=557 participants from BiDirect (HC: n=321; MDD: n=236). Detailed sample characteristics of the three cohorts are described in Table S1 and Table S2. Differences between cohorts in sample characteristics are shown in Table S3, while differences in clinical characteristics between the MDD subsamples of the cohorts are shown in Table S4.

Of note, findings regarding gray matter correlates of CM have been previously published using MNC data at earlier stages of data assessments. However, these analyses only included a fraction of the data available for the current analysis (the largest sample including n=170 subjects; Dannlowski et al., 2012; Opel et al., 2014).

Assessment of childhood maltreatment and clinical characterization

CM was assessed using the German version of the Childhood Trauma Questionnaire (CTQ) (Bernstein et al., 1994; Wingenfeld et al., 2010). The CTQ is a 25-item retrospective self-report questionnaire capturing five different subtypes of CM, namely emotional abuse, physical abuse, sexual abuse, emotional neglect, and physical neglect. Each of these subtypes can be scored separately using a sum score of the five corresponding items, while the sum of these subscale scores amounts to the total CTQ score (expressing the total severity or load of experienced maltreatment). In addition, a categorical scoring of the CTQ has been introduced based on validated subscale cutoff values,

dividing scores in each subscale into different severity categories from 'none to minimal' to 'severe' (Bernstein et al., 1998). As described in detail below, here we utilize the total and subscale sum scores, as well as categorical cutoffs for group comparisons. The CTQ has been used in several hundreds of studies across different nationalities and validated on clinical and non-clinical populations (Viola et al., 2016). It has been extensively tested for its psychometric properties in several languages and geographic contexts (Dovran et al., 2013; Gerdner & Allgulander, 2009; Karos et al., 2014; Kim et al., 2013; Kluwe-Schiavon et al., 2016; Paivio & Cramer, 2004; Spinhoven et al., 2014; Viola et al., 2016; Wingenfeld et al., 2010).

Lifetime clinical diagnosis was assessed using structured clinical interviews by trained study personnel in each cohort. Within MACS and MNC the German version of the Structured Clinical Interview for the DSM-IV (SCID; Wittchen et al., 1997) was used. Within the BiDirect cohort the Mini International Neuropsychiatric Interview (MINI; Decubrier et al., 1997) was used, also based on the DSM-IV criteria. Further clinical characterization was done using a variety of standardized clinical interviews, rating scales and self-report questionnaires, capturing information on current remission status, current depression severity and previous course of disease (see supplements).

Structural image acquisition and processing

T1-weighted high-resolution anatomical brain images were acquired using 3T MRI scanner using highly harmonized scanning protocols across all three cohorts. 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). MNC and BiDirect samples were both scanned using a Gyroscan Intera scanner with Achieva update (both by 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 equally for all cohorts. 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 0.1 was used for all second level analyses as recommended for VBM analyses (http://www.neuro.uni-jena.de/cat/). Image quality was assessed by visual inspection as well as by using the check for homogeneity function implemented in the CAT12 toolbox. Image acquisition and processing for all study samples were extensively described elsewhere (Opel, Redlich, Dohm, et al., 2019; Teismann et al., 2014; Vogelbacher et al., 2018).

Image harmonization was conducted using the neuroCombat toolbox in python (Fortin et al., 2018) with default parameters to control for differences in scanner hardware and corresponding effects on brain images. This procedure allows the specification of 'biological covariates' that are excluded from harmonization in order to preserve desired variance of potentially confounding variables. We defined age, sex, total intracranial volume (TIV) and MDD diagnosis as such covariates. The harmonization process was conducted across six different scanner groups: MACS Münster scanner, MACS Marburg scanner before and after body coil change, MNC scanner before and after gradient coil change and the BiDirect scanner setting.

Statistical analysis

Brain-wide associations between CM and voxel-wise GMV were tested using general linear models in a mass-univariate VBM approach. The available cohorts were investigated in two different steps: In a first step we pooled all cohorts together in order to harvest the maximum sample size and thus the maximum available statistical power. For this pooled analysis we used a voxel-wise family-wise error (FWE)-corrected significance threshold of $p_{FWE}<.05$.

In a second step we investigated the cross-cohort replicability by analyzing each of the cohorts separately using two liberal uncorrected significance thresholds of $p_{unc}<.001$ and $p_{unc}<.01$. Here, we examined the spatial convergence (i.e., overlap) of significant voxels across the cohorts as conjunctive criteria (convergence either across any subset of two cohorts or across all three cohorts). Note that these liberal significance thresholds should not be used by themselves for statistical inference due to a massively inflated alpha error from the mass-univariate testing. However, we defined these liberal thresholds as minimum thresholds for effects to be recognized as replicable. Importantly, the probability of finding the same voxel in two or even three cohorts, constitutes a higher threshold than a voxel becoming significant just in a single cohort. In numbers, a threshold of $p<.001$ exceeded in each of the three single cohorts, results in an effective false-positive rate of $p<.001^3 = .000000001$. No extent threshold for minimum cluster size was used in any analysis.

The following statistical models were probed for 1) pooled analyses and 2) cross-cohort replicability analyses, to delineate the conditions under which VBM associations may become evident (an overview of all statistical models is presented in Table 1). In all models age, sex and TIV were included as covariates. Additionally, lifetime MDD diagnosis was included as a control variable in all models unless stated otherwise:

- In **Model 1** we used CTQ sum as the main predictor of interest, while additionally controlling for lifetime MDD diagnosis, as this variable is highly confounded with CM. Thus, in this model we tested the effect of CM on GMV beyond any effect of diagnosis.
- The model described above may not be sufficiently sensitive to detect CM effects due to substantial shared variance with the MDD diagnosis effect being partialized out. Therefore, we further tested a second model (**Model 2**) removing MDD diagnosis as a covariate. This allows to obtain a liberal estimate for the association between CM and gray matter, which however is not clearly separable from any MDD diagnosis effects.
- To investigate CM effects independently from a confounding MDD diagnosis effect, we further conducted subgroup analyses within HC (**Model 3**) and MDD (**Model 4**) samples separately.
- In order to delineate effects of specific subtypes of CM, we tested a series of models with each of the five CTQ subscale sum scores as predictors respectively (**Models 5-9**).
- Lastly, we investigated effects of severe forms of CM. This is done by identifying participants exceeding the subscale cutoff score for severe CM in any CTQ subscale, as defined and validated by Bernstein et al. (1998). This group of individuals with severe CM is contrasted with a control group that does not exceed any subscale cutoff (only CM reported within the range of "none to minimal"). With this model (**Model 10**) we accounted for the notion that CM associations with VBM may become evident particularly in individuals with severe experiences of CM. In our sample n=591 individuals (HC: n=109; MDD: n=482) fulfilled the criteria for severe maltreatment in at least one of the CTQ subscales, while n=1226 individuals (HC: n=989; MDD: n=237) fell into the category "none to minimal" maltreatment experiences. The distribution of severe CM across diagnostic groups was significantly uneven ($Chi^2=645.71$; $p<.001$, OR=18.45, 95%-CI: [14.348, 23.732). This strongly unequal distribution

led us to conduct severity group analyses additionally in HC (**Model 11**) and MDD (**Model 12**) subgroups separately.

Table 1

Summary of all conducted statistical models and the respective included sample sizes

Model	CM operationalization	Controlled for MDD diagnosis	Subsamples	Sample size - n			
				MACS	MNC	BiDirect	Pooled
Model 1	CTQ sum	yes	HC/MDD	1752	916	557	3225
Model 2	CTQ sum	no	HC/MDD	1752	916	557	3225
Model 3	CTQ sum	-	HC	930	647	321	1898
Model 4	CTQ sum	-	MDD	822	269	236	1327
Model 5	EA subscale sum	yes	HC/MDD	1752	916	557	3225
Model 6	PA subscale sum	yes	HC/MDD	1752	916	557	3225
Model 7	SA subscale sum	yes	HC/MDD	1752	916	557	3225
Model 8	EN subscale sum	yes	HC/MDD	1752	916	557	3225
Model 9	PN subscale sum	yes	HC/MDD	1752	916	557	3225
Model 10	Extreme groups (none/severe)	yes	HC/MDD	None: 644 Severe: 348	None: 369 Severe: 145	None: 213 Severe: 98	None: 1226 Severe: 591
Model 11	Extreme groups (none/severe)	-	HC	None: 500 Severe: 51	None: 324 Severe: 41	None: 165 Severe: 17	None: 989 Severe: 109
Model 12	Extreme groups (none/severe)	-	MDD	None: 144 Severe: 297	None: 45 Severe: 104	None: 48 Severe: 81	None: 237 Severe: 482

Note. In all models we additionally controlled for age, sex and total intracranial volume. The extreme group comparisons are based on cutoff-based categorizations of severity. For Models with comparison of extreme groups (Models 10-12) sample sizes are given for both the group with 'none to minimal' maltreatment (labelled 'none') and the group with 'severe' maltreatment. CM, childhood maltreatment; CTQ, Childhood Trauma Questionnaire; HC, healthy controls; MDD, major depressive disorder; MACS, Marburg Münster Affective Disorders Cohort Study; MNC, Münster Neuroimaging Cohort; BiDirect, BiDirect study cohort; EA, emotional abuse; PA, physical abuse; SA, sexual abuse; PN, physical neglect; EN, emotional neglect.

One-sided negative contrasts were tested in all models (i.e., CM associated with lower gray matter volume) due to poor evidence for potential positive associations between CM and VBM (Lim et al., 2014; McLaughlin et al., 2019; Paquola et al., 2016; Yang et al., 2023). Partial R^2 was calculated based on t-maps and reported as an effect size for the partialized percentage of explained variance of the respective CM predictor for each analysis. Analyses were conducted using python (version 3.9.12). All analysis scripts and non-thresholded results are available in an Open Science Framework (OSF) repository (https://osf.io/j8d9r/?view_only=9edf436ab18f4e8db9ef4c71c4ac356c), to foster transparency and reproducibility of our analyses (Poldrack et al., 2017).

Results

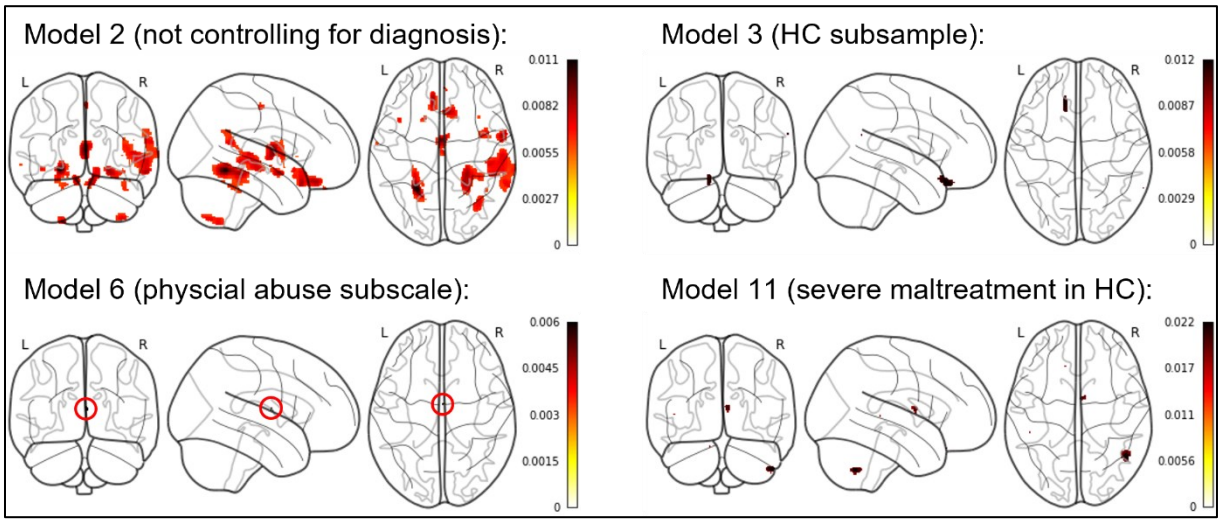
Pooled sample across all cohorts

All results using the full sample from pooling all cohorts together are presented at a conservative significance of $p_{FWE}<.05$, corrected at the voxel-level. Findings from pooled analyses are shown in Table 2.

When controlling for MDD diagnosis (Model 1) no voxels with a significant CM association were found. Dropping MDD diagnosis as covariate (Model 2) yielded significant widespread clusters with an association between CM and gray matter (total k=4970), located mainly within superior and middle temporal areas, a bilateral fusiform and lingual complex, the thalamus, as well as in the orbitofrontal cortex and the insula. Subgroup analyses revealed small significant clusters in HC individuals (Model 3; total k=99) within the medial orbitofrontal cortex, while no clusters survived the FWE-correction within the MDD sample (Model 4). Regarding subtypes of CM, no CTQ subscales were associated with GMV surpassing an FWE-corrected threshold, except a small cluster emerging when using physical abuse as a predictor (Model 6; total k=6 within the thalamus). Within the HC subsample, participants with 'severe' maltreatment showed lower gray matter volume as compared to participants with 'none to minimal' maltreatment in two clusters (Model 11; total k=132), located in the cerebellum and the thalamus. Effect sizes across models when pooling cohorts ranged between partial R^2=.006 and partial R^2=.022.

Figure 1

Significant clusters from pooled analysis at $p_{FWE}<.05$



Note. Glass brains are shown with maximum intensity projections. The cluster in Model 6 is marked with a red circle for visualization purposes. Color bars represent the partial R^2 of the predictor in the respective model. All models are shown yielding significant clusters at $p_{FWE}<.05$. HC, healthy controls.

Table 2

Results summary for pooled cohorts (n=3225) at a significance level of p_{FWE}<.05

Model	k significant	partial R^2 min	partial R^2 max	main regions
Model 1	0	-	-	
Model 2	4970	0.006	0.011	Temporal Mid/Sup R, Fusiform L+R, Lingual L, Parahippocampal R, Thalamus L
Model 3	99	0.010	0.012	OFC med L, Rectus L
Model 4	0	-	-	
Model 5	0	-	-	
Model 6	6	0.006	0.006	Thalamus R[a]
Model 7	0	-	-	
Model 8	0	-	-	
Model 9	0	-	-	
Model 10	0	-	-	
Model 11	132	0.018	0.022	Cerebellum R, Thalamus R[a]
Model 12	0	-	-	

Note. Number of significant voxels at p_{FWE}<.05 are shown, as well as their minimum and maximum effect size for each analysis. Cluster labelling was conducted based on the aal atlas using the atlasreader python package (Notter et al., 2019). Main regions are reported. [a]Cluster labelling using aal resulted in 'no_label' however checking Desikan-Killiany and Harvard-Oxford atlases indicated localization within the thalamus. L, left; R, right; Mid, middle; Sup, superior; OFC, orbitofrontal cortex; med, medial.

Replicability across single cohorts

At both liberal uncorrected thresholds of p_{unc}<.001 and p_{unc}<.01 each statistical model yielded significant voxels in at least one of the three cohorts. In turn, each of the three cohorts produced significant voxels in most of the statistical models. The highest number of significant voxels was observed in model 2, when MDD diagnosis was dropped as a covariate. A detailed summary of cohort-wise results across models is shown in supplementary Table S5-S6. Across probed models and across single cohorts the significant voxels were widespread throughout the brain, including the cerebellum, temporal and frontal areas, subcortical areas and somatosensory cortices.

Regarding replicability there was not one voxel that was congruently significant (i.e., replicable) at a threshold of p_{unc}<.001 in all three cohorts. This finding was consistent across all probed statistical models, including HC and MDD subgroup analyses, testing subtypes of CM and comparing groups with severe CM and no CM. Similarly, comparing pairs of cohorts also yielded no voxels that regionally overlapped between any pairwise cohort combinations, for most of the tested models. The only pairwise convergence in significance at this threshold was found between the MNC and BiDirect cohorts for the physical neglect subscale of the CTQ (Model 9), which was associated with a small overlapping cluster located within the supramarginal gyrus (overlap k=3; Dice=.0019).

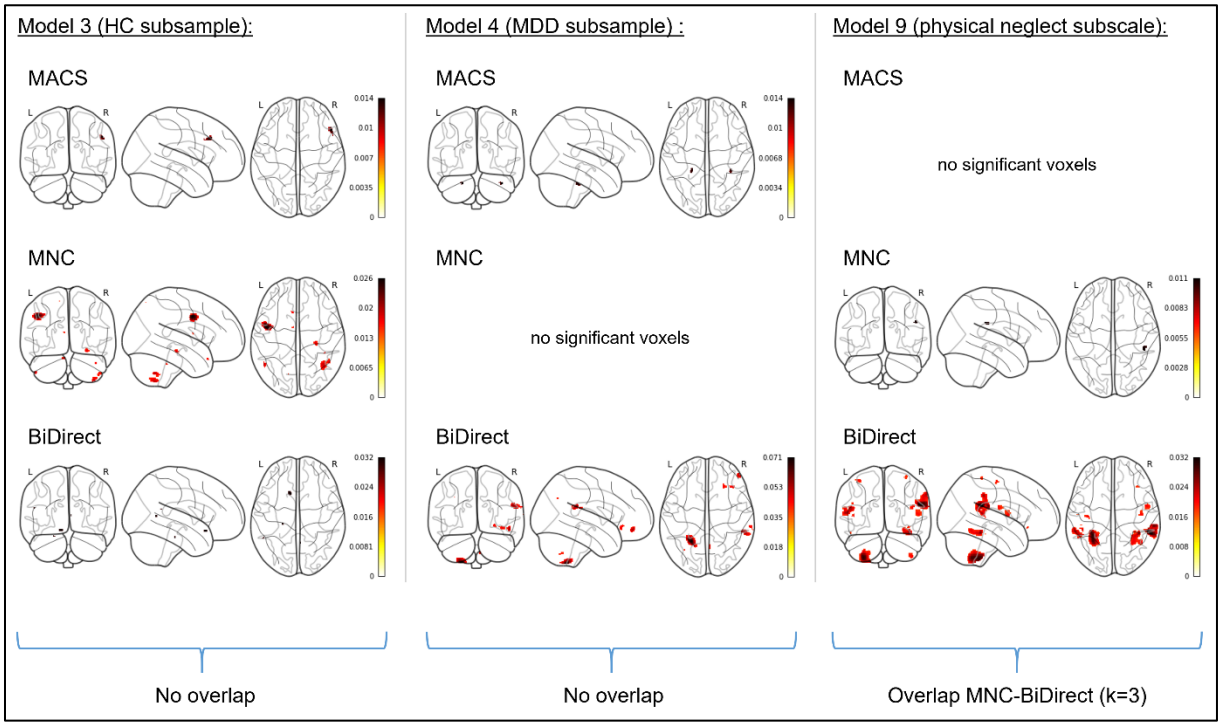
When rerunning the replicability analyses using an even more liberal threshold of p_{unc}<.01 spatial overlap in significant voxels was slightly increased across models. The only model yielding replicability across all three cohorts at this threshold was model 2 (not controlling for MDD diagnosis), with converging significance in k=4 voxels. Pairwise cohort combinations yielded further

spatial congruence in significance across models, with maximum congruence in model 9 (physical neglect, k=609, Dice=0.068).

A summary of the replicability across the three cohorts, in the sense of the extent of spatial convergence of significant effects is shown in Table 3 and supplementary Table S7. Significant clusters across uncorrected thresholds, across single cohorts and across statistical models are additionally shown in supplementary Figures S1-S12.

Figure 2

Significant clusters across cohort-wise analyses at p_{unc}<.001 for exemplary models (replicability analysis)



Note. Glass brains are shown with maximum intensity projections. Color bars represent the partial R^2 of the maltreatment predictor in the respective model. Exemplary model results are shown for two models without any spatial overlap (i.e., replicability), namely the CTQ sum analysis in HC participants (Model 3) and in MDD participants (Model 4), as well as the only model showing some degree of spatial overlap (Model 9 testing the physical neglect subscale as predictor). Plots without color bar yielded no significant clusters. HC, healthy controls; MDD, Major Depressive Disorder; MACS, Marburg Münster Affective Disorders Cohort Study; MNC, Münster Neuroimaging Cohort; BiDirect, BiDirect cohort.

Table 3

Replicability across cohorts indicated by spatial overlap in significance at a level of p_{unc}<.001

	MACS-MNC		MACS-BiDirect		MNC-BiDirect		MACS-MNC-BiDirect	
	k overlap	DICE	k overlap	DICE	k overlap	DICE	k overlap	DICE
Model 1	0	0	0	0	0	0	0	-
Model 2	0	0	0	0	0	0	0	-
Model 3	0	0	0	0	0	0	0	-
Model 4	0	0	0	0	0	0	0	-
Model 5	0	0	0	0	0	0	0	-
Model 6	0	0	0	0	0	0	0	-
Model 7	0	0	0	0	0	0	0	-
Model 8	0	0	0	0	0	0	0	-
Model 9	0	0	0	0	3	0.002	0	-
Model 10	0	0	0	0	0	0	0	-
Model 11	0	0	0	0	0	0	0	-
Model 12	0	0	0	0	0	0	0	-

Note. The overlap in significant voxels is presented across all statistical models and all cohort combinations. The DICE score is presented for pairwise combinations with any voxels overlapping. MACS, Marburg Münster Affective Disorders Cohort Study; MNC, Münster Neuroimaging Cohort; BiDirect, BiDirect study cohort.

Discussion

Findings implicating long-term effects of CM on the structural morphology of the brain have been frequently published over the last decades and are central to a neurobiological model of how environmental risk is conveyed to psychopathology. However, in an unprecedented replication effort we present evidence that localized brain-wide associations between gray matter structure and various operationalizations of CM are essentially non-replicable. This lack of replicability was consistently shown for a wide variety of common statistical approaches, subgroups and operationalizations of CM. In the following we discuss implications for the general underlying hypothesis, as well as for methodological considerations to foster replicability of brain-wide association studies within the field of mental health neuroimaging.

We utilized three large-scale cohorts for an investigation of replicability as well as for pooling samples to harvest the maximum available statistical power. While we found small significant effects within HC subgroups in our pooled analysis, these were not replicable across single cohorts. Furthermore, in the pooled analysis we identified large and widespread clusters associated with CM when dropping MDD diagnosis as a control variable. As described before, this effect is not distinguishable from an MDD effect associated with GMV due to substantial co-occurrence between CM and MDD diagnosis. Notably, the localized clusters seemingly associated with CM in this analysis when not controlling for MDD largely overlap with clusters that were identified to be associated with a lifetime MDD diagnosis in a systematic case-control study using the same cohorts (Dannlowski et al., 2024).

When using liberal uncorrected thresholds, we found a vast array of regions seemingly associated with CM across the different cohorts and statistical models. In isolation, each of these results could easily have been the basis of a publication just like several smaller existing studies in the field, including previous publications from our own group (Dannlowski et al., 2012; Opel et al., 2014).

Importantly, these widespread significant effects for each single cohort should not be interpreted as solid evidence for an association between CM and gray matter. This is vital as these findings resulted from liberal significance thresholds not corrected for multiple tests, thus resulting in a massively inflated alpha error (i.e., higher chance of false-positives) and should therefore not be interpreted alone. Rather, these liberal thresholds were specifically used to investigate if statistically lenient results replicate across cohorts. In our most liberal analyses, some extent of replicability was observed, implying mostly the physical neglect and abuse subtypes, as well as subgroups reporting severe experiences of CM. However, the identified overlap was small even when only investigating pairwise replicability. Overall, our findings suggest that gray matter reductions associated with CM are effectively non-replicable, supporting the notion of a null-finding. Even highly cited previous reports of brain structural correlates of CM, such as associations with lower GMV within the hippocampus (Dannlowski et al., 2012; Opel et al., 2014) could not be confirmed.

Null findings are always difficult to interpret due to a multitude of potential reasons for failing to detect an effect. Potential reasons for false-negatives can stem from the specific measurement and operationalization of the predictor or dependent variable, the specific statistical approach (e.g., inclusion of covariates), insufficient statistical power, the sample selection and, regarding replicability, differences between specific cohort characteristics. In the following we will discuss each of these potential sources of effect variability.

The CTQ we used in our study is one of the most common instruments to assess CM. However, despite its common use, the CTQ has been criticized for neglecting important timing information of maltreatment (Teicher & Parigger, 2015) and for showing low agreement with prospective measures of CM (Baldwin et al., 2019), although the latter may stem from a methodological confounding as prospective and retrospective measures usually differ in their source of information (Goltermann et al., 2019). While the general critique of the CTQ may be valid, it should be mentioned that it is also the instrument which has been used in most of the referenced studies, which have reported gray matter associations with CM (Dannlowski et al., 2012; Gold et al., 2016; Opel et al., 2014; Paquola et al., 2017; Samplin et al., 2013; Whittle et al., 2017). Thus, our methodology is at least comparable to previous studies reporting significant associations.

On the side of the dependent variable, we used state-of-the-art voxel-wise assessment of GMV, which also is highly comparable to similar previous studies. Importantly, it should be noted that previous meta research on replicability of neuroimaging findings has suggested that the myriad of researchers' choices (i.e., *researchers' degrees of freedom*) regarding scanning parameters, preprocessing and quality control pipelines of neuroimaging data is likely to majorly contribute to problems of reproducibility and replicability (Botvinik-Nezer et al., 2020; Zhou et al., 2022). While differences in these pipelines could account for differences in findings it should be noted that we used highly standardized procedures which are closely harmonized across our three included cohorts, thus not posing a sufficient explanation for low replicability between our specific cohorts.

Similarly, several choices have to be made regarding the specific operationalization of predictors and the concrete statistical modelling approach. In previous studies, CM (even when based on the CTQ) was operationalized in different ways (total sum score, subscale scores, different cutoff-based categories), and confounding psychopathology was differently handled. To be able to allow for more compelling and conclusive interpretations we decided to use a comprehensive approach covering an array of the most common statistical models in our analyses. While this list of statistical models is certainly not exhaustive, and our conclusions are limited to these specific models, the main finding of poor replicability proved to be strikingly consistent across all probed models.

Insufficient statistical power could be a potential explanation for low replicability particularly for some of the subgroup analyses. While we were able to conduct the largest replicability investigation of brain alterations associated with CM to date, this could still be insufficient to detect some of the small effects that are common in the field of biological psychology and psychiatry. Recent work from our group (Winter et al., 2022) and others (Marek et al., 2022) have suggested that true effect sizes in this domain typically range up to a maximum of approximately 2.5% explained variance, thus requiring large samples to detect and replicate these effect. Our current results support this estimation of the scope of effect size in this field. Relative to such small effect sizes our analyses may have been statistically underpowered particularly to find effects in subgroups, like the HC subpopulations of each cohort separately (smallest sample size n=129 MDD individuals in BiDirect cohort in model investigating severe CM). However, in comparison to previous studies the sample size and statistical power reached in our current investigation is well within the realm of previous meta-analyses (Lim et al., 2014; Paquola et al., 2016) and large-scale consortia analyses (Schmaal et al., 2017; Tozzi et al., 2020).

The specific sample selection is another very important potential source of variability in findings. While our cohorts include a wide range of maltreatment exposure severity and clinical characteristics (i.e., previous course of disease and current depression severity), they are relatively homogeneous regarding demographic characteristics. All cohorts include German Caucasian individuals, with a relatively high level of education. While this homogeneity across cohorts poses an optimal condition for successful replicability it also limits the generalizability of our null-findings to other demographic circumstances. Despite the described similarities between our three cohorts there are some mentionable differences that could potentially account for low replicability. For example, the BiDirect cohort shows a considerably older age of participants, as compared to the other two cohorts. Similarly, the MNC cohort includes only acutely depressed inpatient MDD patients, while the other two cohorts also include individuals in outpatient care or even remitted patients. Further, differences in current depression severity and previous illness severity characteristics were found between the MDD samples across cohorts. These differences could potentially account for low replicability in some cases. However, due to the extent of non-replicability between any pairwise cohort combinations it seems unlikely that differences between cohort characteristics comprehensively account for our null findings.

In summary, all these aspects pose potential sources of effect variability and could account for false-negative findings. However, previous studies reporting CM associations were highly comparable regarding utilized methodology. It should be noted that it is possible that conventional conceptualizations of clearly localizable gray matter reductions due to CM on a group-level could be too simplistic. Recently, machine learning and normative modelling approaches have been increasingly promoted following the notion that the concrete shape of neurobiological consequences in the brain may be highly individual (Leenings et al., 2022; Marquand et al., 2016).

The absence of evidence cannot directly be interpreted as evidence for the absence of a phenomenon (Marsh, 2019). However, the extent of non-replicability of gray matter correlates of CM still appears disconcerting, particularly due to the vast number of previous publications on this matter. Notably, the observed non-replicability of CM-GMV associations is contrasted by substantial evidence for replicability of GMV reductions associated with a lifetime MDD diagnosis which was found in a highly comparable approach. In a case-control design we found that gray matter reductions in MDD were replicable across the same cohorts (MACS, MNC and BiDirect) in a consistent manner, to some degree even at highly conservative significance thresholds (Dannlowski et al., 2024).

Replicability of research findings is a fundamental principle of the scientific process and is essential for accumulating scientific knowledge (Popper, 1959). However, non-replicability of published findings is a growing concern across disciplines, such as cell biology (Baggerly & Coombes, 2009), genetics (Border et al., 2019; Siontis et al., 2010), oncology (Begley & Ellis, 2012), epidemiology (Lash et al., 2018) and psychology (Open Science Collaboration, 2015). There is broad consensus among researchers that contributing factors for this *replication crisis* are the prevailing pressure to publish and a publication bias towards positive and novel findings, as well as researchers' consequential exploitation of analytic flexibility and selective reporting to enhance chances for publication (Baker, 2016; Nosek et al., 2022). In the field of neuroimaging research these factors could be particularly potent due to a combination of particularly high analytic flexibility, a large number of statistical tests and small underpowered samples in this domain, which leads to an overestimation of effect sizes and to non-replicability (Button et al., 2013; Vul et al., 2009). Recently, a seminal study published by Marek et al. demonstrated this for neuroimaging, showing that samples including thousands of individuals are necessary to obtain robust and replicable brain-wide associations due to small true effect sizes (Marek et al., 2022). These findings have reignited a vibrant controversy about the challenges of replicability in this research domain (e.g., Bandettini et al., 2022; Genon et al., 2022; Nour et al., 2022; Rosenberg & Finn, 2022; Spisak et al., 2023; Tervo-Clemmens et al., 2023). The findings from our current study support this notion and it is conceivable that low replicability may not be a specific issue only concerning the research question at hand. Notably, to date, no consensus exists how to define 'successful' replicability of voxel-wise neuroimaging findings.

Concerns about low replicability have led to the development of so-called *open science* policies, such as the preregistration of hypotheses and analysis plans, as well as comprehensive disclosure of analysis code and results (Nosek et al., 2015), to achieve transparency and reduce biases. Systematic meta-investigations of replicability in other areas of psychology have shown that these practices can be very effective to enhance replicability (Protzko et al., 2020). Accordingly, the use of open science practices has been increasingly demanded by scholars to facilitate replicability in neuroimaging research (Klapwijk et al., 2021). However, to date replications, preregistrations and other open science practices are extremely rare in the neuroimaging literature (Paret et al., 2022). In addition, analysis approaches focusing on the generalizability of statistical findings to independent data, such as cross-validation have been suggested to effectively reveal overestimated effect sizes and non-replicability in smaller samples (Goltermann et al., 2023) and their potential for the domain of neuroimaging should be more closely explored.

Our findings underline the importance of taking a step back and shifting the focus towards investigating the replicability and generalizability of presumably established research findings. Various open science practices are available to increase and systematically assess replicability of research findings. These include 1) preregistrations of hypotheses and analyses, 2) transparent sharing of analysis code, methods and publication of comprehensive (i.e., non-thresholded) results (Taylor et al., 2023), as well as 3) the mere execution of direct and conceptual replication studies and 4) the publication of null findings. Such open science practices should be routinely applied in the field of neuroimaging research of mental disorders to increase replicability and thus maximize the potential for clinical translation of findings.

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

Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, *3*(4), 1309–1334. https://doi.org/10.1214/09-AOAS291

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://doi.org/10.1038/d41586-019-00067-3

Baldwin, J. R., Reuben, A., Newbury, J. B., & Danese, A. (2019). Agreement Between Prospective and Retrospective Measures of Childhood Maltreatment: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *76*(6), 584–593. https://doi.org/10.1001/jamapsychiatry.2019.0097

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.

Begley, C. G., & Ellis, L. M. (2012). Raising standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1111/ebm2.3

Bernstein, D. P., & Fink, L. (1998). *Childhood Trauma Questionnaire. A Retrospective Self-Report Questionnaire and Manual*. The Psychological Corporation.

Bernstein, D. P., Fink, L., Handelsman, L., & Foote, J. (1998). Childhood Trauma Questionnaire. *Assessment of Family Violence: A Handbook for Researchers and Practitioners.*

Bernstein, D. P., Fink, L., Handelsman, L., Foote, J., Lovejoy, M., Wenzel, K., Sapareto, E., & Ruggiero, J. (1994). Initial Reliability and Validity of a New Retrospective Measure of Child Abuse and Neglect. *The American Journal of Psychiatry*, *151*, 1132–1136. https://doi.org/10.1176/ajp.151.8.1132

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. (2019). No Support for Historical Candidate Gene or CandidateGene-by-Interaction Hypotheses for Major DepressionAcross Multiple Large Samples. *American Journal of Psychiatry*, *176*(5), 329–414.

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., … 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., … 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

Dannlowski, U., Stuhrmann, A., Beutelmann, V., Zwanzger, P., Lenzen, T., Grotegerd, D., Domschke, K., Hohoff, C., Ohrmann, P., Bauer, J., Lindner, C., Postert, C., Konrad, C., Arolt, V., Heindel, W., Suslow, T., & Kugel, H. (2012). Limbic Scars: Long-Term Consequences of Childhood Revealed by

Functional and Structural Magnetic Resonance Imaging. *Biological Psychiatry*, *71*(4), 286–293. https://doi.org/10.1016/j.biopsych.2011.10.021

Dannlowski, U., Winter, N. R., Meinert, S., Grotegerd, D., Kraus, A., Flinkenflügel, K., Leehr, E., Böhnlein, J., Borgers, T., Fisch, L., Bauer, M., Pfennig, A., Richter, M., Opel, N., Repple, J., Gruber, M., Minnerup, H., Hermesdorf, M., Nitsch, R., … Goltermann, J. (2024). Replicability and generalizability of gray matter reductions in major depression: a voxel-based investigation of 4021 individuals. *SSRN (Preprint)*. https://doi.org/http://dx.doi.org/10.2139/ssrn.4854882

Dovran, A., Winje, D., Øverland, S. N., Breivik, K., Arefjord, K., Dalsbø, A. S., Jentoft, M. B., Hansen, A. L., & Waage, L. (2013). Psychometric properties of the Norwegian version of the Childhood Trauma Questionnaire in high-risk groups. *Scandinavian Journal of Psychology*, *54*, 286–291. https://doi.org/10.1111/sjop.12052

Fortin, J., Cullen, N., Sheline, Y. I., Taylor, W. D., Cook, P. A., Adams, P., Cooper, C., Fava, M., Patrick, J., Mcinnis, M., Phillips, M. L., Trivedi, M. H., Myrna, M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024.Harmonization

Frodl, T., Janowitz, D., Schmaal, L., Tozzi, L., Dobrowolny, H., Stein, D. J., Veltman, D. J., Wittfeld, K., Erp, T. G. M. Van, Jahanshad, N., Block, A., Hegenscheid, K., Lagopoulos, J., Hatton, S. N., Hickie, I. B., Maria, E., Carballedo, A., Brooks, S. J., Vuletic, D., … Grabe, H. J. (2017). Childhood adversity impacts on brain subcortical structures relevant to depression. *Journal of Psychiatric Research*, *86*, 58–65. https://doi.org/10.1016/j.jpsychires.2016.11.010

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

Gerdner, A., & Allgulander, C. (2009). Psychometric properties of the Swedish version of the Childhood Trauma Questionnaire—Short Form (CTQ-SF). *Nordic Journal of Psychiatry*, *63*(2), 160–170. https://doi.org/10.1080/08039480802514366

Gheorghe, D. A., Li, C., Gallacher, J., & Bauermeister, S. (2021). Associations of perceived adverse lifetime experiences with brain structure in UK Biobank participants. *Journal of Child Psychology and Psychiatry*, *62*(7), 822–830. https://doi.org/10.1111/jcpp.13298

Gold, A. L., Sheridan, M. A., Peverill, M., Busso, D. S., Lambert, H. K., Alves, S., Pine, D. S., & McLaughlin, K. A. (2016). Childhood abuse and reduced cortical thickness in brain regions involved in emotional processing. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *57*(10), 1154–1164. https://doi.org/10.1111/jcpp.12630

Goltermann, J., Opel, N., & Dannlowski, U. (2019). Considering the Source of Information in the Evaluation of Maltreatment Experiences [Letter to the editor]. *JAMA Psychiatry*, *76*(9), 984–985. https://doi.org/10.1001/jamapsychiatry.2019.1688

Goltermann, J., Winter, N. R., Gruber, M., Fisch, L., Richter, M., Grotegerd, D., Dohm, K., Meinert, S., Leehr, E. J., Böhnlein, J., Kraus, A., Thiel, K., Winter, A., Flinkenflügel, K., Leenings, R., Barkhau, C., Ernsting, J., Berger, K., Minnerup, H., … Dannlowski, U. (2023). Cross-validation for the estimation of effect size generalizability in mass-univariate brain-wide association studies. *BioRxiv*. https://doi.org/10.1101/2023.03.29.534696

Goltermann, J., Winter, N. R., Meinert, S. L., Sindermann, L., Lemke, H., Leehr, E. J., Grotegerd, D., Winter, A., Thiel, K., Waltemate, L., Breuer, F., Repple, J., Gruber, M., Richter, M., Teckentrup,

V., Kroemer, N. B., Brosch, K., Meller, T., Pfarr, J.-K. J.-K., … Hahn, T. (2022). Resting-state functional connectivity patterns associated with childhood maltreatment in a large bicentric cohort of adults with and without major depression. *Psychological Medicine*, *53*(10), 4720–4731. https://doi.org/10.1017/S0033291722001623

Karos, K., Niederstrasser, N., Abidi, L., Bernstein, D. P., & Bader, K. (2014). Factor Structure, Reliability, and Known Groups Validity of the German Version of the Childhood Trauma Questionnaire (Short-Form) in Swiss Patients and Non-Patients. *Journal of Child Sexual Abuse*, *23*(4), 418–430. https://doi.org/10.1080/10538712.2014.896840

Kim, D., Bae, H., Han, C., Young, H., & MacDonald, K. (2013). Psychometric properties of the Childhood Trauma Questionnaire-Short Form (CTQ-SF) in Korean patients with schizophrenia. *Schizophrenia Research*, *144*, 93–98. https://doi.org/10.1016/j.schres.2012.12.020

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. (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

Kluwe-Schiavon, B., Wendt, V. T., & Grassi-Oliveira, R. (2016). Cross-cultural adaptation of the Maltreatment and Abuse Chronology of Exposure (MACE) scale to Brazilian Portuguese Adaptação transcultural da escala Maltreatment and Abuse Chronology of Exposure (MACE) para o português brasileiro. *Trends Psychiatry Psychother. Trends Psychiatry Psychother. Trends Psychiatry Psychother*, *383838*(341), 33–39. http://dx.doi.org/10.1590/2237-6089-2015-0051

Lash, T. L., Collin, L. J., & Van Dyke, M. E. (2018). The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports*, *5*(2), 175–183. https://doi.org/10.1007/s40471-018-0148-x

Leenings, R., Winter, N. R., Dannlowski, U., & Hahn, T. (2022). Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage*, *257*(April). https://doi.org/10.1016/j.neuroimage.2022.119298

Lim, L., Radua, J., & Rubia, K. (2014). Gray Matter Abnormalities in Childhood Maltreatment: A Voxel-Wise Meta-Analysis. *American Journal of Psychiatry*, *171*, 854–863.

Lippard, E. T. C., & Nemeroff, C. B. (2020). The Devastating Clinical Consequences of Child Abuse and Neglect: Increased Disease Vulnerability and Poor Treatment Response in Mood Disorders. *The American Journal of Psychiatry*, *177*(1), 20–36. https://doi.org/10.1176/appi.ajp.2019.19010020

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., … 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

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Marsh, O. (2019). Carl Sagan and the circulation fo reputation. *The British Journal for the History of Science*, *52*(3), 467–486.

McLaughlin, K. A., Sheridan, M. A., Winter, W., Fox, N. A., Zeanah, C. H., & Nelson, C. A. (2014). Widespread reductions in cortical thickness following severe early-life deprivation: A neurodevelopmental pathway to attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *76*(8), 629–638. https://doi.org/10.1016/j.biopsych.2013.08.016

McLaughlin, K. A., Weissman, D., & Bitrán, D. (2019). Childhood Adversity and Neural Development: A Systematic Review. *Annual Review of Delevopmental Psychology*, *1*, 277–312.

Meinert, S. L., Repple, J., Nenadic, I., Krug, A., Jansen, A., Grotegerd, D., Förster, K., Enneking, V., Dohm, K., Schmitt, S., Stein, F., Brosch, K., Meller, T., Redlich, R., Böhnlein, J., Sindermann, L., Goltermann, J., Leehr, E. J., Opel, N., … Dannlowski, U. (2019). Reduced fractional anisotropy in depressed patients due to childhood maltreatment rather than diagnosis. *Neuropsychopharmacology*, *44*(12), 2065–2072. https://doi.org/10.1038/s41386-019-0472-y

Nelson, J., Klumparendt, A., Doebler, P., & Ehring, T. (2017). Childhood maltreatment and characteristics of adult depression: meta-analysis. *The British Journal of Psychiatry*, *201*, 96–104. https://doi.org/10.1192/bjp.bp.115.180752

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture: The TOP guidelines. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. grave le B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Notter, M., Gale, D., Herholz, P., Markello, R., Notter-Bielser, M.-L., & Whitaker, K. (2019). AtlasReader: A Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images. *Journal of Open Source Software*, *4*(34), 1257. https://doi.org/10.21105/joss.01257

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., Grotegerd, D., Zaremba, D., Janik Goltermann Jonathan Repple, Kaehler, C., Meinert, S., Leehr, E. J., Böhnlein, J., Förster, K., Emden, D., Enneking, V., Sindermann, L., Dzvonyar, F., Heindel, W., Leenings;, R., Winter, N., Hahn;, T., … 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., 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., … 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., Zwanzger, P., Grotegerd, D., Arolt, V., Heindel, W., Konrad, C., Kugel, H., & Dannlowski, U. (2014). Hippocampal Atrophy in Major Depression: a Function of Childhood Maltreatment Rather than Diagnosis? *Neuropsychopharmacology*, *39*, 2723–2731. https://doi.org/10.1038/npp.2014.145

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*,

349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Paivio, S. C., & Cramer, K. M. (2004). Factor structure and reliability of the Childhood Trauma Questionnaire in a Canadian undergraduate student sample. *Child Abuse & Neglect*, *28*(8), 889–904. https://doi.org/10.1016/j.chiabu.2004.01.011

Paquola, C., Bennett, M. R., Hatton, S. N., Hermens, D. F., Groote, I., & Lagopoulos, J. (2017). Hippocampal development in youth with a history of childhood maltreatment. *Journal of Psychiatric Research*, *91*, 149–155. https://doi.org/10.1016/j.jpsychires.2017.03.019

Paquola, C., Bennett, M. R., & Lagopoulos, J. (2016). Understanding heterogeneity in grey matter research of adults with childhood maltreatment: A meta-analysis and review. *Neuroscience and Biobehavioral Reviews*, *69*, 299–312. https://doi.org/10.1016/j.neubiorev.2016.08.011

Paret, C., Unverhau, N., Feingold, F., Poldrack, R. A., Stirner, M., Schmahl, C., & Sicorello, M. (2022). Survey on Open Science Practices in Functional Neuroimaging. *NeuroImage*, *257*, 119306. https://doi.org/10.1016/j.neuroimage.2022.119306

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J. B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Pollok, T. M., Kaiser, A., Kraaijenvanger, E. J., Monninger, M., Brandeis, D., Banaschewski, T., Eickhoff, S. B., & Holz, N. E. (2022). Neurostructural traces of early life adversities: A meta-analysis exploring age- and adversity-specific effects. *Neuroscience & Biobehavioral Reviews*, *135*, 104589. https://doi.org/10.1016/J.NEUBIOREV.2022.104589

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Routledge.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovski, J. E., Roeder, S., Walleczek, J., & Schooler, J. W. (2020). High replicability of newly-discovered socialbehavioral findings is achievable. *PsyArXiv*. http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203

Ringwald, K. G., Pfarr, J.-K., Schmitt, S., Stein, F., Brosch, K., Meller, T., Andrae, J., Zech, R., Steinsträter, O., Meinert, S., Waltemate, L., Lemke, H., Thiel, K., Winter, A., Opel, N., Goltermann, J., Jansen, A., Dannlowski, U., Krug, A., … Kircher, T. (2022). Interaction of recent stressful life events and childhood abuse on orbitofrontal grey matter volume in adults with depression. *Journal of Affective Disorders*, *312*, 122–127. https://doi.org/10.1016/J.JAD.2022.06.050

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

Samplin, E., Ikuta, T., Malhotra, A. K., Szeszko, P. R., & DeRosse, P. (2013). Sex differences in resilience to childhood maltreatment: Effects of trauma history on hippocampal volume, general cognition and subclinical psychosis in healthy adults. *Journal of Psychiatric Research*, *47*(9), 1174–1179. https://doi.org/10.1016/j.jpsychires.2013.05.008

Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., Cheung, J. W., Erp, T. G. M. Van, Bos, D., Ikram, M. A., Vernooij, M. W., Niessen, W. J., Tiemeier, H., Hofman, A., Wittfeld, K., Grabe, H. J., Janowitz, D., Bülow, R., Selonke, M., … Veltman, D. J. (2017). Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Molecular Psychiatry*, *22*(6), 900–909. https://doi.org/10.1038/mp.2016.60

Sheffield, J. M., Williams, L. E., Woodward, N. D., & Heckers, S. (2013). Reduced gray matter volume in psychotic disorder patients with a history of childhood sexual abuse. *Schizophrenia Research*, *143*(1), 185–191. https://doi.org/10.1016/j.schres.2012.10.032

Siontis, K. C. M., Patsopoulos, N. A., & Ioannidis, J. P. A. (2010). Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics*, *18*(7), 832–837. https://doi.org/10.1038/ejhg.2010.26

Spinhoven, P., Penninx, B. W., Hickendorff, M., van Hemert, A. M., Bernstein, D. P., Elzinga, B. M., Hemert, A. M. Van, David, P., & Elzinga, B. M. (2014). Childhood Trauma Questionnaire: Factor structure, measurement invariance, and validity across emotional disorders. *Psychological Assessment*, *26*(3), 717–729.

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

Struck, N., Krug, A., Yuksel, D., Stein, F., Schmitt, S., Meller, T., Brosch, K., Dannlowski, U., Nenadić, I., Kircher, T., & Brakemeier, E. L. (2020). Childhood maltreatment and adult mental disorders – the prevalence of different types of maltreatment and associations with age of onset and severity of symptoms. *Psychiatry Research*, *293*(August), 113398. https://doi.org/10.1016/j.psychres.2020.113398

Taylor, P. A., Reynolds, R. C., Calhoun, V., Gonzalez-Castillo, J., Handwerker, D. A., Bandettini, P. A., Mejia, A. F., & Chen, G. (2023). Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *NeuroImage*, *274*(October 2022), 120138. https://doi.org/10.1016/j.neuroimage.2023.120138

Teicher, M. H., Anderson, C. M., & Polcari, A. M. (2012). Childhood maltreatment is associated with reduced volume in the hippocampal subfields CA3, dentate gyrus,and subiculum. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9). https://doi.org/10.1073/pnas.1115396109

Teicher, M. H., & Parigger, A. (2015). The "Maltreatment and Abuse Chronology of Exposure" (MACE) scale for the retrospective assessment of abuse and neglect during development. *PLoS ONE*, *10*(2), 1–37. https://doi.org/10.1371/journal.pone.0117423

Teicher, M. H., & Samson, J. A. (2013). Childhood maltreatment and psychopathology: A case for ecophenotypic variants as clinically and neurobiologically distinct subtypes. *The American Journal of Psychiatry*, *170*, 1114–1133. https://doi.org/10.1176/appi.ajp.2013.12070957

Teicher, M. H., Samson, J. A., Anderson, C. M., & Ohashi, K. (2016). The effects of childhood maltreatment on brain structure, function and connectivity. *Nature Reviews Neuroscience*, *17*(10), 652–666. https://doi.org/10.1038/nrn.2016.111

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

Tozzi, L., Garczarek, L., Janowitz, D., Stein, D. J., Wittfeld, K., Dobrowolny, H., Lagopoulos, J., Hatton, S. N., Hickie, I. B., Carballedo, A., Brooks, S. J., Vuletic, D., Uhlmann, A., Veer, I. M., Walter, H., Bülow, R., Völzke, H., Klinger-König, J., Schnell, K., … 'for the ENIGMA-MDD Consortium.' (2020). Interactive impact of childhood maltreatment, depression, and age on cortical brain structure:

mega-analytic findings from a large multi-site cohort. *Psychological Medicine*, *50*(6), 1020–1031. https://doi.org/10.1017/S003329171900093X

Van Harmelen, A. L., Van Tol, M. J., Van Der Wee, N. J. A., Veltman, D. J., Aleman, A., Spinhoven, P., Van Buchem, M. A., Zitman, F. G., Penninx, B. W. J. H., & Elzinga, B. M. (2010). Reduced medial prefrontal cortex volume in adults reporting childhood emotional maltreatment. *Biological Psychiatry*, *68*(9), 832–838. https://doi.org/10.1016/j.biopsych.2010.06.011

Viola, T. W., Salum, G. A., Kluwe-Schiavon, B., Sanvicente-Vieira, B., Levandowski, M. L., & Grassi-Oliveira, R. (2016). The influence of geographical and economic factors in estimates of childhood abuse and neglect using the Childhood Trauma Questionnaire: A worldwide meta-regression analysis. *Child Abuse and Neglect*, *51*(305141), 1–11. https://doi.org/10.1016/j.chiabu.2015.11.019

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

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 319–324. https://doi.org/10.1111/j.1745-6924.2009.01132.x

Whittle, S., Simmons, J. G., Hendriksma, S., Vijayakumar, N., Byrne, M. L., Dennison, M., & Allen, N. B. (2017). Childhood maltreatment, psychopathology, and the development of hippocampal subregions during adolescence. *Brain and Behavior*, *7*(2), 1–9. https://doi.org/10.1002/brb3.607

Wingenfeld, K., Spitzer, C., Mensebach, C., Grabe, H. J., Hill, A., Gast, U., Schlosser, N., Höpp, H., Beblo, T., & Driessen, M. (2010). The German Version of the Childhood Trauma Questionnaire (CTQ): Preliminary Psychometric Properties. *Psychotherapie Psychosomatik Medizinische Psychologie*, *60*, 442–450. https://doi.org/10.1055/s-0030-1253494

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., … 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

Yang, W., Jin, S., Duan, W., Yu, H., Ping, L., Shen, Z., Cheng, Y., Xu, X., & Zhou, C. (2023). The effects of childhood maltreatment on cortical thickness and gray matter volume: A coordinate-based meta-analysis. *Psychological Medicine*, *53*, 1681–1699. https://doi.org/10.1017/S0033291723000661

Zhou, X., Wu, R., Zeng, Y., Qi, Z., Ferraro, S., Xu, L., Zheng, X., Li, J., Fu, M., Yao, S., Kendrick, K. M., & Becker, B. (2022). Choice of Voxel-based Morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. *Communications Biology*, *5*(913). https://doi.org/10.1038/s42003-022-03880-1

Data Sharing Statement

Comprehensive non-thresholded statistical estimates are made openly available via the OSF (https://osf.io/j8d9r/?view_only=9edf436ab18f4e8db9ef4c71c4ac356c). Individual data points are not published due to current EU data protection regulations but can be made available in form of summary statistics or anonymized aggregation of voxel-wise data upon reasonable request to the corresponding author.

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 RE 22 17 07 to JR and the Deanery of the Medical Faculty of the University of Munster. SM received funding by the "Innovative Medizinische Forschung" (IMF) of the medical faculty of Münster (ME122205) and the Else-Kröner Fresenius Stiftung (EKFS; 2023_EKEA.153). JG also received funding by the IMF (GO122301). TH was supported by the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, HA7070/4). 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.