Received: 30 October 2020 | Revised: 22 January 2021 | Accepted: 12 February 2021

DOI: 10.1002/hbm.25386

RESEARCH ARTICLE

WILEY

Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder

Thomas Wolfers[1,2,3] | Jaroslav Rokicki[1,2] | Dag Alnæs[1,2] | Pierre Berthet[1,2] | Ingrid Agartz[2,4,5,6] | Seyed Mostafa Kia[3] | Tobias Kaufmann[2] | Mariam Zabihi[3] | Torgeir Moberget[1,2] | Ingrid Melle[2] | Christian F. Beckmann[3,7] | Ole A. Andreassen[2,4] | Andre F. Marquand[3,7,8] | Lars T. Westlye[1,2,4]

[1]Department of Psychology, University of Oslo, Oslo, Norway

[2]Division of Mental Health and Addiction, Norwegian Center for Mental Disorders Research (NORMENT), University of Oslo and Oslo University Hospital, Oslo, Norway

[3]Donders Center for Cognitive Neuroimaging, Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, The Netherlands

[4]KG Jebsen Center for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway

[5]Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway

[6]Department of Clinical Neuroscience, Center for Psychiatric Research, Stockholm, Sweden

[7]Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, The Netherlands

[8]Department of Neuroimaging, Center for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK

Correspondence
Thomas Wolfers Norment, Department of Psychology, University of Oslo, Oslo, Norway.
Email: thomas.wolfers@psykologi.uio.no

Funding information
H2020 European Research Council, Grant/Award Numbers: 802998, 847776; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 895011; Helse Sør-Øst RHF, Grant/Award Numbers: 2014097, 2015073, 2016083, 2019101; Niels Stensen Fellowship, Grant/Award Number: individual fellowship; Norges Forskningsråd, Grant/Award Numbers: 223273, 249795, 298646, 300768, 276082; Wellcome Trust, Grant/Award Number: 215698/Z/19/Z

Abstract

Identifying brain processes involved in the risk and development of mental disorders is a major aim. We recently reported substantial interindividual heterogeneity in brain structural aberrations among patients with schizophrenia and bipolar disorder. Estimating the normative range of voxel-based morphometry (VBM) data among healthy individuals using a Gaussian process regression (GPR) enables us to map individual deviations from the healthy range in unseen datasets. Here, we aim to replicate our previous results in two independent samples of patients with schizophrenia (*n*1 = 94; *n*2 = 105), bipolar disorder (*n*1 = 116; *n*2 = 61), and healthy individuals (*n*1 = 400; *n*2 = 312). In line with previous findings with exception of the cerebellum our results revealed robust group level differences between patients and healthy individuals, yet only a small proportion of patients with schizophrenia or bipolar disorder exhibited extreme negative deviations from normality in the same brain regions. These direct replications support that group level-differences in brain structure disguise considerable individual differences in brain aberrations, with important implications for the interpretation and generalization of group-level brain imaging findings to the individual with a mental disorder.

Andre F. Marquand and Lars T. Westlye are considered as last authors.

KEYWORDS
bipolar disorder, brain imaging, heterogeneity, individual patient, mental disorders, normative modeling, schizophrenia

1 | INTRODUCTION

Recently, the degree of inter-individual heterogeneity in brain structure was found to be considerably larger than previously anticipated for both schizophrenia and bipolar disorder (Wolfers et al., 2018). As expected, based on the substantial body of literature reporting results from case–control comparisons, patients with schizophrenia or bipolar disorder show evidence of group level deviations from a normative trajectory in brain structure. However, applying normative modeling (Marquand et al., 2016; Marquand et al., 2019) to chart variation in brain anatomy across individual patients showed highly idiosyncratic patterns of deviation, suggesting that such group effects are inaccurate reflections of the brain aberrations found at the individual level (Wolfers et al., 2018). Of note, a similar high level of heterogeneity has recently also been observed in attention-deficit/hyperactivity disorder (Wolfers et al., 2019) and autism spectrum disorder (Zabihi et al., 2019).

Given the existing literature on reproducible group-level differences in brain structure between cases and controls (Van Erp et al., 2016; Moberget et al., 2017), our initial findings of substantial heterogeneity within disorders demonstrated that moving beyond the study of group differences is highly beneficial to understand variability within clinical cohorts and may be required to make inferences at the level of the individual. Due to these important implications, we here report an attempt to replicate and extend our initial findings in two independent samples acquired on different scanners following an identical analytical procedure as in our previous discovery study.

2 | METHODS

2.1 | Participants

Table 1 summarizes the demographic and clinical information of the replication samples and the sample used in the discovery publication (Wolfers et al., 2018). For replication sample 1, we included 94 patients with a schizophrenia diagnosis, 116 patients with a bipolar disorder diagnosis and 400 healthy individuals. As the replication sample 2, we included 312 healthy individuals, 105 with schizophrenia diagnosis and 61 with bipolar diagnosis. As the discovery sample, we selected 256 healthy individuals, 163 patients with schizophrenia and 190 with bipolar disorder. All participants were recruited from the same population and catchment area but there was no overlap between the discovery and replication samples. All participants were recruited as part of the Thematically Organized Psychosis (TOP) study, approved by the Regional Committee for Medical Research Ethics and the Norwegian Data Inspectorate (Doan et al., 2017). The two replication samples were

TABLE 1 Demographics

	Replication 2[a]			Replication 1[a]			Discovery		
Demographics	Healthy	BP	SZ	Healthy	BP	SZ	Healthy	BP	SZ
N	312	61	105	400	116	94	256	190	163
Male (%)	58.01%	45.90%	64.76%	50.25%	35.34%	57.44%	54.70%	41.80%	64.40%
Age (mean ± std)	30 ± 8.0	31 ± 11.8	27 ± 8.8	34 ± 11.3	31 ± 10.8	28 ± 9.2	34 ± 9.5	34 ± 11.3	31 ± 8.7
Years of education (mean ± std)	14.3 ± 2.3	13.8 ± 2.1	11.9 ± 2.1	14.4 ± 2.4	13.8 ± 2.2	12.7 ± 2.3	14.0 ± 2.3	13.6 ± 2.3	12.9 ± 2.6
Symptom scores[b]									
PANSS global (mean ± std)	NA	24.2 ± 4.8	30.9 ± 8.8	NA	25.5 ± 5.3	30.6 ± 7.6	NA	25.4 ± 5.7	32.1 ± 8.6
PANSS negative (mean ± std)	NA	9.3 ± 3.1	16.2 ± 6.2	NA	9.6 ± 5.3	15.7 ± 6.4	NA	10.1 ± 3.5	15.8 ± 6.3
PANSS positive (mean ± std)	NA	9.0 ± 2.7	14.5 ± 5.7	NA	9.3 ± 3.2	13.4 ± 4.3	NA	10.0 ± 3.6	15.1 ± 5.5
PANSS total (mean ± std)	NA	42.6 ± 7.9	61.6 ± 18.2	NA	44.5 ± 9.3	59.8 ± 15.9	NA	45.5 ± 10.0	63.1 ± 16.8

Abbreviations: BP, bipolar disorder; NA, not applicable; PANSS, positive and negative syndrome scale; SZ, schizophrenia.
[a] The participants have been selected from the NoDa (local NORMENT database) database on the September 25, 2020.
[b] Symptom scores have been assessed using PANSS which is a standard clinical instrument for the quantification of positive and negative psychotic symptoms.

selected from the TOP-database on the 25th of September 2019. Patients were recruited from in- and out-patient clinics in the Oslo area, understood and spoke a Scandinavian language, had no history of severe head trauma, and had an IQ above 70. Patients were assessed by trained physicians or clinical psychologists. Psychiatric diagnosis was established using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID). Symptoms were assessed using PANSS (Kay et al., 1987). We used the positive, negative and global summary scores of the PANSS which were combined to the total summary score. Healthy individuals were randomly sampled from national registries and neither they nor their relatives had a psychiatric or alcohol/substance use disorder or cannabis use during the last 3 months. Written informed consent was obtained from all participants.

2.2 | MRI acquisition

Discovery: Structural scans were obtained on 1.5 Tesla Siemens MAGNETOM Sonata scanner at Oslo University Hospital using a standard 32-channel head coil. T1-weighted images were acquired using a MPRAGE sequence with the following parameters: repetition time (TR) = 2,730 ms, echo time (TE) = 3.93 ms, flip angle (FA) = 7°. *Replication 1*: Structural scans were obtained on 3 Tesla GE 750 Discovery scanner at Oslo University Hospital using a standard 32-channel head coil. T1-weighted images were acquired using a BRAVO sequence with the following parameters: repetition time (TR) = 8.16 ms, echo time (TE) = 3.18 ms, flip angle (FA) = 12°. *Replication 2*: Structural scans were obtained on 3 Tesla GE Signa HDxT at Oslo University Hospital one subset with HNS coil the other subset with 8HRBRAIN coil. T1-weighted images were acquired using the following parameters: repetition time (TR) = 7.8 ms, echo time (TE) = 2.956 ms, and flip angle (FA) = 12.

2.3 | Estimation of gray matter volume

In the same way as in our previous study, raw T1-weighted MRI volumes were processed using the computational analysis toolbox version 12 (CAT12; http://www.neuro.uni-jena.de/software/), based on statistical parametric mapping version 12 (SPM12). Images were segmented, normalized, and bias-field-corrected using VBM-SPM12 (http://www.fil.ion.ucl.ac.uk/spm, London, UK) (Ashburner and Friston, 2000, 2003), yielding images containing gray and white matter segments. Prior to the estimation of the normative models, all gray and white matter volumes were smoothed with an 8 mm FWHM Gaussian smoothing kernel and we restricted our analyses to voxels included in the gray matter mask constructed for the discovery study.

2.4 | Normative modeling

As in our previous article, we estimated the normative model using Gaussian Process Regression (GPR) to predict VBM based regional gray matter volumes across the brain from age and sex. To avoid overfitting of the normative models, it is crucial to estimate predictive performance out of sample. Therefore, we estimated the normative range for this model in healthy individuals under 10-fold cross-validation, and then applied one model across all healthy individuals to patients with schizophrenia and bipolar disorder. GPR yields coherent measures of predictive confidence in addition to point estimates. This is important in normative modeling as we need this uncertainty measure to quantify the deviation of each patient from the group mean at each brain locus. Thus, we are able to statistically quantify deviations from the normative model with regional specificity, by computing a Z-score for each voxel reflecting the difference between the predicted and the observed gray matter volume normalized by the uncertainty of the prediction (Marquand et al., 2016).

In line with our previous article, we thresholded the individual normative probability maps at $p < .005$ (i.e., $|Z| > 2.6$) and extreme positive and extreme negative deviations from the normative model were defined based on this threshold. All extreme deviations were combined into scores representing the percentage of extreme positively and negatively deviating voxels for each participant, relative to the total number of voxels in the brain mask. We tested for associations between diagnosis and those scores using a nonparametric test corrected for multiple comparisons using the Bonferroni–Holm method (Holm, 1979) as well as an association with PANSS scores. We repeated these analyses using different thresholds $p < .05$ (i.e., $|Z| > 1.96$) as well as $p < .001$ (i.e., $|Z| > 3.1$) and also modeled extreme deviations using extreme value statistics (Fisher and Tippett, 1928). This is based on the notion that the expected maximum of any random variable converges to an extreme value distribution. Therefore, we estimated a maximum deviation for each subject by taking a trimmed mean of 1% of the top absolute deviations for each subject across all vertices and fit an extreme value distribution to these deviations. Thus in addition to our previous work we checked whether our results remain consistent independent of the thresholding procedure of the normative probability maps that we have introduced in different publications (Marquand et al., 2016; Wolfers et al., 2018, 2019; Zabihi et al., 2019). To assess the spatial extent of those extreme deviations, we created individualized maps and calculated the voxel-wise overlap between individuals from the same groups first by replicating the exact procedure of the discovery study then by introducing different thresholds to check consistency. In the main text we report this overlap for healthy individuals, and people diagnosed with schizophrenia and bipolar disorder. All analyses were performed in python3.6 (www.python.org) and scripts are available on GitHub (https://github.com/RindKind/). Also, in line with our previous article, we fed the normative probability maps into PALM (Winkler et al., 2015) to test for mean differences between groups by means of a general linear model framework and permutation-based inference.

3 | RESULTS

3.1 | Normative modeling

Figure 1 shows the spatial representation of the voxel-wise normative model, characterized by widespread gray matter decreases from

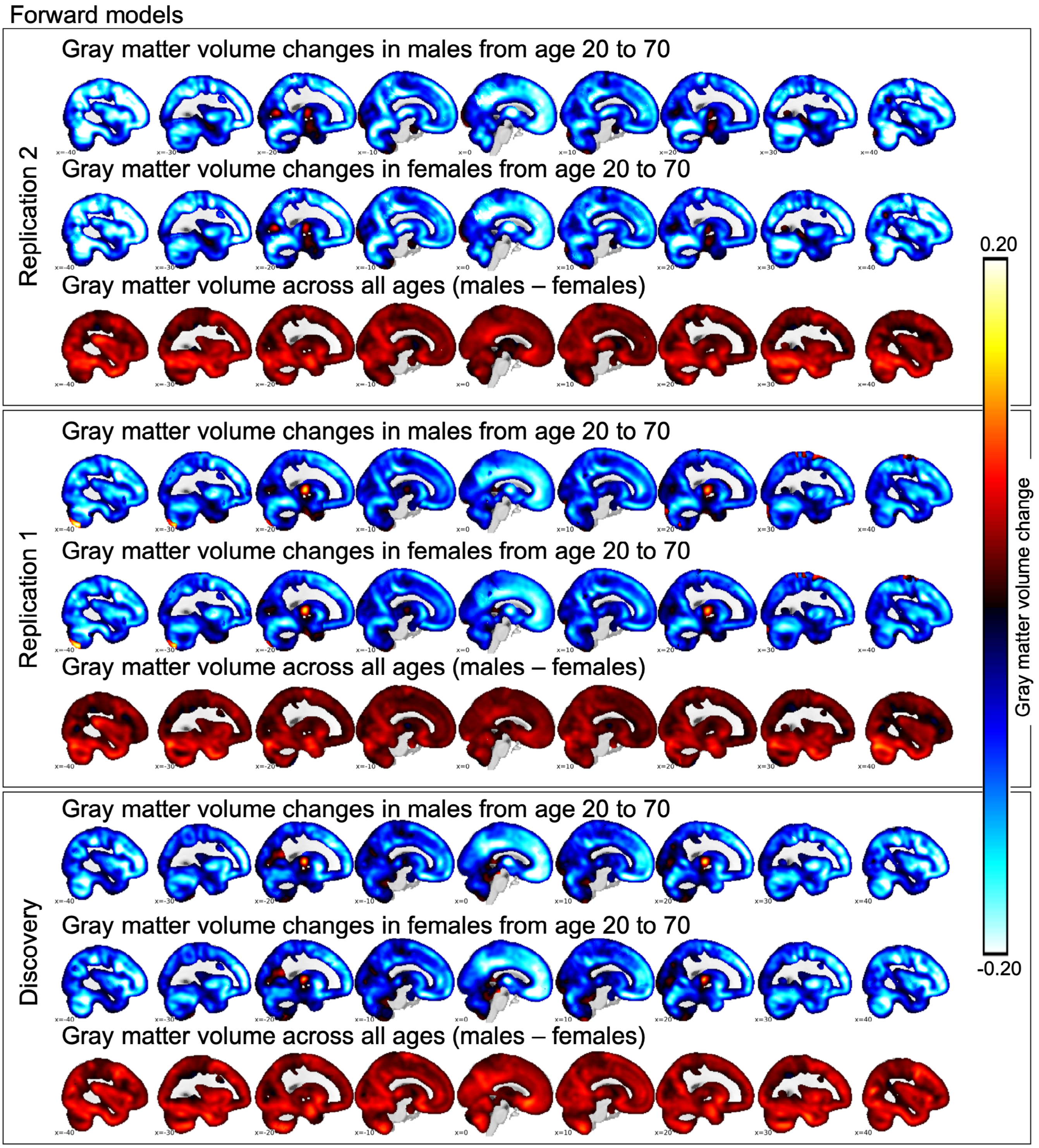
Forward models
Replication 2
Gray matter volume changes in males from age 20 to 70
Gray matter volume changes in females from age 20 to 70
Gray matter volume across all ages (males – females)
Replication 1
Gray matter volume changes in males from age 20 to 70
Gray matter volume changes in females from age 20 to 70
Gray matter volume across all ages (males – females)
Discovery
Gray matter volume changes in males from age 20 to 70
Gray matter volume changes in females from age 20 to 70
Gray matter volume across all ages (males – females)
x=-40
x=-30
x=-20
x=-10
x=0
x=10
x=20
x=30
x=40
0.20
Gray matter volume change
-0.20

FIGURE 1 We depict the slope of a linear approximation of the normative model for males (first row in each panel) and females (second row in each panel) as well as the difference between males and females across the entire age range from 20 to 70 years (third row in each panel). In the lower panel, we depict results based on the data reported in Wolfers et al. 2018, JAMA Psychiatry. In the upper two panels, we depict two replications. *Note:* These approximations are based on the forward model of the estimated normative models

age 20 to 70, with most pronounced age-differences in frontal areas. We depict models for discovery and replication studies separately in this figure. Further, we could show that the models performed well across the whole brain by plotting the correlation of predicted and observed values under 10-fold cross validation. This is depicted in Figure S1.

3.2 | Group comparisons

Figure 2 shows the result from pairwise group comparisons, corrected for multiple comparisons using permutation testing in PALM. In gray matter, patients with schizophrenia show stronger mean negative deviations than healthy individuals in frontal, temporal, and cerebellar regions; mean deviations are also more negative than in patients with bipolar disorder and localized primarily in frontal brain regions (Figure S2). These results replicate well across the three samples. However, for bipolar disorder the replication is not as strong showing

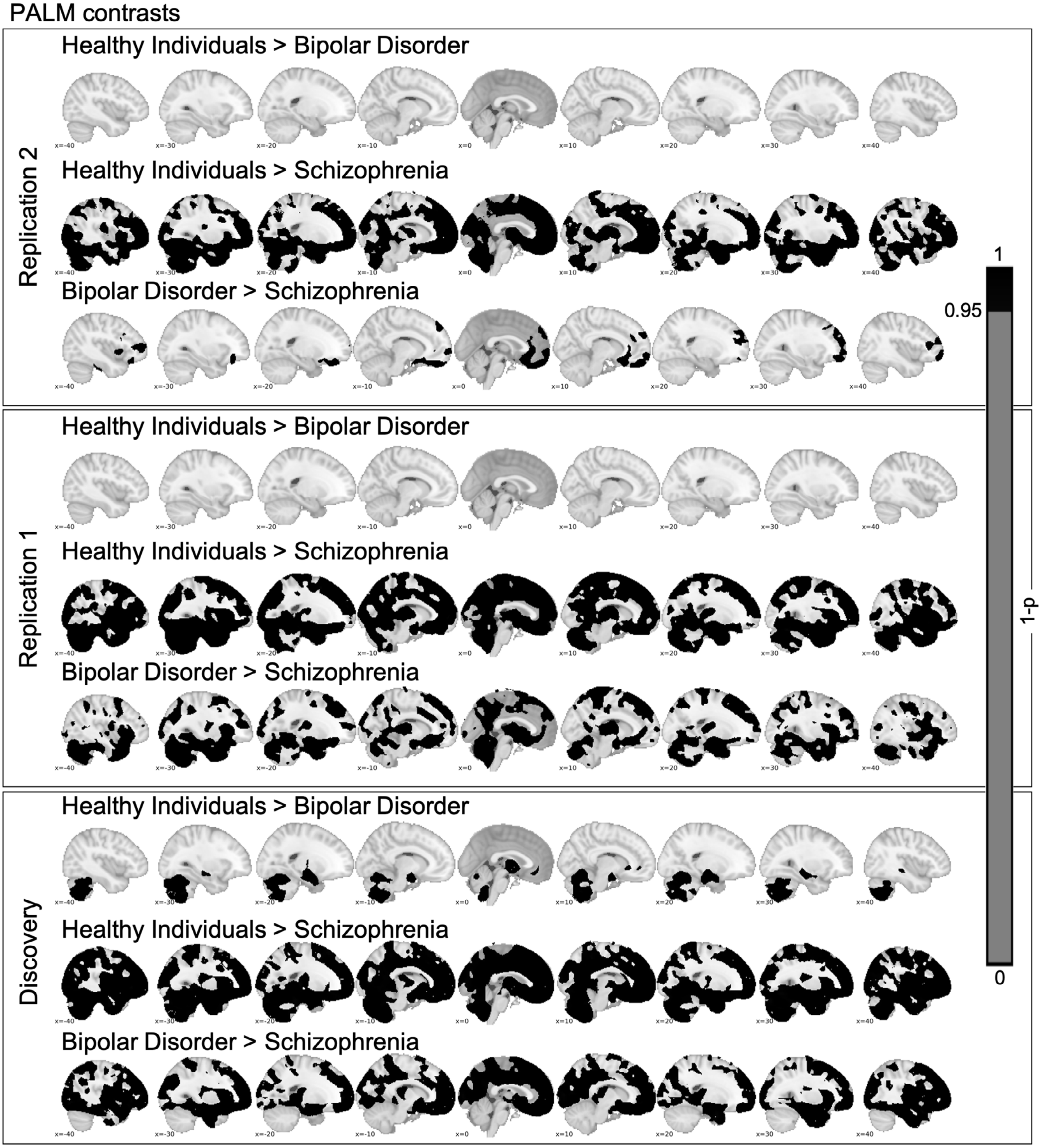
PALM contrasts
Replication 2
Healthy Individuals > Bipolar Disorder
Healthy Individuals > Schizophrenia
Bipolar Disorder > Schizophrenia
Replication 1
Healthy Individuals > Bipolar Disorder
Healthy Individuals > Schizophrenia
Bipolar Disorder > Schizophrenia
Discovery
Healthy Individuals > Bipolar Disorder
Healthy Individuals > Schizophrenia
Bipolar Disorder > Schizophrenia
x=-40
x=-30
x=-20
x=-10
x=0
x=10
x=20
x=30
x=40
1
0.95
1-p
0

FIGURE 2 We depict the contrast between healthy individuals, bipolar disorder and schizophrenia. In the lower panel, we depict results based on the data reported in Wolfers et al. 2018, JAMA Psychiatry. In the upper two panels, we depict two replications. For the PALM-derived mean Z-scores see Figure S2. *Note:* We report one subtracted by multiple comparison corrected *p* values

TABLE 2 Extreme deviations (|Z| > 2.6)

	Replication 2			Replication 1			Discovery		
Case–control	Healthy	BP	SZ	Healthy	BP	SZ	Healthy	BP	SZ
Extreme negative (mean, std)	0.22 ± 0.53%	0.35 ± 1.11%	1.09 ± 4.09%	0.16 ± 0.44%	0.14 + −0.34%	0.64 + −1.15%	0.23 + −0.78%	0.24 + −0.48%	0.90 + −2.15%
Significance	HC = BD HC < SZ (*p* < .001) BP < SZ (*p* < .01)			HC = BD HC < SZ (*p* < .001) BP < SZ (*p* < .001)			HC = BD HC < SZ (*p* < .001) BP < SZ (*p* < .001)		
Extreme positive (mean, std)	1.03 ± 2.06%	0.98 ± 2.59%	0.85 + 1.59%	1.16 + 1.99%	0.88 + 1.44%	0.60 + 0.90%	1.08 + 1.75%	0.79 + −1.25%	0.78 + −1.50%
Significance	HC = BD HC > SZ (*p* < .05) BP = SZ			HC = BD HC > SZ (*p* < .001) BP > SZ (*p* < .05)			HC = BD HC > SZ (*p* < .001) BP > SZ (*p* < .05)		
Dimensional	BP & SZ			BP & SZ			BP & SZ		
Extreme negative PANSS total[a]	*r* = .190 *p* < .05			*r* = .157 *p* < .05			*r* = .241 *p* < .001		
Extreme positive PANSS total[a]	*r* = .071 *p* > .05			*r* = −.072 *p* > .05			*r* = .035 *p* > .05		

Abbreviations: BP, bipolar disorder; PANSS, positive and negative syndrome scale; SZ, schizophrenia.

[a] Symptom scores have been assessed using PANSS which is a standard clinical instrument for the quantification of positive and negative psychotic symptoms.

differences from healthy controls in the cerebellum only in the discovery sample but not in the two replication samples. This may be due to a lower sample size of the bipolar and schizophrenia groups in both replication samples. The differences between patients with schizophrenia diagnosis and healthy individuals are very robust.

3.3 | Spatial distribution and statistical analyses of extreme deviations from normality

In line with our discovery study, in replication study 1 patients with schizophrenia show a higher percentage of extreme negative deviations across the brain (0.64 ± 1.15% of all voxels) compared to healthy individuals (0.16 ± 0.44%, Mann–Whitney $p < .001$) and individuals with bipolar disorder (0.14 ± 0.34%, Mann–Whitney $p < .001$, Table 2). The percentage of extreme positive deviations across participants and groups show that healthy individuals (1.16 ± 1.99%) differed from patients with schizophrenia (0.60 ± 0.90%; Mann–Whitney $p < .001$) and from individuals with bipolar disorder (0.88 ± 1.44%; Mann–Whitney $p < .05$). In replications study 2, patients with schizophrenia show a higher percentage of extreme negative deviations across the brain (1.09 ± 4.09% of all voxels) compared to healthy individuals (0.22 ± 0.53%, Mann–Whitney $p < .001$) and individuals with bipolar disorder (0.35 ± 1.11%, Mann–Whitney $p < .001$, Table 2). The percentage of extreme positive deviations across participants and groups show that healthy individuals (1.03 ± 2.06%) differed from patients with schizophrenia (0.85 ± 1.59%; Mann–Whitney $p < .001$) and from individuals with bipolar disorder (0.98 ± 2.59%; Mann–Whitney $p < .05$). This is an exact replication of the previous study (Table 2). Extreme negative deviations are on average 3.91, 4.00, and 4.95 times more prevalent in individuals with schizophrenia than in healthy controls across discovery and replication samples 1/2, respectively (Figures S3 and S4). All these results replicate for different Z-thresholds on the normative probability maps (Table S1) and also remain consistent with an estimate based on extreme value statistics (Table S1). Further we show that extreme negative deviations correlate significantly with symptom scores as measured by the PANSS across all samples, (discovery sample: $r = .241$, $p < .001$; replication sample 1: $r = .157$, $p < .05$; replication sample 2: $r = .190$, $p < .05$; Table 2). This shows that increasing number of symptoms is associated with more extreme negative deviations. This effect was only found across patient groups not within patient groups (Table S2), which may be due to lower power in individual groups or could potentially reflect a group difference rather than a dimensional effect across groups. Further, we could show an association of extreme negative deviations with the age of onset of both schizophrenia and bipolar disorder but not with other clinical characteristics, the duration of medication or lifetime episodes of psychotic, depressive, manic, or hypomanic events (Tables S3 and S4).

Extreme negative deviations in people with schizophrenia were most pronounced in temporal, medial frontal and posterior cingulate regions (Figure 3). In patients with bipolar disorder, the overlap was strongest in the thalamus. In line with our previous findings, the

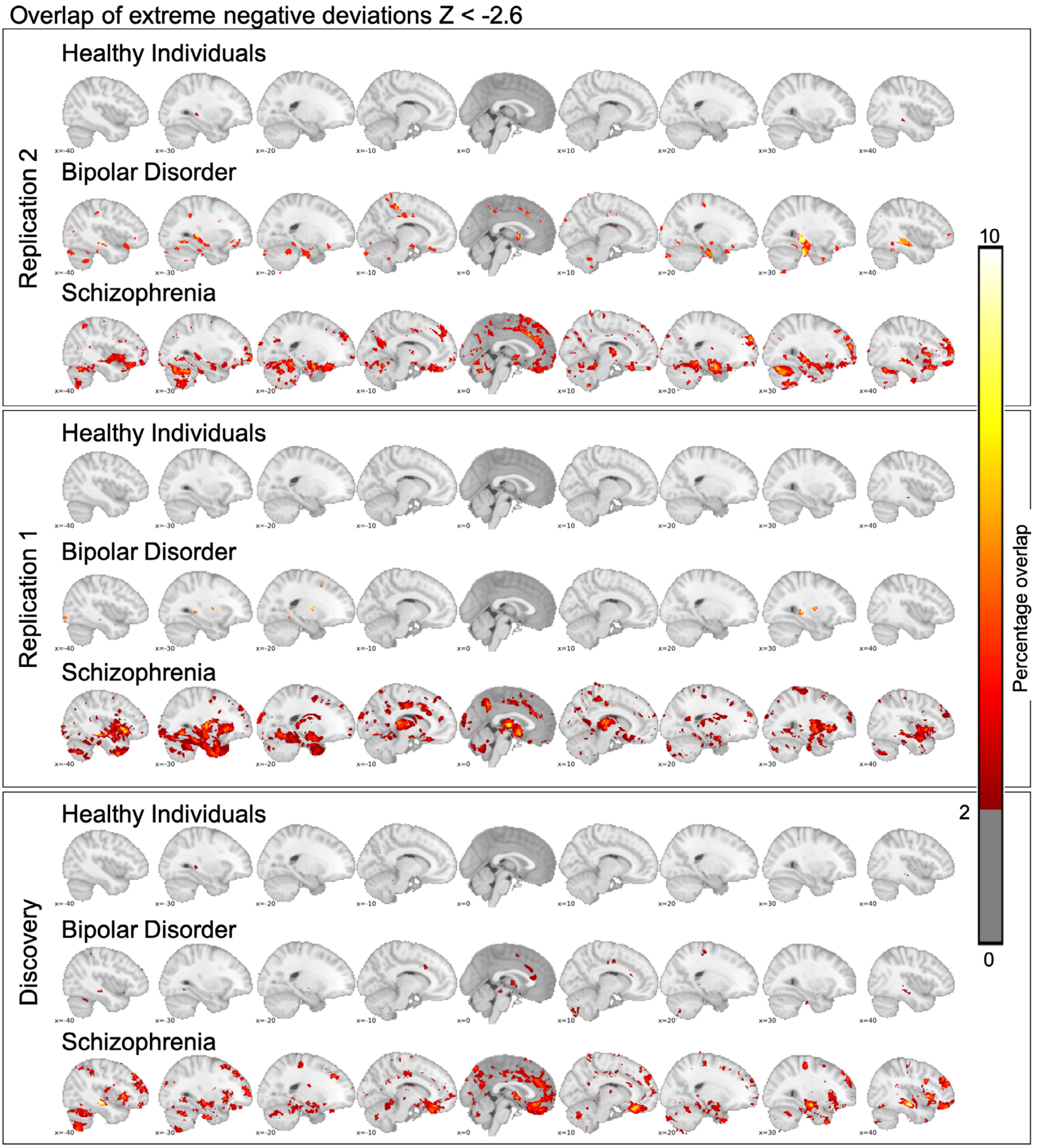

FIGURE 3 We show extreme negative deviations for healthy individuals, bipolar disorder and schizophrenia. In the lower panel, we depict results based on the data reported in Wolfers et al. 2018, JAMA Psychiatry. In the upper two panels, we depict two replications. We show that the overlap across studies is comparable with only a few brain regions showing overlap in more than 2% of the individuals. While the spatial overlap is similar especially for schizophrenia there are also differences. Note that by comparing Figures 2 and 3, it becomes apparent that robust group effects translate only to a relatively sparse overlap of extreme deviations from normality at the level of the individual. This replicates the main conclusion of the previous study. *Note:* Extreme negative deviations here are defined as $Z < -2.6$ at the individual level

overlap of extreme negative (Figure 3) and positive deviations from normality (Figure S5) is sparse across individuals with the same diagnosis, with peak voxels showing extreme negative overlap in 2.75% healthy individuals, 5.17% for individuals with bipolar disorder and 9.57% for schizophrenia in replication sample 1. In replication sample 2, the peak voxel shows extreme negative overlap in 3.52% of the healthy individuals, 8.19% of the individuals with bipolar disorder and 9.52% for individuals with a schizophrenia diagnosis. In expectation, this overlap increased when we applied a lower threshold $|Z| > 1.96$ (Figure S6) but was still sparse and decreased when we utilized a higher threshold $|Z| > 3.1$ (Figure S7) or and FDR threshold equal to 0.05 (Figure S8). Independent of the threshold the findings of the discovery study were replicated across two samples. Interestingly, when we stratified for sex the overlap of extreme negative deviations was stronger in males, which was consistent across samples and true for both disorders (Figure S9).

4 | DISCUSSION

Advanced brain imaging technology has allowed for probing the brain functional and structural correlates of complex human traits and mental disorders. While group-level normative deviations in brain structure in patients with a diagnosis of schizophrenia and bipolar disorder are robust and replicable (Figure 2) we observe substantial interindividual differences in the neuroanatomical distribution of extreme deviations at the individual level (Figures 3 and S4). These findings replicate and extend our previous study (Wolfers et al., 2018) and are largely in line with evidence of large heterogeneity across mental disorders (Wolfers et al., 2019; Zabihi et al., 2019).

Our results confirm that MRI-based brain structural aberrations in patients with severe mental disorders are highly heterogeneous in terms of their neuroanatomical distribution. These findings are in line with recent evidence of substantial brain structural heterogeneity in patients with schizophrenia (Alnæs et al., 2019) and also comply with accumulating evidence from psychiatric genetics strongly suggesting that mental illnesses are complex and heterogeneous disorders associated with a large number of genetic variants as well as environmental risk factors (Sullivan and Geschwind, 2019). Along with documented clinical heterogeneity (Insel, 2009) and large interindividual variability in treatment response and outcome (Kapur et al., 2012), our successful replication of considerable neuroanatomical heterogeneity supports the need for statistical approaches that allow for inferences at the level of the individual. Characterizing the magnitude and distribution of brain aberrations in individual patients is key for identifying neuronal correlates of specific symptoms across diagnostic categories and would represent an important step towards increasing the utility of brain imaging in a clinical context.

While the present findings are robust, it must be considered that other data modalities beyond those provided by structural brain imaging may be more able to capture any common pathophysiological processes in patients with schizophrenia or bipolar disorder. Thus, we may observe larger overlaps across individuals with the same mental disorders in other data domains, such as those measuring brain function, cognition or specific behaviors, on the network-level or relevant biological assays. While this possibility cannot be ruled out the present results indicate that multiple pathophysiological processes and pathways are at play, which is also supported by the large number of identified genetic variants (Ripke et al., 2014; Smoller et al., 2013; Stahl et al., 2019).

Over the last decades it has become increasingly apparent that replication attempts in psychology, psychiatry, neuroscience, and related fields frequently fail (Avinun et al., 2018; Dinga et al., 2019; Eklund et al., 2016; Hong et al., 2019; Ioannidis, 2005; Open Science Collaboration*, 2015; Tackett et al., 2019), which has fueled initiatives promoting reproducible science (Munafò et al., 2017; Poldrack et al., 2017; Schooler, 2014). The neuroimaging field is no exception, and lack of reproducibility in brain imaging studies have been attributed to the high researcher degree of freedom in terms of the many and sometimes arbitrary analytical choices (Eklund et al., 2016). Here, we strictly adhered to the analytical protocols as specified in our original study (Wolfers et al., 2018). The entire analytical pipeline is made publicly available to ease replication by independent researchers and to allow for application to different cohorts and disorders. Note here, that if you replicate these findings in a sample on multiple scanners using different scanning sequences your interpretation might be misguided due to scanning confounds. Currently, we are working on methods to improve normative modeling across sites (Kia et al., 2020). While we are convinced that the here used analytical protocols are appropriate for testing the reproducibility of our original report, the approaches will be improved in future studies and are under continuous development (Kia et al., 2020; Kia and Marquand, 2018). Moving forward, we will scale up this work towards larger samples covering a wider age range including neurodevelopment, incorporate different modalities and levels of information for example, brain network level, including genetics, and link different experimental designs to the normative modeling framework.

Our replications support that group level-differences in brain structure disguise considerable individual differences in brain aberrations. While we find additional similarity across discovery and replication study (Figures S2 and S3), such as extreme negative deviations are on average 3.91, 4.00, and 4.95 times more prevalent in individuals with schizophrenia than in healthy controls, we also find differences. Especially with respect to extreme positive deviations the pattern of overlap is as similar as it is different across studies (Figure S4). However, when we look at the same pattern with a Z-threshold of 1.96 the overlap of extreme positive deviations shows striking similarities (Figure S5, right panel). Further, we could not replicate a main group effect of bipolar disorder in comparison with healthy individuals in the cerebellum while this effect was present in the discovery sample (Figure 2). This may have been caused by differences in sample size. In addition, we want to point out that we worked on a predominantly adult sample, however, the onset of schizophrenia is primarily in adolescence or early adulthood. Therefore, it is important to investigate individual differences in this age group in future studies. Finally, we show results in addition to our previous study such as the correlation of extreme positive and negative deviations with PANSS scores. These results show that extreme negative deviations were associated with higher PANSS scores across all three samples but that this effect was only present when we pooled the bipolar and schizophrenia groups suggesting that it was driven by an increased power or by differences between the bipolar and schizophrenia patients rather than higher symptom scores. This interpretation is in line with the fact that we could replicate all previous findings of extreme negative deviations from normality across the two replication samples (Table 2). With low reproducibility rates across various scientific disciplines (Baker and Penny, 2016) building confidence through replication is critical.

CONCLUSION

Individuals with a mental disorder are sampled from a heterogenous general population based on their clinical and symptom profiles. One would expect a higher degree of similarity in terms of normative deviations in patients with the same diagnosis than in healthy individuals on measures affected by the disorder. In other words, these deviations would be enriched in the clinical as opposed to the general population. This is exactly what we observe and replicate across three samples. However, we do not detect it to the degree that group studies would suggest which generally show significant differences between patients and healthy individuals in terms of averages. Consequentially, these group differences say little about the individual patient with a mental disorder and his/her deviation from a norm, pointing out that we need individualized analyses instead of a focus on group studies in psychiatry. Therefore, the main conclusion of the discovery study is supported by replications across two samples, namely that group level-differences in brain structure captured by a classical case–control paradigm, disguises considerable individual differences in brain aberrations when we map deviations from normality.

ACKNOWLEDGMENTS

This study was performed on the TSD (Tjenester for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT) with resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway. The study is supported by the Research Council of Norway (223273, 249795, 298646, 300768, 276082), the South-Eastern Norway Regional Health Authority (2014097, 2015073, 2016083, 2019101), a Wellcome Trust Innovator award ("BRAINCHART", 215698/Z/19/Z), and the European Research Council under the European Union's Horizon 2020 research and Innovation program (ERC StG Grants 802998 and 847776). TW gratefully acknowledges the Niels Stensen Fellowship as well as the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant agreement No. 895011. We would like to thank the participants of these studies for their contribution. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

Ole A. Andreassen is a consultant to HealthLytix and received speaker's honorarium from Lundbeck. Christian F. Beckmann is shareholder of and director of SBG Neuro. The other authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Thomas Wolfers https://orcid.org/0000-0002-7693-0621
Lars T. Westlye https://orcid.org/0000-0001-8644-956X

REFERENCES

Alnæs, D., Kaufmann, T., van der Meer, D., Córdova-Palomera, A., Rokicki, J., Moberget, T., … Westlye, L. T. (2019). Brain heterogeneity in schizophrenia and its association with polygenic risk. *JAMA Psychiatry*, *76*, 739–748.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry: The methods. *NeuroImage*, *11*, 805–821. http://linkinghub.elsevier.com/retrieve/pii/S1053811900905822

Ashburner, J., & Friston, K. J. (2003). Spatial normalization using basis functions. In *Human Brain Function* (pp. 1–26). London, England: Academice Press.

Avinun, R., Nevo, A., Knodt, A. R., Elliott, M. L., & Hariri, A. R. (2018). Replication in imaging genetics: The case of threat-related amygdala reactivity. *Biological Psychiatry*, *84*, 148–159. https://doi.org/10.1016/j.biopsych.2017.11.010

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, *533*, 3–5.

Dinga, R., Schmaal, L., Penninx, B. W. J. H., Jose, M., Tol, V., Veltman, D. J., … Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage Clinical*, *22*, 101796. https://doi.org/10.1016/j.nicl.2019.101796

Doan, N. T., Kaufmann, T., Bettella, F., Jørgensen, K. N., Brandt, C. L., Moberget, T., … Westlye, L. T. (2017). Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. *NeuroImage Clinical*, *15*, 719–731.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences USA*, *113*, 7900–7905.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*, 180–190.

Holm, S. (1979). A simple sequentially Rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. https://doi.org/10.2307/4615733

Hong, Y. W., Yoo, Y., Han, J., Wager, T. D., & Woo, C. W. (2019). False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. *NeuroImage*, *195*, 384–395. https://doi.org/10.1016/j.neuroimage.2019.03.070

Insel, T. R. (2009). Disruptive insights in psychiatry: Transforming a clinical discipline. *The Journal of Clinical Investigation*, *119*, 700–705.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *The Journal of the American Medical Association*, *294*, 218–227.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it. *Molecular Psychiatry*, *17*, 1174–1179.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*, 261–276.

Kia SM, Marquand AF (2018): Neural processes mixed-effect models for deep normative modeling of clinical neuroimaging data. pp. 1–18. http://arxiv.org/abs/1812.04998.

Kia SM, Huijsdens H, Dinga R, Wolfers T, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF (2020): Hierarchical Bayesian regression for multi-site normative modeling of neuroimaging data, pp. 699–709.

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*, 552–561.

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, *24*, 1415–1424. http://www.nature.com/articles/s41380-019-0441-1

Moberget, T., Doan, N. T., Alnæs, D., Kaufmann, T., Córdova-Palomera, A., Lagerberg, T. V., … Westlye, L. T. (2017). Cerebellar volume and cerebellocerebral structural covariance in schizophrenia: A multisite mega-analysis of 983 patients and 1349 healthy controls. *Molecular Psychiatry*, *23*, 1–9. https://doi.org/10.1038/mp.2017.106

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie, N., … Wagenmakers, E. (2017). A manifesto for reproducible science Marcus. *Nature Human Behaviour*, *1*, 1–9. https://doi.org/10.1038/s41562-016-0021

Open Science Collaboration*. (2015). Estimating the reproducibility of psychological science. *Science (80–)*, *349*, 6251.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munaf, M. R., … Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Review Neuroscience*, *18*, 115–126.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427.

Schooler, J. W. (2014). Metascience could rescue the "replication crisis". *Nature*, *515*, 9.

Smoller, J. W., Kendler, K., Craddock, N., Lee, P. H., Neale, B. M., Nurnberger, J. N., … O'Donovan, M. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet*, *381*, 1371–1379.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., … Sklar, P. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature Genetics*, *51*, 793–803.

Sullivan, P. F., & Geschwind, D. H. (2019). Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell*, *177*, 162–183. https://doi.org/10.1016/j.cell.2019.01.015

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *The Annual Review of Clinical Psychology*, *15*, 579–604.

van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., … Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*, 547–553.

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., & Smith, S. M. (2015). Multi-level block permutation. *NeuroImage*, *123*, 253–268.

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., … Marquand, A. F. (2018). Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry*, *11*, 1146–1155.

Wolfers, T., Beckmann, C. F., Hoogman, M., Buitelaar, J. K., Franke, B., & Marquand, A. F. (2019). Individual differences v the average patient: mapping the heterogeneity in ADHD using normative models. *Psychological Medicine*, *50*(2), 314–323.

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., … Marquand, A. F. (2019). Dissecting the heterogeneous cortical anatomy of autism Spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*, 1–12. https://doi.org/10.1016/j.bpsc.2018.11.013

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wolfers T, Rokicki J, Alnæs D, et al. Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder. *Hum Brain Mapp*. 2021;42:2546–2555. https://doi.org/10.1002/hbm.25386