2-1-2022

Mental Health in the UK Biobank: A Roadmap to Self-Report Measures and Neuroimaging Correlates

Rosie K. Dutt
Washington University School of Medicine in St. Louis, rdutt@smith.edu

Kayla Hannon
Washington University School of Medicine in St. Louis

Ty O. Easley
Washington University School of Medicine in St. Louis

Joseph C. Griffis
Washington University School of Medicine in St. Louis

Wei Zhang
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: https://scholarworks.smith.edu/sds_facpubs

Part of the Data Science Commons, and the Statistics and Probability Commons

Recommended Citation

Authors

Rosie K. Dutt, Kayla Hannon, Ty O. Easley, Joseph C. Griffis, Wei Zhang, and Janine D. Bijsterbosch

WILEY

Mental health in the UK Biobank: A roadmap to self-report measures and neuroimaging correlates

Rosie K Dutt [ID] | Kayla Hannon | Ty O Easley [ID] | Joseph C Griffis [ID] |
Wei Zhang | Janine D Bijsterbosch

Department of Radiology, Washington
University School of Medicine, Saint Louis,
Missouri, USA

Correspondence
Janine D Bijsterbosch, Department of
Radiology, Washington University School of
Medicine, Saint Louis, MO 63110, USA.
Email: janine.bijsterbosch@wustl.edu

Abstract

The UK Biobank (UKB) is a highly promising dataset for brain biomarker research into population mental health due to its unprecedented sample size and extensive phenotypic, imaging, and biological measurements. In this study, we aimed to provide a shared foundation for UKB neuroimaging research into mental health with a focus on anxiety and depression. We compared UKB self-report measures and revealed important timing effects between scan acquisition and separate online acquisition of some mental health measures. To overcome these timing effects, we introduced and validated the Recent Depressive Symptoms (RDS-4) score which we recommend for state-dependent and longitudinal research in the UKB. We furthermore tested univariate and multivariate associations between brain imaging-derived phenotypes (IDPs) and mental health. Our results showed a significant multivariate relationship between IDPs and mental health, which was replicable. Conversely, effect sizes for individual IDPs were small. Test–retest reliability of IDPs was stronger for measures of brain structure than for measures of brain function. Taken together, these results provide benchmarks and guidelines for future UKB research into brain biomarkers of mental health.

KEYWORDS

brain correlates, depression, mental health, replication, test–retest, UK Biobank

1 | INTRODUCTION

Over the years, there has been a multitude of neuroimaging studies that aimed to investigate alterations in the brain in relation to affect-based mental health (e.g., anxiety and depression). The Major Depressive Disorder (MDD) literature reports structural changes in the cortico-limbic network (Klauser et al., 2015), insula and hippocampus (Stratmann et al., 2014), as well as functional changes in the Default Mode Network (DMN; Tozzi et al., 2021; Yu et al., 2019), medial

temporal gyrus, and caudate (Ma et al., 2012). In Generalized Anxiety Disorder (GAD), similar functional changes are seen in the DMN (Andreescu et al., 2011) and ventromedial prefrontal cortex (Cha et al., 2014), as well as structural changes in the DMN (Wolf et al., 2016) and amygdala (He, Xu, Zhang, & Zuo, 2016). However, the literature on neural correlates of MDD contains some inconsistent findings. For example, some studies report greater functional connectivity in the DMN (Greicius et al., 2007; Sheline, Price, Yan, & Mintun, 2010) while others report lesser functional connectivity in the same network (Bluhm et al., 2009; Tozzi et al., 2021; Yan et al., 2019). A potential reason for inconsistent findings is the small sample size of most of these

---

studies. The broader fields of psychology and neuroimaging are recognizing that small sample sizes lead to inflated effect sizes that often result from sampling variability and therefore do not replicate in new data (Button et al., 2013; Grady, Rieck, Nichol, Rodrigue, & Kennedy, 2021; Marek et al., 2020; Poldrack et al., 2017; Yarkoni, 2009). Larger sample sizes are therefore needed to obtain reliable insights into the neural correlates of mental health.

One option to achieve larger sample sizes is to conduct meta-analyses. Meta-analyses use results from prior studies as their input and employ quantitative methods to pool data across studies and test for consensus (Müller et al., 2018). A meta-analysis on resting-state functional connectivity in MDD showed hypo-connectivity in frontoparietal and salience networks and hyper-connectivity in the DMN (Kaiser, Andrews-Hanna, Wager, & Pizzagalli, 2015). Another meta-analysis showed that there are common gray-matter volume changes in MDD which are also seen in bipolar disorder (Wise et al., 2017). In GAD, meta-analyses have also been able to confirm consistent dysregulation of affective control related to numerous networks, which provides support for an integrated model of brain network changes (Xu et al., 2019). While these meta-analyses aid to establish consensus on brain correlates of mental health (Wager, Lindquist, & Kaplan, 2007), they can be limited in their scope. This is because the input studies surveyed in meta-analyses often adopt narrow inclusion and exclusion criteria for the patient sample, which limits cross-diagnostic mental health research. Additionally, due to the lack of availability of whole-brain statistical result images from prior studies, coordinate-based meta-analyses are often undertaken which are limited in their spatial precision (Müller et al., 2018). Furthermore, meta-analyses suffer from publication bias (only including effect sizes from published significant studies; Thornton & Lee, 2000), language bias (only including papers written in English; Egger et al., 1997), and selective outcome reporting (input-papers selectively publish only significant variables; Hutton & Williamson, 2000; Kirkham et al., 2010), which can lead to inflated meta-analytical results (Sterne, Egger, & Smith, 2001). These inherent limitations of meta-analyses may explain why disagreement persists within even meta-analytical work, with a recent study showing hypo-connectivity (rather than hyperconnectivity) in the core DMN in patients with depression (Tozzi et al., 2021).

Consequently, in recent years, there has been a move to accrue larger neuroimaging datasets such as the Young Adult and Lifespan Human Connectome Projects (HCP; Harms et al., 2018; Van Essen et al., 2013), Connectomes Related to Human Disease studies (CRHD; Tozzi et al., 2020), UK Biobank (UKB; Miller et al., 2016; Sudlow et al., 2015), Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA; Schmaal et al., 2017), and Adolescent Brain Cognitive Development study (ABCD; Casey et al., 2018). The increased statistical power afforded by these datasets enables studies to approximate the true effect (Marek et al., 2020). Currently, the UKB is the largest neuroimaging dataset, encompassing data from extensive questionnaires, physical and cognitive measures, and biological samples (including genotyping) in addition to multimodal neuroimaging scans (Sudlow et al., 2015). The UKB is a prospective epidemiological study that recruited a cohort of 500,000 participants, of which 100,000 subjects will take part in one round of imaging, and 10,000 of those subjects will undergo a further second round of scanning (Sudlow et al., 2015). Health outcomes for all participants will be tracked over future years until participants' decease, including full primary health and hospital records. Therefore, the UKB offers a valuable resource to study mental health and other disorders. The goal of our study is to establish a foundation for future mental health biomarker research in the UKB.

The UKB includes multiple rich self-report measures of mental health. However, the organization and abundance of this information can make it somewhat challenging for researchers to navigate. For data pertaining to mental health, there are three sources within the UKB. The first is the assessment center questions (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100060) which participants complete via a touch screen on the day they were scanned. The second is a separately administered online mental health questionnaire (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=136), which is completed by a subset of UKB participants at a time-independent from the scanning date (median absolute number of days between scan 1 and online questionnaire completion: 742, range: −1,185 to +964 days in exploratory sample). The third is the health records available in the UKB which encompass the date of the first experience of specific ICD-10 diagnoses obtained from primary care (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=3000) and hospital inpatient data (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=2000). In this study, we tabulate and compare different mental health measures available in the UKB, with a focus on self-reported symptom scores from the assessment center information and online questionnaire. We test their relationship with brain measures, thereby providing a benchmark for using UKB mental health variables in future research.

This study aims to achieve four key goals. First, we aim to clearly tabulate the different self-report measures of mental health available in the UKB and discern the relationships between summary scores to enable future studies to make an informed decision on which measure is most appropriate to use. Second, we propose and validate a new summary measure (Recent Depressive Symptoms; RDS-4) that uses depression questions which were asked on the day of scanning in the UKB study. The RDS-4 score, therefore, enables research into current depressive symptoms and changes in symptomatology over time. Third, we aim to establish realistic and robust univariate and multivariate effect sizes of commonly reported brain correlates of mental health based on population data. Finally, we aim to determine the test–retest reliability of imaging variables alongside their effect size as both reliability and sensitivity are critical requirements for biomarker research. Large-scale imaging datasets such as the UKB play a critical role in the long-term goal of finding brain biomarkers of mental health, and we hope to provide a foundation that future studies can build on.

2 | METHODS

2.1 | Dataset

Imaging data from 32,420 UKB participants were available at the time the study was performed. From this, we selected multiple

independent test cohorts (Figure 1 and Table 1). Subjects with a mean head motion greater than 0.2 mm were removed resulting in the exclusion of 5,265 subjects. Subjects with any missing online questionnaire or scan 1 assessment center mental health data were also removed, resulting in the exclusion of additional 10,848 subjects (largely because the online questionnaire was only performed in a subset of UKB participants). From the remaining 16,307 subjects, we selected individuals who had undergone brain scans at two timepoints. These subjects make up the test–retest sample.

Late-onset depression (first episode at age 60+) is associated with different brain correlates (e.g., white matter hyperintensities) and different risk factors (e.g., vascular risk) compared with recurrent early-onset depression (age of first episode before 60; Salo, Scharfen, Wilden, Schubotz, & Holling, 2019). Therefore, we assessed subjects for probable late-onset depression based on self-reported age at the first episode of depression (Data-field: 20433).

Subjects who reported their first episode at 60 or older (N = 418) were excluded.

The majority of individuals within the UKB cohort are expected to have no mental health conditions because it is a population sample. To ensure sufficient power to identify neural correlates of mental health, we wanted to reduce the expected over-representation of healthy individuals and ensure that our samples richly capture mental health variability. This was achieved by including equal numbers of participants with and without a history of mental health. From the UKB showcase we used: Seen doctor (GP) for nerves, anxiety, tension, or depression (Data-field: 2090) to ensure our samples included an equal number of subjects who experienced mental health issues on at least one occasion, and those who have not. For each subject who had seen a GP for nerves, anxiety, tension, or depression (N = 4,531), we paired a matched subject from those who had never seen a GP for nerves, anxiety, tension, or depression (i.e., subject pairs were identically matched for sex and age, and
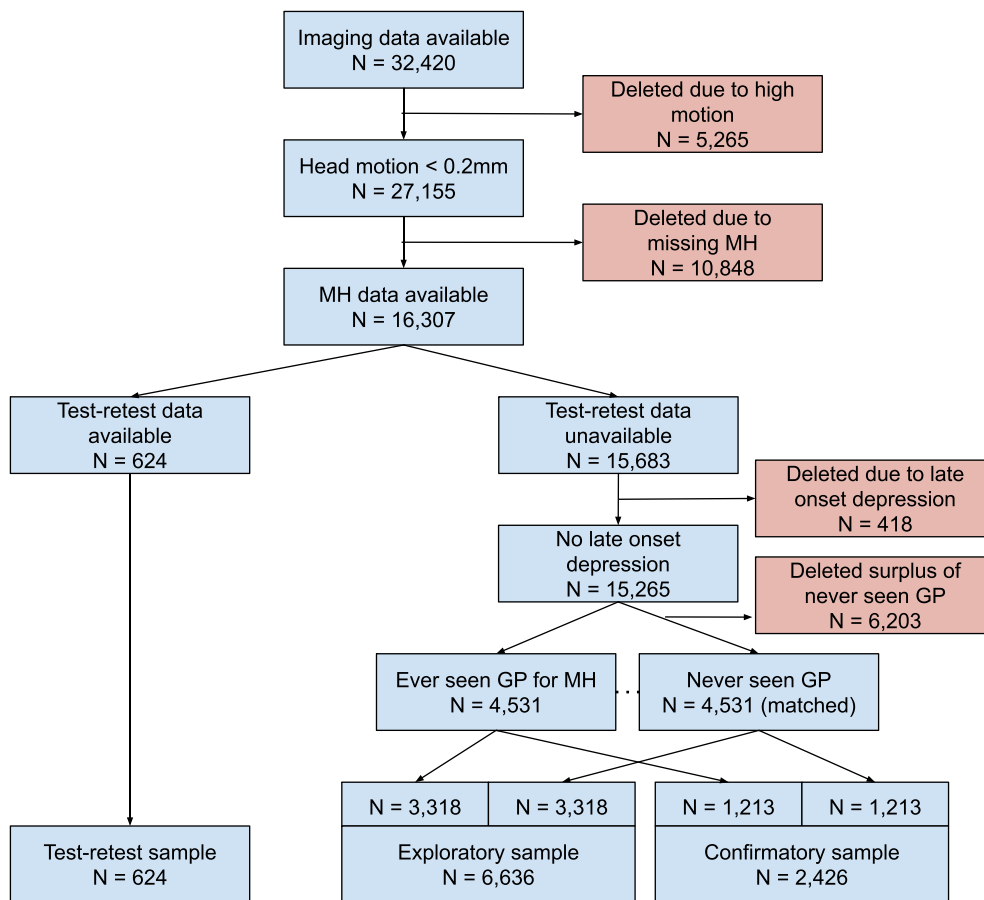
FIGURE 1 UK Biobank subject inclusion chart

TABLE 1 Demographics for samples

Sample	N	Sex (n male)	Age (mean ± SD)	Time between scans (mean absolute days ± SD)
Exploratory	6,636	2,258	61.9 ± 7.2	N.A.
Confirmatory	2,426	796	60.6 ± 7.1	N.A.
Test–retest	624	300	61.7 ± 7.04	823.7 ± 44.8

Note: The "ever seen GP for mental health" and "never seen GP for mental health" subjects were matched, such that the same male-to-female ratio and mean age applies to these groups.

minimal difference in head motion). Subsequently, approximately two-thirds of the "never seen GP" subjects together with their matched "seen GP" subjects were randomly assigned to the exploratory sample, and the remaining subjects were assigned to the confirmatory sample. During a subject assignment to groups, we preserved the matched characteristics within each resulting sample (Figure 1). No subjects overlapped between the exploratory and confirmatory samples.

2.2 | Mental health measures

The set of self-report questions related to mental health included in the UKB were informed by standardized measures, but did not simply cover a list of previously validated scales. Table 2 summarizes the five different UKB mental health measures, which will be used for neuroimaging and questionnaire comparison analyses and

TABLE 2 Measures of affect-based mental health available in the UK Biobank

	Scan day	Online	Range	Questions	Variable IDs
PHQ-9		✔	0–27	Little interest or pleasure in doing things	20514
				Feeling down, depressed, or hopeless	20510
				Trouble sleeping	20517
				Feeling tired	20519
				Poor appetite or overeating	20511
				Feeling bad about yourself	20507
				Trouble concentrating	20508
				Moving or speaking slowly or fidgety or restless	20518
				Thoughts that you would be better off dead	20513
RDS-4	✔		4–16	Frequency of depressed mood in last 2 weeks	2050
				Frequency of unenthusiasm/disinterest in last 2 weeks	2060
				Frequency of tenseness/restlessness in last 2 weeks	2070
				Frequency of tiredness/lethargy in last 2 weeks	2080
GAD-7		✔	0–21	Feeling nervous, anxious, or on edge	20506
				Not being able to stop or control worrying	20509
				Worrying too much about different things	20520
				Trouble relaxing	20515
				Being so restless that it is hard to sit still	20516
				Becoming easily annoyed or irritable	20505
				Feeling afraid as if something awful might happen	20512
N-12	✔		0–12	Mood swings	1920
				Miserableness	1930
				Irritability	1940
				Sensitivity/hurt feelings	1950
				Fed-up feelings	1960
				Nervous feelings	1970
				Worrier/anxious feelings	1980
				Tense/"highly strung"	1990
				Worry too long after embarrassment	2000
				Suffer from "nerves"	2010
				Loneliness, isolation	2020
				Guilty feelings	2030
Probable depression status	✔		0/1	Ever depressed	4598
				Ever unenthusiastic/disinterested	4631
				Duration of the longest period of depression	4609
				Duration of the longest period of unenthusiasm/disinterest	5375
				Seen doctor (GP) for nerves, anxiety, tension, and depression	2090
				Seen psychiatrist for nerves, anxiety, tension, and depression	2100

Abbreviations: GAD-7, general anxiety disorder-7; N-12, neuroticism-12; PHQ-9, patient health questionnaire-9; RDS-4, recent depressive symptoms-4.
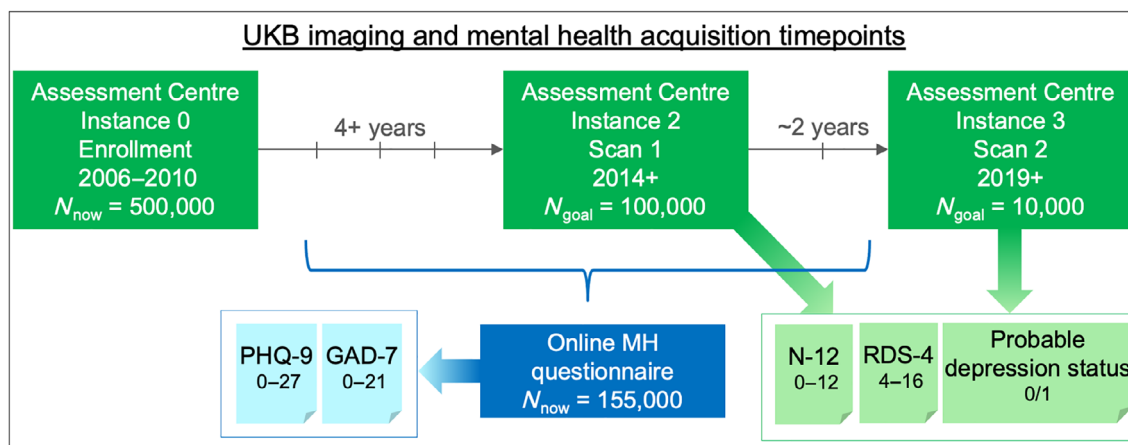
FIGURE 2 Schematic overview of the acquisition timing of UK Biobank mental health measures in relation to imaging acquisition. Mental health measures in light green were obtained on the day of scanning, whereas mental health measures in light blue were obtained at an independent time point that varied from 1,185 days before to 964 days after scan 1 across participants. The range of possible scores for each mental health measure is included. All five measures were included in neuroimaging and questionnaire comparison analyses in this article

Figure 2 provides an overview of the acquisition timing of these mental health measures relative to the scan days. The questions included in the online questionnaire enable calculation of the Generalized Anxiety Disorder (GAD-7) and Patient Health Questionnaire (PHQ-9) scores (Davis et al., 2020). Using the Assessment center information, the Eysenck Neuroticism (N-12) score was calculated. Smith et al. (2013) used questions from the assessment information to develop a categorical (case–control) measure of depression. For the purposes of our study, we adopted similar definitions to obtain a categorical assignment of Probable Depression Status, but we did not differentiate between single and recurrent episodes of depression. Depression status was set to 1 if subjects responded yes to variable IDs 4598 or 4631 (ever depressed|ever unenthusiastic/disinterested), *and* reported a duration of at least 1 week to variable IDs 4609 or 5375 (depression|unenthusiasm/ disinterest), *and* had seen either a GP or psychiatrist for nerves, anxiety, tension, depression (i.e., responded yes to variable IDs 2090 or 2100).

For our study, we proposed a new summary measure of state depression using UKB questions included in the Assessment center information: Recent Depressive Symptoms (RDS-4), which is a continuous measure of depression symptomatology obtained on the day of scanning. The four self-report questions used for the RDS-4 assess depressed mood, disinterest, restlessness, and tiredness. Each question asks about recent experiences of symptoms (past 2 weeks). The response options for the four questions are (a) not at all, (b) several days, (c) more than half the days, and (d) nearly every day. The summed score across these four variables, therefore, has a range of 4–16. Moreover, the RDS-4 questions correspond with several DSM-V diagnostic criteria for major depressive disorder and cover depression domains that are also considered in other measures such as the Hamilton and Montgomery–Åsberg scales.

There are a number of important differences between the RDS-4 and the other mental health measures. Compared to PHQ-9,

the RDS-4 was obtained on the day of the imaging scan, whereas the PHQ-9 was undertaken at a time point that was independent of the scan date. Compared to probable depression status, the RDS-4 provides a continuous measure of recent symptom severity, whereas probable depression status is a categorical (case–control) measure of lifetime occurrence of depression. Compared to N-12, RDS-4 is a measure of recent (state) depressive symptoms, whereas N-12 is a more general measure of personality (trait). Compared to GAD-7, the RDS-4 focuses on depression and the GAD-7 focuses on anxiety.

2.3 | Imaging acquisition

UKB structural modalities include T1-weighted (T1), T2-weighted (T2), susceptibility-weighted MRI (swMRI), diffusion MRI (dMRI), and functional modalities: task-based fMRI (tfMRI) and resting-state fMRI (rsfMRI). MRI data were obtained using a Siemens Magnetom Skyra 3 T scanner. For T1 structural scans, 3D MPRAGE acquisition was used to acquire 1 mm isotropic resolution. For T2 scans, fluid-attenuated inversion recovery (FLAIR) contrast was used with the 3D SPACE optimized readout providing a strong contrast for white matter hyperintensities. For swMRI, a 3D gradient echo acquisition was used (resolution: 0.8 × 0.8 × 3 mm), obtaining two echo times (TE = 9.4 and TE = 20 ms). Diffusion data was acquired with *b*-values of 1,000 and 2,000 s/mm^2, at 2 mm spatial resolution, with a factor 3 multiband acceleration and 50 distinct diffusion-encoding directions. Both tfMRI and rs-fMRI used identical acquisition parameters (spatial resolution = 2.4 mm, TR = 0.735 s, factor = 8 multiband accelerator). Task fMRI used the Hariri faces/shapes "emotion" task as employed in the HCP (Barch et al., 2013; Hariri, Tessitore, Mattay, Fera, & Weinberger, 2002), with a shorter total length and reduced repeats of the total stimulus block. For further information on UKB imaging, please refer to Miller et al. (2016).

2.4 | Imaging derived phenotypes

In addition to raw and processed imaging data, image-derived phenotypes (IDPs) are available for download. IDPs are derived from calculations that combine many images and/or voxels to produce a scalar quantity from the processed imaging data (Miller et al., 2016). Examples of IDPs include regional volumes from structural MRI and "edges" from resting-state functional MRI (i.e., connectivity between a pair of networks).

The IDPs included in this paper are summarized in Table 3, and further information can be found in (Miller et al., 2016) as well as the UKB showcase brain imaging documentation resource (https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/brain_mri.pdf). Briefly, resting-state IDPs were obtained using Independent Components Analysis performed at two different dimensionalities (25 and 100), which resulted in 21 and 55 signal networks, respectively. Subject-specific BOLD time series for each network were calculated using dual regression (Nickerson, Smith, Öngür, & Beckmann, 2017), and the amplitude for each network (temporal standard deviation) and functional connectivity between pairs of networks (full or partial correlation coefficients) were calculated. Resting-state IDPs from both ICA dimensionalities were included as they may offer complementary information at different levels of functional organization. From T1-weighted images, gray matter volumes were obtained with FSL FIRST and FAST, and cortical area and thickness were calculated with Freesurfer. Total volume of white matter hyperintensities was estimated based on T1-weighted and T2-flair images using FSL's BIANCA algorithm (Griffanti et al., 2016). From the diffusion data,

weighted mean fractional anisotropy (FA) and mean diffusivity (MD) were obtained using FSL's DTIFIT tool. Task fMRI IDPs reflect summary measures of activation (the median and 90th percentile for both the percent signal change and the z-statistic) in regions selected from the group-level activation map. Susceptibility weighted IDPs were generated from the signal decay times predicted from the magnitude images at the two TEs such that the IDPs equate to the median signal decay times.

2.5 | Confound variables

All analyses were corrected for the "simple" set of confounds described in (Alfaro-Almagro et al., 2021), namely, scanning site, age, age squared, sex, age * sex, head size, head motion in resting fMRI and in task fMRI scans, date, and date squared. This confound set was previously shown to explain 4.4% of variance in UKB imaging variables on average and captured the most important sources of confound variation (Alfaro-Almagro et al., 2021).

2.6 | Correlations among mental health variables in the UKB

To characterize the degree of overlapping information between mental health measures, Spearman rank correlations were computed between all measures of mental health using data from the exploratory sample ($N = 6,636$).

TABLE 3 Full set of IDPs considered for canonical correlation analysis

	# IDPs	UKB ID	Description
Resting state	21	25754	rfMRI network amplitudes from 21 signal components
	55	25755	rfMRI network amplitudes from 55 signal components
	210	25750	Pairwise full correlation edges between 21 components
	210	25752	Pairwise partial correlation edges between 21 components
	1485	25751	Pairwise full correlation edges between 55 components
Total = 3,466	1485	25753	Pairwise partial correlation edges between 55 components
Structural	139	1101	FAST gray matter volumes
	14	1102	FIRST gray matter volumes
	62	196	Cortical surface area from Freesurfer DKT atlas
	62	196	Cortical thickness from Freesurfer DKT atlas
	1	25781	Total volume of white matter hyperintensity
	27	107	Weighted-mean FA
	27	107	Weighted-mean MD
Total = 346	14	107	Median T2-star from susceptibility-weighted imaging
Task	16	106	Task fMRI median + 90th percentile of BOLD effect and z

Abbreviations: IDP, imaging-derived phenotype; UKB, UK Biobank.

Data used to compute RDS-4 and N-12 were collected at the scan date (assessment center information), whereas GAD-7 and PHQ-9 were computed from data obtained from the online questionnaire. The absolute number of days elapsed between the two data collections ranged from 0 to 1,185 days. To investigate the effects of measurement latency on mental health measure correlation, Spearman rank correlations between the RDS-4 and PHQ-9 (both measures of depression) were computed as a function of elapsed time between measurement (see Supplementary Materials Section S1 and Figure S1).

To test whether self-report measures differed significantly based on probable depression status, a two-sample Kolmogorov–Smirnov test was performed to ascertain whether subjects with a positive depression status had different distributions of depression scores than subjects with no depression status.

2.7 | Mapping between mental health variables in the UKB

To gain insights into how the different measures of mental health included in the UKB relate to each other, we used equipercentile linking in the exploratory sample. Here, the stepwise percentiles for each measure were calculated, and for each score in one measure, the equivalent percentile rank in a different measure was mapped (Kolen & Brennan, 2014). We further calculated the Cronbach alpha for the newly proposed RDS-4 score to measure internal consistency in the exploratory sample.

2.8 | Mechanical Turk study to validate RDS-4

To further validate the proposed RDS-4 score, we performed an independent study using the Amazon Mechanical Turk platform via CloudResearch.com (Litman, Robinson, & Abberbock, 2017). Participants were paid a nominal compensation for questionnaire completion. One hundred thirty-four participants aged 60+ completed the study. This study was reviewed by the Washington University in St Louis IRB board and approved as exempt (IRB #201909165) because participants were fully anonymous (the option of anonymized worker IDs in CloudResearch was adopted) and no participant key was available to any member of the research team.

Participants completed the same set of mental health questionnaires at two time points 7 days apart using the Qualtrics software (Qualtrics, Provo, UT). The following questionnaires were presented in randomized order: RDS-4, PHQ-9, CES-D (Center for Epidemiological Studies—Depression; Radloff, 1977) and MASQ-30 (short-form Mood and Anxiety Symptoms Questionnaire; Wardenaar et al., 2010; Watson & Clark, 1991). The latter two measures were included because they are commonly used measures of depression that can be considered "gold standard" for self-report. Although these measures are not available in the UKB, our goal was to validate the RDS-4 against these standardized measures.

We undertook multiple steps to avoid low-quality responses, which can be a concern in Mechanical Turk questionnaire research. First, we adopted premium options in CloudResearch, such as only including "CloudResearch approved participants" who undergo more extensive vetting. Second, we included two questions to assess the attention levels of the participants while performing the study ("If you are still paying attention, please select 'yes'" & "Please answer this question with the 'Most or all of the time' option"). Participants who failed to answer these questions appropriately were excluded. Third, we imposed a minimum duration for questionnaire completion at 172.5 s (which equals 2.5 s per question). Participants who completed the questionnaire in less than 172.5 s were excluded.

Spearman rank correlation was used to compare scores between the RDS-4, PHQ-9, CES-D, and MASQ-30 using data from time point 1. Intraclass correlation coefficient (ICC A,1; also known as criterion-referenced reliability; Koo & Li, 2016; McGraw & Wong, 1996) was used to calculate the test–retest reliability between time point 1 and time point 2 separately for each measure.

2.9 | Exploratory brain–mental health analysis

We used Canonical Correlation Analysis (CCA) as a data-driven approach to identify joint multivariate relationships between mental health measures and brain imaging variables (Hotelling, 1936). Following nuisance regression to remove variance explained by nuisance regressors, dimensionality reduction was performed separately for resting state, structural, and task IDPs (Table 3) using Principal Component Analysis (PCA). The substantial differences in IDP numbers between resting-state IDPs (3,466), structural IDPs (346), and task fMRI IDPs (16) were the reason for performing the dimensionality reduction separately to ensure that all classes of IDPs were represented in the input components. The top components explaining at least 50% of variance were retained for each of resting state, structural, and task IDPs. This threshold was chosen as a good trade-off between retaining a substantial amount of IDP variance for the CCA while limiting the number of input variables to the CCA to ensure a sufficient subject-to-variable ratio required for stable CCA results (Helmer et al., 2021). The structural and task IDP matrix included a small number of missing values, which were excluded for the nuisance regression and then imputed using nearest neighbor imputation (MATLAB's knnimpute.m). The combined set of IDP eigenvectors were entered into the CCA against five mental health input variables corresponding to summary scores from GAD-7, N-12, PHQ-9, RDS-4, and probable depression status (residuals after regressing out confound variables). CCA was performed on $N = 6,636$ subjects in the exploratory sample. Permutation testing with 2,000 permutations was used to obtain p values for the resulting canonical correlations. Here, the subject order of IDP component inputs and mental health inputs were independently shuffled to break subject correspondence. This is especially important for CCA because the canonical correlation is explicitly maximized and therefore it is important to compare the canonical correlation to the empirical null distribution obtained with permutation

testing (which does not center around zero but shows relatively high null correlations; Smith et al., 2015).

To calculate the univariate contributions (or "loadings") from individual IDPs to the CCA result, we correlated subject scores against original IDPs. For this purpose, the "U" and "V" canonical subject scores from the strongest CCA result were averaged within each subject to obtain a CCA summary subject score (UV). Here, $U = XA$ and $V = YB$, where X is the IDP principal component inputs and Y is the mental health inputs. A and B are the canonical coefficients for IDP eigenvectors and mental health variables respectively, which are optimized such that the correlation between U and V is maximized. We could calculate IDP contributions by correlating U with the IDPs, but the resulting correlations would potentially be inflated because U is optimized for X. Therefore, using the averaged UV subject score for correlations with the IDPs provides a more realistic and unbiased measure of individual IDP correlations (Bijsterbosch et al., 2018). Bonferroni correction for multiple comparisons was performed for these post-hoc correlations that were used to estimate univariate contributions from each original IDP (i.e., p-value below $.05/(3,466 + 346 + 16) = 1.3 \times 10^{-5}$, where 3,466 is the number of resting-state IDPs, 346 is the number of structural IDPs, and 16 is the number of task IDPs). IDPs that survived correction were selected for subsequent tests of effect size in the confirmatory sample. These IDPs are referred to as "selected brain variables" in subsequent confirmatory analyses.

The multivariate CCA results were also replicated in the independent confirmatory sample by projecting the resting state, structural, and task IDPs onto the same PCA subspace (i.e., not repeating the PCA, but using the weights from the exploratory sample), and multiplying brain eigenvectors as well as mental health scores by their respective canonical coefficients (i.e., A & B as estimated from the exploratory sample). The CCA replication was tested based on the correlation between the resulting U and V (i.e., the canonical correlation). We also performed the same post-hoc univariate correlations between averaged UV and individual IDPs as described above to assess the replicability of IDP contributions to the CCA.

2.10 | Confirmatory analysis of effect size

The independent confirmatory sample ($N = 2,426$) was used to test univariate effect sizes of selected brain variables from CCA analysis (i.e., significant IDPs after Bonferroni correction). Specifically, we performed a Cohen's d test based on probable depression status, and calculated the Pearson's r from the correlations between the selected brain variables and each of the four mental health variables (i.e., RDS-4, PHQ-9, N-12, and GAD-7), respectively. These analyses were repeated for each imaging modality including surface area, gray matter volume, cortical thickness, white matter hyperintensity, fractional anisotropy, median T2*, task activity, resting-state network amplitude, and edge connectivity at both dimensionalities (i.e., 25 and 100). We de-confounded both the brain variables and the mental health variables before running the aforementioned analyses. The only

exception from de-confounding is the binary grouping based on probable depression status, as de-confounding would result in subject-specific values that are noncategorical, which is unsuitable for the Cohen's d test.

2.11 | Test–retest reliability of imaging measures

To assess the stability of IDPs across time, we performed test–retest reliability analyses using data from $N = 624$ subjects that were scanned twice at separate time-points, with an inter-scan interval of approximately 2 years (Table 1). Data were de-confounded for this sample using the same approach employed for the exploratory CCA analysis. After data were de-confounded, intra-class correlations were computed between the IDPs collected at each scan time-point using the ICC(A,1) formulation to quantify the agreement between measurements collected at each timepoint (McGraw & Wong, 1996). Test–retest reliability measures were grouped according to IDP measurement modality (e.g., cortical area, cortical volume, etc.) to allow for assessment of the ICC distributions for different modalities.

We also assessed the effect of inter-scan interval length on the test–retest correlation strengths by computing ICCs for each IDP after including regressing out the inter-scan interval (in days) from each IDP, thus removing any additional variance attributable to inter-subject differences in inter-scan interval lengths. Finally, we assessed whether ICCs were affected by mental health changes as indicated by the difference in the RDS-4 scores between timepoints. Of the $N = 624$ subjects included in the test–retest analyses, $n = 336$ exhibited no change in RDS-4 scores between timepoints, while $n = 288$ exhibited changes in RDS-4 scores between time-points (i.e., at least 1 point difference in the RDS-4 scores). For these analyses, we separately computed ICCs for each mental health sub-group and then plotted the ICC distributions for each modality between the sub-groups. We also computed ICCs after regressing out mental health change values from each IDP.

3 | RESULTS

3.1 | Correlations among mental health variables in the UKB

Mental health measures showed moderate correlations with one another, indicating redundancy between these metrics (Figure 3a). RDS-4 and N-12, which are both measured from questions administered on the scan date, had a Spearman rank correlation coefficient (SRCC) of $\rho = .57 \pm 0.01 \left(p \approx 10^{-199} \right)$; PHQ-9 and GAD-7, which were both taken from the online questionnaire, have SRCC $\rho = .69 \pm 0.01 \left(p \approx 10^{-267} \right)$. Correlations between PHQ-9 and GAD-7 scores were significantly higher than between any other pairs of scores ($p < 10^{-9}$).

Both the RDS-4 and N-12 measures were collected at each scan time, which allows for an assessment of the within-measure 2-year
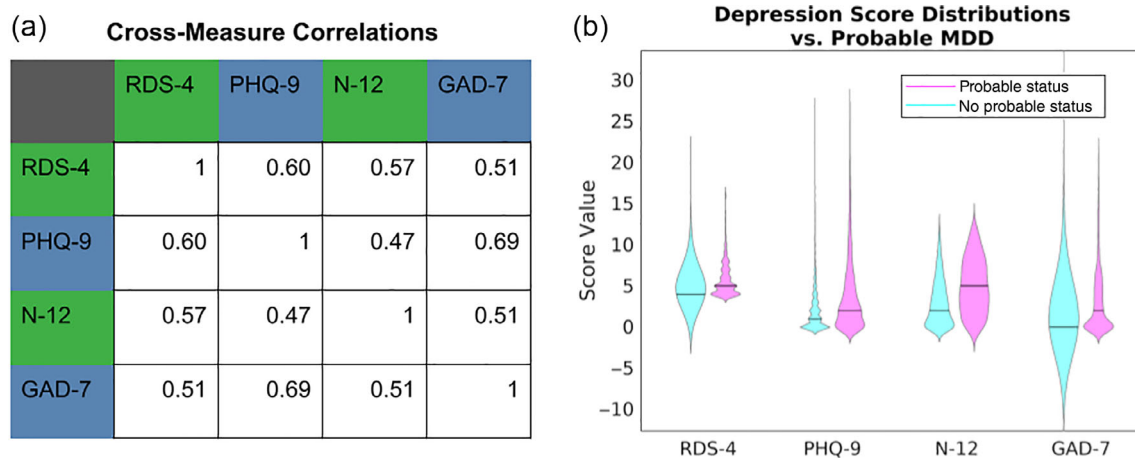
(a) **Cross-Measure Correlations**

	RDS-4	PHQ-9	N-12	GAD-7
RDS-4	1	0.60	0.57	0.51
PHQ-9	0.60	1	0.47	0.69
N-12	0.57	0.47	1	0.51
GAD-7	0.51	0.69	0.51	1

(b)

FIGURE 3 (a) Spearman rank correlation coefficients between each pair of mental health measures. Variables measured on the same date are labeled the same color (green = assessment center day-of-scan information; blue = online questionnaire). (b) Distributions of scores for subjects with probable depression status (pink) and without probable depression status (cyan). Subjects with probably depression status scores significantly higher on all mental health measures (KS-statistic $\chi \geq 0.19, p \leq 10^{-48}$)

correlation of these measures on a sample of $N = 555$ subjects from the test–retest sample (69 subjects were removed from the full $N = 624$ test–retest sample due to missing mental health assessment center information on scan 2). Within this subgroup, the subjects' RDS-4 measures showed a 2-year Spearman rank correlation coefficient of $\rho = .57$ between initial and follow-up scans, and N-12 showed a 2-year correlation of $\rho = .85$. It should be noted that this reflects the correlation between scan timepoints between 761 and 980 days apart. Therefore, a given metric's 2-year correlation (i.e., self-correlation over a long time period) effectively establishes an approximate upper bound on any correlation value between it and other metrics collected over the same time frame. Because anxiety and depression are not fixed states and scores may meaningfully differ between the two timepoints available in the UKB, we also performed a separate Mechanical Turk study to test the short-term (7-day) test–retest reliability of RDS-4 (see Section 3.3).

We performed a two-sided, two-sample Kolmogorov–Smirnov test on RDS-4, PHQ-9, N-12, and GAD-7 scores over subjects with and without probable depression status. Subjects with probable depression scored significantly higher than subjects with no probable depression status on all measures (KS-statistic $\chi \geq 0.19, p \leq 10^{-48}$; Figure 2b).

3.2 | Mapping between mental health variables in the UKB

Given that this is a largely healthy sample, as expected, the distributions for PHQ-9, RDS-4, and GAD-7 all reveal a large number of participants with scores on the lower end of the mental health measure, with a sharp decline seen in the number of participants scoring on the upper end of the mental health measures (Figure 4a–d). Notably, the distribution of N-12 is relatively less skewed than PHQ-9, RDS-4, and GAD-7.

Equipercentile linkage was used to map between different measures of mental health. The results show a stable and approximately linear mapping between RDS-4 and PHQ-9 (Figure 4e). Additionally, our results show the stable mapping between RDS-4 and N-12 (Figure 4f), and between N-12 and GAD-7 (Figure 4g). These results are in line with the literature showing that the personality trait of neuroticism is closely associated with mental health (Lahey, 2009).

We calculated Cronbach's internal consistency alpha for RDS-4, which measures the internal consistency. The Cronbach alpha for RDS-4 was .78, which indicates a moderate to strong internal reliability. This was similar to N-12 (Cronbach alpha = .83).

3.3 | Mechanical Turk study to validate RDS-4

Out of 134 subjects who completed our separate validation study, three subjects were removed because they failed the attention questions and a further 44 subjects were removed because they completed the surveys too fast, resulting in $N = 87$ subjects (53 female and 34 male; mean age 66.0 ± 4.8). The results showed that RDS-4 was highly correlated with other depression scales and achieved test–retest reliability comparable to other depression scales (Table 4).

3.4 | Exploratory brain–mental health analysis

Prior to performing the CCA, the data reduction of resting-state IDPs resulted in 100 components which explained 50.1% of variance. The data reduction of the structural IDPs resulted in 24 components which explained 50.6% of variance. The data reduction of the task IDPs resulted in two components which explained 51.2% of variance. Therefore, the total number of brain variables input into the CCA was 126 and this was tested against the five mental health variables. The
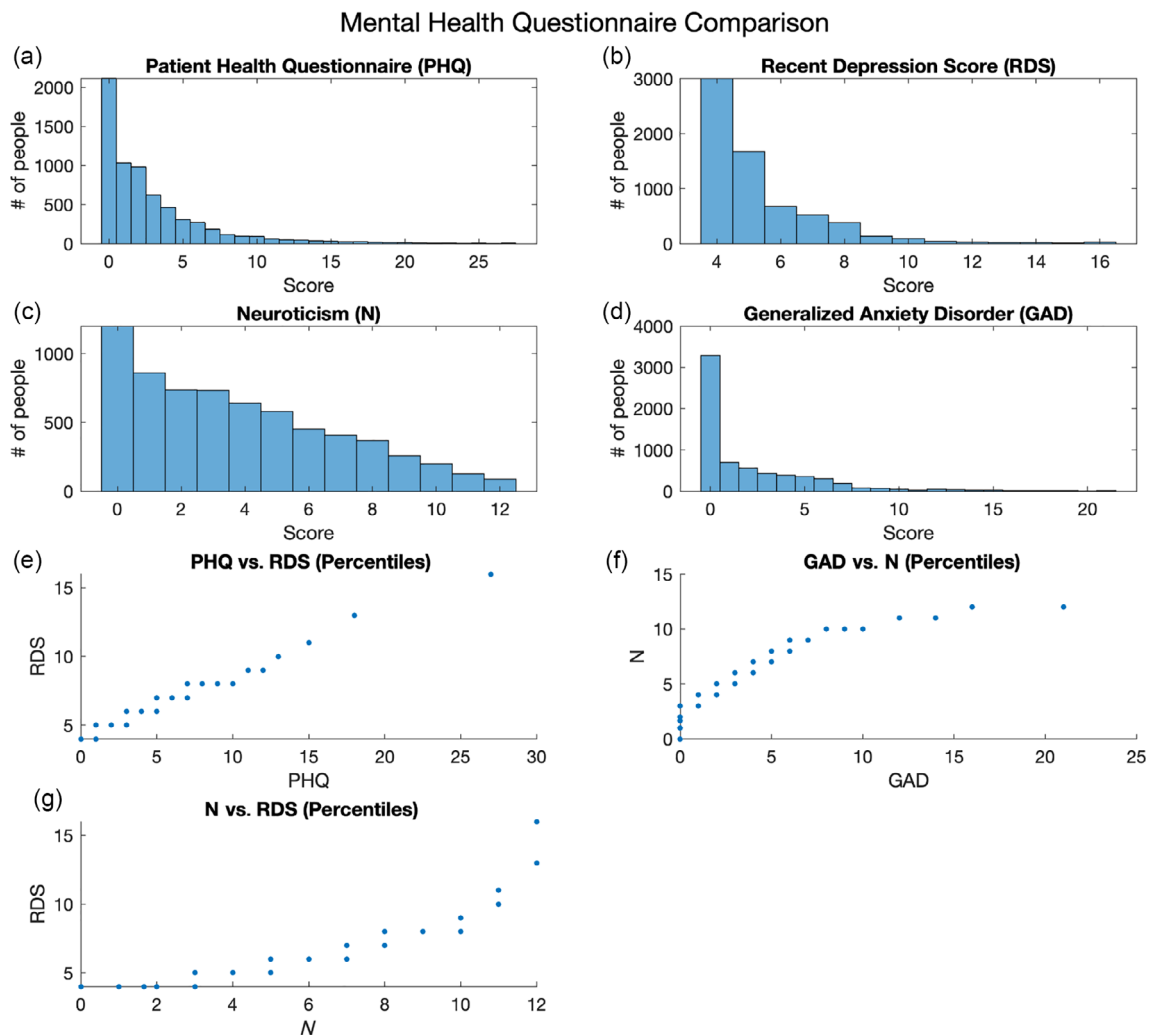
Mental Health Questionnaire Comparison

FIGURE 4 Panels (a)–(d) are the distributions of scores for participant responses to each questionnaire. Panels (e)–(g) depict the equipercentile linkages of the scores for each questionnaire, mapping the equivalence of a score from one questionnaire to the score of the other questionnaire

TABLE 4 Comparison of RDS-4 to other depression scales from MTurk study

	Test–retest reliability (ICC)	Correlation with RDS-4 (ρ)
RDS-4	0.88	–
CES-D	0.91	.89
PHQ-9	0.94	.91
MASQ general distress	0.87	.78
MASQ anhedonic depression	0.82	.67
MASQ anxious arousal	0.92	.71

CCA resulted in two significant canonical covariates ($R1_{UV} = 0.207$, $p = .0005$, and $R2_{UV} = .174$, $p = .015$). The first multivariate canonical correlation partly replicated in the independent confirmatory sample ($R1_{UV[confirmatory]} = 0.125$, $p = 3.7 \times 10^{-9}$, where the p-value was Bonferroni corrected for the maximum of five canonical correlations). Although the second canonical correlation also reached significance in the confirmatory sample ($R2_{UV[confirmatory]} = 0.06$, $p_{Bonferroni} = .02$), we did not perform post-hoc analysis for this finding due to the low canonical correlation in the replication sample. There are a number of factors that may have contributed to the replicability of the first canonical correlation. First, the CCA was relatively well-powered with 50.7 subjects per input variable leading to relatively stable estimates (Helmer et al., 2021). Second, the exploratory and confirmatory samples were well matched in terms of sample characteristics. Third, data reduction of IDPs before CCA likely reduces measurement noise. Post-hoc correlations between the averaged UV subject scores and the mental health variables and IDPs also replicated well (Figures 5 and S2).

In terms of post-hoc correlations with IDPs, 770 resting-state IDPs and 86 structural IDPs, and 1 task IDP were significantly correlated with the canonical covariate (UV) after Bonferroni correction for multiple comparisons. The post-hoc CCA results confirm many regions previously highlighted in the literature such as prefrontal and orbitofrontal cortices.
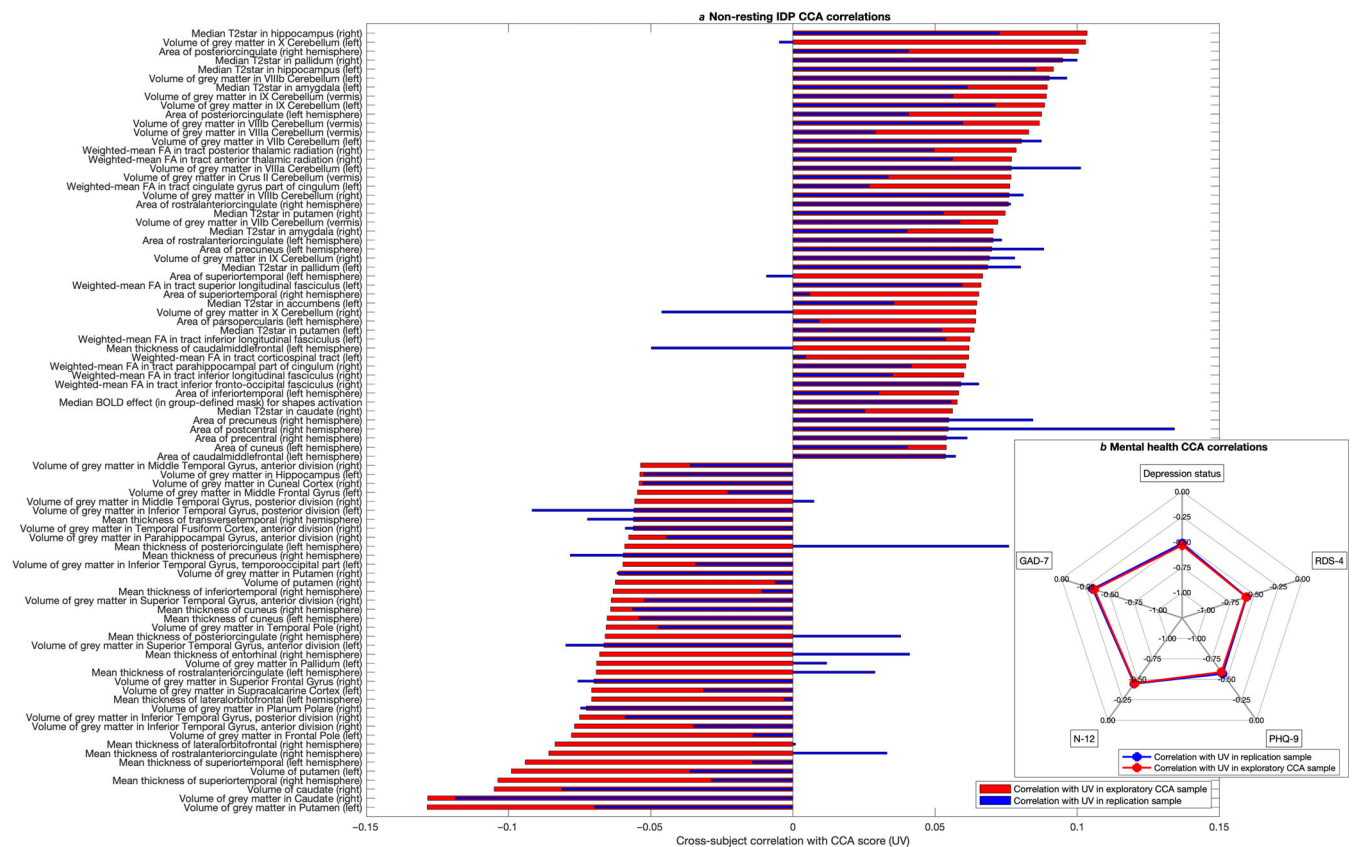
FIGURE 5 Canonical correlation results. (a) Post-hoc correlations for nonresting (structural and task) IDPs, showing only significant IDPs after Bonferroni correction. A similar figure for the resting state IDPs is included in Figure S2. (b) (inset): Post-hoc CCA relations for mental health show that the first canonical covariate is broadly linked to affect-based mental health

IDPs that contributed significantly to the CCA were also tested for univariate direct correlations with individual mental health variables in the independent confirmatory sample (see next section for the results). For these follow-up univariate tests, we furthermore supplemented the target IDPs with a literature-curated list (Table S1) that partly overlaps with the data-driven IDP identification.

3.5 | Confirmatory analysis of effect size

Our findings showed that univariate effect sizes of the relationship between IDPs and mental health determined in our robust population sample were very low. Overall, effect sizes of the differences in the brain variables (i.e., IDPs), indicated by Cohens' d, based on probable depression status, were larger than the Pearson's r values from correlations between IDPs and continuous mental health measures (Figure 6). On average, resting-state node amplitude and edge connectivity derived from partial correlation matrices appeared to have the higher effect sizes in most mental health measures, and task activity and fractional anisotropy ranked high in some mental health measures. At the level of individual IDPs, edges derived from both partial and full correlation matrices emerged as the best "predictors" in explaining data variance in all mental health variables except for

PHQ-9 where amplitude of a few resting-state nodes ranked at top (Figures S3–S7). These findings together suggest an overall higher effect size of resting-state in contrast to nonresting state measures on the investigated mental health variables.

3.6 | Test–retest reliability of imaging measures

We next assessed the stability of IDPs over time in 624 subjects who had data from two separate scan sessions conducted approximately 2–2.5 years apart. Figure 7a shows the distribution of inter-scan intervals for all 624 subjects. To assess test–retest reliability, ICCs were computed between the scan 1 measurements for each IDP and the corresponding scan 2 measurements for the same IDP. Then, the ICCs were assigned to categories based on the measurement modality of the corresponding IDPs: brain surface area (62 measures), brain volume (154 measures), cortical thickness (CT: 62 measures), fractional anisotropy (FA: 27 measures), mean diffusivity (MD: 27 measures), T2* value (T2: 14 measures), task activation (TA: 16 measures), resting-state time-series amplitudes (AMP: 76 measures), full correlation-based resting-state networks (FNT: 1,695 measures), and partial correlation-based resting-state networks (PNT: 1,695 measures).
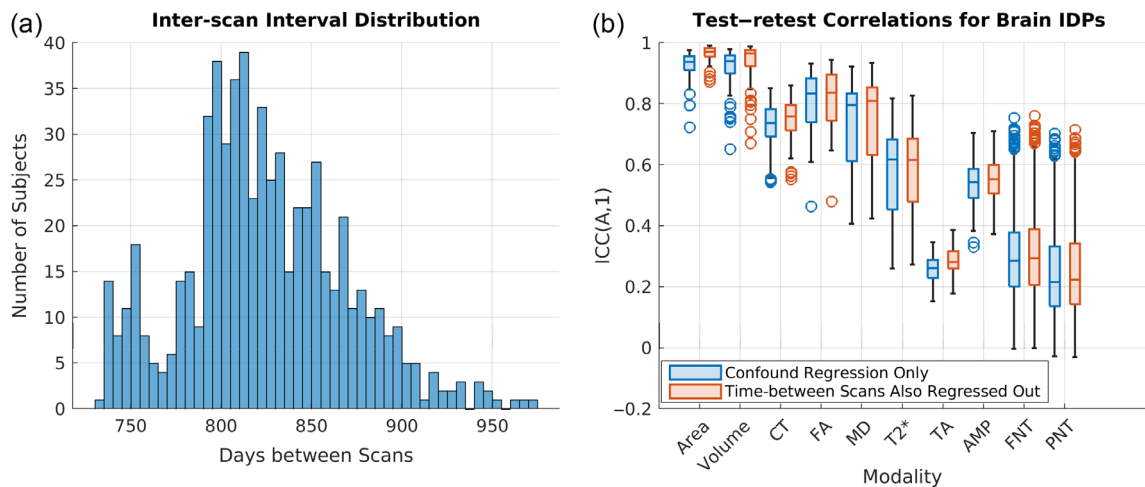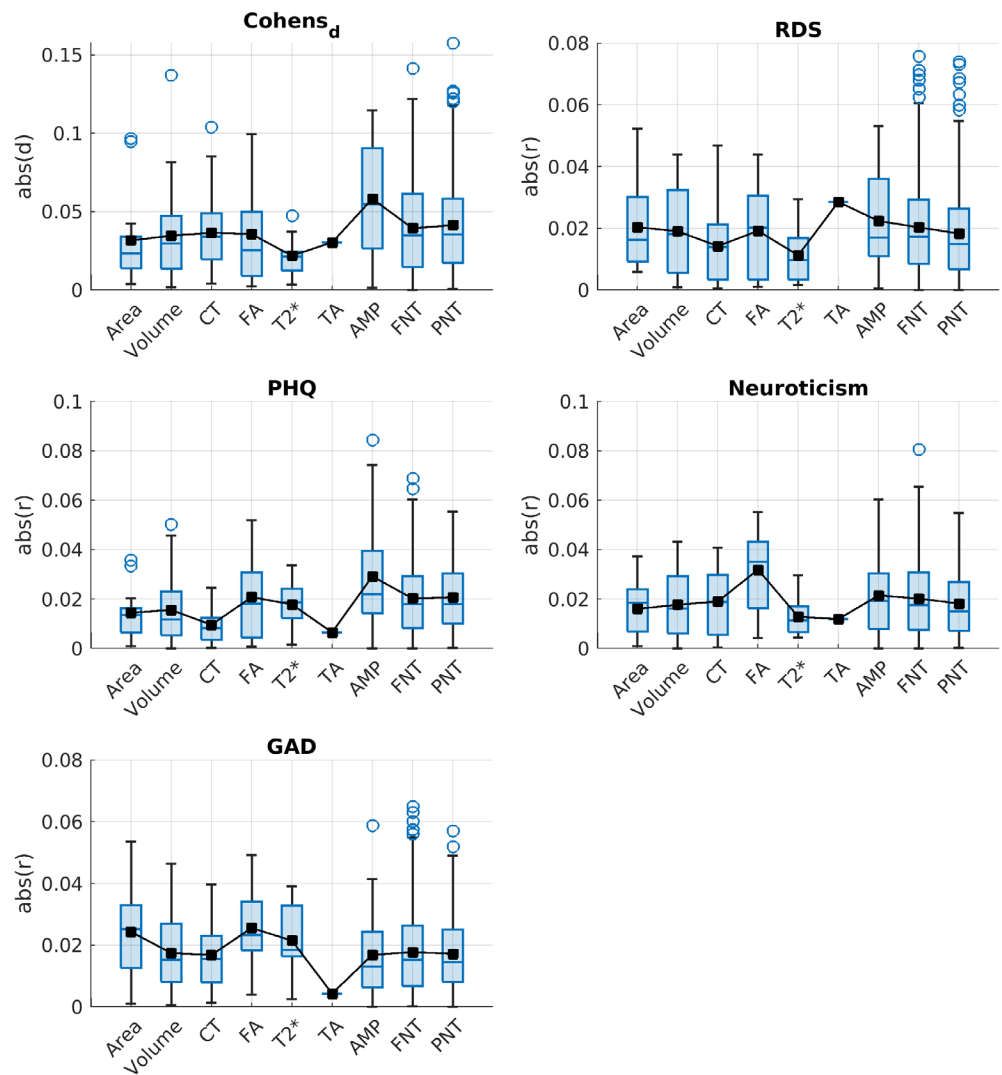
FIGURE 7 Test–retest analyses. (a) The histogram shows the inter-scan interval distribution for the 624 subjects included in these analyses. The x-axis shows days between scans, and the y-axis shows the number of subjects. (b) The boxplots show the ICCs obtained using brain IDPs after standard confound regression (blue) versus ICCs obtained using brain IDPs after standard confound regression plus regressing out effects of inter-scan interval length (orange). IDP measurement modality categories are organized along the x-axis, and the y-axis shows ICC values (see also Figure S8)

Figure 7b depicts the distributions of ICCs for each IDP measurement modality obtained using the confound-regressed data from both scan time-points, along with those obtained after additionally regressing out the effects of inter-scan interval length (i.e., days between scans). Notably, ICC distributions were highly similar for both analyses. In general, IDPs corresponding to measures of brain structure had higher ICCs than IDPs corresponding to measures of brain function. The highest ICCs were observed for IDPs corresponding to brain volume/brain area measures and the lowest ICCs were observed for IDPs corresponding to task measures. This pattern of results is not particularly surprising since macro-scale structural properties like regional volume are expected to be relatively stable over time, especially when considering relative between-subject correlations. Macro-scale functional properties like task activation magnitudes or network connectivity patterns exhibit higher variability over time due to influences of factors such as the level of task engagement (during task), cognitive state (during rest), and physiological state (e.g., hungry vs. sated, sleepy vs. alert), and therefore are expected to have somewhat reduced test–retest stability.

Analyses performed for sub-groups of patients that did ($n = 288$) versus did not ($n = 336$) exhibit changes in mental health between time-points as determined by the difference between RDS-4 measures obtained at each time point yielded highly similar results, as did those obtained after regressing out the change in RDS-4 score (See Supplementary Material Section S5 and Figure S8). Overall, these results suggest that the test–retest reliability of the IDPs is largely independent of mental health change as indicated by the RDS-4.

4 | DISCUSSION

In the present study, we aimed to tabulate mental health questionnaires available in the UKB and investigate their neural correlates. We summarize five different UKB measures of mental health: PHQ-9, GAD-7, RDS-4, N-12, and probable depression status. Our results show that all measures were moderately correlated with one another (Figure 3). CCA analyses to identify multivariate associations between these mental health measures and IDPs indicated a significant CCA mode of covariation which linked brain IDPs to mental health scores (Figure 5). The multivariate CCA analysis indicated a significant correlation between mental health and imaging that was largely reproducible in the independent confirmatory sample. All mental health measures contributed to the CCA result indicating a "trait-like" multivariate brain–mental health association. In a separate test of univariate effect sizes, modalities with the strongest modality-mean effect sizes included amplitude and edge connectivity of resting-state networks, but univariate effect sizes were generally very low (Figure 6). All IDPs showed moderate to high test–retest reliability, with IDPs of brain structure showing higher reliability than IDPs of brain function (Figure 7). Together, these findings provide the foundation for future biomarkers research into mental health using the UKB.

We highlighted a difference in acquisition timing of mental health questionnaires in the UKB study relative to neuroimaging data

acquisition. Two well-validated measures of mental health (GAD-7 and PHQ-9) were obtained as part of the online questionnaire, which is acquired independently of scan days such that they were obtained 742 days apart (median across exploratory subjects) from scan 1 (range −1,185 to +964 days). Because of this time discrepancy (which is highly inconsistent across subjects), the PHQ-9 (which tests recent depressive symptoms over a 2-week period) is not well-suited as a state depression measure for UKB neuroimaging research despite its validity for lifetime depression (Cannon et al., 2007), and its sensitivity to depression in older populations (Levis, Benedetti, Thombs, & DEPRESsion Screening Data (DEPRESSD) Collaboration, 2019). Therefore, we introduced the RDS-4 (obtained on each day of scanning) as a new UKB measure of recently experienced depressive symptoms. We propose the RDS-4 as a more appropriate measure for any UKB neuroimaging research that aims to study acute (state) depression severity or track symptom fluctuations over time. Our results from the independent Mechanical Turk study show that the correlation between the RDS-4 and the PHQ-9 is high when obtained concurrently (0.9, Table 4), whereas a lower "trait-level" correlation between RDS-4 and PHQ-9 is observed in the UKB data (0.6; Figure 3a) due to the gap in acquisition times (Figure S1). Furthermore, RDS-4 has high internal consistency and its scores map closely onto established measures of depression (Figure 4 and Table 4)— further confirming its validity. The RDS-4 questions cover four different depression domains (mood, disinterest, restlessness, and tiredness) that are also considered in other measures such as the Hamilton and Montgomery–Åsberg scales (Hamilton, 1967; Montgomery & Asberg, 1979). Hence, by asking questions in different domains, the RDS-4 inventory reflects overall depression severity relatively well, despite the comparatively small number of items. The Neuroticism-12 index—also obtained on each day of scanning—is a personality trait (Eysenck & Eysenck, 1975) that is strongly related to an increased risk in depression (Hirschfeld et al., 1983; Shaw & Hare, 1969). N-12 items assess generic traits as opposed to recently experienced clinical symptoms (RDS-4 and PHQ-9). Our results confirm that N-12 is more stable over time compared with RDS-4 and PHQ-9 as assessed by the 2-year correlation. We, therefore, suggest that N-12 can be used as a measure of trait-level susceptibility to depression in UKB neuroimaging research.

In terms of neuroimaging correlates of mental health, our findings show that multivariate associations explain more variance in mental health effects than univariate associations, which is supported by previous work (Marek et al., 2020). It should be noted that our estimated effect sizes are derived from a large sample ($N > 2,000$) and are therefore expected to capture true effect sizes that are uninfluenced by sampling variability (Marek et al., 2020). The literature to date is dominated by underpowered studies which, by design, only report high effect sizes because the significance threshold is itself high due to limited power. We have to adjust our expectations to value realistic effect sizes from well-powered samples, which may be lower but, importantly, reproducible. The observed increase in explained variance when using multivariate methods is consistent with the proposal of complex macroscopic patterns of psychopathology in mental health

patients (Williams, 2016; Wise et al., 2017). Future biomarker research will therefore need to focus on multivariate techniques such as CCA, connectome fingerprinting (Finn et al., 2015), topological network properties (Zhu et al., 2017), or machine learning (Dinga et al., 2018).

One reason why multivariate methods may have higher effect sizes than univariate methods could be due to the relatively low signal-to-noise ratio and high measurement noise of individual univariate IDPs and the effective averaging that occurs in multivariate combinations of IDPs and during the dimensionality reduction before CCA, which reduces noise. For example, previous work showed substantial increases in heritability when combining connectivity IDPs with independent component analysis compared with univariate IDPs (Elliott et al., 2018). Given the low SNR of individual IDPs and the risk of overfitting in multivariate methods, robust cross-validation (Poldrack, Huckins, & Varoquaux, 2020) and independent replication of findings (in a split-half group and/or in a fully independently acquired dataset) are essential requirements for future biomarker research (Dinga et al., 2019; Dinga, Schmaal, & Marquand, 2020).

A second potential reason for limited effect sizes (even with the use of multivariate methods like CCA) is between-subject heterogeneity. One type of heterogeneity is diversity in symptoms, such that two patients with depression may present with largely nonoverlapping symptom profiles (Drysdale et al., 2017; Feczko et al., 2019; Feczko & Fair, 2020; Kaczkurkin et al., 2020). Another type of heterogeneity is diversity in psychophysiological disease mechanisms. Here, it is possible that the same symptom may be caused by a number of different patterns of brain changes (Feczko & Fair, 2020), which we refer to as "many-to-one mechanistic mapping". Notably, both types of heterogeneity are potentially more prominent in large-scale population studies such as the UKB compared with smaller studies. This is because studies with smaller samples often implement stricter exclusion criteria in relation to comorbidities and medication to control for known sources of heterogeneity. Reducing the exclusion criteria in the UKB is likely advantageous for mental health research because the UKB and other large-scale studies provide a more accurate representation of "real-life" mental health as it occurs across the population. This makes the findings more likely to be generalizable. However, gaining a better understanding of both symptom heterogeneity and many-to-one mechanistic heterogeneity is critically important for the effective clinical translation of mental health biomarkers. Computational methods are available to account for heterogeneity, such as subtyping analyses to reveal any distinct sub-groups (Drysdale et al., 2017; Kaczkurkin et al., 2020) and normative modeling analysis to compare each individual against the normative range (Marquand, Rezek, Buitelaar, & Beckmann, 2016). These models of heterogeneity benefit from the large sample size available in the UKB which enables stringent cross-validation.

In summary, this article provides a guide for future neuroimaging biomarker research into effect-based mental health in the UKB. We recommend using RDS-4 for imaging-based research into state depression (i.e., currently experienced symptoms) and N-12 for imaging-based research into personality traits associated with depression (Lahey, 2009). Our results regarding the brain correlates of mental health show low effect sizes of individual IDPs, but higher effect sizes and replicability of multivariate associations and relatively high test–retest reliability. Therefore, we recommend the use of approaches that capture multivariate patterns and parse patient heterogeneity in combination with stringent out-of-sample replication to avoid overfitting.

DATA AVAILABILITY STATEMENT

All analysis code for this article is available at: https://github.com/PersonomicsLab/MH_in_UKB. UKB data (Miller et al., 2016; Sudlow et al., 2015) are available following an access application process, for more information please see: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. In accordance with the UKB regulations, newly derived variables in this article (e.g., RDS-4) will be made available to other researchers via UKB data access post-publication.

ORCID

Rosie K Dutt https://orcid.org/0000-0002-9887-9513
Ty O Easley https://orcid.org/0000-0002-8215-2598
Joseph C Griffis https://orcid.org/0000-0002-5548-1169

REFERENCES

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L. R., Bastiani, M., Miller, K. L., ... Smith, S. M. (2021). Confound modelling in UK Biobank brain imaging. *NeuroImage*, *224*, 117002.

Andreescu, C., Gross, J. J., Lenze, E., Edelman, K. D., Snyder, S., Tanase, C., & Aizenstein, H. (2011). Altered cerebral blood flow patterns associated with pathologic worry in the elderly. *Depression and Anxiety*, *28*, 202–209. https://doi.org/10.1002/da.20799

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., ... WU-Minn HCP Consortium. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Bijsterbosch, J. D., Woolrich, M. W., Glasser, M. F., Robinson, E. C., Beckmann, C. F., Van Essen, D. C., ... Smith, S. M. (2018). The relationship between spatial configuration and functional connectivity of brain regions. *eLife*, *7*, e32992. https://doi.org/10.7554/eLife.32992

Bluhm, R., Williamson, P., Lanius, R., Théberge, J., Densmore, M., Bartha, R., ... Osuch, E. (2009). Resting state default-mode network connectivity in early depression using a seed region-of-interest analysis: Decreased connectivity with caudate nucleus. *Psychiatry and Clinical Neurosciences*, *63*, 754–761. https://doi.org/10.1111/j.1440-1819.2009.02030.x

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*, 365–376. https://doi.org/10.1038/nrn3475

Cannon, D. S., Tiffany, S. T., Coon, H., Scholand, M. B., McMahon, W. M., & Leppert, M. F. (2007). The PHQ-9 as a brief assessment of lifetime major depression. *Psychological Assessment, 19*, 247–251. https://doi.org/10.1037/1040-3590.19.2.247

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., ... Imaging Acquisition Workgroup, A. B. C. D. (2018). The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience, 32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Cha, J., Greenberg, T., Carlson, J. M., Dedora, D. J., Hajcak, G., & Mujica-Parodi, L. R. (2014). Circuit-wide structural and functional measures predict ventromedial prefrontal cortex fear generalization: Implications for generalized anxiety disorder. *The Journal of Neuroscience, 34*, 4043–4053. https://doi.org/10.1523/JNEUROSCI.3372-13.2014

Davis, K. A. S., Coleman, J. R. I., Adams, M., Allen, N., Breen, G., Cullen, B., ... Hotopf, M. (2020). Mental health in UK Biobank—Development, implementation and results from an online questionnaire completed by 157 366 participants: A reanalysis. *BJPsych Open, 6*, e18. https://doi.org/10.1192/bjo.2019.100

Dinga, R., Marquand, A. F., Veltman, D. J., Beekman, A. T. F., Schoevers, R. A., van Hemert, A. M., ... Schmaal, L. (2018). Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: A machine learning approach. *Translational Psychiatry, 8*, 241. https://doi.org/10.1038/s41398-018-0289-1

Dinga, R., Schmaal, L., & Marquand, A. F. (2020). A closer look at depression biotypes: Correspondence relating to Grosenick et al. (2019). *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 5*(5), 554–555. https://doi.org/10.1016/j.bpsc.2019.09.011

Dinga, R., Schmaal, L., Penninx, B. W. J. H., van Tol, M. J., Veltman, D. J., van Velzen, L., ... Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage: Clinical, 22*, 101796.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., ... Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine, 23*, 28–38. https://doi.org/10.1038/nm.4246

Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *Lancet, 350*, 326–329. https://doi.org/10.1016/S0140-6736(97)02419-7

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., ... Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature, 562*, 210–216. https://doi.org/10.1038/s41586-018-0571-7

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Eysenck personality questionnaire manual*. San Diego, CA: Educational and Industrial Testing Service.

Feczko, E., & Fair, D. A. (2020). Methods and challenges for assessing heterogeneity. *Biological Psychiatry, 88*, 9–17.

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The heterogeneity problem: Approaches to identify psychiatric subtypes. *Trends in Cognitive Sciences, 23*, 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., ... Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience, 18*, 1664–1671. https://doi.org/10.1038/nn.4135

Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M., & Kennedy, K. M. (2021). Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Human Brain Mapping, 42*, 204–219. https://doi.org/10.1002/hbm.25217

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., ... Schatzberg, A. F. (2007). Resting-state functional connectivity in major depression: Abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry, 62*, 429–437. https://doi.org/10.1016/j.biopsych.2006.09.020

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., ... Jenkinson, M. (2016). BIANCA (brain intensity AbNormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage, 141*, 191–205. https://doi.org/10.1016/j.neuroimage.2016.07.018

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *The British Journal of Social and Clinical Psychology, 6*, 278–296. https://doi.org/10.1111/j.2044-8260.1967.tb00530.x

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage, 17*, 317–323.

Harms, M. P., Somerville, L. H., Ances, B. M., Andersson, J., Barch, D. M., Bastiani, M., ... Yacoub, E. (2018). Extending the Human Connectome Project across ages: Imaging protocols for the lifespan development and aging projects. *NeuroImage, 183*, 972–984. https://doi.org/10.1016/j.neuroimage.2018.09.060

He, Y., Xu, T., Zhang, W., & Zuo, X.-N. (2016). Lifespan anxiety is reflected in human amygdala cortical connectivity. *Human Brain Mapping, 37*, 1178–1193. https://doi.org/10.1002/hbm.23094

Helmer, M., Warrington, S., Mohammadi-Nejad, A.-R., Ji, J. L., Howell, A., Rosand, B., ... Murray, J. D. (2021). On stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *bioRxiv*. Retrieved from. https://www.biorxiv.org/content/10.1101/2020.08.25.265546v2

Hirschfeld, R. M., Klerman, G. L., Clayton, P. J., Keller, M. B., McDonald-Scott, P., & Larkin, B. H. (1983). Assessing personality: Effects of the depressive state on trait measurement. *The American Journal of Psychiatry, 140*, 695–699. https://doi.org/10.1176/ajp.140.6.695

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*, 321–377.

Hutton, J. L., & Williamson, P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society. Series C, Applied Statistics, 49*, 359–370. https://doi.org/10.1111/1467-9876.00197

Kaczkurkin, A. N., Sotiras, A., Baller, E. B., Barzilay, R., Calkins, M. E., Chand, G. B., ... Satterthwaite, T. D. (2020). Neurostructural heterogeneity in youths with internalizing symptoms. *Biological Psychiatry, 87*, 473–482. https://doi.org/10.1016/j.biopsych.2019.09.005

Kaiser, R. H., Andrews-Hanna, J. R., Wager, T. D., & Pizzagalli, D. A. (2015). Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity. *JAMA Psychiatry, 72*, 603–611. https://doi.org/10.1001/jamapsychiatry.2015.0071

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ, 340*, c365. https://doi.org/10.1136/bmj.c365

Klauser, P., Fornito, A., Lorenzetti, V., Davey, C. G., Dwyer, D. B., Allen, N. B., & Yücel, M. (2015). Cortico-limbic network abnormalities in individuals with current and past major depressive disorder. *Journal of Affective Disorders, 173*, 45–52. https://doi.org/10.1016/j.jad.2014.10.041

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lahey, B. B. (2009). Public health significance of neuroticism. *The American Psychologist, 64*, 241–256. https://doi.org/10.1037/a0015309

Levis, B., Benedetti, A., Thombs, B. D., & DEPRESsion Screening Data (DEPRESSD) Collaboration. (2019). Accuracy of patient health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ*, *365*, l1476. https://doi.org/10.1136/bmj.l1476

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z

Ma, C., Ding, J., Li, J., Guo, W., Long, Z., Liu, F., ... Chen, H. (2012). Resting-state functional connectivity bias of middle temporal gyrus and caudate with altered gray matter volume in major depression. *PLoS One*, *7*, e45263. https://doi.org/10.1371/journal.pone.0045263

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., ..., Dosenbach, N. U. F. (2020): Towards reproducible brain-wide association studies. Retrieved from https://www.biorxiv.org/content/10.1101/2020.08.21.257758v1?s=03.

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. https://doi.org/10.1016/j.biopsych.2015.12.023

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., ... Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*, 1523–1536. https://doi.org/10.1038/nn.4393

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, *134*, 382–389. https://doi.org/10.1192/bjp.134.4.382

Müller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., ... Eickhoff, S. B. (2018). Ten simple rules for neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *84*, 151–161. https://doi.org/10.1016/j.neubiorev.2017.11.012

Nickerson, L. D., Smith, S. M., Öngür, D., & Beckmann, C. F. (2017). Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Frontiers in Neuroscience*, *11*, 115. https://doi.org/10.3389/fnins.2017.00115

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

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*, *77*, 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401.

Salo, K. I., Scharfen, J., Wilden, I. D., Schubotz, R. I., & Holling, H. (2019). Confining the concept of vascular depression to late-onset depression: A meta-analysis of MRI-defined Hyperintensity burden in major depressive disorder and bipolar disorder. *Frontiers in Psychology*, *10*, 1241. https://doi.org/10.3389/fpsyg.2019.01241

Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., ... 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*, 900–909. https://doi.org/10.1038/mp.2016.60

Shaw, G. K., & Hare, E. H. (1969). Eysenck personality inventory scores of patients with depressive illness. *The British Journal of Psychiatry*, *115*, 253–255. https://doi.org/10.1192/bjp.115.519.253

Sheline, Y. I., Price, J. L., Yan, Z., & Mintun, M. A. (2010). Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 11020–11025. https://doi.org/10.1073/pnas.1000446107

Smith, D. J., Nicholl, B. I., Cullen, B., Martin, D., Ul-Haq, Z., Evans, J., ... Pell, J. P. (2013). Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: Cross-sectional study of 172,751 participants. *PLoS One*, *8*, e75362. https://doi.org/10.1371/journal.pone.0075362

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E. J., Glasser, M. F., ... Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, *18*, 1565–1567. https://doi.org/10.1038/nn.4125

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, *323*, 101–105. https://doi.org/10.1136/bmj.323.7304.101

Stratmann, M., Konrad, C., Kugel, H., Krug, A., Schöning, S., Ohrmann, P., ... Dannlowski, U. (2014). Insular and hippocampal gray matter volume reductions in patients with major depressive disorder. *PLoS One*, *9*, e102692. https://doi.org/10.1371/journal.pone.0102692

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ... Collins, R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*, 207–216. https://doi.org/10.1016/s0895-4356(99)00161-4

Tozzi, L., Staveland, B., Holt-Gosselin, B., Chesnut, M., Chang, S. E., Choi, D., ... Williams, L. M. (2020). The human connectome project for disordered emotional states: Protocol and rationale for a research domain criteria study of brain connectivity in young adult anxiety and depression. *NeuroImage*, *214*, 116715. https://doi.org/10.1016/j.neuroimage.2020.116715

Tozzi, L., Zhang, X., Chesnut, M., Holt-Gosselin, B., Ramirez, C. A., & Williams, L. M. (2021). Reduced functional connectivity of default mode network subsystems in depression: Meta-analytic evidence and relationship with trait rumination. *Neuroimage Clin*, *30*, 102570. https://doi.org/10.1016/j.nicl.2021.102570

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: Current and future directions. *Social Cognitive and Affective Neuroscience*, *2*, 150–158. https://doi.org/10.1093/scan/nsm015

Wardenaar, K. J., van Veen, T., Giltay, E. J., de Beurs, E., Penninx, B. W. J. H., & Zitman, F. G. (2010). Development and validation of a 30-item short adaptation of the mood and anxiety symptoms questionnaire (MASQ). *Psychiatry Research*, *179*, 101–106. https://doi.org/10.1016/j.psychres.2009.03.005

Watson, D., & Clark, L. A. (1991). *The mood and anxiety symptom questionnaire*. Madison, WI: Addiction Research Center.

Williams, L. M. (2016). Precision psychiatry: A neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry*, *3*, 472–480. https://doi.org/10.1016/S2215-0366(15)00579-9

Wise, T., Marwood, L., Perkins, A. M., Herane-Vives, A., Joules, R., Lythgoe, D. J., ... Arnone, D. (2017). Instability of default mode network connectivity in major depression: A two-sample confirmation study. *Translational Psychiatry*, *7*, e1105. https://doi.org/10.1038/tp.2017.40

Wise, T., Radua, J., Via, E., Cardoner, N., Abe, O., Adams, T. M., ... Arnone, D. (2017). Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: Evidence

from voxel-based meta-analysis. *Molecular Psychiatry*, 22, 1455–1463. https://doi.org/10.1038/mp.2016.72

Wolf, R. C., Nolte, H. M., Hirjak, D., Hofer, S., Seidl, U., Depping, M. S., … Thomann, P. A. (2016). Structural network changes in patients with major depression and schizophrenia treated with electroconvulsive therapy. *European Neuropsychopharmacology*, 26, 1465–1474. https://doi.org/10.1016/j.euroneuro.2016.06.008

Xu, J., Van Dam, N. T., Feng, C., Luo, Y., Ai, H., Gu, R., & Xu, P. (2019). Anxious brain networks: A coordinate-based activation likelihood estimation meta-analysis of resting-state functional connectivity studies in anxiety. *Neuroscience and Biobehavioral Reviews*, 96, 21–30. https://doi.org/10.1016/j.neubiorev.2018.11.005

Yan, C.-G., Chen, X., Li, L., Castellanos, F. X., Bai, T.-J., Bo, Q.-J., … Zang, Y.-F. (2019). Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 9078–9083. https://doi.org/10.1073/pnas.1900390116

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power-commentary on Vul et al. (2009). *Perspectives on Psychological Science*, 4, 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x

Yu, M., Linn, K. A., Shinohara, R. T., Oathes, D. J., Cook, P. A., Duprat, R., … Sheline, Y. I. (2019). Childhood trauma history is linked to abnormal brain connectivity in major depression. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 8582–8590. https://doi.org/10.1073/pnas.1900801116

Zhu, H., Qiu, C., Meng, Y., Yuan, M., Zhang, Y., Ren, Z., … Zhang, W. (2017). Altered topological properties of brain networks in social anxiety disorder: A resting-state functional MRI study. *Scientific Reports*, 7, 43089. https://doi.org/10.1038/srep43089

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.