medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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

Running title: Mental health in the UK Biobank

Rosie Dutt*[1], Kayla Hannon*[1], Ty Easley[1], Joseph Griffis[1], Wei Zhang[1], Janine Bijsterbosch[1]

*Joint first authors

1. Department of Radiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA.

Corresponding author: Janine Bijsterbosch, janine.bijsterbosch@wustl.edu

Acknowledgements: We are grateful to UK Biobank and the UK Biobank participants for making the resource data possible, and to the data processing team at Oxford University for producing the shared processed data. This research was performed under UK Biobank application number 47267. This research was supported by the NIH (1 R34 NS118618-01) and the McDonnell Center for Systems Neuroscience.

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: Depression; Mental Health; Brain Correlates; Test-Retest; Replication; UK Biobank

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

1. Introduction

Over the years there have 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 et al., 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 et al., 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 et al., 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 fronto-parietal and salience networks and hyper-connectivity in the DMN [Kaiser et al., 2015]. Another meta-analysis showed that there are common grey-matter volume changes in MDD which are also seen in bipolar disorder [Wise et al., 2017b]. 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]. Whilst these meta-analyses aid to establish consensus on brain correlates of mental health [Wager et al., 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 and 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 and Williamson, 2000; Kirkham et al., 2010], which can lead to inflated meta analytical results [Sterne et al., 2001]. These inherent limitations of meta-analyses may explain why disagreement persists within even meta-analytical work, with a recent study showing hypo- (rather than hyper-) connectivity in the core DMN in patients with depression [Tozzi et al., 2021].

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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 UK Biobank 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 UK Biobank. The first are 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 are 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. Firstly, 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. Secondly, 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 UK Biobank study. The RDS-4 score therefore enables research into current depressive symptoms and changes in symptomatology over time. Thirdly, we aim to establish realistic and robust univariate and multivariate effect sizes of commonly reported brain

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

correlates of mental health based on population data. Lastly, 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 our hope is to provide a foundation that future studies can build on.

2. Methods

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 (Fig. 1; Table 1). Subjects with a mean head motion greater than 0.2mm 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 et al., 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 minimal difference in head motion). Subsequently, approximately two-thirds of the 'never seen GP' subjects together with their matched 'seen GP' subjects was randomly assigned to the exploratory sample, and the remaining subjects were assigned to the confirmatory sample. During subject assignment to groups, we preserved the matched characteristics within each resulting sample (Fig. 1). No subjects overlapped between the exploratory and confirmatory samples.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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 5 different UKB mental health measures, which will be used for neuroimaging and questionnaire comparison analyses and Fig. 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 and colleagues used questions from the Assessment information to develop a categorical (case-control) measure of depression [Smith et al., 2013]. 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: not at all (1), several days (2), more than half the days (3), and nearly every day (4). 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 from 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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 3T scanner. For T1 structural scans, 3D MPRAGE acquisition was used to acquire 1mm 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: 0.8x0.8x3mm), obtaining two echo times (TE=9.4 and TE=20 ms). Diffusion data was acquired with b-values of 1000 and 2000 s/mm2, at 2mm 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= 2.4mm, TR= 0.735s, factor = 8 multiband accelerator). Task fMRI used the Hariri faces/shapes “emotion” task as employed in the HCP [Barch et al., 2013; Hariri et al., 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].

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 et al., 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 complimentary 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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].

Correlations amongst 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).

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 Fig. 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.

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 and Brennan, 2014]. We further calculated the Cronbach alpha for the newly proposed RDS-4 score to measure internal consistency in the exploratory sample.

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 et al., 2017]. Participants were paid a nominal compensation for questionnaire completion. 134 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Participants completed the same set of mental health questionnaires at two timepoints 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 and 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. Firstly, we adopted premium options in CloudResearch, such as only including 'CloudResearch approved participants' who undergo more extensive vetting. Secondly, 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. Thirdly, we imposed a minimum duration for questionnaire completion at 172.5 seconds (which equals 2.5 seconds per question). Participants who completed the questionnaire in less than 172.5 seconds 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 and Li, 2016; McGraw and Wong, 1996]) was used to calculate the test-retest reliability between time point 1 and time point 2 separately for each measure.

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) was 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 5 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

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

subjects in the exploratory sample. Permutation testing with 2000 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 are the IDP principal component inputs and Y are 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 $0.05/(3{,}466+346+16)=1.3*10^{-5}$, 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.

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

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

running the aforementioned analyses. The only exception from de-confounding is the binary grouping based on probable depression status, as deconfounding would result in subject-specific values that are non-categorical, which is unsuitable for the Cohen's d test.

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 and 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 time-points. Of the *N*=624 subjects included in the test-retest analyses, *n*=336 exhibited no change in RDS-4 scores between time-points, 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

Correlations amongst mental health variables in the UKB

Mental health measures showed moderate correlations with one another, indicating redundancy between these metrics (Fig. 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 ρ=0.57±0.01 ($p \approx 10^{-199}$); PHQ-9 and GAD-7, which were both taken from the online questionnaire, have SRCC ρ=0.69±0.01 ($p \approx 10^{-267}$). 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 two-year 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-

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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 ρ = 0.57 between initial and follow-up scans, and N-12 showed a 2-year correlation of ρ = 0.85. It should be noted that this reflects 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 'Mechanical Turk study to validate RDS-4' section).

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 $\chi \geq 0.19, p \leq 10^{-48}$; Fig. 2B).

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 (Fig. 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 (Fig. 4E). Additionally, our results show stable mapping between RDS-4 and N-12 (Fig. 4F), and between N-12 and GAD-7 (Fig. 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 0.78, which indicates a moderate to strong internal reliability. This was similar to N-12 (Cronbach alpha = 0.83).

Mechanical Turk study to validate RDS-4

Out of 134 subjects who completed our separate validation study, 3 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).

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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 2 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 5 mental health variables. The CCA resulted in two significant canonical covariates ($R1_{UV}$ = 0.207, p = 0.0005 & $R2_{UV}$ = 0.174, p = 0.015). The first multivariate canonical correlation partly replicated in the independent confirmatory sample ($R1_{UV(confirmatory)}$ = 0.125, p = $3.7*10^{-9}$, where the p-value was Bonferroni corrected for the maximum of 5 canonical correlations). Although the second canonical correlation also reached significance in the confirmatory sample ($R2_{UV(confirmatory)}$ = 0.06, $p_{Bonferroni}$ = 0.02) 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. Firstly, the CCA was relatively well-powered with 50.7 subjects per input variable leading to relatively stable estimates [Helmer et al., 2021]. Secondly, the exploratory and confirmatory samples were well matched in terms of sample characteristics. Thirdly, data reduction of IDPs prior to CCA likely reduces measurement noise. Post-hoc correlations between the averaged UV subject scores and the mental health variables and IDPs also replicated well (Fig. 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.

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 (supplementary table 1) that partly overlaps with the data-driven IDP identification.

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, was larger than the Pearson's r values from correlations between IDPs and continuous mental health measures (Fig. 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

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

at top (Figs. S3-S7). These findings together suggest an overall higher effect size of resting-state in contrast to non-resting state measures on the investigated mental health variables.

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. Fig. 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 - 1695 measures), and partial correlation-based resting-state networks (PNT - 1695 measures).

Fig. 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) vs. 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 5 and Fig. 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

4. Discussion

In the present study we aimed to tabulate mental health questionnaires available in the UK Biobank 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 (Fig. 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 (Fig. 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 (Fig. 6). All IDPs showed moderate to high test-retest reliability, with IDPs of brain structure showing higher reliability than IDPs of brain function (Fig. 7). Together, these findings provide the foundation for future biomarkers research into mental health using the UK Biobank.

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 et al., 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; Fig. 3A) due to the gap in acquisition times (Fig. S1). Furthermore, RDS-4 has high internal consistency and its scores map closely onto established measures of depression (Fig. 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 and 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 and Eysenck, 1975] that is strongly related to an increased risk in depression [Hirschfeld et al., 1983; Shaw and Hare, 1969]. N-12 items assess generic traits as opposed to recently experienced clinical symptoms (RDS-4 and PHQ-9). Our results confirm

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

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>2000) 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., 2017a]. Future biomarker research will therefore need to focus on multivariate techniques such as canonical correlation analysis, 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 prior to 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 et al., 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 et al., 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 non-overlapping symptom profiles [Drysdale et al., 2017; Feczko et al., 2019; Feczko and 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 and 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 UK Biobank 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

However, gaining a better understanding of both symptom heterogeneity and many-to-one mechanistic heterogeneity is critically important for 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 modelling analysis to compare each individual against the normative range [Marquand et al., 2016]. These models of heterogeneity benefit from the large sample size available in the UK Biobank which enables stringent cross-validation.

In summary, this paper provides a guide for future neuroimaging biomarker research into affect-based mental health in the UK Biobank. 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.

5. Code and Data Availability

All analysis code for this article is available at: https://github.com/PersonomicsLab/MH_in_UKB. UK Biobank 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.

6. References

Alfaro-Almagro F, McCarthy P, Afyouni S, Andersson JLR, Bastiani M, Miller KL, Nichols TE, Smith SM (2021): Confound modelling in UK Biobank brain imaging. Neuroimage 224:117002. http://www.sciencedirect.com/science/article/pii/S1053811920304882.

Andreescu C, Gross JJ, Lenze E, Edelman KD, Snyder S, Tanase C, Aizenstein H (2011): Altered cerebral blood flow patterns associated with pathologic worry in the elderly. Depress Anxiety 28:202–209. http://dx.doi.org/10.1002/da.20799.

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, Nolan D, Bryant E, Hartley T, Footer O, Bjork JM, Poldrack R, Smith S, Johansen-Berg H, Snyder AZ, Van Essen DC, WU-Minn HCP Consortium (2013): Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80:169–189. http://dx.doi.org/10.1016/j.neuroimage.2013.05.033.

Bijsterbosch JD, Woolrich MW, Glasser MF, Robinson EC, Beckmann CF, Van Essen DC, Harrison SJ, Smith SM (2018): The relationship between spatial configuration and functional connectivity of brain regions. Elife 7. http://dx.doi.org/10.7554/eLife.32992.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Bluhm R, Williamson P, Lanius R, Théberge J, Densmore M, Bartha R, Neufeld 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 Clin Neurosci 63:754–761. http://dx.doi.org/10.1111/j.1440-1819.2009.02030.x.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013): Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376. http://dx.doi.org/10.1038/nrn3475.

Cannon DS, Tiffany ST, Coon H, Scholand MB, McMahon WM, Leppert MF (2007): The PHQ-9 as a brief assessment of lifetime major depression. Psychol Assess 19:247–251. http://dx.doi.org/10.1037/1040-3590.19.2.247.

Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, Orr CA, Wager TD, Banich MT, Speer NK, Sutherland MT, Riedel MC, Dick AS, Bjork JM, Thomas KM, Chaarani B, Mejia MH, Hagler DJ Jr, Daniela Cornejo M, Sicat CS, Harms MP, Dosenbach NUF, Rosenberg M, Earl E, Bartsch H, Watts R, Polimeni JR, Kuperman JM, Fair DA, Dale AM, ABCD Imaging Acquisition Workgroup (2018): The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. Dev Cogn Neurosci 32:43–54. http://dx.doi.org/10.1016/j.dcn.2018.03.001.

Cha J, Greenberg T, Carlson JM, Dedora DJ, Hajcak G, Mujica-Parodi LR (2014): Circuit-wide structural and functional measures predict ventromedial prefrontal cortex fear generalization: implications for generalized anxiety disorder. J Neurosci 34:4043–4053. http://dx.doi.org/10.1523/JNEUROSCI.3372-13.2014.

Davis KAS, Coleman JRI, Adams M, Allen N, Breen G, Cullen B, Dickens C, Fox E, Graham N, Holliday J, Howard LM, John A, Lee W, McCabe R, McIntosh A, Pearsall R, Smith DJ, Sudlow C, Ward J, Zammit S, 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. http://dx.doi.org/10.1192/bjo.2019.100.

Dinga R, Marquand AF, Veltman DJ, Beekman ATF, Schoevers RA, van Hemert AM, Penninx BWJH, Schmaal L (2018): Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Transl Psychiatry 8:241. http://dx.doi.org/10.1038/s41398-018-0289-1.

Dinga R, Schmaal L, Marquand AF (2020): A Closer Look at Depression Biotypes: Correspondence Relating to Grosenick et al. (2019). Biological psychiatry. Cognitive neuroscience and neuroimaging. http://dx.doi.org/10.1016/j.bpsc.2019.09.011.

Dinga R, Schmaal L, Penninx BWJH, van Tol MJ, Veltman DJ, van Velzen L, Mennes M, van der Wee NJA, Marquand AF (2019): Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). NeuroImage: Clinical 22:101796. http://www.sciencedirect.com/science/article/pii/S2213158219301469.

Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, Fetcho RN, Zebley B, Oathes DJ, Etkin A, Schatzberg AF, Sudheimer K, Keller J, Mayberg HS, Gunning FM, Alexopoulos GS, Fox MD, Pascual-Leone A, Voss HU, Casey BJ, Dubin MJ, Liston C (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28–38. http://dx.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. http://dx.doi.org/10.1016/S0140-6736(97)02419-7.

Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, Marchini J, Smith SM (2018): Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562:210–216. http://dx.doi.org/10.1038/s41586-018-0571-7.

Eysenck HJ, Eysenck SBG (1975): Eysenck Personality Questionnaire Manual. San Diego, CA: Educational and Industrial Testing Service.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Feczko E, Fair DA (2020): Methods and Challenges for Assessing Heterogeneity. Biol Psychiatry. http://www.sciencedirect.com/science/article/pii/S0006322320301104.

Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, Fair DA (2019): The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. Trends Cogn Sci 23:584–601. http://dx.doi.org/10.1016/j.tics.2019.03.009.

Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT (2015): Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci 18:1664–1671. http://dx.doi.org/10.1038/nn.4135.

Grady CL, Rieck JR, Nichol D, Rodrigue KM, Kennedy KM (2021): Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. Hum Brain Mapp 42:204–219. http://dx.doi.org/10.1002/hbm.25217.

Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H, Reiss AL, Schatzberg AF (2007): Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. Biol Psychiatry 62:429–437. http://dx.doi.org/10.1016/j.biopsych.2006.09.020.

Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, Schulz UG, Kuker W, Battaglini M, Rothwell PM, Jenkinson M (2016): BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. Neuroimage 141:191–205. http://dx.doi.org/10.1016/j.neuroimage.2016.07.018.

Hamilton M (1967): Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 6:278–296. http://dx.doi.org/10.1111/j.2044-8260.1967.tb00530.x.

Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR (2002): The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage 17:317–323. https://www.ncbi.nlm.nih.gov/pubmed/12482086.

Harms MP, Somerville LH, Ances BM, Andersson J, Barch DM, Bastiani M, Bookheimer SY, Brown TB, Buckner RL, Burgess GC, Coalson TS, Chappell MA, Dapretto M, Douaud G, Fischl B, Glasser MF, Greve DN, Hodge C, Jamison KW, Jbabdi S, Kandala S, Li X, Mair RW, Mangia S, Marcus D, Mascali D, Moeller S, Nichols TE, Robinson EC, Salat DH, Smith SM, Sotiropoulos SN, Terpstra M, Thomas KM, Tisdall MD, Ugurbil K, van der Kouwe A, Woods RP, Zöllei L, Van Essen DC, Yacoub E (2018): Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. Neuroimage 183:972–984. http://dx.doi.org/10.1016/j.neuroimage.2018.09.060.

Helmer M, Warrington S, Mohammadi-Nejad A-R, Ji JL, Howell A, Rosand B, Anticevic A, Sotiropoulos SN, Murray JD (2021): On stability of Canonical Correlation Analysis and Partial Least Squares with application to brain-behavior associations. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.08.25.265546v2.

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

Hirschfeld RM, Klerman GL, Clayton PJ, Keller MB, McDonald-Scott P, Larkin BH (1983): Assessing personality: effects of the depressive state on trait measurement. Am J Psychiatry 140:695–699. http://dx.doi.org/10.1176/ajp.140.6.695.

Hotelling H (1936): RELATIONS BETWEEN TWO SETS OF VARIATES. Biometrika 28:321–377. https://academic.oup.com/biomet/article/28/3-4/321/220073.

Hutton JL, Williamson PR (2000): Bias in meta□analysis due to outcome variable selection within studies. J R Stat Soc Ser C Appl Stat 49:359–370. https://onlinelibrary.wiley.com/doi/10.1111/1467-9876.00197.

Kaczkurkin AN, Sotiras A, Baller EB, Barzilay R, Calkins ME, Chand GB, Cui Z, Erus G, Fan Y, Gur RE, Gur RC, Moore TM, Roalf DR, Rosen AFG, Ruparel K, Shinohara RT, Varol E, Wolf DH, Davatzikos C, Satterthwaite TD (2020): Neurostructural Heterogeneity in Youths With Internalizing Symptoms. Biol Psychiatry 87:473–482.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

http://dx.doi.org/10.1016/j.biopsych.2019.09.005.

Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA (2015): Large-Scale Network Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity. JAMA Psychiatry 72:603–611. http://dx.doi.org/10.1001/jamapsychiatry.2015.0071.

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR (2010): The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 340:c365. http://dx.doi.org/10.1136/bmj.c365.

Klauser P, Fornito A, Lorenzetti V, Davey CG, Dwyer DB, Allen NB, Yücel M (2015): Cortico-limbic network abnormalities in individuals with current and past major depressive disorder. J Affect Disord 173:45–52. http://dx.doi.org/10.1016/j.jad.2014.10.041.

Kolen MJ, Brennan RL (2014): Test Equating, Scaling, and Linking: Methods and Practices. Springer, New York, NY. https://link.springer.com/10.1007/978-1-4939-0317-7.

Koo TK, Li MY (2016): A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15:155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012.

Lahey BB (2009): Public health significance of neuroticism. Am Psychol 64:241–256. http://dx.doi.org/10.1037/a0015309.

Levis B, Benedetti A, Thombs BD, 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. http://dx.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. Behav Res Methods 49:433–442. http://dx.doi.org/10.3758/s13428-016-0727-z.

Ma C, Ding J, Li J, Guo W, Long Z, Liu F, Gao Q, Zeng L, Zhao J, 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. http://dx.doi.org/10.1371/journal.pone.0045263.

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, Conan G, Uriarte J, Snider K, Tam A, Chen J, Newbold DJ, Zheng A, Seider NA, Van AN, Laumann TO, Thompson WK, Greene DJ, Petersen SE, Nichols TE, Thomas Yeo BT, Barch DM, Garavan H, Luna B, Fair DA, Dosenbach NUF (2020): Towards Reproducible Brain-Wide Association Studies. https://www.biorxiv.org/content/10.1101/2020.08.21.257758v1?s=03.

Marquand AF, Rezek I, Buitelaar J, Beckmann CF (2016): Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biol Psychiatry 80:552–561. http://dx.doi.org/10.1016/j.biopsych.2015.12.023.

McGraw KO, Wong SP (1996): Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46. https://psycnet.apa.org/fulltext/1996-03170-003.pdf.

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016): Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523–1536. http://dx.doi.org/10.1038/nn.4393.

Montgomery SA, Asberg M (1979): A new depression scale designed to be sensitive to change. Br J Psychiatry 134:382–389. http://dx.doi.org/10.1192/bjp.134.4.382.

Müller VI, Cieslik EC, Laird AR, Fox PT, Radua J, Mataix-Cols D, Tench CR, Yarkoni T, Nichols TE, Turkeltaub PE, Wager TD, Eickhoff SB (2018): Ten simple rules for neuroimaging meta-analysis. Neurosci Biobehav Rev 84:151–161.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

http://dx.doi.org/10.1016/j.neubiorev.2017.11.012.

Nickerson LD, Smith SM, Öngür D, Beckmann CF (2017): Using Dual Regression to Investigate Network Shape and Amplitude in Functional Connectivity Analyses. Front Neurosci 11:115. http://dx.doi.org/10.3389/fnins.2017.00115.

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline J-B, Vul E, Yarkoni T (2017): Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126. http://dx.doi.org/10.1038/nrn.2016.167.

Poldrack RA, Huckins G, Varoquaux G (2020): Establishment of Best Practices for Evidence for Prediction: A Review. JAMA Psychiatry 77:534–540. http://dx.doi.org/10.1001/jamapsychiatry.2019.3671.

Radloff LS (1977): The CES-D Scale: A self-report depression scale for research in the general poplulation. Appl Psychol Meas 1:385–401.

Salo KI, Scharfen J, Wilden ID, Schubotz RI, 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. Front Psychol 10:1241. http://dx.doi.org/10.3389/fpsyg.2019.01241.

Schmaal L, Hibar DP, Sämann PG, Hall GB, Baune BT, Jahanshad N, Cheung JW, van Erp TGM, Bos D, Ikram MA, Vernooij MW, Niessen WJ, Tiemeier H, Hofman A, Wittfeld K, Grabe HJ, Janowitz D, Bülow R, Selonke M, Völzke H, Grotegerd D, Dannlowski U, Arolt V, Opel N, Heindel W, Kugel H, Hoehn D, Czisch M, Couvy-Duchesne B, Rentería ME, Strike LT, Wright MJ, Mills NT, de Zubicaray GI, McMahon KL, Medland SE, Martin NG, Gillespie NA, Goya-Maldonado R, Gruber O, Krämer B, Hatton SN, Lagopoulos J, Hickie IB, Frodl T, Carballedo A, Frey EM, van Velzen LS, Penninx BWJH, van Tol M-J, van der Wee NJ, Davey CG, Harrison BJ, Mwangi B, Cao B, Soares JC, Veer IM, Walter H, Schoepf D, Zurowski B, Konrad C, Schramm E, Normann C, Schnell K, Sacchet MD, Gotlib IH, MacQueen GM, Godlewska BR, Nickson T, McIntosh AM, Papmeyer M, Whalley HC, Hall J, Sussmann JE, Li M, Walter M, Aftanas L, Brack I, Bokhan NA, Thompson PM, Veltman DJ (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. Mol Psychiatry 22:900–909. http://dx.doi.org/10.1038/mp.2016.60.

Shaw GK, Hare EH (1969): Eysenck personality inventory scores of patients with depressive illness. Br J Psychiatry 115:253–255. http://dx.doi.org/10.1192/bjp.115.519.253.

Sheline YI, Price JL, Yan Z, Mintun MA (2010): Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. Proc Natl Acad Sci U S A 107:11020–11025. http://dx.doi.org/10.1073/pnas.1000446107.

Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans J, Gill JMR, Roberts B, Gallacher J, Mackay D, Hotopf M, Deary I, Craddock N, Pell JP (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. http://dx.doi.org/10.1371/journal.pone.0075362.

Smith SM, Nichols TE, Vidaurre D, Winkler AM, Behrens TEJ, Glasser MF, Ugurbil K, Barch DM, Van Essen DC, Miller KL (2015): A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat Neurosci 18:1565–1567. http://dx.doi.org/10.1038/nn.4125.

Sterne JA, Egger M, Smith GD (2001): Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. BMJ 323:101–105. http://dx.doi.org/10.1136/bmj.323.7304.101.

Stratmann M, Konrad C, Kugel H, Krug A, Schöning S, Ohrmann P, Uhlmann C, Postert C, Suslow T, Heindel W, Arolt V, Kircher T, Dannlowski U (2014): Insular and hippocampal gray matter volume reductions in patients with major depressive disorder. PLoS One

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

9:e102692. http://dx.doi.org/10.1371/journal.pone.0102692.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, 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 Med 12:e1001779. http://dx.doi.org/10.1371/journal.pmed.1001779.

Thornton A, Lee P (2000): Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 53:207–216. http://dx.doi.org/10.1016/s0895-4356(99)00161-4.

Tozzi L, Staveland B, Holt-Gosselin B, Chesnut M, Chang SE, Choi D, Shiner M, Wu H, Lerma-Usabiaga G, Sporns O, Barch DM, Gotlib IH, Hastie TJ, Kerr AB, Poldrack RA, Wandell BA, Wintermark M, Williams LM (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. http://dx.doi.org/10.1016/j.neuroimage.2020.116715.

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

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, WU-Minn HCP Consortium (2013): The WU-Minn Human Connectome Project: an overview. Neuroimage 80:62–79. http://dx.doi.org/10.1016/j.neuroimage.2013.05.041.

Wager TD, Lindquist M, Kaplan L (2007): Meta-analysis of functional neuroimaging data: current and future directions. Soc Cogn Affect Neurosci 2:150–158. http://dx.doi.org/10.1093/scan/nsm015.

Wardenaar KJ, van Veen T, Giltay EJ, de Beurs E, Penninx BWJH, Zitman FG (2010): Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). Psychiatry Res 179:101–106. http://dx.doi.org/10.1016/j.psychres.2009.03.005.

Watson D, Clark LA (1991): The Mood and Anxiety Symptom Questionnaire.

Williams LM (2016): Precision psychiatry: a neural circuit taxonomy for depression and anxiety. Lancet Psychiatry 3:472–480. http://dx.doi.org/10.1016/S2215-0366(15)00579-9.

Wise T, Marwood L, Perkins AM, Herane-Vives A, Joules R, Lythgoe DJ, Luh W-M, Williams SCR, Young AH, Cleare AJ, Arnone D (2017a): Instability of default mode network connectivity in major depression: a two-sample confirmation study. Transl Psychiatry 7:e1105. http://dx.doi.org/10.1038/tp.2017.40.

Wise T, Radua J, Via E, Cardoner N, Abe O, Adams TM, Amico F, Cheng Y, Cole JH, de Azevedo Marques Périco C, Dickstein DP, Farrow TFD, Frodl T, Wagner G, Gotlib IH, Gruber O, Ham BJ, Job DE, Kempton MJ, Kim MJ, Koolschijn PCMP, Malhi GS, Mataix-Cols D, McIntosh AM, Nugent AC, O'Brien JT, Pezzoli S, Phillips ML, Sachdev PS, Salvadore G, Selvaraj S, Stanfield AC, Thomas AJ, van Tol MJ, van der Wee NJA, Veltman DJ, Young AH, Fu CH, Cleare AJ, Arnone D (2017b): Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. Mol Psychiatry 22:1455–1463. http://dx.doi.org/10.1038/mp.2016.72.

Wolf RC, Nolte HM, Hirjak D, Hofer S, Seidl U, Depping MS, Stieltjes B, Maier-Hein K, Sambataro F, Thomann PA (2016): Structural network changes in patients with major depression and schizophrenia treated with electroconvulsive therapy. Eur Neuropsychopharmacol 26:1465–1474. http://dx.doi.org/10.1016/j.euroneuro.2016.06.008.

Xu J, Van Dam NT, 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. Neurosci Biobehav Rev 96:21–30.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

http://dx.doi.org/10.1016/j.neubiorev.2018.11.005.

Yan C-G, Chen X, Li L, Castellanos FX, Bai T-J, Bo Q-J, Cao J, Chen G-M, Chen N-X, Chen W, Cheng C, Cheng Y-Q, Cui X-L, Duan J, Fang Y-R, Gong Q-Y, Guo W-B, Hou Z-H, Hu L, Kuang L, Li F, Li K-M, Li T, Liu Y-S, Liu Z-N, Long Y-C, Luo Q-H, Meng H-Q, Peng D-H, Qiu H-T, Qiu J, Shen Y-D, Shi Y-S, Wang C-Y, Wang F, Wang K, Wang L, Wang X, Wang Y, Wu X-P, Wu X-R, Xie C-M, Xie G-R, Xie H-Y, Xie P, Xu X-F, Yang H, Yang J, Yao J-S, Yao S-Q, Yin Y-Y, Yuan Y-G, Zhang A-X, Zhang H, Zhang K-R, Zhang L, Zhang Z-J, Zhou R-B, Zhou Y-T, Zhu J-J, Zou C-J, Si T-M, Zuo X-N, Zhao J-P, Zang Y-F (2019): Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. Proc Natl Acad Sci U S A 116:9078–9083. http://dx.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). Perspect Psychol Sci 4:294–298. http://dx.doi.org/10.1111/j.1745-6924.2009.01127.x.

Yu M, Linn KA, Shinohara RT, Oathes DJ, Cook PA, Duprat R, Moore TM, Oquendo MA, Phillips ML, McInnis M, Fava M, Trivedi MH, McGrath P, Parsey R, Weissman MM, Sheline YI (2019): Childhood trauma history is linked to abnormal brain connectivity in major depression. Proc Natl Acad Sci U S A 116:8582–8590. http://dx.doi.org/10.1073/pnas.1900801116.

Zhu H, Qiu C, Meng Y, Yuan M, Zhang Y, Ren Z, Li Y, Huang X, Gong Q, Lui S, Zhang W (2017): Altered Topological Properties of Brain Networks in Social Anxiety Disorder: A Resting-state Functional MRI Study. Sci Rep 7:43089. http://dx.doi.org/10.1038/srep43089.

7. Tables

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

Table 1. Demographics for samples. SD = Standard Deviation. 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.

	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

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

				Thoughts that you would be better off dead	20513
RDS-4	✓		4-16	Frequency of depressed mood in last 2 weeks Frequency of unenthusiasm / disinterest in last 2 weeks Frequency of tenseness / restlessness in last 2 weeks Frequency of tiredness / lethargy in last 2 weeks	2050 2060 2070 2080
GAD-7		✓	0-21	Feeling nervous, anxious or on edge Not being able to stop or control worrying Worrying too much about different things Trouble relaxing Being so restless that it is hard to sit still Becoming easily annoyed or irritable Feeling afraid as if something awful might happen	20506 20509 20520 20515 20516 20505 20512
N-12	✓		0-12	Mood swings Miserableness Irritability Sensitivity / hurt feelings Fed-up feelings Nervous feelings Worrier / anxious feelings Tense / 'highly strung' Worry too long after embarrassment Suffer from 'nerves' Loneliness, isolation Guilty feelings	1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030
Probable depression status	✓		0/1	Ever depressed Ever unenthusiastic/disinterested Duration of longest period of depression Duration of longest period of unenthusiasm/disinterest Seen Doctor (GP) for nerves, anxiety, tension, depression Seen psychiatrist for nerves, anxiety, tension, depression	4598 4631 4609 5375 2090 2100

Table 2: Measures of affect-based mental health available in the UK Biobank. PHQ-9 = Patient Health Questionnaire-9, RDS-4 = Recent Depressive Symptoms-4, GAD-7 = General Anxiety Disorder-7, N-12 = Neuroticism-12.

	# IDPs	UKB ID	Description
Resting state	21 55 210 210 1,485 1,485	25754 25755 25750 25752 25751 25753	rfMRI network amplitudes from 21 signal components rfMRI network amplitudes from 55 signal components Pairwise full correlation edges between 21 components Pairwise partial correlation edges between 21 components Pairwise full correlation edges between 55 components Pairwise partial correlation edges between 55 components
Total 3,466			
Structural	139 14	1101 1102	FAST gray matter volumes FIRST gray matter volumes

medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	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
Total	27	107	Weighted-mean MD
346	14	109	Median T2-star from susceptibility weighted imaging
Task	16	106	Task fMRI median + 90th percentile of BOLD effect and z

Table 3: Full set of IDPs considered for canonical correlation analysis. IDP = Imaging Derived Phenotype, UKB = UK Biobank.

	Test-retest reliability (ICC)	Correlation with RDS-4 (ρ)
RDS-4	0.88	-
CES-D	0.91	0.89
PHQ-9	0.94	0.91
MASQ general distress	0.87	0.78
MASQ anhedonic depression	0.82	0.67
MASQ anxious arousal	0.92	0.71

Table 4: Comparison of RDS-4 to other depression scales from MTurk study.

8. Figure Legends

Figure 1: UK Biobank subject inclusion chart.

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 paper.

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* $\chi \geq 0.19, p \leq 10^{-48}$*).*

Figure 4. A-D are the distributions of scores for participant responses to each questionnaire. 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.

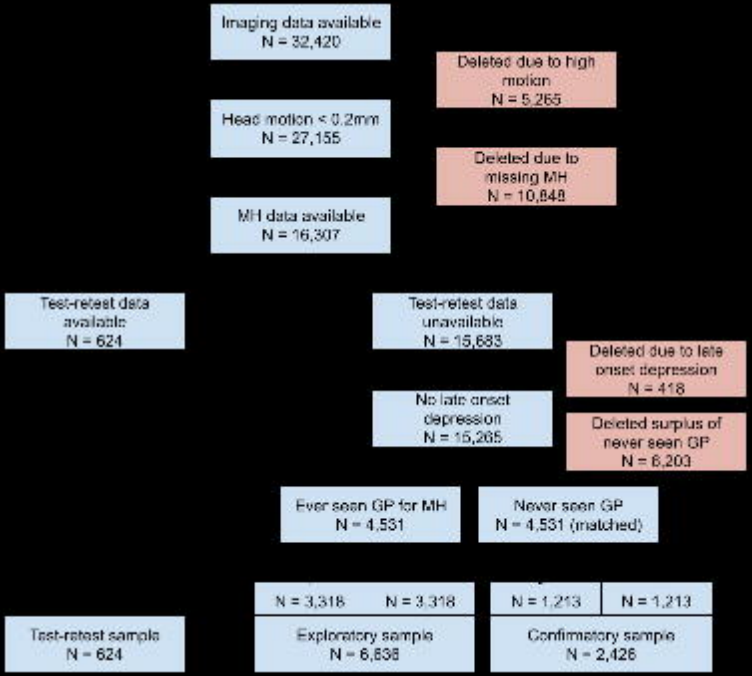
medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

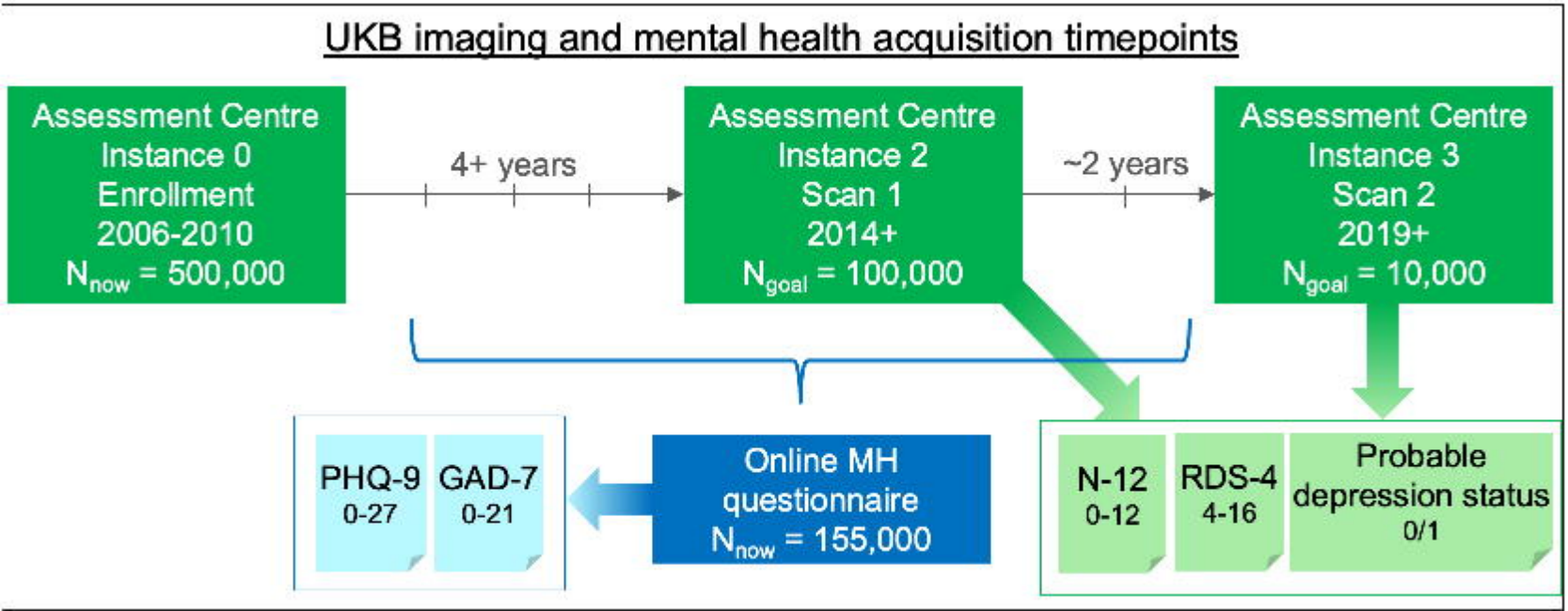
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 5. Canonical correlation results. A: post-hoc correlations for non-resting (structural & task) IDPs, showing only significant IDPs after Bonferroni correction. A similar figure for the resting state IDPs is included in the supplementary material (Fig. S2). B (insert): post-hoc CCA relations for mental health show that the first canonical covariate is broadly linked to affect-based mental health.

Figure 6. Effect sizes are shown for the grouped brain variables of structural (Area, Volume, Cortical Thickness, Fractional Anisotropy and T2) and functional (Task Activity, Amplitude, Full Network connectivity matrix and Partial Network Connectivity Matrix) modalities. Blue boxes indicate the middle 50% of the data (i.e., the range between the first and third quartile) and small black squares and blue lines inside each box represent the mean and median values, respectively. Outliers for each grouped brain IDP are shown as blue circles, which are above the 1.5 times of inter-quartile range (IQR), indicated by the whiskers extending from the boxes. For detailed assessments of effect sizes in specific IDPs see supplementary figures S3-S7.*

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) vs. 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 Supplemental Figure S8.

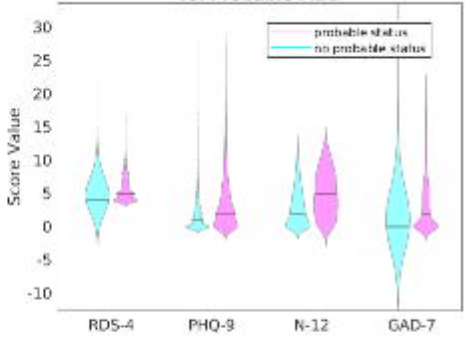

Imaging data available
N = 32,420
Deleted due to high motion
N = 5,265
Head motion < 0.2mm
N = 27,155
Deleted due to missing MH
N = 10,848
MH data available
N = 16,307
Test-retest data available
N = 624
Test-retest data unavailable
N = 15,683
Deleted due to late onset depression
N = 418
No late onset depression
N = 15,265
Deleted surplus of never seen GP
N = 6,203
Ever seen GP for MH
N = 4,531
Never seen GP
N = 4,531 (matched)
N = 3,318
N = 3,318
N = 1,213
N = 1,213
Test-retest sample
N = 624
Exploratory sample
N = 6,636
Confirmatory sample
N = 2,426


UKB imaging and mental health acquisition timepoints
Assessment Centre
Instance 0
Enrollment
2006-2010
N now = 500,000
4+ years
Assessment Centre
Instance 2
Scan 1
2014+
N goal = 100,000
~2 years
Assessment Centre
Instance 3
Scan 2
2019+
N goal = 10,000
PHQ-9
0-27
GAD-7
0-21
Online MH
questionnaire
N now = 155,000
N-12
0-12
RDS-4
4-16
Probable
depression status
0/1

(A) **Cross-Measure Correlations**

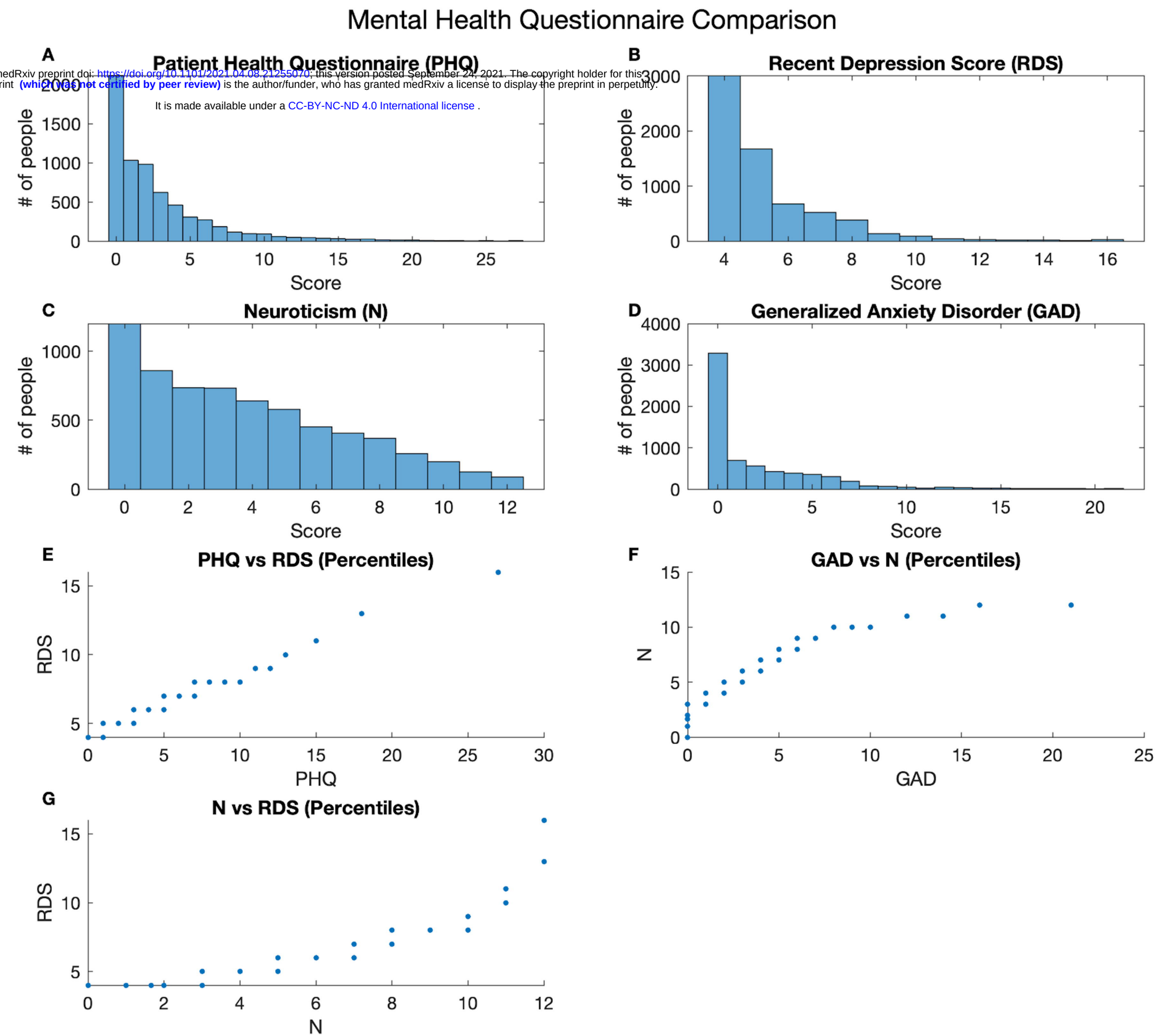
	RDS-4	PHQ-9	N-12	GAD-7
RDS-4	1	0.60	0.57	0.51
PHQ-9	0.60	1	0.47	0.69
N-12	0.57	0.47	1	0.51
GAD-7	0.51	0.69	0.51	1

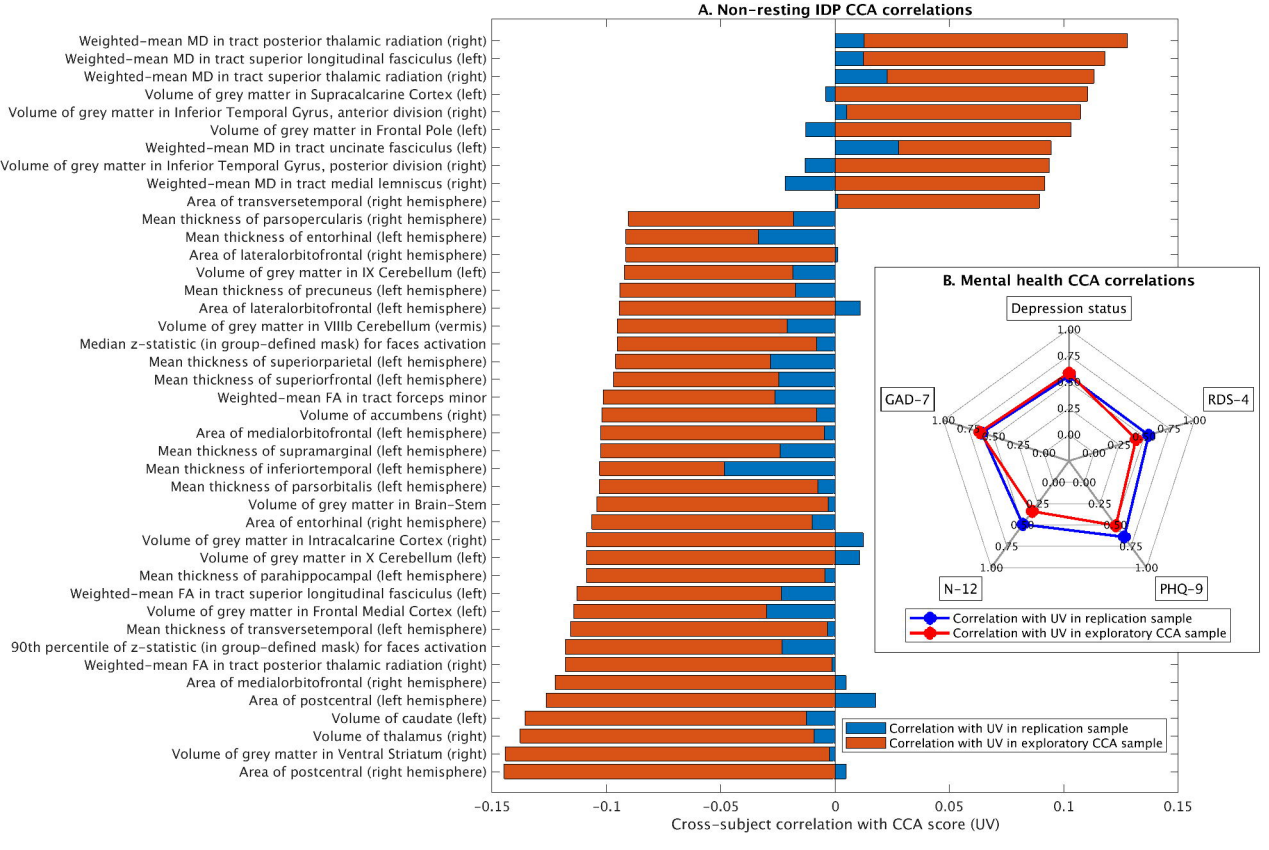
(B) **Depression Score Distributions vs. Probable MDD**

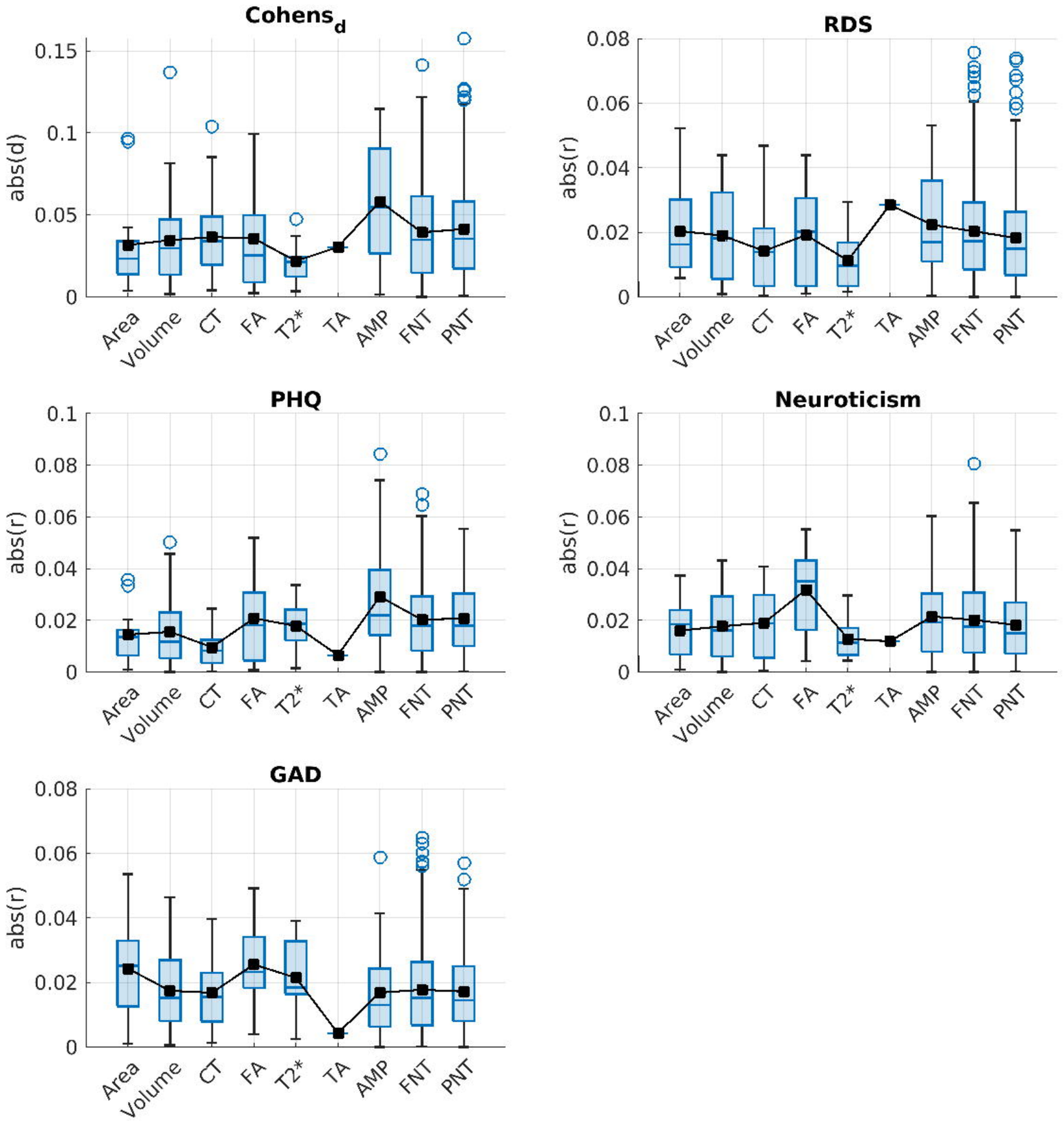


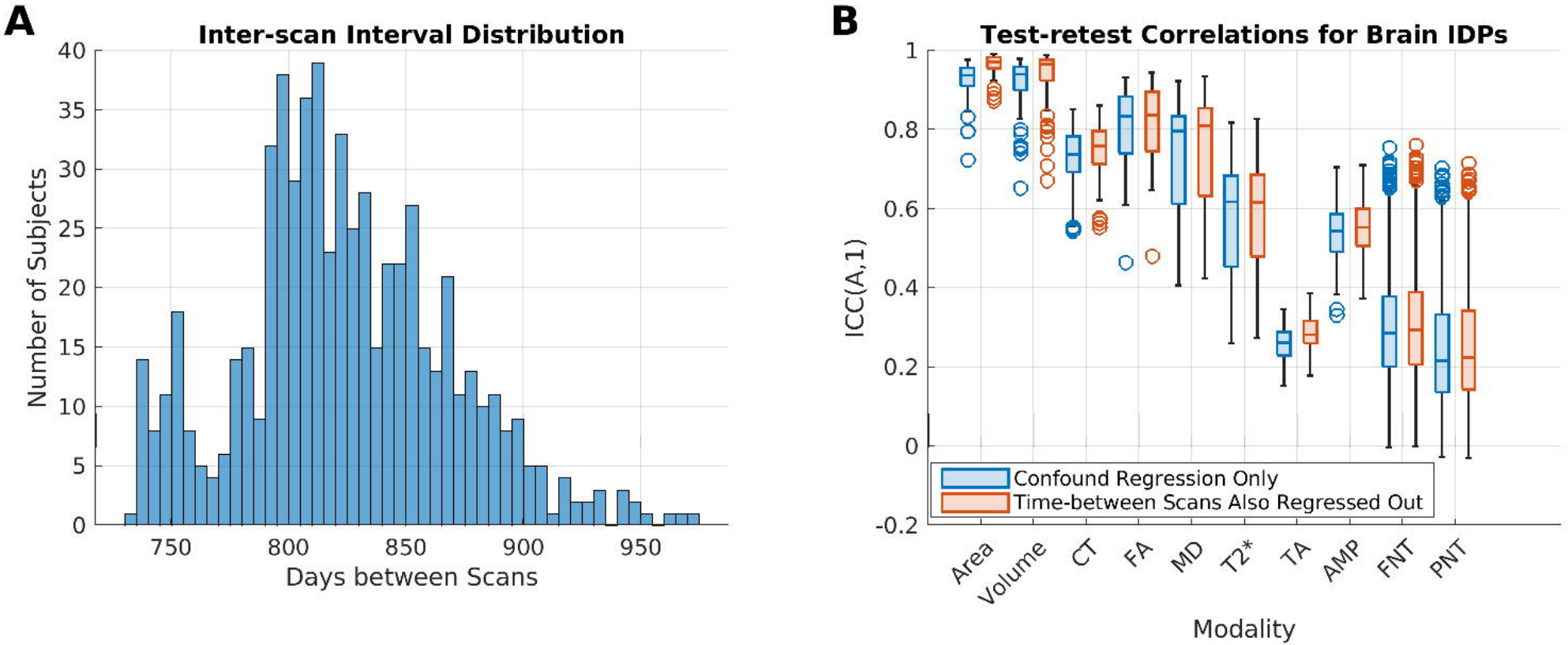
medRxiv preprint doi: https://doi.org/10.1101/2021.04.08.21255070; this version posted September 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .




A. Non-resting IDP CCA correlations
Weighted-mean MD in tract posterior thalamic radiation (right)
Weighted-mean MD in tract superior longitudinal fasciculus (left)
Weighted-mean MD in tract superior thalamic radiation (right)
Volume of grey matter in Supracalcarine Cortex (left)
Volume of grey matter in Inferior Temporal Gyrus, anterior division (right)
Volume of grey matter in Frontal Pole (left)
Weighted-mean MD in tract uncinate fasciculus (left)
Volume of grey matter in Inferior Temporal Gyrus, posterior division (right)
Weighted-mean MD in tract medial lemniscus (right)
Area of transversetemporal (right hemisphere)
Mean thickness of parsopercularis (right hemisphere)
Mean thickness of entorhinal (left hemisphere)
Area of lateralorbitofrontal (right hemisphere)
Volume of grey matter in IX Cerebellum (left)
Mean thickness of precuneus (left hemisphere)
Area of lateralorbitofrontal (left hemisphere)
Volume of grey matter in VIIIb Cerebellum (vermis)
Median z-statistic (in group-defined mask) for faces activation
Mean thickness of superiorparietal (left hemisphere)
Mean thickness of superiorfrontal (left hemisphere)
Weighted-mean FA in tract forceps minor
Volume of accumbens (right)
Area of medialorbitofrontal (left hemisphere)
Mean thickness of supramarginal (left hemisphere)
Mean thickness of inferiortemporal (left hemisphere)
Mean thickness of parsorbitalis (left hemisphere)
Volume of grey matter in Brain-Stem
Area of entorhinal (right hemisphere)
Volume of grey matter in Intracalcarine Cortex (right)
Volume of grey matter in X Cerebellum (left)
Mean thickness of parahippocampal (left hemisphere)
Weighted-mean FA in tract superior longitudinal fasciculus (left)
Volume of grey matter in Frontal Medial Cortex (left)
Mean thickness of transversetemporal (left hemisphere)
90th percentile of z-statistic (in group-defined mask) for faces activation
Weighted-mean FA in tract posterior thalamic radiation (right)
Area of medialorbitofrontal (right hemisphere)
Area of postcentral (left hemisphere)
Volume of caudate (left)
Volume of thalamus (right)
Volume of grey matter in Ventral Striatum (right)
Area of postcentral (right hemisphere)
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
Cross-subject correlation with CCA score (UV)
Correlation with UV in replication sample
Correlation with UV in exploratory CCA sample
B. Mental health CCA correlations
Depression status
RDS-4
PHQ-9
N-12
GAD-7
1.00
0.75
0.50
0.25
0.00
Correlation with UV in replication sample
Correlation with UV in exploratory CCA sample


Cohens$_d$
abs(d)
0.15
0.1
0.05
0
RDS
abs(r)
0.08
0.06
0.04
0.02
0
PHQ
abs(r)
0.1
0.08
0.06
0.04
0.02
0
Neuroticism
abs(r)
0.1
0.08
0.06
0.04
0.02
0
GAD
abs(r)
0.08
0.06
0.04
0.02
0
Area
Volume
CT
FA
T2*
TA
AMP
FNT
PNT


A
Inter-scan Interval Distribution
Number of Subjects
Days between Scans
0
5
10
15
20
25
30
35
40
750
800
850
900
950
B
Test-retest Correlations for Brain IDPs
ICC(A,1)
1
0.8
0.6
0.4
0.2
0
-0.2
Confound Regression Only
Time-between Scans Also Regressed Out
Area
Volume
CT
FA
MD
T2*
TA
AMP
FNT
PNT
Modality